Biosafety laboratory procedures

K. Nirmal Babu and B. Chempakam

Over the past century, chemistry has made great contributions towards our understanding of the physical and biological world. The chemical laboratory has become the center for acquiring knowledge and developing new materials for future use, as well as for monitoring and controlling those chemicals currently used routinely in thousands of commercial processes. Many of these chemicals are beneficial, but others have the potential to cause damage to human health and the environment.

Since the age of alchemy, some chemicals have demonstrated dramatic and dangerous properties, which have required the development of special techniques for safe handling. We also now know that many more are insidious poisons. Until recently, the chemical hazards in many laboratories were not accepted and taken into account by those working in them, and, accordingly, the necessity of putting "safety first" was not fully appreciated. In sharp contrast, a growing recognition of moral responsibility and mounting public pressure have made institutions housing chemical laboratories accountable for providing safe working environments for those employed in them and complying with extensive regulation of the transport of chemicals to the laboratories and removal of waste from them. The "old days" of easygoing attitudes toward laboratory safety and down-the-sink disposal are over! Laboratories have become safe places to work.

A new culture of safety consciousness, accountability, organization, and education has developed in the laboratories of the chemical industry, government, and academic institutions. Now, programs are being implemented to train laboratory personnel and to monitor the handling of chemicals from the moment they are ordered until their departure for ultimate treatment or disposal. Learning to participate in this culture of habitual risk assessment, experiment planning, and consideration of worst-case possibilities for oneself and one's fellow workers is as much a part of scientific education as learning the theoretical background of experiments or the step-by-step protocols for doing them in a professional and craftsman like manner.

Forming the foundation for a lifelong attitude of safety consciousness, risk assessment, and prudent laboratory practices should be an integral part of every stage of scientific education--in the classroom, in textbooks, and in the laboratory, from the earliest exposures in primary or secondary school through graduate and postdoctoral training. Teaching and academic institutions have this essential and unique responsibility.

Advanced training in safety is an important component of education through research. Safety training must be a continuing process; it should become an integral part of the daily activities of laboratory workers and those who are accountable for them. As a student or laboratory worker learns a new protocol, safe practices relevant to it should also be emphasized in the normal setting of the laboratory, with the careful guidance of a mentor and the shared responsibility of colleagues.

There are laboratory safety guidelines that are the foundation for biological safety in any laboratory using infectious agents, recombinant DNA, or biological toxins. Additions to the basic practices are required to enhance containment of particular agents or toxins that pose greater risks. Individual research plans that are submitted to the Institutional Biosafety Committee are reviewed to determine whether the safety measures that are proposed are appropriate to the risks inherent in the research procedures.

All the biotechnological laboratories must have a biosafety committee approved by the Department of Biotechnology, Government of India. Many of the standard laboratories have their biosafety practices.

General guidelines while working with chemicals

Four fundamental principles underlie all of the work practices:

1. **Plan ahead.** Determine the potential hazards associated with an experiment before beginning it.

2. Careful planning needs to precede any experiment involving a highly toxic substance whenever the substance is to be used for the first time or whenever an experienced user carries out a new protocol that increases substantially the risk of exposure.

3. **Minimize exposure to chemicals.** Do not allow laboratory chemicals to come in contact with skin. Use laboratory hoods and other ventilation devices to prevent exposure to airborne substances whenever possible.

4. **Do not underestimate risks.** Assume that any mixture of chemicals will be more toxic than its most toxic component. Treat all new compounds and substances of unknown toxicity as toxic substances.

5. **Be prepared for accidents.** Before beginning an experiment, know what specific action to take in the event of the accidental release of any hazardous substance. Know the location of all safety equipment and the nearest fire alarm and telephone.

General procedures for working with hazardous chemicals

1. Personal behavior

Professional standards of personal behavior are required in any laboratory.

- Avoid distracting or startling other workers.
- Do not allow practical jokes and horseplay at any time.
- Use laboratory equipment only for its designated purpose.
- Do not allow visitors, including children and pets, in laboratories where hazardous substances are stored or are in use or hazardous activities are in progress.

2. Minimizing Exposure to Chemicals

Precautions should be taken to avoid exposure by the principal routes, that is, contact with eyes and skin, inhalation, and ingestion.

a) Avoiding Eye Injury

Eye protection should be required for all personnel and visitors in all locations where chemicals are stored or used. Researchers should assess the risks associated with an experiment and use the appropriate level of eye protection:

- Safety glasses with side shields provide the minimum protection acceptable for regular use.

- Safety splash goggles or face shields should be worn when carrying out operations in which there is any danger from splashing chemicals or flying particles.

b) Minimizing Skin Contact

Wear gloves whenever handling hazardous chemicals, sharp-edged objects, very hot or very cold materials, toxic chemicals, and substances of unknown toxicity.

c) Avoiding Ingestion of Hazardous Chemicals

Eating, drinking, smoking, gum chewing, applying cosmetics, and taking medicine in laboratories where hazardous chemicals are used should be strictly prohibited. Laboratory water sources and deionized laboratory water should not be used for drinking water.

- Laboratory chemicals should never be tasted.
- A pipette bulb or aspirator should be used to pipette chemicals or to start a siphon; pipetting should never be done by mouth.
- Hands should be washed with soap and water immediately after working with any laboratory chemicals, even if gloves have been worn.

d) Avoiding Inhalation of Hazardous Chemicals

- Toxic chemicals or compounds of unknown toxicity should never be smelled.
- Procedures involving volatile toxic substances and operations involving solid or liquid toxic substances that may result in the generation of aerosols should be conducted in a laboratory hood.
- Dusts should be recognized as potentially contaminated and hazardous.

3. Clothing and Protective Apparel

It is advisable to wear a laboratory coat when working with hazardous chemicals. This is particularly important if personal clothing leaves skin exposed. Apparel giving additional protection (e.g., nonpermeable laboratory aprons) is required for work with certain hazardous substances.

4. Housekeeping-Rules

- ➢ Never obstruct access to exits and emergency equipment such as fire extinguishers and safety showers.
- ➢ Clean work areas (including floors) regularly. Properly label and store all chemicals.

6. Disposal of Chemicals

The overriding principle governing the handling of waste in prudent laboratory practice is that *no activity should begin unless a plan for the disposal of nonhazardous and hazardous waste has been formulated.* Each category of waste has certain appropriate disposal methods. In choosing among these methods, several general principles apply, but local considerations can strongly influence the application of these rules:

7. Handling Flammable Substances

Flammable substances present one of the most widespread hazards encountered in the laboratory. Because flammable materials are employed in so many common laboratory operations, basic prudent laboratory practice should always assume the presence of fire hazard unless a review of the materials and operations in the laboratory verifies the absence of significant hazard.

8. Working with biohazardous and radioactive materials

8.1 Biohazardous Materials

For even the most experienced laboratory worker, a careful review of the publication *Biosafety in Microbiological and Biomedical Laboratorie*) should be a prerequisite for beginning any laboratory activity involving a microorganism. It defines four levels of control that are appropriate for safe laboratory work with microorganisms that present occupational risks ranging from no risk of disease for normal healthy individuals to high individual risk of life-threatening disease, and it recommends guidelines for handling specific agents. The four levels of control, referred to as biosafety levels 1 through 4, describe microbiological practices, safety equipment, and features of laboratory facilities for the corresponding level of risk associated with handling a particular agent. The selection of a biosafety level is influenced by several characteristics of the infectious agent, the most important of which are the severity of the disease, the documented mode of transmission of the infectious agent, the availability of protective immunization or effective therapy, and the relative risk of exposure created by manipulations used in handling the agent.

i) Biosafety Level 1 (BSL-1)

Biosafety level 1 is the basic level of protection appropriate only for agents that are not known to cause disease in normal, healthy humans and of minimal potential hazard to laboratory personnel and the environment. The laboratory is not necessarily separated from the general traffic patterns in the building. Work is generally conducted on open bench tops using standard microbiological practices. Special containment equipment or facility design is neither required nor generally used. Laboratory personnel have specific training in the procedures conducted in the laboratory and are supervised by a scientist with general training in microbiology or a related science.

ii) Biosafety Level 2 (BSL-2)

Biosafety Level 2 is similar to Biosafety Level 1 and is suitable for work involving agents of moderate potential hazard to personnel and the environment. It differs from BSL-1 in that (1) laboratory personnel have specific training in handling pathogenic agents and are directed by competent scientists; (2) access to the laboratory is limited when work is being conducted; (3) extreme precautions are taken with contaminated sharp items; and (4) certain procedures in which infectious aerosols or splashes may be created are conducted in biological safety cabinets or other physical containment equipment.

Biosafety level 2 is appropriate for handling a broad spectrum of moderate-risk agents that cause human disease by ingestion or through percutaneous or mucous membrane exposure. Hepatitis B virus, human immunodeficiency virus (HIV), and salmonellae and toxoplasma spp. are representative of agents assigned to this biosafety level. Extreme precaution with needles or sharp instruments is emphasized at this level. A higher level of control may be indicated when some of these agents, especially HIV, are grown and concentrated.

iii) Biosafety Level 3 (BSL-3)

Biosafety Level 3 is applicable to clinical, diagnostic, teaching, research, or production facilities in which work is done with indigenous or exotic agents which may cause serious or potentially lethal disease as a result of exposure by the inhalation route. Laboratory personnel have specific training in handling pathogenic and potentially lethal agents, and are supervised by competent scientists who are experienced in working with these agents.

Biosafety level 3 is appropriate for agents with a potential for respiratory transmission and for agents that may cause serious and potentially lethal infections. Emphasis is placed on the control of aerosols by containing all manipulations. At this level, the facility is designed to control access to the laboratory and includes a specialized ventilation system, such as a biological safety cabinet, that minimizes the release of infectious aerosols from the laboratory. The bacterium *Mycobacterium tuberculosis* is an example of an agent for which this higher level of control is appropriate.

All procedures involving the manipulation of infectious materials are conducted within biological safety cabinets or other physical containment devices, or by personnel wearing appropriate personal protective clothing and equipment. The laboratory has special engineering and design features.

It is recognized, however, that some existing facilities may not have all the facility features recommended for Biosafety Level 3 (i.e., double-door access zone and sealed penetrations). In this circumstance, an acceptable level of safety for the conduct of routine procedures, (e.g., diagnostic procedures involving the propagation of an agent for identification, typing, susceptibility testing, etc.), may be achieved in a Biosafety Level 2 facility, providing 1) the exhaust air from the laboratory room is discharged to the outdoors, 2) the ventilation to the laboratory is balanced to provide directional airflow into the room, 3) access to the laboratory is restricted when work is in progress, and 4) the recommended Standard Microbiological Practices, Special Practices, and Safety Equipment for Biosafety Level 3 are rigorously followed. The decision to implement this modification of Biosafety Level 3 recommendations should be made only by the laboratory director.

iv) Biosafety Level 4 (BSL-4)

Biosafety Level 4 is required for work with dangerous and exotic agents that pose a high individual risk of aerosol-transmitted laboratory infections and life-threatening disease. Agents with a close or identical antigenic relationship to Biosafety Level 4 agents are handled at this level until sufficient data are obtained either to confirm continued work at this level, or to work with them at a lower level. Members of the laboratory staff have specific and thorough training in handling extremely hazardous infectious agents and they understand the primary and secondary containment functions of the standard and special practices, the containment equipment, and the laboratory design characteristics. They are supervised by competent scientists who are trained and experienced in working with these agents. Access to the laboratory is strictly controlled by the laboratory director. The facility is either in a separate building or in a controlled area within a building, which is completely isolated from all other areas of the building. A specific facility operations manual is prepared or adopted.

Within work areas of the facility, all activities are confined to Class III biological safety cabinets, or Class II biological safety cabinets used with one-piece positive pressure personnel suits ventilated by a life support system. The Biosafety Level 4 laboratory has special engineering and design features to prevent microorganisms from being disseminated into the environment.

There are two models for Biosafety Level 4 laboratories: (A) the Cabinet Laboratory where all handling of the agent is performed in a Class III Biological Safety Cabinet, and (B) the Suit Laboratory where personnel wear a protective suit. Biosafety Level-4 laboratories may be based on either model or a combination of both models in the same facility. If a combination is used, each type must me*et al*l the requirements identified for that type.

Exotic agents that pose a high individual risk of life-threatening disease by the aerosol route and for which no treatment is available are restricted to high containment laboratories that meet biosafety level 4 standards. Worker protection in these laboratories is provided by the use of physically sealed glove boxes or fully enclosed barrier suits that supply breathing air.

8.2 Radioactive materials

Prudent practices for working with radioactive materials are similar to those needed to reduce the risk of exposure to toxic chemicals and to biohazards:

1. Know the characteristics of the radioisotopes that are being used, including half-life, types and energies of emitted radiations, the potential for exposure, how to detect contamination, and the annual limit on intake.
2. Protect against exposure to airborne and ingestible radioactive materials.
3. Never eat, drink, smoke, handle contact lenses, apply cosmetics, or take or apply medicine in the laboratory, and keep food, drinks, cosmetics, and tobacco products out of the laboratory entirely so that they cannot become contaminated.
4. Do not pipet by mouth.
5. Provide for safe disposal of waste radionuclides and their solutions.
6. Use protective equipment to minimize exposures.
7. Use equipment that can be manipulated remotely, as well as shielding, glove boxes, and personal protective equipment, including gloves, clothing, and respirators, as appropriate.
8. Plan experiments so as to minimize exposure by reducing the time of exposure, using shielding against exposure, increasing your distance from the radiation, and paying attention to monitoring and decontamination.
9. Keep an accurate inventory of radioisotopes.
10. Record all receipts, transfers, and disposals of radioisotopes.
11. Record surveys.
12. Check workers and the work area each day that radioisotopes are used.
13. Minimize radioactive waste.
14. Plan procedures to use the smallest amount of radioisotope possible.
15. Ceck waste materials for contamination before discarding.
16. Place only materials with known or suspected radioactive contamination in appropriate radioactive waste containers.
17. Do not generate multihazardous waste (combinations of radioactive, biological, and chemical waste) without first consulting with the designated radiation and chemical safety officers.

8.3 Liquefied gases and cryogenic liquids

Cryogenic liquids are materials with boiling points of less than 73° C (100° F). Liquid nitrogen, helium, and argon, and slush mixtures of dry ice with isopropanol are the materials most commonly used in cold traps to condense volatile vapors from a system. In addition, oxygen, hydrogen, and helium are often used in the liquid state.

The primary hazards of cryogenic liquids are fire or explosion, pressure buildup (either slowly or due to rapid conversion of the liquid to the gaseous state), embrittlement of structural materials, frostbite, and asphyxiation. The extreme cold of cryogenic liquids requires special care in their use. The vapor that boils off from a liquid can cause the same problems as the liquid itself.

A special risk to personnel is skin or eye contact with the cryogenic liquid. Because these liquids are prone to splash in use owing to the large volume expansion ratio when the liquid warms up, eye protection, preferably a face shield, should be worn when handling liquefied gases and other cryogenic fluids.

Extreme caution should be exercised in using liquid nitrogen as a coolant for a cold trap. If such a system is opened while the cooling bath is still in contact with the trap, oxygen may condense from the atmosphere. The oxygen could then combine with any organic material in the trap to create a highly explosive mixture. Thus, a system that is connected to a liquid nitrogen trap should not be opened to the atmosphere until the trap has been removed. Also, if the system is closed after even a brief exposure to the atmosphere, some oxygen (or argon) may have already condensed. Then, when the liquid nitrogen bath is removed or when it evaporates, the condensed gases will vaporize, producing a pressure buildup and the potential for explosion. The same explosion hazard can be created if liquid nitrogen is used to cool a flammable mixture that is exposed

8.4. Precautions while using microorganisms

Infection followed by disease will depend on the microorganism's ability to multiply in the host and on the host's ability to resist or control the infection. It has proved useful to categorise all microorganisms into 4 groups which define their pathogenicity to humans; only the first group are non-pathogens.

Hazard Group 1: Organisms that is most unlikely to cause human disease.

Hazard Group 2: Organisms capable of causing human disease and which may be a hazard to laboratory workers, but are unlikely to spread to the community. Laboratory exposure rarely produces infection and effective prophylaxis or effective treatment is usually available.

Hazard Group 3: Organisms that may cause severe human disease and present a serious hazard to laboratory workers. They may present a risk of spread to the community, but there is usually effective prophylaxis or treatment available.

Hazard Group 4: Organisms that cause severe human disease and are a serious hazard to laboratory workers. They may present a high risk of spread to the community, and there is usually no effective prophylaxis or treatment

The intention of this categorisation, which applies to non-modified organisms as well, is to identify appropriate containment, which would be required to protect those working with the organisms. The higher the hazard group, the greater the containment required to control the organism and ensure that it does not infect those working with it.

8.5 Examples of some biohazardous chemicals – precautions

i) Ethidium bromide

1. Ethidium bromide is a potent mutagen that has been used for many years as a nucleic acid stain. This material fluoresces a red-orange color under ultraviolet light and with increased fluorescence when bound to double-stranded DNA.

2. The powder form is considered an irtant to the upper respiratory tract, eyes, and skin. Ethidium bromide is strongly mutagenic, causing living cell mutations. Even though there is no evidence at this time of human carcinogenicity or teratogenicity, this material should be considered a possible carcinogen or teratogen.

3. Wear a lab coat, eye protection, and nitrile gloves when working with ethidium bromide. Leave lab coats, gloves, and other PPE in the lab, when your work is complete, to prevent the spread of this or other chemicals outside of the lab.

4. When an ultraviolet light source is used in your work with ethidium bromide, added caution is required. As a general rule, avoid exposing unprotected skin and eyes to intense UV sources. If the UV light is aimed upwards, wear a UV protective face shield when you are standing near the source. For prolonged work close to UV light boxes or other intense sources, it may be useful to wrap the end of the lab coat sleeves loosely with masking tape to prevent gaps where the wrist could be exposed.

5. When working with ethidium bromide, minimize the potential for spills. Perform all processes that generate ethidium bromide dusts or mists inside the fume hood to minimize inhalation exposures.

6. Spills of ethidium bromide solutions should be absorbed and decontaminated with soap and water. Avoid raising dust when cleaning up solid spills by mixing with water and then absorbing the solution. All spill cleanup materials and absorbents should be bagged or placed in a sealed container with a hazardous waste label.

ii) Decomposition of Ethidium bromide solutions

Method 1

1. Add sufficient water to reduce the concentration.
2. Add 0.2 vol. hypophosphorous acid and 0.12 vol. of fresh 0.5 M sodium nitrate.
3. Incubate for 24 hrs at R.T, add 1 M Sodium bicarbonate in excess.
4. Discard

Method 2

1. Add sufficient water to reduce concentration.
2. Add 1 vol. of 0.5 M $KMnO_4$. Add 1 vol. of 2.5 N HCl.
3. Mix carefully to stand at R.T for hours.
4. Add 1 vol. 2.5 N NaOH.
5. Mix and discard.

ii) Acryl amide

1. Warning! Harmful if swallowed, inhaled or absorbed through skin.
2. Affects central and peripheral nervous systems and reproductive system.
3. Causes irritation to skin, eyes and respiratory tract. Suspect cancer hazard.
4. May cause cancer. Risk of cancer depends on level and duration of exposure.
5. Possible birth defect hazard. May cause birth defects based on animal data.
6. Thermally unstable. May polymerize explosively if heated to the melting point.
7. Keep in a tightly closed container. Store in a cool, dry, ventilated area away from sources of heat or ignition. Protect against physical damage. Store separately from reactive or combustible materials, and out of direct sunlight. Isolate from oxidizing materials and peroxides. Store away from acids and alkalies
8. Wash hands, face, forearms and neck when exiting restricted areas. Shower, dispose of outer clothing, change to clean garments at the end of the day. Avoid cross-contamination of street clothes.
9. Containers of this material may be hazardous when empty since they retain product residues (dust, solids); observe all warnings and precautions listed for the product.

iii) Isopropanol

1. This compound contains no ingredients at concentrations of 0.1% or greater that is carcinogens or suspect carcinogens.
2. Signs/symptoms of overexp: nausea, narcosis, dermatitis, skin irritation, eye irritation, lung irritation, mucous membrane irritation, GI tract irritation.
3. Emergency/first aid proc: Eyes: flush with water for 15 minutes while holding eyelids open. Get medical attention. Skin: Remove contaminated clothing. Wash with soap and water. Other protective equipment: Chemical resistant lab coat is necessary to prevent skin contact.

4. Special hazard precautions: irritation to skin, eyes, lungs, mucous membranes and gi tract. Prolonged exposure may cause dermatitis.

References

CRC Handbook of Laboratory Safety, Third Edition. A. K. Furr, Ed. Chemical Rubber Company. 1990, 704p.

Prudent Practices for Handling Hazardous Chemicals in Laboratories. Prepared by the National Research Council. 1981, 291p

Prudent Practices in the Laboratory. National Research Council. 1995, 427p

Safe Storage and Handling of Laboratory Chemicals - A Review of Safe Storage and Handling Practices for Laboratory Chemicals. Nancy Magnussen. Texas A&M University Chemistry Safety Coordinator

Internet basics, search engines and search strategies

Santhosh J Eapen and V. Sathyanath

Introduction

Information technology is the biggest achievement of 20th century. The Internet, which is the finest flower of Information Technology (IT) has reduced the world to a global information society, where information is the lifeblood, technology is the tool and quality is the objective. In this society information is treated as a strategic global resource and is attributed with many terms.

The Internet is a tremendously useful information source for biological research. Data is now digitized and distributed from centralized databases. There are online and e-journals for scientific publications. Any research group has a web page giving details on their research activities, publications; software downloads to automated data-processing services.

The world changed by the arrival of Internet. It started quietly and has grown to involve much of the world. On an average day, the following events occur.

1. A group of company executives hold a meeting. One executive in New York, another in Florida, and a third is on vacation in Colorado. Each sits in front of a computer that has both a camera and a microphone attached. They see pictures of one another on the screen and hear each others' voice.
2. A college student in India uses a computer to see and hear a tour of the campus at the University of Hawaii.
3. A Professor puts his class schedule and class assignments in his website and students visits the site frequently and answers through e-mail.

The Internet probably has already affected you. Some statistics:

1. The Internet currently reaches millions of people in over 134 countries.
2. Scientists have been using Internet since 1980.
3. The U.S. military has been using the Internet technology for over 16 years; it played a role in military actions such as Operation Desert in the early 1990s.

What is Internet?

Internet is a network of networks. Internet is a rapidly growing global digital library built on a remarkably flexible communication technology. Internet is a platform where millions of people communicate electronically on a one-to one basis or in groups and share ideas of information. Internet is a vast collection of information that can be searched and retrieved electronically.

There are four basic building blocks to the Internet, Hosts, Routers and Clients and Connections. In most cases your computer falls under the "Client" category. Data is sent from your computer in the form of a "packet". It is similar to an envelope; it surrounds your data and contains both a return and destination address. Your computer handles the packets for you; it's all done in the background, without your knowledge.

How the Internet grew from its humble beginnings to the largest network in the world?

In the late 1960s and early 70s, new networking technologies LAN (Local Area Network) and WAN (Wide Area Network) emerged. Even if both are having its own advantages the disappointing news about these technologies is that both are incompatible with each other. So there was a necessity of a single network.

By the late 1960s, the U.S. Department of Defense became interested in using computer networks. Through the Advanced Research Project Agency (ARPA), the military funded research on networking. ARPA projects included a wide area network called the ARPANET.

A key idea in the ARPA research was a new approach to interconnecting LANs and WANs that became known as an internetwork, usually abbreviated as Internet.

Computer software forms an important part of technology that makes it possible to interconnect networks. Two pieces of Internet protocols standout as innovative- the Internet Protocol (IP) and Transmission Control Protocol (TCP). By 1982 a prototype Internet was in place and the TCP/IP technology was used widely .In the beginning of 1983 ARPA expanded the Internet to include all the military sites that connected to the ARPANET.

The group responsible for guiding the research and development of the Internet formed, known as IAB (Internet Activities Board). The primary subgroup responsible for technical matters is known as the Internet Engineering Task Force (IETF). Internet Research Task Force (IRTF) looks into long-term problems, which may become critical in future. By the mid 1980s the National Science Foundation (NSF) took a leadership role and formed NSFNET. ANSNET and vBNS were another backbones formed. By early 1997, the Internet was growing so fast that on an average, a new computer was added to the Internet every 4 seconds. Furthermore the rate continues to increase.

Underlying technologies and basic capabilities of Internet

a) Packet switching

The fundamental technique that computer networks use to ensure fair access to shared network resources is known as packet switching. Before data can be transferred across a network it must be divided into packets. A sender divides a message into packets and transfers the packet across a network. Each packet contains a header that specifies the computer to which the packet should be delivered; the destination is specified by a number known as computer’s address. A receiver reassembles the original message from the packets that arrive. Computers that share access to a network take turns sending packets. On each turn, a given computer sends one packet.

b) Routers

The Internet is not a conventional computer network. It consists of thousands computer networks interconnected by a dedicated special-purpose computers called routers. Because a router can interconnect networks that use different hardware technologies, the router architecture permits the Internet to contain multiple types of networks. The least expensive way to connect a new network to the Internet involves finding the closest Internet site and connecting a router between one of its networks and the new network.

Many people who can’t afford a direct network connection use the telephone system to access Internet services. To do so, their computer must have a dial-up modem, which plugs into the phone system. In addition their computer must have software to use the modem. When the software runs, it uses the modem to place a telephone call to a modem that connects to a computer that attaches to the Internet. The remote computer answers the call and provides access to the Internet services.

c) Internet Protocol (IP)

The Internet Protocol, IP, specifies the basic rules that a computer must follow to communicate across the Internet. IP defines the format of Internet packets, which are called IP datagrams. IP also defines an address scheme that assigns each computer a unique number used in all communications. IP software makes an interconnected set of networks and routers operate like a single, large network.

Each computer on the Internet must have IP software that allows it to create and send IP datagrams. Each router also has IP software that knows how to forward datagrams to their destination. When a datagram arrives at a router, the IP software chooses the path that will lead to the datagram’s destination.

d) Transmission Control Protocol (TCP)

To handle communication errors, a computer must also have TCP software. TCP eliminates duplicate data, ensures that data is resembled in exactly the order it was sent and resends data when a datagram is lost.

TCP uses acknowledgements and timeouts to handle the problem of data loss. The sender retransmits data unless an acknowledgement arrives before the timer expires. TCP's scheme of timeout works across the Internet because TCP changes the timeout automatically depending on whether the destination is close or far away from the source.

e) Client- Server networking

The Internet offers variety of services. Despite apparent differences among the services, all software on the computer uses the same general structure known as client-server computing. In a client –server environment each program must be classified as a client or as a server. A server program offers a service. Usually, computers offers services start automatically when the computer is powered on. The server remains running, ready to accept an incoming request anytime. Users usually run client software when they access a service. A client program contacts a server, sends a request, and displays the server's response.

f) IP address & Domain Name Systems (DNS)

The unique number assigned to a computer is called its Internet Address (IP address). Each computer, including router needs to be assigned an IP address before the computer can communicate on the Internet. Internally, a computer stores an IP address in four binary bits called bytes (e.g.: 202.41.105.25).

Because humans prefer using names instead of numbers, the Intern*et al*lows people to assign a name called Domain names to each computer, and provides an automated system (Domain Name System) that can translate a name into an equivalent IP address. Domain name consists of multiple alphabetic strings separated by periods (e.g. vishak.reccal.ernet.in). Although both the domain names and their IP addresses are written in same formats they are unrelated just as a person's name & his telephone number.

Internet Access

In the late 1980s and early 1990s, about the only option for Internet access was some kind of dedicated or leased line connection. However, individual dial-up accounts became available from a number of commercial Internet access providers as early as 1992. Usually access to Internet is available in two forms: Dial-up access and dedicated access.

Dialup access

In this case a user modem dials a host computer to access the Internet. This access is of two types-terminal dial-up and TCP/IP based. Terminal dial-up access doesn't permit graphics and thus restricts the range of available services. However it allows E-mail, FTP, and other non-graphic applications. Dial-up connections based on several line Internet Protocol (SLIP) and point-to-point protocol (PPP) can connect one's computer directly to Internet and one can run networking applications.

Dedicated access

Dial up users must wait for their modems to dial and connect while users with dedicated access have Internet connectivity all the time. In case of dedicated access in LAN, at least one computer on the LAN is configured with TCP/IP software the programme that instructs the computer how to communicate with the Internet, however dedicated access is not right for every one, besides being expensive it take time to set it up and requires specialized manpower for operations and maintenance.

World Wide Web (WWW)

WWW has become the most widely used information service on the Internet. WWW is a large set of hypermedia documents stored on computers throughout the Internet. A given WWW

Document can contain embedded menu items that refer to WWW documents stored on other computers, a user can follow a link from a document on one computer to a document on a another computer as easily as a user can follow a link between documents on the same computer.

A browser is a computer program used to access the WWW. To make it possible to record the location of a particular piece of information, each WWW document is assigned a unique name known as a Uniform Resource Locator (URL) that contain the information a browser needs to obtain the document.

Web pages are written in a computer language known as Hyper Text Markup Language (HTML). There are advanced web technologies like forms, frames, CGI programs and Java.

Website

Website is an addressed location, which consists of a web server and a set of web pages. It is a combination of textual, graphics, movie, clips and sound files. The resources in a website are usually located on one server, but they don't need to be in all cases. A Web page has links, so that user can move easily between pages in a website, or jump to another related site.

URL

Internet addresses are URLs (Uniform Resource Locations) and understanding them can help the user navigate the Internet better. Most URLs consist of four basic parts: 1. The protocol, 2. The server of domain name, 3. The directory path, and 4. The file name. Each part of a URL helps direct browser through the internet to the file, the user wants to retrieve.

URLs begin by designating the protocol to be used. The WWW protocol is hypertext transport protocol (HTTP), which enables hypertext transmissions across the network. Most Web clients can also connect to resources using other protocols such as gopher, FTP and Telnet.

Gopher

Gopher is menu-based programme that enables us to take for information without having to know where material is specifically located. It was developed at the University of Minnesota. In this programme the information is organized in such a hierarchy that the intermediate nodes are directories or indexes, and lead nodes are documents.

A gopher server can link to local text files, FTP sites, and telnet connections. USENET newsgroups, graphic files, sound files, and moving picture files. A client with compliant graphic viewers and sound players can make gopher multimedia-capable. Gopher clients are even more widely available than gopher servers. Different clients offer slightly different features, with some having very powerful capabilities. One of the most useful client features is the capability for establishing a list of "bookmark". The gopher bookmark greatly enables resource discovery by providing quick access to specific resources.

File Transfer Protocol (FTP)

FTP permits a user to transfer a copy of a data file across the Internet from one computer to another. When using FTP, a user establishes communication with a remote computer and obtains authorization by sending a login and password. The user can list the files on the remote computer, request a copy of a particular file, or send a copy of a local file.

On the Internet, most computers that offer files for access via FTP support anonymous FTP. Anonymous FTP allows any user on an arbitrary computer to access public files; the user does not need a valid login to retrieve a public file.

Remote Login (TELNET)

A remote login facility permits a user who is using one computer to interact with a program on another computer. The Internet's remote login service is called TELNET. To use this service, one must invoke a local application program and specify a remote machine. The local program becomes a client, which forms a connection to a server on the remote computer. The client passes keystrokes and mouse movements to the remote machine, and displays output from the remote machine on the user's display screen.

Browsers

Most internet services operate according to a scheme called client/server. A server on one computer starts a programme that contacts another remote computer system, the client is the programme the user is running, and server is running on a remote system. The user gives command to the client, which passes them on to the server. The server interprets those commands and returns information to client, which passes information to the user.

Browser allows the user to view Web sites. There are many types of browsers available today. Depending on the type of content a browser can handle, they are divided into two types.

- Graphic-based
- Text-based

Among graphical browsers, Netscape Communicator and Internet Explorer are the most popular. In the US some people use the America Online (AOL) browser while several others use Mosaic, HotJava and other browsers. Text-based browsers do not allow much to be seen on the web page other than the plain text. Hence they are not so popular, though some people use these browsers, as they are the cheapest. Among other text-based browsers, Lynx is by far the most popular.

Lynx

Lynx is one of the Web browsers, which can only display the textual data. The graphical or multimedia elements cannot be delivered through this shareware web browser. Lynx was developed at the University of Kansas for their own campus, then was released to the Internet community. Lynx uses full-screen capabilities, and it can easily move back and forth through hypertext links.

Mosaic

Mosaic is the first popularly used graphical Web browser designed and developed at the National Centre for Super Computing Application (NCSA). Mosaic provides a powerful user-friendly graphical interface where hyper textual links appear to connect other web pages, or highlighted words in the text. Colour is used to highlight text links and other elements.

Netscape Navigator

Netscape navigator is the most widely used graphical web browser and popularly known as Netscape. The salient features of Netscape are:

1. It combines text, pictures, graphics, and animation in lots of attractive ways.
2. It runs reasonably well even on inexpensive dialup connections.
3. Consecutive version are being updated to add new internet services and
4. Its authors give it away for free.

Netscape is the best organized, best performing and most efficient web tool. Netscape recognizes the advanced HTML tags. Netscape Navigator supports a host at widely used enhancements to HTML. A clear interface, easy-to-use controls and a straightforward set-up also add to this browser's appeal. The latest version is named as Netscape Communicator, which has the facilities of both editing the HTML document and browsing it.

Internet Explorer

The Microsoft goes long way towards its goal to integrate all Internet tools with Windows 95 under the folder name Internet Explorer for browsing the web. It is distributed with Windows 95. This web browser provides all the features needed to explore the Internet thoroughly.

Search Engines

These are softwares that help us to find our way around the billion or more pages that make up the Web. It is a service that indexes, organizes, and often rates and reviews web sites. There are significant differences in how search engines build their databases and rank sites. Some rely on people to maintain a catalog of web sites or pages. They are called General Topical Search Engines. Some use software called a spider or crawler to identify key information on sites across the Internet. They are known as Keyword Search Engines. Some others combine both types of services.

These differences make some search engines far more useful than others for searching science and technology websites. Key features to look at in a web search engines database building and indexing strategies are free URL submission, full-text indexing, automated, comprehensive web crawling, a fast ‘refresh’ rate and a sensible ranking strategy for results.

AltaVista: You can search for keywords or type your query in the form of a question. Other features include a facility to translate sites in several languages; photo and media finder, which searches for images; and family filter, which helps you to limit the types of results, produced by the search.

Lycos: It is a search engine with a directory, organized by subject.

Google: Google is one of the best search engines, which is extremely comprehensive. Pages are ranked based on how many times they are linked from other pages. Google catches copies of web pages, so pages can be accessible even if the server is offline. It returns only pages that contain all the relevant search terms.

HotBot: For a new user HotBot is probably the best search engine. It is also relatively comprehensive, regularly updated and offers form-based query tools.

Excite: A hybrid search engine. It has the traditional search capacity because of using a software program but also has a directory.

Yahoo: This is actually a directory rather than a search engine, meaning human compile and categorize the sites it searches.

If you are looking for a broad, common topic use a search engine that tends to give you fewer results of higher quality 9usually the type that relies on people to catalog sites) eg. Yahoo.
If you are looking for a rare topic a site that may be less discriminating but yields more results (usually the type compiled by a software program) eg. Excite, Google.

As scientists we should know how to structure a query to weed out the majority of the junk that will come up in a search. Each search engine makes different default assumption when we enter a few keywords into its little box. If your first search doesn’t produce the results you want, try searching with at least one more search engine. Many search engines have a link that leads to detailed information about how it compiles and searches through information, and how to get the best results from it. Reading this page can save you a lot of time.

Effective searching is done when most of the planning, analyzing and keyword refining is done away from the computer. In order to find information effectively, you should plan your search by creating word webs or concept maps or idea maps on the topic. Combine these concepts or ideas depending on the kind of search engine that you will be using. Include distinctive names, acronyms, and abbreviations within your strategies. Include synonyms and variant spellings within your strategies.

Topical Search
With this kind of search, you begin with a broad topic, then you narrow it down. Eg. Science-Biology-Molecular Biology. This search is used with topical search engines such as Yahoo.

Keyword Search (boolean searching)
In order to search effectively, you need to use Boolean logic, which is an extremely simple way of stating how a group of things should be divided or combined into sets. Boolean queries restrict the results that are returned from a database by joining a series of search terms with the operators AND, OR and NOT; and mathematical signs + or -. The word **and** and the sign **+** will connect your concepts together by requiring that both of these words appear in the results. The word **or** will allow either word to appear in the results. The words **and not** and the minus sign (**-**) will exclude words to appear in the results. The word **adj** means "adjacent" and will generate results showing where those two words are next to each other in any order. The word **near** will generate results showing where those two words are close to each other. Quotation marks will generate results showing those web pages that show the exact phrasing contained within the quotation marks. If you want to broaden your search or you are unsure of the spelling for a particular word, many search engines will allow by suffixing the word with signs like ***, ?** and **$.**

Joining two keywords with AND finds documents that contain only *keyword 1* and *keyword 2*; using OR finds documents that contain either *keyword 1* or *keyword 2* (or both); and using NOT finds documents that contain *keyword 1* but not *keyword 2*. However, search engines differ in how they interpret a space or an implied operator. Some consider a space an OR, while others defaults to AND. Therefore, find out how the search engine you are using works before you formulate your query.

Boolean queries are read from left to right, just like text. Parentheses can structure more complex boolean queries. If you look for documents that contain *keyword 1* and one of either *keyword 2* or *keyword 3*, but not *keyword 4*, the query should be like: (*keyword 1* AND (*keyword 2* OR *keyword 3*)) NOT *keyword 4*. By using quotation marks you can specify a phrase in many search engines.

Finding Scientific Articles
Print journals in the sciences are increasingly publishing their content in an electronic format in addition to hardcopy. Almost all journals have their own website, most of which are accessible only to subscribers. But abstracts and table of contents usually are free. E-journals have no print counterpart.

There are several scientific literature databases available on the Net, but most of them are not free. Medline database is an excellent resource for searching scientific literature in the biological sciences. It is a free server sponsored by the National Center for Biotechnology Information (NCBI) at the National Library of Medicine.

Technical reports of several research groups are available on their web sites. These are not peer reviewed or published elsewhere, and your only clue to their quality is the reputation and expertise of the authors.

PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi is one of the most valuable web resources available to biologists. It indexes over 4000 journals in cell and molecular biology, biochemistry, genetics and related fields. It uses a keyword-based search strategy and allows boolean operators in query statements. Users can specify which database fields to check for each search term by following the search term with a field name enclosed in square brackets. In addition, users can search PubMed using Medical Subjects Heading (MeSH) terms. MeSH is a library of standardized terms that may help manuscripts that use alternate terms to refer to the same concept. PubMed again allows you to use a web interface to narrow your search. By using the Limits form below the query box you can limit your search to particular fields in the PubMed database, such as the Author Name or Substance Name field or language, content and date. The Preview/Index menu allows you to build a detailed query interactively. You can also use the options in the History form to access results from earlier searches and to narrow a search by adding new terms to the query. If you want to collect results from multiple queries and save them into one big file, the Clipboard will help you. If you find a search strategy that works for you, you can save that strategy in the form of a URL, and repeat the same search at any time. The URL has to be bookmarked for this.

Public Biological Databases

Storage of macromolecular data like a gene, protein, an enzyme or a functional RNA in electronic databases is a complicated task. At present each entry into the database is given a serial number and then stored in a relational database that knows the proper linkages between that serial number and all information pertaining to that entry. DNA sequence, protein sequence and protein structure, though well related are currently maintained as separate database projects and in unconnected data formats, because biological sequence and structure databases have evolved in parallel.

The first molecular biology database was the Protein Data Bank (PDB), the central repository for X-ray crystal structures of protein molecules (at the Brookhaven National Laboratory in 1971). From 15 sets of co-ordinates in 1973, now it has about 14,000 entries. Journals that publish crystallographic results now require submission to the PDB as a condition of publication.

Sequence databases generally specialize in one type of sequence data: DNA, RNA or protein. There are major sequence data collections and deposition sites in Europe, Japan and the United States. The first DNA sequence database was the Gene Sequence Database (GSDB) at Los Alamos National Lab in 1979. Up-to-date gene sequence information is available from GSDB through the National Centre for Genome Resources. The European Molecular Biology Laboratory, the DNA Database of Japan and the National Institutes of Health co-operated to support the database Gene Bank. Various biological database and information sources available on the web are summarized in Table 1.

Table. Major Biological Data and Information Sources

Subject	Source	Link
Biomedical literature	PubMed	*http://www.ncbi.nlm.nih.gov/entrez/query.fcgi*
Nucleic acid sequence	GenBank SRS at EMBL/EBI	*http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=Nucleotide* *http://srs.ebi.ac.uk*
Genome sequence	Entrez Genome TIGR databases	*http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=Genome* *http://www.tigr.org/tdb/*
Protein sequence	GenBank SWISS-PROT at ExPASy PIR	*http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=Protein* *http://www.expasy.ch/spro/* *http://www-nbrf.georgetown.edu*
Protein structure	Protein Data Bank	*http://www.rcsb.org/pdb/*
Entrez Structure DB Protein and peptide mass spectroscopy	PROWL	*http://prowl.rockefeller.edu*
Post-translational modifications	RESID	*http://www-nbrf.georgetown.edu/pirwww/search/ textresid.Html*
Biochemical and bio-physical information	ENZYME BIND	*http://www.expasy.ch/enzyme/* *http://www.ncbi.nlm.nih.gov: 80/entrez/query.fcgi? db=Structure*

Biochemical pathways	PathDB KEGG WIT	*http://www.ncgr.org/software/pathdb/* *http://www.genome.ad.jp/kegg/* *http://wit.mcs.anl.gov/WIT2/*
Microarray	Gene Expression Links	*http://industry.ebi.ac.uk/~alan/MicroArray/*
2D-PAGE	SWISS-2DPAGE	*http://www.expasy.ch/ch2d/ch2d-top.html*
Web resources	The EBI Biocatalog IUBio Archive	*http://www.ebi.ac.uk/biocat/* *http://iubio.bio.indiana.edu*

Finding Software

Journals like Bioinformatics, Nucleic Acids Research, Journal of Molecular Biology and Protein Science cover new developments in the field of Bioinformatics. Proven software for a particular application can be obtained from reliable web resources that link to computational biology software sites. Most of the major biological databases have listed their sofrwares eg. PDB, TIGR, NCBI, etc. Many organizations and groups provide web implementations of their software. They are also available as web servers. You can use the web server versions to get you started and understand the inputs, outputs and options for the program. If you are good in programming, you can develop scripts that allow you to hit a web server with multiple requests without entering them manually into a form. By this you will be able to download a local copy of the software and run it on your own machine so that you avoid slow data transfer to and from remote sites.

Judging the Quality of Information

Some simple guidelines can help you to screen the information you find on web. Always approach the information offered on the web with healthy skepticism.

Authority: One of the first things to consider when evaluating a software, data or information formed on the Internet is the source. The authors, their reputation, their affiliation and credentials, expertise etc. are all important in judging a site. Many commercial companies offer a stripped-down version of their software or services to make you buy their more complete package. A site of an individual researcher may not always all inclusive as a publicly funded database does.

Transparency: You should have access to the source code (the human readable version of a computer program) for whatever the web & server is doing. Others may have online help pages and other documentation. Read and understand the method and its results you use it.

Timeliness: Many sites on the web are not out-of-data sites and so be skeptical. But timeliness is not always an issue with software. If they are no longer supported or code libraries have ceased to be developed them you should be careful.

Biological Databases

Santhosh J. Eapen

Computers and WWW has brought a paradigm shift in biological science, affecting every aspect of it. Research that used to start in the laboratory now starts at the computer. Besides helping in the data collection, storage and access processes in the laboratory, the computers are powerful devices for understanding any system that can be described mathematically. This remarkable feature of computers has given rise to the discipline of Bioinformatics.

What is Bioinformatics?

First and foremost, bioinformatics is a biological science, the science of using information to understand biology. In a literal sense Bioinformatics is the integration of life sciences and information science. Being a young science everyone defines it differently. A common definition is the ' Science of organizing and analyzing increasingly complex biological data resulting from modern molecular and biochemical techniques'. But a classical definition describes it as 'the mathematical, statistical and computing methods that aim to solve biological problems using DNA and amino acid sequences and related information'. Bioinformatics is conceptualizing biology in terms of molecules and applying 'informatics techniques' to understand and to organize the information associated with these molecules, on a large scale. In short, information science has been applied to manage the information science has been applied to manage the information generated in molecular biology to produce the field called **Bioinformatics.**

The functional roles of Bioinformatics are many. This branch of science developed because of the need to handle staggering amount of biological data generated due to the technical advances in genome sequencing (genomics) and protein identification (proteomics). Bioinformatics employs computational power to catalog, organize and structure these data into biologically meaningful information. This includes design of intelligent data formats and databases, creation and development of query tools and user interfaces to access these databases. In short, it is the science of creating and managing biological databases to keep track of and eventually simulate the complexity of living organisms. Thus, biology, computer science and information technology merge into a single discipline in Bioinformatics.

The second most important role is to develop analytical tools to discover new insights in the data stored. This comprises of comparing sequences to predict the function of a newly discovered gene, understanding how a protein folds based on known 3D protein structures in the databases or modeling the function of proteins and metabolites in a cell. The ultimate goal of Bioinformatics is to evolve predictive methods that allow modeling the function of an organism based purely on its genetic makeup. So essentially Bioinformatics is a tool that helps us to determine the function of new proteins, to predict how a protein will look like and to identify new genes and their functions in a cell with precision and accuracy. Therefore, Bioinformatics is theoretical modeling that provides testable hypotheses, not definitive answers to researchers.

The ultimate goal of the field is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned, Ultimately, however, all of this information must be combined to form a comprehensive picture of normal cellular activities so that researchers may study how these activities are altered in different disease states. Therefore, the field of Bioinformatics has evolved such that the most pressing task now involves the analysis and interpretation of various types of data, including nucleotide and amino acid sequences, protein domains, and protein structures.

Public biological databases

At the beginning of the "genomic revolution" a Bioinformatics concern was the creation and maintenance of a database to store biological information, such as nucleotide and amino acid sequences. Development of this type of database involved not only design issues, but also the development of complex interfaces whereby researchers could both access existing data as well as submit new or revised data.

Storage of macromolecular data like a gene, protein, an enzyme or a functional RNA in electronic databases is a complicated task. At present each entry into the databases is given a serial number and then stored in a relational database that knows the proper linkages between that serial number and all information pertaining to that entry. DNA sequence, protein sequence and protein structure, though well related are currently maintained as separate database projects and in unconnected data formats, because biological sequence and structure databases have evolved in parallel.

The first molecular biology database was the Protein Data Bank (PDB), the central repository for X-ray crystal structures of protein molecules (at the Brookhaven National Laboratory in 1971). From 15 sets of co-ordinates in 1973, now it has about 64623 entries. Journals that publish crystallographic results now require submission to the PDB as a condition of publication.

Sequence databases generally specialize in one type of sequence data: DNA, RNA or protein. There are major sequence data collections and deposition sites in Europe, Japan and the United States. The first DNA sequence database was the Gene Sequence Database (GSDB) at Los Alamos National Lab in 1979. Now up-to-date gene sequence information is available from GSDB through the National Centre for Genome Resources. The European Molecular Biology Laboratory, the DNA Database of Japan and the National Institutes of Health co-operated to support the database GenBank. Numerous biological databases and information sources are available on the web.

Nucleotide sequence databases

GenBank (http://www.ncbi.nlm.nih.gov/genbank/): GenBank is the nucleotide sequence database built and distributed by the National Centre for Biotechnology Information (NCBI) at the National Institutes of Health. As on today Biotechnology contains more than 85 billion bases from over 100,000 species, and is growing exponentially and over 30,000 people per day access it. The data are obtained through direct submission of sequence data from individual laboratories, from large-scale sequencing projects, and from the US Patent and Trademark Office. A little more than half of the total sequences in the database are from *Homo sapiens.*

There are two ways to search GenBank: a text-based query can be submitted through the Entrez system at *www.ncbi.nlm.nih.gov/Entrez/* or a sequence query can be submitted through the **BLAST** family of programs (see http://www.ncbl.nlm.nih.gov/BLAST/). To search GenBank through the **Entrez system** you would select the Nucleotides database from the menu. The **Entrez Nucleotides Database** is a collection of sequences from several sources, including GenBank, RefSeq, and the Protein Databank, so you don't actually search GenBank exclusively. Searches of the Entrez Nucleotides database query the text and numeric fields in the record, such as the accession number, definition, keyword, gene name, and organism fields to name just a few. Nucleotide sequence records in the Nucleotides database are linked to the PubMed citation of the article in which the sequences were published. Protein sequence records are linked to the nucleotide sequence from which the protein was translated.

If you have obtained a record through a text-based Entrez Nucleotides Database search you can read the nucleotide sequence in the record. However, most researchers wish to submit a nucleotide sequence of interest to find the sequences that are most similar to theirs. This is done using the BLAST (Basic Local Alignment Search tool) programs. You select the BLAST program you wish to use depending upon the type of comparison you are doing (nucleotide to nucleotide, or nucleotide to protein sequence, etc.) and then you select the database to run the query in (any of several nucleotide or protein databases).

EMBL Nucleotide Sequence Database (http://www.ebi.ac.uk/embl/): “The EMBL Nucleotide Sequence Database constitutes Europe's primary nucleotide sequence resource. Main sources for DNA and RNA sequences are direct submissions from individual

researchers, genome sequencing projects and patent applications. The database is produced in an international collaboration with GenBank (USA) and the DNA Database of Japan (DDBJ). Each of the three groups collects a portion of the total sequence data reported worldwide, and all new and updated database entries are exchanged between the groups on a daily basis."
From the home page you can submit simple text searches to the EMBL Nucleotide Sequence Database, or to the Protein Databank (what you search when you select protein structures from the menu) or to a protein sequence database called Swall. For more complex searches, they recommend accessing the databases through the Sequence Retrieval System (SRS) server (littp://srs.ebi.ac.uk/). SRS is a database querying / navigation system, similar in function to the Entrez system. It allows you to simultaneously search across several databases and to display the results in many ways. SRS can be used to access a large number of databases, including EMBL, SWISS-PROT and the Protein Databank, depending upon the configuration of the particular SRS server you are using. The structure and content of an EMBL Nucleotide record is very similar to that of an NCBI Entrez Nucleotide database record.

DDBJ DNA Data Bank of Japan: DDBJ is the sole DNA data bank in Japan, which is officially certified to collect DNA sequences from researchers and to issue the internationally recognized accession number to data submitters. It collects data mainly from Japanese researchers, but of course accepts data and issues the accession number to researchers in any other countries and exchange the collected data with EMBL/EBI and GenBank/NCBI on a daily basis. The three data banks share virtually the same data at any given time.

Protein sequences databases

SWISS-PROT (http://us.expasy.org/sprot/): SWISS-PROT is a curated protein sequence database which strives to provide a high level of annotation (such as the description of the function of a protein, its domain structure, post-translational modifications, variants, etc.), a minimal level of redundancy and a high level of integration with other databases. The data in Swiss-Prot are derived from translations of DNA sequences from the EMBL Nucleotide Sequence Database, adapted from the Protein Identification Resource (PIR) collection, extracted from the literature and directly submitted by researchers. It contains high-quality annotations, is nonredundant, and cross-referenced to several other databases, notably the EMBL nucleotide sequence database, PROSITE pattern database and PDB.
From the home page, a quick text search can be done by accession or ID number, description, gene name, or organism. By searching SWISS-PROT through the Sequence Retrieval System (SRS) more sophisticated searches can be performed and the format of the results can be customized. Access to SWISS-PROT (directly or via SRS) and links too many other proteomics resources are available from the ExPASy (Expert Protein Analysis System) proteomics server of the Swiss Institute of Bioinformatics (SIB) at http://ca.expasy.org/. The SWISS-PROT records are quite detailed. Be advised that other databases or search systems that import SWISS-PROT data may not always provide access to the entire SWISS-PROT record.

Entrez Protein Database (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Protein): The Protein database contains sequence data from the translated coding regions from DNA sequences in GenBank, EMBL and DDBJ as well as protein sequences submitted to PIR, SWISSPROT, PRF, and the Protein Data Bank (PDB) (sequences from solved structures). The native SWISS-PROT records usually contain more detailed annotations than will be obtained from Entrez Protein Database records derived from SWISS-PROT records. In typical Entrez fashion, results from a search of the Protein database link to PubMed, to the taxonomy database, to related sequences, and in some cases to pre-computed BLAST search results (look for BLink links).

Protein Information Resource - International Protein Sequence Database (PIRPSD) (http://pir.georgetown.edu/): In 1988 the Protein Information Resource (PIR), which is affiliated with Georgetown University Medical Center, established a cooperative effort with the Munich Information Center for Protein Sequences (MIPS) and the Japan International Protein Information Database (JIPID) to collect, publish and distribute the PIR-International Protein Sequence Database (PIR-PSD). They describe the database as "a comprehensive, non-redundant, expertly annotated, fully classified and extensively cross-referenced protein sequence database in the public domain". Text searches can be done in the title, species, author, citation, keyword, superfamily, feature and gene name fields. Gapped-BLAST sequence similarity searches are also an option. Note that both SWISS-PROT and the Entrez Protein database contain data adapted from the PIR.

Genome databases

Entrez Genome (http://www.ncbi.nlm.nih.gov/sites/genome): The whole genomes of over 6272 organisms can be found in Entrez Genomes. The genomes represent both completely sequenced organisms and those for which sequencing is in progress. All three main domains of life - bacteria, archaea, and eukaryota - are represented, as well as many viruses and organelles. Text searches can be done from the main page. Data can also be accessed alphabetically by species (http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/allorg.html), or hierarchically by drilling down through a taxonomic list to a graphical overview for the genome of that organism, then to specific chromosomes, then on to specific genes. At each level are maps, pre-computed summaries, and analysis appropriate to that level, and links to related records from a variety of other Entrez databases. BLAST searches of some genomes are also possible.

Very useful pages for some of the most commonly studied species (e.g., human, mouse, fruit fly, malaria parasite) can be found on the Genomic Biology page under "organism-specific resources" (http://www.ncbi.nlm.nih.gov/Genomes/). These pages are so detailed that each could be classified as a comprehensive web site in itself. Each one brings together links to the genomic data, useful tools, related data sources and news about the genome of that species.

The Genome Database (GDB) (http://www.gdb.org/): The Genome Database is the official central repository for genomic mapping data resulting from the Human Genome Initiative. The database contains three types of data: (1) regions of the human genome, including genes, clones, and ESTs, (2) maps of the human genome, including cytogenetic maps, linkage maps, radiation hybrid maps, content contig maps, and integrated maps (these maps can be displayed graphically via the Web), and (3) variations within the human genome including mutations and polymorphisms, plus allele frequency data. There are options to browse genes by chromosome, genes by symbol name, and genetic diseases by chromosome. There are multiple ways to search, including text-based searches for people, citations, segment names or accession numbers, and sequence searching via BLAST.

KEGG - Kyoto Encyclopedia of Genes and Genomes (http://www.genome.jp/kegg/): Despite the name, this is actually a biochemical pathway database and gene catalog, not an encyclopedia in the book sense. The primary objective of KEGG is to computerize the current knowledge of molecular interactions; namely, metabolic pathways, regulatory pathways, and molecular assemblies. At the same time, KEGG maintains gene catalogs for all the organisms that have been sequenced and links each gene product to a component on the pathway. Because we need an additional catalog of building blocks, KEGG also organizes a database of all chemical compounds in living cells and links each compound to a pathway component.

Genome mapping in crop plants (http://ukcrop.net): UK CropNet was established in 1996 to apply a bioinformatics approach to the study of genome evolution in crop plants. It contains genome information for *Arabidopsis,* Barley, Brassica, Forage grasses, Millet and comparative analysis. All UK CropNet databases are freely available for anyone to download (though you will also need to download the necessary ACEDB software).

The TIGR *Arabidopsis thaliana* Database (http://www.tigr.org/tdb/e2kl/athl/): Provides access to *Arabidopsis* genomic sequence data and annotation generated at TIGR and assemblies of *Arabidopsis* ESTs from world-wide sequencing projects.

Arabidopsis Information Resource (TAIR) (www.arabidopsis.org): The Arabidopsis Information Resource (TAIR) provides a comprehensive resource for the scientific community working with *Arabidopsis thaliana,* a widely used model plant. TAIR consists of a searchable relational database, which includes many different datatypes. The data can be viewed using our interactive MapViewer, and analyzed.

Oryzabase (http://www.shigen.nig.ac.jp/rice/oryzabase/): The Oryzabase is a comprehensive rice science database established in 2000 by rice researcher's committee in Japan. The database is originally aimed to gather as much knowledge as possible ranging from classical rice genetics to recent genomics and from fundamental information.

GOBASE - Organelle genome database (http://megasun.bch.umontreal.ca/gobase): GOBASE is a taxonomically broad organelle genome database that organizes and integrates diverse data related to organelles. The new version focuses on the mitochondrial and chloroplast subset of data. In its third phase, GOBASE will also include information on representative bacteria that are thought to be specifically related to the bacterial ancestors of mitochondria and chloroplasts.

Comprehensive Microbial Resource (CMR) (http://cmr.jcvi.org/cgi-bin/CMR/CmrHomePage.cgi): CMR is a tool that allows the researcher to access all of the bacterial genome sequences completed to date. It contains the sequence and annotation of each of the completed genomes as well as associated information about the organisms (such as taxon and gram stain pattern), the structure and composition of their DNA molecules (such as plasmid vs. chromosome and GC content), and many attributes of the protein sequences predicted from the DNA sequence (such as pI and molecular weight).

Further reading:

1. Baxevanis, A.D. (2002) The Molecular Biology Database Collection: 2002 update, *Nucleic Acids Research,* 30(1), 1-12.
2. Baxevanis, A.D. and Ouellette, B.F.F. (Eds.) (2001). Bioinformatics: A Practical Guide to Analysis of Genes and Proteins. John Wiley & Sons, New York, USA.
3. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A. and Wheeler, D.L. (2002) GenBank, *Nucl. Acids. Res.* 30: 17-20.
4. Galperin, M. Y. (2008). The Molecular Biology Database Collection: 2008 update. Nucleic Acids Research, 2008, Vol. 36, D2-D4.
5. Gibas, C. and Jambeck, P. (2001). Developing Bioinformatics Computer Skills. O'Reilly, California, USA.
6. http://home.uchicago.edu/-ebetran/bioinformatics.html
7. http://bioinforrnatics.org/faq/
8. Misener, S. and Krawetz, S.A. (2000). Bioinformatics: Methods and Protocols. Humana Press, New Jersey, USA.
9. Mount, D.W. 2001. Bioinformatics: Sequence and Genome Analysis. Cold Spring Harbor Laboratory Press, New York, USA.
10. National Center for Biotechnology Information (NCBI) Education Site. [Online]. Available: http://www.ncbi.nlm.nih.gov/Education/

11. Westbrook, J., Feng, Z., Jain, S., Bhat, T.N., Thanki, N., Ravichandran, V., Gilliland, G.L., Bluhm, W., Weissig, H., Greer, D.S., Bourne, P.E. and Berman, H.M. (2002). The Protein Data Bank: unifying the archive. *Nucl. Acids. Res.* 30: 245-248.
12. Wheeler, D.L., Church, D.M., Lash, A.E., Leipe, D.D., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Tatusova, T.A., Wagner, L. and Rapp, B.A. (2002). Database resources of the National Center for Biotechnology Information: 2002 update. Nucl. Acids. Res. 30: 13-16.

Intellectual property rights

D. Prasath

With the emerging intellectual property rights regimes under new and upcoming legislations, in conformity with the global developments and inter governmental agreements, there is a need to protect the intellectual property developed in various areas of agricultural sector. Similarly the major achievements of a scientific research organisation involved in agriculture are primarily reflected in its viable technologies developed, research publications and the intellectual property rights it holds.

Intellectual Property (IP)

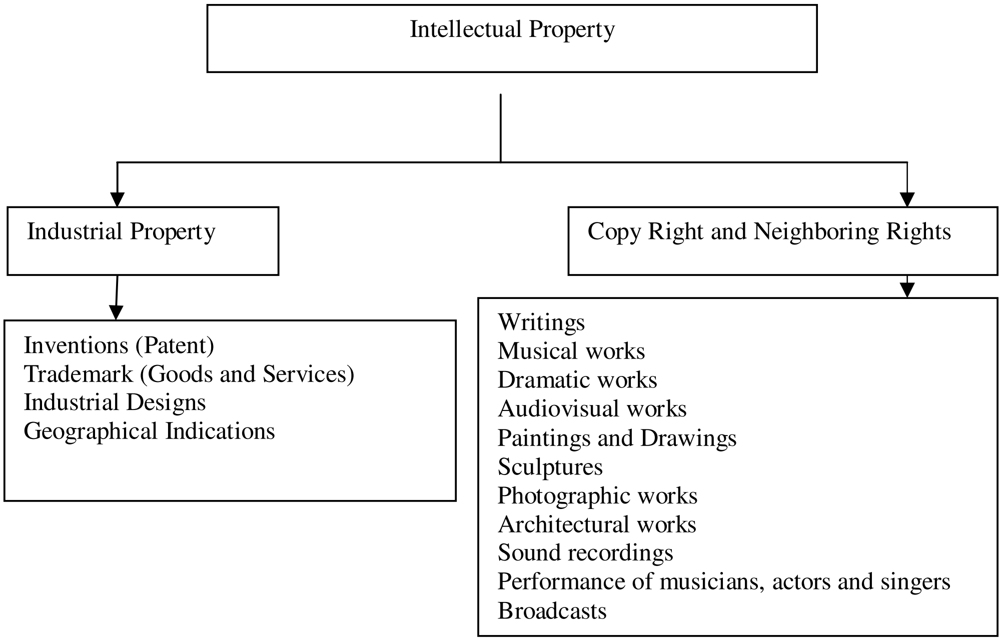
Intellectual property refers to creations of the mind such as inventions, literary and artistic works, and symbols, names, images, and designs used in commerce. There is no uniform definition for IP. The domain of IP is expanding fast as knowledge and information become key drives of techno-economic growth and of societal progress in general.
Intellectual property is divided into two main categories,

i. Inventions (patents), trademarks, industrial designs, and geographic indications of source
ii. Copyright, which includes literary and artistic works such as novels, poems and plays, films, musical works, artistic works such as drawings, paintings, photographs and sculptures, and architectural designs. Rights related to copyright include those of performing artists in their performances, producers of phonograms in their recordings, and those of broadcasters in their radio and television programs.

The TRIPS (Trade Related Intellectual Property Rights) Agreement recognizes seven types of IP rights. They are

- Copyright and Related Rights
- Trade marks, Trade names and service marks
- Geographical indications
- Industrial designs
- Patents
- Layout designs of integrated circuits
- Undisclosed information

Article 7 of the TRIPS agreement states the objective of Intellectual Property protection as, "*the protection and enforcement of intellectual property rights (with the objective of contributing) to the promotion of technological innovation and to transfer and (disseminate) technology, to the mutual advantage of producers and users of knowledge and in a manner conducive to social and economical welfare, and to balance the rights and obligations*".



Different forms of Intellectual Property

Patent

A patent is an exclusive right granted for an invention, which is a product or a process that provides, in general, a new way of doing something, or offers a new technical solution to a problem. In order to be patentable, the invention must fulfill criteria of novelty, involving an inventive step and being capable of industrial applications.

A patent provides protection for the invention to the owner of the patent. The protection is granted for a limited period, generally 20 years. Patent protection means that the invention cannot be commercially made, used, distributed or sold without the patent owner's consent. These patent rights are usually enforced in a court, which, in most systems, holds the authority to stop patent infringement. Conversely, a court can also declare a patent invalid upon a successful challenge by a third party.

India's history with IP law is deeper than that of many developing countries. The first patent law was adopted in 1856, and by 1911. The Indian patent law conformed to the standards of developing countries. Patent law was strengthened in 1999 and again in 2002 to become compliant with TRIPS. Legislation passed in 2005 reinstated the patenting of pharmaceutical compounds.

In order to ensure the interests of society, as per Indian patent Act, certain things have been excluded from the preview of the patentability (Anon, 2007).

- A method of agriculture or horticulture (Section 3. h)
- Any process of medicinal, surgical, curative, prophylactic (diagnostic, therapeutic or other treatment of human beings or any process for a similar treatment of animals to render them free of disease or to increase their economic value or that of their products (Section 3. i)
- Plants and animals in whole or any part thereof other than microorganisms but including seeds, varieties and species and essentially biological processes for production or propagation of plants and animals (Section 3. j).
- An invention which in effect, is traditional knowledge or which is an aggregation or duplication of known properties of traditionally known component or components (Section 3. p).

Patenting of Plant varieties

Patent law was originally considered unsuitable for protecting new plant varieties developed by traditional breeding methods. Some countries therefore introduced special national laws for PVR in the 1960s, as did the International Union for the Protection of New Varieties of Plants (UPOV), established in 1961. These rights are granted by the state to plant breeders to exclude others from producing or commercializing material of a specific plant variety for a minimum of 15 to 20 years.

In the United States, the Plant patent Act (PPA), the Plant variety protection Act (PVPA) and Utility Plant variety protection Act (UPP) are the statuary forms of plant germplasm rights available. Utility patents provide the most extensive coverage for invention. In agricultural biotechnology, the utility patent may be obtained to protect everything from genetically modified seeds and genetically modified plants, to transformation methods. For example, Monsanto's utility patent Roundup Ready® soybeans, allow the company to claim protection not only for the methods of producing the Roundup Ready® soybeans, but also for the DNA molecule that encodes the herbicide resistant trait, for the herbicide resistant plant cell, and for the final Roundup Ready® soybean plant itself. While utility patents provide the most robust protection for plant innovation, only few countries offer utility patent protection for agriculture. A more common regime is plant variety protection, a *sui generis* form of IP protection to breeders of new varieties of plants.

India does not protect plants by strict patenting system and has enacted 'The Protection of Plant Varieties and Farmers Rights Act, 2001' a *sui generis* system of plant variety protection. The idea behind this Act is to establish an effective and efficient system for protecting plant varieties, farmers' rights and the right of the plant breeders and developing new plant varieties. The PPV&FR Act is TRIPS complaint and compatible with UPOV system of plant variety protection.

The PPV&FR Act, 2001 provides protection to following types of plant varieties.

- Newly bred varieties
- Extant varieties (The varieties which were released under Indian Seeds Act, 1966 and have not completed 15 years as on the date of application for their protection.
- Farmer's varieties
- Essentially derived varieties
- Transgenic varieties

Patenting of microorganism

The TRIPS agreement makes it obligatory to provide patents for microorganisms and microbiological processes. The word microorganism is not defined in the agreement and does not specify any parameters concerning the scope of protection. Microorganism *per se* can be claimed for protection provided they are not mere discovery of organism (Chawla, 2007). The first patent on living organism was granted to microorganism, *Pseudomonas*. The bacterium manipulated to contain more than one plasmid controlling the breakdown of hydrocarbons, thus a new bacterium with markedly different characteristics from any found in nature. If the microorganism is to form the subject of patent application, the microorganism should be deposited under the terms of the Budapest Treaty on the International recognition of the deposit microorganisms for the purpose of patent procedure. It the requirements of this treaty are satisfied, the microorganism will be deemed to satisfy the enablement requirements of countries that are members of the treaty. Once the samples has been deposited, the International depository Authority (IDA) is obliged to check the viability of the deposited microorganism and then to issue a deposit receipt and viability statement, copies of which are often required by the patent offices in order to satisfy themselves that the organism has been deposited. Currently there are 60 IDAs, and 60 countries have signed Budapest treaty (Webber, 2006).

The Indian Patent Act has now a specific provision in regard to patenting of microorganisms and microbiological processes. It is now possible to get a patent for a microbiological process and also products emanating from such processes. Substances prepared or produced by chemical process, which include biochemical, biotechnological and microbiological, no patent shall be granted in respect of claim for the substances themselves, but claim for the methods or process of manufacturing shall be patentable. Investor has to deposit the new strain in any recognized international repository. In India, Microbial Type Culture Collection and Gene bank (MTCC) at the Institute of Microbial Technology (IMTECH), Chandigarh, is a recognized international repository of microorganisms (Chawla, 2005). An international treaty called "Budapest Treaty" was signed in Budapest in 1973 and later on amended in 1980 and India became a member of this Treaty, with effect from December 17, 2001. This is an international convention governing the recognition of deposits in officially approved culture collections for the purpose of patent applications in any country that is a party to this treaty. The examples of the patent details obtained area are given below.

Area		**Patent details**	
Biotech.	Gene sequence	DNA sequence for root preferred gene expression in plants	USA Patent 7173166
	Transgenics	Transgenic tea through biolistic using leaf explants	USA Patent 7129394
		Transgenic plants with enhanced chlorophyll content and salt tolerance	USA Patent 6791009
	Tissue culture	Tissue culture process for producing a large number of viable mint plants in vitro	USA Patent 6323394
	Culture medium	A nutrient medium composition useful for enhancing shoot sprouting and multiplication from mature bamboo species	Indian Patent 176550
Biochem.	Biosynthesis	Process for the synthesis of curcumin-related compounds	United States Patent 5679864
Microbes	Biological control	Strain of Trichoderma harzianum useful as nematode inhibitor, fungicide and plant growth promoter and a process for the isolation thereof	USA Patent 6475772
Horticultural techniques	Growth regulators	Plant growth stimulator	USA Patent 7053025
	Seed germination	Method for inducing improved seed germination in Podophyllum hexandrum Royle	USA Patent 6449899
Post harvest technology	Essential oil extraction	Formulation and method for increasing the essential oil content in aromatic crops	USA Patent 6720289
	Processing	An improved process for the preparation of bright juices from noncitrus fruits	Indian Patent 191361
Devices	Processing	An improved honey processing device	Indian Patent 189380
	Soil testing	A portable digital soil salinity tester	Indian Patent 181263
	Agricultural implements	Multifunctional digging tool to function as spade-cum-hoe	Indian Patent 167940
		Paddy transplanting machine	Indian Patent 120490

Trademarks

A trademark provides protection to the owner of the mark by ensuring the exclusive right to use it to identify goods or services, or to authorize another to use it in return for payment. The period of protection varies, but a trademark can be renewed indefinitely beyond the time limit on payment of additional fees. Trademark protection is enforced by the courts, which in most systems have the authority to block trademark infringement.

In a larger sense, trademarks promote initiative and enterprise worldwide by rewarding the owners of trademarks with recognition and financial profit. Trademark protection also hinders the efforts of unfair competitors, such as counterfeiters, to use similar distinctive signs to market inferior or different products or services. The system enables people with skill and enterprise to produce and market goods and services in the fairest possible conditions, thereby facilitating international trade.

Almost all countries in the world register and protect trademarks. Each national or regional office maintains a Register of Trademarks, which contains full application information on all registrations and renewals, facilitating examination, search, and potential opposition by third parties. The effects of such a registration are, however, limited to the country (or, in the case of a regional registration, countries) concerned.

The principal form of Intellectual Property protection for agriculture is plant patents, plant variety protection patents and utility patents. However, trademarks can also provide long lasting and significant protection for plant varieties. The advantage of trade mark over the statutory forms of IP protection for plants is that trademarks can be protected indefinitely as long as the product is marketed and the trademark enforced. Trademarks have helped create value for agricultural products (Tracker and Ross, 2007).

Trade mark	Features
Roundup Ready®	Crops developed by Monsanto that contains transgenes that encode tolerance to herbicide glyphosate
Horticulture crop varieties Pink lady® (apples), superior seedless® (grapes), camarosa® (strawberry)	To emphasize distinctive and attractive attributes of varieties

The designation ™ indicates that a particular word or symbol or logo is considered by its user to function as a trademark. When a mark becomes registered, for example with the US patent office, the designation should change from ™ to the registers mark symbol ®.

The most important agreements dealing with international registration of trademark are Madrid system and Madrid protocol. Licensing of a trademark can either stand alone or be combined with other forms of IP rights protection.

Enactment of the Indian Trademarks Act 1999 is a big step forward from the Trade and Merchandise Marks Act 1958 and the Trademark Act 1940.

Industrial designs

An Industrial design is the ornamental or aesthetic aspect of an article hence, they add to the commercial value of a product and increase its marketability. An industrial design is distinguished from trade mark primarily it is constituted by the appearance of a product, which is not necessarily distinctive.

When an industrial design is protected, the owner - the person or entity that has registered the design - is assured an exclusive right against unauthorized copying or imitation of the design by third parties. This helps to ensure a fair return on investment. An effective system of protection also benefits consumers and the public at large, by promoting fair competition and honest trade practices, encouraging creativity, and promoting more aesthetically attractive products.

Geographical indication

A Geographical Indication (GI) is a sign used on goods that have a specific geographical origin and possess qualities or a reputation that are due to that place of origin. Most commonly, a geographical indication consists of the name of the place of origin of the goods. Agricultural products typically have qualities that derive from their place of production and are influenced by specific local factors, such as climate and soil. Whether a sign functions as a geographical indication is a matter of national law and consumer perception. Geographical indications may be used for a wide variety of agricultural and horticultural products.

The Paris Convention uses two terms in the context of geographical indications ie appellation of origin and indication of sources. Indication of source on a product merely indicates that the product originates in the place indicated. Appellation of origin indicated not only the place of origin but also the essential quality link between the product and the area of its origin.

An effective protection for GIs was of considerable importance for a country like India, which was richly endowed with natural and agricultural products and which already had in its possession renowned geographical names such as 'Darjeeling' (tea), 'Alphonso' (mango), 'Coorg' (mandarin), etc., there was no separate legislation on GIs until the enactment of 'The Geographical Indications of Goods (Registration and Protection) Act, 1999'. Under this Act, which has come into force with effect from 15th September 2003, the Central Government has established the 'Geographical Indications Registry' with all India jurisdictions at Chennai. The registration of GIs in India has been depicted in the following table.

Geographical Indications	Products	State	Date of
Registration			
Darjeeling Tea	Tea	West Bengal	29.10.04
Mysore Agarbathi	Incense sticks	Karnataka	02.06.05
Kangra Tea	Tea	Himachal Pradesh	12.12.05
Coorg Orange	Horticulture	Karnataka	30.01.06
Mysore Betel Leaf	Horticulture	Karnataka	30.01.06
Nanjanagud Banana	Horticulture	Karnataka	30.01.06
Mysore Sandalwood Oil	Essential Oil	Karnataka	30.01.06
Mysore Sandal Soap	Soap	Karnataka	30.01.06

(Hirwade and Hirwade, 2007)

Copyright and Related Rights

Copyright and related rights are legal concepts and instruments which, while respecting and protecting the rights of creators in their works, also contribute to the cultural and economic development of nations. Copyright law fulfills a decisive role in articulating the contributions and rights of the different stakeholders taking part in the cultural industries and the relation between them and the public.

The original creators of works protected by copyright, and their heirs, have certain basic rights. They hold the exclusive right to use or authorize others to use the work on agreed terms. The creator of a work can prohibit or authorize:

- its reproduction in various forms, such as printed publication or sound recording;
- its public performance, as in a play or musical work;
- recordings of it, for example, in the form of compact discs, cassettes or videotapes;
- its broadcasting, by radio, cable or satellite;
- its translation into other languages, or its adaptation, such as a novel into a screenplay.

Many creative works protected by copyright require mass distribution, communication and financial investment for their dissemination (for example, publications, sound recordings and films); hence, creators often sell the rights to their works to individuals or companies best able to market the works in return for payment. These payments are often made dependent on the actual use of the work, and are then referred to as royalties.

Copyright protection extends only to expressions, and not to ideas, procedures, methods of operation or mathematical concepts as such. Copyright and its related rights are essential to human creativity, by giving creators incentives in the form of recognition and fair economic rewards. Under this system of rights, creators are assured that their works can be disseminated without fear of unauthorized copying or piracy. This in turn helps increase access to and enhances the enjoyment of culture, knowledge, and entertainment all over the world.

The copyright symbol, © or the designation Copr. and Copyright are the proper legal notices for copyright protection. The Copyright Act, 1957 has been enacted to amend and consolidate the law relating to copyrights in India. The 1957 Act was recently amended by the Copyright (Amendment) Act, 1999 seeking to extend the term of protection of performer's rights from twenty five years to fifty years.

IPR and Biotechnology

Under biotechnology, one of the most important examples of intellectual property is the processes and products, which result from the development of genetic engineering techniques through the use of restriction enzymes to create recombinant DNA (http://www.molecular-plant-biotechnology.info/biotechnology-intellectual-property rights- IPR-Intellec). The IPR issues has special importance in Biotechnology since most of the products and processes in Biotechnology can be copied (Mayer, 2003), for eg., genes can be copied once their sequences are unveiled, seeds can be replanted, methods can be repeated using standardized protocols. Under the World Trade Organization's Agreement on Trade Related Aspects of Intellectual Property Rights (TRIPS), most processes and many products of biotechnology research are patentable. India has been practicing conventional Biotechnology and there are mostly issues pertaining to unprotectable intellectual property. The forms of protection with relevance to Biotechnology and Agriculture are copyright, trade secrets, trademarks, plant breeder's rights, and patents. Trade secrets in the area of biotechnology may include material like (i) hybridization conditions (ii) cell lines (iii) corporate merchandising plans or (iv) customer lists. Unlike patents, trade secrets have an unlimited duration. Disclosure of a trade secret and its unauthorized use can be punished by the court. Due to large degree of research component in biotechnology, there is increased risk in maintaining trade secrets. The results are published and discussed in conferences and disclosed.

In biotechnology, the copyright may cover DNA sequence data which may be published. However, an alternative sequence coding for same protein may be prepared using wobble in the genetic code, so that the copyright is not infringed. Computer data bases and photomicrographs of DNA instruction manuals related to biotechnology can also be copyrighted. The protection to copyright is limited, however, since although one may not be allowed to photocopy the present book on biotechnology due to copyright, but the ideas given in the book can be used for any purpose. This is not so in case of patents and trade secrets. An instruction manual can be copyrighted and also protected as a trade secret, but can not be patented.

In biotechnology research, laboratory equipment bears trademarks that are well known to workers in this field. Certain vectors useful in recombinant research are also known by their trademarks. Although ordinarily this intellectual property may not raise any legal issue, it docs make a part of intellectual property of any industry (http://www.molecular-plant-biotechnology.info/biotechnology-intellectual-property-rights-IPR-Intellec).

Patents are the most critical form of protection for agricultural biotechnology and considered to be the most powerful in the IP system. Patents are temporary, generally about 20 years, and are country specific (Kowalski *et al*., 2002). In agricultural biotechnology, patents may cover, for example, plant transformation methods, vectors, genes, etc. and in countries that allow patenting of higher life forms, transgenic plants or animals.

MTAs are preferred mechanism for transfer of cloned genes and plasmid constructs in countries where IPR enforcement is not appropriate (Kowalski *et al.*, 2002). MTAs may have provisions that deal with IPR issues– such as, a requirement for the recipient to get an OK or agreement from the provider before any IPRs are sought over improvements associated with the use of the materials, or provisions that prohibit any IPRs being taken out over such improvements/innovations, or mandated ownership or licensing of any improvements of IPRs (eg., automatic joint ownership). However, there is no requirement that an MTA must include provisions that deal with IPR-type issues.

Patentable issues in biotechnology

The Biotechnology sector is dealt quite closely by Article 27 of TRIPS. Article 27 makes a distinction between biological materials produced by biological processes and those by essentially a non-biological routes- the later only being a subject matter of patents. Natural resources biodiversity is not protectable as an IPR of individuals by its mere possession as per article 27(3) (b) (Tripp *et al.*, 2007 www.elsevier.com/locate/foodpol) under the IPR provisions of the Country. In conformity with the provisions of TRIPS the naturally occurring substances like proteins, glycoproteins, carbohydrates, genes, DNA's and RNA's are all exempted from IPR. A mere DNA fragment without indication of a function or specific asserted utility is not patentable invention. A DNA fragment, of which specific utility, e.g. use as a probe to diagnose a specific disease, is disclosed, is a patentable invention as long as there is no other reasons for rejection. A DNA fragment showing no unexpected effect, obtained by conventional method, which is assumed to be part of a certain structural gene based on its high homology with a known DNA encoding protein with a known function, is not a patentable invention. The mere fact that DNA fragments are derived from the same source is not sufficient to meet the requirement for unity of invention. Areas where patents of DNA and Proteins can be filed are: Gene Expression Analysis, eg., sequence specific probes, chip hybridization, sequence specific amplification, differential display, etc. Protein Arrays eg., antibody chip, 2-D gel, antigen chip, etc. Genomic Analysis (individual/subpopulation variation) eg., SNP, VNTR, STR, RFLP etc (http://www.ris.org.in/v7n2_editorial.pdf). In order to patent a protein/nucleic acid sequence, must know something about function of gene/protein. Similarly, in case of essentially derived varieties, e.g. a variety into which one gene construct has been introduced by genetic manipulation while maintaining the rest of the genome practically unscathed does not come under the IPR. According to the Article 14 of the UPOV a newly introduced gene does not make a variety. The registration of a genetically modified variety requires the permission of PBR holder. However, research tools needed for further downstream research and development can be patented. Some of the research tools that come under the purview are expressed sequence tags, expression systems, restriction enzymes and techniques related to DNA sequencing and single nucleotide polymorphisms

Patents in Biotechnology can be generally claimed for products, processes or devices involving: microbial transformations; fermentation; enzymes; novel organisms; antibodies; cell fusion; hybridomas; immunodiagnostics; vaccines; recombinant DNA; gene and protein sequences; polynucleotide hybridization; DNA probes and primers; bioreactors; cell and tissue culture; blood products; and other biotechnology applications and products. Patents for biotech products subjected to subsequent chemical modification or derivatization, such as chemically modified antibiotic derivatives, were generally not classed as biotech, especially where non-naturally occurring chemical structures can be claimed. As far as genetic engineering is considered patents are involve claim involving: i) recombinant DNA (including genetic recombinations and gene sequences); ii) hybridomas/monoclonal antibodies (including cell fusions) and; iii) DNA probes (including oligonucleotide hybridizations, PCR).

References

Anonymous, (2007). The Patent Act, 1970, Universal Law Publishing Co. Pvt. Ltd., Delhi.

Anonymous, (2007). The Copy Right Act, 1957, Commercial Law Publishers (India) Pvt. Ltd., Delhi

Anonymous, (2007). The Designs Act, 2000, Universal Law Publishing Co. Pvt. Ltd., Delhi.

Chawla, H.S. (2005). Patenting of biological material and biotechnology. *Journal of Intellectual Property Rights*. 10: 44-51.

Chawla, H.S. (2007). Review: Intellectual Property Rights. *Journal of Eco-friendly Agriculture*. 2(2): 103-112.

Kowalski, S.P., Ebora, R.V., Kryder, R.D., Reichman, J. (2002). Transgenic crops, biotechnology and ownership rights: what scientists need to know. *Plant Journal,* 31: 407-421.

Mittal, R and Singh, G. (2005). Patenting activities in agriculture from India. *Journal of Intellectual Property Rights*. 10: 315-320.

Nagarajan, S., Kannaiyan, S., Yada, S.P., Singh, S.P., Trivedi, R.K and Suresh Pal. (2007). A white paper on Agro-biodiversity hot spots. In: National level consultation on Agro biodiversity hot spots, 19-20 July, 2007, Annamalai Univeristy, Tamil Nadu.

Ramamohan Rao, T.V.S. (2007). Biotechnology inventions and patent regime. *Asian Biotechnology and Development Review*, 9(2): 109-121.

The Biological Diversity Act, 2002, The Gazette of India. Extra pt II, Sec 1 dated 5th February, 2003.

The Protection of Plant Varieties and Farmers Rights Act, 2001, The Gazette of India. Extra pt II, Sec 3(i) dated 30th October, 2001.

The Protection of Plant Varieties and Farmers Rights Rules, 2003, The Gazette of India. Extra pt II, Sec 3(i) dated 12th September, 2003.

The Geographical Indications of Goods (Registration and Protection) Act, 1999, The Gazette of India. Extra pt II, Sec 1 dated 30th December, 1999.

Tripp, R., Louwaars, N. and Eaton, D. (2007). Plant variety protection in developing countries – A report from the field. *Food Policy*, 32(3): 354-371.

Tripathi, K. K. Biotechnology and IPR regime in the context of India and other developing countries. Asian Biotechnology and Development Review. 1-24 (Available online at www.ris.org.in/article1_v7n2.pdf)

Webber, P.M. (2006). Patenting of microorganisms. Nature reviews, 5: 4.

WIPO@Internet. WIPO Database of Intellectual Property - Legislative Texts. The World Intellectual Property Organization, Geneva. http://www.wipo.int

GIS-A tool for biodiversity study

Utpala Parthasarathy

Biological Diversity or Biodiversity is the variety of life in all its forms, levels and combinations, which includes ecosystem diversity, species diversity, and genetic diversity. Biodiversity studies require cataloguing of collections to establish relationships among different datasets.

The Geographic differentiation of plant populations reflects the dynamics of gene flow and natural selection. Sampling of geographically distinct populations is a practical approach to understand biodiversity or genetic variation. Frankel et al. (1995) have considered this as the “cardinal principle” of the acquisition of crop genetic resources. Sampling among diverse eco-geographic areas is specifically recommended for conserving the genetic diversity of rare and wild species. The application of GA (geographic analysis) to these germplasm collection data helps to predict where species naturally occur, or may be successfully introduced.

Geographic analysis is based on the deployment of three basic technologies; global positioning systems (GPS), remote sensing devices and GIS software. GPS provides a quick and accurate method for recording locations on the surface of the earth. Remote sensing methods, such as aerial photography and orbital satellite systems provide extensive geographic and temporal coverage in spatial and spectral combinations that describe landscapes from a far. Most useful to biologists are combinations that describe topographical features and vegetation coverage. The analysis of spatial information with GIS tools introduces new strategies for understanding and exploiting patterns of geographic diversity and can be carried out efficiently with personal computers and GIS software.

A GIS is mapping software that links information about where things are with information about what things are like**.** A GIS map can combine many layers of information to give a full view of the area or the crop concern. So if the bio-diversity data takes the help of GIS to arrange its collection data then it will be easy to accommodate all the necessary information in one point. A GIS makes it possible to link, or integrate, information that is difficult to associate through any other means. Thus, a GIS can use combinations of mapped variables to build and analyze new variables.

A common strategy for sampling intra specific genetic diversity is to maximize the sampling of geographically distinct populations. The objective of this paper is to illustrate how geographic information coupled with geographic information system (GIS) can provide a new level of precision for establishing frameworks for sampling germplasm occurring in eco geographically diverse sites

Geographic information System (GIS) is a computer based tool for mapping and analyzing geographic phenomenon that exists, and event that occur, on the earth. Overall GIS should be viewed as a technology, not simply as a computer system. It is an integrated set of hardware and software tools used for the manipulation and management of digital spatial (geographic) and related attribute data.

The power of GIS is remarkable, but often overlooked. It can assemble, store, manipulate and analyze any information that comes with a spatial location attached, i.e., ‘geo-referenced data’. – A gene bank, for instance, can use GIS to make sense of all its passport and other data, linking each germplasm accession and its associated characterization and evaluation data with a particular spot on the globe. After feeding into a GIS, these data can dramatically transform the daunting task of planning collecting programme as well as to decide where to site *in situ* reserves.

What is GIS?

A geographic information system (GIS) integrates hardware, software, and data for capturing, managing, analyzing, and displaying all forms of geographically referenced information. GIS allows us to view, understand, question, interpret, and visualize data in many ways that reveal relationships, patterns, and trends in the form of maps, globes, reports, and charts. Geography plays a roll in nearly every decision we make. Choosing sites, targeting market segments, planning distribution networks, responding to emergencies, or redrawing country boundaries—all of these problems involve questions of geography. A GIS helps to answer questions and solve problems by looking at the data in a way that is quickly understood and easily shared. GIS technology can be integrated into any enterprise information system framework.

Three Views of a GIS

The Database View: A GIS is a unique kind of database of the world—a geographic database (geodatabase). It is an "Information System for Geography." Fundamentally, a GIS is based on a structured database that describes the world in geographic terms.

The Map View: A GIS is a set of intelligent maps and other views that show features and feature relationships on the earth's surface. Maps of the underlying geographic information can be constructed and used as "windows into the database" to support queries, analysis, and editing of the information. .

The Model View: A GIS is a set of information transformation tools that derive new geographic datasets from existing datasets. These geoprocessing functions take information from existing datasets, apply analytic functions, and write results into new derived datasets.

One of the main tasks of conservation biologists is to evaluate the viability of endangered and threatened species under different natural conditions, and under alternative options for wildlife management, reserve design, and habitat protection plans. These evaluations usually ask questions about the predicted future abundance, risk of extinction, or chance of recovery of the species; and are addressed by population viability analysis (PVA), which is a systematic examination of interacting factors that place a population or species at risk of extinction (Gilpin and Soul, 1986, Shaffer 1990). The factors that a PVA examines may be both natural and anthropogenic in origin, and their analysis often involves mathematical or computer models that predict the future changes in the abundance and distribution of the species in question, given information about its ecology and demography (Burgman et al. 1993).

Habitat loss and fragmentation are among the most common threats facing endangered species, making GIS-based evaluations an essential component of population viability analyses. Often habitat loss and fragmentation, combined with the natural heterogeneity of landscapes, forces species to exist in multiple populations inhabiting relatively isolated habitat patches. Such a collection of populations of the same species is called a metapopulation. The existence of multiple populations usually introduces complexities that make it impossible to evaluate the viability of the species based on PVAs performed on separate populations.

GIS is a rapidly growing technological field that incorporates graphical features with tabular data in order to assess real-world problems. What is now the GIS field began around 1960, with the discovery those maps could be programmed using simple code and then stored in a computer allowing for future modification when necessary. This was a welcome change from the era of hand cartography when maps had to be painstakingly created by hand; even small changes required the creation of a new map. The earliest version of a GIS was known as computer cartography and involved simple linework to represent land features. From that evolved the concept of overlaying different mapped features on top of each other to determine patterns and causes of spatial phenomenon.

The capabilities of GIS are a far cry from the simple beginnings of computer cartography. At the simplest level, GIS can be thought of as a high-tech equivalent of a map. However, not only can paper maps be produced far quicker and more efficiently, the storage of data in an easily accessible digital format enables complex analysis and modeling not previously possible. The reach of GIS expands into all disciplines and has been used for such widely ranged problems as prioritizing sensitive species habitat to determining optimal real estate locations for new businesses.

The key word to this technology is Geography - this usually means that the data (or at least some proportion of the data) is spatial, in other words, data that is in some way referenced to locations on the earth. Coupled with this data is usually data known as attribute data. Attribute data generally defined as additional information, which can then be tied to spatial data. An example of this would be schools. The actual location of the schools is the spatial data. Additional data such as the school name, level of education taught, school capacity would make up the attribute data. It is the partnership of these two data types that enables GIS to be such an effective problem solving tool.

GIS operates on many levels. On the most basic level, GIS is used as computer cartography, i.e. mapping. The real power in GIS is through using spatial and statistical methods to analyze attribute and geographic information. The end result of the analysis can be derivative information, interpolated information or prioritized information.

"In the strictest sense, a GIS is a computer system capable of assembling, storing, manipulating, and displaying geographically referenced information, i.e. data identified according to their locations. Practitioners also regard the total GIS as including operating personnel and the data that go into the system." *USGS*"A geographic information system (GIS) is a computer-based tool for mapping and analyzing things that exist and events that happen on earth. GIS technology integrates common database operations such as query and statistical analysis with the unique visualization and geographic analysis benefits offered by maps".

"*GIS is an integrated system of computer hardware, software, and trained personnel linking topographic, demographic, utility, facility, image and other resource data that is geographically referenced*." GIS has already affected most of us in some way without us even realizing it. If you've ever using an Internet mapping program to find directions, congratulations, you've personally used GIS. The new supermarket chain on the corner was probably located using GIS to determine the most effective place to meet customer demand.

Components of GIS

This article has briefly explained what GIS is. The next step in understanding GIS is to look at each component of GIS and how they work together. These components are:

Hardware

Hardware comprises the equipment needed to support the many activities of GIS ranging from data collection to data analysis. The central piece of equipment is the workstation, which runs the GIS software and is the attachment point for ancillary equipment. Data collection efforts can also require the use of a digitizer for conversion of hard copy data to digital data and a GPS data logger to collect data in the field. The use of handheld field technology is also becoming an important data collection tool in GIS. With the advent of web-enabled GIS, web servers have also become an important piece of equipment for GIS.

Software

Different software packages are important for GIS. Central to this is the GIS application package. Such software is essential for creating, editing and analyzing spatial and attribute data, therefore these packages contain a myriad of GIS functions inherent to them. Extensions or add-ons are software that extends the capabilities of the GIS software package. For example, Xtools is an ArcView extension that adds more editing capabilities to ArcView 3.x. Component GIS software is the opposite of application software. Component GIS seeks to build software applications that meet a specific purpose and thus are limited in their spatial analysis capabilities. Utilities are stand-alone programs that perform a specific function. For example, a file format utility that converts from on type of GIS file to another. There is also web-GIS software that helps serve data through Internet browsers.

Data

Data is the core of any GIS. There are two primary types of data that are used in GIS. A geodatabase is a database that is in some way referenced to locations on the earth. Geodatabases are grouped into two different types: vector and raster. Coupled with this data is usually data known as attribute data. Attribute data generally defined as additional information, which can then be tied to spatial data. Documentation of GIS datasets is known as metadata.

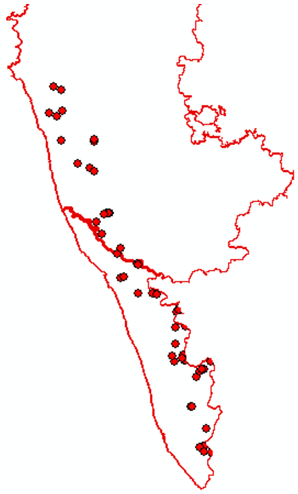
People

Well-trained people knowledgeable in spatial analysis and skilled in using GIS software are essential to the GIS process. There are three factors to the people component: education, career path, and networking. The right education is key; taking the right combination of classes. Selecting the right type of GIS job is important. A person highly skilled in GIS analysis should not seek a job as a GIS developer if they haven't taken the necessary programming classes. Finally, continuous networking with other GIS professionals is essential for the exchange of ideas as well as a support community.

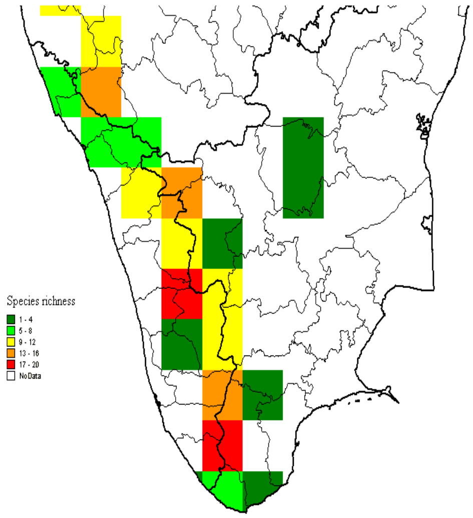
There are two types of data generally used in GIS viz., Vector data and Raster data. Vector data is the reality data represents with the help of points, lines, area the discreet objects or the view of reality. While in Raster data the reality is represented in the form of grids. Raster data files can be manipulated quickly by the computer, but they are often less detailed and may appear less visually appealing than vector data files, which can approximate the appearance of more traditional hand-drafted maps. Image data from a satellite that has been interpreted by a computer to produce a land-use map can be "read into" the GIS in raster format. Raster data files consist of columns and rows of uniform cells coded according to data values.

Making Maps

Maps have a special place in GIS. The process of making maps with GIS is much more flexible than are traditional manual or automated cartography approaches. It begins with database creation. Existing paper maps can be digitized and geo-coded, computer compatible information can be translated into the GIS with the help of suitable software like DIVA, Flora Map, ARC-GIS etc. The GIS based cartographic database can be both continuous and scale free. Map products can then be created centered on any location, at any scale and showing selected information symbolized effectively to highlight specific characteristics.



Vector data



Raster data

Helpful Software to study Genetic resource data

Mapping and spatial analysis of gene bank data can be carried out with off –the –shelf geographic information system software like GRAM++, IDRISI, Arc-Info, Arc Map ERDAS etc. However some of these software packages are too expensive for small programs or institutes and they do not provide specific options that enable rapid and uncomplicated analysis of biological diversity data.

The DIVA-GIS software allows analysis of gene bank and herbarium database to elucidate genetic, ecological and geographic pattern in the distribution of crops and wild species. DIVA-GIS helps improve data quality by assigning coordinates, using a large gazetteer. DIVA can also be used to check existing coordinates using overlays of the collecting site and administrative boundary databases. Maps can then be made of the collection sites. Analytical functions implemented in DIVA include mapping of richness and diversity, distribution of useful traits and location of areas with complementary diversity. DIVA can also extract climate data for all terrestrial locations, which can be used to describe the environment of collection sites

The most popular diversity index is *Shannon's diversity index* (SHDI) based on information theory (Shannon and Weaver 1949). The value of this index represents the amount of "information" per individual (or patch,). Information is a somewhat abstract mathematical concept. The absolute magnitude of Shannon's diversity index is not particularly meaningful; therefore, it is used as a relative index for comparing different landscapes or the same landscape at different times.

Shannon's diversity index

$H' = -\sum p_1 ln\ p_1$

n_1 = number of individual in the i-th class

p_1 =proportional abundance of the i-th class=n_1 / N

The Richness function offers two output variables; Number of different classes and number of observations. Number of different classes counts the different classes of a variable of (e.g. the different species names in a dataset covering a genepool) present in each grid cell. A variable could be selected from the gene bank database and then one or more values of those variables.

Species Richness is partially a function of scale. Larger areas are generally richer because there is generally greater heterogeneity than over comparable smaller areas. This contributes to the species-area relationship or the Species Richness. The number of species (or other object) observed in an area depends to some extent on the effort invested in recording there. Because a complete census is rarely feasible, in most cases only a sample of the area was surveyed.

Global Positioning System (GPS) is a very useful device to record the longitude and latitude of the places of collection of the sample. The data collected in this system easily could transfer to the computer system and get an overall idea of the collection site.Analysis of such eco-geographic patterns can make considerable contributions to several plant genetic resources research activities, including planning collecting programs, targeting genetic resources for breeding programs, developing core conditions, selecting and designing sites for *in situ* conservation and assessing the potential impact of the products arising from the use of plant genetic resources.

What are the main fields that could be explained with the help of GIS?

1. Bio-diversity

   a Diversity of plant &animal species of a given place.

   b. Diversity of a particular plant or animal species.

2. Geographical distribution of genetic variation and species migration.

3. It could help in Genetic resource conservation and predicting resource conservation.

4. On plotting the data on the map the cropping pattern and the crop suitability could be seen easily.

Documentation

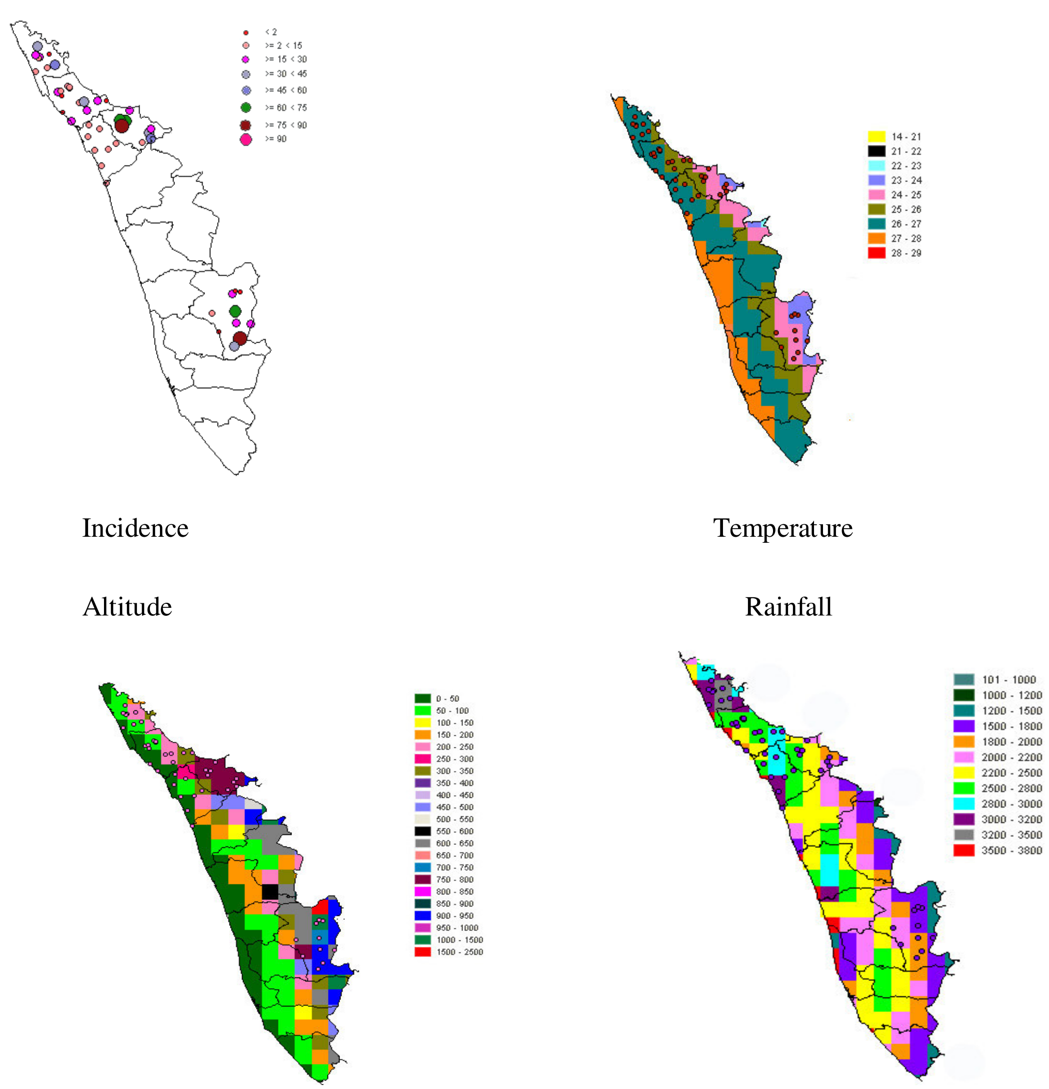
i) Fuller and more accurate passport data can be obtained. Using GIS, the collector would be able to record latitude, longitude, altitude, soil type, vegetation and other attributes for collecting sites automatically by overlaying their locations on different digitized base maps. The capability to generate isopleth maps, and in particular digital terrain models, allow the values of particular attributes (e.g., climatic factors) at collecting sites to be accurately estimated .(fig-1)

ii) GIS can assist in data verification-For example, a GIS can spot outliers due to mistake of latitude, longitude or other attribute data.

iii) GIS can produce listings of samples satisfying different criteria-All collecting sites with particular attributes (of the environment and/or germplasm) or combinations of attributes can be picked out. For example, a map of collecting site distribution may be overlaid on soil and rainfall maps to identify those collecting sites where the material collected might be expected to be adapted to particular combinations of edaphic and climatic conditions.

The biodiversity is not uniform but is heterogeneous with respect to a number of parameters. These are often restricted to three types; genetic diversity, species diversity and ecosystem diversity. The eco-system Diversity is very easy to understand with the help of GIS map. At a sight at the map it becomes clear what should be the optimum temperature and rainfall for the growth of any particular crop. Once these types of maps are made with certain collection, the prediction for the further collection becomes easy. For instance Fig.1 shows the collection site of stunted disease in pepper in Kerala along with the temperature, rainfall and altitude.

After plotting the collection places on the map clustering of the species also become easy. Important species with their special characters also could be shown. A better understanding of the geographical distribution of genetic variation within species is likely to increase sampling efficiency. In many instances, species exhibit a spatial structure of genetic variation.



Incidence

Temperature

Altitude

Rainfall

Fig-1. Showing percentages of stunted disease and the influence of temperature, rainfall and altitude on the impact of the disease

With the help of this process the species richness or the hot spot of genetic diversity becomes very easy to understand. G.Ravikanth *et al (*2002) tried to understand the “hotspot” species richness and genetic variability of population of rattan. They found that three regions, south Karnataka and south and north of Palghat gap were highly rich in rattan species. At the inter specific level population of the southern latitude and eastern longitude were characterized by high heterozygisity and therefore could be considered as“hot spot “of rattan genetic diversity. They had developed a contour map of rattan species richness and the hot spot of diversity. (Fig-2).

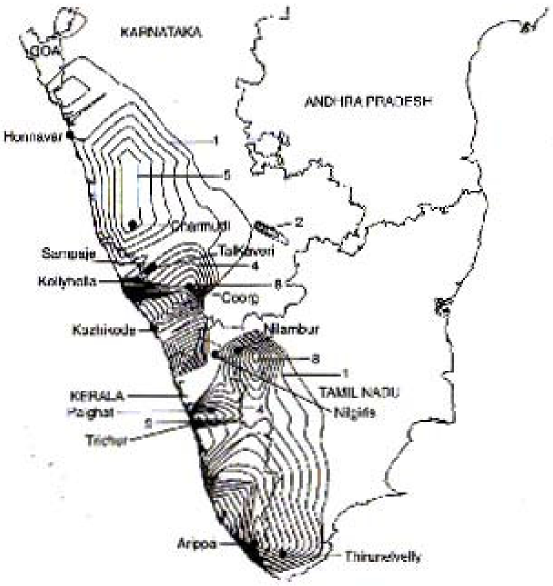


Fig. 2. Contour map of rattan species richness.

With the help of Dendogram and the spatial location on the map, the migration of the species and the climatic influence on the species and their variation could be clearly understood. Molecular markers, complemented by appropriate GIS software packages have been found useful in mapping the geographical distribution of genetic variation. The association of genetic diversity with environmental variables was assessed through hierarchical cluster analysis of climate data using Flora Map (Jones and Glakov1999).

Green et al (2002) examined the correlation between the distance matrices generated from morphological, RAPD marker and GIS-derived data .A total eight different GIS data matrices representing various combinations of GIS- derived attributes were compared against the matrices obtained from morphological and RAPD markers data. GIS is becoming an important tool in plant genetic resources conservation, including mapping the eco geographical distribution and predicting potential distribution of species using geo-referenced data from diverse sources. It has been recognized that the distribution of plants is primarily constrained by climate. This concept has underpinned many attempts to predict the potential distribution of individual species through correlation of climate with spatial distribution of individual species. In a study a combination of Flora Map (Jones Gladkov 1999) and multivariate statistical methods has been used as an appropriate GIS methodology to interpolate areas with the highest probability of species occurrence.

The diversity between populations is unrelated to geographic distance, which rules out simple isolation-by-distance models and suggests intense local differentiation. The analysis of diversity over sub-regions clearly indicated that the total amount of genetic diversity in a given set of germplasm accessions or geographic locations does not necessarily depend on the number of populations but more on the environmental conditions in which they are found or were collected.

The approach of using GIS as a tool for bio-diversity is not simply identifying the co-occurrence of conditions, but actually combining different kinds of data in some sort of mathematical model. A GIS can delimit areas where particular conditions or combinations of conditions are found. It can combine data sets to generate derived data sets of predictive or enhanced descriptive value. It can calculate climatic 'envelopes' for species, landraces or genotypes.

Sustainable agriculture and GIS

Productivity, equity, sustainability and stability have been identified as key goals of agricultural policies. The question of agricultural sustainability can be considered at a farm, a village and regional ,national or global level .There is a general concern that as agricultural production is increasing globally, and this gain of production are becoming less sustainable and further increase in production are becoming difficult to obtain because of the environmental constrains. In this context assessment of agricultural sustainability has become important issue for maintaining long term productivity, without damaging the natural resources base of agricultural production. GIS is an important tool for this measure.

Layman and Herdt (1985) proposed the definition of sustainability as: A sustainable system is one with non negative trends in measured out put" provides opportunity to develop quantitative measure to assess agricultural sustainability. Monteiln (1991) suggested the trends in per capita availability, say foodgrain in relation to trends in increasing population as an indicator of sustainable agriculture. This trend can be quantified by estimating the growth rate (%) in per-capita availability of food grains in a given period.

For example, the production of food grains per –capita is

$C=YA/P$

Where C is per capita availability of food grains, Y is the yield per hector, A is the area under food grains and P is the population of the region.

The fractional growth rate in per-capita availability of food grains is the sum of the fractional growth rate of productivity of food grains and area under food grains after subtracting the fractional growth rate of population. Thus prepared database could plot on the map and can get a visual idea of the sustainability of the required place and could be compared with the environmental set up as well as the suitability of crops.

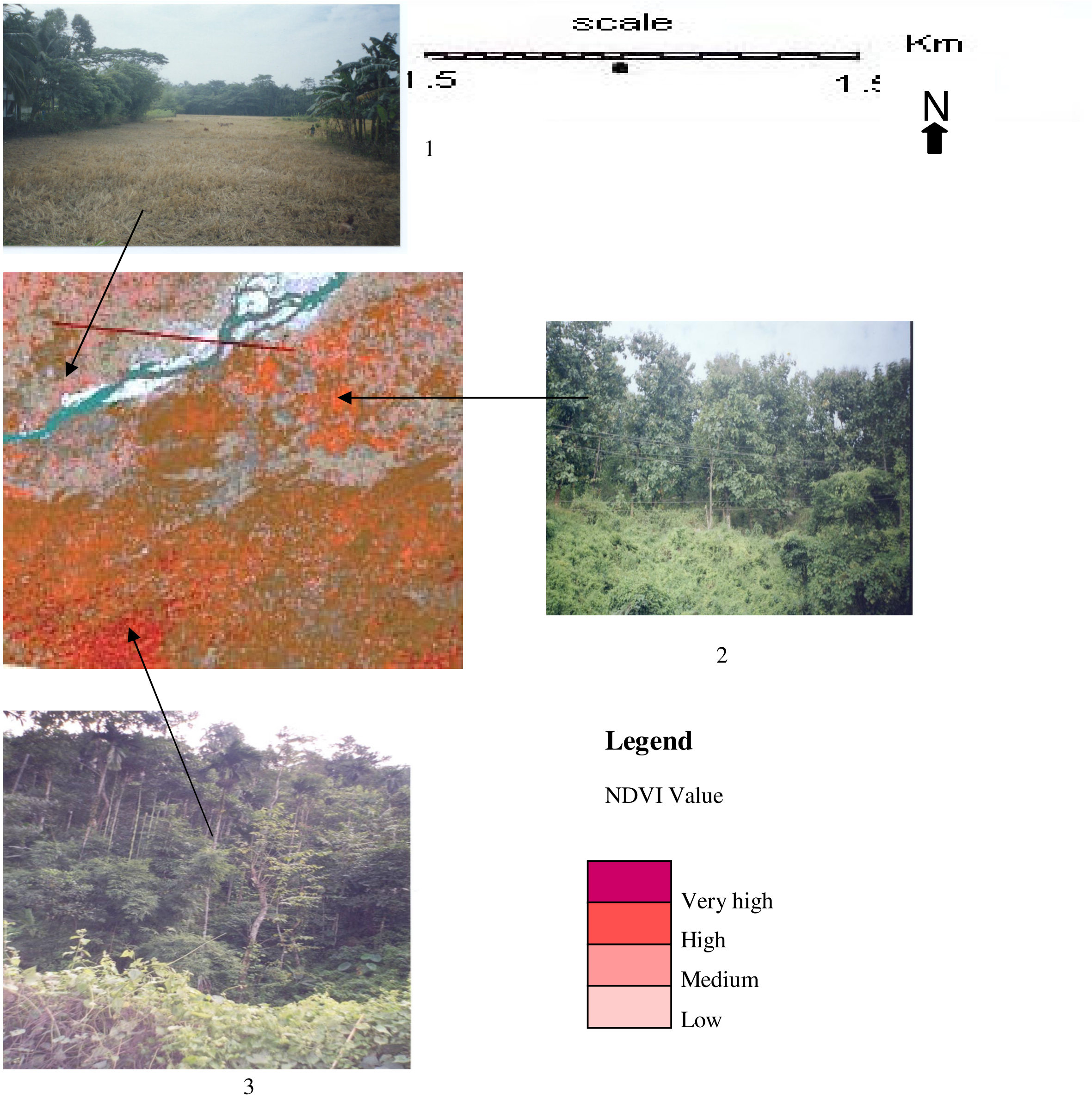
Concept of Remote Sensing

Remote Sensing is defined as the science and technology by which the characteristics of objects of interest can be identified, measured or analyzed the characteristics without direct contact. Electro-magnetic radiation which is reflected or emitted from an object is the usual source of remote sensing data. However any media such as gravity or magnetic fields can be utilized in remote sensing. A device to detect the electro-magnetic radiation reflected or emitted from an object is called a "remote sensor" or "sensor". Cameras or scanners are examples of remote sensors. A vehicle to carry the sensor is called a "platform". Aircraft or satellites are used as platforms. The technical term "remote sensing" was first used in the United States in the 1960's, and encompassed photogrammetry, photo-interpretation, photo-geology etc. Since Landsat-1, the first earth observation satellite was launched in 1972; remote sensing has become widely used.

The characteristics of an object can be determined; using reflected or emitted electro-magnetic radiation, from the object. That is, "each object has a unique and different characteristics of reflection or emission if the type of deject or the environmental condition is different.”Remote sensing is a technology to identify and understand the object or the environmental condition through the uniqueness of the reflection or emission

The land cover class vegetation can be made up of several land use classes. It can be natural vegetation, forestry or agricultural vegetation. Within these classes vegetation can be separated based on species, biomass, diseases and other. In all cases the most interesting data can be found in the visible, near infrared and middle infrared regions. Species distribution and health conditions are some examples of vegetation remote sensing applications (CEOS, 1995). Different surface types such as water, bare ground or vegetation reflects radiation differently in various channels. The radiation reflected as a function of the wavelength is called the spectral signature of the surface. Normalized vegetation Index NDVI is an index which helps to understand types of vegetation in the satellite image.

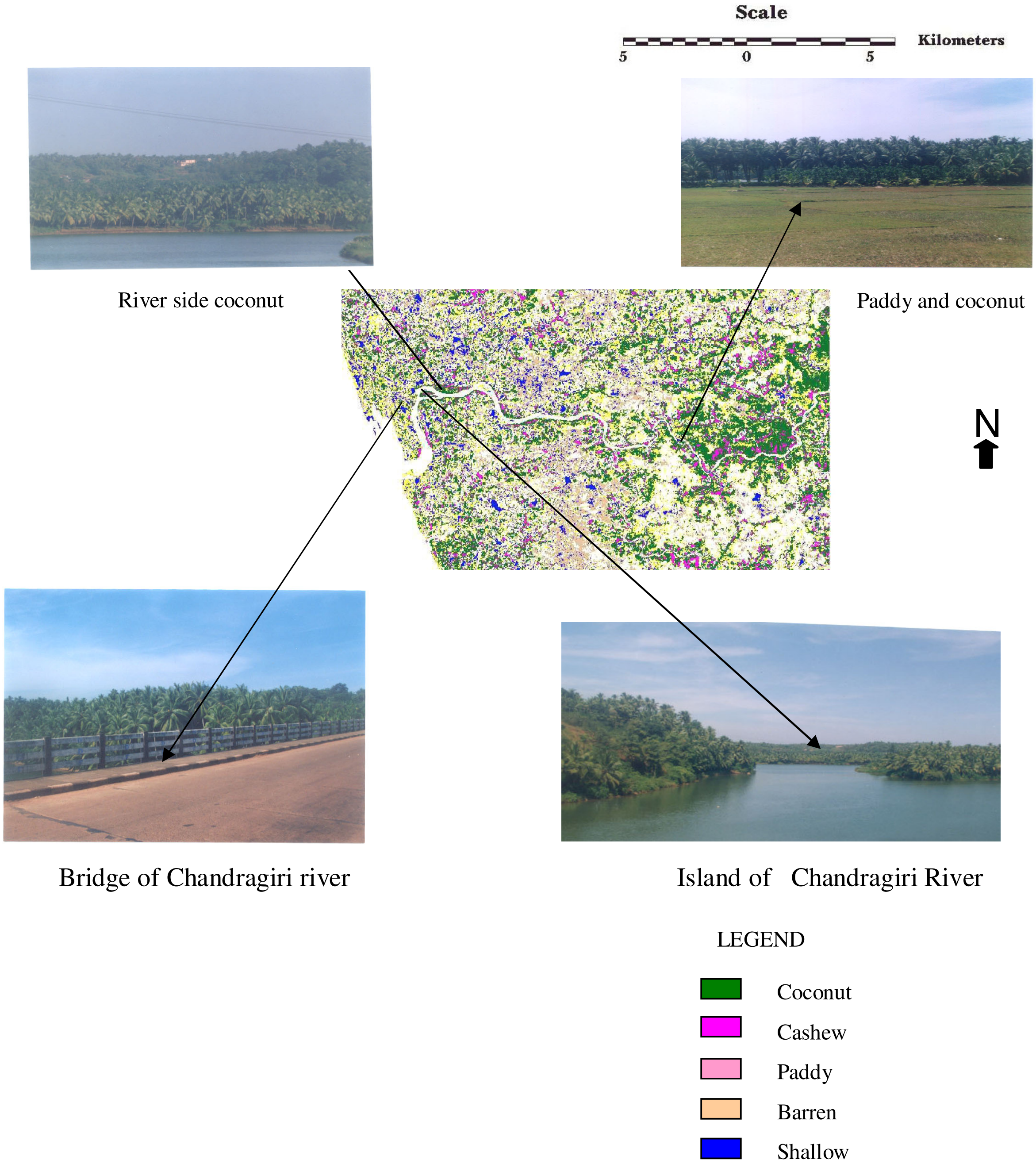
NDVI image and land truth for Kamrup district



NDVI image of a part of Kamrup district of Assam showing clear distinction of plantation forest and dense forest

1. Harvested paddy field with other vegetation; 2. Forest plantation; 3.Dense forest

Digital mosaic of a part of Kasaragod district with Ground Picture (Signature classified image)



River side coconut

Paddy and coconut

Bridge of Chandragiri river

Island of Chandragiri River

The FCC image has classified with the signature which gives a much better idea of the land truth and makes the survey easy

References

Engels JMM, Rao Ramanatha, Brown, A.H.D, Jackson, M.T. (2002). *Managing Plant Genetic Diversity, IPGRI Publication.*

Hijmans, R.J., Guarino., L, Cruz, M., and Rojas, R.(2001)*Computer Tools For Spatial Analysis of Plant Genetic Resource Data; 1. DIVA-GIS.* Plant genetic Resources Newsletter, 127.1 5-19.

Guarino, L.,, Ramanatha Rao,V. Reid R. (eds) (1995). *Collecting Plant Genetic Diversity.*IPGRI Publication

Singh,I.P, and Singh,S. (Ed).2003.*Exploration, Collection and Mapping of Citrus Genetic Diversity In India.*

Robert J. Hijmans, Luigi Guarino, Coen Bussink, Prem Mathur, Mariana Cruz, Israel Barrentes, and Edwin Rojas. (2003) *A geographic information system for the analysis of species distribution data. DIVA-GIS Manual.*

H.Resit.Akcakaya (1996) Linking GIS with Models. Third International Conference on Integrating GIS and Environmental Modeling, at Santa Fe, New Mexico, on January 23, .

Bioactive compounds from spices

N K Leela

Spices have attracted attention of mankind right from the early days. Spices are primarily used for flavoring and as food preservatives. But, these are now gaining importance for their diversified uses. The nutritional, anti-oxidant, anti-microbial and medicinal properties of spices have far-reaching implications. In the present scenario, the anti-diabetic, anti-hypercholesterolemic, anti-carcinogenic and anti-inflammatory effects of spices have paramount importance, as the key health issues of mankind nowadays are diabetes, cardio-vascular diseases, arthritis and cancer. Here comes the use of the components of spices as possible ameliorative or preventive agents for these health disorders.

Definition and classification

Bioactive compounds are compounds that interfere directly or indirectly with normal biological functions. These may be additive, synergistic or antagonistic in nature. The compounds that act against the natural biological processes are called antagonistic or inhibitory; those which enhance the normal biological functions are known as additive and those which add to the biological activity of other compounds are synergistic. We come across all these type of compounds in spices.

Black pepper

In *Ayurveda* pepper fruits are considered as acrid, bitter, anthelmintic, carminative, diuretic, digestive, stimulant and stomachic. They are useful in arthritis, asthma, fever, cough, dysentery, dyspepsia, flatulence, hiccough, haemorrhoids etc. In Indian home remedies, pepper decoction is used as a curative against cough and chest congestion.

Piperine is the major pungent principle of black pepper berries. It is an alkaloid. It has antioxidant and antiinflamamtory properties. An important use of piperine is that it is a bioavailability enhancer. Making use of this property of piperine Sabinsa Corporation has a standardized an extract from the fruits of black pepper and is traded as Bioperine®. It contains 95 percent of piperine. It is used to increase the bioavailability of nutritional compounds.Antiepilepsirine (1-(3-benzodioxol-5-yl)-1-oxo-2-propenyl-piperidide) is another interesting compound isolated from pepper that possesses potent antiepileptic properties. It is used in Chinese hospitals for the treatment of epilepsy.

Cardamom

Cardamom oil is an anaesthetic, antiseptic, antipharingitic, hepatotoxic and hypotensive. Major components of the oil are 1, 8- cineole and a- terpinyl acetate**.** 1, 8-cineole exhibits antibacterial activity against *Bacillus subtilis* and *Escherichia coli* and antifungal activity against various fungi. It possesses fumigant toxicity towards storage insects especially the Tribolium species. a- Terpinyl acetate, is present to the extent of 25-45% in cardamom oil. It has a fruity, floral odour. It is used as a semiochemical in the integrated management of pests.

Ginger

Ginger is used as a digestive aid and stimulant. It helps to relieve pain and reduce inflammation associated with arthritis, rheumatism and muscle spasms. Other uses are in the treatment of asthma, bronchitis, cough etc. These properties are due to its essential oil components and pungent principles. The main constituents of essential oil are sesquiterpenoids with (-)-zingiberene (~35%), as the main component. Gingerols are the pungent principles present in ginger rhizomes. Among gingerols, 6-gingerol is the most important. 6-gingerol and 6-shogaol, present in ginger rhizomes show antioxidant activity. These compounds also possess analgesic and hypotensive properties. Zingerone, a component of ginger is an antioxidant. Zingerone and dihydrogingerone isolated from ginger show antifungal activity against *Rhizoctonia solani* and *Candida albicans* respectively. *Ginger Soother* is a unique beverage made from ginger, honey, lemon and natural flavor. It is reported that a daily dose of this natural refreshing drink will help to keep one in harmony through out the day.

Turmeric

The medicinal properties of turmeric are due to the natural pigments namely curcuminoids present in the rhizomes. Three diarylheptanoids – curcumin, demethoxy curcumin, bis-demethoxy curcumin make up the yellow dye of turmeric rhizomes. The three curcuminoids present in turmeric are strong antioxidants. Another component reported from turmeric, a water-soluble peptide, namely turmerin, is a more potent anti-oxidant than curcuminoids and butylated hydroxy anisole (BHA).The antiinflammatory action of turmeric is also due to the curcuminoids. The eye drops prepared from the decoction of turmeric, known as *Haridra Eye Drops*, used in bacterial conjunctivitis demonstrates its anti-inflammatory activity.

Curcuminoids have been found to be cancer-preventing compounds in different tumor models. Curcumin also possesses anti-HIV activity and antiallergent properties. It is found to exhibit antithrombotic effect. It has been shown that curcumin at concentrations of 2.5-50 mg/100ml inhibited *in vitro* growth of *Staphylococcus aureus.* Curcuminoids consisting of curcumin, demethoxycurcumin, bisdemethoxycurcumin and cyclocurcumin have been shown to be effective against nematode parasites *in vitro.*

Garcinia

Hydroxy citric acid (HCA) is a major component in *Garcinia cambogia* fruit. 14-16% of fruit extract is hydroxy citric acid. Bioactivity studies have shown that HCA regulates lipid metabolism and is an ingredient in some weight loss products and dietary supplements. A number of products namely, citrin (50% HCA, Sabinsa Corporation), citrimax, Super citrimax (InterHealth), etc. are available in the market for treatment obesity, which contain HCA.

Essential oil components

Essential oils and their components have many functional properties and exert their action in human beings as well as in other organisms (insects, fungi, bacteria and viruses). Terpinene-4-ol, a constituent of turmeric leaf oil possesses antiseptic properties. Terpineol, a component of turmeric leaf oil possesses antiseptic properties. It shows antibacterial activity against *Bacillus subtilis*.

ar-Turmerone another component of turmeric oil has anticancer and antiinflammatory properties. It is shown to inhibit the lethal action of rat-snake venom. Turmeronol-A and turmeronol-B have potent antioxidant activity. Eugenol, the major component of clove oil, cinnamon leaf oil etc possesses appreciable antioxidant activity. It is fungitoxic to *Phytophthora capsici* and *Aspergillus flavus.*Cinnamaldehyde, the major constituent of cinnamon bark oil, possesses anti diabetic and antifungal properties. Nutmeg oil contains myristicin, elemicin and safrole which are known for their hallucinogenic effect. Likewise, wide spectrum of bioactivities possessed by many spice ingredients has been reported. Only a few of these are quoted here. The natural product industry is actively seeking natural therapeutics, preservatives, repellents and other agents that can replace synthetic compounds. Therefore many interesting products of this kind are likely to emerge and revolutionise the market in near future.

References

Cherdchu C, Karlsson E (1983) Proteolytic-independent cobra neurotoxin inhibiting activity of *Curcuma* Sp. (*Zingiberaceae*) *Southeast Asian J. Trop..Med. Pub.Health* 14, 176-178.

Ferreira L A F, Henriques O B, Andreoni A A S, Viotal G R F, Campos M M C, Habermehl G G, de Moraes V L G (1992) Antivenom and biological effects of ar-turmerone isolated from *Curcuma longa* (*Zingiberaceae*) Toxicon 30, 1211-1218.

Amman H P T, Wahl M A (1990) Pharmacology of *Curcuma longa, Planta Medica* 57, 1-7.

Kiso Y, Suzuki Y, Watanabe N, Oshima Y, Hikino H (1983) Antihepatotoxic principles of *Curcuma longa* rhizomes, *Planta Medica* 49, 185-187.

Srimal R C (1997) Turmeric: a brief review of medicinal properties, *Fitoterapia* 68, 483-493.

Tonnesen H H (1992) Chemistry of curcuminoids In: Ho C T, Lee Y, Huang M T (Eds.) Phenolic compounds in food and their effects on Health I. A C S symposium e 566, 143-153.

Chemical profiling of spices

T. John Zachariah

Spices and spice products have become dispensable as the most important natural flavoring materials available to the food processor. It provides individuality to otherwise insipid dishes and for adding tang and zest to beverages and appetizers. It has application in a variety of food and non food items including meat and fish, baked foods, sugar confectionery, beverages, tobacco, pharmaceuticals and fragrance.

Most of the herbs and spices are composed of three different types of ingredients

1. Volatile compounds - which can be recovered as essential oils by various distillation techniques which impart aroma
2. Non volatile components - which may be extracted as oleoresin or resinoids by using selected solvents which impart taste, flavour and pungency.
3. Inert matter which is mostly cellulose comprising the basic cellular plant structure and of value as natural carrier or diluent as well as contributing bulk or weight.

The major spices which we discuss here are black pepper, cardamom, ginger, turmeric, cinnamon, clove, nutmeg and allspice.

Black pepper

Black pepper and white pepper are the major products from pepper. Major countries involved in cultivation and trade of pepper are India, Indonesia, Sarawak (Malaysia), Brazil, Sri Lanka, Vietnam and China.

As in the case of other spices, the aroma, flavour and medicinal property of black pepper is attributed to the volatile oils present in black pepper. Pepper oil contains a range of organic compounds, which belong to the group of terpenes and hydrocarbons. The major aroma compounds of black pepper are β-caryophyllene, sabinene, myrcene, limonene and linalool.

The non-volatile part of black pepper which is extracted using organic solvents is called black pepper oleoresin. The major pungent alkaloid present in black pepper is piperine. In general, black pepper contains about 3-5% volatile oil and 8-16% oleoresin and 2-6% piperine.

However, starch, is the predominant constituent ranging from 35% to 40% in black pepper and 53% to 58% in white pepper. Some of the traditional cultivars like Kottanadan and Kumbhakodi are rich in oleoresin and piperine. Some of the new varieties with high quality are Malabar Excel, Sreekara and Subhakara.

Factors influencing the chemical profile

Volatile oil content of pepper reaches a maximum at a relatively early stage, about four and a half months after fruit setting and then diminishes, while piperine content continues to increase for a period. (Piperine, major alkaloid of pepper, which is responsible for its pungency constitute about 35% of the oleoresin of pepper). Fully matured pepper is harvested and dried to a moisture level of 10-12%. Dried pepper is stored in airtight containers to prevent moisture absorption and mould attack during storage. Bulk density of dried pepper (weight per litre) varies from 450gm to 650 gm. Pepper harvested at tender stage is used for various products like pepper in brine, canned pepper, dehydrated green pepper etc while fully matured pepper is used for preparing pepper powder, oil, oleoresin and white pepper.

Pepper oleoresin

Extraction of black pepper powder with organic solvents like acetone, ethylene dichloride etc provide a resinous extract called oleoresin possessing the odour, flavour and pungent principles of the spice. The organoleptic property of the oleoresin is determined by its volatile oil and piperine contents and the abundance of these components depend on the raw material used for extraction. Good quality oleoresin contains 15-20% volatile oil and 35-55% piperine. It is generally claimed that 1kg of oleoresin obtained from 8 kg of black pepper, when suitably dispersed on an inert base, can replace up to 25 kg of the spice for flavouring purpose.

Extraction of oleoresin may be done by different methods by circulation of hot or cold solvent through a bed of ground spice, or a method based on Soxhlet extraction, or by a counter current process using several extractors. In counter current process, the pepper remains in the extractors, and the solvent is transferred from one extractor to the next during which time, the concentration of the material extracted increases, until the solvent meets fresh pepper, after which it is passed on to a still where most of the solvent is removed. The concentrated extract is then subjected to carefully controlled vacuum distillation in which the solvent in the extract is reduced to trace levels. Normally pepper yields about 8-15% oleoresin depending upon the variety and maturity at harvest.

Black pepper oil

Dried pepper is crushed to a coarse powder and on steam distillation, yields 2-4% colourless or pale green essential oil.

Pepper oil is used in perfumery and in flavorings. In certain parts of world, pepper oil is distilled from siftings and less- attractive berries and very often the oil is of good quality. Two other grades are produced by fractionation - light oil or forerun and heavy oil, consisting of the higher fractions of the oil.

The oil from white pepper can be, distilled in much the same way, but the limiting factors are the higher cost of the raw material and the lower oil yield - some oil is lost when the mesocarp is removed.

Super critical extraction

This is a process for extracting spice to obtain oleoresin and fractionating the oleoresin into fixed and essential oil components. It uses CO_2 and high pressure to separate out (or "Fractionate") all of the natural essence of plant, spice, bean, etc. without heat or chemical solvents to do it.

The extraction is performed on the ground spice using supercritical fluid carbon dioxide under a pressure of about 400 bar to about 600 bar and at a temperature of about 80^0 C. to about 120^0 C. A cascading continuous extraction method is preferred. Oleoresin fractions are precipitated from the supercritical fluid at reduced pressures in the ranges 280-380 bar and 100-200 bar at 80^0-100^0. C., while a last fraction is obtained from the non-critical fluid at a pressure of 30-50 bar and a temperature of 0-30^0 C.

Chemistry

The pepper berry contains steam volatile oil, fixed (fatty) oil, alkaloids, resin, proteins, cellulose, pentosans, starch, minerals etc. About 25% of the dry berry weight of black pepper is comprised of the outer skin, which consists mainly of fibre and some of the essential oil cells. This outer skin is removed during preparation of white pepper and the general oil composition of white pepper differs principally from black pepper in that it contains less fibre (3-5% compared to 10-15%) and more of starch (55-60% compared to 34-48%). General composition of dried green pepper is similar to that of black pepper.

Dried pepper contains 2-5% volatile oil and the value for black pepper is generally higher than that for white pepper. A commercial consignment of black pepper generally has 0.6-2.6% oil while that of white pepper is usually 0.5-1.0%. The pungent alkaloid principle piperine content range from 3-9% in pepper berries. HPLC method of estimating piperine is more realistic as the UV method give absorption value of all pigments. Pepper harvested before maturity give higher value for oil, oleoresin and piperine.

Changes in volatile oil and piperine content are usually of a small order of magnitude during the conversion of fresh green pepper to dried black pepper by the traditional sun-drying process. However, in the preparation of white pepper there is an inevitable loss of some of the volatile oil during decorticating process. Losses of volatile oil can also be encountered during artificial drying of both black and white pepper if care is not exercised in control of the drying temperature.

The organoleptic properties of the spice and oleoresin can deteriorate during common forms of utilization involving heat treatment. Gas chromatographic analysis revealed that major qualitative changes occur to volatile oil upon heating, and the monoterpene hydrocarbons are particularly susceptible.

Chemistry of volatile oil

The aroma and flavour of pepper is determined by the composition of steam volatile oil. It is composed of monoterpene hydrocarbons with smaller amounts of sesquiterpene hydrocarbons. Steam distilled black pepper oil contain about 70-80% monoterpene hydrocarbons, 20-30% sesquiterpene hydrocarbons and less than 4% oxygenated compounds. The oil obtained by vacuum distillation of oleoresin extract are considered to be closer in composition, aroma and flavour to the natural oil present in the spice than those obtained by steam distillation. White pepper oil is qualitatively and quantitatively similar to black pepper oil.

Major chemical constituents of oil

- *Monoterpene hydrocarbons:*α-Pinene, β-pinene, sabinene, limonene, phellandrene, α-terpinene
- *Sesquiterpene hydrocarbons:*β-Caryophyllene, bisabolene, bergamotene, humulene, isocaryophyllene
- *Oxygenated monoterpenes:*Borneol, camphor, linalool, α-terpineol

Cultivar specific profiling

Four major cultivars of black pepper from Kerala viz. Aimpiriyan, Narayakodi Neelamundi and Uthirankotta were analyzed for a period of three consecutive seasons for their oil composition. Fifty-five compounds were identified in the oils by GC and GC/MS. The main components of Aimpiriyan oil were limonene (19.8%-22.5%), β-pinene (9.3-23.9%), and β-caryophyllene (20.3-34.7%). Narayakodi oil contained as major compounds sabinene (4.4-24.6 %) limonene (9.5-19.5%), β-pinene (4.8-15.6%), β-caryophyllene (29.8-52.9%) and caryophyllene oxide (2.3-3.9%). Neelamundi oil contained α-pinene (4.7-6.5%), sabinene (23.2-27.3%), β-pinene (7.8-11.3%), limonene (12.9-18.6%) and β-caryophyllene (17.0-31.0%). The oil from Uthirankotta contained α-pinene (9.1-14.6%), β-pinene (9.3-12.5%), δ-3-carene (6.7-8.5%), limonene (13.3-19.5%), β-caryophyllene (25.1-37.8%) and caryophyllene oxide (0.6-2.7 %)

Aroma of pepper

Utilising present day technology like solid-phase microextraction (SPME) the odorous target components responsible for the characteristic odor of these valuable spices and food flavoring products were identified. By means of GC-flame ionization detection (FID) and GC-MS (using different polar columns) the main compounds of the SPME headspace samples of *P. nigrum* (black) and *P. guineense* (black and white) were found to be *P. nigrum* (black)-germacrene D (11.01%), limonene (10.26%), β-pinene (10.02%), α-phellandrene (8.56%), β-caryophyllene (7.29%), α-pinene (6.40%) and cis-β-ocimene (3.19%); *P. guineense* (black)-β-caryophyllene (57.59%), β-elemene (5.10%), bicyclogermacrene (5.05%) and α-humulene (4.86%); and *P. guineense* (white)-β-caryophyllene (51.75%), cis-β-ocimene (6.61%), limonene (5.88%), β -pinene (4.56%), linalool (3.97%) and α -humulene (3.29%).

The pungent principles

The most pungent alkaloid piperine was first isolated by Oersted (1820) and its structure was later identified as the trans - trans form of 1-piperoyl piperidine. It was revealed later that piperine is not the only pungent principle of pepper. Chavicine is another alkaloid. Piperine is undoubtedly the major pungent principle in pepper, which probably comprises over 95% of the total pungent alkaloids present, and the determination of its content is a good measure of the pungency of the spice.

Ginger

Ginger rhizome contains a little volatile oil, fixed (fatty) oil, pungent compounds, resin, proteins, cellulose, pentosans, starch and mineral elements. Starch is the most abundant and comprises 40-60% of rhizome on dry weight basis.

The crude fibre, volatile oil contents and the pungency level are most important criteria in assessing suitability of ginger for various purposes. The relative abundance of these constituents in the fresh rhizome is governed by its state of maturity at harvest. Tender ginger harvested at 5 to 7 months after planting are ideal for the manufacture of preserved ginger since the fibre content is less in that. At about 9 months after planting, the volatile oil and pungent principles reach a maximum and thereafter their relative abundance falls as the fibre content continues to increase.

Most ginger varieties have about 3-6% crude fibre, 1.5-3.0% volatile oil and 4-10% oleoresin.The major pungent principle in ginger oleoresin is gingerol. Among the gingerols 6-gingerol, 8-gingerol and 10-gingerol are very important in contributing the pungency. Gingerol on storage gets converted to another compound called shogaol. The age of ginger oleoresin can be assessed based on the shogaol content.

Gingerol are susceptible to chemical transformation to less pungent degradation products, these degradations can occur by poor handling during the preparation, storage and utilization of dried ginger and its oleoresin with consequent deterioration of quality.

The level of gingerols and shogaols in ginger oleoresin can be determined using vanillin standard by HPLC. Good quality ginger oleoresin will have 22-25% total gingerol. Using HPLC method ginger oleoresin can be fractionated into 6-gingerol, 8-gingerol, 10-gingerol, 6-shogaol, 8-shogaol and 10-shogaol.

Volatile oil

The aroma and flavour of ginger are determined by the composition of steam volatile oil, which is comprised, mainly of sesquiterpene hydrocarbons, monoterpene hydrocarbons and oxygenated compounds. The monoterpene constituents are believed to contribute to the aroma of ginger and they are relatively more abundant in the fresh rhizome rather than that of dry ginger. The major sesquiterpene hydrocarbon constituent of the oil is α-zingiberene. Other major constituents are sesquiphellandrene and ar-curcumene.

Scientists have identified seventy-two components in the volatile oil extracted from the air-dried rhizomes. The main components were α-zingiberene (21.8%), geranial (9.9%), geraniol (9.4%), β-bisabolene (7.9%), nerol (7.1%), 1,8-cineol (6.2%), α-terpineol (5.6%), borneol (5.4%), β-phellandrene (3.1%), linalool (1.7%), methyl nonyl ketone (1.6%) and camphene (1.4%); the other components accounted for ~1% each of the volatile oil.

Turmeric

It is an essential item in curry powder. Curry powder usually contains about 24% of turmeric powder and may only be exceeded in quantity by coriander seeds. Turmeric is valued for its yellow- orange colouring powder, which is determined by the content of curcuminoid pigments. Turmeric also possesses highly aromatic volatile oil, which has the peculiar turmeric aroma.

Curcumin the colouring principle is composed of three component pigments viz., 1) Curcumin 2) De methoxy curcumin and 3) bis-demethoxy curcumin.

Curcuminoids are polyphenolic pigments found in the spice turmeric. The term turmeric is used both for the plant *Curcuma longa* L. and the spice derived from the rhizomes of the plant. The major curcuminoids are curcumin, demethoxycurcumin and bisdemethoxycurcumin. These substances comprise 3 to 6% of *Curcuma longa*. Curcumin makes up 70 to 75% of the curcuminoids, demethoxycurcumin 15 to 20% and bisdemethoxycurcumin about 3%.

Curcuminoids are responsible for the yellow color of turmeric, as well as the yellow color of curry. Curcuminoids are derived from turmeric by extraction with ethanol. Curcumin is the most studied of the curcuminoids. In pure form, it is an orange-yellow, crystalline powder that is insoluble in water. It is also known as diferuloylmethane and turmeric yellow. Its chemical name is (*E*, *E*) -1, 7-bis (4-hydroxy-3- methoxyphenyl)-1, 6-heptadiene-3, 5 dione. The molecular formula of curcumin is $C_{21}H_{20}O_6$ and its molecular weight is 368.39 daltons.

The curcuminoids have been found to have a number of antioxidant activities, including scavenging of such reactive oxygen species as superoxide anions and hydrogen peroxide, inhibition of lipid peroxidation and inhibition of the oxidation of low-density lipoprotein (LDL). The reduced derivative of curcumin, tetrahydrocurcumin, has been found to have even stronger antioxidant activity.

The possible anticarcinogenic activity of curcumin and the other curcuminoids may be accounted for by a few mechanisms. These include inhibition of angiogenesis, upregulation of apoptosis, interference with certain signal transduction pathways that are critical for cell growth and proliferation, inhibition of colonic mucosa cyclooxygenase (COX) and lipoxygenase (LOX) activities and inhibition of farnesyl protein transferase. In addition to its possible activity in preventing malignant transformation and inhibiting tumor growth, curcumin may have antimetastatic potential, as well.

Curcumin has modest anti HIV-1 activity. It has been found to inhibit HIV-1 and HIV-2 proteases, HIV-1 LTR (long terminal repeat)-directed gene expression, Tat-mediated transactivation of HIV-1-LTR and HIV-1 integrase. All of these actions have been demonstrated *in vitro*. There is no evidence that curcumin or the other curcuminoids significantly inhibit the replication of HIV-1 *in vivo*. The mechanism of the possible hypocholesterolemic effect of the curcuminoids is unclear.

Cardamom

The characteristic odour and flavour of true cardamom is determined by the composition of its aromatic steam volatile oil. The main chemical components of cardamom oil are α-pinene, β-pinene, sabinene, myrcene, α-phellandrene, limonene, 1,8-cineole, γ-terpinene, p-cymene, terpinolene, linalool, linalyl acetate, terpinen-4-oil, α-terpineol, α-terpinyl acetate, citronellol, nerol, geraniol, methyl eugenol and trans-nerolidol.

Two constituents of significance in cardamom oil are 1,8-cineole and α-terpinyl acetate. High 1,8-cineole in the oil gives more of a camphoraceous note and high terpinyl acetate give more sweet, fruity note.

Paprika

Paprika is defined in the United States as a sweet, dried, red powder. This mild powder can be made from any type of *C. annuum* that is non-pungent and has brilliant red color. Paprika may be pungent in Hungary, but paprika is always non-pungent in international trade. Paprika comes from milling dry fruits of different varieties of *Capsicum annum* L. In Europe, it is principally produced in Hungary, Turkey and Spain. Spanish powdered paprika is again classified based on its pungency in to three categories such as hot ('Picante'), sweet ('dulce') and an intermediate one called 'ocal' or agridulce.

The paprika type chilli which was evolved in the temperate regions around Mediterranean and some parts of U.S is widely used as a table spice and also in the meat processing industry as a natural colorant. This is valued principally for the brilliant red colour it gives to pale foods and also for its delicate aroma. Color is very important in paprika and chilli powder. Paprika and paprika oleoresin is currently used in a wide assortment of foods, drugs, and cosmetics, as well as for improving the feather color of flamingoes in zoos. Technological innovations have already made great strides in separating colour components and pungent constituents.

Paprika is in some cases classified as non-pungent, even though it may contain low or high levels of pungent compounds. There are many hot pungent types that vary in pungency; for example New Mexico, Jalapeno, Cayenne etc. Thai and Habanero types increase in pungency in this order. Chillies are generally smaller and lower in red colour than non-pungent fruit. Paprika like chillies and capsicums is always a ground product. Oleoresin paprika is prepared from varieties of *C.annuum* L. from which paprika is produced.

The colour of chilli spice powder is due to the presence of red-pigmented carotenoids. The main pigments are Capsanthin, capsorubin, zeaxanthin and cryptoxanthin. Carotenoids are very stable in intact plant tissue.

Carotenoids control pod color with approximately 20 carotenoids contributing to the color of the powder. Carotenoid compounds are yellow to red pigments of aliphatic or alicyclic structures composed of isoprene units, which are normally fat-soluble colors. The keto-carotenoids, capsanthin, capsorubin, and cryptocapsin are unique *Capsicum* carotenoids. The major red color in chilli comes from the carotenoids capsanthin and capsorubin, while the yellow-orange color is from beta-carotene and violaxanthin. Capsanthin, the major carotenoid in ripe fruits, contributes up to 60% of the total carotenoids. Capsanthin and capsorubin increase proportionally with advanced stages of ripeness; with capsanthin being the more stable of the two. The amount of carotenoids in fruit tissue depends on factors such as cultivar, maturity stage, and growing conditions In ripe fruits, capsanthin, capsorubin, zeaxanthin, cucurbitaxanthin A and beta-carotene were the main carotenoids, the remainder being capsanthin 5,6-epoxide, capsanthin 3,6-epoxide, karpoxanthin, cucurbitaxanthin B, violaxanthin, cycloviolaxanthin, antheraxanthin, capsanthone, nigroxanthin, beta-cryptoxanthin and several cis isomers and furanoid oxides.

Pungency in chilli is produced by the capsaicinoids, group of alkaloid compounds that are found only in the plant genus, *Capsicum*. The nature of the pungency has been established as a mixture of seven homologous branched-chain alkyl vanillylamides. They often are called capsaicin after the most prevalent one. Dihydrocapsaicin is usually the second most prevalent capsaicinoid, while the other five compounds, nor capsaicin, nordihydrocapsaicin, nornordihydrocapsaicin, homocapsaicin and homodihydrocapsaicin are considered minor capsaicinoids because of their relative low abundance in most natural products. Capsaicin is a powerful and stable alkaloid that can be detected by human taste buds in solutions of ten parts per million. Capsaicin's composition ($C_{18}H_{27}NO_3$) is similar to piperine ($C_{17}H_{19}NO_3$) that gives black pepper its bite.

The capsaicinoids present in the capsicum fruit are predominantly capsaicin and dihydrocapsaicin, making up 80 to 90%. Dihydrocapsaicin accounts for about 22% of the total capsaicinoids mixture and has about the same pungency as capsaicin. .Capsaicinoids is mainly ingested as naturally occurring pungency-producing components of Capsicum species (chilli, cayenne pepper, red pepper). They typically range from 0.1 mg/g in chilli pepper to 2.5 mg/g in red pepper and 60 mg/g in oleoresin red pepper.

Functional Foods and Nutraceuticals

Shamina Azeez

The effect of dietary factors on health status has been recognized since antiquity. The name nutraceuticals was coined by Dr. Stephen Defelice in 1989. Many such nutraceuticals are phytonutrients. While nutraceuticals, sometimes referred to as natural health products, are often used in medicinal forms as tablets, capsules or liquid in nutrient premixes or nutrient systems in the food and pharmaceutical industries, functional foods are those that resemble the traditional food, as such, but confer benefits beyond their nutritional role.

Definitions

A ***functional food*** is similar in appearance to, or may be, a conventional food that is consumed as part of a usual diet, and is demonstrated to have physiological benefits and/or reduce the risk of chronic disease beyond basic nutritional functions, i.e. they contain bioactive compounds.

A ***nutraceutical*** is a product isolated or purified from foods that is generally sold in medicinal forms not usually associated with foods. Nutraceuticals are natural, bioactive chemical compounds that have health promoting, disease preventing or medicinal properties *(http://foodsci.rutgers.edu)*Although some of the dietary phytochemicals and bioactives have traditionally been considered as "antinutrients", their beneficial effects in human health promotion and disease prevention has recently been acknowledged. Examples of claims made for nutraceuticals are red wine (resveratrol) as an antioxidant and an anticholesteremic, broccoli (sulforaphane) as a cancer preventative, and soy and clover (isoflavonoids) to improve arterial health in women. Several nutraceuticals are known. Some examples are flavanoids, Antioxidants such as γ-linolenic acid, β-carotenes, anthocyanins, etc. With the Dietary Supplement Health and Education Act (DSHEA), USA, several other compounds were added to the list of supplements originally mentioned in FDA notification. Thus many botanical and herbal extracts such as Ginseng, garlic oil, etc., have found glory as nutraceuticals.
Among diseases of concern, cancer and coronary heart disease (CHD) are high on the list. In this respect, phenolics of plant origin, as an example, have been found to act as free radical scavengers, inhibitors of cholesterol oxidation and DNA breakage, among others, thus serving as potential cancer preventing agents. The type of phenolics involved depends on the species as well as other variables. On the other hand, marine foods have often been considered as "heart food" because of the role of their ώ-3-fatty acid constituents in lowering of triacylglycerol and cholesterol levels and hence the incidence of CHD. (Shahidi. F. Nutraceuticals and Functional Foods in Health Promotion and Disease Prevention. ISHS Acta Horticulturae 680: III WOCMAP Congress on Medicinal and Aromatic Plants - Volume 6: Traditional Medicine and Nutraceuticals).

Table 1. Examples of Functional Food Components *(Source: International Food Information Council)*

Functional components	Source	Potential benefits
Carotenoids		
Alpha-carotene β-carotene	Carrots, Fruits, Vegetables	Neutralize free radicals, which may cause damage to cells
Lutein	Green vegetables	Reduce the risk of macular degeneration
Lycopene	Tomato products (ketchup, sauces)	Reduce the risk of prostate cancer

Functional components	**Source**	**Potential benefits**
Dietary Fibre		
Insoluble Fibre	Wheat Bran	Reduce risk of breast or colon cancer
β-Glucan	Oats, barley	Reduce risk of cardiovascular disease. Protect against heart disease and some cancers; lower LDL and total cholesterol
Soluble Fibre	Psyllium	
Fatty Acids		
Long chain ώ-3 Fatty Acids-DHA/EPA	Salmon and other fish oils	Reduce risk of cardiovascular disease. Improve mental, visual functions
Conjugated Linoleic Acid (CLA)	Cheese, meat products	Improve body composition. Decrease risk of certain cancers
Phenolics		
Anthocyanidins	Fruits	Neutralize free radicals; reduce risk of cancer
Catechins	Tea	
Flavonones	Citrus	
Flavones	Fruits/vegetables	
Lignans	Flax, rye, vegetables	Prevention of cancer, renal failure
Tannins (proanthocyanidines)	Cranberries, cranberry products, cocoa, chocolate	Improve urinary tract health. Reduce risk of cardiovascular disease
Plant Sterols		
Stanol ester	Corn, soy, wheat, wood oils	Lower blood cholesterol levels by inhibiting cholesterol absorption
Prebiotics/Probiotics		
Fructo-oligosaccharides (FOS)	Jerusalem artichokes, shallots, onion powder	Improve quality of intestinal microflora; gastrointestinal health
Lactobacillus	Yogurt, Other dairy	
Soy Phytoestrogens		
Isoflavones: Daidzein Genistein	Soybeans and soy-based foods	Menopause symptoms, such as hot flashes Protect against heart disease and some cancers; lower LDL and total cholesterol

Categories of Functional Foods/Nutraceuticals (*Source: Shambrock Consulting Group Inc. and Kelwin Management Consulting)*

Basic Foods: Carrots (containing the natural level of the antioxidant β-carotene). Foods - oat bran cereal (containing the natural level of β-glucan)

Processed Foods with Added Ingredients: Calcium-enriched fruit juice; Foods enhanced to have more of a functional component (via traditional breeding, special livestock feeding or genetic engineering); Tomatoes with higher levels of lycopene (an antioxidant carotenoid); Oat bran with higher levels of β-glucan; Eggs with ώ-3 from flax.

Isolated, purified preparations of active food ingredients (dosage form): Isoflavones from soy; ώ-3 from fish oils (DHA and ALA).

Other Definitions *(Source: Shambrock Consulting Group Inc. and Kelwin Management Consulting).*

Bioactive Compounds are naturally occurring chemical compounds contained in or derived from a plant, animal or marine source, that exert the desired health/wellness benefit (e.g. ώ-3 fatty acids in flax or fish oils and β-glucans from oats and barley).

Functional Ingredients are the standardized and characterized preparations, fractions or extracts containing bioactive compounds of varying purity, that are used as ingredients by manufacturers in the food (human and pet) preparations, fractions or extracts containing bioactive compounds of varying purity, which are used as ingredients by manufacturers in the cosmetics and pharmaceutical sectors. Examples include probiotics, prebiotics, vitamin enriched drinks and food.

Herbs are leaves, roots and flowers of plants grown and processed for culinary, cosmetic, industrial, medicinal, landscaping, decorative and fragrant purposes. Much of the early interest in functional foods and nutraceuticals was based on the medicinal uses of herbs. Examples: mint, thyme, saffron etc.

Industrial Ingredients are the standardized and characterized preparations, fractions or extracts of agri-commodities of varying purity, that are used as ingredients by manufacturers of non-food products. Example: malt.

Natural Health Products (NHP) includes homeopathic preparations; substances used in traditional medicines; minerals or trace elements; vitamins; amino acid; essential fatty acids; or other botanical, or animal or microorganism derived substances. These products are generally sold in medicinal or "dosage" form to diagnose, treat, or prevent disease; restore or correct function; or to maintain or promote health. As a product group, NHPs include nutraceuticals.

Novel Foods are defined by Health Canada as products that have never been used as food, foods that result from a process that has not previously been used for food, or, foods that have been modified by genetic manipulation (genetically modified foods).

Spices are seeds, root, bark and flowers of plants that are grown, harvested and processed for use as food or beverage flavoring. Examples include caraway, coriander, dill and mustard. Recently there has been interest in bioactive compounds identified in spices.

Traditional Food Ingredients are the standardized and characterized preparations, fractions or extracts of agri-commodities of varying purity, that originate from plant, animal or marine sources and are used as ingredients by manufacturers in the food (human and pets) and NHP sectors. They are used for a variety of reasons in food products including consistency, adding flavor or coloring, modifying texture or stabilizing mixtures. They are not marketed on the basis of any bioactive compounds that they may contain.

Traditional Processed Foods are the conventional foods that have been manufactured by the traditional food processing industry and sold to the public through established distribution systems for generations. These foods are consumed as part of a usual diet and are consumed primarily for basic nutritional purposes. Examples are processed meats, bottled fruit juice, yogurt, and breakfast cereal. They are not marketed on the basis of any bioactive compounds that they may contain.

Traditional Whole Foods are the conventional foods that have been grown by agricultural producers for generations. They are subjected to minimal processing (e.g. sorting, cleaning and bulk packaging) before being sold to the public through established distribution systems. These foods are consumed as part of a usual diet and are consumed primarily for basic nutritional purposes. Examples include raw vegetables (broccoli and spinach), fresh fruit (blueberries and apples) and fresh meat (tray packed beef, pork or chicken).

Spices

Spices are esoteric food adjuncts that have been used as flavoring and coloring agents and as preservatives for thousands of years. Spices have also been recognized to possess medicinal properties and their use in traditional systems of medicine has been on record for a long time. With the advancement in the technology of spices and on knowledge of the chemistry and pharmacology of their active principles, their health benefit effects were investigated more thoroughly in recent decades. Much health benefit attributes of these common food adjuncts have been recognized in the past few decades by pioneering experimental research involving both animal studies and human trials. These include digestive stimulant action, hypolipidemic effect, antidiabetic influence, antilithogenic property, antioxidant potential, anti-inflammatory property, antimutagenic, and anticarcinogenic potential of spices. Among these, the hypocholesterolemic and antioxidant properties of a few specific spices have far-reaching nutraceutical value. These beneficial physiological effects also have the potential of possible therapeutic application in a variety of disease conditions (Srinivasan, K. 2005. Role of Spices Beyond Food Flavoring: Nutraceuticals with Multiple Health Effects. *Food Reviews International*. 21(2): 167 – 188.)

In recent years there has been increased awareness about the detrimental effects of synthetic food supplements and food additives and the search for natural alternatives in plants has intensified. In this regard, spices appeal more to consumers, not only for the flavor and aroma they impart, but also to their diverse nutraceutical effects, listed above. Most of these properties are attributed to the secondary metabolites – the essential oils and oleoresins – present in spices, and of which a good deal of studies have been undertaken in India and abroad. The various phytochemicals include flavonoids, terpenoid, lignans, sulfides, polyphenolics, carotenoids, coumarins, saponins, plant sterols, curcumins, phthalides

Table 2. The major secondary metabolites identified in spices.

Black pepper	Piperine
Ginger	Zingiberene, AR-curcumene, Farnesene, Gingerole, Zingerone
Turmeric	Curcuminoids (curcumin, demethoxycurcumin and bisdemethoxycurcumin)
Cinnamon	Cinnamic acid, Cinnamaldehyde
Clove	Eugenol, Acetyl eugenol
Nutmeg, mace	Pinenes, Terpinene, Limonene, Myrcene, Linalool, Terpineol, Safrol, Myristicine
Allspice	Eugenol

The medicinal value of spices has been exploited by the traditional systems of medicine like Ayurveda, folklore medicines and 'kitchen remedies' since time immemorial. Some of the reported pharmaceutical effects of spices are given in the table below:

Table 3. Some documented pharmaceutical effects of spices.

Cancer preventive	Ginger, Black pepper, Nutmeg, Cinnamon, Clove, Turmeric, Cardamom, Vanilla, Allspice, Mace
Antimicrobial	Ginger, Nutmeg, Black pepper, Cinnamon, Vanilla, Turmeric, Clove, Allspice, Cardamom, Mace
Anti-inflammatory	Black pepper, Cinnamon, Clove, Turmeric, Allspice, Cardamom
Spasmolytic	Cinnamon, Black pepper, Clove, Ginger, Nutmeg, Turmeric
Antioxidant	Vanilla, Ginger, Black pepper, Clove, Turmeric
Antiulcer	Ginger, Black pepper, Turmeric, Cinnamon, Clove, Nutmeg, Vanilla, Allspice, Mace
Sedative	Ginger, Cinnamon, Black pepper, Nutmeg, Clove
Anaesthetic	Black pepper, Nutmeg, Vanilla, Turmeric, Clove
Hypocholesteremic	Cardamom, Mace
Hypotensive	Ginger, Black pepper, Cinnamon, Clove, Turmeric, Mace
Analgesic	Ginger, Black pepper, Cinnamon, Nutmeg, Turmeric, Clove
Antidepressant	Ginger, Clove, Mace, Allspice, Cardamom
Antiasthmatic	Allspice
Candidicide	Nutmeg, Cinnamon
Hypoglycemic	Cardamom
Antihepatotoxic	Vanilla
Antiallergic	Allspice
Antimigraine	Turmeric, Allspice, Cardamom, Mace
Antiosteoporotic	Black pepper, Allspice, Clove, Cardamom, Mace
Estrogenic/ Androgenic	Cardamom
Immunostimulant	Turmeric, Mace
Antilithic	Allspice
Anti-insomniac	Allspice, Clove, Mace
Antiedemic	Vanilla

Robert Ronzio has reviewed the antioxidants in spices, and examined how they function as Nutraceuticals and Functional Foods (http://www.tldp.com/issue/191/Antioxidants191.htm)

Some case studies on clinical effects of plants

Botanical antioxidants may counter the toxicity of certain chemotherapeutic agents

Tinospora cordifolia is widely used in botanical formulations in India and China. *Tinospora cordifolia* extract (TCE) quenched superoxide radicals and hydroxyl radicals *in vitro* assays. The extract also inhibited lipid peroxidation in isolated liver microsomal fractions. When incubated with activated macrophages cell cultures, TCE inhibited the production of superoxide generated by the oxidative burst of these immune cells.

A variety of chemotherapeutic agents such as cyclophosphamide cause free radical-induced cytotoxicity as a side effect when they are activated by mixed function oxidases. Mice administered 10 doses of cyclophosphamide had a drastically suppressed bone marrow cell population. Treatment with cyclophosphamide and together with TCE (50 mg iv) entirely blocked the cyclophosphamide-induced cytotoxicity in bone marrow cells and in lymphocytes of mice. Furthermore, TCE reduced lipid peroxidation due to drug exposure in the liver of test animals.

Thus *Tinospora cordifolia* is reported to benefit the immune system in a variety of ways. These experiments suggest that its ability to scavenge free radicals and to block free radicals and to inhibit radical-induced membrane damage may be the underlying feature of many of the biological activities of this herb. Cyclophosphamide is inert until activated by cytochrome P450 to produce a DNA alkylating agent capable of killing cells. This process produces free radicals as a by-product and TCE may quench them, and thus limit oxidative damage associated with this agent. The key question remains, do these radical quenching effects occur when the herb is ingested orally? (Mathew, S.K.G. 1997. Antioxidant activity of *Tinospora cordifolia* and its usefulness in the amelioration of cyclophosphamide induced toxicity. J Exp Clin Cancer Res; 16:407-11)

Taxol for cancer treatment

Paclitaxel is a mitotic inhibitor used in cancer chemotherapy, which works by interfering with normal microtubule breakdown during cell division. It was discovered in a National Cancer Institute program at the Research Triangle Institute in 1967 when Monroe E. Wall and Mansukh C. Wani isolated it from the bark of the Pacific yew tree, *Taxus brevifolia* and named it 'taxol'. The generic name was changed to 'paclitaxel', when it was developed commercially by Bristol-Myers Squibb (BMS). It is now used to treat patients with lung, ovarian, breast cancer, head and neck cancer, and advanced forms of Kaposi's sarcoma and also used for the prevention of restenosis.

From 1967 to 1993, almost all paclitaxel produced was derived from bark of the Pacific yew, the harvesting of which kills the tree in the process. Gordon Cragg of the NCI's Natural Product Branch calculated that the synthesis of enough taxol to treat all the ovarian cancer and melanoma cases in the US would require the destruction of 360,000 trees annually. Total chemical synthesis and semi-synthetic production route to taxol production were attempted in the US and France. Currently, all paclitaxel production for BMS uses plant cell fermentation technology developed by the biotechnology company Phyton Biotech, Inc.

Paclitaxel interferes with the normal function of microtubule breakdown by binding to the β subunit of tubulin (the "building block" of microtubules), destroying the cell's ability to use its cytoskeleton in a flexible manner. The resulting microtubule/paclitaxel complex cannot disassemble, adversely affecting cell function, because the shortening and lengthening of microtubules is necessary for their function as a mechanism to transport other cellular components. Paclitaxel induces programmed cell death (apoptosis) in cancer cells by binding to an apoptosis stopping protein called Bcl-2 (B-cell leukemia 2) and thus arresting its function.

In addition to stabilizing microtubules paclitaxel may act as a molecular mop by sequestering free tubulin effectively depleting the cells supply of tubulin monomers and/or dimers. This activity may trigger the aforementioned apoptosis.

The French paradox

The French paradox is the observation that the French suffer a relatively low incidence of coronary heart disease, despite having a diet relatively rich in saturated fats. The phenomenon was first noted by Irish physician Samuel Black in 1819, and the term was coined by Dr. Serge Renaud, a scientist from Bordeaux University in France in 1992. However, some health researchers question the validity of this paradox.

According to FAO data, the average French person consumed 108 grams per day of fat from animal sources in 2002 while the average American consumed only 72. The French eat four times as much butter, 60 percent more cheese and nearly three times as much pork. Although the French consume only slightly more total fat (171 g/d vs 157 g/d), they consume much more saturated fat because Americans consume a much larger proportion of fat in the form of vegetable oil, with most of that being soybean oil. However, according to data from the British Heart foundation, in 1999, rates of death from coronary heart disease among males aged 35–74 years was 115 per 100,000 people in the US but only 83 per 100,000 in France.

France's high red wine consumption has been suggested as a primary factor in the trend. Wines, particularly red wines, are a source of resveratrol, one of the ingredients in red wine is the cause for the effect, and it has also been linked to longevity and cancer prevention in other species. Research suggests that moderate drinkers are less likely to suffer heart attacks than are abstainers or heavy drinkers.

Professor Roger Corder and team believe that a group of polyphenols, known as oligomeric procyanidins, not resveratrol alone, offer the greatest degree of protection to human blood-vessel cells, and the highest procyanidins are found in wines grown in the southwest France. Other research suggests that polyphenols in wine reduce the absorption of malondialdehyde, which is implicated in arteriosclerosis, cancer, diabetes and other diseases.

In his book, *The Fat Fallacy*, Dr. Will Clower suggests the French paradox may be narrowed down:

- Good fats versus bad fats — French people get up to 80% of their fat intake from dairy and vegetable sources including milk and milk products, they consume very little animal fat.
- Higher quantities of fish (three times a week).
- Smaller portions, eaten more slowly and divided among courses that let the body begin to digest food already consumed before more food is added.
- Lower sugar intake — American low-fat and no-fat foods often contain high concentrations of sugar. French diets avoid these products preferring full-fat versions without added sugar.
- Low incidence of snacks between meals.
- Avoidance of common American food items, such as soda, deep-fried foods, snack foods, and especially pre-prepared foods which can typically make for a large percentage of the foods found in American grocery stores.

Estrogen-like and progesterone-like properties

Extracts of herbs and spices with the highest estrogen receptor binding activity in cultured human breast cancers (that is, with high levels of phytoestrogens) were soy, thyme, licorice, red clover, turmeric, hops and verbena. The six highest progesterone-binding herbs and spices (ie, highest in phytoprogestins) were thyme, oregano, verbena, turmeric, red clover and damiana. The abilities of extracts of these plants to regulate the growth of estrogen receptor-positive and estrogen receptor-negative breast cancer cell lines was also examined. Their abilities to regulate production of progestin-induced alkaline phosphatase, in progesterone receptor-positive cells were also compared. Herbal extracts, high in phytoestrogens that bound to estrogen receptors, either stimulated growth (that is, they functioned as estrogen agonists) or they had no activity (they were neutral). In contrast, extracts that were high in phytoprogestins either inhibited alkaline phosphatase (ie, they functioned as progesterone antagonists) or they had no effect (they were neutral) in these test systems.

The results show that of the plant materials tested, culinary herbs and spices such as thyme, oregano, turmeric and hops contained relatively large amounts of estrogen receptor binding activity. Consequently, these herbs and spices may act as phytoestrogens *in vivo*, and foods containing them would be expected to possess mild estrogenic activity.

The most potent estrogen-binding products were soy milk, licorice and red clover. All three have been proposed to have chemopreventive properties. In addition, licorice, red clover and motherwort could be direct-acting estrogen agonists. Interestingly, dong quai, damiana, and black cohosh exhibited little estrogen receptor binding activity in these studies. However, these three herbs are often incorporated into protocols employed to normalize female hormone imbalances. It is not clear whether these latter herbs contain precursors that are converted to estrogens *in vivo*, or whether they mediate other receptor systems or modulate metabolism of endogenous hormones (Zava, D.T. *et al*. 1998. Estrogen and progestin bioactivity of foods, herbs and spices. PSEBM; 217:369-78).

Culinary spices can inhibit several food-borne pathogens

The antibacterial properties of 21 plant essential oils were studied using five different pathogenic strains, including *Campylobacter jejuni, Salmonella enteritis, Escherichia coli, Staphylococcus aureus* and *Listeria monocytogenes*. Generally, the Gram-positive bacteria were more sensitive to inhibition by the plant oils than Gram-negative species. The oils prepared from bay, cinnamon, clove and thyme possessed the highest bacteriostatic properties, with effective concentrations of 0.075% or less.

Bacterial pathogens account for a large percentage of food-borne illness. For example, *Campylobacter jejuni* is a recently recognized and most common food pathogen. *Listeria monocytogenes* is especially troublesome because this bacterium is able to multiply in refrigerated foods, unlike most other food-borne pathogens. While it is unlikely that concentrations of these oils of culinary herbs and spices can be routinely used in foods at sufficiently high concentrations to kill bacteria, this study suggests that concentrations sufficient to inhibit bacterial growth could improve food safety if the pathogen load is low (Smith-Palmer A, et al. 1998. Antimicrobial properties of plant essential oils and essences against five important food-borne pathogens. Letters Applied Micobiol; 26:118-22.)

To conclude, phytochemicals have always been recognized as potential nutraceuticals and have also been used in alternative systems of medicine for ages. The search for the bioactive principles in plants, trees and microbes which make up their chemistry and their clinical effects remain a vast promising area for research.

Basic Instruments Used for Chemical Analysis of Soil, Water, Manure and Plant Samples

S. Hamza and V. Srinivasan

1. pH Meter

A **pH meter** is a specific type of voltmeter. pH is defined as the negative logarithm of hydrogen ion concentration. pH = - log (H+). **pH meter** works on Nerns't equation. The common pH meter has several inputs for indicator (ion-sensitive or redox) and reference electrode and temperature sensors. Two electrodes are used in the determination of pH. One is the reference electrode, which provides standard voltage. It contains two layers, first layer consist of a strip of platinum sealed in glass and dipped into paste of calomel (Hg_2Cl_2, 0.1 M) and the second layer filled with saturated KCl solution. Second electrode is a glass electrode that consists of high resistance glass tube with a thin, low resistance glass bulb at the bottom. It encloses silver coated silver wire with wax insulation. When the electrodes are dipped in the solution, the saturated solution of KCl comes out of the reference electrode through the small holes and forms an invisible ionic bridge between electrodes through which current passes. The H ions absorbed by glass electrode and electric potential develops between electrodes. This potential difference is measured in terms of pH by suitable galvanometer. The pH scale of the device should be calibrated by at least two buffer solutions. Usually one of the buffers used for calibration has pH 7.00 and the second is selected depending on the range where the measurements are to be taken - 9.20 for basic solutions and 4.01 for acidic solutions. This correlates the measured potential of the indicator ectrode with the pH scale. While measuring the electrode should always remain immersed in the water. Allow the instrument to stabilize for a period of at least 30 minute before use.

2. Conductivity Meter

A conductivity meter measures the ionic conductivity (or conversely, the resistance) of a liquid. The number it gives the total ion content of the liquid. The device usually consists of is a probe which usually has two platinum electrode plates parallel to each other and separated by some small distance. The meter is a Wheatstone bridge apparatus with a small oscillator and readout. The oscillator generates small amplitude (about 5 to 10 mV peak-to-peak) sin wave. The meter is zeroed by dialing in a resistance on one side of the Wheatstone bridge. The resistance dialed in is then the resistance between the two plates in the probe. Relating this value to a calibration curve for the probe will give you the concentration of ionic species in solution.

Conductivity is normally reported in micro-mhos per centimeter (µmho/cm). In the international system of units the conductivity is reported as simens (S) (reciprocal of the ohm) milli-siemens or desi-siemens per meter (mS/m or dS/m).

Conductivity (µS/cm) X 0.5 = TDS mg/l as NaCl

Conductivity meter is calibrated using standard KCl solution of following concentrations

1.0 M KCl	Conductivity	111.3 mS/cm	at 25°C
0.1 M.	„	85 mS/cm	,,
0.01 M	„	1408 µS/cm	,,
0.001 M	„	146.1 µS/cm	,,

3. UV-Visible Spectrophotometer

In this instrument to obtain absorption measurments, a sample is placed in the spectrophotometer and ultraviolet or visible light at a certain wavelength, or range of wavelengths, is transmitted through the sample. The beam of light from a visible and/or UV light source is separated into its component wavelengths by a prism or diffraction grating. Each monochromatic (single wavelength) beam in turn is split into two equal intensity beams by a half-mirrored device. One beam, the sample beam (colored magenta), passes through a small transparent container (cuvette) containing a solution of the compound being studied in a transparent solvent. The other beam, the reference (colored blue), passes through an identical cuvette containing only the solvent. The intensities of these light beams are then measured by electronic detectors and compared. The spectrophotometer actually measures how much of the light is absorbed by the sample. The intensity of the reference beam, which should have suffered little or no light absorption, is defined as $\mathbf{I_0}$. The intensity of the sample beam is defined as **I**. Over a short period of time, the spectrometer automatically scans all the component wavelengths in the manner described. The ultraviolet (UV) region scanned is normally from 200 to 400 nm, and the visible portion is from 400 to 800 nm.
Absorption may be presented as **transmittance** ($\mathbf{T = I/I_0}$) or **absorbance** (A= log $\mathbf{I_0/I}$). If no absorption has occurred, T = 1.0 and A= 0. Most spectrometers display absorbance on the vertical axis, and the commonly observed range is from 0 (100% transmittance) to 2 (1% transmittance). The wavelength of maximum absorbance is a characteristic value, designated as **λmax**. Different compounds may have very different absorption maxima and absorbances. The absorbance of a sample will be proportional to its molar concentration in the sample cuvette and follows **Beer-Lambert law,** the concentration of the desired compound varies linearly with the absorbance. Thus UV/VIS spectroscopy can determine the concentration of an unknown solution, based on reference molar extinction coefficients or more accurately, using a calibration curve **Molar Absorptivity, ε** =A/C × I
(where A= absorbance, **C** = sample concentration in moles/liter & **I** = length of light path through the cuvette in cm.).

4. Kjeldhal Nitrogen Analyzer

Johan Kjeldahl first introduced the Kjeldahl nitrogen method in 1883. Kjeldahl nitrogen determinations are performed on food and beverages, meat, feed, grain, waste water, soil and many other samples. The auto Kjeldhal nitrogen analyzer consists of two parts, block digester and distillation, titration assembly. In block digester sample of known weight is mixed with catalytic mixture (K_2SO_4:$CuSO_4$, 20:1) and Conc.H_2SO_4 and digestion is performed at 420°C The digested samples after dilution with distilled water transferred into distillation titration assembly and steam distilled after neutralization of sulphuric acid with sodium hydroxide solution (40%) The liberated ammonia will be collected in boric acid solution containing bromo cresol green,methyl red mixed indicator which is titrated against standard sulphuric acid(0.1N). After the end of process result can be read out on display or manually in burette.

Digestion - the decomposition of nitrogen in organic samples utilizing a concentrated sulfuric acid. The end result is an ammonium sulfate solution.
Organic N + $H_2SO_4 \rightarrow (NH_4)_2SO_4 + H_2O + CO_2$ + other sample matrix by-products.

Distillation - Adding excess base to the acid digestion mixture to convert $(NH_4)_2SO_4$ to NH_3, followed by boiling and condensation of the NH_3 gas in a receiving solution.
$(NH_4)_2SO_4 + 2NaOH \rightarrow 2NH_3 + Na_2SO_4 + 2H_2O$
$NH_3 + H_3BO_3 \rightarrow NH_4^+$: $H2BO_3^- + H_3BO_3$

Titration - To quantify the amount of ammonia in the receiving solution, the amount of nitrogen in a sample can be calculated from the quantified amount of ammonia ions in the receiving solution.

NH_4+: $H_2BO_3^- + HCl \rightarrow NH_4Cl + H_3BO_3$

N (%) =(S - B) × N × 1.407/ Sample weight (g)

Where

S= Volume of acid used against sample.

B= Volume of acid used against blank.

N= Normality of acid.

If it is desired to determine % protein instead of % nitrogen, the calculated % N is multiplied by a factor, the magnitude of the factor depending on the sample matrix. Many protein factors have been developed by AACC and AOAC for use with various types of samples such as, 6.38 for milk and dairy, 5.95 for rice, 5.70 for wheat flour and 6.25 for other grains.

5. Flame Photometer

A flame photometer is an instrument used for measuring the spectral emission intensity of metals present in the metallic salt. Generally it is used for the routine estimation of metal with low ionization potentials especially sodium (Na), potassium (K), lithium (Li) and calcium (Ca) It is also called *flame atomic emission spectrometry*. As an analytical method, atomic emission is a fast, simple, and sensitive method for the determination of trace metal ions in solution. The sample solution is sprayed or aspirated as fine mist into a flame (LPG gas as fuel). Conversion of sample solution into an aerosol is made by the atomizer. The Heat of the flame vaporizes sample constituents. Here the molecules & ions of the sample species are decomposed and reduced to give atoms. The energy provided by the flame excites the orbital electron to a higher energy levels. When these electrons returned to their ground state they emit characteristic radiations of the element and the spectrometer measures the intensity of the emitted line. Atoms in the vapour state give line spectra. Coloured glass filter usually able to isolate the line of analyte element if well separated from other emission lines. Now the developments of AAS and ICP despite their high cost have replaced flame photometer.

6. Atomic Absorption Spectrophotometer

Atomic absorption spectrophotometer (AAS) is analytical equipment based on atomic absorption spectro-photometry and is used to measure concentrations of any metals in the sample. When a sample is aspirated into the instrument, solvent will vaporize and atoms in vapor phase subjected to a heavy thermal environment and as a result, ‘the ground state’ atoms absorb light energy of a specific wavelength from specific source (Hollow cathode lamp) and enter in to the excited sate. As the number of atoms in the light path increases, the amount of light absorbed increases in a predictable way. By measuring the intensity of light absorbed, a quantitative determination of the amount of analyte element present can be made. Normally AAS consist of five basic components, *viz.* a light Source (cathode lamp), a sample cell (absorption cell), Monochromometer, Detector; and out put unit, spectra. A continuous spectrum light source, even with high quality monochromator cannot achieve sufficiently narrow band pass width for absorption line spectra. So special lamps, each emitting line spectrum matched to the line spectrum of the analyte atoms, hollow cathode lamp are used. Characteristic concentration in atomic absorption (sometimes called "sensitivity") is defined as the concentration of an element (expressed in mg/l) required to produce a signal of 1% absorption (0.0044 absorbance units). As long as measurements are made in the linear working range, characteristic concentration can be determined by reading the absorbance produced by a known concentration of the element (standard solutions).

7. Microwave Digestion System

Microwave digestion system is an advanced and highly sophisticated system for sample preparation by using the electromagnetic radiation to achieve higher temperature for the reaction. It provides high performance, reliable quality and unrivaled safety, which is required in sample preparation in order to achieve superior analytical results. Its closed vessel technique helps to speed up reactions by allowing higher temperatures, while preventing the loss of volatile analytes. The resulting low reagent consumption saves time and money and also helps to minimize exposure to corrosive gases and hazardous solvent vapors. The magnetron of the system generates microwaves of 0.3 mm – 1.0 m wavelength having frequency of 100GHz to 300 MHz, which has strong penetrating power on the matrix to extract the analyte of interest. This instrument is now widely used in digesting of various abiotic and biotic matrices to extract organic and inorganic chemicals.

Extraction of plant enzymes and proteins and isozyme analysis

K S Krishnamurthy and B Chempakam

Enzymes are biological catalysts that regulate the rate of a specific biochemical reaction in living organism. When it comes to the hand of a biotechnologist, it becomes the "*tailor's scissors and the surgeon's scalpel*." Enzymes are originally discovered in yeast by Buchner brothers (En= in; zyma=yeast). They are structurally proteins.

How do enzymes work?

The chemicals at which an enzyme operates are called **substrates.** Enzymes are **specific** ie; they catalyse only a few closely related chemical reactions –or in many cases only one particular reaction. They form temporary chemical compounds with their substrates (E-S complex), which then release the products and regenerating the original enzyme molecule for reuse.

The prosthetic group is the non-protein group of the enzyme, which is bound by the chemical bonds. Haem, biotin.pyridoxal phosphate etc. function as prosthetic groups. **Apoenzyme** is the remaining protein part of the enzyme, which is inert. **Coenzyme** is the coworker of prosthetic group and apoenzyme, since without coenzyme, the enzyme is non-reactive. They combine with the enzyme and leave it after catalysis.

Eg; DPN, NAD, ATP, ADP etc

Cofactor: When a prosthetic group consists of single atoms of some metals like Mg^{++}, Fe^{++}, Cu^{++}, Mo^{++} then it s known as **cofactor** and can be easily separated from the protein part.

There are two ways an enzyme can work;

1. Lock and key hypothesis
2. Induced fit model

Lock and Key model

Enzymes have one or more regions called *active sites,* which are located close to one another on the enzyme surface. During the course of a reaction, the substrate molecules occupy the sites and are temporarily brought together. At this the shape of the enzyme molecule changes resulting in strain in critical bonds in the substrate molecule. The new chemical compound produced has little affinity for the substrate and moves away.

Induced fit model

Here the active sites of the enzyme are not rigid. When a substrate binds to the enzyme, it induces a change in the shape of the enzyme molecule resulting in optimum fit of the substrate- enzyme reaction. This puts a strain on the substrate, promoting the reaction.

Regulation of enzyme action

The chemical reactions are controlled by the enzymes. But what controls the enzymes?

1. The mechanism of enzyme regulation depends on the amount of enzyme produced, which is directed by a specific gene, which in turn is switched on by a signal from the hormone or any other cellular product.
2. *Allosteric control*: Enzymes which are present in the inactive form in the cytoplasm may be activated. In the inactive form, the active sites of the enzymes have no shape, so the substrate does not fit. In some cases, active site and substrate site are also present. When a molecular activator or cyclic AMP attaches to the site, the shape of the enzyme molecule changes, making the active site available for substrate combination.

Isozymes and allozymes

Underlying principles

1. Certain enzymes are well-characterized as to their genetic basis and inheritance
2. Individuals of particular plant populations or species may harbor one or more alleles of each enzyme
3. Genetic variation and various parameters of population subdivision, relationships and Mendelian inheritance are calculated based on the proportion of alleles within and among individuals, populations and species

Important assumptions of isozymes and allozymes

1. Enzyme variation is a representative reflection the overall DNA variation
2. Two comigrating "identical" alleles for the same enzyme in two different individuals/ populations/species are coded for by the exact same mutation

What are isozymes?

Isozymes (Isoenzymes) were discovered 30 years ago and since then lot of information has accrued on the subject. These studies have helped us in understanding the metabolic regulation in plants, bacteria and animals and in making available a multitude of highly sensitive markers. Isozymes are also being used in diagnostic clinical biochemistry.

Thus isozymes are ***multiple molecular forms of enzymes, which can be separated from one another using several techniques***. They are also proteins, carrying a net negative charge. Relative migration rate of isozymes under polarised field is a function of their MW, configuration, Amino acid composition etc. Isozyme structure is the direct expression of genetic information contained in m RNA and provide direct link to DNA. And structural genes encoded therein. Hence isozyme polymorphism is especially rewarding to plant and animal geneticists and breeders.

Nomenclature & terminology

1. In many cases, the normal enzyme name should be used followed by a no. Which is given according to the electrophoretic mobility under defined conditions.
2. Increase of complex enzymes where more than one zone is seen, numbers may be used to designate major groups. Thus, 1a, 1b, 2a, 2b etc.

What are allozymes?

While isozymes refer to *isomers of functional types of enzymes regardless of locus*, allozymes refers to isomers *within a genetic locus.* Starch gel electrophoresis is used to assay the allozymes of numerous gene loci .The enzyme migrate through the gel matrix in response to the electric field at a rate proportional to the net surface charge. After electrophoresis, the gels are cut into thin horizontal slices and each slice is put into an enzyme specific staining solution.

Multi locus enzyme information from a single tree can provide a biochemical fingerprint, a genotype of that tree. Allozyme data from populations of trees can form a powerful database for analysing the genetic characteristics of that tree.

Why enzyme multiplicity does occurs?

1. Genetic or primary causes
2. Post translational or secondary causes

In the former, the organism carries multiple genes, each one encoding for a different enzyme sub unit, while in the latter, homogeneous enzyme sub units are modified differentially from a single gene.

Isozymes as genetic markers

Isozyme analysis is an important experimental technique in Genetics. Each isozyme sub unit, by definition, is the result of a different gene, whether encoded by multiple alleles at a single locus or multiple loci. Use of isoenzymes as markers is wide spread and the hypothesis 'one-gene one-isozyme sub unit 'has been proposed in 1980.

Well-set example of isozymes - LDH

In glycolysis, glucose molecule is metabolised to form pyruvate. The final reaction of anaerobic glycolysis is the conversion of pyruvate to lactate calalysed by the enzyme Lactic De Hydrogenase (LDH). In skeletal muscle where oxygen deprivation is common during exercise, this reaction is efficient and large amount of lactate can be formed. In tissues that preferentially oxidise glucose aerobically to CO_2 and water, such as cardiac muscle the reaction is not efficient and pyruvate is preferentially converted to Acetyl CoA and then enters TCA cycle.

Five major LDH isozymes are found in different vertebrate tissues. Each LDH molecule is composed of 4 polypeptide chains, but the sub unit composition of the 5 LDH isozymes are different. There are two types of polypeptide chains in LDH called **M** (for skeletal muscle) and **H** (for heart muscle), which can be combined in 5 different ways.

Different genes encode the H and M polypeptide chains and the two genes are expressed to different degrees inn different tissues. Thus in heart muscle the gene for the H sub unit is more active than the gene for the M sub unit. Thus **LDH 1** is the predominant one in cardiac muscle, while **LDH5** is more active in skeletal muscle.

Applications

Measurement of LDH isozymes in serum has been extensively used to find the site and nature of tissue injury in humans. Thus when blood supply to the heart muscle is severely reduced as during a heart attack, muscle cells die and liberate **LDH 1** to the blood stream. In contrast, **LDH 5** increase in serum is indicative of muscular dystrophy.

Isozyme markers

Isozyme markers have been developed in several horticultural crops for characterisation, varietal identification, sex determination etc. as follows:

Crop	**Objective**	**Reference**
Amaranthus	Classification	Oken *et al*, 1996
Citrus	Classification	Dequin *et al*, 1994
Tomato	Nematode resistance	Rick and Fobes, 1994
Asparagus	Sex determination	Maestri *et al*; 1991

Applications

1. Widely applicable for population-level and species-level comparisons among closely related organisms
2. Still the "work horse" for population-level studies, although many have switched to DNA-based approaches (restriction site variation, RAPDs and other anonymous markers)

Advantages

1. At least some enzymes will show variation in most plant studies; no need to do extensive screening beyond preliminary tests for successful enzyme/gel/buffer systems
2. Purportedly cheaper to do lots of samples than other methods, including DNA
3. Initial cash outlay for equipment not huge, no particularly expensive pieces of equipment beyond gel rigs and electrophoresis power supplies
4. Disadvantages
5. Must use fresh or recently frozen tissue; cannot store material for long term beyond freezing
6. Relatively few data points per unit of effort--at most a couple of dozen alleles possible even with extensive screening

7. Must have some understanding of genetic situation for each plant group--polyploid taxa will show complex patterns, difficult to interpret
8. Some plant species or species groups will display little or no variation; must use more sensitive, DNA-based methods to access variation in those
9. Availability of limited no of enzyme loci
10. Developmental and season –dependant expression of activity.

Reagents for Native gel electrophoresis:

- Acrylamide - bisacrylamide mix
- Acryl amide 30%
- Bisacryl amide 0.8%-Strore at 4^0C, away from light
- 1.5M Tris pH 8.8 (4x resolving gel buffer)-pH adjusted to 8.8 with 4N HCl
- 1.0M Tris pH 6.8-pH adjusted to 6.8 with 4N HCl
- 10% Ammonium persulphate (APS)-Should be prepared fresh as APS decomposes slowly during storage
- 10 % Sodium Dodecyl Sulphate (SDS)-(Only for SDS PAGE)
- Electrophoresis (Reservoir) buffer (10x) for Native gel- 7.56 g TRIS + 38 g glycine, pH ajusted to 8.3, volume made up tp 250 ml with deionised water.
- Electrophoresis (Reservoir) buffer (5x) for SDS PAGE- 15.1 g TRIS + 94 g glycine, pH adjusted to 8.3, 0.1% SDS., volume made upto 1000 ml with deionised water

(*Note: It is advisable to use deionised water for the preparation of reagents. In case of non-availability, double distilled water may be used*)

Sample buffer for SDS PAGE

- 4x stacking gel buffer 2.5 ml
- 10 % SDS 0.4 ml
- Glycerol 2 ml
- Bromophenol blue 2 mg
- Mercapto ethanol 5 % (final conc.)

Make up to 10 ml with deionised water

Ingredients for Native gel (10% gel), Volume 7.5 ml

- Acryl amide bisacryl amide mix 2.5 ml
- 1.5M Tris (8.8 pH)
- Double distilled water
- 10 % APS
- TEMED

Ingredients for SDS (10% gel), Volume 10 ml

- Acryl amide bisacryl amide mix 3.3 ml
- 1.5M Tris (8.8 pH)
- Double distilled water
- 10 % SDS
- 10 % APS
- TEMED

Ingredients for stacking gel (5 % gel), Volume 6 ml

- Acryl amide bisacryl amide mix 1 ml
- 1.0M Tris (6.8 pH)
- Double distilled water
- 10 % APS
- TEMED 10 µl

Gel casting

1. Rub the glass plates thoroughly with ethanol, dry and assemble them. The notched glass plate should be facing out wards.
2. In an erlenmeyer flask, prepare the appropriate volume of the solution containing the desired concentration of acrylamide for the resolving gel, using the values given earlier. Mix the components in the order shown. Polymerization will begin as soon as the TEMED is added. Without delay, swirl the mixture rapidly and proceed to the next step.
3. Pour the acrylamide solution into the gap between the glass plates. Leave sufficient space for the stacking gel (length of the teeth of the comb plus 1 cm). Using a pasteur pipette, carefully overlay the acrylamide solution with 0.1% SDS (for gels containing ≤ 8 % acrylamide) or isobutanol (for gels containing ≥ 10 % acrylamide). Place the gel in a vertical position at room temperature. The overlay prevents oxygen from diffusing into the gel and inhibiting polymerization.
4. After polymerization is complete (30 minutes), pour off the overlay and wash the top of the gel several times with deionised water to remove any unpolymerised acrylamide. Drain as much fluid as possible from the top of the gel, and then remove any remaining water with the edge of a paper towel.
5. Prepare the stacking gel in a beaker/flask, prepare the appropriate volume of solution containing the desired concentration of acrylamide, using the values given earlier. Mix the components in the order shown. Polymerization will begin as soon as the TEMED is added. Without delay, swirl the mixture rapidly and proceed to the next step.
6. Pour the stacking gel solution directly onto the surface of the polymerized resolving gel. Immediately insert a clean teflon comb into the stacking gel solution, being careful to avoid trapping air bubbles. Add more stacking gel solution to fill the spaces of the comb completely. Place the gel in a vertical position at room temperature.
7. Teflon comb should be cleaned with water and dried with ethanol before use.
8. After polymerization is complete (30 minutes), remove the Telfon comb carefully. Using a squirt bottle, wash the wells immediately with deionised water to remove any unpolymerized acrylamide. If necessary, straighten the teeth of the stacking gel with a blunt hypodermic needle attached to a syringe. Mount the gel in the electrophoresis apparatus. Add Tris-glycine electrophoresis buffer to the top and bottom reservoirs. Remove any bubbles that become trapped at the bottom of the gel between the glass plates using a bent hypodermic needle attached to syringe.
9. Load up to 50 µl of each of the samples in to the bottom of the wells using a micropipette or a syringe. Attach the electrophoresis apparatus to an electric power supply (the positive electrode should be connected to the bottom buffer reservoir). Apply current at the rate of 1.5-milli amp per gel. After the dye front has moved into the resolving gel, increase the current to 2-milli amp per gel and run the gel until the bromophenol blue reaches the bottom of the resolving gel. Then turn off the power supply.
10. Remove the glass plates from the electrophoresis apparatus and place them on a paper towel. Using a spatula, pry the plates apart. Mark the orientation of the gel by cutting a corner from the bottom of the gel that is closet to the left most well.
11. The gel now can be fixed and stained.

Staining Procedure (For Isozymes)

1. Conduct electrophoresis in starch or polyacrylamide disc/slab buffer gels (with no SDS) at low temperature
2. Immediately after electrophoresis, incubate the gel in the substrate solution(s). The zones where the enzymes are located in the gel are visualized due to the appearance of coloured reaction products. After sufficient incubation period, stop the reaction by adding appropriate stop solution and photograph the zymogram. Otherwise, the relative position of each visualized band in the gel may be drawn schematically for easy reference.

3. Enzyme extraction and stain for various isoenzymes are given below:

Esterase
Enzyme Extract
Homogenize the sample material in 5 fold volume of 10 mM sodium phosphate buffer (pH 9.5), 1 mM EDTA Na^{2+}, 1 mM 2-mercaptoethanol. Centrifuge the homogenate at 10,000 x g for 10 min and use the supernatant as enzyme source. All operations are at 0-4ºC. Incubate the gel in a solution given below at 37ºC for 20-30 min, preferably in dark.

• Sodium dihydrogen phosphate	2.8 g
• Disodium hydrogen phosphate	1.1 g
• Fast blue RR salt	0.2 g
• Alpha-naphthyl acetate	0.03 g
• Water to	200 ml

Stop the enzyme reaction by adding a mixture of methanol: water: acetic acid: ethyl alcohol in the ration 10:10:2:1.

Polyphenol oxidase (PPO)
Enzyme Extract
The PPO is extracted by homogenizing the material in 0.01 M potassium phosphate buffer (pH 7.0) containing 1% non-ionic detergent (Tween 80) at 0ºC for 15 min. Centrifuge the homogenate at 20,000-x g for 15 min at 0ºC in a refrigerated centrifuge. Use the supernatant as enzyme source. Equilibrate the gel for 30 min in 0.1% p-phenylenediamine in 0.1 M potassium phosphate buffer (pH 7.0) followed by 10 mM catechol in the same buffer.

Peroxidase
Enzyme Extract: Same as above
Incubate the gel in the following solutions

- Benzidine 2.08 g
- Acetic acid 18 ml
- Hydrogen peroxide (3%) 100 ml
- Water 80 ml

Bright blue coloured bands appear in gel. When the bands are stained sufficiently, arrest the reaction by immersing the gel into a large volume of 0.67% sodium hydroxide or 7% acetic acid solution for 10 min.

Staining SDS polyacrylamide gels with Coomassie brilliant blue

1. Dissolve 0.25 g of Coomassie Brilliant Blue R250 in 90 ml of methanol:DD H_2O (1:1 v/v:) and 10 ml of glacial acetic acid. Filter the solution through a whatman No. 1 filter paper to remove any particulate matter.
2. Immerse the gel in at least 5 volumes of staining solution and place on a slowly rotating platform for a minimum of 4 h at room temperature.
3. Remove the stain and save it for future use. Destain the gel by soaking it in the methanol: DD H_2O (1:1 v/v:) and glacial acetic acid as in step 1 without the dye on a slowly rocking platform for 4-8 h, changing the destaining solution 3-4 times.
4. The more thoroughly the gel is destained, the smaller the amount of protein that can be detected by staining with Coomassie Brilliant Blue. Destaining for 24 hr usually allows as little as 0.1 µg of protein to be detected in a single band.
5. Store fixed gels in 7 % acetic acid or in 20 % glycerol. Stained gels should not be stored in destaining buffer, which will cause the stained protein bands to fade.

Protein estimation by Lowry's method

Protein can be estimated by different methods as described by Lowry's and also by estimating the total nitrogen content. No method is 100% sensitive. Hydrolyzing the protein and estimating the amino acids alone will give the exact quantification. The method developed by Lowry *et al* is sensitive enough to give a moderately constant value and hence largely followed. Protein content of enzyme extracts is usually determined by this method.

Principle

The blue colour developed by the reduction of the phosphomolybdic-phosphotungstic components in the Folin-Ciocalteau reagent by the amino acids tyrosine and tryptophan present in the protein pus the colour developed by the biuret reaction of the protein with thealkaline cupric tartrate are measured in the Lowry's method.

Materials

- 2% Sodium Carbonate in 0.1 N Sodium Hydroxide (Reagent A)
- 0.5% Copper Sulphate ($CuSO_4$. $5H_2O$) in 1% potassium sodium tartrate (Reagent B)
- Alkaline Copper solution: Mix 50 ml of A and 1 ml of B prior to use (Reagent C)
- Folin-Ciocalteau Reagent (Reagent D) - Reflux gently for 10 hrs a mixture consisting of 100 g sodium tungstate ($Na_2WoO_4.2H_2O$), 25 g sodium molybdate ($Na_2MoO_4.2H_2O$), 700 ml water, 50 ml of 85% phosphoric acid and 100 ml of conc. HCl in a 1.5 l flask. Add 150 g lithium sulfate, 50 ml water and a few drops of bromine water. Boil the mixture for 15 min without condenser to remove excess bromine. Cool, dilute to 1 l and filter. The reagent should have no greenish tint (Determine the acid concentration of the reagent by titration with 1 N NaOH to a phenolphthalein end point).
- Protein solution (stock standard): Weigh accurately 50 mg of bovine serum albumin (Fraction V) and dissolve in distilled water and make up to 50 ml in a standard flask.
- Working standard: Dilute 10 ml of the stock solution to 50 ml with distilled water in a standard flask. One ml of this solution contains 200 µg.

Procedure

Extraction of protein from sample: Extraction is usually carried out with buffers used for the enzymes assay. Weigh 500 mg of the sample and grind well with a pestle and mortar in 5-10 ml of the buffer. Centrifuge and use the supernatant for protein estimation.

Estimation of protein

1. Pipette out 0.2, 0.4, 0.6, 0.8 and 1.0 ml of the working standard into a series of test tubes.
2. Pipette out 0.1 ml and 0.2 ml of the sample extract in two other test tubes.
3. Make up the volume to 1 ml in all the test tubes. A tube with 1 ml of water serves as the blank.
4. Add 5 ml of reagent C to each tube including the blank. Mix well and allow to stand for 10 min.
5. Then add 0.5 ml of reagent D, mix well and incubate at room temp in the dark for 30 min. Blue colour is developed.
6. Take the readings at 660 nm.
7. Draw a standard graph and calculate the amount of protein in the sample.

Calculation

Express the amount of protein mg/g or 100 g sample.

References

Isozymes in Plant Biology: Douglas E Soltis. Kluwer Acad. Publishers, Dordrecht, ISBN 0-412-38500-6

Lowry, OH; Rosebrough NJ; Farr, AL and Randall, RJ. (1951) J Biol Chem. 193:265

Hanes BD and Rickwood D: Gel Electrophoresis of proteins- A practical approach II Edition, OIRL Press, New York, 1994, 383p

Plant tissue culture and micropropagation

K Nirmal Babu and TE Sheeja

Introduction

Plant tissue culture is the science or art of growing plant cells, tissues or organs isolated from the mother plant, on artificial media. The culture initiation and plant regeneration are still accomplished empirically by varying conditions until the desired response is obtained. Schleiden and Schwann suggested the totipotency of cells and which states that cells are autonomic and are capable of regenerating a complete plant was in fact the foundation of plant cell and tissue culture. However it was Haberlandt took the first real steps in *in vitro* culture. Since then many investigators have studied plant regeneration under *in vitro* conditions and have obtained further understanding of cell totipotency.

Tissue culture is a collective term commonly used to describe all kinds of *in vitro* plant cultures. The tissue cultures can be cultures of an organized tissue, *viz.,* callus cultures, suspension or cell cultures, protoplast cultures, pollen cultures etc., or cultures of organized structures like meristem cultures, shoot and root cultures, embryo cultures, inflorescence cultures, ovule cultures etc.

Plant tissue culture has direct commercial applications as well as value in basic research into cell biology, genetics and biochemistry. The techniques include culture of cells, anthers, ovules and embryos on experimental to industrial scales, protoplast isolation and fusion, cell selection and meristem and bud culture.

Plant tissue culture involves growing of different tissues cells and organs on culture media containing essential nutrients, growth regulators and physical conditions conducive to growth and differentiation. Growth and differentiation is influenced by several factors, the most important being genotype and environment (George. 1993).

The Components of Culture media

Plant tissues are grown *in vitro* on artificial media with the nutrients necessary for growth. The success of plant tissue culture as means of plant propagation is greatly influenced by the nature of the culture medium used. The components of media for growth and development of plant callus and suspension cultures are classified in to six groups, namely i) major inorganic nutrients, ii) trace elements, iii) iron source, iv) organic supplements (vitamins), v) carbon source and vi) organic supplements/ plant growth regulators (Table.2).

Table 1. Composition of various tissue culture media (mg/l-1)

	Murashige & Skoog	**Gamborg**	**Nitsch**	**White**	**SH**
Macro nutrients					
$NH_4 NO_3$	1650				
$(NH_4)_2 SO_4$		134			
$NH_4H_2PO_4$					300
$CaCl_2 2H_2O$	440	150			200
Ca $(N0_3)_2$ $4H_2O$			500	300	
$MgSO_4$ 7H2O	370	250	250	720	400
KCl				65	
KNO_3	1900	2500	125	80	2500
KH_2PO_4	170		125		
$NaH_2PO_4H_2O$		150		16.5	
Na_2SO_4				200	

Micro nutrients					
$CoCl_26H_2O$	0.025	0.025			0.1
$CuSO_45H_2O$	0.025	0.025	0.025		0.2
$FeSO_47H_2O$	27.85	27.85			15.0
Na_2-EDTA	37.25	37.25			20.0
$Fe_2(SO_4)_3$				2.5	
$MnSO_4$ $4H_2O$	22.3	10.0	3.0	7.0	10.0
KI	0.83	0.75	0.5	0.75	1.0
$Na_2MoO_42H_2O$	0.25	0.25	0.025		0.1
$ZnSO_47H_20$	8.6	2.0	0.5	3.0	1.0
H_3BO_3	6.2	3.0	0.5	1.5	5.0
Fe-Citrate			10.0		
Vitamins					
Inositol	100	100.0			1000
Nicotinic acid	0.5	1.0		0.1	5.0
Pyridoxine-HCl	0.5	1.0		0.1	0.5
Thiamine-HCl	0.1	10.0	1.0	0.5	5.0
Glycine	2.0			3.0	
L-Cysteine-HCl			10		

Plant organs and tissues are most suitably retained above the surface of a culture medium by increasing its viscosity with some kind of gelling agent. It should be capable of being sterilized by autoclaving; thickened medium should be liquid when hot, viscous when cool. Agar is a preferred gelling agent for tissue cultures.

Plant growth regulators:

Growth and morphogenesis *in vitro* are determined by the interaction and balance between growth regulators supplied in the medium, and the growth substances produced endogenously. Five groups of plant hormones are generally recognized: auxins, gibberellins, cytokinins, abscisic acid and ethylene. These differ in their chemical nature and their physiological function (Table 2). Plant growth regulator levels are usually the most critical factor for successful de-differentiated growth of cultured plant cells. The optimum growth regulator requirement is species specific.

Table 2. Commonly used plant growth regulators and their function in plant tissue culture

Hormone	Name	Function in Plant Tissue Culture
Auxins	Indole-3-Acetic Acid Indole-3-Butyric Acid Indole-3-Butyric Acid, Potassium Salt á-Naphthaleneacetic Acid 2,4-Dichlorophenoxyacetic Acid p-Chlorophenoxyacetic acid Picloram Dicamba	Adventitious root formation (high concentration) Adventitious shoot formation (low concentration) Induction of somatic embryos Cell Division Callus formation and growth Inhibition of axillary buds Inhibition of root elongation
Cytokinins	6-Benzylaminopurine 6-ã,ã-Dimethylallylaminopurine (2iP) Kinetin Thidiazuron (TDZ) N-(2-chloro-4-pyridyl)-	Adventitious shoot formation Inhibition of root formation Promotes cell division Modulates callus initiation and growth Stimulation of axillary bud breaking and growth Inhibition of shoot elongation

	N'Phenylurea Zeatin Zeatin Riboside	Inhibition of leaf senescence
Gibberellins	Gibberellic Acid	Stimulates shoot elongation Release seeds, embryos, and apical buds from dormancy Inhibits adventitious root formation Paclobutrazol and ancymidol inhibit gibberellin synthesis thus resulting in shorter shoots, and promoting tuber, corm, and bulb formation.
Abscisic Acid	Abscisic Acid	Stimulates bulb and tuber formation Stimulates the maturation of embryos Promotes the start of dormancy
Polyamines	Putrescine Spermidine	Promotes adventitious root formation Promotes somatic embryogenesis Promotes shoot formation

Micro propagation

The main advantage of micropropagation is rapid clonal propagation and vegetative propagation where it is difficult or impossible by conventional means. In addition the modern techniques for genetic modification of plants depend heavily on micropropagation for plant regeneration and multiplication. Murashige (1974) has developed the concept of three developmental stages which was further expanded and is now accepted that there are five stages critical for successful micropropagation. (Debergh & Zimmerman, 1993).

Stage 0, the preparative stage, which emphasizes on raising mother plants under more hygienic and controlled conditions. The most often manipulated parameters are light, temperature and growth regulators. This helps in the development of reliable and repeatable micropropagation scheme and can reduce some of the contamination problems.

Stage 1: focuses on establishment of axenic and viable cultures. The age of the stock plant, developmental and the physiological stage of an explant as well as its size can determine the success of a procedure.

Stage 2: Multiplication – in this stage proliferation and maintenance of cultures is give more importance. For most species, the most desirable method of multiplication would be adventitious caulogenesis, because it can facilitate a substantially faster increase in propgules. However, one has to determine whether the adventitious bud system produces true-to-type plants. Regardless of the technique (adventitious or axillary caulogenesis, or even somatic embryogenesis) and plant employed, one should always evaluate the impact of mutations and variations. Sometimes a low frequency is desired, especially for variegated forms of ornamental foliage plants. Non-interesting mutations can occasionally be tolerated, depending on the efficiency of the system. The number of sub-cultures needs to be standardized for each species, as it may contribute to somaclonal variation.

The cultural environment, which is the result of the interaction between the plant material (which is in turn codetermined by the culture medium), the culture container and the external environment of the culture room, has without doubt a tremendous influence on a tissue culture system

Stage 3: In this stage one produces plantlets. Factors influencing this choice are commercial implications, available know-how and infrastructure. Only the quality and quantity of material established in the field will ultimately determine the approach to be used in this stage. Cultures showing visible abnormalities e.g. Vitrification, apex necrosis, absence of visible roots is eliminated to reduce variation. Rooting can be done both *in vivo* as well as in vitro depending on the crop and need.

Stage 4 - Transfer to green house conditions

It depends primarily upon the quality and the type of material produced in the previous stage. Co_2 enrichment, increase in light intensity, use of mist or fog, plastic tunnels, application of fungicides, etc. can help in acclimatization and hardening of plants. A gradual return to normal characteristics occurs during the acclimatization period in the green house. Because of the 'soft' nature, shoots and plantlets, which result from *in vitro* propagation, suffer high mortality if they are planted directly from culture vessel into the natural environment. Survival depends upon the ability of the cultured material to withstand water loss and carry out photosynthesis. It is greatly enhanced by gradual acclimatization and hardening (George, 1993). The delicate nature of plant material raised *in vitro* arises because, when cultured in conditions of comparatively low light fluency and high relative humidity, the anatomy and physiology of tissues are different to those of plants grown in the green house or field. In most species, leaves, which grow in vitro produce much less, wax in and on their outer cuticle, with abnormal stomata. For successful acclimatization, high humidity is essential. A minimum of 85% relative humidity should be maintained during the first few weeks. At the same time it is necessary to prevent plantlets from becoming overheated, or from receiving very high levels of solar radiation.

Micropropagation - a case study - ginger

Explant Source and Collection

For ginger the explant source and stage is the most important step for establishing contamination free cultures with good *in vitro* responses. The following steps are important.

1. Select the genotype, which is to be multiplied. Collect sufficient quantities of unadulterated pure clones. Ensure while collection the clumps from germinated seedlings in and around the bush are not collected.
2. The source plants are to be collected from disease and virus-free plantations and the individual plants are also checked to be free symptoms.
3. Plant the rhizome bits in greenhouse under protected conditions with 4-5 inches of sand on top and periodically (once in 20 days) spray / drench copper oxychloride/ bavistin/ Dithane M-45 at 0.3% to minimize the contamination.
4. Select the newly sprouting buds and use it as a source of explant for the culture.
5. Collect the explants from field directly into copper 0.3%oxychloride / bavistin solution and keep for 10-15 minutes before surface sterilization.
6. Uses of vegetative bud explants from rhizomes are most suited for direct clonal multiplication

.

Fig: 1 Sprouting buds used as explant.

Glassware

Borosil culture tubes are used for culture initiation and 500 ml borosil conical flasks as well as borosilicate glass bottles are ideal for multiplication and plantlet growth before harvesting plants for hardening. The culture vessels, tubes, bottles and flasks are closed with cotton plugs (made of non- absorbent cotton covered with cheese cloth), aluminum foil or polypropylene caps.

Instruments

Inoculations are made under aseptic conditions, in Klenzaids horizontal laminar airflow chamber. ‘Nat steel’ horizontal autoclave is used for sterilizing the culture media, stock solutions and other instruments like blades, forceps, needles etc needed to carry out the sterile operations.

For filter sterilization of thermo labile chemicals and enzymes, Millipore filter sterilization system with 0.22µ pore size filter was used.

Distilled water

Double glass distilled water or Millipore ‘Milli Q’ water can be used for stocks preparation, growth regulators, enzyme solutions, media, buffers etc.

Carbon source

Sucrose (Qualigens, Bombay) is used as the carbon source at the rate of 30 gml^{-1}.

Gelling agents

For solidifying the culture medium, bacteriological grade agar agar was used at the concentration of 7.0 gl^{-1}.

Basal nutrient medium

MS (Murashige and Skoog, 1962) was used in all experiments. For MS medium, separate stocks were prepared for macronutrients, micronutrients, vitamins, amino acid and growth regulators (Tables 1 and 2). Sucrose at a concentration of 30 gl^{-1} and agar agar (7.0 gl^{-1}) was added directly to the medium. pH was adjusted to 5.8 before adding agar. The agar was melted to ensure uniform distribution in the medium. The media were autoclaved at 121^0C at 16 psi for 20 minutes.

Growth regulators

Auxins: Two major auxins, namely α-naphthalene acetic acid (NAA) and Indole-3-butyric acid (IBA) were used in this study at concentrations of 0-1.0 mgl^{-1}.

Cytokinins: 2 cytokinins namely 6-Benzylaminopurine (BA) and 6-fufurylamino purine (kinetin) were used at concentration of 0-1.0 mgl^{-1}. These were supplemented to the basal medium.

Preparation of culture media

Use Murashige and Skoog basal medium. Prepare separate stocks for macronutrients, micronutrients, vitamins, amino acid and growth regulators (Tables1 and 2).

1. Mix the stocks as mentioned in table add 30 gl^{-1} sucrose dissolve well make up to 1 Lit.
2. Adjust the pH 5.8 and add 6 gl^{-1} of agar, mix well melt it on heater to ensure uniform distribution.
3. Transfer it to culture tubes and flasks in the required quantity and autoclave at 121^0C at for 20 minutes.

Incubation conditions

The cultures were incubated at 22 ± 2^0C and with a photoperiod of 14 hours with a light intensity of 3000 lux, provided by ‘Philips’ cool white fluorescent tubes.

Culture Establishment

Rhizomes are to be collected from potted plats, after drenching with fungicide 3 days before collection for reducing rhizome-borne contaminants.

1. Select the newly sprouting buds (Fig.1) and use it as a source of explant for the culture. The buds are to be collected directly into 0.3% copper oxychloride and Teepol for 10-15 minutes followed by through wash in running tap water.
2. Transfer to clean flask with 0.1% mercuric chloride solution stir for 6-7 minutes or apply vacuum. in the laminar flow hood wash the explants thoroughly with sterile double distilled water and dry on sterile filter paper. Cut to required size usually 1 cm with 2-3 dormant buds or 3-4 mm with growing vegetative bud (Fig.2) and inoculate to the initiation medium.

Culture Initiation

For micropropagation, the explant chosen is mostly the apical or axillary bud. This is to maintain the genetic fidelity of micro propagated plants and to minimize the chances of obtaining off type plants. Plants regenerated from shoot apex cultures may occasionally include variants (Denton *et al.*, 1977).

Fig: 2 various stages of culture initiation

In vitro *multiplication and rooting*

Contamination free cultures, which exhibited initial bud growth, are transferred to MS medium supplemented with BAP (1.0mgl^{-1}) and IBA (0.5mgl^{-1}). About 100 ml of medium in 500 ml borosil flasks or glass bottles of same size are used. The rhizome buds produced shoots and roots simultaneously. 8-12 shoots per culture could be produced with an average of 2 roots per shoot, after 90 days of culture in this stage. 75% of the explants showed good shoot elongation and developed into full-fledged plants (Fig 3).

Fig: 3 *in vitro* Multiplication of ginger

The plantlets from the vegetative bud cultures were healthy, robust, 4- 8 cm tall with roots, depending on the culture vessel. The culture vessel has significant effect on the size of the plant let. The bigger the vessels bigger the plantlets.

Hardening and acclimatization

High water loss was recorded from leaves of transplanted tissue cultured plants immediately after transplantation. This is due to reduced quantity of epicuticular wax, high volumes of mesophyll intercellular space and slowness of stomatal response to water stress.

1. Carefully remove the clusters of plantlets from culture vessel with out damaging the leaves and roots. The leaves have a tendency to stick to culture vessel while taking them out in laminar airflow.

2. Select suitable well-grown plants for hardening. The remaining smaller plantlets can go for next round of multiplication. At this stage even liquid medium can be used for better response.
3. Well developed plantlets (each unit should have at least 2 suckers and 3-4 roots) are taken out into treys with water and carefully remove the traces of culture medium and are placed in water with 0.3% copper oxychloride / bavistin for 5 minutes before planting out. Care should be taken not to damage the roots.
4. Place the plantlets in m in trays suitable cups or micro pots or polybags and add a mixture of sand garden soil and vermiculite in equal proportions carefully nor to damage the roots and keep in humid chamber for hardening. The plantlets produce new leaves and roots within 30 days in humid chamber. Hardening was easy with 90% establishment due to the well-developed root system. The plants were maintained in nursery for 6 months before transferring them to field.
5. Place them in humid chamber and maintain high humidity for the first 15 days and slowly reducing the humidity to ambient.
6. The plants are ready for transfer to bigger polybags in to secondary nursery after 20 – 30 days.

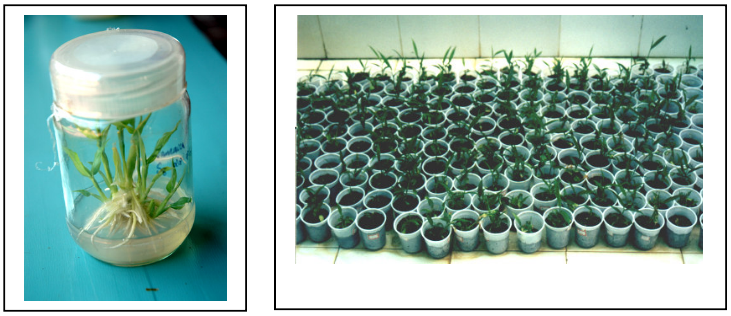


Fig: 4 a. Multiplied culture b. Hardening of in vitro derived plants.

Fig: 5 Field grown *in vitro* derived plants

A period (2–4 weeks) of acclimatization helps newly transferred plants to undergo morphological and physiological adaptations (Grout and Aston, 1977). Maintain them in secondary nursery for 6 months before planting in the field. A new and healthy sprout must germinate before they are planted in the field.

The *in vitro* grown rooted plantlets with 3- 4 roots per shoot can be transferred to suitable cups or micro pots or polybags containing a mixture of sand garden soil and vermiculite in equal proportions and keep in humid chamber for hardening. The plantlets produced new leaves and roots within 30 days in humid chamber. Hardening was easy with 90% establishment due to the well-developed root system. The plants were maintained in nursery for 6 months before transferring them to field.

Uses

Micropropagation is an important technology for large scale cloning of disease free plant material of elite genotypes. It also forms an important technology for in vitro conservation and Cryo preservation of crop genetic resources by slow growth as well as storage in liquid nitrogen. Somaclonal variation produced from cell and callus cultures forms an important source of variation especially in crops with narrow genetic variability. Micropropagation is an essential component in protoplast fusion and genetic transformation technologies.

References

Debergh P.C and Zimmerman R.H (1993) Micropropagation- *Technology and Application,* Kluwer Academic Publishers, The Netherlands.

George E.F (1993*) Plant Propagation by Tissue Culture Part 1 & 2*, Exegetics Limited, U.K.

Pierik R.L.M (1987) *In vitro culture of higher plants*. Martinus Nijhoff, The Netherlands.

Anonymous (1994, 1998). The Wealth of India, Publications and Information Directorate, New Delhi.

Debergh, P.C. and Read, P.E. 1991. Micropropagation. pp: 1–13. In. *Micropropagation. Technology and Application*. Debergh, P.C. and Zimmerman, R.H. (eds.). Kluwer Academic Publishers, Dordrecht. The Netherlands.

Denton, I.R., Westcott, R.J. and Ford–Lloyd, B.V. 1977. Phenotypic variation of *Solanum tuberosum* L. cv. Dr. Mclntosh regenerated directly from shoot–tip culture. *Potato Research*. 131–136.

Grout, B.W.W. and Aston, M.J. 1977. Transplanting of cauliflower plants regenerated from meristem culture I. Water loss and water transfer related to changes in leaf wax and to xylem regeneration. *Hort. Res*. **17**: 1–7.

Spices Board 1998. Large Cardamom. *A Decade of Research and Development*. Spices Board, Cochin. pp. 17-18.

Madhusoodanan, K. J. and Rao, Y. S. (2001). Cardamom (large). In K. V. Peter (ed.) *Handbook of herbs and spices.* pp - 134-142.

Mukherjee, D.K. 1972. Large Cardamom Cultivation in Darjeeling Dist. West Bengal, Cardamom News. 2 (11), 1-8.

Murashige T and Skoog F. 1962. A revised medium for rapid growth and bioassays with tobacco tissue cultures. *Physiol Plant* 15: 473–497.

Murashige, T. 1974. Plant propagation through tissue cultures. *Ann. Rev. Plant. Physiol.* **25**: 135–166.

Nirmal Babu K, Ravindran PN and Peter KV (eds.). 1997. Protocols for micropropagation of Spices and Aromatic Crops. Indian Institute of Spices Research. p.35.

Pangtey, US & Azad Thakur N.S 1986. Insect Pests of large cardamom in Sikkim,. Indian Forming 35 (12): 17-21.

Pruthi J.S., 1993. Spices and condiments NBT, New Delhi.

Roy, B.J. 1988.*Amomum cardamom* - the main cash crop of Sikkim and Himalayan region. Spices News letter, 22 (12): 32-39.

Sajina A, Mini PM, John CZ, Nirmal Babu K, Ravindran PN and Peter KV. 1997. Micropropagation of large cardamom (*Amomum subulatum* Roxb.). J. Spices and Aromatic Crops. 6 (2) : 145-148.

Singh, G.B. 1978. Large Cardamom, Cardamom. 10 (5): 3-13.

Analysis of Biological Sequences

S. J. Eapen & Reena N

Sequences and the concept of information

Protein and DNA can be similar with respect to their function, their structure, or their primary sequence of amino or nucleic acids. The general rule is that sequence determines shape, and shape determines function. So when we study sequence similarity we eventually hope to discover or validate similarity in shape and function. Sequences, for convenience of understanding and performing various mathematical functions, are seen as strings of characters.

Similarity in the sequences has both a quantitative and a qualitative aspect. A similarity measure gives a quantitative answer, saying that two sequences show a certain degree of similarity. An alignment is a mutual arrangements of two sequences, which is short of qualitative answer, it exhibits where the two sequences are similar, and where they differ. An optimal alignment, of course, is one that exhibits the most correspondences, and the least differences.

Biological motivations of sequence analysis

A large variety of the biologically motivated problems involve sequences or strings. Sequence alignment can be useful for discovering information related to functions, structure and evolution. For instance:

- Reconstructing long sequences of DNA from overlapping string fragments
- Determining physical and genetic maps from probe data under various experiments protocols.
- Storing and retrieving and comparing DNA strings
- Comparing two or more strings for similarities to find related proteins.
- Exploring frequently occurring patterns of nucleotides
- Finding informative elements in protein and DNA sequences
- Identify a known sequence
- Find other members of multigene families.

Many of these research problems aim at learning about functionality or the structure of protein without performing any experiments and without having to physically construct the protein itself. The basic idea is that similar sequences produce similar proteins. Thus, in order to predict the characteristics of a protein using only its sequence data. We can use the structure/function information on known protein with similar sequences available in databases.

For instance, when considering protein folding, it usually suffices that two protein sequences are identical at 25% of their positions for three dimensional structures to be almost identical. Classical example is the establishment of an association between cancer and uncontrolled cells growth. This discovery was enabled by comparing the sequence of a cancer-associated gene against the sequence of a protein that had already been known to be influencing cell growth. The correlation between these two sequences was very high, proving the connection between cancer and cellular growth.

Terminology

Identical: When a corresponding character is shared between two species or populations, that character is said to be identical.

Similar: The degree to which two species or populations share identities.

Homologous: When characters are similar due to common ancestry, they are homologous.

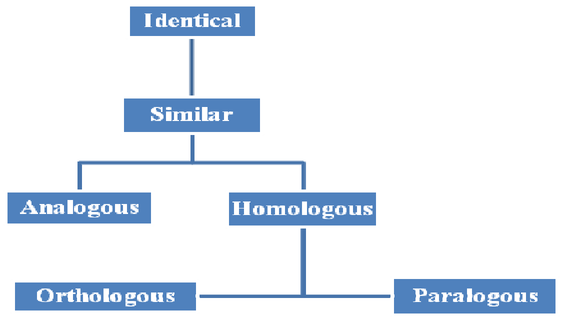
Analogous – When characters are similar due to convergent evolution, they are analogous.

Orthologous – When characters are homologous with conserved function, they are orthologous.

Paralogous – when characters are homologous with divergent function, they are paralogous.

Homology is therefore NOT synonymous with similarity, Homology is a judgment, and similarity is a measurement.

Similarity and Homology



Dot Plots

The dotplot is a simple, graphical way of displaying similarity of two sequences. To construct a simple dot plot, the first sequence to be compared is assigned to the horizontal axis of a plot space and the second is then assigned to the vertical axis.

A window length is fixed, together with a criterion when two sequence windows are deemed to be similar. Whenever one window in one sequence resembles another window in the other sequence, a dot or short diagonal is drawn at the corresponding position of the array. Thus, when two sequences share similarity over their entire length a diagonal line will extend from one corner of the dot plot to the diagonally opposite corner. If two sequences share only patches of similarity, this will be revealed by diagonal stretches. Such plots quickly become overly complex and crowded when large, similar sequences are compared. Sliding windows that consider more than just one position at a time are an effective way to deal with this problem.

Assume the window size of 10 with a similarity cutoff

- Window Size is number of bases in a sliding window that is moved along each sequence and compared to generate a single data point on the plot. *Window size must be an odd number*.
- Mismatch Limit (or cut off limit) determines how similar the two sequences in a window must be to "match". For example, if window size is 10 and mismatch limit is 2, then up to 2 mismatches in a 10 base window will still be classified as a match. At first, nucleotides 1-10 of the X-axis sequence are compared with nucleotides 1-10 of the sequence along the Y axis. If 8 or more of the 10 nucleotides (nt) in the first comparison are identical, a dot is placed in position (1,1) of the plot space. Next the window is advanced one nucleotide on the X axis, so that nucleotides 2-11 of the X axis sequence are now compared with 1-10 of the sequence along the Y axis. This procedure is repeated until each 10nt subsequence of the X axis has been compared.

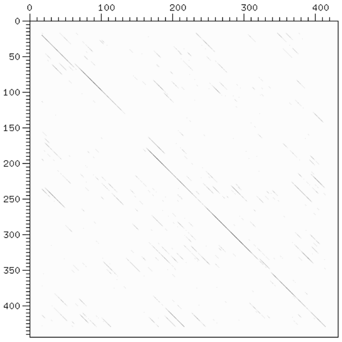
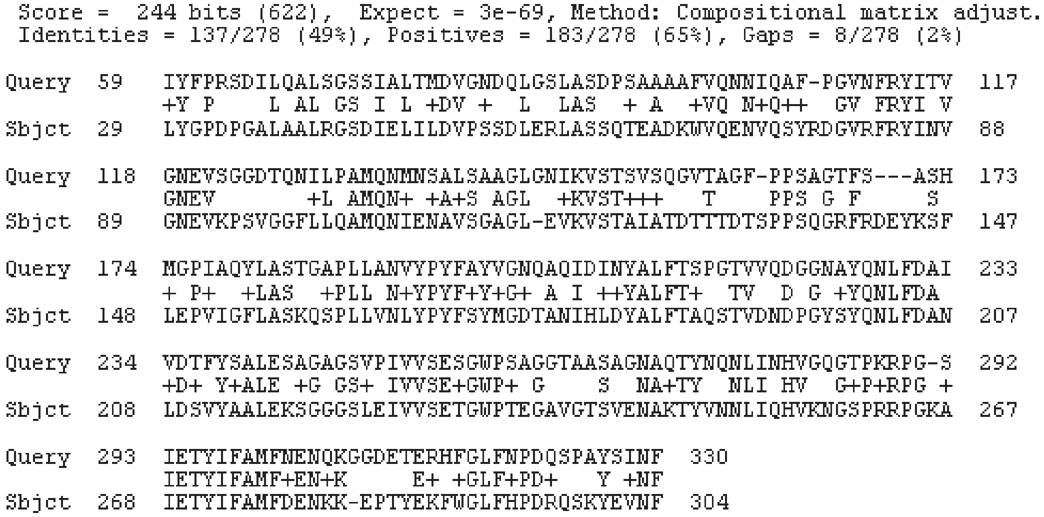


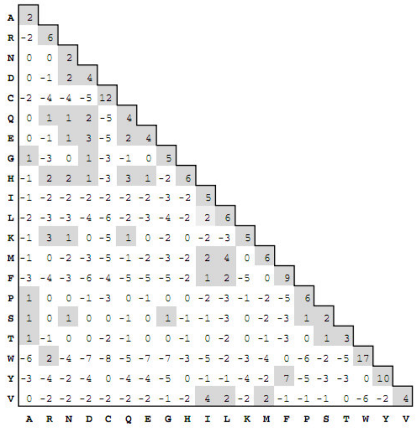
Fig 1: dot plot of two coding DNA-sequences: the alpha chain of human hemoglobin is assigned to the horizontal axis as the beta chain of human hemoglobin is assigned to the vertical axis.

Sequence Alignment:
Early in the days of protein and gene sequence analysis, it was discovered that the sequences from related proteins or genes were similar, in the sense that one could align the sequences so that many corresponding residues match. This discovery was very important, since strong similarity between two genes is a strong argument for their homology. Sequence alignment is one of the important sequence analyses in Bioinformatics. Here is an example of a sequence alignment for two proteins (beta-1,3-glucanase from Arabidopsis thaliana and beta-1,3-glucanase from Oryza sativa) using the BLAST program.



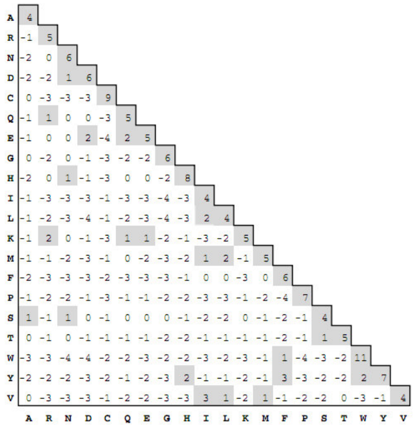
```
 Score =  244 bits (622),  Expect = 3e-69, Method: Compositional matrix adjust.
 Identities = 137/278 (49%), Positives = 183/278 (65%), Gaps = 8/278 (2%)

Query  59   IYFPRSDILQALSGSSIALTMDVGNDQLGSLASDPSAAAAFVQNNIQAF-PGVNFRYITV  117
            +Y P    L AL GS I L +DV +  L  LAS  + A  +VQ N+Q++  GV FRYI V
Sbjct  29   LYGPDPGALAALRGSDIELILDVPSSDLERLASSQTEADKWVQENVQSYRDGVRFRYINV  88

Query  118  GNEVSGGDTQNILPAMQNMNSALSAAGLGNIKVSTSVSQGVTAGF-PPSAGTFS---ASH  173
            GNEV       +L AMQN+ +A+S AGL  +KVST+++   T    PPS G F     S
Sbjct  89   GNEVKPSVGGFLLQAMQNIENAVSGAGL-EVKVSTAIATDTTTDTSPPSQGRFRDEYKSF  147

Query  174  MGPIAQYLASTGAPLLANVYPYFAYVGNQAQIDINYALFTSPGTVVQDGGNAYQNLFDAI  233
            + P+  +LAS  +PLL N+YPYF+Y+G+ A I ++YALFT+  TV  D G +YQNLFDA
Sbjct  148  LEPVIGFLASKQSPLLVNLYPYFSYMGDTANIHLDYALFTAQSTVDNDPGYSYQNLFDAN  207

Query  234  VDTFYSALESAGAGSVPIVVSESGWPSAGGTAASAGNAQTYNQNLINHVGQGTPKRPG-S  292
            +D+ Y+ALE +G GS+ IVVSE+GWP+ G    S  NA+TY  NLI HV  G+P+RPG +
Sbjct  208  LDSVYAALEKSGGGSLEIVVSETGWPTEGAVGTSVENAKTYVNNLIQHVKNGSPRRPGKA  267

Query  293  IETYIFAMFNENQKGGDETERHFGLFNPDQSPAYSINF  330
            IETYIFAMF+EN+K     E+ +GLF+PD+   Y +NF
Sbjct  268  IETYIFAMFDENKK-EPTYEKFWGLFHPDRQSKYEVNF  304
```

Scoring system is based on simple match/ mismatch scheme, we can increase sensitivity to weak alignments through the use of a substitution matrix. Protein sequences are frequently aligned using substitution matrices that reflect the probabilities of given character-to-character substitutions. A series of matrices called PAM matrices (Point Accepted Mutation matrices, originally defined by Margaret Dayhoff and sometimes referred to as "Dayhoff matrices") explicitly encode evolutionary approximations regarding the rates and probabilities of particular amino acid mutations. Another common series of scoring matrices, known as BLOSUM (Blocks Substitution Matrix), encodes empirically derived substitution probabilities. When calculating alignment scores, identical amino acids should be given greater value than substitutions, but conservative substitutions should also be greater than nonconservative changes.



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	12															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	6											
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5										
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6									
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5								
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6							
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9						
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6					
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2				
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3			
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17		
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4

PAM250 scoring matrix

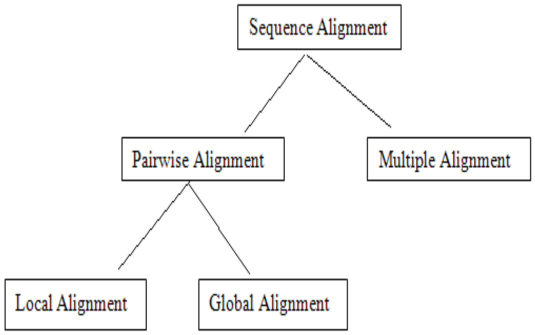


	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4																			
R	-1	5																		
N	-2	0	6																	
D	-2	-2	1	6																
C	0	-3	-3	-3	9															
Q	-1	1	0	0	-3	5														
E	-1	0	0	2	-4	2	5													
G	0	-2	0	-1	-3	-2	-2	6												
H	-2	0	1	-1	-3	0	0	-2	8											
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

BLOSUM62 scoring matrix

Types of sequence alignment:

Sequence alignment can be classified as pairwise and multiple. Pairwise alignment is of two types global and local sequence alignment.



Classification of sequence alignment

Global Alignment:

The global alignment is stretched over the entire sequence length to include as many matching amino acids as possible up to and including the sequence ends. Vertical bars between the sequences indicate the presence of identical residues.

```
V M N D W V I G K S S P G
J     I     I I I I     J
V N _  D G V I  G K I  T P  A
```

```
- - - - - V I G K - - - - -
          | | | |
- - - - - V I  G K - - - - -
```

Difference between global and local alignment

Local Alignment:

In a local alignment, the alignment stops at the ends of regions of identity or strong similarity, and a much higher priority is given to finding these local regions. Sequence alignment is useful for discovering functional, structural, and evolutionary information in biological sequences.

FASTA

FASTA is a powerful tool for scanning databases to find sequences that are similar to a query sequence. It is generally best to make protein – protein comparisons, but FASTA can also be used to compare DNA sequence to DNA databanks.

FASTA starts by making a generalization from the idea of dot plots. In a dot plot, regions of similarity between two sequences show up as diagonals. FASTA essentially calculates the sum of the dots along each diagonal. The "FAST" in fasta comes from the method of calculating those diagonal sums. If it were necessary to actually construct a dot plot matrix and then add along the diagonals for every sequence in the database, then FASTA wouldn't be any quicker than Smith-Waterman. Instead, FASTA uses a "word" based method. It looks for matching sequence patterns or words, called k-tubles, and then attempts to build a local alignment based upon these word matches. It makes a list of all words, (1 or 2 amino acids, or 5 or 6 nucleotides) in each sequences. It matches identical words from each list, and then creates diagonals by joining adjacent matches, but it only counts non-overlapping words. It then re-scores the highest scoring regions using a replacement matrix such as the PAM250, the best of these scores is called "init1". It then tries to join together the high scoring diagonals, allowing for gaps.

FASTA calculates an E() – value (expectation of significance). The final output plots the initial scores of each library sequences in a histogram ranked by the z-score which is derived from the opt score corrected for differences in sequence length. The general idea of this graph is to show a normal curve of Z-scores and E() values that allows you to see the typical values of these statistics for random matches versus the more significant matches at the very bottom of the graph. A list of the most significant scores follows the histogram and then the optimal alignments are displayed for these matches (the cutoff can be set by the user). The list of matches also contains, for each database sequence, the beginning and end positions of the region of significant similarity to your query sequence. It is also possible to force FASTA to show global alignments between the best hits and your query sequence rather than the local alignments used in its similarity calculations. There are several implementations and extensions of FASTA.

FASTA – compares a protein sequence to another protein sequence or a protein library or a DNA sequence to another DNA sequence or to a DNA sequence library.

TFASTA – compares a protein sequence to a DNA sequence or DNA sequence library by translating each DNA sequence into all 6 possible reading frames and then comparing each frame to the protein sequence.

LFASTA- identifies one or more regions of similarity between two sequences

PLFASTA – presents a dot matrix plot of regions of sequence similarity between two sequences.

FASTX and FASTY – translate a probe DNA sequence in three reading frames and compares all tree frames to a protein sequence database.

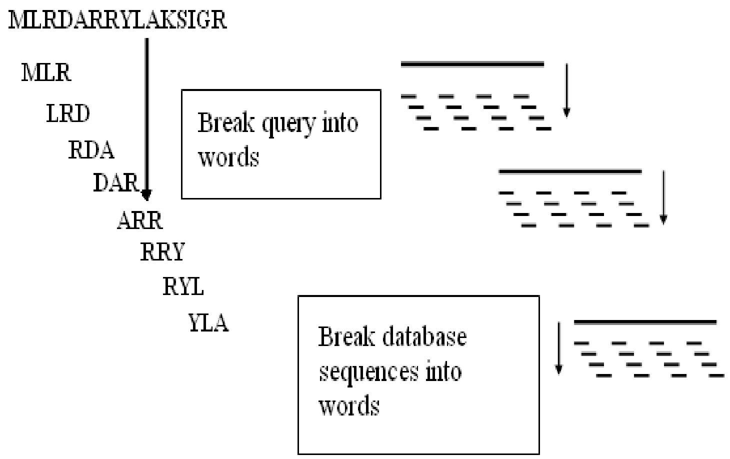
TFASTX and TFASTY – compare a probe protein sequences to a DNA sequence database, calculating similarities with frame shifts to the forward and reverse orientations.

BLAST

BLAST (Basic Local Alignment Search Tool) is a similarity search program developed by the research staff at NCBI\GenBank. It is available as a free service over the internet that provides very fast, accurate, and sensitive database searching.

How blast works: The basics

BLAST program uses a heuristic approach (a mathematical problem solving method). BLAST first creates a dictionary of three or four letter words from the query sequence (depending on Word size) and also creates a dictionary of works for database and searches for matching words in the database sequence. During the initial BLAST seeding, the algorithm finds all common words between the query sequence and the hit sequence(s). Only regions with a word hit will be used to build on an alignment. A substitution matrix is then consulted for assigning scores. It matches its dictionary words one after the other and looks for match score. After initial finding of words (seeding), the BLAST algorithm will extend the (only 3 residues long) alignment in both directions.



Word match method of BLAST

The steps can be summarized as follows:

1. Create dictionary of words for query sequence (According to the word size selected)
2. Scan the database and make a dictionary of database sequences
3. Once the program finds an initial match, it extends the analysis in both directions
4. Find maximum scoring pairs in sequence and prepare respective scores
5. From score S, the E (expectation) value is calculated.
6. Biologically significant hits are selected by selecting hits with lowest E value and high score.

Different BLAST programs are listed below.
Nucleotide Blast – Search a **nucleotide** database using a **nucleotide** query - *Algorithms:* blastn, megablast, discontiguous megablast.
Protein Blast - Search **protein** database using a **protein** query - *Algorithms:* blastp, psi-blast, phi-blast.
blastx - Search **protein** database using a **translated nucleotide** query
tblastn - Search **translated nucleotide** database using a **protein** query
tblastx - Search **translated nucleotide** database using a **translated nucleotide** query
Specialized Blast:

- Make specific primers with **Primer-BLAST**
- Search **trace archives**
- Find **conserved domains** in your sequence (cds)
- Find sequences with similar **conserved domain architecture** (cdart)
- Search sequences that have **gene expression profiles** (GEO)
- Search **immunoglobulins** (IgBLAST)
- Search for **SNPs** (snp)
- Screen sequence for **vector contamination** (vecscreen)
- **Align** two (or more) sequences using BLAST (bl2seq)
- Search **protein** or **nucleotide** targets in PubChem BioAssay

The output from BLAST can be enormous. For first run, you might want to restrict the number of hits returned with /LISTSIZE and /SEGMENTS. The first restricts the summary of hits at the top of the BLAST output, the second restricts the detailed alignments that follow. BLAST and FASTA produce similar output files. First there is a short description of the program and a list of the databases and program options chosen. Then there is a list of all of the database sequences that matched your query sequences. Several numbers are assigned to each of these sequences that represent the quality of the match. The list is presented in descending order so that the best matches are at the top of the list. However, the most biologically significant matches are not always the ones ranked highest in the list. The P-value is an indicator for the quality of a match. P-values smaller than e-100 (represents negative exponents) are exact matches (same gene, same species). P-value in the range of e-50 to e-100 is nearly identical genes (alleles, mutations, related species). P-value in the range of e-10 to e-50 is interesting closely related sequences. P-values between 0.1 and e-5 are still interesting but usually represent distant relationships. P-values greater than 0.1 are generally uninteresting or totally bogus matches. (Fig.2a,b,c,d)

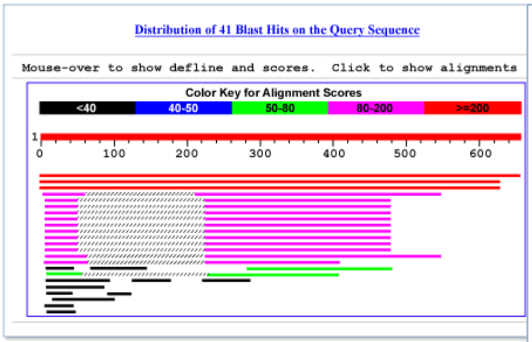
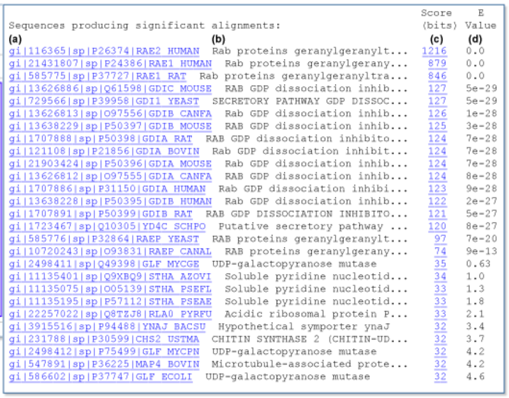
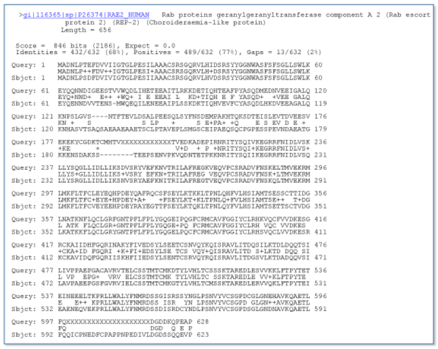

Distribution of 41 Blast Hits on the Query Sequence
Mouse-over to show defline and scores. Click to show alignments
Color Key for Alignment Scores
<40
40-50
50-80
80-200
>=200
1
0
100
200
300
400
500
600

Fig 2a.Graphical overview of BLAST results



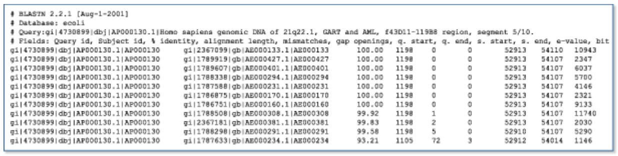
```
                                                                   Score    E
Sequences producing significant alignments:                       (bits)  Value
(a)                                  (b)                           (c)     (d)
gi|116365|sp|P26374|RAE2 HUMAN  Rab proteins geranylgeranylt...   1216    0.0
gi|21431807|sp|P24386|RAE1 HUMAN  Rab proteins geranylgerany...    879    0.0
gi|585775|sp|P37727|RAE1 RAT  Rab proteins geranylgeranyltra...    846    0.0
gi|13626886|sp|Q61598|GDIC MOUSE  RAB GDP dissociation inhib...    127    5e-29
gi|729566|sp|P39958|GDI1 YEAST  SECRETORY PATHWAY GDP DISSOC...    127    5e-29
gi|13626813|sp|O97556|GDIB CANFA  Rab GDP dissociation inhib...    126    1e-28
gi|13638229|sp|P50397|GDIB MOUSE  RAB GDP dissociation inhib...    125    3e-28
gi|1707888|sp|P50398|GDIA RAT  RAB GDP dissociation inhibito...    124    7e-28
gi|121108|sp|P21856|GDIA BOVIN  Rab GDP dissociation inhibit...    124    7e-28
gi|21903424|sp|P50396|GDIA MOUSE  Rab GDP dissociation inhib...    124    7e-28
gi|13626812|sp|O97555|GDIA CANFA  RAB GDP dissociation inhib...    124    8e-28
gi|1707886|sp|P31150|GDIA HUMAN  Rab GDP dissociation inhibi...    123    9e-28
gi|13638228|sp|P50395|GDIB HUMAN  Rab GDP dissociation inhib...    122    2e-27
gi|1707891|sp|P50399|GDIB RAT  RAB GDP DISSOCIATION INHIBITO...    121    5e-27
gi|1723467|sp|Q10305|YD4C SCHPO  Putative secretory pathway ...    120    8e-27
gi|585776|sp|P32864|RAEP YEAST  RAB proteins geranylgeranylt...     97    7e-20
gi|10720243|sp|O93831|RAEP CANAL  RAB proteins geranylgerany...     74    9e-13
gi|2498411|sp|Q49398|GLF MYCGE  UDP-galactopyranose mutase          35    0.63
gi|11135401|sp|Q9XBQ9|STHA AZOVI  Soluble pyridine nucleotid...     34    1.0
gi|11135075|sp|O05139|STHA PSEFL  Soluble pyridine nucleotid...     33    1.3
gi|11135195|sp|P57112|STHA PSEAE  Soluble pyridine nucleotid...     33    1.8
gi|22257022|sp|Q8TZJ8|RLA0 PYRFU  Acidic ribosomal protein P...     33    2.1
gi|3915516|sp|P94488|YNAJ BACSU  Hypothetical symporter ynaJ        32    3.4
gi|231788|sp|P30599|CHS2 USTMA  CHITIN SYNTHASE 2 (CHITIN-UD...     32    3.7
gi|2498412|sp|P75499|GLF MYCPN  UDP-galactopyranose mutase          32    4.2
gi|547891|sp|P36225|MAP4 BOVIN  Microtubule-associated prote...     32    4.2
gi|586602|sp|P37747|GLF ECOLI  UDP-galactopyranose mutase           32    4.6
```

Fig 2b.One-line descriptions in the BLAST report



```
 >gi|116365|sp|P26374|RAE2_HUMAN  Rab proteins geranylgeranyltransferase component A 2 (Rab escort
           protein 2) (REP-2) (Choroideraemia-like protein)
          Length = 656

 Score =  846 bits (2186), Expect = 0.0
 Identities = 432/632 (68%), Positives = 489/632 (77%), Gaps = 13/632 (2%)

Query: 1   MADNLPTEFDVVIIGTGLPESILAAACSRSGQRVLHIDSRSYYGGNWASFSFSGLLSWLK 60
            MADNLP++FDV++IGTGLPESI+AAACSRSGQRVLH+DSRSYYGGNWASFSFSGLLSWLK
Sbjct: 1   MADNLPSDFDVIVIGTGLPESIIAAACSRSGQRVLHVDSRSYYGGNWASFSFSGLLSWLK 60

Query: 61  EYQQNNDIGEESTVVWQDLIHETEEAITLRKKDETIQHTEAFPYASQDMEDNVEEIGALQ 120
            EYQ+NND+  E++ +WQ+ I E EEAI L  KD+TIQH E F YASQD+  +VEE GALQ
Sbjct: 61  EYQENNDVVTENS-MWQEQILENEEAIPLSSKDKTIQHVEVFCYASQDLHKDVEEAGALQ 119

Query: 121 KNPSLGVS----NTFTEVLDSALPEESQLSYFNSDEMPAKHTQKSDTEISLEVTDVEESV 176
            KN +   S             S LP   +      S E+PA+ +Q    E S EV D E +
Sbjct: 120 KNHASVTSAQSAEAAEAAETSCLPTAVEPLSMGSCEIPAEQSQCPGPESSPEVNDAEATG 179

Query: 177 EKEKYCGDKTCMHTVXXXXXXXXXXXTVEDKADEPIRNRITYSQIVKEGRRFNIDLVSK 236
            +KE      +                V+D  + P +NRITYSQI+KEGRRFNIDLVS+
Sbjct: 180 KKENSDAKSS--------TEEPSENVPKVQDNTETPKKNRITYSQIIKEGRRFNIDLVSQ 231

Query: 237 LLYSQGLLIDLLIKSDVSRYVEFKNVTRILAFREGKVEQVPCSRADVFNSKELTMVEKRM 296
            LLYS+GLLIDLLIKS+VSRY EFKN+TRILAFREG VEQVPCSRADVFNSK+LTMVEKRM
Sbjct: 232 LLYSRGLLIDLLIKSNVSRYAEFKNITRILAFREGTVEQVPCSRADVFNSKQLTMVEKRM 291

Query: 297 LMKFLTFCLEYEQHPDEYQAFRQCSFSEYLKTKKLTPNLQHFVLHSIAMTSESSCTTIDG 356
            LMKFLTFC+EYE+HPDEY+A+   +FSEYLKT+KLTPNLQ+FVLHSIAMTSE++  T+DG
Sbjct: 292 LMKFLTFCVEYEEHPDEYRAYEGTTFSEYLKTQKLTPNLQYFVLHSIAMTSETTSCTVDG 351

Query: 357 LNATKNFLQCLGRFGNTPFLFPLYGQGEIPQGFCRMCAVFGGIYCLRHKVQCFVVDKESG 416
            L ATK FLQCLGR+GNTPFLFPLYGQGE+PQ FCRMCAVFGGIYCLRH VQC VVDKES
Sbjct: 352 LKATKKFLQCLGRYGNTPFLFPLYGQGELPQCFCRMCAVFGGIYCLRHSVQCLVVDKESR 411

Query: 417 RCKAIIDHFGQRINAKYFIVEDSYLSEETCSNVQYKQISRAVLITDQSILKTDLDQQTSI 476
            +CKA+ID FGQRI +K+FI+EDSYLSE TCS VQY+QISRAVLITD S+LKTD DQQ SI
Sbjct: 412 KCKAVIDQFGQRIISKHFIIEDSYLSENTCSRVQYRQISRAVLITDGSVLKTDADQQVSI 471

Query: 477 LIVPPAEFGACAVRVTELCSSTMTCMKDTYLVHLTCSSSKTAREDLESVVKKLFTPYTET 536
            L VP  EPG+  VRV ELCSSTMTCMK TYLVHLTC SSKTAREDLE VV+KLFTPYTE
Sbjct: 472 LAVPAEEPGSFGVRVIELCSSTMTCMKGTYLVHLTCMSSKTAREDLERVVQKLFTPYTEI 531

Query: 537 EINEEELTKPRLLWALYFNMRDSSGISRSSYNGLPSNVYVCSGPDCGLGNEHAVKQAETL 596
            E   E++ KPRLLWALYFNMRDSS ISR  YN LPSNVYVCSGPD GLGN++AVKQAETL
Sbjct: 532 EAENEQVEKPRLLWALYFNMRDSSDISRDCYNDLPSNVYVCSGPDSGLGNDNAVKQAETL 591

Query: 597 FQXXXXXXXXXXXXXXXXXXXXXDGDDKQPEAP 628
            FQ                     DGD  Q E P
Sbjct: 592 FQQICPNEDFCPAPPNPEDIVLDGDSSQQEVP 623
```

Fig 2c.A pairwise sequence alignment from a BLAST report



```
# BLASTN 2.2.1 [Aug-1-2001]
# Database: ecoli
# Query:gi|4730899|dbj|AP000130.1|Homo sapiens genomic DNA of 21q22.1, GART and AML, f43D11-119B8 region, segment 5/10.
# Fields: Query id, Subject id, % identity, alignment length, mismatches, gap openings, q. start, q. end, s. start, s. end, e-value, bit
gi|4730899|dbj|AP000130.1|AP000130    gi|2367099|gb|AE000133.1|AE000133    100.00  1198   0    0     52913  54110  10943
gi|4730899|dbj|AP000130.1|AP000130    gi|1789919|gb|AE000427.1|AE000427    100.00  1198   0    0     52913  54107  2347
gi|4730899|dbj|AP000130.1|AP000130    gi|1789607|gb|AE000401.1|AE000401    100.00  1198   0    0     52913  54107  6037
gi|4730899|dbj|AP000130.1|AP000130    gi|1788338|gb|AE000294.1|AE000294    100.00  1198   0    0     52913  54107  5700
gi|4730899|dbj|AP000130.1|AP000130    gi|1787588|gb|AE000231.1|AE000231    100.00  1198   0    0     52913  54107  4146
gi|4730899|dbj|AP000130.1|AP000130    gi|1786875|gb|AE000170.1|AE000170    100.00  1198   0    0     52913  54107  2321
gi|4730899|dbj|AP000130.1|AP000130    gi|1786751|gb|AE000160.1|AE000160    100.00  1198   0    0     52913  54107  9133
gi|4730899|dbj|AP000130.1|AP000130    gi|1788508|gb|AE000308.1|AE000308    99.92   1198   1    0     52913  54107  11740
gi|4730899|dbj|AP000130.1|AP000130    gi|2367181|gb|AE000381.1|AE000381    99.83   1198   2    0     52913  54107  2030
gi|4730899|dbj|AP000130.1|AP000130    gi|1788298|gb|AE000291.1|AE000291    99.58   1198   5    0     52910  54107  5290
gi|4730899|dbj|AP000130.1|AP000130    gi|1787633|gb|AE000234.1|AE000234    93.21   1105   72   3     52912  54014  1146
```

Fig 2d.BLAST output in hit table format

E-value

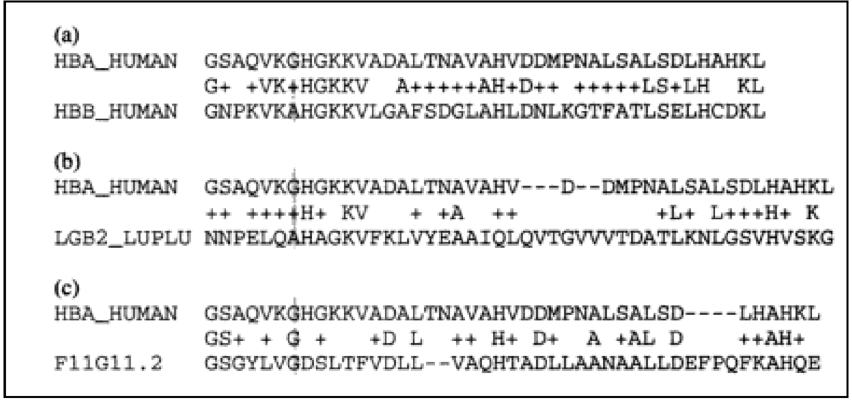
E-value, is the expectation value. The lower the E-value, the better the hit. The E-value is dependent on the length of the query sequence and the size of the database. For example, an alignment obtaining an E-value of 0.05 means that there is a 5 in 100 chance of occurring by chance alone.Short identical sequence may have a high E-value. Lower the e-value, more the sequences are biologically related.

- E-value < 10e-100 Identical sequences.
- 10e-50 < E-value < 10e-100 Almost identical sequences.
- 10e-10 < E-value < 10e-50 Closely related sequences, could be a domain match or similar.
- 1 < E-value < 10e-6 Could be a true homologue.
- E-value > 1 Proteins are most likely not related
- E-value > 10 Hits are most likely junk unless the query sequence is very short.

Pairwise Alignment

The most basic sequence analysis task id is to ask if two sequences are related. This is usually done by first aligning the sequences (or parts of them) and then deciding whether that the alignment is more likely to have occurred because the sequence are related, or just by a chance. The key issues are: 1) what sorts of alignment should be considered; 2) The scoring system used to rank alignments; 3) the algorithm used to find optimal (or good) scoring alignments; and 4) the statistical methods used to evaluate the significance of an alignment score.

Figure 3 shows an example of three pair wise alignments, all to the same region of the human alpha globin protein sequence (SWISS – PROT database identifier HBA_HUMAN). The central line in each alignment indicates identical positions with letters and ‘similar’ positions with a plus sign. (‘similar’ pairs of residues are those which have a positive score in substitution matrix used to score the alignment) (Fig. 3).



```
(a)
HBA_HUMAN  GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKL
           G+ +VK+HGKKV  A+++++AH+D++ +++++LS+LH  KL
HBB_HUMAN  GNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKL

(b)
HBA_HUMAN  GSAQVKGHGKKVADALTNAVAHV---D--DMPNALSALSDLHAHKL
           ++ ++++H+ KV   + +A  ++          +L+ L+++H+ K
LGB2_LUPLU NNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG

(c)
HBA_HUMAN  GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSD----LHAHKL
           GS+ + G +   +D L  ++ H+ D+  A +AL D    ++AH+
F11G11.2   GSGYLVGDSLTFVDLL--VAQHTADLLAANAALLDEFPQFKAHQE
```

Fig: 3: Three sequence alignments to a fragment of human alpha globin. a) Clear similarity of human beta globin. b) A structurally possible alignment to leghaemoglobin from yellow lupin c) A spurious high -scoring alignment to a nematode glutathione s-transferase homologue named F11G11.2.

Multiple sequence alignment methods

Multiple alignment is a powerful method for identifying and visualizing conserved regions in family of homologous proteins. Moreover, all phylogenetic methods need multiple alignment to start with. The 'correctness' of this alignment determines the 'correctness' of resulting phylogenetic tree. We have difficulties with exhaustive alignment of large number of sequences. Therefore, in 1987 an alternative approach was suggested. This method utilizes the progressive pairwise alignment algorithm iteratively to achieve the multiple alignment of a set of protein sequences and to construct an evolutionary tree depicting their relationship. The closest sequences according to this evolutionary tree are aligned first. Then this pairwise alignment is aligned to other sequences or to other alignments and so on until the final multiple alignment is put together. The sequences are assumed *a priori* to share a common ancestor. This approach is fast and relatively meaningful. By the nature, progressive alignment does not optimize any score and does not guarantee the optimal alignment. Nevertheless, the method is fast and with proper settings gives nice alignments with all kind of sequences. The masterpiece among the multiple alignment programs is ClustalW (and more recently ClustalX with graphical interface). More details on this are given in the chapter on ‘Multiple Sequence Alignment’.

Gene structure and DNA sequences

Eukaryotic genomes have introns, exons, coding region (CDS), untranslated region and phylogenetic analysis gives family relationship amongst species. The analysis is carried on small sections of aligned DNA taken from the same gene in different organisms. DNA is used, because the pattern of mutations, insertions and deletions at gene level is definitive. The silent mutations that do not result in mutations in amino acids level, are automatically introduced. Phylogenetic relationships are often represented graphically. A series of program packages are available. More details are given in the Chapter on 'Phylogenetic Analysis'.

Detecting open reading frame (ORF)

Finding a correct reading frame is not an easy task. It is considered normally as the longest reading frame uninterrupted by the stop codon (TGA, TAA or TAG). Such frame is called the open reading frame (ORF). Generally the initial codon is that of methionine (ATG). However, it is also a very common amino acid. It is necessary to use additional techniques to detect 5' untranslated sequence end. Recognition of flanking sequence is useful. In eukaryotic systems exons are part of transcribed sequences. Introns are transcribed but not a part of the coding sequence or CDs. If there is a gene made up of two introns and three exons, the exons are terminated by intron-exon boundaries. Thus at 5' junction we have sequences as: 'AGGTAAGT' while at 3' we might have Py Py Py Py Py PyNCAG (Py – Pyridine). Pattern of codon usage is important. Codon usage varies with the species. Codon usage pattern can be used to identify 5' and 3'. One can study forward translation and backward translation and come to the conclusion of the right gene.

There are several packages that translate DNA sequence to protein. Some examples are given below.

Translate – Translate a nucleotide sequence to a protein sequence.

Tanseq – Nucleotide to protein translation for EMBOSS package.

Graphical Codon Usage Analyser – Displays the codon bias in a graphical manner.

BCM search launcher – Six frame translation of nucleotide sequences

Back translation – Translates a protein sequence back to a nucleotide sequence.

Genewise – Compares a protein sequence to a genomic DNA sequences, allowing for introns and frame shifting errors.

FSED – Frame shift error detection.

LabOnWeb – Elongation, expression profiles and sequence analysis of ESTs using Compugen LEADS clusters.

For the computers, a reading frame is any stretch of a sequence which starts with a start codon and ends with a stop codon. In between, the protein is assumed. The situation is rather trivial if we analyse cDNA samples. However, keep in mind that side effects and sequencing errors might also be an issue in cDNA analysis.

Genomic sequence analysis: Detection of coding regions

The systematic sequencing analysis of genomes will result in long sequences which are unknown, whether or not these translate at all into a protein. Therefore, one of the prime targets of genomic sequence analysis will be to spot the location of splicing sites, coding regions, and intron/exon boundaries. Software for performing detailed analyses of codon usage variation is available from number of sites.

GCUA – General Codon Usage Analysis is available from ftp://ftp.nhm.ac.uk/pub/gcua.

A website dedicated to multivariate analysis of DNA sequences is available at http://acnuc.univ-lyon1.fr/mva/coa.html

A program Codon W is being released by John Peden at ftp://www.molbiol.ox.ac.uk/cu.

FORTRAN source code for a variety of programs is available at ftp://acer.gen.tcd.ie/pub/cod/

DNA sequence Assembly

An important aspect of DNA sequence analysis is to determine the sequence of a CLONE. Cloning errors can arise as a result of incorrect primer in PCR experiment. cDNA is synthesized using mRNA template. The clone is sequenced by designing primers to the known oligonucleotides present in the cloning vector. It is terminated by ddATP, ddGTP, ddCTP and ddTTP. Not all chains terminate at the same base. This results in a series of fragments for the same primer. ASSEMBLER programs build a consensus sequence.

EST Analysis

The major issue in DNA sequence analysis is that we do not have a full length DNA sequence. The available data is made up of partial sequences, majority of them are expressed sequence tags (ESTs). EST alphabet is five character- ACGTN N-is an ambiguous base.

- There may be phantom indels
- EST will often be a subsequence of other sequence in the data base.
- EST may not represent CDS of any gene.

There are various approaches to create EST libraries. Much of the public data is created in EST sections of GeneBank (dbEST) and EMBL. Suppliers of EST data submit their data to rigorous filtering processes prior to database submission.

Further reading

Altschul, SF., Gish, W, Miller, W, Myers, EW and Lipman DJ (1990). Basic local alignment search tool. J. Mol. Biol. 215(3): 403-10.

Birney, E (2000). Sequence Alignment in Bioinformatics, Ph.D. thesis, The Sanger Centre, Cambridge, U.K.

Birney, E (2001). Hidden Markov models in biological sequence analysis. IBM J. Res & Dev 45: 3/4

Eddy, SR (1996). Hidden Markov models. Curr. Opinion Structural Biol. 6: 361-5.

Hinxton (2004). Evolutionary analysis biological sequences. EMBL Research Report 363 – 366.Raffaele, G, Alessandro, S, Enrico, S and Filippo, U (2007). A basic analysis toolkit for biological sequences, Algorithms for Mol. Biol. 2:10.

Richard, D, Sean, RE, Anders, K and Graeme, M (1998). Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge University Press. pp. 356.

Thompson, JD, Higgins, DG and Gibson, TJ (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22(22): 4673-80.

Yang, Z (1996). Maximum-likelihood models for combined analysis for multiple sequence analysis data. J. Mol. Evol. 42: 587-96.

Zhu, J, Liu, JS and Lawrence, CE (1998). “Bayesian Adaptive Sequence Alignment Algorithms,” Bioinformatics 14: 25–39.

DNA microarray technology

A I Bhat

During the last three decades due to development of newer technologies significant progress has been made in the study of plant and animal genomes. In 1990s the DNA microarray technology was developed in the research area of genomics. This technology is popularly described with its various other names like Oligomer chip technology or Gene chip technology or DNA chip or Genomic chip technology. Microarrays can be used for analysis of entire genome of an organism in a single assay and believed to play in future the same role in the field of genomics, which in the past semiconductors played in the field of electronics. It is important to realize that biochips are not a single product but rather a family of product that form a technology platform. Many developments over the last two decades have contributed to its evolution. In early and mid 1990s companies such as Hyseq and Affymetrix (Santa Clara, CA) were formed to develop DNA array technologies. Micro arrays can be sued for expression analysis, single polymorphism detection, DNA sequencing by hybridization and genotyping on a genomic scale. Microarray based characterization of plant genome has the potential to revolutionize plant breeding and Agricultural biotechnology.

DNA Chip

DNA chip are microscopic arrays of single stranded DNA molecule immobilized on solid surface for biochemical analysis.

Principle

Hybridization of an unknown sample to an ordered array of immobilized DNA molecules of known sequence produces a specific hybridization patterns. That can be analysed or compared to a given sample.

Characteristic features of microarrays

1. Parallelism

Microarray analysis allows parallel acquisition and analysis of massive data. This greatly increases the speed of experimental work. It allows meaningful comparison among genes or gene products represented in microarrays and may be eventually allow the analysis of entire genome of any organisms in a single reaction.

2. Miniaturization

Microarray analysis involves miniaturization of DNA probes and reaction volumes thus reducing time and reagent consumption.

3. Speed

Microarray analysis is highly sensitive and allows rapid data acquisition with either confocal scanner or camera equipped with charged coupled device (CCD).

4. Multiplexing

Multiplexing is a process by which multiple samples are analysed in a single assay and also increase the accuracy of comparative anlaysis by eliminating complicating factors such as chip-to-chip variation, discrepancies in reaction condition.

5. Automation

Automation leads to proliferation of microarray analysis by ensuring their quality, availability and affordability.

Types of DNA chips

1. Oligonucleotide based chip

This type of DNA chip contains a high density of short oligonucleotides (10-25 bases) microarrays which are prepared by photolithography, a process in which light is used to direct the simultaneous synthesis of many different chemical compounds, synthesis occurs on a solid support. Arrays contain 1 lakh to 4 lakh oligonucleotides immobilized within area of 1.6 cm^2. This is used in large-scale analysis of single nucleotide polymorphism and sequencing by hybridization.

2. DNA based chip

These types of DNA chips contain a high density of DNA microarrays most often, derived from cDNA. Usually composed of 0.5 to 2.0 kb cDNA amplified by PCR. These chips are currently made by robotically spotting a large number of PCR amplified DNA fragments onto glass or nylon surface. It is mainly used in analysis of differential gene expression.

Production of microarrays

Currently two support media are being used for DNA arrays. They are nylon filter and glass. Three approaches are presently used for the production of microarrays. They include:

1. Photolithography

This is an *insitu* synthesis technique used for the production of oligonucleotide microarrays. Glass substrate is used as synthesis support. In this method, a glass wafer modified with photolabile protecting groups (X) is selectively activated for DNA synthesis by shining light through a photomask (M1). The wafer is then flooded with a photoprotected DNA base (A-X), resulting is spatially defined coupling on the chip surface. A second photomask (M2) is used to deprotect defined regions of the wafer. Repeated deprotection and coupling cycles enable the preparation of high-density oligonucleotide microarrays (see Fig 1).

2. Mechanical microspotting

It is used in the production of cDNA microarrays. In this a biochemical sample is loaded into a spotting pin by capillary action, and a small volume is transferred to a solid surface by physical contact between the pin and the solid substrate. After the first spotting cycle, the pin is washed and a second sample is loaded and deposited to an adjacent address. Robotic control systems and multiplexed printheads allow automated microarray fabrication (Fig 1).

3. Injetting

In this a biochemical sample is loaded into a miniature nozzle equipped with a piezoelectric fitting (rectangles) and an electrical current is used to expel a precise amount of liquid from the jet onto the substrate. After the first jetting step, the jet is washed and a second sample is loaded and deposited to an adjacent address. A repeated series of cycles with multiple jets enables rapid microarray production.

Target Preparation

Labeling of the target nucleic acid is often fluorescent (Cyanine dyes such as Cy3 and Cy 5) which have different excitation wavelengths. The use of 2, 3 and 4 colour detection to detect 2, 3 and 4 samples simultaneously has also been reported. Labeling is achieved without amplification of the target nucleic acid by incorporating label into cDNA using either flrourescently lableled nucleotides or primers in the reverse transcription reaction. Biotin can also be used to label target nucleic acid which can then be detected following hybridization by staining with streptavidin-conjugated Cy dye (Boonham et al., 2007). Several procedures are available for signal amplification to increase the sensitivity of detection.

Hybridization, washing and scanning of arrays

Hybridization of target nucleic acids in solution to immobilized capture probes is relatively slow process, which can take more than 24h to reach equilibrium. Hybridization temperature, the salt concentration, ph of hybridization buffer and stringency of subsequent washing steps must be optimized to achieve best sensitivity as well as reproducibility between experiments. The use of an automated hybridization station to agitate the sample can substantially reduce the hybridization time, as well as reducing the volume of sample required and automating the washing process.

The availability of high resolution laser confocal scanning has been instrumental in the development of microarray methods. It is important to ensure uniform spot morphology and avoid spotting artifacts in order to simplify image processing, ensuring high quality data, and potentially allow automated data acquisition.

Factors that affect Microarray hybridization

1. Concentration of targets and probes

When the probe concentration is ten fold greater than it's cognate species in the target mixture pseudo first order reaction kinetics ensue such that the hybridization rate is determined largely by target concentration. A two-fold increase in target concentration produces two fold increases in signal. Hybridization reaction involving probes concentration that is equal to or less than the target concentration display second order hybridization kinetics.

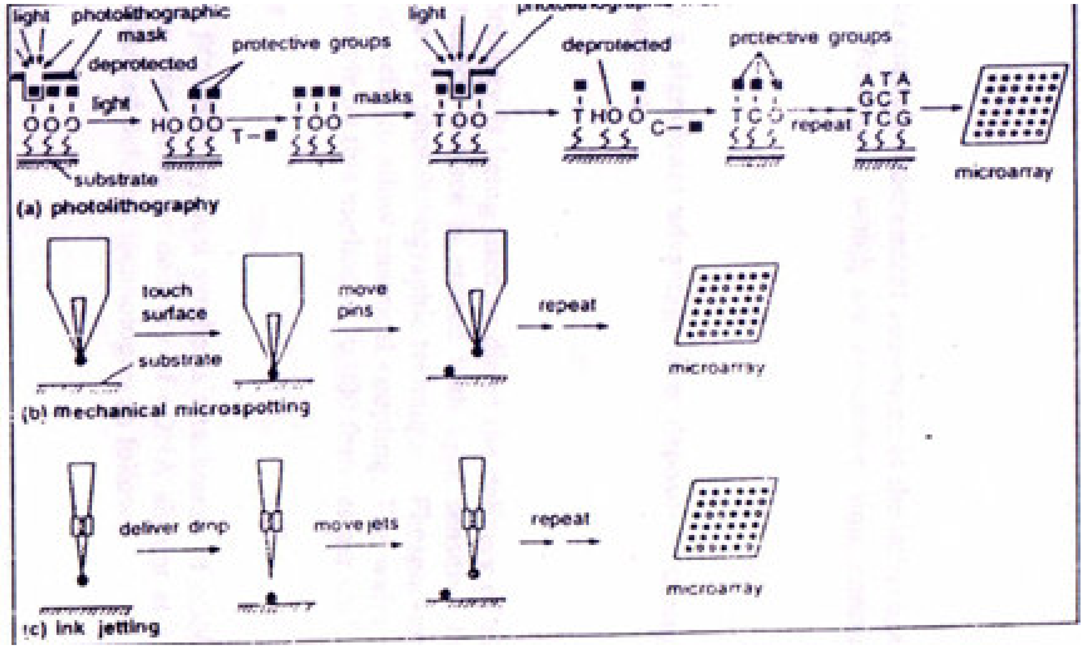


Fig. 1. Line diagram depicting principal steps in the production different kinds of microarrays

2. Concentration of salts

The presence of monovalent cations such as Na increases the rate of heteroduplex formation by shielding the negative charged phosphate backbone that would otherwise hinder base pairing interaction between targets and probe molecules. Typically Na ion concentration of 1 molar is used for microarrays experiment.

3. Temperature

Temperature exerts a positive effect on hybridization rate constant providing that the hybridization temperature is sufficiently below the melting temperature (Tm) of the heteroduplex. Hybridization temperature of 25 to 42^{o}C and 55 to 70^{o}C are typically used for oligonucleotides and cDNA microarrays experiments respectively.

4. Sequence composition

Sequence composition is a parameter over which the experiment has least control and is a much greater concerned with oligonucleotide than cDNA. Hybridization involving oligonucleotides are more sensitive to the GC content of individual heteroduplex than experiment involving longer DNA sequence. To minimize this problem tetra methyl ammonium chloride (TMAC) is used to equalize the binding energy of G: C and A: T.

Applications of DNA Microarray

Microarrays have a large number of applications some of which are given below:

1. Quantitative monitoring and differential gene expression

Microarrays can be used to monitor gene expression in samples of any biological origin including bacteria, fungi, higher plants and animals. Gene expression microarray technologies were developed for the first time using the small flowering plant, *Arabidopsis thaliana* as a model system. Microarrays for gene expression analysis provide an integrated platform for functional genomics. Changes in the physiological state of the cells and tissues used for microarray analysis lead to specific changes in gene expression patterns. Messenger RNA from samples of interest (inputs) is isolated, labeled and analysed by hybridization based microarray analysis, yielding quantitative expression information for thousands of cellular genes.

2. Sequencing by hybridization

Large number of oligonucleotides is immobilized on a solid surface, which can then be hybridized with a labeled target sequence. The method involves manufacturing the sequencing DNA chips that contain complete set of immobilized oligonucleotides of a particular size (eg. 8 mers) and hybridization of a target DNA of unknown sequence on to this DNA chips. The hybridization patterns are then recorded using confocal microscope. Identification and analysis of the overlapping oligomers that form perfect duplexes with a DNA of interest leads to reconstruction of the target DNA sequence.

3. Single nucleotide polymorphism (SNP) detection

The approach used for this purpose relies on the capacity to distinguish a perfect match from a single base mismatch. SNPs offer great promise for rapid and highly automated genotyping leading to rapid advancement in developing human genetic map.

4. Detection of mutants and polymorphism

DNA chips can be used for characterization of mutant populations. Mutation in cystic fibrosis transmembrane conductance regulator (CFTR) was studied using the Affymetrix chip. An array was designed to detect known deletions, insertions or base substitution mutations in exon 10 and 11 of CFTR.

5. Potential impact on agribusiness

Experiments involving the use of cDNA microarrays for expression monitoring in strawberry and petunia, indicate the immediate applicability of cDNA chips in agricultural biotechnology.

If the expression of genes on hormone application is known by DNA chips, then the hormone application can be monitored.

Transgenic plants can also be rapidly analsyed-using microarray on DNA chip and expression patterns and different environmental conditions can be predicted at gene level.

Microarray could be used to identify DNA polymorphism, which in turn could be used to expedite breeding.

Detection of difference in activity of various genes in healthy and infected plants with a particular pathogen.

Problems with microarray

Some of the problems associated with DNA microarray technology include:

1. Probe selection and target preparation
2. Equipment to be used to make and analyse microarrays
3. Problems in informatics

Conclusions

DNA microarray technology is used to analyse entire genome of an organism in a single assay. The application of this technology included sequencing by hybridization, detection of single nucleotide polymorphism, quantitative monitoring of gene expression, diagnostic chips, proteomics and also play a vial role in the field of agriculture. The DNA chip technology provides a variety of advantage so that the currently used gel based and filter based assay will be replaced in future by DNA chip based microarray analysis.

References

Agindotan B, Perry KL. 2002. Macroarray detection of plant RNA viruses using randomly primed and amplified complementary DNAs from infected plants. *Phytopathology* **97:** 119-127.

Boonham N, Walsh K, Smith P, Madagan K, Graham I, Barker I. 2003. Detection of potato viruses using microarray technology; towards a generic method for plant viral disease diagnosis. *J Virol Methods* **108:**181-187

Boonham N., Tomlinson J and Mumford R, 2007. Microarrays for rapid identification of plant viruses. *Annu Rev Phytopath*. **45:** 307-328.

Granheaud, S., Bertucci, F and Jordon, B.R. 1999. Expression profiling: DNA microarrays in many guises. BioEssays, 21: 781-790.

Heller, R.A., Schena, M., Chai, A., Shelon, D and Bedilion, T. 1997. Discovery and analysis of inflammatory disease related genes using cDNA microarrays. *Proc. Natl. Sci., USA* 94: 2150-2155.

Lemieux, B., Aharoni, A and Schena, M. 1998. Overview of DNA chips technology. *Molecular Breeding* 4: 277-289.

Ramsay, G. 1998. DNA chips: State of art. *Nature Biotech*, 16: 40-44.

Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. 1995. Quantitative monitoring of gene expression patterns with complementary DNA microarray. *Science* 270: 467-470.

Southern M.E.D. 1996. DNA chips: Analysing sequence by hybridization to oligonucloetides on a large scale. *Trends Genetics* 12: 110-115.

DNA markers in plant improvement - Amplified Fragment Length Polymorphisms (AFLPs) and Simple Sequence Repeats (SSRs)/Micro satellites

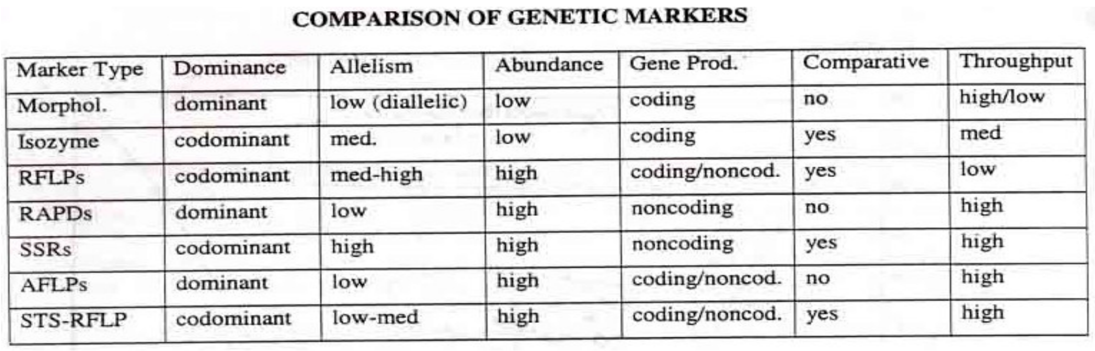
K. Nirmal Babu

In recent years the fast growing field of molecular biology has provided tools suitable for rapid and detailed genetic analysis of higher organisms. The most fundamental of these tools are DNA markers – detect differences in the genetic information carried by two or more individuals. Information from DNA markers serves many divergent purposes – forensic science, paternity testing, identifying the genes responsible for diseases and inferring evolutionary relationships, the most wide spread being in construction of genetic maps (Patterson *et al*., 1991).

A series of techniques and genetic markers have been developed to estimate genetic diversity, but no single technique is universally ideal; each available technique exhibits both strengths and weaknesses. Of the many molecular marker techniques available today, polymerase chain reaction (PCR)-based approaches are in demand because of their simplicity and requirement for only small quantities of sample DNA. Three widely-used PCR-based markers are RAPDs, SSRs ormicrosatellites, and AFLPs.

Each marker technique has its own advantages and disadvantages. RAPD markers are very quickand easy to develop (because of the arbitrary sequence of the primers) but lack reproducibility. AFLP has medium reproducibility but is labour intensive and has high operational and development costs. Microsatellites are specific and highly polymorphous, but they require knowledge of the genomic sequence to design specific primers and, thus, are limited primarily to economically important species. The choice of a molecular marker technique depends on its reproducibility and simplicity. The best markers for genome mapping, marker assisted selection, phylogenic studies, and crop conservation have low cost and labour requirements and high reliability (Fig 1). The choice of technique is often a compromise that depends on the research question pursued and the genetic resolution needed, as well as on financial constraints and the technical expertise available.

AFLP methods rapidly generate hundreds of highly replicable markers from DNA of any organism; thus, they allow high-resolution genotyping of fingerprinting quality. The time and cost efficiency, replicability and resolution of AFLPs are superior or equal to those of other markers except that AFLP methods primarily generate dominant rather than codominant markers. Because of their high replicability and ease of use, AFLP markers have emerged as a major new type of genetic marker with broad application in systematics, pathotyping, population genetics, DNA fingerprinting and quantitative trait loci (QTL) mapping.



COMPARISON OF GENETIC MARKERS

Marker Type	Dominance	Allelism	Abundance	Gene Prod.	Comparative	Throughput
Morphol.	dominant	low (diallelic)	low	coding	no	high/low
Isozyme	codominant	med.	low	coding	yes	med
RFLPs	codominant	med-high	high	coding/noncod.	yes	low
RAPDs	dominant	low	high	noncoding	no	high
SSRs	codominant	high	high	noncoding	yes	high
AFLPs	dominant	low	high	coding/noncod.	no	high
STS-RFLP	codominant	low-med	high	coding/noncod.	yes	high

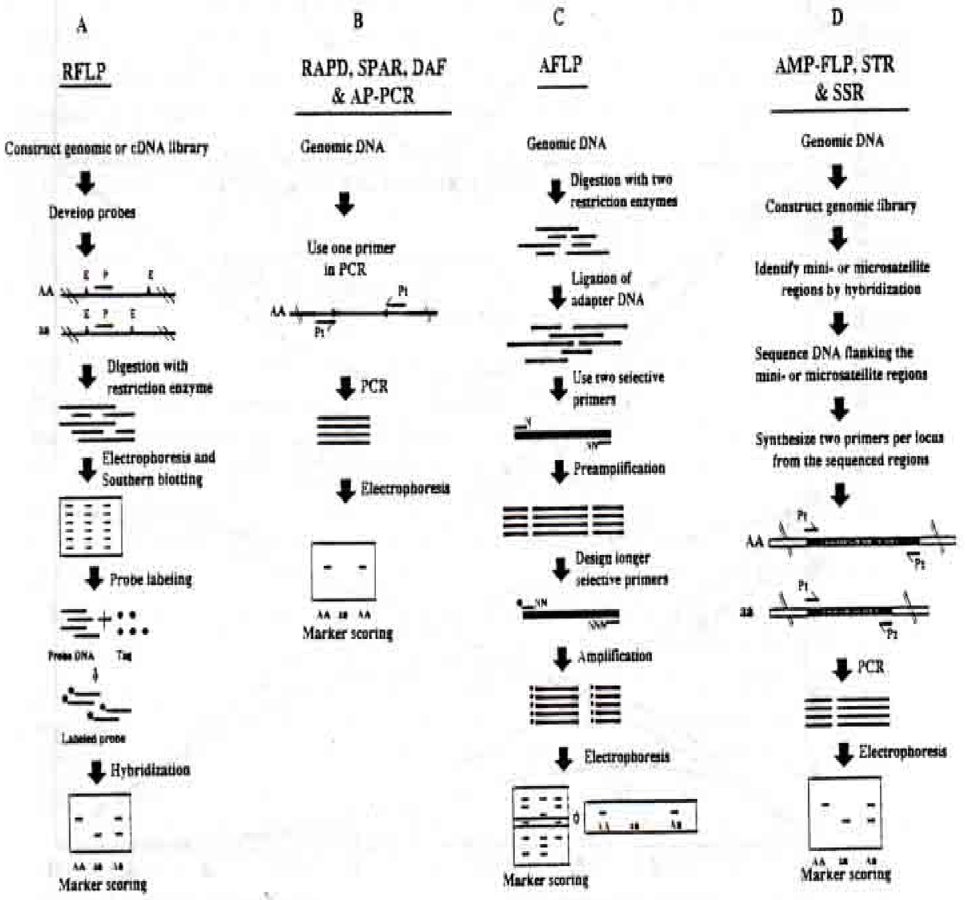

A
RFLP
Construct genomic or cDNA library
Develop probes
E P E
AA
aa
Digestion with restriction enzyme
Electrophoresis and Southern blotting
Probe labeling
Probe DNA
Tag
Labeled probe
Hybridization
AA aa Aa
Marker scoring
B
RAPD, SPAR, DAF & AP-PCR
Genomic DNA
Use one primer in PCR
P1
AA
PCR
Electrophoresis
AA aa AA
Marker scoring
C
AFLP
Genomic DNA
Digestion with two restriction enzymes
Ligation of adapter DNA
Use two selective primers
Preamplification
Design longer selective primers
Amplification
Electrophoresis
AA aa Aa
Marker scoring
D
AMP-FLP, STR & SSR
Genomic DNA
Construct genomic library
Identify mini- or microsatellite regions by hybridization
Sequence DNA flanking the mini- or microsatellite regions
Synthesize two primers per locus from the sequenced regions
P1
P2
AA
aa
PCR
Electrophoresis
AA aa Aa
Marker scoring

Fig. 1. Comparisons of various molecular marker technologies. In contrast to technologies presented in panels B and C, those in panels A and D require an expensive and time-consuming set-up phase for probe and primer development. Based on their availability, probes and primers can also be obtained from institutions or universities. DNA markers in panels B and C can be generated within a short time and require minor development. (A) E = restriction enzyme, P = probe, tag = either radioactive or nonradioactive for detection of hybridized probes. (B) P_1 = single primer of polymerase chain reaction (PCR). Although most PCR-based technologies amplify few to multiple bands, examples of only one band are given to illustrate scoring. (C) Adapter DNA = short double-stranded DNA molecules, 18 to 20 base pairs in length, representing a mixture of two types of molecules. Each type is compatible with one E generated DNA ends. Selective primers = primers for preamplification designed to contain an adapter DNA sequence plus one to two random bases at the 3' end for reading into the genomic fragments. Primers for amplification have the preamplification primer sequence plus one to two additional bases at the 3' end. A tag is attached at the 5' end of one of the amplification primers for detecting amplified molecules (multiple bands are also possible). Differences in scoring exist between bands in this panel and those of panel B (also see Fig. 2). (D) P1 and P2 = two PCR primers.

AFLP - Principles and applications

AFLP is a DNA fingerprinting technique, which detects DNA restriction fragments by means of PCR amplification. The AFLP technology usually comprises of the following steps

1. The restriction of the DNA with two restriction enzymes, preferably a hexa-cutter and a tetra-cutter;
2. The ligation of double-stranded (ds) adapters to the ends of the restriction fragments;
3. The amplification of a subset of the restriction fragments using two primers complementary to the adapter and restriction site sequences, and extended at their 3' ends by "selective" nucleotides
4. Gel electrophoresis of the amplified restriction fragments on denaturing polyacrylamide gels ("sequence gels");
5. The visualization of the DNA fingerprints by means of autoradiography, phospho-imaging, or other methods.

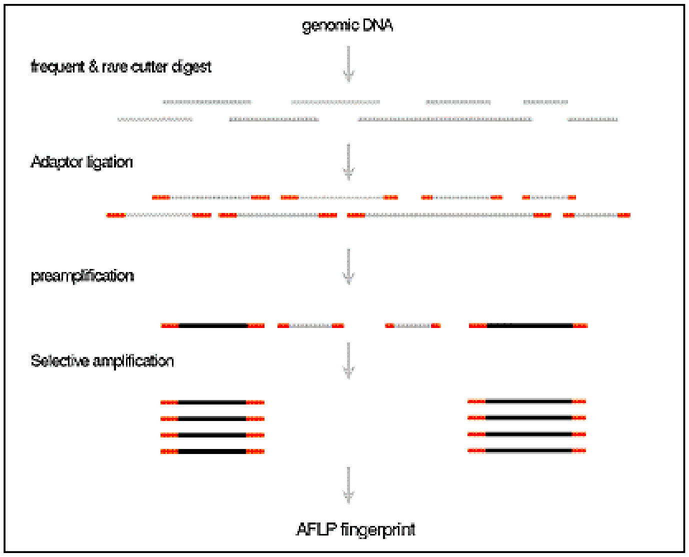
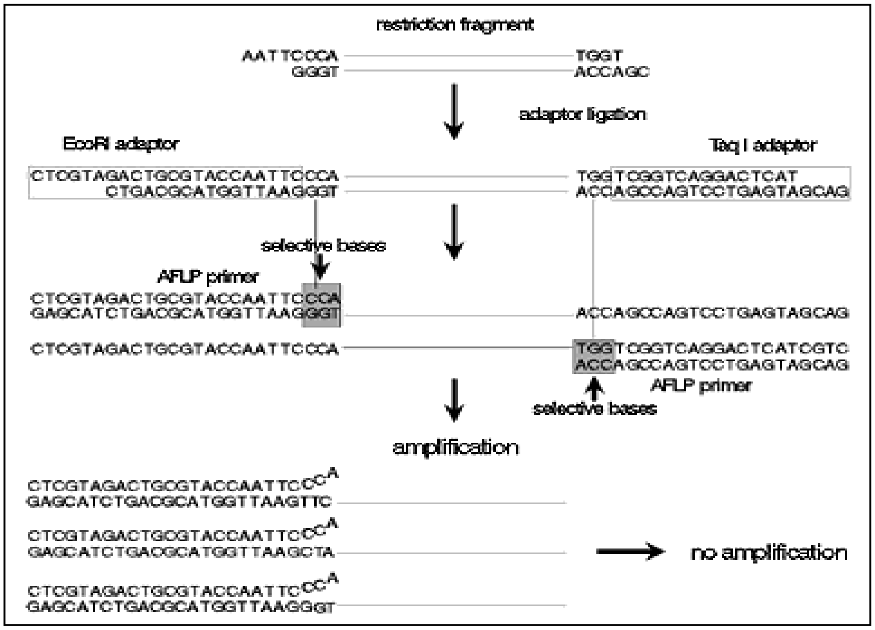


Fig. 2.

The AFLP technology is a random amplification technique, which, in contrast to most other random amplification techniques, makes use of stringent PCR conditions. The amplification primers, known as AFLP primers, are generally 17 - 21 nucleotides in length and anneal perfectly to their target sequences; i.e. the adapter and restriction sites, and a small number of nucleotides adjacent to the restriction sites *(*Fig. 3). This renders AFLP a very reliable and robust technique, which is unaffected by small variations in amplification parameters (e.g. thermal cyclers, template concentration, PCR cycle profile). The high marker densities that can be obtained with AFLP are an essential characteristic of the technology: a typical AFLP fingerprint contains between 50 and 100 amplified fragments, of which up to 80% may serve as genetic markers. Moreover AFLP technology requires no sequence information or probe collections prior to the generation of AFLP fingerprints. This is of particular benefit when studying organisms where very little DNA marker information is available. AFLP markers usually exhibit Mendelian inheritance, indicating that they are unique DNA fragments.

Restriction fragment patterns generated by means of the AFLP technology are known as AFLP fingerprints. These AFLP fingerprints are a rich source of restriction fragment polymorphisms, termed AFLP markers. The frequency with which AFLP markers are detected depends on the level of sequence polymorphism between the tested DNA samples. The molecular basis of AFLP polymorphisms will usually be sequence polymorphisms at the nucleotide level. Single nucleotide changes will be detected by AFLP when: **Fig. 3.**



1. The actual restriction sites are affected, 2. Nucleotides adjacent to the restriction sites are affected, which cause the AFLP primers to mispair at the 3' end, thus preventing amplification: the selective nucleotides are not an exact match of the sequence next to the restriction site. In addition, deletions, insertions and rearrangements affecting the presence or size of restriction fragments will result in polymorphisms detected by AFLP. Most AFLP markers will be mono-allelic markers; the corresponding allele is not detected. At a low frequency bi-allelic markers will be identified, as a result of small insertions or deletions in the restriction fragments.

Applications of AFLP technique

The AFLP technique can be used in a large number of applications, such as

1. The use of AFLP markers in genetic studies, such as
   a. Biodiversity studies,
   b. The analysis of germplasm collections,
   c. The genotyping of individuals, and genetic distance analyses,
   d. The identification of closely-linked DNA markers,
2. The construction of genetic DNA marker maps.
3. The use of AFLP in the precision mapping of genes, and the subsequent isolation of these genes.

AFLP markers have proved useful for assessing genetic differences among individuals, populations and independently evolving lineages.

Microsatellites

Simple sequence repeats (SSRs) also called microsatellites are stretches of DNA consisting of tandemly repeating mono, di, tri, tetra or penta nucleotide units that are arranged throughout the genomes of most eukaryotic species (Powell *et al.*, 1996). SSR loci harbour considerable length variation and are extremely abundant. They are individually amplflied by polymerase chain reaction from total genomic DNA, using a pair of oligonucleotide primers specific to the DNA flanking the SSR sequence and hence define the microsatellite locus. Amplification products obtained from different individuals can be resolved on gels to reveal polymorphism. The amplified products usually exhibit high levels of length polymorphisms, which result from variation between alleles in the number of tandemly repeating units of the locus (Tautz, 1989; Weber and May, 1989). The advantages of microsatellites are that they are relatively abundant with uniform genome coverage, high variable codominant, robust and reproducible, easy to detection by PCR, represent sequence tagged sites and require only small amount of starting DNA. Their high information content, which is directly related to the effective number of alleles at each locus and the ease of automating the PCR assays for identifying the Simple Sequnece Repeat polymorphisms make SSRs ideal genetic markers. But there is considerable difficulty in generating SSR markers compared to others as cloning and sequence information is necessary.

The most common method for isolation of SSRs involve - Creation of small insert genomic library, Screening of library for presence of microsatellites, sequencing of the positive clones, primer design and locus specific analysis and identification of polymorphisms (Rafalski *et al.*, 1996).

ATTTGTATTT TACAACACCT CACATGCTCA GTTATTTGGT TCATATGCAA
Forward Primer
GTCTCGGTTT TGGTCTCTGC TCAGAAAAAG **AGAGAGAGAG AGAGAGAGAG**
Reverse Primer
AGAGAGAGAG AGAGAGAGAA GAAATTTGCA GTTAATTGTC AAGTAGAAGT

Fig. 4. Soyabean library derived microsatelllite (AG) $_{20}$

Because of their sensitivity to minor genetic differences, PCR-based markers such as AFLPs and microsatellites are likely to remain key molecular tools for sometime to come.

Protocol for development of AFLP profiles

AFLP profiles were developed as per the method suggested by Vos *et al.,* (1995). The various materials, primers and stock solutions used are as follows.

Primers

EcoRI and *Mse*I primers, use at 50 ng /μL concentration.
*Eco*RI-primers +0 (E): 5'-AGACTGCGTACCAATTC-3'
*Eco*RI-primers +2 (E-AC): 5'-AGACTGCGTACCAATTCAC-3'
*Eco*RI-primers +2 (E-GG): 5'-AGACTGCGTACCAATTCGG-3'
*Eco*RI-primers +2 (E-TG): 5'-AGACTGCGTACCAATTCTG-3'
*Mse*I-primers +0 (M): 5'-GATGAGTCCTGAGTAA-3'
*Mse*I-primers +2 (M-AC): 5'-GATGAGTCCTGAGTAAAC-3'
*Mse*I-primers +2 (M-GC): 5'-GATGAGTCCTGAGTAAGC-3'
*Mse*I-primers +2 (M-TC): 5'-GATGAGTCCTGAGTAATC-3'
*Mse*I-primers +2 (M-TG): 5'-GATGAGTCCTGAGTAATG-3'
The numbers depict the number of selective nucleotides at the 3' terminus of the individual primers. The +0 nprimers were used in first amplification of restricted products and the +2 primers were used for second amplification in 4 different combinations, *i.e*., EAC-MTG, EGG-MTC, EGG-MGC and ETG-MAC.

Corresponding adaptors

EcoRI and MseI adapters, use at 5 and 50 pmol /μL concentration respectively.
*Eco*RI-adapter 5'-CTCGTAGACTGCGTACC
CATCTGACGCATGGTTAA-5'
*Mse*I-adapter 5'-GACGATGAGTCCTGAG
TACTCAGGACTCAT-5'

Others

Gamma-33P-ATP (~2000 Ci/mmol) - for radiolabeling.
Double-distilled water (ddH$_2$O).
Buffers: 1 M Tris.HAc pH 7.5; 1 M Tris.HCl pH 8.0 and pH 8.3.
Magnesium: 0.1 mM MgCl2.
TE (10x): 100 mM Tris.HCl, 10 mM EDTA pH 8.0.
100 mM DTT.
11. 5 mM of dNTPs
10 mM ATP.
T4-buffer (10x): 250 mM Tris.HCl pH 7.5, 100 mM MgCl2, 50 mM DTT, 5 mM spermidine
Restriction-ligation buffer (10x): 50 mM Tris-HAc, 50 mM MgCl$_2$, 250 mM KAc, 25 mM DTT, 250 ng/μL, pH 7.5.
Restriction endonucleases: *Eco*RI (20 U / μL), *Mse*I (4 U/ μL) (New England Biolabs).
Enzymes: T4 DNA ligase, T4 polynucleotide kinase, *Taq* DNA polymerase
PCR buffer (10x): 100 mM Tris.HCl pH 8.3, 15 mM MgCl$_2$, 500 mM KCl.
Molecular weight standards
General reagents for polyacrylamide gel electrophoresis.
10X TBE: Tris base 108 g, Boric Acid 55g, Na$_2$ EDTA 9.3 g in 1 L of dd H$_2$O, adjust pH 8.3 and autoclave.
Loading dye: same as for RAPD

Methods

Digestion of DNA

Genomic DNA (about 10 ng /μL) was digested with restriction endonucleases in a 50 μL reaction containing 25 μL DNA, 5 μL 10x RL buffer, 5 units *Eco*RI and 2 units *Mse*I, and 16 μL of dd H_2O, mixed well and incubated for 3 h at 37°C to obtain complete digestion with all the fragments below 500 bp.

Adaptor ligation

Adaptors were ligated to the digested genomic DNA by adding 1 μL *Eco*RI adapter (5 pmol), 1 μL *Mse*I adapter (50 pmol), 1.2 μL 10 mM ATP, 1 μL RL buffer, 1 unit T4 DNA ligase, 0.1M dTT and 3.79 μL water to the digestion mix. Incubate another 3 h at 37°C. Overall, DNA is incubated for a total of 6 h with endonucleases, the last 3 h in the presence of T4 DNA ligase and oligonucleotide adapters.

Dilution

The ligation reaction mixture was diluted 10 times with TE (usually 10 μL in 100 μL) and the diluted reaction mixture was used directly as template DNA for the AFLP reactions. Store diluted DNA at -20°C.

AFLP Pre-amplification

Preamplification was done (50 μL total volume) with following components
5 μL ligated DNA, 1.5 μL *Eco*RI-primer +0 (75 ng), 1.5 μL *Mse*I-primer +0 (75 ng), 2 μL 5 mM dNTPs, 0.2 μL Taq polymerase (1 unit), 5 μL 10x PCR-buffer and 34.8 μL water.
This mix was preamplified for 30 cycles in PCR machine using the following regime:
30 s at 94°C; 30 s at 60°C; 60 s at 72°C with a final extension of 120 s at 72°C.

Dilution

Preamplification, 10 μL of the reaction is diluted with 190 μL of TE0.1 to 100 μL, which is sufficient for 40 AFLP-reactions +2/+2. The diluted reaction mix and the rest of the preamplification reaction are stored at -20°C. If necessary new dilutions of the preamplification reactions may be made to give additional template for the AFLP reactions.

AFLP amplification

Preparation of labelling mix

Label primers for selective AFLP amplification by phosphorylating the 5' end of the primers with gamma-33P-ATP and polynucleotide kinase. Check above for the right primer combination to use in this step. Only one of the two primers of the AFLP reaction should be labelled (e.g., the EcoRI-primer). When possible use the more expensive 33P-labelled primers because they give better product resolution in polyacrylamide gels, and are less prone to degradation due to autoradiolysis. Prepare the following primer labelling mixes (40 μL) for 100 AFLP reactions. 10 μl gamma- 33P-ATP (~2,000 Ci/mmol), 5 μL 10xT4-buffer, 2 μL T4-kinase (10 units/μL) and water to 40 μL.

Primer labeling

Add 10 μl of primer (either *Eco*RI- or *Pst*I-primers at 50 ng/μL) to 40 μL labelling mix and incubate 60 min at 37°C, followed by incubation at 70°C for 10 min for the inactivation of the kinase. This gives a labelled primer with a concentration of 10 ng/μL.

Preparation of AFLP reaction mixes

Primer and *dNTPs* mix (50 μL): 5 μL labelled primer (10 ng/μL), 6 μL unlabelled primer (50 ng/μL), 8 μL 5 mM dNTPs and 31 μL water.
Taq polymerase mix (100 μL): 20 μL 10x PCR-buffer, 0.8 μL *Taq* polymerase (4 units) and 79.2 μL water.

AFLP amplification

Assemble the reaction by adding 5 µL of the primers and dNTPs mix and 10 µL of the *Taq* polymerase mix to 5 µL of pre-amplified ligated DNA. The template DNA should be pipetted first followed by the two mixes. The reagents should be mixed by tapping the base of the tubes on the bench. Pipetting mixes is essential for the rapid start of the AFLP reactions that are assembled at room temperature (to avoid loss of AFLP fingerprint quality). Tubes are amplified in a thermocycler with the following cycle regime a first cycle of 30 s at 94°C, 30 s at 65°C and 60 s at72°C, followed by 12 cycles with a stepwise decrease of the annealing temperature in each subsequent cycle by 0.7°C, and 23 cycles of 30 s at 94°C, 30 s at 56°C and 60 s at 72°C. The reaction is started at a high annealing temperature to obtain optimal primer selectivity. In the following steps the annealing temperature is lowered gradually to a temperature for optimal primer annealing.

Polyacrylamide gel electrophoresis of AFLP products

General: Amplification products were analyzed on 5% denaturing polyacrylamide-sequencing gels. The gels were cast at least 2 h before use and should be pre run for 30 m just before loading the samples Pre-running and running electrophoretic steps were done at 110 W. TBE (1x) was used as running buffer.

Sample loading: Mix AFLP reaction products with an equal volume (20 µL) of loading dye. Heat the samples for 3 min at 90°C, and then quickly cool on ice. Rinse the the gel wells with running buffer and push carefully two 24-well sharktooth combs about 0.5 mm into the gel surface to create the gel slots. Rinse the gel slots formed in this way with TBE and load 2 µL of each sample per well.

Post-electrophoretic procedures: Disassemble the gel cassette and remove the front glass plate with the silane-attached gel to the front dry it on gel drier. Autoradiographic exposure of the 33P-gels to standard X-ray film for 2 to 3 days with intensifying screens give good autoradiogrammes.

In the absence of radiographic facility silver staining can also be used. The protocol gets slightly modified from second amplification onwards (Promega, ….)

Protocol for developing microsatellites

Microsatellites can be amplified with specifically designed primers, if available for crop in question, using PCR and can be resolved on either acrylamide or high quality agarose gels both radioactive as well as non-radioactive methods can be used.

A simple method using non-radioactive PCR and agarose gel electrophoresis is given below;

PCR Amplification: Select the DNA of the population that is to be studied. Prepare PCR for each of the genotypes in the following method:

Per reaction	x 10	
10x PCR buffer:	2µl	20µl
1.25 mM dNTPs;	2µl	20µl
Forward Primer 5µM	1µl	10µl
Reverse Primer 5µM	1µl	10µl
Sterile H_2O	9.5µl	95µl
Taq Polymerase 1 Unit	0.5µl	5µl

Mix thoroughly, distribute 16µl into each PCR tube and add

DNA: 15-25ng 4 µl (5ng /µl) to each of the tube

Total reaction volume 20μl

Follow standard PCR with

Initial denaturation:	94^0C for 2 min	1 cylce
Denaturation, annealing and primer extension:	94^0C for 30 seconds	35 cycles
	50-55^0C for 30seconds	
	72^0C for 1 min	
Final Extension	72^0C for 5 min	1 cycle

Electrophoresis

Resolve the amplification product in 2% agarose gel in 1x TBE. Use the profile for analysis.

References

Jones CJ, Edwards KJ, Castaglione S, Winfield MO, Sale F, Van de Wiel C, BredemeijerG, Buiatti M, Maestri E, Malcevshi A, Marmiroli N, Aert R, Volckaert G, Rueda J, Linacero R, Vazquez A and Karp A (1997) Reproducibility testing of RAPD, AFLPand SSR markers in plants by a network of European laboratories. Mol Breed 3:381-390.

Karp, A., Seberg, O. and Buiatti, M. (1996) Molecular techniquesin the assessment of botanical diversity, *Ann. Bot*. 78,143–149

Lu, Z.X. *et al*. (1998) Construction of a genetic linkage map and identification of AFLP markers for resistance to root-knot nematodes in peach rootstocks, *Genome* 41, 199–207

Lynch, M. and Walsh, B. (1998) *Genetic Analysis of Quantitative Traits*, Sinauer

Patterson AH, Tanksley SD and Sorrels ME. (1991). DNA markers in plant improvement. Advances in Agronomy. Vol. 46. Academic Press. pp 40-90.

Powell W, Gordon CM and Provan J. (1996). Polymorphism revelaed by simple sequence repeats. Elsevier Publishers 1 (7) : 215.

Powell, W. *et al*. (1996) The comparison of RFLP, RAPD, AFLP andSSR (microsatellite) markers for germplasm analysis, *Mol. Breed*. 2,225–238.

Rafalski JA *et al*., (1996). Generating and using DNA markers implants, In, Analysis of Non mammelian Genomes – A Practical Guide (Birren E and Lai E eds.) Academic Press.

Rosendahl, S. and Taylor, J.W. (1997) Development of multiplegenetic markers for studies of genetic variation in arbuscular mycorrhizal fungi using AFLP, *Mol. Ecol*. 6, 821–829

Semblat, J.P. *et al*. (1998) High-resolution DNA fingerprinting of parthenogenetic root-knot nematodes using AFLP analysis, *Mol. Ecol.* 7, 119–125

Tautz D (1989) Hypervariability of simple sequences as a general source for polymorphic DNA markers. Nucleic Acids Res 17(16): 6463-6471.

Vos P, Hogers R, Bleeker M, Reijans M, Van der Lee T, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M and Zabeau M (1995) AFLP: a new technique for DNA fingerprinting.Nucleic Acids Res 23: 4407-4414.

Primer design and validation

A. Chandrasekar

The most critical parameter for successful PCR is the design of primers. All things being equal, a poorly designed primer can result in a PCR reaction that will not work. The primer sequence determines several things such as the length of the product, its melting temperature and ultimately the yield. A poorly designed primer can result in little or no product due to non-specific amplification and/or primer-dimer formation, which can become competitive enough to suppress product formation. This application note is provided to give rules that should be taken into account when designing primers for PCR.

Primer selection

Several variables must be taken into account when designing PCR primers. Among them, most critical are:

1. Primers should be 17-28 bases in length;
2. Base composition should be 50-60% (G+C);
3. Primers should end (3') in a G or C, or CG or GC: this prevents "breathing" of ends and increases efficiency of priming;
4. Tms between 55-80°C are preferred;
5. 3'-ends of primers should not be complementary (ie. base pair), as otherwise primer dimers will be synthesised preferentially to any other product;
6. Primer self-complementarity (ability to form 2^o structures such as hairpins) should be avoided;
7. Runs of three or more Cs or Gs at the 3'-ends of primers may promote mispriming at G or C-rich sequences (because of stability of annealing), and should be avoided.

Also keep in mind that most oligonucleotide synthesis reactions are only 98% efficient. This means that each time a base is added, only 98% of the oligos will receive the base. This is not often critical with shorter oligos, but as length increases, so does the probability that a primer will be missing a base. This is very important in mutagenesis or cloning reactions. Purification by HPLC or PAGE is recommended in some cases.

Oligonucleotide length	Percent with correct sequence
10 bases	$(0.98)^{10} = 81.7\%$
20 bases	$(0.98)^{20} = 66.7\%$
30 bases	$(0.98)^{30} = 54.6\%$
40 bases	$(0.98)^{40} = 44.6\%$

Primer length

Since both specificity and the temperature and time of annealing are at least partly dependent on primer length, this parameter is critical for successful PCR. In general, oligonucleotides between 18 and 24 bases are extremely sequence specific, provided that the annealing temperature is optimal. Primer length is also proportional to annealing efficiency: in general, the longer the primer, and the more inefficient the annealing. With fewer templates primed at each step, this can result in a significant decrease in amplified product. The primers should not be too short, however, unless the application specifically calls for it. As discussed below, the goal should be to design a primer with an annealing temperature of at least 50°C.

The relationship between annealing temperature and melting temperature is one of the “Black Boxes” of PCR. A general rule-of-thumb is to use an annealing temperature that is 5°C lower than the melting temperature. Thus, when aiming for an annealing temperature of at least 50°C, this corresponds to a primer with a calculated melting temperature(T_m) ~55°C. Often, the annealing temperature determined in this fashion will not be optimal and empirical experiments will have to be performed to determine the optimal temperature. This is most easily accomplished using a gradient thermal cycler like Eppendorf's Mastercycler gradient.

Melting Temperature (*T_m*)

It is important to keep in mind that there are two primers added to a PCR reaction. Both of the oligonucleotide primers should be designed such that they have similar melting temperatures. If primers are mismatched in terms of T_m, amplification will be less efficient or may not work at all since the primer with the higher T_m will mis-prime at lower temperatures and the primer with the lower T_m may not work at higher temperatures.

The melting temperatures of oligos are most accurately calculated using nearest neighbor thermodynamic calculations with the formula:

$T_m^{primer} = \Delta H\ [\Delta S + R \ln (c/4)] - 273.15°C + 16.6 \log_{10} [K^+]$

where H is the enthalpy and S is the entropy for helix formation, R is the molar gas constant and c is the concentration of primer. This is most easily accomplished using any of a number of primer design software packages on the market. Fortunately, a good working approximation of this value (generally valid for oligos in the 18–24 base range) can be calculated using the formula:

$T_m = 2(A+T) + 4(G+C)$.

Specificity

Primer specificity is at least partly dependent on primer length. It is evident that there are many more unique 24 base oligos than there are 15 base pair oligos. That being said, primers must be chosen so that they have a unique sequence within the template DNA that is to be amplified. A primer designed with a highly repetitive sequence will result in a smear when amplifying genomic DNA. However, the same primer may give a single band if a single clone from a genomic library is amplified.

Because *Taq* polymerase is active over a broad range of temperatures, primer extension will occur at the lower temperatures of annealing. If the temperature is too low, non-specific priming may occur which can be extended by the polymerase if there is a short homology at the 3' end. In general, a melting temperature of 55°C –72°C gives the best results (Note that this corresponds to a primer length of 18–24 bases using Wallace's rule above).

Complementary Primer Sequences

Primers need to be designed with absolutely no intra-primer homology beyond 3 base pairs. If a primer has such a region of self-homology, “snap back”, partially double-stranded structures, can occur which will interfere with annealing to the template.

Another related danger is inter-primer homology. Partial homology in the middle regions of two primers can interfere with hybridization. If the homology should occur at the 3' end of either primer, primer dimer formation will occur which, more often than not; will prevent the formation of the desired product via competition.

G/C content and polypyrimidine (T, C) or polypurine (A, G) stretches

The base composition of primers should be between 45% and 55% GC. The primer sequence must be chosen such that there is no PolyG or PolyC stretches that can promote non-specific annealing. Poly A and Poly T stretches are also to be avoided as these will “breath” and open up stretches of the primer-template complex. This can lower the efficiency of amplification. Polypyrimidine (T, C) and polypurine (A, G) stretches should also be avoided. Ideally the primer will have a near random mix of nucleotides, a 50% GC content and be ~20 bases long. This will put the T_m in the range of 56°C – 62°C.

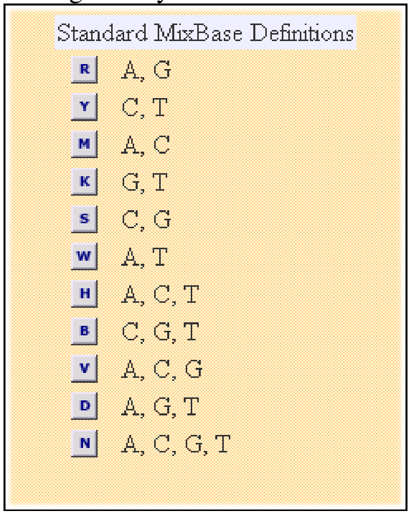
3’-end sequence

It is well established that the 3' terminal position in PCR primers is essential for the control of mis-priming. We have already explored the problem of primer homologies occurring at these regions. Another variable to look at is the inclusion of a G or C residue at the 3' end of primers. This “GC Clamp” helps to ensure correct binding at the 3' end due to the stronger hydrogen bonding of G/C residues. It also helps to improve the efficiency of the reaction by minimizing any “breathing” that might occur.

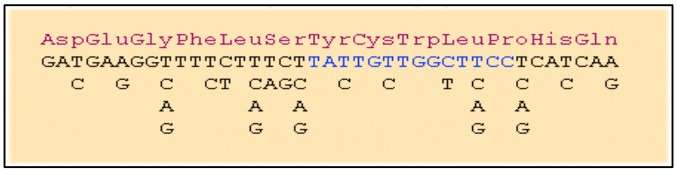
There are several web-based services or stand-alone software provided to the public for primer design, such as PRIDE, PRIMER MASTER, PRIMO, Primer3, Primer3Plus, Prime and Web Primer (http://genome-www2.stanford.edu/cgi-bin/SGD/web-primer), and Primer Design Assistant (PDA). Users can define the parameters listed in the menu of these tools and then get several pairs of primers for the target template sequence. However, most of them only take a single sequence query.

Degenerate Primer Design via Clustering

A nucleotide sequence is called **degenerate** if one or more of its positions can be occupied by one of several possible nucleotides. For example, AYGCNY is a sequence written down using IUPAC ambiguity codes, where Y stands for one of C or T, and N stands for A, C, G, or T. The *degeneracy* of a sequence is the number of different sequences that it represents. Thus, the degeneracy of AYGCNY is 16.



Standard MixBase Definitions	
R	A, G
Y	C, T
M	A, C
K	G, T
S	C, G
W	A, T
H	A, C, T
B	C, G, T
V	A, C, G
D	A, G, T
N	A, C, G, T



```
AspGluGlyPheLeuSerTyrCysTrpLeuProHisGln
GATGAAGGTTTTCTTTCTTATTGTTGGCTTCCTCATCAA
  C  G  C  CT CAGC  C  C   T C  C  C  G
        A     A  A           A  A
        G     G  G           G  G
```

One way to amplify a specific target sequence in a genome is to design a pair of primers (one forward and one reverse) that flank either end of the target sequence and to use the Polymerase Chain Reaction (PCR). If some pair of primer sequences is strongly conserved in several genomes, then the same pair of primers can be used to amplify the target region from all the genomes. However, when the primer sequences are only weakly conserved, “degenerate” primers (i.e., primers whose sequences are degenerate) are needed. Degenerate primers are particularly useful in amplifying homologous genes from different organisms. In the “candidate gene approach”, known genes that affect similar processes in one organism could have their homologues amplified in other related organisms by the use of a well-designed degenerate primer pair. Homologous genes display regions where they are highly conserved and also regions where they have evolved and are divergent. Primers can be found by searching in the highly conserved regions. In order to account for small mutations in the conserved regions, degenerate primers are used to match a large collection of similar sequences. Thus, degenerate primers can be used to isolate genes encoding proteins that belong to known protein families.

The first step in identifying primers for a set of homologous genes is to compute a sequence alignment of the genes or the protein sequences (or both). The next step is to identify at least two conserved regions in the alignment such that the target region to be amplified is contained in between them. Primers need to satisfy several properties. These include a feasible annealing temperature, an appropriate range for its GC-content, reasonably sticky ends that avoid degeneracy at the end of the primers, low degeneracy, and reasonable distance between two conserved regions. Therefore the last step involves finding a primer pair that best satisfies all the constraints.

Programs for the design of degenerate primers include GeneFisher, CODEHOP and HYDEN. GeneFisher uses a straightforward algorithm of starting with an amino acid sequence alignment and then back- translating conserved portions to obtain degenerate primers. CODEHOP designs “hybrid” degenerate primers that contain a short 3’ degenerate core region (about 11-12 bp) and a longer 5’ consensus clamp region (about 18-25 bp). It requires the input to contain a set of conserved amino acid blocks. CODEHOP then uses position-specific scoring matrices of aligned nucleotide sequences to design primers with low degeneracy. Both GeneFisher and CODEHOP work well for small sets of sequences that have strong consensus blocks. HYDEN, the most recent of the three, designs primers for aligned DNA sequences.

Primer validation

Designed primers can be validated using FAST PCR. In FAST PCR program check the primers length, quality, GC content and percentage. *In silico* PCR is used to check the primer amplification and product size. After checking the quality, primers are validated in wet lab condition for its amplification and compare the product size of both *in silico* wise and wet lab condition.

Primer3 plus:

Primer3 plus is advance software from primer3 which is useful to design and validate the primer. It is having so many options when compare to primer3. Primer check option will fulfill the need of primer validation. Just paste your primer in to primer test and click the check primer button if it is showing left primer is acceptable it is good primer otherwise it will show **Tm too low/High end self complementarity that primer is not** acceptable. This is one of the method for validating the primer.

Conclusion

It is essential that care is taken in the design of primers for PCR. Several parameters including the length of the primer, %GC content and the 3' sequence need to be optimized for successful PCR. Certain of these parameters can be easily manually optimized while others are best done with commercial computer programs. In any event, careful observance of the general rules of primer design will help ensure successful experiments.

Further reading:

1. Dieffenbach, C.W., Lowe, T.M.J., Dveksler, G.S., General Concepts for PCR Primer Design, in *PCR Primer, A Laboratory Manual*, Dieffenbach, C.W, and Dveksler, G.S., Ed., Cold Spring Harbor Laboratory Press, New York, 1995, 133-155.
2. Innis, M.A., and Gelfand, D.H., Optimization of PCRs, in *PCR protocols, A Guide to Methods and Applications*, Innis, M.A., Gelfand, D.H., Sninsky, J.J., and White, T.J., Ed., CRC Press, London, 1994, 5-11.
3. Sharrocks, A.D., The design of primers for PCR, in *PCR Technology, Current Innovations*, Griffin, H.G., and Griffin, A.M, Ed., CRC Press, London, 1994, 5-11.
4. Kwok, S., Kellog, D.E. McKinney, N., Spasic, D., Goda, L., Levenson, C., and Sninsky, J.J., Effects of primer-template mismatches on the polymerase chain reaction: Human Immunodeficiency Virus 1 model studies. *Nucleic Acids Res.* 18:999-1005, 1990
5. Linhart, C. and R. Shamir, The degenerate primer design problem. Bioinformatics, 2002. 18 Suppl 1: p. S172-81
6. Suggs, S.V., Hirose, T., Miyake, E.H., Kawashima, M.J., Johnson, K.I., and Wallace, R.B., Using Purified Genes, in *ICN-UCLA Symp. Developmental Biology*, Vol. 23, Brown, D.D. Ed., Academic Press, New York, 1981, 683.
7. Primer3 - http://frodo.wi.mit.edu/

8. Primerfinder - http://eatworms.swmed.edu/~tim/primerfinder/
9. Primer Premier 5 - http://www.PremierBiosoft.com/
10. CODEHOP- http://blocks.fhcrc.org/blocks/codehop.html
11. Fast PCR - http://www.biocenter.helsinki.fi/bi/bare-1_html/oligos.htm
12. PRIDE - http://pride.molgen.mpg.de/pride.html
13. Prime and Web Primer - http://genome-www2.stanford.edu/cgi-bin/SGD/web-primer
14. Proutski. V, Holmes. E.C. (1996) Primer Master: a new program for the design and analysis of PCR primers. *Comput Appl Biosci*. 12 (3): 253-5.
    PRIMO - http://www.changbioscience.com/primo/primo.html

Multiple sequence alignment

Santhosh J. Eapen and Reena N

Pair wise alignments are fundamental and useful, but there are some problems with them. For instance, when using one of the popular sequence searching programs (FASTA, BLAST) which perform pair wise alignments to find similar sequences in a database, one very often obtains many sequences that are significantly similar to the query sequence. Comparing each and every sequence to every other may be possible when one has just a few sequences, but it quickly becomes impractical as the number of sequences increases.

What we need is a multiple sequence alignment, where all similar sequences can be compared in one single figure or table. The basic idea is that the sequences are aligned on top of each other, so that a coordinate system is set up, where each row is the sequence for one protein, and each column is the 'same' position in each sequence. Each column corresponds to a specific residue in the 'prototypical' protein.

As with pairwise alignment, there will be gaps in some sequences, most often shown by the dash '-' or dot '.' character. Note that to construct a multiple alignment one may have to introduce gaps in sequences at positions where there were no gaps in the corresponding pair wise alignment. This means that multiple alignments typically contain more gaps than any given pair of aligned sequences.

As new nucleotide and protein sequences become available, by determining how these sequences are 'related' to known proteins, one can make predictions of their structural, functional and evolutionary features. The 'relation' between sequences can be level of sequence similarity or a common feature or domain in the sequences or associated structure etc.

Definition: Multiple Sequence Alignment (MSA) can simply be defined as an alignment of more than two sequences. Uses of MSA are plenty and some of them are listed below:

1. Detecting similarities between sequences (closely or distantly related)
2. Detecting conserved regions / motifs in sequences
3. Detection of structural homologies; patterns of hydrophobicity/ hydrophilicity , gaps etc thus assisting improved prediction of secondary and tertiary structures and loops and variable regions
4. Predict features of aligned sequences like conserved positions which may have structural or functional importance
5. Making patterns or profiles that can be further used to predict new sequences falling in a given family
6. Computing consensus sequence
7. Inferring evolutionary trees / linkage - Phylogenetic Analysis etc.
8. Deriving profiles or Hidden Markov Models that can be used to remove distant sequences (outliers) from protein families

Given so many uses MSA has become an essential tool in molecular biology. As such both nucleotide (DNA / RNA) and amino acid sequences can be used for multiple alignments but it’s preferable to use amino acid sequences when inferring phylogenetic relationship / building trees or building profiles or looking for structural homologies. If you have a nucleotide sequence then it can be translated (use PERL recipe or any web-based tool). The general procedure for doing a multiple sequence alignment comprises of following steps:

1. Searching for similar sequences in databases to identify all potential homologs (this can be done by the BLAST program)
2. Making an input file set (collection of similar sequences under investigation)
3. Computing the alignments with appropriate software.

4. Checking, analyzing and editing the alignment with any competitive alignment viewer.

A number of programs, which use different approaches, have been made for the purpose of multiple sequence alignment. Those who have ever done pair wise sequence alignment earlier using any program might have heard of the 'dynamic programming method' at some or the other time. If not, don't worry at the moment. The only point here is that although theoretically dynamic programming algorithms can be used to obtain mathematically optimum alignment (as in the case of pair wise alignment methods like Smith Waterman, Needleman-Wunsch), the computing power and time required become major limiting factors when larger alignments are involved. Therefore, for multiple sequence alignments different heuristic approaches (and different strategies like progressive, iterative, mixed etc.) have been used in various programs. Some of the most popular programs for multiple sequence alignment are MULTALIGN, Multal, PILEUP (Wisconsin Package), ClustalW, T-Coffee, Muscle etc. The performance of these different programs varies with the number of sequences, the degree of identity of sequences and the number of indels in the alignment.

How to generate a multiple alignment?

How do we generate a multiple alignment? There is an obvious solution: Given a pair wise alignment, just add the third, then the fourth, and so on, until all have been aligned. It depends not only on the various parameters (insertion/deletion penalties, substitution coefficients) but also on the order in which sequences are added to the multiple alignment.

In pair wise alignments, one has a two-dimensional matrix with the sequences on each axis, and the elements in the matrix are initially the substitution coefficients, which are then operated on to locate the best "path" through the matrix. The number of operations required to do this is approximately proportional to the product of the lengths of the two sequences. A possible general method would be to extend the pair wise alignment method into a simultaneous N-wise alignment, using a complete dynamical-programming algorithm in N dimensions. Algorithmically, this is not difficult to do. In the case of three sequences to be aligned, one can visualize this reasonably easy: One would set up a three-dimensional matrix (a cube) instead of the two-dimensional matrix for the pair wise comparison. Then one basically performs the same procedure as for the two-sequence case. This time, the result is a path that goes diagonally through the cube from one corner to the opposite.

The problem here is that the time to compute this N-wise alignment becomes prohibitive as the number of sequences grows. The algorithmic complexity is something like $O(c^{2n})$, where c is a constant, and n is the number of sequences. This is disastrous, as may be seen in a simple example: if a pair wise alignment of two sequences takes 1 second, then four sequences would take 10^4 seconds (2.8 hours), five sequences 10^6 seconds (11.6 days), six sequences 10^8 seconds (3.2 years), seven sequences 10^{10} seconds (317 years), and so on.

Algorithmic complexity

An algorithm is supposed to do two things:

1. Compute the correct answer given valid input data.
2. Perform the computation in a reasonable time.

The second point is of course essential. A correct algorithm is useless if it is too slow. In such cases, we must find another algorithm, or maybe we will have to accept some approximation which allows us to use another algorithm, which runs in reasonable time, but only generates an approximately correct result (whatever that means). One of the most important properties of an algorithm is how its execution time increases as the problem is made larger. By a larger problem, we mean e.g. more sequences to align, or longer sequences to align. This is the so-called algorithmic (or computational) complexity of the algorithm.

Let's say we have two alternative algorithms (A and B) for solving the same problem. Algorithm A is the fastest when we have a small number of input data points. But what happens when we have larger input data sets? If the time require by A and B increases in the same way, then A will always be the best. But what if the time to execute A is proportional to the square of the input data size, while B is linear? Then clearly there will be a point at which B becomes the better choice.

Alternative algorithms for solving the same problem may differ in how much computation must be done to set up the initial data structure, the initialization stage. There are examples of algorithms where the setup stage is expensive, but the computations that follow it are cheap, compared to some other method. In these cases, the choice of the best algorithm depends on how often the problem arises, and whether the setup can be saved and maintained between runs.

Because of this complexity, MSA are constructed by ‘progrssive sequence alignments’. This method builds an alignment up stepwise, starting with the most similar sequences and progressively adding the more dissimilar (‘divergent’) ones (Fig. 1a). The process begins with the construction of crude ‘guide tree’ (Fig. 1a). This tree then determines the order in which the sequences are progressively added to build the alignment (Fig. 1b). Note that the guide tree is included as part of the alignment output, but only to show the user how the alignment was assembled. The cardinal rule of progressive sequence alignment is ‘once a gap always a gap’; gaps can only be added or enlarged, never moved or removed. This is based on the assumption that the best information on gap placement will be found among the most similar sequences, but also for practical purposes; if gaps were adjusted at every step the alignment process would be tremendously slower. However, the once-a-gap rule can also be the source of some obvious silliness, because there often is better information in the full alignment on where the gaps really belong. This can be particularly apparent for small deletions, which might clearly be shared by several sequences but nonetheless placed at slightly different positions in each (Fig. 2a). These types of error are among the reasons for the widespread and fairly well accepted practice of ‘adjusting’ alignments ‘by eye’ to ‘minimize insertion/deletion events’ (Fig. 2b), using a program such as BioEdit (see below).

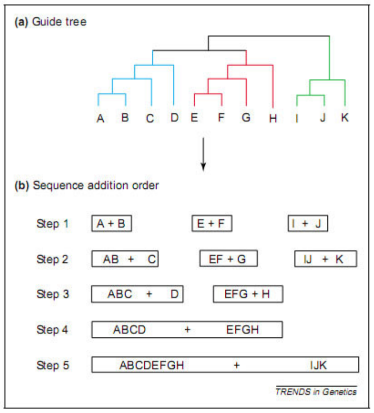
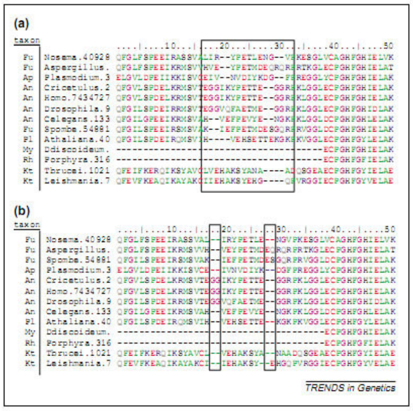


Fig. 1. Steps in progressive sequence alignment. (a) The first step is to calculate the guide tree. (b) This determines the order in which sequences are added to the growing alignment.



```
                ....|....10...|....20...|....30...|....40...|....50
Fu Nosema.40928 QFGLFSPEEIRASSVALIR--YPETLENG--VEKESGLVCAGHFGHIELVK
Fu Aspergillus. QFGLFSPEEIKRMSVVHVE--YPETMDEQRQRERTKGLECPGHFGHIELAT
Ap Plasmodium.3 ELGVLDPEIIKKISVCEIV--NVDIYKDG--FEREGGLYCPGHFGHIELAK
An Cricetulus.2 QFGVLSPDELKRMSVTEGGIKYPETTE--GGREKLGGLECPGHFGHIELAK
An Homo.7434727 QFGVLSPDELKRMSVTEGGIKYPETTE--GGREKLGGLECPGHFGHIELAK
An Drosophila.9 QFGILSPDEIRRMSVTEGGVQFAETME--GGREKLGGLECPGHFGHIDLAK
An Celegans.133 QFGILGPEEIKRMSVAH--VEFPEVYE--NGKEKLGGLDCPGHFGHLELAK
Fu Spombe.54981 QFGILSPEEIRSMSVAK--IEFPETMDESGQRERVGGLDCPGHFGHIELAK
Pl Athaliana.40 QFGILSPDEIRQMSVIH----VEHSETTEKGKEKVGGLECPGHFGYLELAK
My Ddiscoideum. --------------------------------------ECPGHFGHIELAK
Rh Porphyra.316 --------------------------------------ECPGHFGFIELAK
Kt Tbrucei.1021 QFEIFKERQIKSYAVCLVEHAKSYANA----ADQSGEAECPGHFGYIELAE
Kt Leishmania.7 QFEVFKEAQIKAYAKCIIEHAKSYEHG----QEVRGGIECPGHFGYVELAE
```

```
                ....|....10...|....20...|....30...|....40...|....50
Fu Nosema.40928 QFGLFSPEEIRASSVAL--IRYPETLE--NGVPKESGLVCAGHFGHIELVK
Fu Aspergillus. QFGLFSPEEIKRMSVVH--VEYPETMDEQRQRPRTKGLECPGHFGHIELAT
Fu Spombe.54981 QFGILSPEEIRSMSVAK--IEFPETMDESGQRPRVGGLDCPGHFGHIELAK
Ap Plasmodium.3 ELGVLDPEIIKKISVCE--IVNVDIYK--DGFPREGGLYCPGHFGHIELAK
An Cricetulus.2 QFGVLSPDELKRMSVTEGGIKYPETTE--GGRPKLGGLECPGHFGHIELAK
An Homo.7434727 QFGVLSPDELKRMSVTEGGIKYPETTE--GGRPKLGGLECPGHFGHIELAK
An Drosophila.9 QFGILSPDEIRRMSVTEGGVQFAETME--GGRPKLGGLECPGHFGHIDLAK
An Celegans.133 QFGILGPEEIKRMSVAH--VEFPEVYE--NGKPKLGGLDCPGHFGHLELAK
Pl Athaliana.40 QFGILSPDEIRQMSVIH--VEHSETTE--KGKPKVGGLECPGHFGYLELAK
My Ddiscoideum. --------------------------------------ECPGHFGHIELAK
Rh Porphyra.316 --------------------------------------ECPGHFGFIELAK
Kt Tbrucei.1021 QFEIFKERQIKSYAVCL--VEHAKSYA--NAADQSGEAECPGHFGYIELAE
Kt Leishmania.7 QFEVFKEAQIKAYAKCI--IEHAKSY--EHGQFVRGGIECPGHFGYVELAE
```

Fig. 2. Refining an alignment. (a) The raw output from a ClustalX alignment of rpb1 sequences, which predicts six insertion/deletion events (boxed), some of which are blatantly inconsistent with known taxonomy. (b) The refined alignment makes much better evolutionary sense, because it shows only two insertion events in well-defined taxonomic groups (animals and higher fungi). Taxon labels are Fu (fungi), An (animals), Pl (green plant), Ap (apicomplexan), Rh (rhodophyte), My (mycetozoan), Kt (kinetoplastids). In (b), the sequence from Saccharomyces pombe has

been placed adjacent to the other fungi to make these relationships more obvious.

Alignments are about gaps – where to put them and how big to make them. These are two different issues. Genes do not generally take insertions and deletions lightly. One out of three changes the reading frame, not to mention adding new stop codons or unwieldy junk to a protein’s structure. However, the size of a gap is much less important than the fact that it is there at all, so alignment programs have separate penalties for inserting a gap (which is costly) and for making it bigger (relatively cheap). Ideally gap penalties should differ for closely related versus distantly related sequences, for different kinds of sequence, and for different regions of the same sequence, but this is mostly impractical. Therefore, all gap penalties are compromises, and an alignment can look very different depending on the penalties that are used. In the end, the user might need to try a range of penalties, compare these by eye and pick the most logical combination.

Multiple alignment tools

Program	WWW location	Email server	Program source
Global multiple alignment			
ClustalW	"http://www2.ebi.ac.uk/clustalw/"	-	"ftp://ftp.ebi.ac.uk/pub/software/ unix/clustalw" (UNIX) vax/clustalw" (VAX) mac/clustalw" (Mac) dos/clustalw" (DOS)
Local multiple alignment			
BlockMaker	"http://bioinfo.weizmann.ac.il/blocks"	"Blockmaker@blocks.fhcrc.org"	"/ftp://ncbi.nlm.nih.gov/repository/blocks/unix/protomat" (UNIX)
MEME	"http://www.sdsc.edu/MEME"		"ftp://ftp.sdsc.edu/pub/sdsc/biology/meme/" (UNIX)
MACAW			"ftp://ncbi.nlm.nih.gov/pub/macaw" (Windows and Mac)

ClustalW: A standard multiple alignment program

From what we have learned in previous sections, doing a simultaneous N-wise alignment is not a realistic option if we have, say, 50 sequences to align. What to do? The obvious alternative is to use a so-called progressive alignment method: The alignment is built up in stages where a new sequence is added to an existing alignment, using some rules to determine in which order the sequences should be added, and how. ClustalW is one of the standard programs implementing one variant of the progressive method in wide use today for multiple sequence alignment. The W denotes a specific version that has been developed from the original Clustal program.

The basic steps of the algorithm implemented in ClustalW are:

1. Compute the pair wise alignments for all against all sequences. The similarities are stored in a matrix (sequences versus sequences).
2. Convert the sequence similarity matrix values to distance measures, reflecting evolutionary distance between each pair of sequences.
3. Construct a tree (the so-called guide tree) for the order in which pairs of sequences are to be aligned and combined with previous alignments. This is done using a

neighbour-joining clustering algorithm. In the case of ClustalW, a method by Saitou & Nei is used.

4. Progressively align the sequences/alignments together into each branch point of the guide tree, starting with the least distant pairs of sequences. At each branch point, one must do either a sequence-sequence, sequence-profile, or profile-profile alignment.

A number of rules (tricks) are used to increase the success rate of the procedure:

- Each sequence is weighted according to how different it is from the other sequences. This accounts for the case where one specific subfamily is overrepresented in the data set.
- The substitution matrix used for each alignment step depends on the similarity of the sequences (a somewhat circular argument, but what the hell...).
- Position-specific gap-open penalties are modified according to residue type using empirical observations in a set of alignments based on 3D structures. In general, hydrophobic residues have higher gap penalties than hydrophilic, since they are more likely to be in the hydrophobic core, where gaps should not occur.
- Gap-open penalties are decreased if the position is spanned by a consecutive stretch of five or more hydrophilic residues.
- Both gap-open and gap-extend penalties are increased if there are no gaps in a column, but gaps occur nearby in the alignment.
- The guide tree can in some circumstances be overridden, for instance by deferring joining two branches if they are too dissimilar, until more information has been added by processing other branches.

There are some specific cases where ClustalW is known to have problems.

- If the sequences are similar only in some smaller regions, while the larger parts are not recognizably similar, then ClustalW may have problems aligning all sequences properly. This is because ClustalW tries to find global alignments, not local. In such a case, it may be wise to cut out the similar parts with some other tool (text editor).
- If one sequence contains a large insertion compared to the rest, then there may be problems, for much the same reason as the previous point.
- If one sequence contains a repetitive element (such as a domain), while another sequence only contains one copy of the element, then ClustalW may split the single domain into two half-domains to try to align the first half with the first the domain in the first sequence, and the other half to the second domain in the first sequence. There are many proteins that contain multiple, very similar copies of a domain, so one should watch out for this.

ClustalW is an example of an algorithm that has given up on trying to be perfect (because it takes too much time), and instead uses an approximation strategy, combined with more-or-less intelligent tricks that guide the computation towards a successful (but not necessarily optimal) result. This is called a heuristic algorithm. One important point to keep in mind is that since ClustalW is a heuristic algorithm, it cannot produce a solution that is guaranteed to be optimal. But in practice, the results it produces are good enough, and one should perhaps worry more about the quality of the input data. For example, if one has sequences that are just barely significantly similar, one should worry more about if all of them really belong in the alignment at all, rather than if the alignment is perfect or not (which it in such a case almost certainly isn't). ClustalW has a number of parameters that the user can change. This will affect the exact manner in which the computation proceeds and it may be useful to compare runs with different parameters; the near-perfect parameter set varies with the specific case.

Databases of multiple alignments - domains

Very early in the days of protein sequence analysis, it was observed that some protein sequences contained long segments that were very similar to other proteins, while the rest of the sequence in that protein had no detectable similarity. Today, we take more or less for granted that proteins are composed of domains, segments of sequence which have been joined together by genetic events during evolution so that the new protein has a function that is based on the activities of the domains it contains. Often the domains detectable by sequence analysis correspond to structural domains in the 3D structure as well. There are now many well-documented cases where it has been shown that domains can exists perfectly well in isolation, when excised from the original protein. Surprisingly often, a domain can be expressed and folded all on its own.

There are today several databases that keep track of which domains have been discovered, which proteins are involved, and that store the multiple sequence alignments of the relevant segments of the protein sequences. A good example of such a database is Pfam. Also, several of the primary sequence databases now contain information about the domains in the sequence entries.

The idea behind Pfam is twofold:

1. Create and maintain good-quality multiple sequence alignments of well-defined protein sequence domains from proteins in SWISS-PROT.
2. Use these multiple alignments for creating so-called HMMs that can be used in profile searches of sequence databases.

The multiple alignment used to define a domain (protein family) in Pfam are called the seed alignment. It is created by a curator, or taken from the literature. It is used to generate a profile HMM for identifying other sequences in the databases (SWISS-PROT and TREMBL) that contain the domain. The search results are inspected to decide which cutoff should be used for that particular Pfam entry. The search hits are then aligned automatically into a so-called full alignment. There are a number of other useful databases of multiple sequence alignments, such as:

- PRINTS, multiple motifs consisting of ungapped, aligned segments of sequences, which serve as fingerprints for a protein family.
- BLOCKS, multiple motifs of ungapped, locally aligned segments created automatically.

These databases allow analysis of new sequences in terms of which domains can be detected in the sequence. This is often more useful, and sometimes also more sensitive (although this is somewhat controversial) than doing sequence-to-sequence comparisons. For instance, if a new protein has a kinase domain, then it is more helpful to use a domain database (with some appropriate search software, such as HMMER for Pfam) to identify it directly in the sequence. The alternative, using BLAST or FASTA to find similar sequences, would return thousands of sequences, and it would require some work to sort out that this is because the query sequence contains a very common kinase domain.

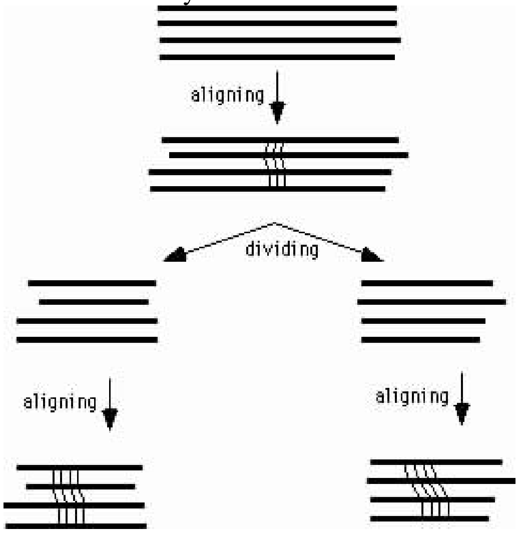


Fig. 1. Divide and conquer approach for MSA

MSA is the process of aligning three or more sequences with each other so as to bring as many similar sequences characters (nucleotides or amino acids) into register as possible. The resulting alignment can be used for two purposes.

1. To find regions of similar sequence in all of the sequences that defines a conserved consensus pattern or domain.
2. To use the aligned positions to try and derive the possible evolutionary relationships among the sequences, if the alignment is particularly strong.

When dealing with a sequence of unknown function, the presence of similar domains in several similar sequences implies a similar biochemical function or structural fold that may become the basis of further experimental investigation. A group of similar sequences may define a protein family that may share a common biochemical function or evolutionary origin. Similar proteins have been organized by several methods in to protein families. The major ones are as follows.

1. The sequence families used by PAM matrices used for sequence comparisons.
2. Patterns called MOTIFS used by PROSITE tool. This approach is based on identification of a large number of protein families called the PROSITE catalog and which defines the active sites of these proteins. These motifs may have more than one amino acid at each position and may include gaps. Once motifs have been defined the sequence databases may be searched for additional sequence entries with the same motif.
3. BLOSUM approach involved the alignment all of these families, addition of more members from the databases, and then definition of conserved patterns of amino acids called blocks. Blocks are present in all members of the family and are approximately 4 -60 amino acids long with no gaps or substitutions. The BLOSUM amino acid comparison tables were derived from these aligned blocks.

Searching

Multiple alignments are powerful tools for identifying new members of the aligned group. It is possible to query databases of multiple alignments with single sequences and to query sequences databases with multiple alignments. It has been shown that such searches are more sensitive and selective than sequence –to-sequence searches. A simple 'hybrid' approach is to use a properly made consensus sequences. Some of the automated tools are BLLMPS, MAST, LAMA etc.

Viewing MSA

Multiple alignments of many sequences and those with different sequences weights are difficult to visualize. Sequence logos are a graphical way for presentation multiple alignments. A different graphical view of MSA is by a tree relating their sequence similarity. This is very useful when the aligned sequences are of several functional subtypes and we wish to know to which one our sequence/s belongs. A way to estimate the significance of a tree is by bootstrap values. Simply put, these values show how many times each bifurcation (branching point) was observed with different models of the input data. The higher the fraction of the bootstrap value (number of observations/number of trials) the more confident we can be that the sequences emerging from that branch point cluster together.

References

Altschul, SF and Erickson, BW 1986. Optimal sequence alignment using affine gap costs. Bull. Math. Biol. 48: 603-66.

Altschul, SF, Gish, W, Miller, W, Myers, EW and Lipman DJ 1990. Basic local alignment search tool. J. Mol. Biol. 215(3): 403-10.

Gotoh, O 1982. An improved algorithm for matching biological sequences. J. Mol. Biol. 162: 705-8.

Feng, DF, Johnson, MS and Doolittle, RF 1984. Aligning amino acid sequences: comparison of commonly used methods. J. Mol. Evol. 21: 112-25.

Mott, R 1992. Maximum-likelihood estimation of the statistical distribution of Smith-Waterman local sequence similarity scores. Bull. Math. Biol. 54: 59-75.

Pearson, WR 1995. Comparison of methods for searching protein sequence databases. Prot. Sci. 4: 1145-60.

Thompson, JD, Higgins, DG and Gibson, TJ 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22 (22): 4673-80.

Isolation of plant DNA and RAPD analysis

T.E.Sheeja

Principle

For plant cells with a rigid cell wall, the disruption of cells usually requires the tissue to be ground using a mortar and pestle in a pool of liquid nitrogen. The powdered plant tissue is then transferred to an extraction buffer that contains detergent to disrupt the cell membranes. Cetyltrimethyl ammonium bromide (CTAB) is commonly used for this purpose. The extraction buffer also contains a reducing agent (2 mercaptoethanol) and a chelating agent (ethylenediamine tetra acetic acid, EDTA). This helps to inactivate nucleases that are released from the plant cell and can cause serious degradation of genomic DNA. Phenolic compounds may also be released on disruption of plant tissues and these may interfere with subsequent uses of DNA (e.g. if it is to be used in the PCR). Polyvinyl pyrroidone (PVP) can be added to the extraction buffer to remove phenolic compounds.

Phenol extraction can be used to remove any traces of proteins and the genomic DNA can be precipitated using either ethanol or isopropanol. Precipitated DNA can be hooked out of the solution or collected by centrifugation. It is important that DNA is not sheared, for this reason the DNA should not be vortexed or pipetted repeatedly using fine tipped pipette and all manipulations should be as gentle as possible.

All DNA extraction have three steps:

1. Lysis of cell walls and membranes to free DNA into solution.
2. Purification of DNA by precipitating proteins and polysaccharides.
3. Precipitation of DNA and resuspension in a buffer or sterile nuclease free water.

Protocol:

1. Weigh 3 g of clean young leaf tissue and grind to fine powder with a pestle and mortar after freezing in liquid nitrogen.
2. Transfer to 50 ml Oakridge tube with 12 ml CTAB buffer maintained at 60°C in a water bath. Mix thoroughly or vortex.
3. Incubate at 60°C for one hour. Mix intermittently.
4. Add equal volume of chloroform: isoamyl alcohol. Mix gently by inverting for 10 minutes.
5. Spin at 12,000 rpm for 15 minutes at 25 °C.
6. Transfer aqueous phase to a fresh Oakridge tube. Add 0.6 volume of 100% ethanol or isopropanol and let the DNA to settle down for 30 minutes in a freezer.
7. Spool out the DNA
8. Wash the DNA with 0.5 ml of 70% ethanol. Decant off and dry the pellet under vacuum.
9. Dissolve the DNA in minimum volume of nuclease free water or 10:1 TE.
10. Add RNAse (10µg/ml) and incubate at 37°C for one hour.
11. Add equal volume of phenol: chloroform: isoamyl alcohol (25:24:1) mix properly for 5 minutes and centrifuge at 25°C for 10 minutes at 8,000 rpm.
12. Take the aqueous layer and add equal volume of chloroform: isoamyl alcohol (24:1) mix properly for 5 minutes and centrifuge at 25°C for 10 minutes at 8,000 rpm.
13. Take out the DNA supernatant and add equal volume of ice cold 100% ethanol and keep at -20°C for one hour to precipitate the DNA. Spool out the DNA or collected by centrifugation and dry under vacuum.
14. Dissolve the DNA in minimum volume of nuclease free water or TE (10:1). Store frozen at -20°C.

Quantification of DNA

DNA quantification is an essential step in many procedures where it is necessary to know the amount of DNA that is present when performing different techniques such as PCR and RAPDs. There are several methods for quantifying DNA, the most widespread being: i) the comparison of an aliquot of the extracted sample with standard DNAs of known concentration using gel electrophoresis, ii) spectrophotometric determination, and iii) flourimetric determination. In our lab we are using the first two methods for the quantification of genomic DNA.

I) Electrophoresis of a DNA sample of unknown concentration with a known standard

- Attach tape to the ends of the gel tray. Position the well-forming comb and ensure that the gel tray is horizontal.
- Prepare 0.8% agarose gel by adding 0.8 gm agarose in 100 ml of 1x TAE and gently boil the solution in microwave oven with occasional mixing until all agarose particles are completely dissolved. Allow it to cool to 60°C and add 0.1µg/ml to 0.5µg/ml ethidium bromide. Pour agarose onto the gel tray and allow the gel to set for 20 minutes.
- Remove the comb and tape. Place the gel into the electrophoresis tank and pour 1x TAE until the gel is fully immersed.
- Load the DNA sample wit 6 x loading dye in to the wells. In one well load a standard marker.
- Carry out the electrophoresis at 5-6V/cm gel until the dye is 4-5 cm from the wells.
- Visualize the DNA bands on a UV transilluminator. Document the gel using Gel Documentation System.
- Compare the intensity of the DNA bands of the samples with the intensity of the standard DNA bands. As the amount of DNA present in each standard DNA bands is known, the amount of DNA of each sample can be calculated by comparing the fluorescent yield of the sample with that of the standard.

II) DNA quantification by UV spectroscopy:

- Take 5µl of the DNA samples in a quartz cuvette. Make up the volume to 1 ml with distilled water.
- Measure absorbance of the solution at wavelengths 260 and 280 nm.
- Calculate the ratio A280/A260.
- A good DNA preparation exhibits this ratio < 0.55 O.D units.
- Calculate DNA concentration using the relationships for soluble standard DNA, 1 O.D at 260 nm =50µg/ml. This estimate is influenced by the contaminating substances like RNA and very low molecular weight DNA in the solution.

Randomly Amplified polymorphic DNA (RAPD) analysis

Randomly amplified polymorphic DNA's (RAPD's) are well suited to high through put system, required for plant genetic analysis because of its simplicity, speed, low cost requirement of smaller quantities of genomic DNA and relative abundance of the marker in the genome. This is a PCR based technique in which single PCR primer of ten nucleotides in length will find homologous sequences in the genome, by chance and will amplify several regions of the genome, if the primer is annealed within the reasonable distance that can be amplified by Taq DNA polymerase and also in correct orientation. RAPD's are dominant marker, which cannot differentiate the homozygotes from the heterozygotes.

The primers used in the RAPD reaction possess the base sequences, which is arbitrarily defined. In this marker system the investigator have no idea to which, if any gene or repeated sequence in the plant genome, the primer may have homology. Any band after the RAPD reaction resolved in an ethidium bromide stained agarose gel or silver stained polyacrylamide gel can be used as the raw data for comparison of plant genome.

Optimization of reaction conditions should precede the actual RAPD analysis to get repeatable results. Following optimizations are essential:

1. Template DNA concentration.
2. *Taq* DNA polymerase concentration.
3. Mg $^{2+}$ ion concentration.
4. Primer concentration.
5. Primer annealing temperature.
6. Primers suitable for detection of polymorphic loci in the taxa to be analysed.

Material and reagents:

Protocol:

To a 0.2 ml thin walled PCR tube add the following in order:

Sterile distilled water	9.17µl
10 X reaction buffer	2µl
$MgCl_2$ (50 mM)	0.5µl
dNTP mix (2.5 mM)	2µl
Primer (5µM)	2µl
Template DNA (5 ng/µl)	4 µl
Taq DNA polymerase (3U/µl)	0.33 µl (1 U)
Total reaction volume	20.00µl

Mix well and spins down for 10 seconds.

Since the pipetting of small volumes is difficult and often inaccurate, a master mix is prepared where constituents common to all the reactions are combined in one tube multiplying the volume for one reaction with the total number of samples. Latter, the appropriate amount of the master mix is aliquot to each tube and DNA template (or the variable constituent) is added separately in each tube.

Carry out the amplification with the following programme:

Step 1. Denaturation at 94°C for 4 minutes

Step 2. Denaturation at 94 °C for 1 minute

Step 3. Annealing at 37°C for 1 minute

Step 4. Polymerisation at 72°C for 1 minute

Step 5. Repeat from step 2 to 4: 40 times

Step 6. Extended polymerization at 72°C for 6 minutes

Step 7. Refrigerate at 4°C.

At the end of the run take out the tubes and add 2µl of 10 x loading dye, spin for 10 seconds and store at 4°C till electrophoresis.

Agarose Gel Electrophoresis:

The amplification products in RAPD analysis are usually smaller than 4 kb size. Hence they are separated by electrophoresis in 1.4% to 1.8% Agarose gels and visualized by staining with ethidium bromide and viewing under UV light.

Protocol:

1. Attach tape to the ends of the gel tray. Position the well-forming comb and ensure that the gel tray is horizontal.
2. Prepare 1.4% agarose gel by adding 1.4 gm agarose in 100 ml of 1x TAE and gently boil the solution in microwave oven with occasional mixing until all agarose particles are completely dissolved. Allow it to cool to 60°C and add 0.1µg/ml to 0.5µg/ml ethidium bromide. Pour agarose onto the gel tray and allow the gel to set for 20 minutes.
3. Remove the comb and tape. Place the gel into the electrophoresis tank and pour 1x TAE until the gel is fully immersed.

4. Load the samples carefully. Take care to load suitable DNA size markers. Connect the leads and start electrophoresis run at constant 60 V.
5. Stop the run when bromophenol blue dye has traveled less than 2/3 the length of gel.
6. Visualize the DNA bands on a UV transilluminator. Document the gel using Gel Documentation System.
7. The amplification products are scored across the lanes comparing their respective molecular weights. The data can be analyzed stastically using appropriate software packages.

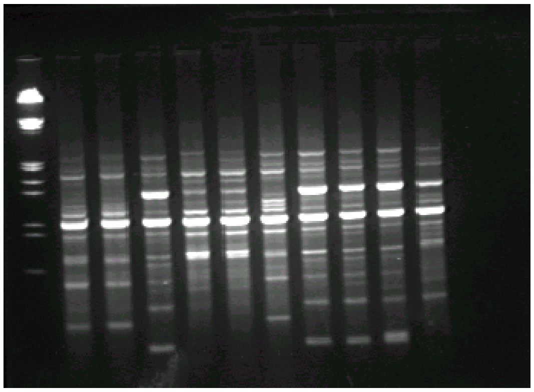


Fig. RAPD profile of ginger accessions

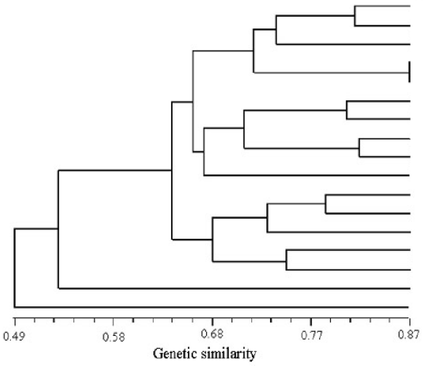


Fig. Dendrogram showing similarity among ginger accessions

References

Sambrook, J., Fritsch, E.F. and Maniaties, T (1989) *Molecular cloning: A Laboratory Manual* Cold spring Harbor Laf. Cold spring Harbor, NY.

Williams J.G.K, Rafalsky, J.A and Tingey, S.V (1993a). Genetic analysis using RAPD markers in methods in enzymology. (R.Wu.ed) 218:704-740. Academic press Sandiego.CA

Development of molecular maps in plants

K. Nirmal Babu

Introduction

In the past decade genome mapping has emerged as a powerful new approach for research in Botany, Agriculture as well as in many other fields. Many of the concepts associated with genome mapping are more than a century old. However recent advances in Molecular Biology have resulted in description of the structure and function of plant genome in unprecedented detail. Molecular level understanding of the inheritance of Agriculturally important traits creates new opportunities to streamline plant breeding. Further this understanding will enable us to identify specific DNA element responsible for particular plant characteristics.

A “Genome map” can be like a road map, reflecting the relative proximity of different landmarks to one another. It is made possible by the fact that the nuclear genome in higher organism is organized and transmitted as linear units called chromosomes. Molecular tools enable the geneticists to establish specific DNA markers at defined places along each chromosome.

DNA markers can then be used to delineate when one has reached particular gene of interest. DNA mapping encompasses a wide range of techniques, useful for studying DNA at different levels of magnification. The beginning of genome mapping lies in genetic linkage
Because the chromosomes not the genes are the unit of transmission during meiosis, linked genes is not free to undergo independent assortment, *i.e.* all alleles at all loci of one chromosome should be transmitted as a unit during gamete formation. And crossing over, during meiosis will result in the reshuffling or recombination of the alleles between the homologs. The degree of crossing over between any two loci on single chromosomes is proportional to the distance between them known as inter-locus distance. Thus the percentage of recombination gametes varies, depending on which loci are considered. This correlation serves the basis for the construction of chromosome map.

Conventional genetic mapping involves the crosses between individuals heterozygous for two or more genes. This allows as to estimates the relative likelihood that a cross over will occur between them. The principle underlying the mapping is, if two genes are very close together on the same chromosomes a cross over is unlikely to begin in the region between them. If two genes are far apart a cross over is more likely to initiate in this region and there by recombine the alleles of the two genes. The presence of recombinant offspring’s is correlated with the distance between the two genes. The map distance of two genes under considerations can be calculated by

$$\frac{\textit{Number of recombinant offspring}}{\text{Total number of progenies}} \times 100$$

The distance is expressed in mapunit or centimorgan. When the distance between the two genes is large the chance of multiple cross over in the region between the two genes is large. So multiple crossover sets a quantitative limits on the relationship between map distance and the percentage of recombinant off springs. Trihybrid crosses can be used to determine the gene order and the correct distance between the genes.

For example In a three point test cross:

Parents XYZ/xyz X xyz/xyz

Gametes XYZ xyz xyz
XyZ xYZ
Xyz xYz
XyZ Xyz

Progenies will be

Sl no	Genotype	Number of observation	Recombination between X and Y	Recombination between X and Z	Recombination between Y and Z
1	XYZ/xyz	853	-	-	-
2	xyz/xyz	926	-	-	-
3	Xyz/xyz	51	Yes	Yes	Yes
4	XYZ/xyz	42	Yes	Yes	-
5	xYz/xyz	52	Yes	-	Yes
6	XyZ/xyz	60	Yes	-	Yes
7	XyZ/xyz	7	-	Yes	Yes
8	Xyz/xyz	9	-	Yes	Yes

Recombination between X and Y

$$\frac{51+42+52+60}{2000} \text{ X } 100 = 10.25\%$$

Recombination between Y and Z

$$\frac{52+60+7+9}{2000} \text{ X } 100 = 6.4\ \%$$

Recombination between X and Z

$$\frac{51+2+7+9}{2000} \text{ X } 100 = 5.45\%$$

Then the Gene order will be:

X ----------- Z -----------------Y

5.45 cM 6.4 cM

Genetic linkage mapping using molecular markers

A molecular marker is a segment of DNA that is found in the specific sites in the genome and has properties that enables it to be uniquely recognized using molecular tools. As with the alleles, the characteristics of molecular markers varies from individual to individuals. DNA markers arise from different classes of mutations such as single base pair change, alters the restriction site for an enzyme or a primer binding site in a PCR, rearrangement in the DNA intervening between two restriction site or primer targets or change in the number of tandem repeats viz. minisatellites and microsatellites The term polymorphism refers to the idea that the individuals with in a population differ with regard to a particular fragments. The DNA segment is identical among all the members of a population is said to be monomorphic. The polymorphic markers allow the experimenter to follow a character as in conventional mapping.

In an actual analysis with molecular marker data, linkage is determined on the likelihood of linkage between two markers using statistical test called the lod (Logarithm of odds)

lod score is calculated as:

$$\text{lod} = \log_{10} \frac{\text{probability of a certain degree of linkage}}{\text{Probability of independent assortment}}$$

For e .g. If the lod score +3, $\log_{10}$ of 3 =1000 and so there is a 1000 fold greater probability that the two markers are linked then they are assorting independently. Geneticists accepect that two markers are linked if the lod score is +3 or higher. Mapping of genome by using molecular marker is done by making crosses and analyzing the offspring, but importance is given to the DNA fragments visualized in an agarose gel rather than phenotypical characters as with the conventional mapping.

The major steps involved in preparation of molecular maps are

1. Prepare a suitable mapping population
2. Develop molecular profiles (RFLP, RAPD, SSR, AFLP, EST and Isoenzyme)
3. Score the data of the progenies as parentals and recombinants for each loci

4. Analyse the data to prepare a molecular map. Usually readymade softwares like Mapmaker, Joinmap etc are used.
5. Prepare primary linkage groups – ‘preliminary map’.
6. Add data on more markers to saturate the map so that the markers are distributed uniformly in the genome.
7. Collect segregation and recombination data on important single gene controlled agronomic characters. Add to the map as before.
8. Locate the DNA markers associated with important agronomic characters

For QTLs a slightly different way of scoring to assign the different levels of expressin to different multiple allells need to be done. Use QTL soft ware for analysis and add the characters on the molecular map. These markers can be usedas probes for isolating the genes from DNA libraries and in marker assisted selection.

Genetic distances versus physical distance

Linear arrangements of the data’s obtained from the meiotic recombination’s cannot pinpoint the physical whereabouts of the gene. Cytogenetic mapping provides a way to determine the locations of the genes by microscopically examining the chromosomes. Among the several techniques are available Fluorescence In Situ Hybridization is the most commonly used method, using the cloned and labeled gene as a probe.

Mapping of the polygenes (Quantitative Trait Loci): Characters whose phenotypic variation are continuous and are determined by the segregation of multiple loci have often been referred to quantitative traits and the inheritance as polygenic. The individual loci controlling a quantitative trait are referred to as polygenes or quantitative trait loci (QTLs). Phenotypic variation, which is due to the segregation at a single genetic locus, is very easy to map. Quantitative traits show a continuum of the phenotypic variations with in a group of individuals, for such traits it is impossible to place organisms in to a discrete phenotypic class. With polymorphic molecular markers and linkage maps as tools, mapping QTL is simply a matter of growing and evaluating large populations of plants, and of applying the appropriate statistical tools. Much software are designed to assist map making. The main stages of QTL mapping are Detection: Detection of a QTL depends on a statistical test. The power of this statistical test, defined as the probability of detecting a QTL, is dependent on the number of progeny in the population, the strength of the QTL, the type of cross, and the dominance of the QTL. For additive QTLs, an F2 is somewhat more powerful than a backcross, but for dominant QTLs a backcross can be twice as powerful as an F2. The number of progeny required for detecting a QTL is, roughly speaking, proportional to the variance of the non-genetic (environmental) contributions and inversely proportional to the square of the strength of the QTL. The likelihood ratio statistics needed for significance is about 20 for an F2 and about 15 for a backcross. Weak QTLs, which account for a few percent of the trait variance, may require several hundred progeny for a 50% chance of detection.

QTL Localization: When several QTLs appear to contribute to a trait, evaluate them one at a time, starting with the most significant. Choose a marker locus near the most significant and copy it to a new chromosome called "Background". Use this chromosome as the background chromosome for composite interval mapping to localize the next QTL. If the next QTL is also significant, add a copy of a nearby locus to the Background chromosome and try to localize a third QTL. The process stops when the next QTL is not statistically significant.

QTL Fine Mapping: Fine mapping of QTLs requires the construction of special populations with large numbers of recombinations in the region identified by the genome scan. The underlying assumption of using marker loci to detect polygenes is that linkage disequilibrium exists between alleles at the marker locus and alleles of the linked polygenes. Linkage disequilibrium can be defined as the nonrandom association of alleles at differenet loci in a population and can be caused by a number of factors including selection and genetic drift.

To discover a marker/QTL linkage, one must have a segregating (i.e. experimental) population. In plants, experimental populations such as F2, backcross (BC) recombinant inbred (RI), and doubled haploid (DH) are easy to produce. Arguably, the most useful type of population is the DH. A simple ANOVA or t-test would be used to establish whether or not there is a significant difference in plant size between the two marker classes. It is possible to perform the same test using simple linear regression.

Softwares used in QTL Mapping

1. MapMaker/QTL. It will analyze F2 or backcross data using standard interval mapping.
2. QTL Cartographer It permits analysis from F2 or backcross populations. It displays map positions of QTLs using the GNUPLOT software.
3. QGene Eleven different population types (all derived from inbreeding) can be analyzed. It has a number of other features for displaying and analyzing molecular marker and phenotypic data.
4. MapQTL It can analyze a variety of predigree types including outbred pedigrees (cross pollinators).
5. Map Manager QT It will conduct single-marker regression analysis, regression-based simple interval mapping, composite interval mapping, and permutation tests.
6. MultiCrossQTL It can deal with a wide variety of simple mapping populations for inbred and outbred species.

References

George, A.W., P.M. Visscher, and C.S. Haley.(2000). Mapping quantitative trait loci in complex pedigrees: a two-step variance component approach. Genetics 156: 2081-2092.

Kao, C.H., Z.B. Zeng, and R.D. Teasdale. (999) Multiple interval mapping for quantitative trait loci. Genetics 152: 1203-1216.

Lander, E.S. and D. Botstein. (989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121: 185-199.

Patterson AH,(1996). Genome mapping in Plants. Academic Press. pp 40-90.

Seaton, G., C.S. Haley, S.A.Knott, M. Kearsey, and P.M. Visscher.(2002). QTL express: mapping quantitative trait loci in simple and complex pedigrees. Bioinformatics 18: 339-340.

Tanksley SD (1993). Mapping Polygenes, Ann. Rev. Genet. 27: 205-211

Zeng, Z. B. (993) Theoretical basis os separation of multiple linked gene effects on mapping quantitative trait loci. Proc. Natl. Acad. Sci. USA 90: 10972-10976.

Zou, F, B.S. Yandell, and J.P. Fine. (2001). Statistical issues in the analysis of quantitative traits in combined crosses. Genetics 158:1339-1346.

Mechanisms of biological control of plant pathogens

R. Suseela Bhai

The biological world is a vast interacting network of living organisms in a state of dynamic equilibrium, reflecting changes in their physical environment and their relations to each other. Individual species follow their normal cyclic changes without significant effect on the whole network, because the compensating changes in other components maintain the balance.
Organisms exhibit many type of interactions in the process of achieving biological balance. A microorganism may exert no effect on another or it may produce one or more of the following effects.

1. It may stimulate or inhibit the growth or development of the associate
2. It may stimulate or inhibit the formation of resting spores of the associate or
3. It may enforce dormancy or cause lysis of the associate

Some of these interactions produce harmful effect directly on a population of a given plant (weed competitor, insect pests, plant pathogens, organisms that produce phytotoxins). An indirect harmful effect is produced by others such as organisms beneficial to weeds, or plant pathogens, insect vectors of plant pathogens, organisms harmful to other organisms that are beneficial to plants, such as fungi that inhibit legume nodule bacteria. Many organisms are directly beneficial e.g. the nodule bacteria on legume root, insect that pollinates higher plants, Mycorrhizal roots etc.

Others may be indirectly beneficial e.g. nitrifying bacteria that convert ammonium to nitrate nitrogen, organisms that decompose organic matter, pathogens of weeds or harmful insects, organisms antagonistic to plant pathogens or those that decompose phytotoxins. An organism may thus be beneficial by interfering with the activity of a harmful one, as well as, by direct assistance and one may be harmful by interfering with beneficial organism or by direct entry.

The occurrence of a disease indicates that some aspect of the biological balance is not in equilibrium, the greater the imbalance, the more severe the disease is. A plant disease develops when one or more of the following condition prevails.

1. The pathogen is highly virulent, in high inoculum density or not in equilibrium with the antagonists,
2. The abiotic environment is especially Favourable for the pathogen, is unfavorable to the host or antagonists or both.
3. The host plant is highly susceptible or continuously and extensively grown.
4. The antagonists are absent or in low population, lack something to function as antagonists are inhibited by other microorganisms, or antibiotic produced or sorbed by the soil or inactivated by other microorganisms

Any disease control in which antagonists are involved is biological control. It is the use of natural or modified organisms, genes, or gene products, to reduce the effect of undesirable organisms. Biocontrol agents are also known as antagonists and antagonism is the generalized mechanism by which they reduce the survival or disease causing activities of plant pathogens. This is accomplished by destroying existing pathogen, inoculum, excluding the pathogen from the host plant or suppressing or displacing the pathogen after infection has occurred. Antagonism is considered to include three types of activity namely antibiosis and lysis, competition, parasitism and Predation.

1. Antibiosis and lysis: Antibiosis is the inhibition of one organism by a metabolic product of another. Antagonists may produce powerful growth inhibitory compounds that are lethal to a wide array of microorganisms. Such compounds are referred to as broad spectrum antibiotics. They may penetrate a cell and inhibit its activity by chemical toxicity. Lysis is the destruction, disintegration, dissolution or decomposition of biological materials. The lysis of biological material or organisms occurs in two ways viz. endolysis and exolysis.

Endolysis is the breakdown of the cytoplasm of a cell by own enzymes following death, which may be caused by nutrient starvation or by antibiotics or other toxins. Internal metabolic changes produced by aging or senescence, nutrient deficiency, inability to use nutrient or the accumulation of self generated toxic metabolite. have been refereed to us autolysis. It can also result form microorganisms especially bacteria, that commonly proliferate around the mycelia, particularly when leakage of content is increased. This is referred to as heterolysis. Endolysis of bacteria is termed as bacterial necrosis. Here bacteria may be inhibited by antibiotics produced by other bacteria (bacitracin), actinomycetes (streptomycin) or fungi (penicillin)

Exolysis is the partial enzymatic digestion of the walls of living cells by other organisms external to them Typically exolysis is the destruction of the walls of organisms by chitinases, cellulases etc. and is frequently results in the death of the attacked cells. The amount of antibiotic produced by an antagonist depends on its population density and also on environmental conditions. Two factors limit the utilization of antibiotics. First the antibiotics must continuously be present on the surface to protect from new infections. Secondly, the pathogen may build up resistance against the antibiotics.

2. Competition: Competition is for nutrients (high energy carbohydrates and N_2) and certain growth factors. It may also be for oxygen and space. Here two or more organisms try to utilize the same food or mineral source, or occupy the same niche or infection site. The successful competitor excludes the others due to faster growth or reproductive rate or is more efficient in obtaining nutrients from food sources. Cross protection is a form of competition in which an avirulent or weakly virulent strain of a pathogen is used to protect against infection from a more virulent strain of the same or closely related pathogen. This is more applicable in case of viral diseases that cannot be controlled using chemicals. There is no host response in cross protection and is not transmissible. But under certain conditions weakly virulent strains can become more virulent or there may be unexpected synergistic effect that enhances disease.

3. Parasitism and Predation: Parasitism is the feeding of one organism on another organism. **Mycoparasitism:** is described as fungus- fungus relationship. Enzymes degrading fungal cell walls such as chitinases and β- glucanases are commonly produced by hyperparasites. Parasitism depends on the close contact between antagonist and the host, on the secretion of enzymes and on the active growth of hyper parasites into the host. It is often exploited for biocontrol of biotrophic pathogens such as *Verticillium lecanii* parasitizing rust or downy mildews. Mycelial development of necrotrophic pathogens can also be controlled by mycoparasites. Four stages can be distinguished in mycoparasitism.

1. Chemotrophic growth in which a chemical stimulus from the pathogenic fungus attract the parasite.
2. Recognition of the pathogen. Lectins may play an important role in this specific interaction.
3. Attachment.: Fungal hyphae can grow along side the host hyphae or coil around it.
4. Degradation of host wall by production of lytic enzymes such as chitinases and glucan 1, 3 glucosidases.

Other mechanisms of biocontrol include Hypovirulence, Induced Systemic Resistance and Increased Growth Response.

1. **Hypovirulence:** Biological control by Hypovirulence occurs when a weakly virulent strain (hypovirulent) of a fungal pathogen fuses (anastamosis) with a virulent strain of the pathogen and transmits the hypovirulent condition to the virulent strain. Anastamosis is fusion of touching hyphae and represents vegetative compatibility. Transmissible hypovirulence results from the infection of the pathogen with one or more dsRNA of viral origin.

2. **Induced Systemic Resistance:** Colonization of plants with non-pathogenic Plant Growth Promoting Rhizobacteria (PGPR) can cause Induced Systemic Resistance (ISR) in the host plant. ISR is a plant mediated mechanism of biocontrol in which the biocontrol agent and the pathogen do not come in contact with one another. In ISR host plant defenses are stimulated and plants are protected systematically. The level of host response is modulated by jasmonic acid and ethylene. Because many PGPR also produces antibiotics, iron chelating siderophores or lytic enzymes their ability to suppress diseases may involve more than one mechanism.

3. **Increased Growth Response:** Micro organisms used in biological control can be associated with enhanced plant growth. In some cases enhanced growth of the host plant is due to a reduction of viable inoculum of undetected pathogens which cause only slight vigor or yield. In other cases enhanced plant growth particularly in the absence of pathogens may be due to plant growth promoting compounds of microbial origin

Mechanisms of biocontrol by *Trichoderma* and *Pseudomonas fluorescens*

Trichoderma and *Pseudomonas fluorescens* represent two common biocontrol agents used world wide for the control of many diseases including soil born and Phylloplane diseases. Mode of action of these organisms is described below.

Mode of action of fungal antagonist Trichoderma

Trichoderma species have many advantages as biocontrol agents. First, they have a faster growth and a great arsenal of inducible polysaccharide degrading enzymes. Hence the fungi can be propagated on a wide variety of carbon sources, so it is easy to find a reasonably cheap substrate for biomass production of the biocontrol agent. Another advantage is the wide range of environmental conditions tolerated by the various *Trichoderma* sp. and isolates, by which it is possible to select isolates suited for most of the environmental conditions under which the plant pathogens can cause diseases. For e.g. some isolates can grow at low temperature others at high temperature. Isolates also vary in their tolerance to various chemicals and many show remarkable resistance to fungicides, either inherently or through mutation or adaptation. It is therefore possible to select fungicide tolerant or resistant biocontrol agents for use in integrated control. This great variation among *Trichoderma* isolates provide us with a large arsenal from which to choose in the design of biocontrol programme. Many *Trichoderma* isolates possess three antagonistic mechanisms viz.: competition, antibiosis and mycoparasitism.

Competition is the most potent mechanism employed by *Trichoderma* in the control of many diseases. Early colonization in the rhizosphere and even fresh wounds or stumps helps in preventing the pathogen to enter and occupy the host niche. This competition leads to disease control. Antibiosis occurs when the production of toxic metabolites or antibiotics produced by *Trichoderma* has a direct effect on the pathogen. Even though many antagonist are able to produce antibiotics or toxins in pure culture there is little proof of a major effect of such compounds in biocontrol under field condition with *Trichoderma*.Mycoparasitism is the phenomenon of one fungus parasitizing another covers a multitude of different interactions. Four stages can be distinguished in mycoparasitism.

1. Chemotropic growth in which a chemical stimulus from the pathogenic fungus attract the parasite.

2. Recognition of the pathogen. A given *Trichoderma* antagonist attacks only a few fungi. Lectins may play an important role in this specific interaction.
3. Attachment.: *Trichoderma* hyphae can either grow along side the host hyphae or coil around it.
4. Degradation of host wall by production of lytic enzymes such as chitinases endochitinase, chitobiosidase, N-acetyl-β-glucosaminidase and glucan 1, 3 glucosidases. These antifungal compounds are inhibitory to many fungi except *Pythium* which does not have chitin as a structural component in the cell wall.

The importance of a given biocontrol mechanism is dependant on the isolate used, the target organism and the environmental conditions.

Mode of action of bacterial antagonists e.g. Pseudomonas fluorescens

Species of fluorescent pseudomonas produce siderophores called pseudobactines or pyoverdin which are very efficient competitors for iron and that competition for iron is one of the mechanisms responsible for soil suppressiveness. Reducing iron availability by addition of ethylene diamine di-o-hydroxy phenyl acetic acid (EDDHA) resulted in an increased level of soil suppressiveness and on the contrary, soil suppressiveness was decreased by addition of Fe-ethylene diamine tetra acetic acid (Fe EDTA) providing iron available for the pathogenic *Fusarium* sp. The pyoverdin produced by the fluorescent pseudomonas species was able to reduce the rate of chlamydospore germination and growth of germ tubes arising from chlamydospores

Fluorescent pseudomonas species produce several types of metabolites having antifungal activities. The lipopolysaccharides present in the outer membrane of Fluorescent pseudomonas induce systemic resistance. Lipopolysaccharides are complex molecules composed of a lipid part which is anchored in the outer lipid layer of the bilayer outer membrane and a polysaccharide part which protrudes from the cell surface. The O-antigenic side chain of the LPS is responsible for the induction of systemic resistance. Fluorescent pseudomonads also produce salicylic acid at low Fe availability which in turn is also involved in the induction of resistance.Knowledge of mechanisms involved in biological control is important for estimating and predicting its reliability and selection of better strains. Besides other criteria, the choice of an antagonist with characteristic mechanisms depends on the stage of the life cycle of the pathogen the antagonist is aimed at. Allowable interaction times and niche characteristics determine the suitability of certain modes-at action during different developmental stages of the pathogen.

References:

Bonnie H. Ownley & Mark T. Windham 2003. Biological control of plant pathogens. In: Plant Pathology Concepts and Laboratory exercises (Eds. Robert N. Trigiano, Mark T. Windham & Alan S. Windham .CRC press New York Washington D.C.

Claude Alabouvette, Bob Schippers, P. Lemanceau and Peter A.H.M Bakker 1997 Biological control of *Fusarium* wilts: Toward Development of Commercial products. In : Plant Microbe Interactions and Biological Control (Eds. GregJ Boland &L .David Kuykeendall) MARCEL DEKKER, Inc.

Kenneth F Baker & R James Cook 1973 (Eds.) Biological control of plant pathogens W H Freeman & Co, San Francisco.

Transgenic crops

T E Sheeja

A transgenic has a DNA from another organism inserted into it to impart some desired properties. Identifying and locating genes for agriculturally important traits is currently the most limiting step in the transgenic process. The isolated gene is altered so that it function effectively in a host organism and cloned in an appropriate vector, at which point it is known as a transgene. A transgenic organism sometimes called a chimera is one that contains a transgene introduced by technological methods rather than through selective breeding.

Transgenic technology enables plant breeders to bring together in one plant, useful genes from a wide range of living sources including within the crop species or from closely related plants. This technology provides the means for identifying and isolating genes controlling specific characteristics in one kind of organism, and for moving copies of those genes into another quite different organism, imparting these characteristics. It allows scientists to develop organisms that express a novel trait not normally found in the species. It is expected that transgenic plants probably hold the greatest promise in augmenting agricultural production in the first decades of the next millennium.

Possible transgenic combinations:

Plant-animal-human combinations: the DNA of mouse and human tumor fragments inserted into tobacco DNA- the harvested plants contain a potential vaccine against non-Hodgkiins lymphoma; other transgenic plants have been used to cerate edible vaccines by incorporating a human protein into bananas, potatoes and tomatoes.

Animal-animal combinations: Biosteel is a product of animal animal transgenic combination. A gene for silk protein was isolated from a spider capable of spinning silk fibres and inserted into goats egg prior to fertilization, when the transgenic female goats matured they produced milk containing the protein from which spider silk is made. It is lightweight and strong yet supple suitable for bulletproof vests.

Animal-human combinations: Pig organs can be used for organ transplantation alleviating the shortage of human hearts and kidneys, which are scarce supply. Temporary skin substitute for healing wounds or burns or as replacement cartilage, heart valves, cerebrospinal shunts or even collagen tubes to guide regrowth of nerves that have been injured. Thereapeutic proteins such a monoclonal antibodies, from the milk of transgenic cows, goats, rabbits and mice and use them to administer drugs in treatment of rheumatoid arthritis, cancer another autoimmune diseases.

Transgenic vs. traditional breeding

The basic idea behind any genetic manipulation is increasing crop yields. Two approaches have traditionally been taken to increase the quality and quantity of the crop yields by standard genetic manipulations and inhibiting deleterious effects of insects and weeds by spraying crops with pesticides and herbicides. However, these methods are cumbersome, time consuming, and have little and limited control over the process of gene recombination. Hence transgenics were identified as a method that overcomes most of these drawbacks and plant system was identified as an excellent system for genetic transformation.

The special characters of plants that makes them suitable for transgenesis are-

- Plant tissue can redifferentiate (a transformed piece of leaf may be regenerated to a whole plant)
- Plant transformation and regeneration are relatively easy for a variety of plants.
- Plants cannot move
- Plants undergo self-fertilization

- Produce a large number of progenies
- Are easy to harvest
- A natural transformation system exists for plants (the bacterium *Agrobacterium tumefaciens*).

Methods for plant transformation

Agrobacterium mediated gene transfer

Agrobacterium were first identified as the pathogen responsible for the appearance of crown gall tumors in areas of plant wounding. The best-characterized strain of *Agrobacterium* is *A. tumefaciens*, it transfers a ~25 kb segment of DNA called T-DNA into plant cells by a technique similar to conjugation. *Agrobacterium* proteins recognize and cleave Ti plasmid DNA at short repeat sequence elements called the right border and left border, which define the termini of T-DNA. *Agrobacterium* mediated plant transformation has been exploited as an efficient gene transfer method in a variety of dicot plants such as soybean, squash and tomato.

The "Gene Gun" method

DNA can also be delivered into the cells by small, µm-sized tungsten or gold particles coated with the DNA. The bullets are fired from a device that works similar to a shotgun. The modernized device uses a sudden change in pressure of He gas to propel the particles, but the principle of "shooting" the DNA into the cell remains the same. This DNA-delivery device is nicknamed "gene gun", and has been shown to work for DNA delivery into chloroplasts as well. Over the last several years, use of the "gene gun" has become a very common method to transform plants, and has been shown to be applicable to virtually all species investigated. Classical examples are transformation of monocots represented by rice, corn and wheat.

Electroporation

Another method to get foreign genes into cereals is by electroporation: a jolt of electricity is used to puncture self-repairing holes in protoplasts (i.e., the cell without the cell wall), and DNA can get in through these holes. However, it is often very difficult to regenerate fertile plants from protoplasts of cereals. Nonetheless, significant advances in overcoming these practical difficulties have been made over the years. Now even transgenic trees have been created: for example, the gene for a coat protein of the plum poxvirus has been introduced into apricot. The resulting transgenic tree shows a markedly decreased sensitivity to this virus. Thus, now there are a number of different techniques to introduce foreign genes into plants. Essentially all major crop plants can be (and have been or are being) genetically engineered, the procedures are now routine and the frequency of success is very high.

Major transgenic breakthroughs

Traits that have been introduced by transformation include herbicide resistance, improved quality characteristics, increased virus tolerance, or decreased sensitivity to insect or pathogen attack. Traditionally, most of such genetically engineered plants were tobacco, petunia, or similar species with a relatively limited agricultural application. However, during the past decade research has made it possible to transform major staples such as corn and rice and to regenerate fertile plants therefrom. The most important traits imparted into crops by transgenesis include-

- Resistance against biotic (pathogen) and abiotic (environmental) stress: Transgenic plants tolerant to increased levels of glyphosate, a widely used broad-spectrum herbicide developed by transfer of gene for EPSP synthase that overproduces the enzyme.
- Bt insect resistance: Resistance imparted to crops such as corn, potato, cotton for control of tobacco budworm and cotton bollworm etc: *Bacillus thuringiensis* is a soil bacterium whose spores contain a crystalline (Cry) protein. In the insect gut, the protein breaks down to release a toxin, delta endotoxin, which binds to and creates pores in the intestinal lining, resulting in ion imbalance, paralysis of the digestive system and insect death. In Bt

crops, a modified version of the bacterial Cry genes are incorporated into the plants own DNA, so that the plant's cellular machinery produces the toxin imparting resistance to the transgenic plant.

- Genes providing resistance against plant viruses: Successfully introduced into crop plants like tobacco, tomato and potato. Coat protein gene from TMV, classified as a positive strand RNA virus, has been transferred to tobacco and in the transgenic plant, expression of the coat protein (CP) was observed and on infection with TMV showed low and delayed infection. Transgenics have also been created with genes for nucleocapsid (N) protein from tomato spotted wilt virus (TSWV). Satellite RNA has also been used for transformation as it depends on virus for its replication and transmission even though it is unrelated to the viral genome.
- Transgenic sunflowers resistant to mildew.
- Production of pharm crops- plants that can produce antigens and antibodies and vaccines (edible plant vaccine against diarrhoea in potato and antibody against dental caries in tobacco)
- Increased production of nutrients, desired flavours etc.: Transgenic golden rice with elevated levels of vitamin A.

Agricultural impact of the transgenic crops

It is inevitable that within the next decade, almost all economically important plant crops will be derived from transgenic strains. In the last 5 years alone, transgenic varieties of over 25 commercially grown plants have been approved for agricultural purposes by the USDA. In the US 75% of the cotton crop, 50% of the soybean crop, 20% of the corn crops are transgenics. Other crops on small scale include canola, cherry tomatoes, squash, sugar beets and potatoes. Besides a number of food products have also been launched. This rapid introduction of foreign genes into food producing plants has not been without controversy. The term GMO is being used to label transgenic plants and animals. One of the first GMO to hit the market was the Flavr Savr tomato developed by Calgene, a Californian biotechnology company that has since been bought by Monsanto. The idea behind this product was to produce a hearty tomato that could vine ripen and yet still withstand the bumps and bruises of shipping.

India has allowed production of three GM cotton hybrids; developed by Mahyco in collaboration with U.S. biotechnology giant Monsanto Co, and the first commercial crops are expected to be planted soon. Bt cabbage was created recently at IARI, New Delhi. In a nation of over a billion people, GM technology offers India the tool to boost abysmally low farm yields, raising hopes for more such crops like potato, rice and mustard. Research is under way in plants like Brassica, Cabbage, Cauliflower and several transgenes are at their final stages of field trial. Transgenic technology also offers a number of ways for making agricultural products safer they are:

- By imparting resistance to postharvest diseases the levels of fungal or bacterial toxins like aflatoxin can be reduced (aflatoxins are carcinogenic in humans)
- It may be possible to modify known protein allergens to eliminate allergenicity (eg. Peanut)
- Disease resistant varieties may allow decreased use of fungicides and insecticides
- Modification of seed oils can help in decreasing levels of saturated fats
- Modification of plants can help in restricting heavy metal deposition into seed or fruit
- Bioremediation technique can be done with genetically engineered plants to remove heavy metals from soils.

Transgenic hazards

Although there is considerable optimism about the future prospects for transgenic crops among the scientific and plant breeding community, public concerns are being expressed in some countries over the following possible hazards-

- The movement of the transgene itself with subsequent expression in a different organism or species.
- The transgenic plant itself.
- Non-target hazards associated with the transgene product outside the plant.
- Resistance evolution in the targeted pest population.
- Seed dispersal due to spillage during travel or from the plant itself.
- Horizontal transfer: the nonsexual transfer of genetic material from one organism to the genome of another.
- Pollen dispersal to other varieties of that crop, related crops and wild relatives.

Transgenic controversies?

In spite of being one of the first major product groups in agriculture biotechnology, transgenics have always been surrounded with controversies. Lot of apprehensions exists on the utility, profitability, public acceptance and health and environment risks related to transgenics. The assessment and potential impact of transgenic crops has been discussed in scientific publications and international fora, but now the subject is a major focus of pressure groups, with vandalism of transgenic field plots in some countries and substantial amounts of money being spent against it. However, these allegations lack sufficient scientific proof. Some of the important controversies are discussed-

- Golden rice: Contains a gene from daffodils. It produces high levels of beta-carotene a vitamin A precursor. Claims to be adequate to supplement diets of people having vitamin A deficiency. The controversy is that the amounts of rice normally eaten provide only a fraction of the vitamin A needed and for maximum absorption of vitamin A other food supplements also need be taken. Hence the ultimate aim of producing the transgenic is not fulfilled unless problem of food distribution is solved. Hence the mere production of transgenic does not solve the problem of malnutrition or vit A deficiency.
- Herbicide resistant transgenic crops: Herbicide resistance genes can be inserted into a crop and thus we can use herbicides all over the land where the crop is grown. But some weeds vary in heribicide tolerance and they may dominate the area by natural selection. The microbial bar and pat genes confer tolerance to the non-selective herbicide phosphinothricin (PPT; sold as Basta or Finale). This tolerance in plants could provide an environmental gain compared to current-day herbicide cocktails, but the safety of such a transgene approach is questioned by many. It is argued that the introduction of transgenic PPT-tolerant crops could result in acquired PPT tolerance in weedy relatives of these crops.
- Bt transgenics: In case of Bt gene, *B. thuringensis* can produce a toxin deadly to insect larvae especially the caterpillars, while it is nontoxic to other wild life. It can be applied to the plant surface. However, in the due course bacterial gene can get incorporated into plants. This may increase the selection for resistant insects. Pollen may contain Bt toxin, and if they contaminate other plants, it may kill harmless or beneficial insects.

  The Bt protein is toxic to the European corn borer. As well as other insects, including the Monarch butterfly. However, studies shows benefits outweigh the risk. Another apprehension about Bt corn is that there is a possibility that Bt may be an allergen in humans and therefore Bt corn is not approved for human food. However, Bt cotton developed for use in areas where cotton bollworm causes crop damage is a big business.
- Escape of transgenic plants: Some plants can naturally breed with closely related species, which helps in gene flow. This happens in GM plants also. GM canola and GM barley genes were reported to have transferred some genes to wild relatives and non-GM crops in Canada and Finland. Apprehensions are that they may be harmful or can affect the ability of weeds to survive and reproduce. Commercial crops and wild relatives face a threat of getting contaminating by transgenics

- Allergens: Though we eat food-containing lot of DNA and hence it can be presumed that DNA is harmless there are apprehensions about the use of GMO's for food. Though the proteins produced by GMO's are similar to those in other foods they are likely to produce allergens like in the case of transgenic soyabeans having Brazil nut protein gene that caused serious reactions to people allergic to nut.

How to ensure safety in transgenic food crops:

Tremendous efforts were put in to solve the problems and controversies associated with transgenics. One of the most important step towards promoting universal acceptance of this technology is to make it environmentally safe. Ways and means suggested by environmentalists and scientists however, leaves many a question unanswered. However, some of these may help to minimize the hazards (?) associated with transgenics.

- Use gene knockout wherever possible
- Modify regulation of a gene whose protein is already present
- Use genes for proteins that already appear in other food crops
- Limit expression of a novel protein to specific non-food tissues when a novel protein must be present in a food, it must be demonstrated to be safe.
- Modifying crops by inserting genes into chloroplasts that are inherited maternally in many species, they have also prevented transmission of the gene by pollen--closing a potential escape route for transgenes into the environment. (eg. engineering tolerance to the herbicide glyphosate into the tobacco chloroplast genome)
- Deletion of non-essential genes (antibiotic markers) by Cre-lox system: transform plant with T-DNA containing target gene plus a selectable marker. The selectable marker is flanked by lox sites. Cross with another transgenic line containing the Cre recombinase gene. In progeny, Cre will act in trans to excise the selectable marker. Since the Cre gene is not linked to the target gene it can be crossed out by normal Mendelian segregation target +Cre progeny can be detected by PCR.

Transgenic crops may interbreed with nearby weeds, increasing their competitiveness, and may themselves become a `volunteer' weed in the following crop. The desired transgene can be coupled in tandem with genes that would render hybrid offspring or volunteer weeds less able to compete with crops, weeds and wild species. Genes that prevent seed shatter or secondary dormancy, or that dwarf the recipient could all be useful for mitigation and may have value to the crop. Many such genes have been isolated in the past few years.

Molecular approaches for studying the genetic diversity of fungal pathogens and biocontrol agents

M Anandaraj and Sheji Chandran

Genetic diversity – a molecular approach

Genetic diversity of microorganisms particularly of beneficial microbes is important for selecting the most suitable one for effective biocontrol of plant diseases. Molecular markers are useful tools for developing detailed linkage maps of species or even at lower taxonomic levels among isolates of a single species from different populations (which are likely to be similar and to have overlapping traits) suitable distinguishing feature which may be difficult or impossible to find out using conventional methods. Molecular markers, however, can be applied at these levels with great reliability and they allow simultaneous measurement of variability at multiple loci of each individual tested.

Genetic Markers: Isozymes

Isozymes are alternative enzyme forms encoded by different alleles at the same locus, which can be used as informative genetic markers. Multiple polymorphic loci can be surveyed and typically 2 or 3 are detected at each locus. Randomly sampled Isozymes loci are generally accepted to be of independent genetic origin. However, Isozyme variations in many pathogenic fungi are low or Non-existent.

DNA based markers:

RFLP: These are the first DNA based markers developed. They may be the result of length mutation and /or point mutation at a restriction digests of genomic DNA, cDNA or mitochondrial DNA fragments from specific DNA segments amplified using PCR. Thus, depending on the probe used, RFLPs can be used to analyze mtDNA variation, ribosomal (r)DNA region variation, repetitive and single-copy sequence variations. RFLPs are co-dominant markers. This makes them suitable for population genetic studies as well as for linkage map construction. By employing probes that detect multiple loci and dispersed repetitive sequences, the sensitivity of the RFLP can be enhanced to fingerprinting resolution.

AFLP

RFLPs can be converted to AFLPs by ligating adaptors for PCR amplification. The method offers the potential to detect large numbers of amplification products. Although this method does not target specific areas of the genome for marker identification, the large number of loci that can be analyzed in a single experiment greatly improves the chance of identifying markers linked to the chosen locus

PCR with Arbitrary primers

Direct Amplification Fingerprinting (DAF) (Caetano- Anolles *et al,* 1991) involves the use of short primers, five to eight nucleotides long, with low or high stringency annealing steps and a two-temperature instead of standard three temperature cycling program. Resulting fragments are separated on polyacrylamide gels and visualized by silver staining. AP-PCR (Welsh and McClelland, 1990) is performed with oligonucleotides with 20 or more nucleotides, were used as primers. Two cycles with low stringency (allowing for mismatches) were followed by 3- to 40 cycles with high stringency. This method is employed to assay variation "within species" (Williams *et al*., 1990), in which arbitrary short oligonucleotide primers, target unknown sequences in the genome, are used to generate amplification products that often show size polymorphism within species.RAPD analysis offers the possibility of creating polymorphisms without any prior knowledge of the DNA sequences of the organism investigated. The patterns produced are highly polymorphic, allowing discrimination between isolates of a species if sufficient numbers of primers are screened. RAPDs data can contain artifacts are not fully reproducible. However, RAPDs have been used to generate large numbers of genetic markers useful for linkage mapping quickly and cheaply.

Uses:

RFLP analysis is a laborious and time-consuming procedure that requires previous information of the genome. All these limitations can be overcome by RAPD analysis.

RAPD is useful in generating large numbers of DNA markers for species whose genomes are relatively unknown.Specific primer designing: RAPD fragments can be used to clone and sequence data, to design PCR primers specific for particular organism which then can function as sequence tagged sites (STSs). This method eliminates the problems associated with reproducibility associated with RAPD analysis.

I. Purification and Culture of *Trichoderma*

Trichderma Source

The *Trichoderma* isolates can be obtained from wide variety of agro-ecosystems where black pepper, ginger, cardamom and other plantation crops are grown, the soil, plant parts, rhizosphere soil and. These isolates have been collected from a Many of them have been tested as effective biocontrol agents against a number of plant pathogens especially *Phytophthora* and *Pythium*. Some of them have remarkable resistance against metals and a few have very good commercial prospects.

2. Purification of *Trichoderma*

With the help of a sterile loop, fungal spores were taken from agar slants and resuspended in 1ml sterile distilled water in a microfuge tube. The suspension was serially diluted to 10^{-4} dilution. Isolated colonies were obtained by pour plating and/or by streaking on *Trichoderma* specific medium (TSM). The plates were incubated at 28^0C.

3. Production of Fungal Mycelium

For submerged culture, well-isolated colonies are used for inoculation in 50 ml of TSM basal broth and the fungus is cultivated with constant shaking at 180 rpm and 28^0C.

4. Preparation of mycelium for DNA isolation

Four days old submerged culture mycelium is used for DNA isolation. The mycelium is filtered out on a sterile Whatman No. 1 filter paper and wash thrice with sterile distilled water to remove traces of the medium. The fungal mass is air dried to remove excess of moisture. Mycelium is used either fresh or stored at -80^0C for subsequent use.

II. Standardization of DNA isolation

Fungal cells contain a thick rigid chitinous cell wall. When isolating DNA from fungus, the yield and quality of DNA depends on the extent of rupturing cell wall and the degree of carbohydrate and protein contamination. The best protocol is the one that effectively ruptures the cell wall and minimizes contamination.

A) CTAB Method

1 gm of four days old fungal mycelium is ground in liquid N_2 and immediately transferred to 5ml CTAB extraction buffer (2%w/v CTAB, 100mM Tris-Cl, 20mM EDTA pH 8.0, 1.4M NaCl) preheated at 65^0C. The contents are mixed well and incubated at 65^0C for 30 minutes. DNA is extracted with equal volume of chloroform: isoamyl alcohol (24:1) for 15 minutes on a rocking platform. The two phases are separated by centrifugation at 12,000 rpm for 5 minutes. The aqueous phase is separated and the DNA precipitated with 2/3rd volume of isopropanol and centrifuging at 12,000rpm for 5 minutes pellets DNA. The pellet is air dried

and dissolved in 500µl of sterile distilled water.

B) SDS Method

1 gm of mycelium is ground using liquid N_2 and transferred to a sterile centrifuge tube. To this 10ml of STE extraction buffer (320mM Sucrose, 10mM Tris.Cl, 20mM EDTA, 75mM NaCl and 2.5ml of 20% SDS is added and incubated for 10 minutes in water bath followed by extraction with equal volume of phenol: chloroform (1:1). The aqueous phase is taken and DNA precipitated with 2/3rd volume of isopropanol. DNA is pelleted by centrifugation at 14,000 rpm for 15 minutes and washed with 70% ethanol, air-dried and dissolved in 500µl of sterile distilled water.

C) SDS Method using Powdered Glass

In this method, fine powdered glass made from glass wool is employed. 1gm of mycelium is ground with powdered glass with enough STE buffer to saturate the powdered glass. The rest of the method is same as in the above protocol.

D) SDS Method using Minigrinder

This method is the most successful one giving good quality DNA in high yield. To effect disruption of rigid chitinous cell wall a minigrinder (Sigma) is employed. This minigrinder is designed such that it fits in the microfuge tube and effectively disrupts the cell wall which is reflected in the high DNA yield.

DNA isolation from fungi by SDS method *(eg. Trichoderma harzianum)*

1. Transfer 200mg mycelium in a microfuge tube.
2. Add 100µl STE extraction buffer (see Annex. II for composition) and grind well for 3 minutes using minigrinder.
3. Add 400µl STE buffer and grind again for 2 minutes.
4. Finally add 500µl of STE buffer and 250µl of 20% SDS.
5. Mix well and incubate at 65^0C for 10 minutes
6. After incubation divide the contents of the tube into two microfuge tubes. Add equal volume of phenol: chloroform (1:1) and mixed gently.
7. Centrifuge at 5000 rpm for 5 minutes to separate the phases.
8. Re-extract aqueous phase with phenol: chloroform (1:1).
9. Collect the aqueous phase and precipitate DNA with 1/10th volume of 3M Sodium acetate (pH 5.2) and 2 volumes of ethanol.
10. Wash the pellet with 500µl of 70% ethanol.
11. Dry pellet *in vacuuo* for 20 minutes and dissolve in 100µl sterile distilled water.
12. Stored at 4^0C.

III. Quality Analysis of DNA in Agarose Gels

1. The quality of isolated fungal genomic DNA was checked in 0.7% agarose gel.
2. Melt 0.28gm of Agarose in 40 ml of 1x TAE in a 150 ml flask by heating the flask in a microwave oven.
3. Molten agarose was cooled to 50^0C and 2µl of EtBr (10mg/ml) is added. A gel is cast in a gel-casting tray with properly placed combs.
4. 10µl of DNA and 2µl of gel loading buffer is properly mixed and load the mix to each well without spillover.
5. The gel is subjected to 3V/cme electricity for 1 hour in 1X TAE buffer.
6. It is then visualized in Multi-image™ light cabinet (Alpha Imager 2200) and the gel images can be captured by Charge-Coupled Device (CCD) camera. The images are stored as JPEG extension file for analysis.

IV. Quantification of DNA

All the DNA samples were diluted ten fold times and then quantified in UV-VIS spectrophotometer (Cary 50). The A260/A280 ratio is calculated to check the purity of the DNA preparation. The thermal cycler Eppendorf Master Cycler Gradient has an option for setting temperature gradient so that each column of PCR tubes receives temperature in increasing order. It has a special feature that alleviates the mineral oil layering on the reaction mixture.

V. Primer Screening

Primer screening is performed to ascertain the suitability of primers and to select primers for the RAPD analysis. For screening primers PCR reactions are set up using DNA from model organism as template. Label 0.2ml PCR tubes.

1. Prepare reaction stock mixture. (Eg. For 10 reactions prepare stock for 11 Reactions) Add all the components except primers.
2. Dispense 24µl reaction mixture in each PCR tube.
3. Add 1µl of selected primers to each different tube.
4. Keep the tubes in the thermal cycler and run the machine using desired program.

VI. PCR with template DNA as variable

For the RAPD analysis, the primers giving positive amplification during screening are selected. These selected primers are used to amplify DNA from all the 22 isolates. In this reaction, the template DNA is a variable while all other reaction constituents are kept constant.

Procedure:

1. Label twenty-two 0.2-ml PCR tubes.
2. Prepare reaction stock mixture according to table 1. Add each component in the given order except the template DNA.
3. Dispense 24µl of stock in each PCR tube.
4. Add 1µl of respective template to each tube.
5. Place the tubes into the thermal cycler and amplify DNA using the required conditions.

VII. Post PCR analysis of data

Resolve the PCR products in 1.4% agarose gel using protocols mentioned earlier and record the results for analysis using suitable programme for data interpretation

Molecular approaches for studying the genetic diversity of *P. capsici* black pepper

Culturing of *P. capsici*: *Phytophthora capsici*, the foot rot pathogen of black pepper is used for DNA isolation. To ensure the genetic purity the single zoospore cultures are plated on the *Phytophthora* specific medium (PVPH incorporated Carrot Agar) and after 48 hours the mycelial disc is transferred into the carrot agar medium and after the period of incubation of 24-48 hrs the mycelial disc is aseptically transferred into the liquid medium (GYP).

DNA isolation from *Phytophthora*

1. 0.1gm of four days old mycelium + 750µl odf SDS extraction buffer + 50mg of polyvinyl polypyrolidone + 10mg of glass powder
2. Grind well with minigrinder
3. Centrifuge at 13,000 rpm for 5 minutes
4. Add mixture of phenol:chloroform:isoamyl alcohol (25:24:Supernatant added with the 1) to supernatant and mix well
5. Centrifuge at 13,000 rpm for 5 minutes
6. Add isopropanol (2/3 vol)
7. Centrifuge at 13,000 rpm for 10 minutes.

8. Look for DNA pellet and wash the pellet in 70% ethanol by centrifuging at 13,000 rpm for 2 minutes.
9. Dry the pellet and add 50µl of sterile distilled water with Rnase. (20ug/ml)

RFLP PCR

ITS6 and ITS4 were used to identify different species of *Phytophthora*. The primer sequence for ITS 6 is 5'GAAGGTGAAGTCGTAACAAGG 3' and ITS4 is 5' TCCTCCGCTTATTGATATGC 3'.

The PCR mixture contained

DNA	1.0µl (50ng)
Taq pol (1U)	0.5µl
Reaction buffer (1X)	2.5µl
$MgCl_2$ (2mM)	3.0µl
dNTP (100µM)	2.0 µl
Primes ITS4 and ITS6	1.25µl +1.25µl (each primer)
Water to makeup	25µl

The PCR machine is programmed as follows.

1. Initial denaturation at 94°C for 3 minutes, 40 cycles of denaturation at 94°C for 1 minute, followed by annealing at 55°C for 1 minute and polymerization at 72°C for 2 minutes and final extention at 72° C for 10 minutes
2. Restriction digestion of the amplified product
3. Once the reaction is over, out of 25µl, 5 µl is used to check out the presence of the amplified product of 862bp by agarose gel electrophoresis of 1.4%. Once the presence of the product is confirmed, out of the remaining 20µl was separated into 10 µl each, and the first 10µl is subjected to MSP restriction digestion by adding
4. 0.5µl of restriction enzyme (Msp I or Taq I)
5. 0.2µl of restriction enhancer (BSA),
6. 2µl of 10X restriction buffer and
7. 7.3µl of sterile distilled water, making up the total volume of 20µl.
8. This was subjected to incubation at 37°C for 2 hrs.
9. The same is done with the left out 10µl of the PCR product but with the change in the restriction enzyme, i.e instead of Msp I, Taq I (restriction enzyme) is used and incubated at 65°C for 2 hrs.
10. And the restricted product of Msp I and Taq I are electrophoresed in 2.4% agarose gel, and is documented.

References

Caetano-Anolles, G., Bassam, B.J. and Gresshoff, P.M. (1991). DNA amplification fingerprinting: a strategy for genome analysis. *Plant Mol. Biol. Rep*. 9: 294-307.

Welsh, J. and McClelland, M. (1990). Fingerprinting genomes using PCR with arbitrary primers. *Nucl. Acids Res.* 18: 7213-7218.

Williams, J.G.K., Kubelik, A.R., Livak, K.J., Rafalski, J.A., and Tingey, S.V. (1990). DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nuc. Acids Res.* 18: 6231-6235.

Molecular approaches for detection and identification of viral pathogens

A I Bhat and S. Abraham

A virus is a nucleoprotein that is too small to be seen with a light microscope multiplies only in living cells and has the ability to cause disease. Viruses do not divide or do not produce any kind of specialized reproductive structures such as spores, but they multiply by inducing host cells to form more virus particles. About 1000 different viruses are known to infect various plant species. Viruses are either elongate, spherical or bacilliform in shape and consist of nucleic acid (either RNA or DNA) and a protein, and a few of them contain additional chemical compounds such as polyamines, lipids or enzymes. Viruses are transmitted from plant to plant in a number of ways such as vegetative propagation, mechanically through sap, and by seed, pollen, insects, mites, nematodes, dodder and fungi.

Plant viruses differ greatly from all other plant pathogens in size, shape, composition, physical structure, methods of infection, multiplication, translocation within the host, dissemination and symptoms they produce on the host.

Need for Diagnostics for virus detection

Correct diagnosis is a pre-requisite for effective management of any disease. The causal organism involved in fungal and bacterial diseases can be easily identified with a light microscope. However, this in not the case with viruses as they cannot be seen under light microscope. Besides, in most cases symptoms caused by viruses resemble those caused by mutations, nutrient deficiencies or toxicities. As viruses are obligate parasites and always systemic in nature, once infected virus particles will be present throughout the plant system. Unlike fungi and bacteria, a plant infected with a virus cannot be cured because of lack any effective viricide. Hence planting of virus-free healthy material is very important especially in cases when virus gets transmitted from one generation to another through the use of infected vegetative propagating material or seed. In order to check the virus-free nature of the planting material, use of quick, sensitive and reliable diagnostic methods are essential.

Diagnostic methods

The methods for detecting plant viruses involve primarily the transmission of the virus from diseased to healthy plant by grafting, or by rubbing with plant sap or transmission through dodder or insect vectors. However, the most definitive proof of the presence of virus in a plant is provided by its purification, electron microscopy, serology, nucleic acid hybridization and polymerase chain reaction. Choice of planting material to be sampled is very important for successful detection and diagnosis. Sensitivity (how small an amount of virus can be detected), accuracy, reproducibility, number of samples that can be processed in a given time, adaptability to field conditions, cost and degree of operator training required are some of the factors one need to consider while selecting a method for detection. In the following, a brief of some of the important methods are given.

1. Electron Microscope

The presence of virus particle in the extract of diseased plant can be quickly carried out through electron microscope through negative staining. It reveals size and shape of the virus particle too. Considering the cost of equipment, operational expenses and need for a highly skilled worker, the method has little utility as far as routine diagnosis is concerned. However, it will be of use whenever a new virus has to be identified.

2. Enzyme linked immunosorbent assay (ELISA)

Enzyme linked immunosorbent assay is the most important and widely used methods for detection. It is a serological procedure based on the interaction between viral protein (antigen) with antibodies raised against them in a vertebrate. In this method, sensitivity of detection of the antibody-antigen reaction is increased by attaching a small quantity of enzyme to the antibody molecule. An enzyme substrate is then added resulting in colour reaction which can be quantitatively measured.

The method is very sensitive detecting upto 1-10 ng per ml of virus. The method is also economical in the use of reactant and readily adapted to quantitative measurement. It can be used for the routine detection of viruses, upto 90 samples can be tested in a single plate. It is important to use a known positive and negative controls in the test. Ready to use commercial ELISA kits for various viruses are now available from different manufacturers. ELISA test is performed using microtitre plate referred to as ELISA plate. Each plate will have 96 wells for loading and analyzing the test samples. In this procedure, wells are initially loaded with appropriately diluted antibody solution and allowed to stand for 2h after which contents of the wells are emptied and washed thrice. The extract from the test sample is then added as per the pre-decided plan and incubated for 2h.The wells are then washed thrice and added with enzyme labeled antibody. The plate is incubated again for 2h after which washed thrice and added with substrate. Positive samples can be identified by the formation of coloured product, which can be observed visually or measured using an ELISA reader. An example of ELISA test is provided in Figure 1. Besides, many variations of ELISA procedures are currently available.

3. Dot Immunobinding assay (DIBA)

It is also based on the same principle as that of ELISA. Instead of ELISA plate, in this procedure nitrocellulose or nylon membranes are used as solid support media for spotting the samples. A small quantity of extract from the test sample is directly spotted on the membrane and allowed to air dry. The membrane is later put in a solution containing heterologous (unrelated) protein (ovalbumin or SDM) for 1 h after which required quantity of antibodies coupled with an enzyme is added and incubated for 1h. The membrane is then washed to remove unbound antibodies and transferred to a solution containing substrate. The positive samples are identified by the formation of insoluble coloured product at the spotted area. The intensity of the coloured spot can be assessed by eye or densitometer, which is proportional to the amount of virus present in the sample. DIBA is more sensitive than ELISA and highly suited for field-testing of number of samples. An example of DIBA test is shown in Figure 2.

4. Dot blot hybridization

All viruses have nucleic acid (either RNA/DNA) as their genome. Dot blot hybridization is based on the specific interaction between two nucleic acid molecules (either RNA: RNA or RNA: DNA or DNA: DNA) by complementary base pairing. The method is based on the fact that the two complementary strands in a nucleic acid duplex or double helix can be denatured and separated by heat and then renatured (annealed, hybridized). Dot blot hybridization is now the commonly used procedure for testing of large number of samples. It is more sensitive than ELISA and DIBA detecting upto 1-10 pg of virus. The main steps in dot blot hybridization include: (a) A small amount of sap is extracted from the test plant and is heated to denature the nucleic acid (b) A small quantity of this is applied to the membrane, allowed to air dry and membrane baked to bind the nucleic acid firmly to it (c) The membrane is put in a solution containing heterologous nucleic acid of un-related DNA (d) A small quantity of labeled probe nucleic acid (to test nucleic acid) is added and allowed for hybridization (e) washing off unhybridized probe and estimation of the amount of probe bound by a method appropriate to the kind of label used for the probe.

5. Polymerase Chain Reaction (PCR)

Polymerase chain reaction (PCR) is the most sensitive method (10^2-10^5 times more than ELISA) available presently for the detection. It is important when viruses occur at low concentration. It also has the potential to detect more than one virus in one reaction and diagnosis is amenable to automation. PCR technique involves the specific amplification of the target DNA fragment enzymatically under *invitro* conditions. The method involves the hybridization of synthetic complementary oligonucleotide primers to the target sequence and synthesis of multiple copies of cDNA of the sequence between primers using heat stable DNA polymerase. Each amplification cycle consists of melting the double stranded template DNA at high temperature, annealing (hybridization) of the primers with complementary sequences in the template DNA at low temperature, extension of the primers with DNA polymerase (DNA synthesis). During each cycle, the sequence between the primers is doubled. A normal PCR involves 30-50 such amplification cycles.

As PCR is based on DNA, it is not directly applicable to most plant viruses that have RNA genomes. However, a complementary DNA (cDNA) can be made to the desired region of the RNA using a primer and reverse transcriptase, and this used as the initial template. This procedure now widely used is termed as RT-PCR (Reverse Transcription PCR). PCR and RT-PCR has proved to be a very powerful tool for virus detection and diagnosis. It can be used directly to produce a DNA product of predicted size that can be confirmed by gel electrophoresis.

Besides above listed methods there are several other techniques and their variations available for the detection of viruses in plants. One can choose the best method, depending on the sensitivity required, number of samples to be tested, type of sample, reliability of the technique, equipment, cost and expertise available. General difficulties such as variation in the concentration of the virus in the sample, mixed infections with more than one virus or other diseases may be encountered during tests.

The following procedure describes total DNA isolation from black pepper and PCR for the detection of Piper yellow mottle virus infecting black pepper.

The important steps in the PCR for plant viruses containing either single- or double stranded DNA as their genomes include: (i) extraction of total DNA from virus infected plant (ii) synthesis of two virus specific primers; (iii) set up the PCR reaction in a vial by adding extracted nucleic acid, primers, nucleotides, magnesium chloride and *Taq* polymerase. Use nucleic acid extracted from a known infected and healthy plants to serve as positive and negative controls respectively; (iii) keep the vials in the PCR machine and start the run as per the pre-decided program and (iv) identify positive reactions by running contents of PCR on the agarose gel. The presence of bands at the expected position indicate that sample under test is positive.

DNA isolation from black pepper
Requirements
Materials

Chilled mortar and pestle
Sterile Eppendorff tubes and stands
Sterile 1ml, 200µl, 20 µl tips
Micropipette
Water bath
Centrifuge
Ice flaking machine
Freezer (-20^{o}C)

Reagents

1. Extraction buffer (100mM Tris Hcl (pH8.0), 4mM EDTA,1.4 mM NaCl, 2% CTAB, 1% PVP,0.5% β-Mercaptoethanol)
2. 10%CTAB.
3. Phenol:Chlorofom:Isoamylalcohol (25:24:1)
4. Chlorofom:Isoamylalcohol (24:1)
5. 3M Sodium acetate (pH 5.2)
6. Ice -cold Isopropanol.

7. 70% ethanol
8. HPLC grade water

Procedure

1. Grind 100 mg of leaf tissue in 500 µl extraction buffer using chilled mortar and pestle and collect the filtrate in an Eppendorff tube.
2. Incubate in a water bath at 65°C for 30 min.
3. The homogenate is allowed to cool to room temperature and add equal volume of Phenol:Chlorofom:Isoamylalcohol (25:24:1) and mix well.
4. Centrifuge at 2500g for 10 min at room temperature.
5. Collect the supernatant in a new tube and add 0.1 V of 10% CTAB, equal volume Chloroform:Isoamylalcohol (24:1)and mix well.
6. Centrifuge at 2500g for 10 min at room temperature.
7. Collect the supernatant in a new tube and add 0.1V of 3M sodium acetate (pH 5.2) and add equal volume of ice-cold isopropanol.
8. Mix well and incubate in ice for 30min.
9. Centrifuge the mixture at 10,000 rpm for 15 min at 4°C.
10. Discard the supernatant. Add about 500 µl of 70% ethanol to the pellet and centrifuge for 5 min at 12,000 rpm.
11. Discard the supernatant and air dry the pellet.
12. Dissolve the pellet in 100 µl of HPLC grade water and store the DNA at -20°C.

PCR for amplification of PYMoV (*Piper yellow mottle virus*) infecting black pepper

Materials

Thermal cycler
Sterile PCR tubes (200 µl)
Micropipette
Micro tips
PCR work station
Micro centrifuge
Agarose gel Electrophoresis unit with power pack

Reagents

1. Autoclaved HPLC grade water
2. *Taq* buffer
3. $MgCl_2$
4. Primers (Forward and reverse)
5. dNTPs
6. *Taq* polymerase
7. Template

Procedure

1. In a pre cooled PCR tube, add the following reagents in the order given below.

Components	**Stock**	**Req.Conc**	**volume needed**
PCR water	1x	1x	17.8
Taq buffer	10x	1x	2.5
$MgCl_2$	25mM	2.5mM	2.5
dNTPs	2.5mM	100 µM	0.5
Primer(Forward)	10pm/ µl	5pm	0.5
Primer(Reverse)	10pm/ µl	5pm	0.5
Taq polymerase	3U/ µl	0.5U	0.2
Template	-	-	0.5
Total	-	-	25.0

2. Mix the contents and give a short spin in the micro centrifuge. Place the tube in a thermal cycler and set the temperature profile as given below.

94°C	-	5 min	
94°C	-	30 sec	35 Cycle
50°C	-	1 min	
72°C	-	1 min	
72°C	-	10 min	

3. Take out the tube and add 2.5 µl of 6x gel loading dye.
4. Prepare 0.8% agarose gel in 1X TAE containing ethidium bromide.
5. Load the samples into the wells. Load 2µl of 1kb ladder in a separate well
6. Electrophorase at 60V for 1.30 h in horizontal submarine electrophoresis unit containing 1X TAE
7. Visualize the gel under UV- transilluminator in gel documentation system.

Expected result:

The lane loaded with 1 kb ladder would show up 5 bands (250bp, 500bp, 750bp, 1kb, 1.5kb, 2kb).Compare the band obtained with samples. The infected sample would show up a single band at ~450bp while no band would be visible in lanes with healthy samples.

References:

Agiros, G.N. (2000). "Plant pathology", 4th Ed., Harcourt Asia, Academic press.

Bhat, A.I., Jain, R.K. and Ramiah, M. 2002. Detection of *Tobacco streak virus* from sunflower and other crops by reverse transcription polymerase chain reaction. Indian Phytopath, 55: 216-218.

Hull, R. (2002). "Mathews' Plant Virology", 4th Ed., Academic Press, USA.

Mullis, K.B., Faloona, F., Scharf, S.J., Saiki, R.K., Horn, G.T. and Erlich, H.A. (1986). Specific enzymatic amplification of DNA *in vitro*; the polymerase chain reaction. Cold Spring Harbor Symp. Quant. Biol. 51, 263-273.

Nalasco,G., de Blas, C., Torres, V. and Ponz, F. 1999. A method combining immunocapture and PCR amplification in a microtitre plate for the detectin of plant viruses and sub-viral pathogens. J.Virol Methods, 45: 201-218.

Rowhani, A., Maningas, M.A., Lile, L.S., Daubert, S.D nad Golino, D.A. 1995. Development of a detection system for viruses of woody plants based on PCR analysis of immobilized virions. Phytopathology, 85: 347-352.

Sharman, M. **,** Thomas, J.E. **,** Dietzgen, R.G., 2000. Development of a multiplex immunocapture PCR with colourim

Agrobacterium mediated transformation of plants

K. Nirmal Babu and T. E. Sheeja

Plant genetic engineering is an area in which tremendous progress has been made in the last five years A number of useful traits specifying pest resistance herbicide resistance and virus resistance have been successfully transferred to crop plants .Two factors that contributed to the success or plant genetic engineering are 1. Isolated plant cells can be regenerated into whole plants under tissue culture conditions 2. Foreign genes should be efficiently introduced into several plants using a soil phytopathgenic bacterium called *Agrobacterium tumefaciens*.

Agrobacterium tumefaciens is a common soil bacterium that naturally inserts its genes into plants and uses the machinery of plants to express those genes in the form of compounds that the bacterium uses as nutrients. *Agrobacterium* causes plant tumors commonly seen near the junction of the root and the stem, deriving from it the name of "crown gall disease" in dicot plants .The infected plant cells that grown as tumors exhibit uncontrolled, hormone independent growth. Agrobacterium has a large tumor inducing Ti plasmid s (140 to 235kb) from which a portion called transfer DNA (T-DNA) is transferred into plant cells .The T-DNA gets integrated into the nuclear genome of plants and becomes a part of the plant genome. The foreign DNA that is introduced into the genome is stably inherited.

The genes in the T-DNA region have plant promoters .The expression of T-DNA genes in the transformed plant cells leads to the synthesis of the plant hormones auxins and cytokinins and a group of tumor specific compounds called opines that are generally derived from an amino acid and an α-keto acid .The uncontrolled synthesis of hormones leads to the tumorous growth of the transformed plant cells.

The transfer of T-DNA from the bacterium to plant cells is mediated by proteins coded by two groups of virulence genes, 1. Chromosomal virulence genes (chv) and 2. Ti plasmid vir genes. Therefore the T-DNA transfer system in *Agrobacterium* has three components: 1. Ti plasmid virulence genes 2.Chromosomal virulence genes (chv) and 3. The T-DNA,

Initially attempts were made to introduce useful foreign genes into plants by first inserting the foreign gene in the middle of the T-DNA region of *Agrobacterium and* then infection of the plants. It was realized that such transformed plant cells grow only as tumors and could not be regenerated into whole plants. Therefore studied were carried out to understand the mechanism of T-DNA transfer so that Agrobacterium Ti plasmid based vectors could be developed that would efficiently introduce DNA into plant and would still allow the regeneration of transformed cells.

First it was discovered that T- DNA transfer would occur even after the removal of all tumor (onc) genes. Infact the only structural component of the T-DNA is sets of 25 base pair direct repeat ends of the T-DNA .All the intervening sequences consisting of onc genes would be removed and any foreign genes of interest along with selectable marker could be inserted in between the border sequences. Any sequence of DNA place between the boarder sequences is treated as T-DNA by *Agrobacterium* and transferred to plant cells. This led to the development of Ti plasmid vectors designated as "disarmed vectors" that did not cause tumor formation.

The second major step in the development of the Ti plasmid based vectors is the discovery that the T-DNA is transferred even after the T-DNA and the vir region or the Ti plasmid are separated and placed on two different plasmids. This discovery lead to the development of the most widely used binary vectors against the previously known co-integrate vectors. *A.tumefaciens* harboring a Ti plasmid in which the T-DNA region is deleted but the vir region is intact is used as a helper strain coding for all transfer functions. A binary vector is designed which is a small plasmid containing an E.coli selectable marker, a *A.tumefaciens* origin of replication, T-DNA left and right borders, a plant selectable marker such as kanamycin resistance gene (npt II) and a multiple cloning site into which foreign gene is of interest could be cloned. Because the binary plasmid is small and multiplies in E.coli a foreign gene could be cloned and the recombinant plasmid is subsequently transferred into the *A.tumefaciens* helper strain. Leaf disks are infected with *A. tumifaciens* harbouring the helper and binary plasmids. The DNA sequence between the borders, which consists of plant selectable marker, and the foreign gene is transferred to the plant cell. The transformed plant cells are selected on the basis of kanamycin resistance. Using the appropriate tissue culture conditions, whole plants designated as 'transgenic plants' are regenerated from transformed plant cells.

Selectable markers

Successful plant genetic engineering is based on delivery, integration and expression of defined genes into plant cells, which can be grown to generate transformed plants. Efficiency of stable gene transfer is not high even in the most successful transfer systems and only a fraction of the cells exposed integrate the DNA construct into their genomes. Moreover, a successful gene transfer does not guarantee expression, even by using signals for the regulation of transgene expression. Therefore, systems to select the transformed cells, tissues or organisms from the non-transformed ones are indispensable to regenerate the truly genetically transformed organisms. Antibiotic resistance genes allow transformed cells expressing them to be selected for out of populations of non-transformed cells. As part of this system, a selective toxic agent that interferes with the cellular metabolism is applied to a population of putatively transformed cells. The population of cells that has been transformed with and expresses a resistance gene is able to neutralize the toxic effect of the selective agent, either by detoxification of the antibiotic through enzymatic modification or by evasion of the antibiotic through alteration of the target.

The antibiotic resistance genes can be the genes of interest in their own right or they can be operatively linked to other genes to be transformed into the organisms.

1. An effective antibioti concentration of the toxic compound that suppresses growth of the non-transformed cells and does not cause detrimental effects to the transformed ones is used.
2. The resistance gene. Transcriptional and translational control signals fused to the resistance gene determine to a great extent the expression level of resistance. In addition, the gene sequence plays an important role, as some are more compatible with animal or plant systems or subgroups of animals and plants, such as monocotyledonous or dicotyledonous plants.
3. The material to be selected. In the case of plants, sensitivity to the selective agent depends on many factors, including the explant type, the developmental stage, tissue culture conditions and the genotype

Among the most widely used antibiotic resistance genes as selectable markers are neomycin phosphotransferase II (npt II) and hygromycin phosphotransferase (hpt). There are also other marker genes like gentamycin acetyltransferase (accC3) resistance and bleomycin and phleomycin resistance, but these are not as commonly used. The enzyme NPTII inactivates by phosphorylation of a number of aminoglycoside antibiotics such as kanamycin, neomycin, geneticin (or G418) and paromomycin. Of these, G418 is routinely used for selection of transformed mammalian cells. The other three are used in a diverse range of plant species; however, kanamycin has proved to be ineffective to select legumes and gramineae.

Hygromycin phosphotransferase is a suitable marker system for both plant and animal systems. The HPT enzyme inactivates the antibiotic hygromycin B. Hygromycin is usually more toxic than kanamycin and kills sensitive cells more quickly. It is nowadays one of the preferred antibiotic resistance marker systems for transformation of monocotyledonous plants, particularly gramineae (cereals and forages).

Leaf disc infection method

Day1

Cut off the healthy young leaves from a sterile ginger plant in the laminar flow chamber. Place the leaves in a petridish with sterile water.

Place 3 Whatmann 1 circles (sterile) in a petri dish. Wet them with sterile water and place two leaves over the filter paper. Cut out discs (about 5mm diameter) using a cork borer. Repeat this procedure until sufficient amount of disks were cut. Using a bacterial loop, transfer each disk from water to a piece of dry Whatmann #1 paper. Blot the disc dry and then place it inverted on MS medium. Incubate for two days in a tissue culture room.

In the evening start a 2 ml culture of the *Agrobacterium* strain (pBZ 100) (Kan 50/µml, 28^0C)

Day 2

Inoculate 20ml YEP culture with 0.1,0.2,0.4 and 0.6 ml of the starting culture in separate flasks.

Day 3

Monitor the four cultures by measuring the O.D at 600nm. Use a culture that reads 1 O.D. at 600nm. Transfer 5ml of culture to a sterile petridish. Using a bacterial loop, transfer the leaf disks to *Agrobacterium* culture. Leave for two minutes and then blot dry on a Whatmann1 paper. Place the leaf disks on MS medium and incubate for 2 more days.

Day 5

Transfer the leaf disks on to a shooting/selection medium containing kanamycin in which only plant cells transformed with neomycin phosphotransferase would grow. This medium also contains carbenicillin, which would kill *Agrobacterium*.

Day 25

Callus appears around the edges of the leaf disks.

Notes:

The calli developed on the leaf disc can be selected on the antibiotic amended medium and regeneration of the transformed cells can be carried out using subsequent stages in the shooting and rooting media. After the development of roots, the plants are transferred to moist vermiculite in pots and transferred to growth chambers and subsequently to green house. Transformation can be confirmed by biochemical / histochemical / phenotypic assay/ PCR techniques southern hybridization etc.

References

Liontenstein. C.P. and S. L. Fuller .1987. Vectors for the genetic engineering of plants. In: Genetic Engineering vol. 6ed/ V.J. Rigby. Pp. 102-183. Academic Press

Nester, E.W., M.P. Gordon, R.M. Amasino and M.F. Yanofsky.1984. Crown gall: a molecular and physiological analysis. Annu. Rev. Plant Physiol. 35: 387-413

Plant Molecular Biology _ Manual. 1988. Eds. S. B. Gelvin and R. A. Shilperoort. Kluwer Academic Publishers, London

Plant Genetic Transformation and Gene Expression: A Laboratory Manual. Eds. J. Draper *et al.* 1988. Blackwell Scientific Publications, Oxford.

Cytological techniques and photomicrography

R. Ramakrishnan Nair

Chromosomes being the carriers of genetic material attracted the attention of biologists through out the world. Various techniques have been developed to study the chromosomes at different stages of cell cycle. These techniques are being referred to as 'Cytological techniques'. Cytological techniques have passed through several stages of refinement so as to enable at present both extensive and intensive comparisons of chromosome number, structure and meiotic behavior in large samples within a relatively short period.

Chromosomes were earlier studied from serial sections following microtomy of the tissue embedded in paraffin block. This was later replaced by smear and squash techniques. The squash technique used for root tips, leaf tips and other somatic tissues consists of softening the pretreated and fixed tissue in HCl or a mixture of enzymes. The softened meristematic region is squashed in a drop of stain on a slide, pressed under a cover slip and cells are flattened by gently warming the slide and tapping the cover slip a number of times to spread chromosomes. Smear method is applicable to pollen mother cells in which anthers are squeezed out to release PMCs and stained to observe meiotic stages. A more recent method, splash technique involves dissociating the cells of the tissue with enzymes, obtaining a thick cell suspension by maceration and splashing the suspension on a slide from a distance. The cells and chromosomes spread via surface tension. Several staining solutions using Carmine (Aceto-carmine, Propionic-carmine), Orcein (Aceto-orcein, Propionic-orcein, Lacto-Propionic orcein) and Fuchsin (basic fuchsin) are involved in staining chromosomes for normal observation. With the usual fixation and staining procedures, these solutions stain the whole chromosomes uniformly except at the centromere and secondary constrictions, which remain unstained. As a result chromosomes are identified and compared only on the basis of length and the position of primary and secondary constrictions. A more precise identification of chromosomes is possible by differential staining techniques. These involve the use of special tissues, fixation, treatment or staining schedules by which specific chromosome segments are differentiated either unstained or more intensely stained bands than the rest of the chromosomes.

The stages in the division cycle of a cell at which chromosomes are suitable for observation by **light microscopy** are mitotic metaphase for somatic chromosomes and pachytene to anaphase –I for meiotic chromosomes. The preparation of mitotic metaphase spreads involves following steps:

1. Selection of tissue with high mitotic activity (root tips, young leaves, apical meristems)
2. In vivo or in vitro treatment with a mitotic arresting agent (prefixation)
3. Fixing (and storing) tissues or cells
4. Making chromosome preparations
5. Making temporary preparations permanent.

For meiotic chromosomes, the tissues required are microspore mother cells from anthers or megaspore mother cells at proper stages of division. Except for prefixation, which is skipped, all steps listed for mitotic preparations are followed for meiotic preparations also.

Prefixation

The objectives of pre-fixation are:

1. Clearing of the cytoplasm and softening the tissue
2. Inhibiting the spindle formation
3. Dissolving of middle lamella
4. Contraction and scattering of chromosomes
5. Clearing the surface conditions of the cells and tissues for rapid penetration of the fixative and the stain

6. Removal of extra nuclear structures, separation of chromosomes and clarification of chromosomal structure.

Chemicals commonly used for pre-treatment are colchicine (0.01-0.05%), acenaphthene (Saturated solution), chloral hydrate (0.01-0.05%), 8-hydroxyquinoline (0.001-0.004 M) and α –bromonaphthalene (Saturated solution). Usually pre-treatment for 2-4 hours with the above chemicals is sufficient under the cold condition of a refrigerator. In some cases such as cereals pretreatment in ice cold water for 15-24 hours is sufficient.

Fixation

The objectives of fixation are:

1. To kill the cells suddenly and uniformly so that they retain as near as possible, the same appearance which they possessed in life
2. To preserve the tissues and cells by the inhibition of putrifaction and autolytic changes
3. To induce differences in the refractive indices of certain cell elements. This helps in differential visibility of the different elements.
4. To facilitate proper staining of tissues

For chromosomes, the fixation should:

1. Increase visibility of chromosomes
2. Clarify the details of chromosome morphology such as chromatic and heterochromatic regions and the primary and secondary constriction
3. Maintain the structural integrity of the chromosome intact and it should be able to enhance the basophilia of the chromosomes.

Since all the above properties are rarely found within a single chemical, a fixative is normally a combination of several compatible fluids, which together satisfy all the above requirements. Most commonly used components of fixing solutions are **Ethyl alcohol, Acetic acid, Formaldehyde, Propionic acid and chloroform.** Some commonly used fixing solutions are

1. Carnoy's fluid
   Glacial acetic acid 1 part
   Absolute ethyl alcohol 3 parts
2. Carnoy's fluid –II
   Glacial acetic acid 1 part
   Cloroform 3 parts
   Absolute ethyl alcohol 6 parts
3. Carnoy's and Lebrun's fluid
   Glacial acetic acid 1 part
   Chloroform 1 part
   Absolute ethyl alcohol 1 part
4. Battaglia's 511 mixture
   95% ethyl alcohol 5 parts
   Chloroform 3 parts
   Glacial acetic acid 1 part
   40% aqueous formaldehyde 1 part

Staining

Carmine, orcein, basic fuchsin, methyl green, giemsa etc. are stains used for making chromosome staining solutions. Some of the commonly used solutions are

1. Aceto-carmine: Prepared by dissolving 1 g carmine in 100ml boiling 45% acetic acid in a flask fitted with reflux condenser. Reflux for 1-2 hours. Cool and filter. To make propiono-carmine, substitute acetic acid with propionic acid.

2. Aceto-orcein: Use orcein instead of carmine in the above procedure. Propiono-orcein is prepared by replacing acetic acid with propionic acid.
3. Lacto-propionic orcein: Stock solution: Dissolve 2 g of orcein in 100 ml of a mixture of equal parts of lactic acid and propionic acid at room temperature. Filter. To use, dilute the stock solution to 45% with distilled water.
4. Schiff's reagent (HCl-Leucobasic fucshin): Dissolve 0.5 g basic fuchsin by pouring over it 100 ml boiling distilled water. Shake thoroughly and cool to 80° C. Filter and add 15 ml 1N HCl to the filtrate. Add 1.5 g potassium metabisulphite (K2S2O5) and allow the solution to stand ina well stoppered bottle in the dark overnight or until a light straw or a faint pink colour appears. If not completely decolorized, add 500 mg activated charcoal and shake for 2 minutes; filter, Store in an amber bottle and refrigerate.
5. Methyl green: Stock solution: Dissolve 0.11 g methyl green in 25 ml phosphate buffer, pH 6.8. To use: dilute 2 ml stock solution in 50 ml phosphate buffer, pH 6.8.
6. Giemsa stain: Stock solution: 1 g Giemsa powder + 66 ml glycerine +66 ml methanol. Dissolve Giemsa powder in glycerine at 60°C for 1 hour with constant stirring. Add methanol, continue stirring at 60°C for 1 hour. Store in refrigerator. To use: 5 ml filtered stock + 1.5 methanol + 1.5 ml stock citrate buffer + 60-100 ml distilled water.

Other solutions required for chromosome staining:

Enzyme solution: Solution in water of 2% cellulase (Sigma) and 20% pectinase (Sigma).

Barium hydroxide: 100 ml distilled water + 5 g Ba (OH) 2 shake vigorously and use immediately.

2xSSC: 17.4 g NaCl + 8.8 g sodium citrate, make 1 litre with distilled water, adjust pH to 7.0 with 1N HCl.

Citrate buffer: A: 2.1 g citric acid + 100 ml distilled water. B: 14.2 g sodium phosphate + 500 ml distilled water. Mix 4.55 ml of A+15.45 ml of B for use as stock citrate buffer.

Phosphate buffer: pH 6.8: 0.02 M KH_2PO_4. Titrate to pH 6.8 with 50% Na OH

Some Staining Procedures:

Carmine/Orcein staining to study mitosis:

1. Warm the fixed root tips in 1N HCl in watch glass for five minutes
2. Transfer the rroot tips to another watch glass containing 2% aceto carmine and boil for two to three mnts.
3. Lift a root tip from the watch glass and place it on to a drop of acetocarmine in the centre of a slide.
4. Cut off and remove the older part of the root tip with the help of a needle. Place the cover glass on the tip.
5. Squash the root tip by applying uniform pressure on the cover glass with the thumb, through a piece of blotting paper.
6. If necessary the cells can be spread with the blunt end of a pencil (A slight heating on the flame improves staining).
7. Seal the cover slip with sealing wax or nail polish.
8. Observe the slide under the microscope for stages of mitosis.

If root tips are fixed in propionic acid: ethanol 1:3, then they can be stained in propiono-carmine/lacto-propionic orcein.

Feulgen staining to study mitosis

1. Prefix and fix the root tips
2. Wash the root tips in distilled water
3. Hydrolyse and macerate by placing in 1N H Cl at 60° C in a staining vial for 5- 11 minutes
4. Rinse out HCl in several changes of water

5. Place the root tips in a small quantity of stain in a tightly stoppered vial. Place the vial in dark. Staining may take 30 mnts to several hours.
6. Wash the root tips in several changes of water and squash in 45% acetic acid.
7. Observe under microscope.

Carmine staining to study meiosis:

a. Take flower buds serially from an inflorescence, starting from the smallest and working upto the largest, until a correct bud having divisional stages is found.
b. Dissect out a single anther from the selected bud with the help of a needle.
c. Place the anther in the center of a clean slide and smear it with the help of a scalpel.
d. Place a drop of acetocarmine on the smeared material stir with a needle to spread the cells.
e. Remove the anther debris and add a clean coverslip over the drop
f. Warm the slide by passing it back and forth over a clear flame
g. Press the slide with cover slip gently between layers of filter paper holding it in place at corners by fingers
h. Examine under microscope for required stages with satisfactory spread and staining
i. Seal with paraffin wax and keep in refrigerator.

The above procedure can be adopted for fresh as well as fixed flower buds.

Feulgen staining to study meiosis:

1. Fix the flower buds in Carnoy,s fluid for 2-4 hours
2. Dissect out anthers with proper divisional stages
3. Hydrolyse the anthers in 1 N HCl at 60˚C for 10 mnts
4. Wash in tap water –3 changes
5. Stain in leuco-basic fuchsin
6. Squash in 45% acetic acid or 2% aceto carmine

Single anther should be squashed on one slide.

Giemsa C-banding of somatic chromosomes:

1. Prefix and fix root tips
2. Soften root tips in enzyme solution (0.5-2 h)
3. Wash in water
4. Macerate root tips in 45% acetic acid on a slide. Apply cover slip and press gently.
5. Remove cover slips on dry ice or liquid nitrogen
6. Soak slides in 95% ethanol (5 mnts-2 h), air dry
7. Treat slides with saturated barium hydroxide solution for 5 mnts, wash in water for 10 mnts, 3 changes
8. Place in hot 2xSSC buffer for 1 h at 60˚C. Wash in water, 3 changes
9. Stain in Giemsa (30 sec to a few mnts). Wash in water, air dry.
10. Place overnight in xylene. Mount in a synthetic resin.

Preparation of permanent slides:

Slides of mitotic as well as meiotic preparations can be made permanent using the following procedure.

1. Remove the sealing wax carefully from the edges of coverslip using a sharp blade
2. Carefully separate the cover slip and slide
3. Note the side of cover slip as well as slide containing cells
4. Pass the slide as well as coverslip to 45% acetic acid (5mnts)
5. Pass to Acetic acid : n Butanol 3:1 (5mnts)
6. Pass to Acetic acid : n Butanol 1:1 (5mnts)
7. Pass to Acetic acid : n Butanol 1: 3 (5mnts)
8. Pass to n Butanol (2 changes) (5mnts)

9. Mount the cover slip on a frsh slide with DPX mountant
10. Put a drop of DPX mountant to the slide with cells and add a fresh clean cover slip.

Photomicrograhy

Photomicrography is the technique of making photographs through a compound microscope. It involves coupling a camera to a microscope to produce enlarged photographs of very minute, microscopic detail. In fact, modern photomicroscopes incorporate the camera as an integral part of the microscope design.

To make good photomicrographs, you must know how to use the microscope efficiently, and to use microscope efficiently, you must have a good working knowledge of the instrument. The knowledge includes the selection, capabilities, and limitations of the optical components; how to adjust the microscope; how to illuminate the specimen; and even how to prepare specimens for examination and photography.

Following are some of the tips for better photomicrography under bright light:

1. Prepare the mitotic / meiotic specimens neatly with good staining, chromosome spread and with out debris
2. Mark the cell to be photographed by putting a circle around it under the low power of the microscope
3. Use slides of thickness 1 mm and cover slips of 0.1 7 mm for better results.
4. Use microscopes fitted with achromatic / apochromatic objectives. Apochromatic objectives are always better
5. Use oil immersion objectives for chromosome photography (To reduce magnification, use a photolens of low power).
6. Use of optimum lighting to illuminate specimen is always better than too much light or less light
7. Adjust the aperture to give optimum contrast (Adjust it till the light just start fading)
8. Adjusting the light to an exposure time of 1 to 10 seconds will be ideal.
9. Use green filter for black and white photography and blue filter for colour photography
10. 11.Use of slow speed films (<100 ASA) will give better results for chromosome photography
11. Focus the specimen sharply and ensure that complete specimen is within the frame
12. Ensure that your microscope is placed on a firm table to avoid risks of shaking.
13. Record the exposure details in the specified format. If your photomicrographic system is attached with a databack / computer, you can utilize the same.

With the advancement of technology photomicrographic systems have improved a lot during the last decade. Recently digital cameras are replacing normal cameras in microscopes also. Such advancement is useful in analyzing the chromosome preparations directly using image analysis packages without taking a photographic print.

Genome analysis by karyotyping

Karyotype

A characteristic that identifies a set of chromosomes of an individual or species is called a karyotype. Usually a diploid individual will be having two homologous sets of chromosomes (genomes) received from either parent. It is possible to identify the individual chromosomes of a genome on the basis of following characteristics

1. The total chromosome length (L+S)
2. The arm ratio (ratio of the lengths of the long and the short arms as determined by the position of the centromere)
3. Differences in relative chromosome size
4. Centromeric index (short arm length/total cromosome length x 100)
5. Position of the secondary constrictions and NORs.

Of these, arm ratio, relative chromosome length and centromeric index are constant for a particular chromosome of a particular genome and any significant changes in these parameters indicate chromosome rearrangements.

Classification of chromosomes based on arm ratio (Levan, 1964)

Centromeric position	Arm ratio (r) L/S	Symbol	Designation
Median sensu stricto	1.0	M	Metacentric
Median	1.7	M	Metacentric
Sub median	1.71-3.0	Sm	Sub metacentric
Sub terminal	3.01-7.0	St	Sub telocentric
Terminal	>7	T	Acrocentric
Terminal sensu stricto	-	T	Telocentric

Idiogram

Diagrammatic representation of chromosome morphology used diagnostically for comparison of the karyotypes of different species and varieties is known as idiogram. The construction of idiogram is based on measurements of total chromosome length, arm length ratio (long arm/short arm), centromeric position and position of secondary constriction. Chromosomes are arranged in idiogram in the decreasing order of length. Comparison of idiograms of the members of a family or genera may help to trace the evolutionary relation between the members.

Isolation of plasmid DNA from *E. coli*

A.I.Bhat and M.V.Jiby

Many methods are available to purify plasmids from bacteria. These methods involve: growing of bacterial culture, harvesting and lysis of bacteria and purification of plasmid DNA. Bacteria are harvested from late logarithmic phase grown cultures by centrifugation. The supernatant is removed from cell pellets. Many methods such as nonionic detergents, organic solvents, alkali and heat are available to lyse the bacterial cells. The choice depends on the size, strain of *E. coli* and technique used subsequently to purify DNA. Among all alkali lysis method is largely used. The method In the following protocol, preparation of plasmid DNA by alkaline lysis with SDS is described. The method was developed by Birnboim and Doly (1979). In this procedure, bacterial suspension are exposed to high anionic detergent at high pH which will open cell wall, denature chromosomal DNA and proteins and release plasmid DNA into the supernatant. The denatured material is removed by centrifugation and native plasmid DNA recovered from supernatant s subjected to further purification. Large plasmids (>15 kb) are susceptible to damage during cell lysis and subsequent purification and hence must be handled with care. Alkali lysis method works well with all strains of *E. coli* and with bacterial culture volume ranging from 1 ml to 500 ml. The resulting DNA preparation may be used for PCR, restriction digestion and nucleic acid hybridization test. With further purification by treatment with polyethylene glycol (PEG), the preparation may be used as template in DNA sequencing.

Requirements
Materials

Balance
Beakers
Horizontal gel apparatus and power pack
Ice flakes
Magnetic stirrer
Measuring cylinders
Microcentrifuge
Microcentrifuge tubes

Micropipette (adjustable volume)

Microtips
Microwave oven
pH meter
Shaking waterbath
Transilluminator

Reagents

Master plate containing putative transformants
Agarose
Ethidium bromide
Phenol: chloroform: isoamylalcohol (25:24:1)
Ethanol

Solution I:

20% Glucose ------2.25 ml
0.5M EDTA, pH 8.0---------1.00 ml
Sterile distilled water ------ 45.50 ml

Solution II

10N NaOH	0.4 ml
10% SDS	2.0 ml
Sterile distilled water	18.6 ml

(not be autoclaved)

Solution III

3M sodium acetate pH4.8. Dissolve 40.81 g of sodium acetate.$3H_2O$ in 60 ml water. Adjust pH to 4.8.

Solution IV

1M Tris pH 8.0	2.50 ml
3M NaoAC pH 4.8	1.65 ml
Sterile water	45.85ml

RNase A:

Dissolve pancreatic RNase A at a concentration of 10 mg/ml in 10 mM Tris-HCl pH 7.9 and 15mM NaCl. Heat to 100^{o}C for 15 min and allow to cool slowly to room temperature. Dispense into aliquots and store at -20^{o}C.

TE, pH 8.0:

Tris base 0.1211 g
EDTA 0.0372 g
Adjust pH to 8.0 with HCl and make the volume to 100 ml.

10X TAE buffer:

Tris base: 48.4g
Glacial acetic acid: 11.42 ml
0.5 M EDTA (pH 8.0): 20 ml
water: to make 1000 ml

Gel loadimg dye (for DNA samples)

1% bromophenol blue	200 µl
Glycerol	200 µl
10% SDS	60 µl
0.5 M EDTA	50 µl
10X TAE	60 µl
H^2O	30 µl

Procedure

1. Select colonies which are positive in Rapid and colony PCR and inoculate them individually in 5 ml-broth with appropriate antibiotic at 37^{o}C for 14 h with shaking (200rpm).
2. Collect the growth in a 1.5 ml eppendorf tube and spin for 30 sec in microcentrifuge, pour off supernatant
3. Suspend pellet thoroughly in 100 µl of solution I by vortexing immeadiately. Keep on ice for 5 min.
4. Add 200 µl freshly prepared solution II at room temperature. Mix gently by inverting tubes (do not vortex) and keep on ice for 5 min or until SDS precipitates. Samples should be viscous and stringly when cap opened at completion of this step.
5. Add 150 µl solution III at room temperature and mix gently by inverting tubes several times. Keep on ice for 5 min.

6. Spin 3 min in centrifuge in cold room for 5 min at 14,000 rpm at 4°C. Without disturbing the pellet, carefully remove the supernatant and put into a fresh eppendorf tube. Discard the pellet.
7. Add 0.9 to 1.0 ml cold ethanol. Mix gently and freeze at -70°C for 15 min.
8. Spin in centrifuge for 2-3 min. Remove and discard supernatant.
9. Add 100 µl solution IV to the pellet and gently mix till pellet is dissolved.
10. Add 200 µl cold ethanol and freeze at -70°C for 15 min again.
11. Spin in microfuge for 2-3 min and remove the supernatant. Wash pellet with chilled 70% ethanol to remove excess salts. Dry final pellet for 5 min under vacuum.
12. Suspend pellet in 20 µl of TE and store at -20°C until needed.

References

Brinboim, H.C. and Doly, J. 1979. A rapid alkaline extraction procedure for screening recombinant plasmid DNA. Nucleic acids Res. 7: 1513-1523.

Clewell, D.B. and Helinski, D.R. 1969. Supercoiled circular DNA protein complex in *Escherichia coli*: purification and induced conversion to an open circular DNA form. PNAS, USA 62: 1159-1166.

Sambrook, J. and Russel, D.w. 2001. Molecular cloning (3rd edition). Cold Spring Harbor Laboratory Press, New York.

Isolation of total RNA from plants and RT-PCR

A. I. Bhat and A. Siljo

The extraction procedure for obtaining total RNA consists the following steps. (i) effective disruption of tissue (ii) inactivation of ribonuclease (RNase) activity and (iii) separation of RNA from protein, carbohydrates, polysaccharides etc. Major portion of the isolated total RNA comprises of ribosomal RNA (rRNA) (85%) followed by low molecular weight RNA species (tRNA, siRNA etc). So purity and integrity of isolated total RNA is a critical determinant of its effectiveness to be used in different molecular biology procedures.

Among different methods used to extract total RNA from plant tissues, a combination of guanidium thiocyanate and phenol chloroform extraction followed by precipitation of RNA using isopropanol is the most widely used method. Guanidium thiocynate is chaotropic agent that destroys the 3D structure of proteins and converts most protein in a random coiled state. Along with guanidium thiocynate different reducing agents and detergents are also used. Reducing agents help in breaking down the protein disulfide bonds and detergent such as SDS helps in disrupting hydrophobic interactions and thus inhibiting RNase activity.

The basic principle for purification of RNA from DNA and other protein is based on the pH. At acidic pH, RNA is soluble in the aqueous phase The DNA and proteins are insoluble and are located in the organic layer. At alkaline pH, DNA is soluble in the aqueous phase while RNA is insoluble. After extraction, total RNA is separated from proteins by phenol/chloroform extraction. After centrifugation, RNA selectively portions into the upper aqueous phase which is collected and re-extracted with phenol-chloroform followed by chloroform alone. Finally total RNA in the aqueous phase is precipitated in the presence of salt by adding two to three volumes of ethanol followed by incubation at -20º C overnight or at -70ºC for 30 min. Alternatively, RNA can also be purified after DNase and proteinase treatment followed by column purification. The purified RNA should be analysed by gel electrphoresis to check integrity and quality.

Unlike DNA, extraction of RNA requires special care because of ubiquitous presence of RNases that degrade ssRNA molecules. Ribonucleases are highly stable and active enzymes that require no co-factors to function and are released immediately from cell membrane and membrane bound organelles upon lysis. Hence the extraction buffer in addition to additives should also contain chemicals to prevent RNase contamination. A protocol that works with one plant group may fail with others. Consequently, a number of RNA isolation methods have been developed for different target groups. In the following a protocol based on acid guanidium phenol thiocynate (AGPC) described by Chomczynski and Sacchi (1987) is described.

Requirements

Materials

-20ºC freezer
-70ºC freezer
Balance
Centrifuge tubes
Eppendorf tubes
Gloves (powder free)
Magnetic stirrer
Microcentrifuge
Micropipette

Microtips
Pestle and mortor
pH meter
Vortex
Waterbath or temperature block

Reagents

Infected plant samples
Denaturing solution:

4 M Guanidium thiocyanate, 25 mM sodium citrate (pH 7.0) and 0.5% sacrcocyl. Autoclave and store at room temperature up to three months. Add 0.5% β mercaptoethanol for the required volume just before use.

2 M sodium acetate pH 4.0
Phenol (water saturated)
Chloroform: isoamyl aclcohol (49:1)
Isopropanol
Ethanol 75%

Procedure

1. Harvest the tissue (50 mg) and mince them on ice.
2. Homogenize the tissue with 500 µl of denaturing solution in a pestle and mortor at room temperature.
3. Transfer the homogenate to a centrifuge tube and add sequentially the following reagents with thorough mixing after each addition.
   50 µl of 2 M sodium acetate pH 4.0
   500 µl phenol (water saturated)
   100 µl chloroform: isoamyl aclcohol (49:1)
4. Shake vigorously for 10 sec and cool on ice for 15 min before centrifuging at 10,000 g for 20 min at 4°C,
5. Transfer the aqueous phase to a fresh tube and mix with 500 µl of isopropanol and incubate for 1 to 2 h at –20°C.
6. Collect the precipitate by centrifuging at 10,000 g for 20 min.
7. Discard the supernatant and add 500 µl of 70% ethanol to the tube.
8. Centrifuge at 10000g for 10 min. Discard the supernatant and vacuum dry the RNA pellet
9. Dissolve the RNA pellet in 50 µl RNase free water and store at -70°°C.

Reverse transcription polymerase chain reaction (RT-PCR)

Reverse transcription polymerase chain reaction (RT-PCR) is a method to amplify complementory DNA (cDNA) copies of RNA. In this, the first step is the enzymatic conversion of RNA to a single stranded cDNA template. An oligo deoxynucleotide primer is hybridized to the mRNA and is then extended by an RNA-dependent DNA polymerase (Reverse transcriptase) to create cDNA copy that can be amplified by PCR. The enzyme reverse transcriptase (RT) catalyses the synthesis of cDNA from RNA template. Total RNA isolated from plant can be used as template to initiate cDNA synthesis. The primer for first strand cDNA synthesis can be specifically designed to hybridize to a particular target gene. Amplification of the desired portion of cDNA can be achieved in PCRs primed by sense and antisense oligonucleotide primers corresponding to specific sequences in particular cDNAs. For maximum specificity, the antisense primer should be located upstream of the oligonucleotide used to prime cDNA synthesis. Positive and negative controls should always be included when setting up RT-PCRs.

Requirements

Materials

-20ºC freezer
Balance
Eppendorf tubes
Gel documentation unit
Gloves
Horizontal gel apparatus with power pack
Ice flakes
Microcentrifuge
Micropipette
Microtips
Microwave
PCR work station
Thermal cycler
Thin walled PCR tubes for amplification
Transilluminator
Water bath

Reagents

Agarose
Dithreitol (Commercially available)
dNTP mix (10 mM) (Commercially available)
PCR amplification buffer(10x) (Commercially available)
DNA molecular weight markers (Commercially available)
Ethidium bromide (10 mg/ml)
$MgCl_2$ (25 mM) (Commercially available)
Oligonucleotide primers (both sense and antisense, specific for the target sequence to be amplified)
Reverse transcriptase (Commercially available, thermo sensitive, always store at -20ºC
RNA template (either isolated from purified virus (see page) or total RNA isolated from infected plants (see page) may be used.
RNase free water
RNase inhibitor (RNasin) (Commercially available)
TAE buffer
Taq DNA polymerase (Commercially available)

Procedure

1. Thaw all components (except reverse transcriptase) and give a brief spin to bring all components into the bottom of tube. Assemble RT-PCR reaction components on wet ice and prepare amplification mix into PCR tube in order given below.

10X PCR buffer	10.0 µl
0.1 M DTT	10.0 µl
25 mM $MgCl_2$	5.0 µl
10 mM dNTP mix	1.0 µl
Forward primer	1.0 µl
Reverse primer	1.0 µl
rRNasin (32 units/µl)	0.5 µl
AMV RT (20units/µl)	1.0 µl
Taq DNA polymerase	1.0 µl
Template RNA*	40.0 µl

Water to make	100.0µl

*Should be denatured by heating at 75°C for 10 min and snap cooling on ice for at least 3 min before adding into PCR mix.

2. Mix reaction components assembly by inverting and place the tube on a thermal cycler and proceed with the thermal cycle profile chosen for the reaction.

Teperature programme for RT-PCR reaction

Step	Temp (°C)	Duration
Reverse transcription	42	45 min
Denaturation	90-94	30 –60 sec
Annealing*	37-55	30- 60 sec
Extension$	72	1- 3 min
Final extension	72	10 min

Repeat steps denaturation, annealing and extension for 30 to 40 times

* depend on the Tm of primers.

$ Provide one minute for every 1000 bp to be amplified.

3. Run about 10-20 µl of reaction on 1% agarose gel and visualize the DNA with the help of transilluminator and document the image as explained earlier.

Selected references

Bhat, A.I., Jain, R.K. and Ramiah, M. 2002. Detection of *Tobacco streak virus* from sunflower and other crops by reverse transcription polymerase chain reaction. Indian Phytopath, 55: 216-218.

Pappu S S, Brand R, Pappu H R, Rybicki E, Gough K H *et al*, A polymerase chain reaction method adapted for selective cloning of 3' non translated regions of potyviruses: application to dasheen mosaic virus, *J Virol Methods,* 41 (1993) 9-20.

Chomczynski, P. and Sacchi, N. 1987. Single step method of RNA isolation by acid guanidium thiocynate phenol chloroform extraction. Anal. Biochem. 162: 150-159.

Levy, L., Lee, I.M. and Hadidi, A. 1994. Simple and rapid preparation of infected plant tissue extracts for PCR amplification of virus, virioid, and MLO nucleic acids. J Virol. Methods, 49: 295-304.

MacKenzie, D.J., McLean, M.A., Mukerji, S., Green, M. 1997. Improved RNA extraction from woody plants for the detection of viral pathogens by reverse transcription polymerase chain reaction. Plant Disease 81: 222-226.

Siju, S., Madhubala, R. and Bhat, A.I. 2007. Sodium sulphite enhances RNA isolation and sensitivity of *Cucumber mosaic virus* detection by RT-PCR in black pepper. Journal of Virological Methods 141: 107-110.

Plant molecular farming

A I Bhat

Molecular farming activity has existed since the first higher plant was successfully transformed (Fraley et al. 1983), because any protein has the potential of being a protein product. One of the earliest marker genes that scientist have used in developing transformation system in plants, *uids* (Jefferson et al. 1987), is now a molecular farming product (Kusnadi e al. 1998; Witcher et al. 1998). The first report of human antibodies produced in plants was by During (1998) and was expand to include secretary antibodies by Hiatt et al. (1989). The first report of a protein being produced in plants for the specific purpose of extraction, purification, and sale of that protein was by Hood et al. (1997), which detailed the production of avoiding, an egg protein with several important properties. Aprotinin, one of the first molecular farmed pharmaceutical proteins to be produced in plants (Zhong et al. 1999), may soon be used on medical patients for wound closure and to suppress the systematic inflammatory response during surgery.

Molecular farming system

There are currently four methods of protein production from plants: (1) stable nuclear transformation of a crop species that will be grown in the field or a green house, (2) stable plastid transformation of crop species, (3) transient transformation of crop species, and (4) stable transformation of plant species that is grown hydroponcally such that the transproteins is secreted into the medium and recovered.

Products of molecular farming

Proteins currently being produced in plants for molecular farming purposes can be categorized into four broad areas: (1) parental therapeutics and pharmaceutical in termediates, (2) industrial proteins (e.g, enzymes), (3) monoclonal antibodies (MAbs), and (4) antigens for edible vaccines (see Table).

(1) Parental therapeutics and pharmaceutical intermediates

This group includes all proteins used directly as pharmaceuticals along with those proteins used in the marketing of pharmaceuticals. The list of such proteins is long, growing, and includes such products as thrombin and collagen (therapeutics), and trypsin and aprotin (intermediates). Products in this classification must generally be manufactured under stringent cCMP (current god manufacturing practices) procedures and be of high purity.

(2) Industrial proteins-enzymes

This group includes hydrolases, encompassing both glycosidases and proteases. Oxido-reductase enzymes such as laccase, a fungal enzyme used in fiber bleaching and bioglue of wood products (Hood et al. 2004; Bailey et al. 2004), represent a separate classes of industrial enzymes. Enzymes involved in biomass conversion for the purposes of producing ethanol are candidates for molecular farming. All of these products will usually be characterized by the fact that they are used in very large quantities and must therefore be produced very in expensively (Hood et at al.1999). Regulatory hurdles as they exit today may be a major hindrance to the molecular farming of these products since very large acreage will be required. This may necessitate deregulation by federal regulatory agencies of reduced regulations for this type of product based on safely assessments.

(3) Monoclonal antibodies

This group includes all antibody forms (IgA, IgG, IgM, secretory IgA, etc.) and antibody fragments (Fv). They can be produced in plants in both glycosylated and non glycosylated forms. These plant derived MAbs (plantibodies) have the potential of alleviating the serious production bottleneck that currently exists as dozens of new MAb products attempt to reach the marketplace. Examples of plant –derived MAbs in product development include a-caries for prevention of dental decay and a- herpes for prevention of herepes transmission.

(4) Antigens for edible vaccines

Specific protein antigens can be produced in plants that will induce a humoral immune response when eaten by an animal or human (Streatfield et al. 2003). Protection studies have shown good efficacy when these edible (or oral) vaccines have been used. In some cases, protection has actually been better with the edible vaccines that with the commercially available vaccine (Lamphear et al.2004).

Molecular farming products (to be close to or on the market)

Product	Companuy or companies	Use
Trypsin	ProdiGene	Pharmaceutical intermediate
GUS	ProdiGene	Diagnostic regent
Avidin	ProdiGene	Immunological regent
Aprotinin	ProdiGene, Large Scale Biology	Reduce SIR and bleeding, promote wound closure, mammalian cell culture
Collagen	ProdiGene, Medicago, Meristem Therapeutics	Gel caps, skin sealant, scar treatment
Lipase	Mesristem Therapeutics	Exocrine pancreatic insufficiency, steatorrhea, cystic fibrosis
Lactoferrin	Ventria, Meristem Therapeutics	Natural defense protein against infections, iron repository
Lysozyme	Ventria	Anti-viral, anti-bacterial, anti-fungal
Brazzein	ProdiGene	Natural protein sweetener
TGEV edibleVaccine	ProdiGene	TGEV vaccine in swine
α-Caries MAb	Planet Biotechnology	Prevention of dental caries
α-Herpes MAb	Epicyte Pharmaceutical	Prevention of herepes transmission

Glycosylation

The glycosylation of transproteins in plants differs slightly from those produced in transgenic animals or animal cells *in vitro* (Lerouge et al. 1998). The addition of xylose and change from a β1→6 to a β1→3 linkage of focus are typical in plants. A significant difference with transprotein production in plants is their inability to add sialic acid to glycoproteins (Lerouge et al. 1998). This sugar has been implicated in longer clearance times for proteins in the blood and therefore is a major factor for a select group of pharmaceutical proteins. These difference have had minimal or no effect on the function of the transproteins products from plants to date (Hood et al. 1997; SamynPetit et al. 2001; Woodard et al. 2003).

Public acceptance

Sales are a good measure of the public's perceived benefits of specific products. However today's public also wants to know that not only is there a benefit for the direct end user, but that there are otherwise no significant risks to the general public. This is illustrated by the recent concerns and debates over the use of GMO products produced in plants. While the initial concern involved GMO food products, this now encompasses non-food products as well. The fear is that the non-food products may inadvertently enter the food chain and present an un international risk. A quantitative safety assessment model for non-food products produced in agricultural crops. On the surface, this appears to be an unprecedented use of food crops to produce pharmaceutical and industrial products. However, the use of plants to produce non-food products is not unlike the current use of other food products, such as eggs or yeast, to produce pharmaceuticals. The difference is the later have well established complains programmes, which are in line with the production of pharmaceutical products rather than the production of food.

Such compliance programmes start with the regulatory agencies that represent the public safety concerns. The regulatory agencies take the position that the non-food products are unsafe until proven otherwise. There is a regulatory framework in place specifically targeted to ward the introduction of non-food products when using plants as the production system. There are strict rules on agronomic practices, which are targeted to keep non-food products out of the food chain. Unfortunately, in any system, plants included, it is not possible to eliminate all possibilities of united exposure due to unforeseen circumstances such has an accident, a natural disaster, or an act of non-compliance. In the event of unintended exposure, a quantitative assessment can be used to determine if there is case for alarm. While some of these products may present a significant risk to the public, many of the products would not pose a hazard if they were inadvertently introduced. A quantitative risk assessment similar to that employed for other regulated products can be used to evaluate what type of measures need to be taken for specific products. This approaches as proven useful to realize the benefits of new technology and to reduce or eliminate risks.

Future research

One of the keys to success in the future will be the level of expression of the recombinant protein in plants. This is the one of the most important aspects with regard to economics. The expression level affects the cost of growing, processing, extraction, purification and waste disposal. Clearly there will be a drive towards higher levels expression and there is much more room for improvement compared to other established systems.

Expression is also a major regulatory concern. Whether or not the protein is in specific tissues will enable or nullify exposure to the environment. There as already been work to show that expression can be limited to specific tissues, thus reducing regulatory concerns. As an example, keeping the protein out of pollen can reduce inadvertent exposure to the environment. However, this does not remove the possibility that the pollen will outcross with other plants and intermix with food crops. There are physical isolation requirements imposed by the regulatory agencies to prevent this room occurring. There may become cases where genetic control expression is also warranted either for economic of safety concerns, depending on the product. Possibilities including male sterile crops, included expression, or sequence that prevent germination or the expression on the protein product in on-food products have been discussed. Some combination of these different limitations on expression will most likely find a way into future programmes.

The other regulatory concern is that the pathway to commercialization for human therapeutics has not been proven. The industry will anxiously await the establishment of a clear road map detailing how this process is similar to, and/or different from, the existing protocols today. With the first approved therapeutic products will also come the realization of the many benefits of transgenic plant technology. These real benefits should also help public acceptance and open the way for a much more rapid acceptance of this technology.

Important questions concerning the glycosylation, immunogenicity, accumulation and stability of the transproteins and being answered,. Academic laboratories have been instrumental in elucidating much of the science behind the potential products and will continue to do so. As the industry develops, academic laboratories will need to put more emphasis in to down stream process development research. This will complement their fundamental work on protein expression and will provide the basic knowledge to fuel the industry. However, the marketing and delivery of commercial products will necessarily fall to industry. As with any new industry, there have been hurdles to overcome, both technical and regulatory. However, the experience to date has taught us much and the industry is now poised for rapid growth and profitability.

References

Azzoni AR. Kusnadi AR, Miranda EA, Nikolov Zl (2002). Recombaint aprotinin produced in transgenic corn seed: extraction and purification studies. Biotechnol Bioeng 80: 268-276.

Carter JE III, Langridge WHR (2002). Plant-based vaccines for protection against infectious and autoimmune diseases. Crit Rev Plant Sci 21: 93-109.

Daniell H, Streatfield SJ, wycoff K (2001). Medical molecular farming: production of antibiodies, biopharmaceuticals and edible vaccines in plants. Trends Plant Sci 6: 219-226.

Evangelistra RL, Kusnadi AR, Howard JA, Nikolov ZL (1998). Process and economic evaluation of the extraction and purification of recombinant β-glucuronidase from transgenic corn. Biotechnol Prog 14: 607-614.

Hiatt A, Cafferkey R, Bowdish K (1989) production of antibodies in transgenic plants. Nature 342: 76-78.

Hood EE, Woodard SL, Horn ME (2002). Monoclonal antibody manufacturing in transgenic plants-myths and realities. Curr Opin Biotechnol 13:630-635.

Nandi S, Suzuki YA, Huang J, Yalda D, Pham P., Wu L, Bartely G. Huang N, Lonnerdal B, (2002). Expression of human lactoferrin in transgenic rice grains for the application in infant formula. Plant Sci 163: 713-722.

Stoger E, Sack M, Fischer R, Christou p (2002b) Plantibodies: applications, adavantages and bottle necks. Curr Opin Biotechnol 13: 161-166.

Streatfield SJ, Howard JA (2003). Plant based vaccines. Int J Parasitol 33: 479-493

Zhong G-Y, Paterson D, Delaney DE, Bailey M. Witcher DR, Register JC III, Bond D Li C-P, Marshall L, Kulisek E, Ritland D, Meyer T, Hood EE, Howard JA (1999). Commercial production of aprotinin in transgenic maize seeds. Mol Breed 5: 345-356.

Multilocus Sequence Typing of bacteria

A.Kumar

Background

Multilocus sequence typing (MLST) is a nucleotide sequence based approach for the unambiguous characterization of isolates of bacteria and other organisms via the internet. MLST is a technique in molecular biology for the typing of multiple loci. The procedure characterizes isolates of bacterial species using the DNA sequences of internal fragments of multiple (usually seven or eight) housekeeping genes. Approximately 450-500 bp internal fragments of each gene are used, as these can be accurately sequenced on both strands using an automated DNA sequencer. For each housekeeping gene, the different sequences present within a bacterial species are assigned as distinct alleles and, for each isolate, the alleles at each of the loci define the allelic profile or sequence type (ST). MLST directly measures the DNA sequence variations in a set of housekeeping genes and characterizes strains by their unique allelic profiles. The principle of MLST is simple: the technique involves PCR amplification followed by DNA sequencing. Nucleotide differences between strains can be checked at a variable number of genes (generally seven) depending on the degree of discrimination desired. The first MLST scheme to be developed was for *Neisseria meningitidis*, the causative agent of meningococcal meningitis and septicaemia.

The aim of MLST is to provide a portable, accurate, and highly discriminating typing system that can be used for most bacteria and some other organisms. It is envisaged that this approach will be particularly helpful for the typing of bacterial pathogens. MLST is highly unambiguous and portable. Materials required for ST determination can be exchanged between laboratories. Primer sequences and protocols can be accessed electronically. It is reproducible and scalable. MLST is automated, combines advances in high throughput sequencing and bioinformatics with established population genetics techniques. MLST data can be used to investigate evolutionary relationships among bacteria. MLST provides good discriminatory power to differentiate isolates.

Objective

To tag gram negative bacteria with gene coding for green fluorescent protein

Procedure

Test bacterium

DNA isolated

PCR amplification of housekeeping genes (~800-1000bp) using gene specific primers

Sequencing of internal regions of the genes (~400-500bp) using another set of specific primers

Sequence analysis using sequence analyzer (e.g. Vector NTi, CLC sequence viewer, BioEdit etc)

Blast analysis of internal sequence in MLST database

Identification of allele number for each of the genes. If numbers are not found in the data base, then, the sequence could be new allele

String of allele numbers representing all the genes constitute sequence type (ST)

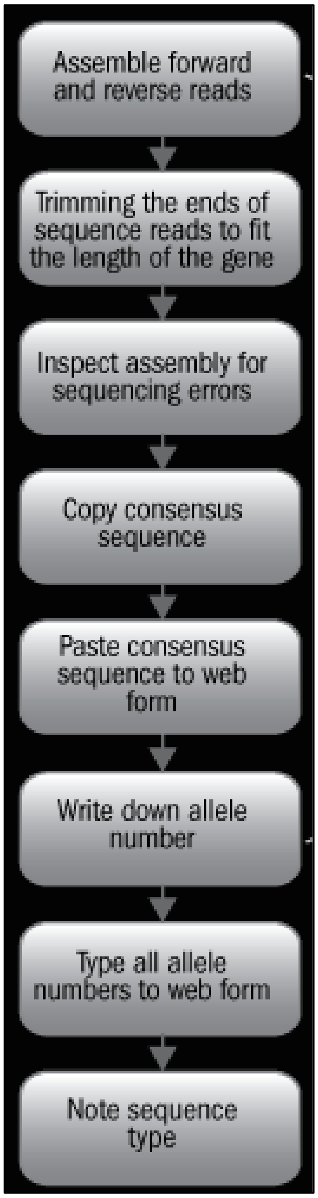
The combination of allele number, ST, is statistically analysed using a novel algorithm called eBurst to identify the position of the strain the global collection of bacterial strains. The following diagram describes various steps involved in the MLST analysis

MLST analysis workflow

MLST analysis of *Pseudomonas aeruginosa*

DNA isolation

1. Cell concentration of the bacterial suspension is adjusted to 1.0
2. Spin down at 14000 rpm at room temperature for 2 min.
3. The supernatant is discarded and pellet was washed three times with sterile distilled water.
4. To the pellet 550µl of TE buffer+lysozyme is added, mixed well and incubated for 30 min at 37C.
5. After incubation 76µl of 10% SDS+Proteinase K is added.
6. The contents are mixed by flipping the tube and incubated for 15 min at 65C.
7. After incubation 100µl of 5M NaCl is added and mixed the contents by flipping the tube.
8. Then 80µl of CTAB/NaCl is added, mixed and incubated for 10 min at 65C.
9. After incubation 660µl of Chloroform+isoamyl alcohol is added.
10. The contents are mixed by flipping the tube about 30sec.
11. Then centrifuge for 5 min at 14000 rpm at room temperature.
12. After centrifugation the aqueous fraction is carefully transferred to a new 1.5ml tube without touching the white middle layer (interface). This step is repeated twice.
13. Equal volume of isopropanol is added and inverted to mix.
14. Then centrifuge for 15 min at 14000 rpm at room temperature.
15. The supernatant is gently drained and mixed with 0.5ml of 70% ice cold ethanol.
16. Centrifuge for 15min at 14000rpm at room temperature.
17. After this the supernatant is carefully removed and evaporated the remaining ethanol in the laminar flow for about one hour.
18. 25µl of 10:1 TE is added to each tube to dissolve the DNA and the tubes are kept at4C for overnight.
19. The DNA from the two tubes is pooled in one tube, so the total is 50µl and added RNase to remove the contaminating RNA at a concentration of200µg/ml
20. The tubes are incubated for 30min at 30C. Then stored the DNA at-20C.


Assemble forward and reverse reads
Trimming the ends of sequence reads to fit the length of the gene
Inspect assembly for sequencing errors
Copy consensus sequence
Paste consensus sequence to web form
Write down allele number
Type all allele numbers to web form
Note sequence type

Quantification of DNA

1. 5µl of stock DNA is diluted 10 times by adding 45µl of MQ water.
2. The quantity of DNA is measured using a Biophotometer
3. Quality assessed by gel electrophoresis
4. DNA concentration is adjusted to 200ng per ul of water
5. Proceed with PCR amplification of genes

Choice of Genes

The *Pseudomonas aeruginosa* MLST scheme uses internal fragments of the following seven house-keeping genes:

1. *acsA* (Acetyl coenzyme A synthetase)
2. *aroE* (Shikimate dehydrogenase)
3. *guaA* (GMP synthase)
4. *mutL* (DNA mismatch repair protein)
5. *nuoD* (NADH dehydrogenase I chain C, D)
6. *ppsA* (Phosphoenolpyruvate synthase)
7. *trpE* (Anthralite synthetase component I)

PCR Amplification

Reaction conditions for all the primers are as follows: initial denaturation at 96°C for 1 min; 30 cycles of denaturation at 96°C for 1 min, primer annealing at 55°C for 1 min, extension at 72°C for 1 min; followed by a final extension step of 72°C for 10 min. Each 50 µl amplification reaction mixture comprised 2.0 µl chromosomal DNA (5-20 ng/µl), 2.0 µl forward primer (10 pmol/µl), 2.0 µl reverse primer (10 pmol/µl), 5.0 µl 10x PCR buffer (contains 15 mM $MgCl_2$), 1.0 µl dNTP solution (10 mM each dNTP), 0.25 µl Taq polymerase (5units/µl) and 37.75 µl PCR-grade water.

Primers used for amplification of gene sequences and their sequencing

Primer pairs are used for PCR amplification of internal fragments of these genes	Primer pairs are used for sequencing
	Sequencing
PCR amplification	
acsA-F ACCTGGTGTACGCCTCGCTGAC	acsA-F GCCACACCTACATCGTCTAT
acsA-R GACATAGATGCCCTGCCCCTTGAT	acsA-R AGGTTGCCGAGGTTGTCCAC
aroE-F TGGGGCTATGACTGGAAACC	aroE-F ATGTCACCGTGCCGTTCAAG
aroE-R TAACCCGGTTTTGTGATTCCTACA	aroE-R TGAAGGCAGTCGGTTCCTTG
guaA-F CGGCCTCGACGTGTGGATGA	guaA-F AGGTCGGTTCCTCCAAGGTC
guaA-R GAACGCCTGGCTGGTCTTGTGGTA	guaA-R GACGTTGTGGTGCGACTTGA
mutL-F CCAGATCGCCGCCGGTGAGGTG	mutL-F AGAAGACCGAGTTCGACCAT
mutL-R CAGGGTGCCATAGAGGAAGTC	mutL-R GGTGCCATAGAGGAAGTCAT
nuoD-F ACCGCCACCCGTACTG	nuoD-F ACGGCGAGAACGAGGACTAC
nuoD-R TCTCGCCCATCTTGACCA	nuoD-R TGGCGGTCGGTGAAGGTGAA
ppsA-F GGTCGCTCGGTCAAGGTAGTGG	ppsA-F GGTGACGACGGCAAGCTGTA
ppsA-R GGGTTCTCTTCTTCCGGCTCGTAG	ppsA-R GTATCGCCTTCGGCACAGGA
trpE-F GCGGCCCAGGGTCGTGAG	trpE-F TTCAACTTCGGCGACTTCCA
trpE-R CCCGGCGCTTGTTGATGGTT	trpE-R GGTGTCCATGTTGCCGTTCC

The amplification product was then purified using gel purification kit following the manufacturers' protocol before being used in a sequencing reaction. Sequencing is carried out on each DNA strand with BigDye Terminator Ready Reaction Mix under the following conditions: initial denaturation at 96°C for 1 min, 30 cycles of 10 sec at 96°C, 5 sec at 51°C and 2 min at 60°C. Unincorporated dye terminators were removed by precipitation with 95% alcohol.

Sequences are carefully analysed and sequence type assigned for each of the strain. The strain relation with the existing collection of strain can be determined by eBurst programme (http://eburst.mlst.net)

Selected Reading

Maiden MC, Bygraves JA, Feil E, et al. (1998). "Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms". Proc. Natl. Acad. Sci. U.S.A. 95 (6): 3140–5.

Identification of gene (s) governing antibiotic biosynthesis in bacteria through genome wide random mutagenesis

A.Kumar

Background

DNA transposition is an important biological phenomenon that mediates genome rearrangements, inheritance of antibiotic resistance determinants, and integration of retroviral DNA. Transposition has also become a powerful tool in genetic analysis, with applications in creating insertional knockout mutations, generating gene–operon fusions to reporter functions, providing physical or genetic landmarks for the cloning of adjacent DNAs, and locating primer binding sites for DNA sequence analysis. A number of in vitro transposition systems (for e.g. Tn*5*, Tn*7*, Mu, *Himar1,* and Ty*1*) have been described. Among them transposon Tn5 is widely used in bacterial genetic experiments. The Tn*5* transposition process involves the following steps: (1) binding of transposase monomers to the 19 bp end sequences; (2) oligomerization of the end-bound transposase monomers, forming a transposition synaptic complex; (3) blunt end cleavage of the transposition synaptic complex from adjoining DNA, resulting in formation of a released transposition complex or Transposome; (4) binding to target DNA; and (5) strand transfer of the transposon 3' ends into a staggered 9 bp target sequence.

Objective

To determine the role of specific antibiotics for metabolites in interaction between antagonistic *Pseudomonas* spp and plant pathogens. Mutants are generated that are defective in the production of antibiotics or other antifungal metabolites. One procedure to generate mutants is random mutagenesis. In this experiment, you will perfume a mutagenesis on *Pseudomonas fluorescens* strain SBW25. Strain SBW 25 produces biosurfactant which kill zoospores of Oomycetes pathogens, including *Phytophthora capsici*. The objective of this experiment is to generate surfactant deficient mutants of strain SBW25

Procedure

Make sure that all the equipment you use is properly sterilized

Grow *E.coli* strains overnight at 37C and strain SBW25 at 25C

E. coli strain consist of a donor strain Eco-TnModKmOlacZ and a helper strain Eco-pRK2013. The donor and helper strains harbour resistance to Kanamycin (25ug per mL). The donor strain contains a plasmid that is necessary for transfer of the donor plasmid into *Pseudomonas*. Strain SBW25 is the recipient strain and harbours resistance to rifamycin (50ug per mL)

Transfer 1.5 mL of the O/N cultures to eppendorf tubes and wash the strains thoroughly 3 times with sterile distilled water; resuspended the cell pellets every time in 0.5mL sterile distilled water.

Prior to the triparental mating, incubate the washed bacterial cells of *Pseudomonas* strain SBW25 at 42C for 5 min. This wills hut down molecular mechanisms in the recipient strain that prevent entry of foreign DNA into cell

Directly after incubation of the recipient strain SBW25 at 42C, mix the donor, helper and recipient in an eppendorf tube in a 1:1:1 ration (V/V): 200:200:200ul; this procedure is called triparental mating. As controls, include recipient strain SBW25 alone and a mix of donor and helper strains. So in total you will have three tubes

1. A mix of 3 strains (200:200:200ul)
2. A mix of the two *E. coli* (200:200ul)
3. The recipient strain SBW25 alone (200ul)

Spin down the mixtures of strains (5 min at 9000rpm), remove supernatant and resuspend pellet in 10-15ul sterile distilled water as a drop to a dry LB agar plate and incubate O/N at 30C

After incubation, resuspend the complete colony in 500ul of sterile distilled water and plate 50ul on to PSA plates with rifamycin (50ug per mL) and Kanamycin (100ug per mL). Rifamycin select for strain SBW25 and Kanamycin select for transformants harbouring the transposable elements (Kanamycin resistance)

Centrifuge the reminder of the bacterial mixtures (5 min, 9000rpm) and resuspend the bacterial pellet in 50ul of sterile distilled water and plate on PSA+Rifamycin+Kanamycin. So in total you will have 6 plates

Incubate for 48-72h at 25C

After incubation, count the number of transformants from the triparental mating (mix of SBW25 and the two *E. coli* strains); plate with the control (mix of two *E. coli*; and SBW25 alone) should have no colonies

Select mutants of strain SBW25 that have lost the ability to produce biosurfactants by a drop collapse assay. For the drop collapse assay, transfer colonies to PSA plates and incubate for 48h at 25C

After incubation, make cell suspension of the mutants in microtiter plates with 50ul of sterile distilled water per well. Spot 5ul on parafilm to visualize the shape of the droplet. Do not forget to include water and drop of bacterial cell suspension of wild-type SBW25 as controls

The transformants that have lost the ability to collapse a droplet of water have the transposon integrated into a gene that is involved in the synthesis or regulation of biosurfactant production. The transformants will be tested for activity against zoospores of *Phytophthora capsici*. Spot 5ul droplets of zoospore suspension of P. capsici on a glass slide and watch the zoospores under microscope at a magnification of 10X or 25X. Then gently add 5ul of droplets of bacterial suspension stay focused on several zoospores and record the response (motility etc). Do not forget to include water and wild-type SBW25 as controls

Selected reading

Igor Y. Goryshin, Jerry Jendrisak, Les M. Hoffman, Ronald Meis & William S. Reznikoff Insertional transposon mutagenesis by electroporation of released Tn5 transposition complexes Nature Biotechnology 18, 97 - 100 (2000)

Insertion of the *gfp* marker gene into Gram-negative bacteria

A.Kumar

Background

Genes coding for green fluorescent protein (gfp) are very useful reporter marker in biology. *gfp* gene is inserted into the genome of bacteria by several vectors aided by variety of transformation methods. The bacterial transposon Tn*7* inserts at a high frequency into a specific intergenic site *att*Tn*7* on the chromosome in a number of Gram-negative bacteria (Craig, 1989). Due to the specific insertion, the Tn*7* transposon is a very useful delivery system for insertion of cloned DNA. Potential applications of Tn*7*-based systems include stable insertion of marker genes to enable specific detection of the tagged bacteria in natural environments and insertion of transcriptional fusions in a single copy on the chromosome for gene expression studies. The gfp gene is inserted in the genome of the test bacteria by one of the several methods such as electroporation, hear shock transformation, or classical bi or tri or four parental mating.

Objective

To tag gram negative bacteria with gene coding for green fluorescent protein

Procedure

The following protocol explains the methods for genetic transformation of bacteria using gfp gene. The method is based on conventional mating of bacteria.

Day 1

Select spontaneous mutants of test bacteria (Recipient- R) that is resistant to rifampicin (Rif) at 50-100ug per mL

Inoculate test bacteria in to 5ml of CPG broth with Rif 100 for36 hours with shaking at 30C

Inoculate pBKmini Tn7gfp2Gm10 (Donor- D) in to 5ml of LB broth with Gm10 at 37C with shaking for 12 hours

Inoculate pUX BF13Amp100 (Helper 1- H1) in to 5ml of LB broth with Amp100 at 37C with shaking for 12 hours

Inoculate pRK2013 (Helper 2-H2) in to 5ml of LB broth with Kan 50 at 37C with shaking for 12 hours

Day 2

After incubation 500µl of test bacteria is inoculated in 5ml of CPG broth without antibiotic and incubated at 37C for 3 hours without shaking

After 3 hours 500µl of donor, helper 1 and helper 2 are separately inoculated in to 5 ml of LB broth without antibiotics, incubated at 30C for 4 hours with shaking

After 4 h 1.5ml of the each cultures is taken in 2 tubes centrifuged at 8000rpm for2min, pellet washed three times with 0.9% NaCl.

The recipient cells are heat treated at 42C for 5 min.

The cultures are mixed in 2:1:1:1: ratio (R×D×H1×H2) and 1:1:1:1 ratio(R×D×H1×H2)

Two controls are kept 1. D×H1×H2 (1:1:1)
2. R

The mixed cultures are centrifuged at 8000rpm for2min

Pellet is resusended in to 10-15µl of NaCl (0.9%)

15µl of the culture is spot inoculated in to the centre of cellulose nitrate membrane filter placed in the middle of LB agar plates.

The plates are incubated at 30C for overnight.

Day 3

The colonies are scraped from the membrane, centrifuged at 8000 rpm for 2min, pellet resuspended in 1.0ml of NaCl.

From this 50µl is spread plated on CPG+Rif100+Gm10 plates.

Remaining 950µl is centrifuged at 8000 rpm for 2min, 900µl of supernatant is discarded and remaining50µl is mixed and spread plated on CPG+Rif100+Gm10 plates.

The plates are incubated at 30C for 36-48 hours.

Day 5-6

Observe the plates for colonies and transfer colonies to new plates with antibiotics. The transformed colonies are screened for gfp insertion by various phenotypic and genotypic methods

Monitoring of GFP-Tagged Bacterial Cells

The GFP phenotype is not natural in the autochthonous microbial populations and does not require any exogenous substrate, complex medium, or expensive equipment. This phenotype allows gfp-tagged cells to be monitored by different means (epifluorescence microscopy, flow cytometry, and spectrofluorometry)

PCR confirmation of *gfp*-tagged *Pseudomonas* strains

Though antibiotic selection confirms the transformants for gfp insertion, additional confirmation for presence of gfp insert in the genome can be done by several methods. PCR analyses can be carried out to verify Tn*7* insertion in *P. fluorescens*. PCR analyses is carried out either on chromosomal DNA or directly on cell lysates obtained by boiling cells for 7 min in sterile water. The PCR solution contained PCR reaction buffer, 0.4 mM dNTPs, 0.2 mM of each primer, 0.4 U DNA polymerase and approximately 20 ng template DNA or cell lysate. After an initial denaturation for 3 min at 94°C, a total of 30 cycles are run using the following program: annealing at 52°C for 1 min, extension at 72°C for 1 min and denaturation at 94°C for 1 min, followed by 1 cycle of 1-min annealing at 52°C and 5-min extension at 72°C. The primers recognize the *cat* gene encoding chloramphenicol acetyltransferase present on the vector (primer sequence: 5′TTT TTC GTC TCA GCC AAT CCC3′) and the chromosomal sequence obtained from *Pseudomonas* (primer sequence 5′GTA ATC TGG CGA AGT CGG TG3′), respectively.

Selected reading

Craig, 1989 N.L. Craig , Transposon Tn*7*. In: D.E. Berg and M.M. Howe, Editors, *Mobile DNA*, American Society for Microbiology, Washington, DC (1989), pp. 211–225.

Genes and their Expression

Johnson George K.

The concept of the gene is and has always been a continuously evolving one. The classical view prevailed into the 1930s, and conceived the gene as an indivisible unit of genetic transmission, recombination, mutation, and function. The discovery of intragenic recombination in the early 1940s and the establishment of DNA as the physical basis of inheritance led to the neoclassical concept of the gene, which prevailed until the 1970s. In this view the gene (or cistron, as it was called then) was subdivided into its constituent parts, mutons and recons, identified as nucleotides. Each cistron was believed to be responsible for the synthesis of a single mRNA and hence for one polypeptide. This colinearity hypothesis prevailed from 1955 to the 1970s. Starting from the early 1970s, DNA technologies have led to the modern period of gene conceptualization, wherein none of the classical or neoclassical criteria are sufficient to define a gene. Modern discoveries include those of repeated genes, split genes and alternative splicing, assembled genes, overlapping genes, transposable genes, complex promoters, multiple polyadenylation sites, polyprotein genes, editing of the primary transcript, and nested genes. In recent years, many investigators have also focused on epigenetic inheritance, in which information is passed from parent to offspring independent of the DNA sequence. Today’s assault on the gene concept is also more far reaching, fuelled largely by studies that show the previously unimagined scope of RNA.. Some recent reports suggests that transcription can start at a DNA sequence associated with one protein and run straight through into the gene for a completely different protein, producing a fused transcript.

The basics

Most genes specify one or more protein molecules, the 'expression' of these genes involving an RNA intermediate, called messenger or mRNA, which is transported from the nucleus to the cytoplasm where it directs synthesis of the protein coded by the gene. Other genes do not specify proteins, the end-products of their expression being non-coding RNA, which plays various roles in the cell.

A gene can be defined as the functional and physical unit of heredity passed from parent to offspring. Intriguingly, a surprisingly small proportion of the human genome is made up of genes (only up to 1.5 % are covered by protein-coding genes). Nevertheless, it is the set of genes and their interactions that define all living beings. Eucaryotic genes map to individual sequence ranges on genomic DNA. Their structure is given by the following elements (listed from 5’ to 3’):

1.) **Promoter**: The part of a gene that contains the information to turn the gene on or off. The process of transcription is initiated at the promoter. The extent of a promoter is often difficult to determine. Proximal and distal promoter elements play a role in controlling the expression level of a gene.

2.) **Exons**: Regions downstream of the promoter of a gene that are transcribed and exported from the nucleus as part of the messenger RNA (mRNA). mRNA contains all information (“the message”) for the formation of the final protein product of a gene.

3.) **Introns** Regions downstream of the promoter of a gene that are also transcribed into RNA but are excised (spliced) from the maturating RNA. Thus, these regions are absent from the messenger RNA.

The biological information in the genes are thus divided into a series of exons separated by non-coding introns. During gene expression, the initial RNA that is synthesized is a copy of the entire gene, including the introns as well as the exons. The process called splicing removes the introns from this pre-mRNA and joins the exons together to make the mRNA which eventually directs protein synthesis. At one time it was thought that splicing was a straightforward process, each exon being joined to its neighbor to produce a single mRNA from each discontinuous gene. Now it is known that many pre-mRNAs undergo alternative or differential splicing, giving rise to a series of mRNAs containing different combinations of exons and each specifying a different protein. As well as the gene itself, the pre-mRNA transcribed from a gene also contains sequences from the regions preceding the first exon and following the last exon. These are called the $5'$-untranslated region ($5'$-UTR) and $3'$-untranslated region ($3'$-UTR), respectively. Plant intron splice sites and branch point sequences, small nuclear RNA (snRNA) sequences, and many small nuclear ribonucleoproteins (snRNPs) and spliceosomal proteins are very similar to yeast and vertebrate intron signals and splicing components. UA-rich sequences that are characteristic of plant introns are believed to act early in intron recognition and definition and direct the association of splicing factors to assemble presplicing complexes. One of RNA binding proteins, UBP-1, that interact with U-rich sequences in plant pre-mRNA transcripts was found to increase the efficiency of splicing of “weak” introns and also increase transcript levels.

Alternative splicing does not appear to be as widespread in plant (Arabidopsis) genomes as it is in the human genome (35% of genes may be alternatively spliced), it is important to define the functions of proteins produced by alternative splicing and to understand the mechanisms by which alternative splicing is regulated in plants. One of the best studied alternatively spliced plant genes is the *N* gene of tobacco that confers resistance to *Tobacco mosaic virus*. Analysis of the splicing behavior and the associated resistance phenotype of a series of *N* gene constructs has identified regions essential for both alternative splicing and resistance . Infection by *Tobacco mosaic virus* induces changes in the splicing pattern of *N* transcripts, and the role of U-rich RNA binding proteins, such as UBP1 is yet to be found out.

Gene expression comprises the following steps.

- **Accessing the genome.** This involves various processes that influence chromatin structure and nucleosome positioning in the parts of the genome that contain active genes, ensuring that these genes are accessible and are not buried deep within highly packaged parts of the chromosomes.
- **Assembly of the transcription initiation complex,** which comprises the set of proteins that work together to copy genes into RNA. Assembly of initiation complexes is a highly targeted process because these complexes must be constructed at precise positions in the genome, adjacent to active genes, and nowhere else.
- **Synthesis of RNA,** during which the gene is transcribed into an RNA copy.
- **Processing of RNA** involves a series of alterations that are made to the sequence of the RNA molecule and to its chemical structure, and which must occur before the RNA molecules can be translated into protein or, in the case of non-coding RNA, before they can carry out their other functions in the cell.
- **RNA degradation** is the controlled turnover of RNA molecules. Degradation is not simply a means of getting rid of unwanted RNAs: it plays an active role in determining the make up of the transcriptome and hence is an integral step in genome expression.
- **Assembly of the translation initiation complex** occurs near the $5'$ termini of coding RNA molecules, and is a prerequisite for translation of these molecules.
- **Protein synthesis** is the synthesis of a protein by translation of an RNA molecule.
- **Protein folding and protein processing** may occur together. Folding results in the protein taking up its correct three-dimensional configuration. Processing involves modification of the protein by addition of chemical groups and, for some proteins, removal of one or more segments of the protein.

- **Protein degradation** has an important influence on the composition of the proteome and, like RNA degradation, is an integral component of genome expression.

The initial product of genome expression is the transcriptome, a collection of RNA molecules derived from those protein-coding genes whose biological information is required by the cell at a particular time. These RNA molecules direct synthesis of the final product of genome expression, the proteome, the cell's repertoire of proteins, which specifies the nature of the biochemical reactions that the cell is able to carry out.

It is important to appreciate that not all RNA constitutes the transcriptome. Most of the cellular RNA does not fall into this category because it is non-coding. The coding RNA comprises the transcriptome and is made up of just one class of molecule: Messenger RNAs (**mRNAs**), which are transcripts of protein-coding genes and hence are translated into protein in the latter stages of genome expression. Messenger RNAs rarely make up more than 4% of the total RNA and are short-lived, being degraded soon after synthesis.

Non-coding RNAs are more diverse than the coding RNA and comprises transcripts with a number of different functions, all of which are performed by the RNA molecules themselves. In both prokaryotes and eukaryotes the two main types of non-coding RNA are: a.) Ribosomal RNAs (**rRNAs**), which are the most abundant RNAs in the cell. These molecules are components of ribosomes, the structures on which protein synthesis takes place. b.) Transfer RNAs (**tRNAs**) are small molecules that are also involved in protein synthesis, carrying amino acids to the ribosome and ensuring that these are linked together in the order specified by the nucleotide sequence of the mRNA that is being translated .

The other non-coding RNA types are more limited in their distribution . Eukaryotes, for example, have a variety of short non-coding RNAs that are usually divided into three categories, the names indicating their primary locations in the cell: a.)Small nuclear RNA (snRNA; also called U-RNA because these molecules are rich in uridine nucleotides), which is involved in mRNA processing. b.) Small nucleolar RNA (snoRNA), which plays a central role in the processing of rRNA molecules . c.) Small cytoplasmic RNA (scRNA), a diverse group including molecules with a range of functions, some understood and others still mysterious. Eukaryotes also possess RNAs like microRNAs (miRNAs) involved in gene regulation.

The enzymes responsible for transcription of DNA into RNA are called DNA-dependent RNA polymerases. Most research attention has been directed at RNA polymerase II, as this is the one that transcribes genes that code for proteins. RNA polymerase is the central component of the transcription initiation complex. Every time a gene is transcribed, a new complex has to be assembled immediately upstream of the gene. The initiation complexes are constructed at the appropriate positions, and not at random points within the genome, because their target sites are marked by specific nucleotide sequences called promoters, which are only found upstream of genes.

It was recognized in the 1950s that a triplet genetic code - one in which each codeword or codon comprises three nucleotides - is required to account for all 20 amino acids found in proteins. Only tryptophan and methionine have just a single codon each: all others are coded by two, three, four or six codons. This feature of the code is called degeneracy The initiation codon is usually 5'-AUG-3', which also specifies methionine (so most newly synthesized polypeptides start with methionine), although with a few mRNAs other codons such as 5'-GUG-3' and 5'-UUG-3' are used. The three termination codons are 5'-UAG-3', 5'-UAA-3' and 5'-UGA-3'; these are sometimes called amber, opal and ochre, respectively. Genes that code for proteins comprise open reading frames (ORFs) consisting of a series of codons that specify the amino acid sequence of the protein that the gene codes for . The ORF begins with an initiation codon - usually (but not always) ATG - and ends with a termination codon: TAA, TAG or TGA.

Codon bias is usually observed. 'Codon bias' refers to the fact that not all codons are used equally frequently in the genes of a particular organism. For example, leucine is specified by six codons in the genetic code TTA, TTG, CTT, CTC, CTA and CTG; , but in human genes leucine is most frequently coded by CTG and is only rarely specified by TTA or CTA. Similarly, of the four valine codons, human genes use GTG four times more frequently than GTA.

Core Promoters

The term 'promoter' is used to describe all the sequences that are important in initiation of transcription of a gene. For some genes these sequences can be numerous and diverse in their functions, including not only the core promoter, sometimes called the basal promoter, which is the site at which the initiation complex is assembled, but also one or more upstream promoter elements which, as their name implies, lie upstream of the core promoter. Assembly of the initiation complex on the core promoter can usually occur in the absence of the upstream elements, but only in an inefficient way. This indicates that the proteins that bind to the upstream elements include at least some that are activators of transcription, and which therefore 'promote' gene expression. Each of the three types of eukaryotic RNA polymerase recognizes a different type of promoter sequence; indeed, it is the difference between the promoters that defines which genes are transcribed by which polymerases.

Typically, the core promoter encompasses the site of transcription initiation and extends either upstream or downstream for an additional~35 nt. Thus, in many instances, the core promoter will comprise only about 40 nt. There are several sequence motifs—which include the TATA box, initiator (Inr), TFIIB recognition element (BRE), and downstream core promoter element (DPE)—that are commonly found in core promoters . These motifs each have specific functions that relate to the transcription process. It is important to note that each of these core promoter elements is found in some but not all core promoters. Genes regulated in parallel with one another would contain common control elements.

In addition to the core promoter, other *cis*-acting DNA sequences that regulate RNA polymerase II transcription include the proximal promoter, enhancers and silencers. These elements contain recognition sites for a variety of sequence-specific DNA-binding factors that are involved in transcriptional regulation. The proximal promoter is the region in the immediate vicinity of the transcription start site (roughly from -250 to +250 nt). Enhancers and silencers can be located many kbp from the transcription start site and act either to activate or to repress transcription. Boundary/insulator elements appear to prevent the spreading of the activating effects of enhancers or the repressive effects of silencers or heterochromatin.

Control of Gene expression

Plants also have unique needs and strategies for responding to changes in their environment. When light strikes an etiolated leaf, numerous genes encoding chloroplastic, mitochondrial, peroxisomal, and cytosolic proteins are activated. Similarly, a number of biotic and abiotic stresses cause a battery of genes to be activated as part of the plant-defense/stress response.

Expression of a gene can be regulated at multiple steps including transcription, posttranscriptional processing, nuclear export and localization, stability, translation of mature mRNA molecules and post-translational events such as protein stability and modification. The DNA strands are not just laid out, they are wrapped around proteins, and that complex is coiled up inside the nucleus. It’s a very complex machinery that interacts with the chromatin and activates genes at the right time. Chromatin structure can influence genome expression in two ways. First, the degree of chromatin packaging displayed by a segment of a chromosome determines whether or not genes within that segment are expressed. Second, if a gene is accessible, then its transcription is influenced by the precise nature and positioning of the nucleosomes in the region where the transcription initiation complex will be assembled. The formation of DNA loops by the binding of proteins and protein complexes at distal DNA sites were also found to play a central role in many cellular processes, such as transcription. Important thermodynamic concepts underlie the assembly of macromolecular complexes on looped DNA.

DNA methylation can repress gene activity. In eukaryotes, cytosine bases in chromosomal DNA molecules are sometimes changed to 5-methylcytosine by the addition of methyl groups by enzymes called DNA methyltransferases . Cytosine methylation is relatively rare in lower eukaryotes but in vertebrates up to 10% of the total number of cytosines in a genome are methylated, and in plants the figure can be as high as 30%. The methylation pattern is not random, instead being limited to the cytosine in some copies of the sequences $5'$–CG–$3'$ and, in plants, $5'$–CNG–$3'$. Housekeeping genes –those that are expressed in all tissues –have unmethylated CpG islands, whereas tissue-specific genes are unmethylated only in those tissues in which the adjacent gene is expressed.

Transcriptional Regulation

In eukaryotic organisms, regulation of gene expression proceeds through mechanisms that are fundamentally different from those in prokaryotes, which explains both the large number and diversity of proteins that are involved in the process, as well as how it can be tightly regulated to facilitate the diversification in expression patterns that is required for biological complexity. Many of the biological processes in a plant are regulated at the level of transcription.

Briefly, the proteins involved in transcription in eukaryotes can be classified into four different functional groups: (1) the basic transcription apparatus and intrinsic associated factors (also known as general transcription factors, or GTFs); (2) large multi-subunit coactivators and other cofactors; (3) sequence-specific DNA-binding transcription factors; and (4) chromatin-related proteins. In contrast to the components of the basal transcription machinery, which in general are highly conserved, coregulators and transcription factors have diverged largely among eukaryotes.

Chromatin-related proteins consists of a group that covalently modify histones (such as histone acetylases and deacetylases), and remodeling complexes that hydrolize ATP for reorganizing chromatin structure such as the SWI/SNF and ISWI complexes. Histone acetylation is generally a characteristic of transcribed chromatin, whereas deacetylation is associated with repression. Accordingly, histone acetyltransferase activities are found in coactivators, and deacetylase activities in corepressors. Chromatin proteins usually form part of multi-subunit complexes.

In eukaryotic organisms, there are three different RNA polymerases, which are responsible for the synthesis of rRNA (Pol I), mRNA (Pol II), and tRNA, 5S rRNA, and other small RNA molecules (Pol III). The transcription of protein-encoding genes is carried out by Pol II exclusively. Pol II is a multi-subunit enzyme that requires accessory factors to recognize promoter sequences and accurately initiate transcription. These general transcription factors (GTFs) include TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH. GTFs carry out a variety of different functions, from positioning the polymerase on the promoter (TFIIB) to unwinding its DNA (TFIIH). TFIID is a multi-subunit complex that is generally responsible for promoter recognition. contains the TATA-box binding protein (TBP) and several TBP-associated factors (TAFs) . The TAF subunits of TFIID are critical for the responsiveness of the basic apparatus to transcriptional activators. However, individual TAFs are not essential for transcription of all genes in a genome. TAFs contribute to the specificity and variety of transcriptional responses: distinct TAFs can be targeted by different classes of activators and individual TAFs can function as promoter selectivity factors. Furthermore, some TAFs can form part of other multi-subunit regulatory complexes, in addition to TFIID, such as the histone acetylation SAGA complex; and whereas most of the TAFs are ubiquitously expressed, some are expressed in a tissue or cell-type specific manner, which can lead to the formation of different TAF-containing complexes. Coactivators, and cofactors involved in gene transcription consists of heterogeneous class of regulatory proteins and cofactors that interact with sequence-specific transcription factors and modulate their DNA binding or interaction with the core machinery, as well as large multisubunit coactivators.

Sequence-specific DNA-binding transcription factors are transcription factors that show sequence-specific DNA binding and are capable of activating and/or repressing transcription. They are responsible for the selectivity in gene regulation, and are often themselves expressed in a tissue, cell-type, temporal, or stimulus-dependent specific manner. Transcription factors are modular proteins, with distinct and functionally separable domains, such as DNA-binding and activation domains. Most known transcription factors can be grouped into families according to their DNA binding domain. Transcription factors can interact directly with different components of general machinery and with coactivators, affecting complex formation. They can also interact with chromatin remodeling complexes.

Study of putative functional domains by mutational and functional analysis has demonstrated that typical plant transcription factors consist of a DNA-binding region, an oligomerization site, a transcription regulation domain and a nuclear localization signal (NLS), although some lack either a transcription regulation domain or a specific DNAbinding region. Some of the general transcription factors are complex, e.g. TFIID is composed of TATA box binding protein (TBP) and a number of TBP-associated factors (TAFs).

WRKY proteins constitute a family of transcriptional regulators identified in higher plants. Nearly all WRKY factors tested show DNA binding selectivity towards a DNA element termed the W box (T/CTGACC/T). Such elements are often found in promoters of plant defense-response genes suggesting a likely role of these factors in mediating plant responses towards biotic stresses.

A major level at which gene expression is regulated is the initiation of transcription, and this is reflected in the percentage of the genome dedicated to transcription factors in plants and other eukaryotes. For example, an analysis of 1.9 Mb of Arabidopsis genomic sequence from chromosome 4 revealed that about 15% of the genes with predicted or known functions were involved in transcription, a percentage similar to what has been found in other eukaryotes. DNA binding transcription factors are thus the principal factors upon which the mechanisms for selectivity of gene activation are built, and the basic (although not the only) protein components of the combinatorial logic of transcription.

In many instances, the correct functioning of a gene requires the termination of the activation of its transcription to be as rapid or precise as its initial triggering. Termination of activation can be accomplished by several mechanisms, among them the targeted destruction of transcription factors after their interaction with the basal transcription machinery. Phosphorylation of a transcription factor molecule by kinases that form part of the Pol II holoenzyme (such as Srb10 or TFIIH) would mark it for ubiquitin mediated destruction, effectively preventing it from engaging into another Pol II initiation event, and freeing the promoter sequence to interact with another transcription factor molecule.

Post-transcriptional gene regulation

The initiation of mRNA translation is regulated at the global and message-specific levels. Differential translation of individual gene transcripts is determined by *cis*-acting mRNA sequences and translation factors. mRNAs compete for discriminatory initiation factors that couple the 5'-7mGpppNcap and the 3'-poly(A) tail of the RNA message. The resultant

circularization of the mRNA promotes the association of the pre-initiation complex that scans the 5’-leader for the initiation codon of the protein coding sequence.
The recruitment of a ribosome to the initiating AUG codon of an mRNA is typically the rate-limiting step in polypeptide synthesis in eukaryotes, although polypeptide elongation can also be regulated . Most mRNAs are translated via a mechanism that depends on the interaction of the 5’-7mGpppN-cap with the 3’-poly(A) tail. The initial step in translation is the assembly of a circular mRNÃprotein complex. In plants, the 5’-7mGpppN-cap of the transcript is recognized by eukaryotic initiation factor 4E (eIF4E) or eIFiso4E, which are bound to their respective partners, eIF4G and eIFiso4G . eIF4G and eIFiso4G are scaffold proteins that also recruit: first, an RNA helicase (i.e. eIF4A, a monomer that facilitates the ATP-dependent unwinding of RNA) ; second, an RNA-binding protein (i.e. eIF4B, a homodimer that binds poly(A), stabilizes the binding of ATP to eIF4A , and promotes the RNA-dependent ATP-hydrolysis of eIF4E/eIF4G/eIF4A but not of eIFiso4E/eIFiso4G/eIF4A) ; and third, poly(A)-binding protein (PABP; which binds poly[A] and enhances 5’-cap binding). The assembly of these proteins circularizes the mRNA and stimulates RNA helicase activity, causing a synergistic enhancement of translation . eIF4G and eIFiso4G also interact with the eIF3 complex (which includes 11 subunits, one of which is a plant-specific subunit) to position the 43S pre-initiation complex (i.e. the 40S subunit and the eIF2α̃-GTP̃-tRNA^{met} ternary complex) near the 5̤end of the mRNA. As secondary structure in the 5’-leader is relaxed, the 43S pre-initiation complex scans in a 5̣’to 3̣’direction until an AUG codon in the correct context is recognized and bound by tRNAmet, with the assistance of eIF1 and eIF1A. Completion of initiation requires the eIF5-assisted release of eIF2α-GDP and the eIF5B-assisted coupling of the 40S and 60S ribosomal subunits. It is thought that the 5’-cap/3’-tail connection may facilitate ribosome re-initiation (i.e. recycling) as well as the primary initiation event.

The post-transciptional gene regulation also relies on specific RNA–protein interactions that either result in the targeted degradation of the mRNA. Among eukaryotes, most progress in understanding mRNA degradation has been made with yeast. At least four pathways have been identified. One of these involves a multiprotein complex called the exosome, which degrades transcripts in the 3′→5′ direction. Rather more is known about two other eukaryotic mRNA degradation processes. The first of these is deadenylation-dependent decapping , which is triggered by removal of the poly(A) tail, possibly by exonuclease cleavage or possibly by loss of the polyadenylate binding protein which stabilizes the tail. Poly(A) tail removal is followed by cleavage of the 5′ cap by the decapping enzyme Dcp1p. Decapping prevents the mRNA from being translated and so ends its functional life. The mRNA then undergoes rapid exonuclease digestion from its 5′ end.

RNA silencing is another remarkable type of gene regulation based on sequence-specific targeting and degradation of RNA. RNA silencing was first discovered in transgenic plants, where it was termed cosuppression or posttranscriptional gene silencing (PTGS). Sequence-specific RNA degradation processes related to PTGS have also been found in fungi, and a variety of animals from *Caenorhabditis elegans* to mice (RNA interference RNAs (siRNAs). . A key feature uniting the RNA silencing pathways in different organisms is the importance of double-stranded RNA (dsRNA) as a trigger or an intermediate. The dsRNA is cleaved into small interfering RNAs (siRNAs) (21 to 25 nucleotides) of both polarities, and these are thought to act as guides to direct the RNA degradation machinery to the target RNAs . The siRNAs are incorporated into a complex termed RNA-induced silencing complex (RISC), and the RNA-protein complex is thought to recognize and destruct the target mRNA.

Conclusion

Plants also have unique needs and strategies for responding to changes in their environment. When light strikes an etiolated leaf, numerous genes encoding chloroplastic, mitochondrial, peroxisomal, and cytosolic proteins are activated. Similarly, a number of biotic and abiotic stresses cause a battery of genes to be activated as part of the plant-defense/stress response. Continuing discoveries of new and surprising mechanisms of gene regulation suggest that our understanding of this complex and ubiquitous biological process remains incomplete. Emerging studies illustrate that many and perhaps all genes are regulated at multiple steps including transcription, posttranscriptional processing, nuclear export and localization, stability, and translation of mature mRNA molecules. The past few years have seen a rapid expansion in the identification and characterization of mRNA regulatory elements and their binding proteins. For the majority of these examples, the mechanism by which translational regulation is achieved is not well understood.

The analysis of transcriptional control in plants will continue to be an exciting field of research. The rapid progress being made on the isolation of important regulatory proteins, the development of in vitro transcription systems, and the use of powerful genetic screening approaches for additional mutants using promoter/reporter gene fusions will facilitate further studies of transcriptional control in plants, which should provide valuable insight into the mechanisms underlying various aspects of plant growth and development and lead to agricultural benefits.

It is also evident that translational control is closely associated with other mechanisms of gene regulation; there are close links between translation, mRNA stability and intracellular localization. Translation itself is regulated by these diverse mechanisms that act not only at the initiation step but also during elongation and termination and even after termination. Detailed analyses of a few examples show that almost every event in the initiation pathway, from binding of the cap complex to the joining of the 60S ribosomal subunit, is subject to regulation. Our growing understanding of post-transcriptional regulatory mechanisms including RNA silencing offer additional tools for bio-technologists seeking to increase or to control more precisely the expression of transgenes in plants. These efforts should also provide new paths to pursue in the quest to improve plants for humanity.

Selected References

Abler M L & Green P J 1996 Control of mRNA stability in higher pants. Plant Mol. Biol. 32:63-78.

Bonifer C 2000 Developmental regulation of eukaryotic gene loci: which cis-regulatory information is required?. Trends Genet. 16: 310-315.

Browning K S 1996 The plant translational apparatus. Plant Mol. Biol. 32: 107-144.

Eulgem, T, Rushton, P J, Robatzek S & Somssich I E 2000 The WRKY superfamily of plant transcription factors. Trends Plant Sci. 5: 199-206.

Flavio Mignone, Carmela Gissi, Sabino Liuni & Graziano Pesole 2002 Untranslated regions of mRNAs Genome Biology 3: 1–10

Hammond, S M, Caudy A A & Hannon, G J 2001 Posttranscriptional gene silencing by double-stranded RNA. Nat. Rev. Genet. 2: 110-119.

Hampsey M & Reinberg D 1999 RNA polymerase II as a control panel for multiple coactivator complexes. Curr. Opin. Genet. Dev. 9: 132-139.

Jennifer E, Butler F & James Kadonaga T 2002 The RNA polymerase II core promoter: a key component in the regulation of gene expression. Genes & Dev 16: 2583-2592.

Ji-Young Lee, Juliette Colinas, Jean Wang Y, Daniel Mace, Uwe Ohler & Philip Benfey N 2006 Transcriptional and posttranscriptional regulation of transcription factor expression in Arabidopsis roots. PNAS 103: 6055–6060.

Karam B Singh 1998 Transcriptional Regulation in Plants: The Importance of Combinatorial Control. Plant Physiol. 118: 1111–1120.

Kornberg R D 1999. Eukaryotic transcriptional control. Trends Cell Biol. 9: 46-49.

Liansen Liu, Michael J, White & Thomas H 1999 Transcription factors and their genes in higher plants: Functional domains, evolution and regulation. Eur. J. Biochem. 262: 247-257.

Ptashne M & Gann A 1997 Transcriptional activation by recruitment. Nature 386: 569–577.

Riechmann J L & Ratcliffe O J 2000. A genomic perspective on plant transcription factors. Curr. Opin. Plant. Biol. 3: 423-434.

Roeder R G 1998 Role of general and gene-specific cofactors in the regulation of eukaryotic transcription. Cold Spring Harb. Symp. Quant. Biol. 58: 201–218.

Rushton, P.J. and Somssich, I.E. 1998. Transcriptional control of plant genes responsive to pathogens. *Curr. Opin. Plant Biol.* 1: 311–315.

Zhu Q, Droge-Laser W, Dixon R A & Lamb C 1996 Transcriptional activation of plant defense genes. Curr. Opin. Genet. Dev. 6: 624-630.

Phylogenetic Analysis

A. Chandrasekar

Phylogenetics is the science of estimating the evolutionary past, in the case of molecular phylogeny, based on the comparison of DNA or protein sequences. A phylogenetic tree is a statement about the evolutionary relationship between a set of homologous characters of one or several organisms. Homology according to Fitch is the relationship of two characters that have descended, usually with divergence, from a common ancestral character. The characters can be any genic (gene sequence, protein sequence), structural (i.e. morphological) or behavioural feature of an organism. The idea of representing these hypotheses as trees probably dates back to Darwin, but the numerical calculation of trees using quantitative methods is relatively recent, and their application to molecular data even more so. In the age of rapid and rampant gene sequencing, molecular phylogeny has truly come into its own, emerging as a major tool for making sense of a sometimes overwhelming amount information. This tutorial aims to introduce the basic principles behind and programs for constructing evolutionary trees (phylogenetic analysis). It is intended primarily for those who want to read other people's trees, but also as a general introduction for those who might wish to begin to try building their own. In the latter case the reader is warned – phylogenetic analysis and evolutionary theory are not trivial pursuits; as with any new methodology, it is advisable to seek expert help before getting in too deep.

Some basics

Terminology

A phylogenetic tree is composed of branches (edges) and nodes. Branches connect nodes; a node is the point at which two (or more) branches diverge. Branches and nodes can be internal or external (terminal). An internal node corresponds to the hypothetical last common ancestor (LCA) of everything arising from it. Terminal nodes correspond to the sequences from which the tree was derived (also referred to as operational taxonomic units or 'OTUs'). Trees can be made up of multigene families (gene trees) or a single gene from many taxa (species trees, at least theoretically) or a combination of the two. In the first case, the internal nodes correspond to gene duplication events, in the second to speciation events.

Groups

Trees are about groupings (Fig. 1). A node and everything arising from it is a 'clade' or a 'monophyletic group'. A monophyletic group is a natural group; all members are derived from a unique common ancestor (with respect to the rest of the tree) and have inherited a set of unique common traits (characters) from it. A group excluding some of its descendents is a paraphyletic group (e.g. animals excluding humans). A hodge-podge of distantly related OTUs, perhaps superficially resembling one another or retaining similar primitive characteristics, is polyphyletic; that is, not a group at all.

Trees

Intuitively we draw trees from the ground up like real trees (Fig. 2a). However, as these trees get larger and more complex, they can become cluttered and difficult to read. As an alternative we can expand the nodes (Fig. 2b) and turn the tree on its side (Fig. 2c). Now the tree grows left to right, and all the labels are horizontal. This makes the tree easier to read and to annotate. Thus, the widths of the nodes have no meaning; they are simply adjusted to give even spacing to the branches. To make things slightly more complicated, all branches can rotate freely about the plane of their nodes, so all trees in Fig. 2 are identical (except that tree F is 'unrooted', see below). Molecular phylogenetic trees are usually drawn with proportional branch lengths; that is, the lengths of the branches correspond to the amount of evolution (roughly, percent sequence difference) between the two nodes they connect (Fig. 2a–f). Thus, the longer the branches the more relatively divergent (highly evolved) are the sequences attached to them. Alternatively, trees can be drawn to display branching patterns only (‘cladograms’), in which case the lengths of the branches have no meaning (Fig. 2g), but this is rare done with molecular sequence trees.

Roots

At the base of a phylogenetic tree is its ‘root’. This is the oldest point in the tree, and it, in turn, implies the order of branching in the rest of the tree; that is, who shares a more recent common ancestor with whom. The only way to root a tree is with an ‘outgroup’, an external point of reference. An outgroup is anything that is not a natural member of the group of interest (i.e. the ‘ingroup’). This might not seem like a difficult concept, but do not be misled. The excluded member of a monophyletic group (i.e. the exclusion that makes it paraphyletic, Fig. 1) is not an outgroup (just an outcast); for example, humans are not an outgroup to animals. In the absence of an outgroup, the best guess is to place the root in the middle of the tree (at its midpoint), or, better yet, not root it at all (Fig. 2f). Alternatively you can use extrinsic, more traditional taxonomic information, such as the fossil record in the case of species trees. This is obviously more difficult with gene trees.

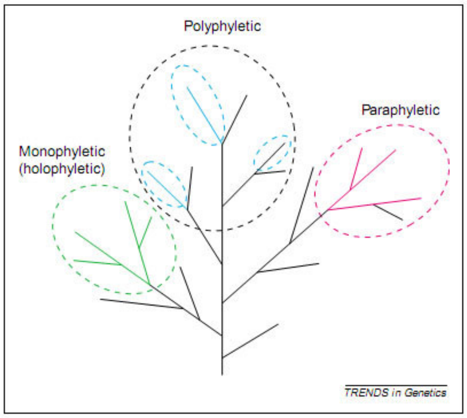


Fig. 1. Trees are about groups: monophyletic (holophyletic), paraphyletic and ‘polyphyletic’.

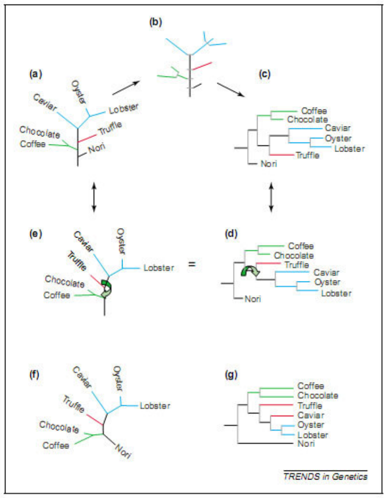


Fig. 2. Phylogenetic tree styles. All these trees have identical branching patterns. The only differences are (f), which is unrooted. (g) is a cladogram, so the branch lengths are right justified and not drawn to scale (i.e. they are not proportional to estimated evolutionary difference

Homology

Evolution is about homology; that is, the similarity due to common ancestry. Homologues can be orthologues or paralogues (Fig. 3). Orthologues only duplicate when their host divides; i.e. along with the rest of the genome (Fig. 3a). They are strictly vertically transmitted (parent to offspring), so their phylogeny traces that of their host lineage (Fig. 3b). Paralogues are members of multigene families; they arise by gene duplication (Fig. 3a). If you try to infer species relationships with paralogues you can run into trouble; if some of the copies are missing, you can be very convincingly misled (Fig. 3c). However, if you have all copies of two paralogues in your tree, then you are fine. Better still, you have two mirror phylogenies (Fig. 3b). In this case, paralogues can serve as each other’s natural outgroup.

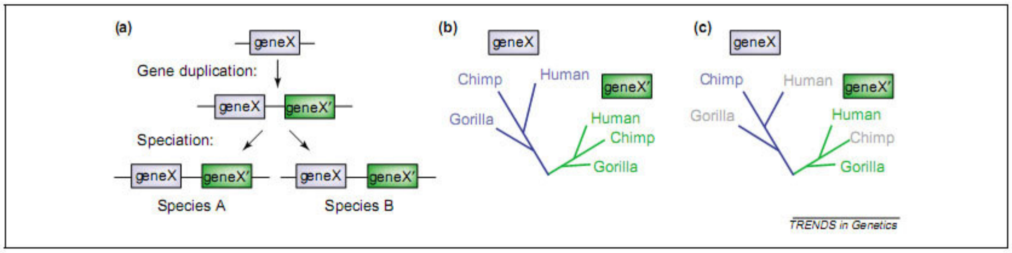


Fig. 3. The problem with paralogues. (a) Paralogous genes are created by gene duplication events. Gene X is duplicated in a common ancestor to species A and B resulting in two paralogous genes, X and X'. All subsequent species inherit both copies of the gene (unless one or the other is lost somewhere along the way). (b) Phylogenetic analysis of the X/X' gene family gives two parallel phylogenies. All sequences of gene X are orthologues of each other, and all the sequences of gene X' are orthologues of each other. However, X and X' are paralogues. Both the X and X' subtrees show the true relationships among the three species. The subtrees are also each other's natural outgroup, and as a result each subtree is rooted with the other (reciprocally rooting). (c) A tree of the X/X' gene family can be misleading if not all the sequences are included (because of incomplete sampling or gene loss). If the broken branches are missing, then the true species relationships are misrepresented.

Assembling a dataset

The first step in constructing a tree is building the dataset. For most of us, this means finding and retrieving sequences from the public domain. The main repository for these data is the public nucleotide database, stored independent in the USA (GenBank), EU (EMBL) and Japan (DDBJ). Primary entries are redundant among them, and they are updated against each other nightly. Some of the most exciting molecular evolutionary data are coming from genome sequencing projects. Much of this data, both in-progress and completed, is deposited in the public database, with some in-progress data partitioned off separately. Other genome project data are available only from their own websites; for example, The Institute for Genomic Research (TIGR) and the Joint Genome Research Institute (DOE). There are two basic kinds of search strategy for finding a set of related sequences – Keywords and similarity. A Keywords search identifies sequences by looking through their written descriptions (i.e. the annotation section of a database file); a similarity search looks at the sequences themselves (e.g. using 'BLAST' software,). Keyword searching is easier and seems more intuitive, but it is far from exhaustive. This is mostly because a lot of data entries are very scantily annotated or even mis-annotated (sometimes quite entertainingly so). This is particularly true for genomic data where high throughput is the priority. The best-annotated data are the painstakingly annotated protein data found in the SwissProt database. This is accessible directly or through the main database sites, but this is only a subset of all that is available. The main search engines for Keywords searching are Entrez (NCBI) and SRS (everywhere else); both have excellent online tutorials. Beginners might find SRS easier, with its simple forms and obvious blanks to fill in. The main search engine for similarity searching is the 'BLAST' software, available at all databanks and most genome websites. The NCBI BLAST server is the most sophisticated with numerous 'flavours' and options such as honing a BLAST search using keywords, searching with alignment profiles to find distant homologues (PSI-BLAST), and much more.

Multiple sequence alignment - The heart of the matter

Molecular trees are based on multiple sequence alignments. Until 1989 these were all assembled by hand (eg.) because the exhaustive alignment of more than six or eight sequences was, and more or less still is, computationally unfeasible. Now, most multiple sequence alignments are constructed by the method known as 'progressive sequence alignment'. This method builds an alignment up stepwise, starting with the most similar sequences and progressively adding the more dissimilar ('divergent') ones (Fig. 4a). The process begins with the construction of a crude ‘guide tree’ (Fig. 4a). This tree then determines the order in which the sequences are progressively added to build the alignment (Fig. 4b). Note that the guide tree is included as part of the alignment output, but only to show the user how the alignment was assembled. The cardinal rule of progressive sequence alignment is ‘once a gap always a gap’; gaps can only be added or enlarged, never moved or removed. This is based on the assumption that the best information on gap placement will be found among the most similar sequences, but also for practical purposes; if gaps were adjusted at every step the alignment process would be tremendously slower. However, the once-a-gap rule can also be the source of some obvious silliness, because there often is better information in the full alignment on where the gaps really belong. This can be particularly apparent for small deletions, which might clearly be shared by several sequences but nonetheless placed at slightly different positions in each (Fig. 5a). These types of error are among the reasons for the widespread and fairly well accepted practice of ‘adjusting’ alignments ‘by eye’ to ‘minimize insertion/deletion events’ (Fig. 5b), using a program such as BioEdit (see below).

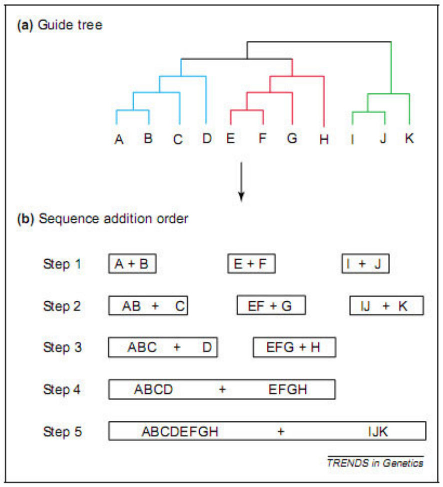


Fig. 4. Steps in progressive sequence alignment. (a) The first step is to calculate the guide tree. (b) This determines the order in which sequences are added to the growing alignment.

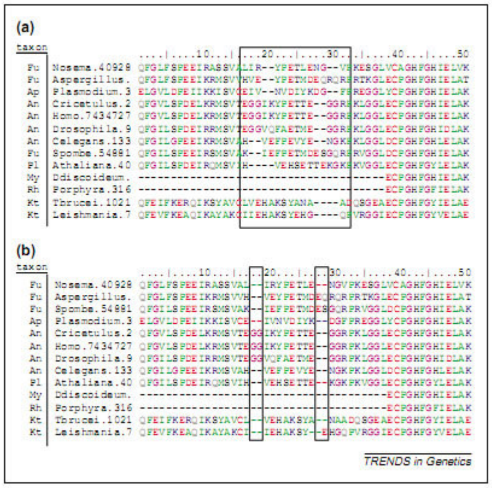


Fig. 5. Refining an alignment. (a) The raw output from a ClustalX alignment of rpb1 sequences, which predicts six insertion/deletion events (boxed), some of which are blatantly inconsistent with known taxonomy. (b) The refined alignment makes much better evolutionary sense, because it shows only two insertion events in well-defined taxonomic groups (animals and higher fungi). Taxon labels are Fu (fungi), An (animals), Pl (green plant), Ap (apicomplexan), Rh (rhodophyte), My (mycetozoan), Kt (kinetoplastids). In (b), the sequence from Saccharomyces pombe has been placed adjacent to the other fungi to make these relationships more obvious.

Alignments are about gaps – where to put them and how big to make them. These are two different issues. Genes do not generally take insertions and deletions lightly. One out of three changes the reading frame, not to mention adding new stop codons or unwieldy junk to a protein’s structure. However, the size of a gap is much less important than the fact that it is there at all, so alignment programs have separate penalties for inserting a gap (which is costly) and for making it bigger (relatively cheap). Ideally gap penalties should differ for closely related versus distantly related sequences, for different kinds of sequence, and for different regions of the same sequence, but this is mostly impractical. Therefore, all gap penalties are compromises, and an alignment can look very different depending on the penalties that are used. In the end, the user might need to try a range of penalties, compare these by eye and pick the most logical combination.

As exhaustive multiple sequence alignment is essentially impossible, new improved methods are something of a cottage industry. However, the oldest program is also one of the easiest, friendliest and most widely used, only partly because it is also free. This is Clustal, now in its W and X incarnations (X being the X-window version of W). Besides basic alignment, the program allows iterative alignment of selected regions, profile alignment (i.e. alignment of alignments), and basic phylogenetic analysis (see below). Although it does not allow you to modify your alignment within the program, Clustal can be run as a subroutine of the BioEdit sequence editor or a Clustal alignment imported into BioEdit for subsequent editing.

Trees – methods, models and madness

The infile

The basic premise of a multiple sequence alignment is that, for each column in the alignment, every residue from every sequence is homologous; that is, has evolved from the same position in a common ancestral sequence without insertion or deletion. When this premise is met, a multiple sequence alignment can hold a wealth of information about protein structure and function, mode of evolution and, of course, phylogeny. However, a molecular phylogeny is only as good as the alignment it's based on. At best, misaligned sequence has no useful phylogenetic information; at worst, it might have convincing misinformation.

Therefore, the first step in tree building is to inspect your alignment carefully and to decide what should and should not be included in your analysis. The general rule is to delete all positions with gaps plus any adjacent, ambiguously aligned positions (i.e. columns in the alignment). This is for two reasons. First, you cannot be confident that these regions are correctly aligned, and if they are misaligned then they will have no real phylogenetic information. Second, even if it is clear that a region containing a gap is correctly aligned, it can have an undue influence on your tree. This is because the larger the gap, the more characters uniting the groups that share it; for example, a 9-nucleotide insertion would be nine shared characters for the OTUs that have it. This is inappropriate because, in reality, a gap is a single evolutionary event, regardless of its size. The importance of a gap is not proportional to its size for the same reason that gap insertion penalties are so much larger than gap extension penalties.

Another important consideration in building molecular trees from protein-coding genes is whether to analyze your sequences at the DNA or the protein level. For closely related sequences, there will be more change (information) at the DNA level, so you'll want to use DNA sequences. For more distant relationships, amino acid sequences hold more information. However, you can still use DNA sequences for distant relationships, but first it is important to remove third codon positions as these will be pure noise or worse.

Methods

The methods for calculating phylogenetic trees fall into two general categories. These are distance-matrix methods, also known as clustering or algorithmic methods (e.g. UPGMA, neighbour-joining, Fitch–Margoliash), and discrete data methods, also known as tree searching methods (e.g. parsimony, maximum likelihood, Bayesian methods). Distance is relatively simple and straightforward – a single statistic, the distance (roughly, the percent sequence difference), is calculated for all pairwise combinations of OTUs, and then the distances are assembled into a tree. Discrete data methods examine each column of the alignment separately and look for the tree that best accommodates all of this information. Unsurprisingly, distance methods are much faster than discrete data methods. However, a distance analysis yields little information other than the tree. Discrete data analyses, however, are information rich; there is an hypothesis for every column in the alignment, so you can trace the evolution at specific sites in the molecule (e.g. catalytic sites or regulatory regions). One way to look at the two classes of methods is to imagine trying to come up with an evolutionary classification of the flowers in your garden. You would start by counting the number of petals, sepals and stamens, etc. for each – that's your dataset. If you used a distance approach, you could sort your flowers simply by the number of characters they share; the flowers with the most characters in common would be presumed to be the most closely related. To use a tree searching method, you would first calculate a set of all possible classification Scheme (i.e. all possible trees), and then measure how each of your characters would have to evolve on each of these trees (e.g. would asymmetrical flowers have to evolve twice on tree 1 versus tree 2, etc). The best possible classification scheme (tree) would be the one that required the simplest set of hypotheses. In the end, in most cases, you would come up with the same groupings. However, if you had any very unusual or highly degenerate flowers, ones that bore little resemblance to any of the rest, then the problem gets harder. Under these conditions, one or other of your methods could fail, although for somewhat different reasons. This is why people generally test their trees with more than one phylogenetic method (see bootstrapping below).

Models

This is all complicated by the fact that molecular evolution is ancient history, a kind of molecular archaeology where we are trying to recover the past by extrapolating backward from a small set of surviving clues. If little evolution has occurred, this is fairly straightforward. However, and quite rapidly, the true evolutionary difference between two sequences becomes obscured by multiple mutations (changes on top of changes), especially at the more rapidly evolving sites. In these cases, a simple count of the differences between two sequences will underestimate how much evolution has actually occurred. Various models (corrections) have been developed to try to estimate the true difference between sequences based on their present states, such as amino acid substitution matrices (e.g. Dayoff, Blossom, etc.) or gamma corrections (giving more weight to changes at slowly evolving sites), etc. However, it is beyond the scope of this tutorial to explain these, and the interested reader should consult one of several excellent texts on molecular evolution for further detail.

Programs

PHYLIP, Mega and PAUPp (pronounced 'pop star') are the most comprehensive and widely used phylogeny packages. All are inexpensive or free, and all allow a variety of models and methods. PHYLIP is the grand daddy of them all, but Mega2 is perhaps the easiest to use (at least on a PC), as it has straightforward pull down menus. Both the PHYLIP and Mega2 manuals are also good all-round primers on phylogenetic theory and practice. PAUPp, a perpetual work in progress, is easily the most sophisticated and versatile of the lot, but also has the steepest learning curve.

An example

Clustal (see above) can also be used to calculate trees using evolutionary distances. Although even the authors do not recommend this for serious phylogeny, quick and easy has its uses. It is also easy to explain and a good way for the beginner to get their feet wet. To calculate a phylogenetic tree using ClustalX involves three or four very simple steps. All the necessary commands are located in the 'Trees' menu. First, remove the regions of the alignment with gaps (select 'delete positions with gaps'). Second, if your sequences are less than 95% identical, you should choose to correct your distance measures for multiple substitutions ('correct for multiple substitutions'). Third, make sure you have the right format for the output (follow the 'output format' link', change 'bootstrap label options' from 'branches' to 'nodes' and close). Finally, calculate the tree by selecting 'bootstrap NJ tree'. This last analysis actually incorporates several steps. First, the program calculates a distance (neighbour-joining) tree, and then it evaluates the tree by a statistical test called bootstrapping (see below). Finally, it puts all the information together into a single output file, consisting of the tree with the appropriate bootstrap values on it. To view these results you need a tree-viewing program, such as Treeview, and you're done. This is about as basic as it gets, but it is often a good place to start; for example, to sort through mountains of sequences quickly to pick a representative set to work with or to decide whether your data look interesting enough to pursue in more detail.

Tests – telling the forest from the trees

Bootstrapping

So how good was that tree? The simplest test of phylogenetic accuracy is the bootstrap; it is rare now to see a tree without it. Bootstrapping essentially tests whether your whole dataset is supporting your tree or if the tree is just a marginal winner among many nearly equal alternatives. This is done by taking random subsamples of the dataset, building trees from each of these and calculating the frequency with which the various parts of your tree are reproduced in each of these random subsamples. If group X is found in every subsample tree, then its bootstrap support is 100%, if it's found in only two-thirds of the subsample trees, its bootstrap support is 67% (Fig. 6). Each of the subsamples is the same size as the original, which is accomplished by allowing repeat sampling of sites; that is, random sampling with replacement. It is a simple test, but bootstrap analyses of known phylogenies (viral populations evolved in the laboratory) show that it is a generally dependable measure of phylogenetic accuracy and that value of 70% or higher are likely to indicate reliable groupings.

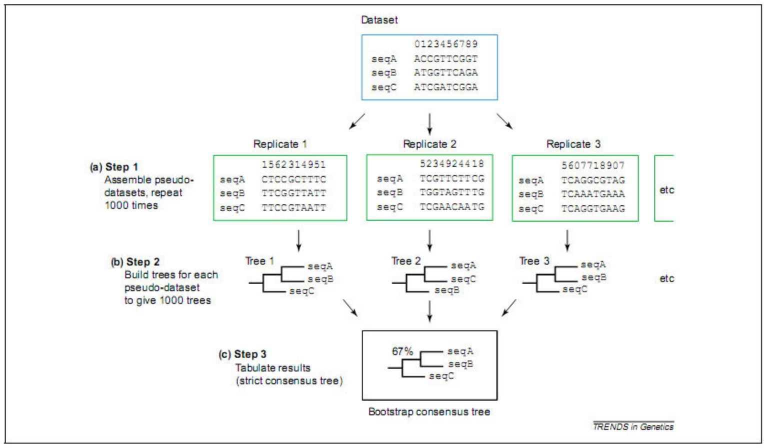


Fig. 6. Bootstrap analysis proceeds in three steps. The dataset is randomly sampled with replacement to create multiple pseudo-datasets of the same size as the original ((a), three are shown in this example). (b) Individual trees are constructed from each of the pseudo-datasets. (c) Each of the pseudo-dataset trees are scored for which nodes (groupings) appear and how often. In this case, a node uniting seqA plus seqB is found in two of the three replicate trees. This gives a bootstrap support for this grouping of 2/3 or 67%.

Long branches

The most problematic and pervasive problem in molecular phylogeny is the problem of 'long branch attraction'. This is the tendency of highly divergent sequences (i.e. those with long terminal branches) to group together in a tree regardless of their true relationships. This is at least partly because rapidly evolving sequences, or sequences without any close relatives, will have numerous unique mutations (with respect to the rest of the tree). Because there are only a limited number of possible states (20 amino acids or 4 nucleotides) for rapidly evolving sites to change to, sequences with a lot of these changes will start to pick up spurious similarities to each other. If their branches are very long (i.e. if there are a lot of these changes), these spurious similarities can override the true phylogenetic signal, and the sequences will be 'attracted' to each other.

This causes all sorts of problems, one of which is that bootstrap values all over the tree tend to deteriorate. In fact, one way to test whether you have a problem with long branches is to remove these sequences from your dataset and see if the bootstrap values go up.

However, there is no easy solution to the problem of long-branch attraction. Methods such as maximum likelihood tend to be less affected than others, but rarely strikingly so. Generally, it is best to break up the long branches, by adding intermediate sequences to a tree. To return to the flower analogy, with highly derived forms, if you can find some of the intermediate relatives it becomes much easier to see how they fit in with the rest of your taxa. If all else fails, and you do not really need the offending sequence, you can either omit it from your dataset, being completely transparent about why, or analyse your data with and without it and see what difference it really makes.

Data presentation

Finally, there are few set rules on how to present phylogenetic trees, but there are some widely accepted conventions. In molecular phylogenetic trees, branch lengths are almost always drawn to scale; that is, proportional to the amount of evolution estimated to have occurred along them. Although the relationship between branch lengths and real time is far from straightforward and probably unreliable for any single gene, lengths still give a good general impression of relative rates of change across a tree. Bootstrap values should be displayed as percentages, not raw values. This makes the tree easier to read and to compare with other trees. By convention, only bootstrap values of 50% or higher are reported; lower values mean that the node in question was found in less than half of the bootstrap replicates. And finally, please, please use meaningful names for OTUs or annotate your tree by indicating important groupings with brackets, color etc.; a tree full of three letter acronyms or bare database accession numbers can be excruciating to interpret.

Phylogenetic analysis is a powerful tool for sorting and interpreting molecular data. With even a very basic understanding of general principles and conventions it is possible to glean valuable information from a phylogenetic tree – on the origin, evolution and possible function of genes and the proteins they might encode. I hope that you will now feel more confident in reading trees and more intrigued than ever to discover what they might have to say.

References

1. Sneath, P.H.A. and Sokal, R.R. (1973) Numerical Taxonomy, W.H. Freeman
2. Zuckerkandl, E. and Pauling, L. (1965) Evolutionary divergence and convergence in proteins. In Evolving Genes and Proteins (Bryson, V. and Vogel, H.J., eds) pp. 97–166, Academic Press
3. Iwabe, N. et al. (1989) Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc. Natl. Acad. Sci. U. S. A. 86, 9355–9359
4. Gogarten, J.P. et al. (1989) Evolution of the vacuolar Hþ-ATPase: implications for the origin of eukaryotes. Proc. Natl. Acad. Sci. U. S. A. 86, 6661–6665
5. Baldauf, S.L. et al. (1996) The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. Proc. Natl. Acad. Sci. U. S. A. 93, 7749–7754
6. Boeckmann, B. et al. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 31, 365–370
7. Altschul, S.F. et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402
8. Roberts, L. (2003) Atussle over the rules for DNA data sharing. Science 298, 1312–1313
9. Schwartz, R.M. and Dayhoff, M.O. (1978) Matrices for detecting distant relationships. In Atlas of Protein Sequence and Structure (Dayhoff, M.O., ed.), pp. 353–358, National Biomedical Research Foundation
10. Feng, D.F. and Doolittle, R.F. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J. Mol. Evol. 25, 351–360
11. Higgins, D.G. and Sharp, P.M. (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene 73, 237–244
12. Hall, B.G. (2000) Phylogenetic Trees Made Easy: a How-To Manual for Molecular Biologists, Sinauer Associates

13. Page, R.D.M. and Holmes, E.C. (1998) Molecular Evolution: a Phylogenetic Approach, Blackwell Science
14. Graur, D. and Li, W-H. (1999) Fundamentals of Molecular Evolution, Sinauer Associates
15. Durbin, R. et al. (2000) Biological Sequence Analysis, Cambridge University Press
16. Nei, M. and Kumar, S. (2000) Molecular Evolution and Phylogenetics, Cambridge University Press
17. Swofford, D.L. et al. (1996) Phylogenetic inference. In Molecular Systematics (Hillis, D.M. et al., eds), Sinauer Associates
18. Felsenstein, J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39, 783–791
19. Hillis, D.M. and Bull, J.J. (1993) An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analyses. Syst. Biol. 42, 182–192
20. Felsenstein, J. (1978) Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27, 401–410
21. Gribaldo, S. and Philippe, H. (2002) Ancient phylogenetic relationships. Theor. Popul. Biol. 61, 391–408
22. Hillis, D.M. et al. (1994) Application and accuracy of molecular phylogenies. Science 264, 671–677
23. Zwickl, D.J. and Hillis, D.M. (2002) Increased taxon sampling greatly reduces phylogenetic error. Syst Biol. 51, 588–598
24. TRENDS in Genetics Vol.19 No.6 June 2003 349 http://tigs.trends.com

Computational Methods in Gene Prediction

Riju

Introduction

A **gene** is a unit of heredity in a living organism. It is normally a stretch of DNA that codes for a type of protein or an RNA chain that has a function in the organism. All proteins and functional RNA chains are specified by genes. All living things depend on genes. Genes hold the information to build and maintain an organism's cells and pass genetic traits to offspring. A modern working definition of a gene is "*a locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions, and or other functional sequence regions* ". Incorrect colloquial usage of the term *gene* may actually refer to an allele: a *gene* is the basic instruction, a sequence of nucleic acid (DNA or, in the case of certain viruses RNA), while an *allele* is one variant of that instruction.

The notion of a gene is evolving with the science of genetics, which began when Gregor Mendel noticed that biological variations are inherited from parent organisms as specific, discrete traits. The biological entity responsible for defining traits was later termed a *gene*, but the biological basis for inheritance remained unknown until DNA was identified as the genetic material in the 1940s. All organisms have many genes corresponding to many different biological traits, some of which are immediately visible, such as eye color or number of limbs, and some of which are not, such as blood type or increased risk for specific diseases, or the thousands of basic biochemical processes that comprise life.

In cells, a gene is a portion of DNA that contains both "coding" sequences that determine what the gene does, and "non-coding" sequences that determine when the gene is active (expressed). When a gene is active, the coding and non-coding sequences are copied in a process called transcription, producing an RNA copy of the gene's information. This piece of RNA can then direct the synthesis of proteins via the genetic code. In other cases, the RNA is used directly, for example as part of the ribosome.

The molecules resulting from gene expression, whether RNA or protein, are known as gene products, and are responsible for the development and functioning of all living things. The physical development and phenotype of organisms can be thought of as a product of genes interacting with each other and with the environment. A concise definition of a gene, taking into account complex patterns of regulation and transcription, genic conservation and non-coding RNA genes, has been proposed by Gerstein et al. "A gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products".

Gene Structure in Eukaryotes

The gene structure and the gene expression mechanism in eukaryotes are far more complicated than in prokaryotes. In typical eukaryotes, the region of the DNA coding for a protein is usually not continuous. Eukaryotic gene is organization as exons and introns. During transcription, both exons and introns are transcribed onto the RNA, in their linear order. Thereafter, a process called splicing takes place, in which, the intron sequences are excised and discarded from the RNA sequence. The remaining RNA segments, the ones corresponding to the exons are ligated to form the mature RNA strand. Generally, all exons can be separated into four classes (fig. 1): 5'exons, internal exons, 3' exons and intronless exons (or, simply, intronless genes). They can be further subdivided into 12 mutually exclusive subclasses, according to their coding content. Almost all gene-prediction papers refer to four types of 'exon'. A typical multi-exon gene starts with the promoter region, which is followed by a transcribed but non-coding region called *5'* untranslated region (5' UTR). Then follows the initial exon which contains the start codon. Following the initial exon, there is an alternating series of introns and internal exons, followed by the terminating exon, which contains the stop codon. It is followed by another non-coding region called the 3' UTR. Ending the eukaryotic gene, there is a polyadenylation (polyA) signal: the nucleotide Adenine repeating several times. The exon-intron boundaries (i.e., the splice sites) are signalled by specific short (2bp long) sequences. The 5'(3') end of an intron (exon) is called the *donor* site, and the 3'(5') end of an intron (exon) is called the *acceptor* site. The problem of gene identification is complicated in the case of eukaryotes by the vast variation that is found in gene structure.

Gene prediction Methods

Genes are predicted by finding the gene structure that has the highest score, given the sequence. The accuracy with which genes can be predicted is still far from satisfactory. Although, at the nucleotide level, 80% of genes are accurately predicted, at the exon level only 45% are predicted, and at the whole-gene level only ~20%. Generally there are six types of gene prediction methods.

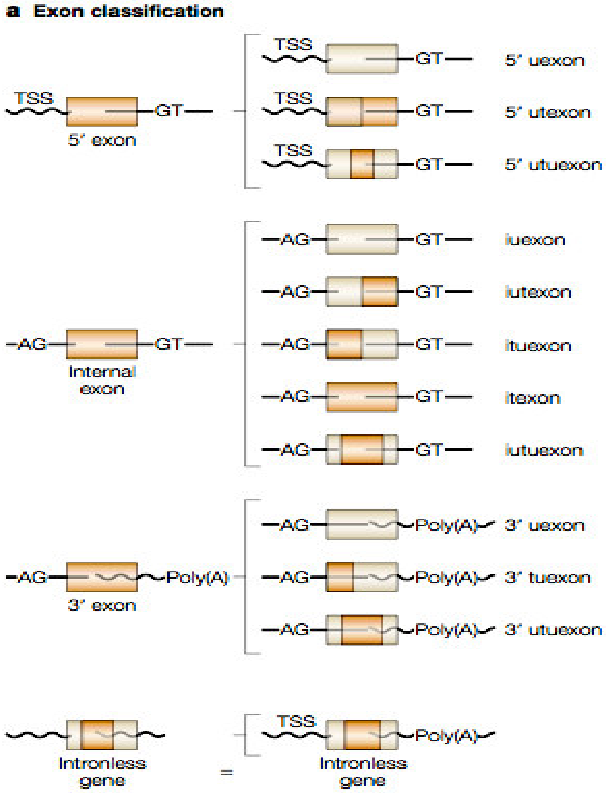


Fig 1. Classification of exons.
Zang 2002, Genetics 3: 698-709

1. Signal - Exon based method

'Exon-definition' model: In vertebrates, the internal exons are small (~140 nucleotides on average), whereas introns are typically much larger (with some being more than 100 kb in length); 'Exon-definition' model postulates that an internal exon is initially recognized by the presence of a chain of interacting splicing factors that span it; The binding of these trans- acting factors to the pre-mRNA is responsible for the non-random nucleotide patterns that form the molecular basis for all exon-recognition algorithms. These sequence features are often divided into two types: 'signals', which correspond to short cis-elements or boundary sites (such as splice sites and branch sites); and 'content', which corresponds to the extended functional regions (such as exons and introns); Also coding statistics and scoring features and integration methods for evaluation.

1.1 *Ab initio* gene prediction

Ab initio method is based on the computational identification of genes is to use gene structure as a template to detect genes. This method rely on two types of sequence information: signal sensors and content sensors. Signal sensors refer to short sequence motifs, such as splice sites, branch points, polypyramidine tracts, start codons and stop codons. Exon detection must rely on the content sensors, which refer to the patterns of codon usage that are unique to a species, and allow coding sequences to be distinguished from the surrounding noncoding sequences by statistical detection algorithms. Many algorithms are applied for modeling gene structure. Based on these models, a great number of *ab initio* gene prediction programs have been developed. Some of the frequently used are GeneParser, Genie and GRAIL. The latest generation of gene prediction programs such as GENESCAN or AUGUSTUS provide an acceptable level of accuracy coupled with ready usability, and are therefore of significant use in the analysis of uncharacterized genomic sequences. These programs essentially rely on a statistical model, the Hidden Markov Model, for predicting either an intronic or exonic state of the sequence in a given region. Thus, the algorithm does not rely directly on sequence homology to known genes. However, the statistical models that underpin the operation of the software have been trained on a set of training data relating to known genes and characteristic structure. Therefore, parts of the genome that are of significantly different structure compared to these training data may be only poorly annotated.

2. Intron based gene structure prediction

Intron-based gene-structure prediction: Lower eukaryotes have better success; short splicing signals were identified in introns: the donor site (5' splice site or 5' ss), which is characterized by the consensus AG|GURAGU; the acceptor site (3'ss), which is characterized by the consensus YYYYYYYYYYNCAG|G; and the less-conserved branch site, which is characterized by CURAY; However these are not sufficient for higher organisms (Lim and Burge, 2001)

3. Frame-specific hexamer frequencies method

Frame-specific hexamer frequencies method: utilizing coding statistics (content-based) codon bias information and codon–codon correlations. They also capture splice-site preferences, which are the most characteristic exon–intron features. For long open reading frames (ORFs), such as in bacterial or intronless genes, frame-specific hexamer frequencies alone can detect most of the CDS regions. Some methods interpret sequences as successions of 'words'— so-called because nucleotides are not independent of each other, but tend to occur together as if in a word — of length k (k-tuples); 6-tuples are called hexamers. In-frame hexamer frequencies in a region of DNA have traditionally been used as a powerful way of discriminating coding regions from non-coding regions, as some 'words' are more likely to be present in either type of DNA. A score s for a hexamer w, such as CAGCAG, can be defined as $s(w) = log(freq(w))$. Because the frequency of CAGCAG is relatively high in exons, its score in exons will be higher than that of, for example, TAATAA.

4. Sequence similarity based method

Sequence similarity search is a conceptually simple approach that is based on finding similarity in gene sequence between EST, proteins, or other genomes to the input sequence. This approach is based on the assumption that functional regions(exons) are more conserved evolutionary than nonfunctional regions(intronic regions). Once there is a similarity between a certain genomic region and an EST, DNA, or protein, the similarity information can be used to infer gene structure or function of that region. PROCRUSTES and GeneWise, use global alignment of a homologous protein to translated ORFs in a genomic sequence for gene prediction. The limitation to this type of approaches is that only about half of the genes being discovered have significant homology to genes in the databases.

5. Comparative genomics methods

By comparing the genomes of several closely related species, conserved regulatory regions can be identified easily. For these reasons, making use of comparative genomic data is an important future challenge for the gene-prediction field; Note: (a) 1/2 conservation between human and mouse are in non-coding region false positives; larger scale of genome environment may be different; (b) comparative gene prediction requires the sequenced genomes of two or more species at the appropriate phylogenetic distance. Some of the frequently used gene prediction programs are shown in Table 1.

Gene prediction using GENSCAN

GENSCAN, developed by Chris Burge and Sam Karlin (Burge and Karlin,1997; Burge and Karlin,1998), is designed to predict complete gene structure. In this section we shall consider how GENSCAN determines various parameters of the model to get meaningful results. The program uses a training set of 238 multi-exon genes and 142 single-exon genes. These are completely sequenced genes from GenBank. On whole, the training set consists of about 2.5 million base pairs.

Initial state probabilities

The initial probabilities of various states in the model should be proportional to the frequencies with which various functional units occur in the actual human genomic data. For example, if the estimated proportion of the non-coding intergenic region is 80%, then initial probability for the state *N* must be around 0.8. But as a matter of fact, the relative bulk of the various functional units is found to vary considerably with the C+G content (*isochore*) of the genomic sequence. Thus, for training GENSCAN the training set is divided into four categories depending on the C+G content of the sequence. The categories are: (1) < 43%, (2) 43 -51% (3) 51 - 57% & (4) > 57. For each of these categories, separate initial state probabilities are computed by estimating the relative frequencies of various functional units in these categories.

Transition probabilities

Transition probabilities are also known to vary quite a bit with the C+G content (although not as much as the initial probabilities). Thus, transition probabilities are also separately computed for each of the categories. Of course, while estimating these probabilities, it is ensured that the transitions are biologically permissible. For example, some transitions are obligatory (like $P^+ \rightarrow F^+$). Such transitions are assigned probability one.

State length distributions

Different functional units on a gene have vastly different lengths. For example, an average internal exon is about 150bp long, while introns of the order of 1Kbp length are not uncommon. Thus, in our probabilistic model of gene structure, different states need to have different length distributions. Intron lengths are known to vary dramatically with the C+G content. For example, the mean intron length for category I (< 43% C+G) of the training set is 2069 bp as opposed to only 518bp for category IV (> 57 % C+G). Thus, the program uses separate distributions for intron states in each category. The learning set shows quite different length distributions for initial exons, internal exons and terminal exons. Consequently, different distributions are used for them. It is important to note here is that the length of an internal exon has to be consistent with the phase of its adjacent introns. For example, if the preceding state is I_2 and the succeeding state is I_1 , then the generated internal exon length (for state E_2 in this case) must be $3n+2$ for some *n*. *n* is therefore generated randomly according to the length distribution and then a string of length $3n+2$ is generated according to the string generating model for that state. For the 5' UTR and 3' UTR states, geometric distributions with mean values of 769bp and 457bp are used.

Signal models

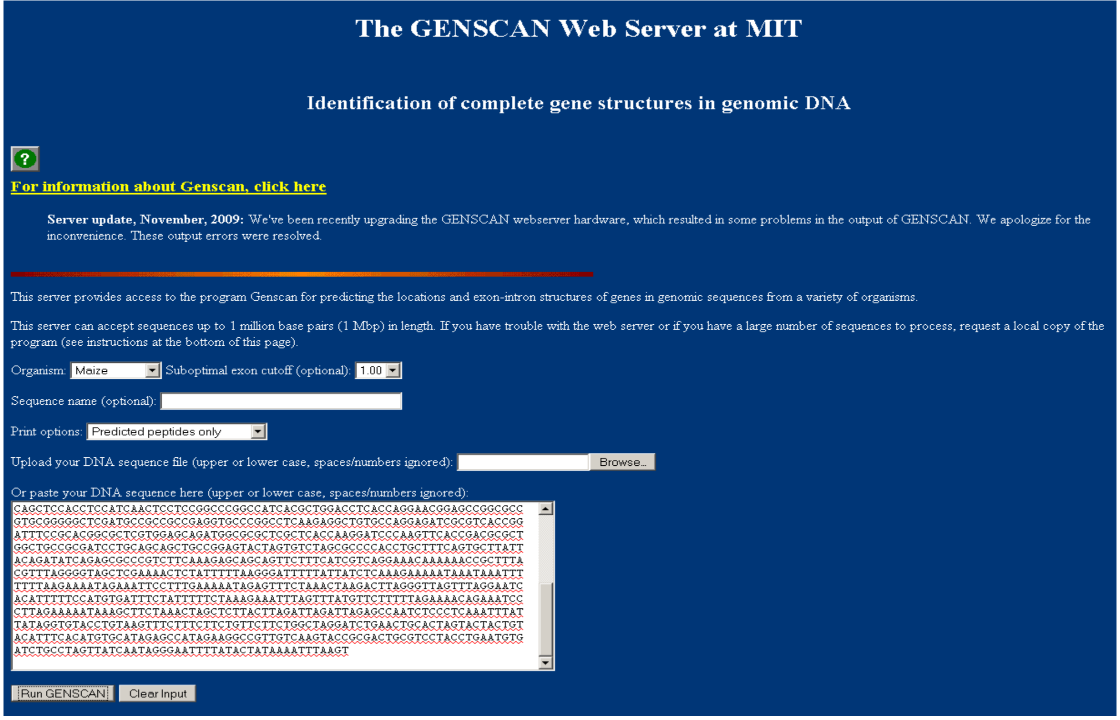
GENSCAN uses 3 kinds of signal models to model different functional units. It uses the weight matrix model (WMM) for modeling polyadenylation signals, translation initiation signal, translation termination signal and promoters. A modified version of the weighted array model (WAM) is used for acceptor splice sites.

Performance of GENSCAN

GENSCAN's accuracy has been significantly better that other programs.
Important features of GENSCAN include

- Identification of complete intron/exon structures of a gene in genomic DNA.
- Ability to predict multiple genes and to deal with partial as well as complete genes.
- Ability to predict consistent sets of genes occurring on either or both strands of the DNA.

Although the results are good they are still not good enough for massive gene finding. GENSCAN has 80% chance of detecting an exon. If a gene has more than one exon the probability of correctly detecting all of them declines rapidly.



The GENSCAN Web Server at MIT

Identification of complete gene structures in genomic DNA

For information about Genscan, click here

Server update, November, 2009: We've been recently upgrading the GENSCAN webserver hardware, which resulted in some problems in the output of GENSCAN. We apologize for the inconvenience. These output errors were resolved.

This server provides access to the program Genscan for predicting the locations and exon-intron structures of genes in genomic sequences from a variety of organisms.

This server can accept sequences up to 1 million base pairs (1 Mbp) in length. If you have trouble with the web server or if you have a large number of sequences to process, request a local copy of the program (see instructions at the bottom of this page).

Organism: Maize Suboptimal exon cutoff (optional): 1.00

Sequence name (optional):

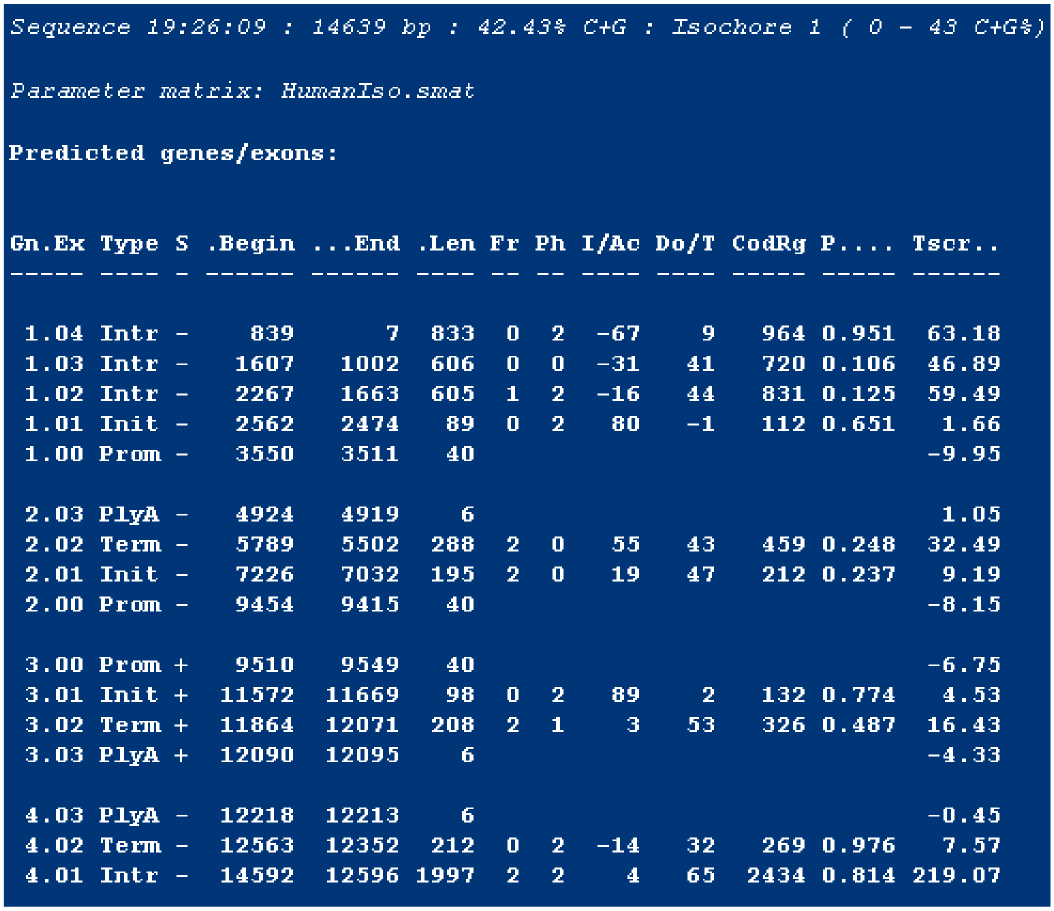
Print options: Predicted peptides only

Upload your DNA sequence file (upper or lower case, spaces/numbers ignored):

Or paste your DNA sequence here (upper or lower case, spaces/numbers ignored):

```
CAGCTCCACCTCCATCAACTCCTCCGGCCCGGCCATCACGCTGGACCTCACCAGGAACGGAGCCGGCGCC
GTGCGGGGGCTCGATGCCGCCGCCGAGGTGCCCGGCCTCAAGAGGCTGTGCCAGGAGATCGCGTCACCGG
ATTTCCGCACGGCGCTCGTGGAGCAGATGGCGCGCTCGCTCACCAAGGATCCCAAGTTCACCGACGCGCT
GGCTGCCGCGATCCTGCAGCAGCTGCCGGAGTACTAGTGTCTAGCGCCCCACCTGCTTTCAGTGCTTATT
ACAGATATCAGAGCGCCCGTCTTCAAAGAGCAGCAGTTCTTTCATCGTCAGGAAACAAAAAACCGCTTTA
CGTTTAGGGGTAGCTCGAAAACTCTATTTTTAAGGGATTTTTATTATCTCAAAGAAAAATAAATAAATTT
TTTTAAGAAAATAGAAATTCCTTTGAAAAATAGAGTTTCTAAACTAAGACTTAGGGTTAGTTTAGGAATC
ACATTTTTCCATGTGATTTCTATTTTTCTAAAGAAATTTAGTTTATGTTCTTTTTAGAAAACAGAAATCC
CTTAGAAAAATAAAGCTTCTAAACTAGCTCTTACTTAGATTAGATTAGAGCCAATCTCCCTCAAATTTAT
TATAGGTGTACCTGTAAGTTTCTTTCTTCTGTTCTTCTGGCTAGGATCTGAACTGCACTAGTACTACTGT
ACATTTCACATGTGCATAGAGCCATAGAAGGCCGTTGTCAAGTACCGCGACTGCGTCCTACCTGAATGTG
ATCTGCCTAGTTATCAATAGGGAATTTTATACTATAAAATTTAAGT
```

GENSCAN Result



```
Sequence 19:26:09 : 14639 bp : 42.43% C+G : Isochore 1 ( 0 - 43 C+G%)

Parameter matrix: HumanIso.smat

Predicted genes/exons:


Gn.Ex Type S .Begin ...End .Len Fr Ph I/Ac Do/T CodRg P.... Tscr..
----- ---- - ------ ------ ---- -- -- ---- ---- ----- ----- ------

 1.04 Intr -    839      7  833  0  2  -67    9   964 0.951  63.18
 1.03 Intr -   1607   1002  606  0  0  -31   41   720 0.106  46.89
 1.02 Intr -   2267   1663  605  1  2  -16   44   831 0.125  59.49
 1.01 Init -   2562   2474   89  0  2   80   -1   112 0.651   1.66
 1.00 Prom -   3550   3511   40                              -9.95

 2.03 PlyA -   4924   4919    6                               1.05
 2.02 Term -   5789   5502  288  2  0   55   43   459 0.248  32.49
 2.01 Init -   7226   7032  195  2  0   19   47   212 0.237   9.19
 2.00 Prom -   9454   9415   40                              -8.15

 3.00 Prom +   9510   9549   40                              -6.75
 3.01 Init +  11572  11669   98  0  2   89    2   132 0.774   4.53
 3.02 Term +  11864  12071  208  2  1    3   53   326 0.487  16.43
 3.03 PlyA +  12090  12095    6                              -4.33

 4.03 PlyA -  12218  12213    6                              -0.45
 4.02 Term -  12563  12352  212  0  2  -14   32   269 0.976   7.57
 4.01 Intr -  14592  12596 1997  2  2    4   65  2434 0.814 219.07
```

Gn.Ex: gene number, exon number (for reference)
Type: Init = Initial exon (ATG to 5' splice site)
Intr = Internal exon (3' splice site to 5' splice site)
Term = Terminal exon (3' splice site to stop codon)
Sngl = Single-exon gene (ATG to stop)
Prom = Promoter (TATA box / initation site)
PlyA = poly-A signal (consensus: AATAAA)
S : DNA strand (+ = input strand; - = opposite strand)
Begin: beginning of exon or signal (numbered on input strand)
End : end point of exon or signal (numbered on input strand)
Len : length of exon or signal (bp)
Fr : reading frame (a forward strand codon ending at x has frame x mod 3)
Ph : net phase of exon (exon length modulo 3)
I/Ac : initiation signal or 3' splice site score (tenth bit units)
Do/T: 5' splice site or termination signal score (tenth bit units)
CodRg: coding region score (tenth bit units)
P : probability of exon (sum over all parses containing exon)
Tscr: exon score (depends on length, I/Ac, Do/T and CodRg scores)

Table 1. *Ab inito* gene prediction programs

PROGRAME	ORGANISM	ALGORITHM*	WEBSITE
GeneID	Vertiberates, Plants	DP	http://www1.imim.es/geneid.html
FGENESH	Humun, Mouse, Drosophila,rice	HMM	http://www.softberry.com/berry.phtml
Gene Parser	Vertebrates	NN	http://beagle.colorado.edu/GeneParse.html
Genie	Drosophila,humun,other	GHMM	http://www.fruitfly.org/seq_tools/genie.html
GenLang	Vertribates,Drosophila,dicots	Grammar rule	http://www.cbil.upenn.edu/genlang/genlang_home.html
GENSCAN	Vertibrates,Arabidopsis,maize	GHMM	http://genes.mit.edu/GENSCAN.html
GlimmerM	Small eukaryotes,Arabidopsis,rice	IMM	http://www.tigr.org/tdb/glimmerm/glmr_form.html
GRAIL	Humun, mouse, Arabidopsis, Drosophila	NN,DP	http://www.compbio.ornl.gov/Grail-bin/EmptyGrail
HMM gene	Vertebrates, *C.elegans*	CHMM	http://www.cbs.dtu.dk./services/HMM gene/
AUGUSTUS	Human, Arabidopsis	IMM,WWAM	http://www.augustus.gobics.de/
MZEF	Human,mouse,Arabidopsis, Fission Yeast	Quadratic discriminanat analysis	http://www.rulai.cshl.org/tools/genefinder/

*DP, dynamic programming; NN, Neural Network; MM, Markov model; HMM, Hidden Markov Model; CHMM, class HMM; GHMM, Generalized HMM; IMM, interpolated MM.

Evaluation of Gene Prediction Programs

The abundance of gene prediction program raises the problem of adequate evaluation of prediction program quality. Sensitivity (Sn) and Specificity (Sp) are probably the two most widely used measures,which are explained by Burset and Guigo. The accuracy of the predictions can be measured at three different levels:coding nucleotide sequences, exonic structure, and the protein product. The nucleotide level accuracy that measures Sn, Sp, CC(Correlation coefficient) and AC(approximate coefficient) gives an overall sense of how closely the predicted and actual coding regions are in a sequence alignment, but does not accurately reflect the identification of precise exon boundaries. Evaluation at the exon level mainly provides how well the sequence signals(splice site, start codon, and stop codon, etc.) are identified. The accuracy can be measured by comparing predicted and real exons along the test sequence.At the protein level, the accuracy is measured by comparing the protein product encoded by the actual gene in the test sequence with the protein product encoded by the predicted gene.

The prediction accuracy of some usual programs has been tested on Burset and Guigo's sequence set, and the results at exon level shows that GENSCAN based on GHMM is significantly more accurate than other programs.

REFERENCE

Burge C. & Karlin S (1997) Prediction of complete gene structure in human genomic DNA. J. Mol. Biol. 268, 78–94

Burge, C. B. and Karlin, S. (1998) Finding the genes in genomic DNA. *Curr. Opin. Struct. Biol.* 8, 346-354

Do JH, Choi D-K (2006) Computational approaches to gene prediction. The journal of microbiology 44(2):137-144

Lim, L. P. & Burge, C. B. (2001). A computational analysis of sequence features involved in recognition of short introns. *Proc. Natl Acad. Sci. USA* 98, 11193–11198

Wang Z, Chen Y, Li Y(2004) A brief review of computational gene prediction methods. Genomics Proteomics Bioinformatics. 2(4): 216-221

Zhang MQ (2002) Computational prediction of eukaryotic protein-coding genes. Nat Rev Genet. 3(9): 698-709

Basics and Applications of Statistics in Biological Research

K. N. Shiva and K. Jayarajan

Statistics is the branch of applied mathematics applied to observation data. It is '*Calculus of Observations'*. It is used to study the population, variation and its causes, methods of reduction of data. Statistics is the study of principles and methods used in the collection, analysis and interpretation of numerical data in any sphere of enquiry (Chandel, 1999).

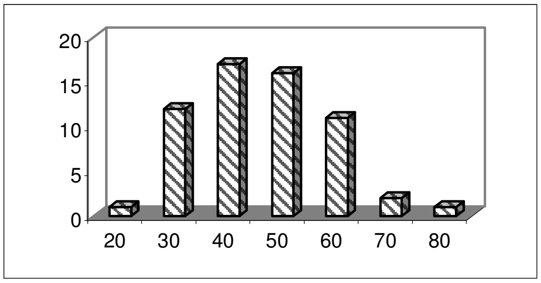
Inference is drawn or conclusions are formulated through a process of statistical reasoning. Analysis of data helps us to schematise this process of reasoning. Suppose we have the data on yield of maize from 60 plots in a farm documented as in the following table.

Yield of maize (kg)				
54	39	45	59	34
49	34	40	54	29
40	25	31	45	20
51	36	42	56	31
52	37	43	57	32
44	29	35	49	24
51	36	42	56	31
43	28	34	48	23
47	32	38	52	27
41	26	31	46	21
64	34	45	74	24
59	29	40	69	19

The first step in the analysis of data is to explore the data to highlight the hidden properties. We first find that the values of yield differ from plot to plot (a few of them may be equal by chance). Here, yield is the variable we are interested to study further. Let yield be represented by the letter 'x'. We observe that x is distributed in the range of 19 to 74. This is a piece of information which we easily observed. This is an attribute of the data in our example.
Similarly we can look for other attributes of the data. We further observe that a few of the values repeat in the distribution. This property can be made use of in presenting the data in a more intelligible way thus we observe that there is only one observation with value less than 20, 12 values within the range 20-30, 17 within the range 30-40 and so on.

<	**Frequency**
20	1
30	12
40	17
50	16
60	11
70	2
80	1

The number of values occurring in a given range is called the frequency in that range and a table representing different ranges (classes) and the corresponding frequencies is called the frequency table. Now this table gives much more information about the distribution than the raw data. This can be made to look further appealing if we plot the data as follows:



Now from the chart we observe there is heavy concentration values in the middle range of the distribution .We say that the variable has got a tendency to assume more often values in the central region of the distribution. This property is termed as ***central tendency***. This is yet another attribute of the data in our example.

We further observe that the data are occurring in the range (19, 74). That is the values are scattered within this range. In our example there is very heavy concentration in the mid range but only a few observations at the tail regions, sometimes the data may be evenly spread out within the range or scattered within the range without any ▯ cognisable pattern. Thus this quality of data being spread out (dispersion) is another attribute which we can follow. This can be illustrated with a small example.

Set1	set2	set3
1	41	50
51	51	51
101	61	52

We find in the first step data are widely scattered in the second set they are moderately scattered, while in the third set the values are very close to each other. But in all the cases the central value remains the same (51) and the other values are scattered about 51, thus the extent of spread of the individual values is a property that reveals information about the data.

Let us see the data in the example given below:

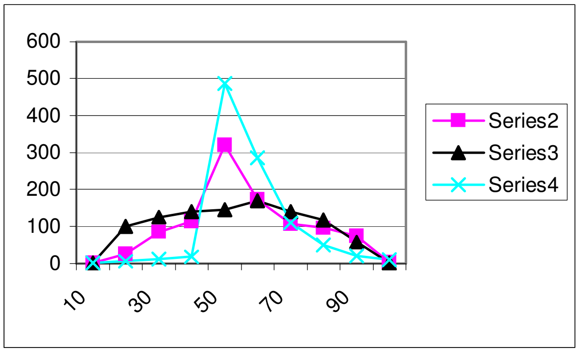
Value less than	frequency 1	frequency 2	frequency 3
10	2	5	2
20	26	465	6
30	85	315	12
40	114	75	18

50	320	67	25
60	173	34	40
70	107	23	110
80	96	9	487
90	74	5	285
100	3	2	15
	100		
Total	0	1000	1000

We observe that there are concentrations of frequencies in the mid range in case 1 and the frequencies tail of in the beginning and the end. In case 2 we observe heavy concentration of values in the beginning while in the higher ranges frequencies tail off. And in the third case it is in the reverse order. If we observe heights of trees in a plantation we may see that the data follow the first pattern. In a questionnaire survey the data on income may follow the second pattern while the expenditure on cosmetics may follow the third pattern. If we plot charts as above we observe the first set will result in a symmetric plot while both other show lack of symmetry. This property of lack of symmetry (Skewness) or other wise is another attribute of a data, which we can look for.

Sometimes, we find that a large portion of observations are concentrated at some point resulting in an unusually high graph while in some other cases data are almost evenly distributed in a wide range. In the example given below we find three cases. Series 1 showing an unusual peakness in the middle, series 2 with peak ness which we normally expect and the third series showing moderate frequencies in a wide range. Thus the level of peak ness (kurtosis) of the "frequency curve" also gives us some information about the distribution.

We have seen above some attributes of data to be assessed, which help us in understanding the structure of the distribution.



The four attributes will describes satisfactorily the structure of the given set of data (distribution). These attributes also help us in distinguishing one distribution from the other fairly satisfactorily. However, it may be borne in mind that we need sixteen attributes to describe a distribution completely**.**

Now the question is 'Is there any yardstick'to assess these properties or attributes?. 'How can we compare two distributions in terms of these attributes'? A solution to these questions lies in quantifying these attributes or simply representing these attributes by numbers, which can be regarded as measurers of theses attributes. Such measures should be easily understood and easily computable, well defined and should satisfy other desirable properties so that we can use them in further analysis of our data. Some of the measures satisfying the above requirements are discussed below.

Measures of Central Tendency

As we have seen above, by central tendency, we mean the property of heavy concentration of values in the central region of the distribution. A *measure* of ***Central Tendency*** is a hypothetical value about which majority of the values is distributed in the data. This may or may not be a value observed and listed in the data. A few measures which are appropriate in many a common situation are discussed here briefly.

1. Arithmetic Mean (or simply mean).

Let us suppose there are N observations listed as X_1, X_2, X_3, ……..,X_N.

Then, **A.M = ($\sum X_i$)/N**

(This is the simple average of all values in the data)

In our example (table 1)

AM = (54+49+…….+59+39+………..+19)/60 = 40.4

(Remember: it is expressed in the same scale and unit as data values are)

Arithmetic mean is commonly referred to as "average" or simply as "mean".

Population Mean:

$$\mu = \frac{\sum x}{N}$$

Sample Mean:

$$x = \frac{\sum x}{n}$$

Frequency Distribution:

$$\bar{x} = \frac{\sum xf}{\sum f}$$

The mean of a frequency distribution is also the weighted mean.

2. Geometric Mean

Geometric mean is defined as our data are on growth of production over a period of years, then average growth rate is appropriately given by geometric mean. If we have two plantations where the first yielded a profit of a given percentage say 40% and the second 30 %, then the average profit percentage is given by harmonic mean. Median is a positional average. It gives the value of individual that lies some where in the middle most position of the observations. It is commonly used where the data are scores on some grading. When we talk about the size of a shirt or that of a shoe that suits a person of a given age, we are necessarily implying mode as the measure.

3. Mode

It is that value of the variable which occurs most frequently or whose frequency is maximum. Consider several samples are drawn from a population in which the important value which appears repeatedly in all the samples is called the mode. There may be no mode if no one value appears more than any other. There may also be two modes (bimodal), three modes (trimodal), or more than three modes (multi-modal).

4. **Median**

The median is the number in the middle. It is the value which divides the series into two equal parts. The data must be ranked (sorted in ascending order) first. To find the depth of the median, there are several formulas that could be used, the one that we will use is: Depth of median = 0.5 * (n + 1). In a symmetrical distribution, Mean = Median = Mode, whereas in moderately skewed or asymmetrical distribution, Mean-Mode = 3 (Mean-Median).

Measures of Dispersion/Variation

Dispersion is an attribute of the data which tells how the data points are scattered. Having defined the measure of central tendency we can easily quantify the deviation of any of the observation from a given measure of central tendency. An average of these deviations is often regarded as a measure of dispersion. However, when certain properties are to be satisfied by measures of dispersion, usually we compute the deviations of individual observations from arithmetic mean though this is not essential.

Thus, the measure of dispersion is defined as follows:

$$\textit{\textbf{Variance}}, \mathrm{V} \ = \frac{1}{N}\sum_i \left(x_i - \bar{x}\right)^2$$

and ***standard deviation***, defined as $\sqrt{\mathrm{var}\,iance}$, is taken as a measure of dispersion.

Range:

The range is the simplest measure of variation to find. It is simply the highest value minus the lowest value. RANGE = MAXIMUM – MINIMUM. Since the range only uses the largest and smallest values, it is greatly affected by extreme values, that is - it is not resistant to change.

Variance:

1. Population Variance

The deviation from the mean is squared and called the "squared deviation from the mean". This "average squared deviation from the mean" is called the variance.

$$\textit{Population Variance} = \sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

2. Unbiased Estimate of the Population Variance

One would expect the sample variance to simply be the population variance with the population mean replaced by the sample mean. However, one of the major uses of statistics is to estimate the corresponding parameter. This formula has the problem that the estimated value isn't the same as the parameter. To counteract this, the sum of the squares of the deviations is divided by one less than the sample size.

$$\textit{Sample Variance} = s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

Standard Deviation (SD):

There is a problem with variances. Think that the deviations are squared. It means that the units are also squared. To get the units back the same as the original data values, the square root must be taken.

$$\textit{Population Standard Deviation} = \sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (x - \mu)^2}{N}}$$

$$\textit{Sample Standard Deviation} = s = \sqrt{s^2} = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

The sample standard deviation is not the unbiased estimator for the population standard deviation.

Coefficient of Variation (CV):

A statistical measure of the dispersion of data points in a data series around the mean. It is calculated as follows:

$$\text{Coefficient of Variation} = \frac{\textbf{Standard Deviation}}{\textbf{Expected Return}}$$

The coefficient of variation represents the ratio of the standard deviation to the mean, and it is a useful statistic for comparing the degree of variation from one data series to another, even if the means are drastically different from each other.

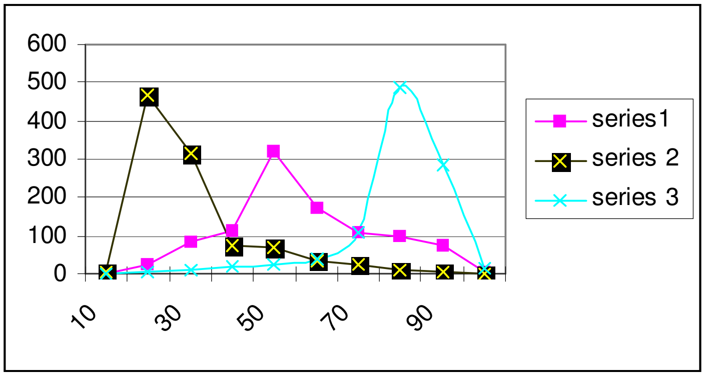
When deviations are taken from the arithmetic mean, the variance has got the property that sum of the squares of deviations is the minimum (if any other measure or value is taken as the origin the sum of the squares of deviations is more than that from mean and this is a desirable property for a measure of dispersion. Though there are few other measures of dispersion, the most commonly used one and the most appropriate in almost all cases is the standard deviation.

a	b	c	d	e
35	56	52	48	40
20	40	34	54	57
46	74	49	40	47
25	40	74	56	45
51	39	25	42	47
31	43	42	41	45
64	49	46	45	57
31	45	40	38	59
31	45	64	34	32
52	51	59	31	45
56	59	51	31	36
52	54	34	31	19
37	74	29	42	69
52	24	54	32	36
31	32	54	20	29

Parameters	**a**	**b**	**c**	**d**	**e**
Mean	41	48	47	39	44
Median	37	45	49	40	45
Mode	31	40	34	42	57
Standard Deviation (SD)	13	14	13	10	13
Range	43	50	48	36	50
Minimum	20	24	25	20	19
Maximum	64	74	74	56	69
Sum	615	723	705	585	661
Coefficient of Variation (CV)	32	29	28	25	29
Count	15	15	15	15	15

When mean, median and mode of a distribution coincide we find that the distribution is symmetric about the mean. If they don't coincide the distribution may deviate from symmetry. We say the distribution is skewed. In case of a moderately skewed distribution mean, median and mode are empirically related as Mean – Mode = 3 (Mean –Median)

So, the position of the mean, median and mode tell about this lack of symmetry.



Let us have a look at the chart above. Here, series 1 is symmetric, whereas series 2 is skewed towards left and series 3 is skewed towards right. We say series 2 is positively skewed and series 3 is negatively skewed. The coefficient of **skewness** as computed above is positive when positively skewed and negative when negatively skewed. For a symmetric distribution, skewness is zero.

We have seen above that **kurtosis** refers to the peakness in the frequency curve corresponding to heavy concentration of values at some points. Some times the vales are evenly distributed in a wide range. Both deviating from what we normally expect as series 2 in the chart above.

Series 4 is said to be leptokurtic and series 3 is platikurtic and series 2 which is normally expected is said to be mesokuritc.

Normal distribution

***Properties of Normal distribution*:**
Normal curve is symmetric. Skewnes = 0, 68.26% of values lie between $\mu \pm \sigma$, 95.44% of values lie between $\mu \pm 2\sigma$, 99.74% of values lie between $\mu \pm 3\sigma$

A normal distribution with μ =0 and SD $\sigma = 1$ is called a ***standard normal distribution***. This allows us to compare different distributions, for, by subtracting mean from and dividing by SD any distribution can be transformed into a standard normal distribution.

In Normal Distribution, Mean = μ ; *Variance* = σ^2 ; $\beta_1 = 0$; $\beta_2 = 3$

HYPOTHESIS TESTING

Definitions:
Null Hypothesis (H_0): Statement of zero or no change. If the original claim includes equality (<=, =, or >=), it is the null hypothesis. If the original claim does not include equality (<, not equal, >) then the null hypothesis is the complement of the original claim. The null hypothesis *always* includes the equal sign. The decision is based on the null hypothesis.
Alternative Hypothesis (H_1 or H_a): Statement which is true if the null hypothesis is false. The type of test (left, right, or two-tail) is based on the alternative hypothesis.
Type I error: Rejecting the null hypothesis when it is true (saying false when true). Usually the more serious error.
Type II error: Failing to reject the null hypothesis when it is false (saying true when false).
Alpha: Probability of committing a Type I error.
Beta: Probability of committing a Type II error.
Test statistic: Sample statistic used to decide whether to reject or fail to reject the null hypothesis.
Critical region: Set of all values which would cause us to reject H_0
Critical value(s): The value(s) which separate the critical region from the non-critical region. The critical values are determined independently of the sample statistics.
Significance level (alpha): The probability of rejecting the null hypothesis when it is true. alpha = 0.05 and alpha = 0.01 are common. If no level of significance is given, use alpha = 0.05. The level of significance is the complement of the level of confidence in estimation.
Decision: A statement based upon the null hypothesis. It is either "reject the null hypothesis" or "fail to reject the null hypothesis". We will never accept the null hypothesis.
Conclusion: A statement which indicates the level of evidence (sufficient or insufficient), at what level of significance, and whether the original claim is rejected (null) or supported (alternative).

Regression and Correlation

Definitions:
Coefficient of Determination: The percent of the variation that can be explained by the regression equation
Correlation: A method used to determine if a relationship between variables exists
Correlation Coefficient: A statistic or parameter which measures the strength and direction of a relationship between two variables
Dependent Variable: A variable in correlation or regression that can not be controlled, that is, it depends on the independent variable.

Independent Variable: A variable in correlation or regression which can be controlled, that is, it is independent of the other variable.
Pearson Product Moment Correlation Coefficient: A measure of the strength and direction of the linear relationship between two variables
***Regression*:** A method used to describe the relationship between two variables.
Regression Line: The best fit line.
Scatter Plot: An plot of the data values on a coordinate system. The independent variable is graphed along the x-axis and the dependent variable along the y-axis
Standard Error of the Estimate: The standard deviation of the observed values about the predicted values

Often we have occasions to collect information on more than one variable on an individual. For example we may collect data on height of the plant, girth of the plant, number of fruits in a bunch etc from on a plant. These lead to bivariate or multivariate distributions.

Suppose we are interested in the interrelationship of two such variables. The relationship may be of the following types:

1. One of the variables may be the cause of the other. Example is yield and dose of fertilizer. We know yield depend on dose of fertiliser.
2. There may be a functional relationship between the two. Example is height and girth of a plant.
3. One variable may explain the variation in the other. Example: the levels of rainfall may explain the variation in yield of the same variety of a crop in different parts of the country.

***Correlation*:**

In a bivariate distribution as above, suppose there is no reason to believe that there exists a cause and effect relation ship but yet the variables may be associated. A statistic that measures this association ship is correlation coefficient ‘ r’ defined as

$$r = \frac{\sum xy}{\sqrt{(\sum x^2 \sum y^2)}}$$

The test of significance of ‘r’ is provided by

$$t = \frac{r(\sqrt{(n-2)}}{\sqrt{(1-r^2)}} \quad (\text{ H}_o\text{: population correlation coefficient =0})$$

t follows t distribution with (n-2) d.f. ‘r^2’ is called coefficient of determination and it gives the strength of the association between the variables.

Regression:

The idea behind regression is that when there is significant linear correlation, you can use a line to estimate the value of the dependent variable for certain values of the independent variable.

The regression equation should only used

- When there is significant linear correlation. That is, when you reject the null hypothesis that rho=0 in a correlation hypothesis test.

- The value of the independent variable being used in the estimation is close to the original values. That is, you should not use a regression equation obtained using x's between 10 and 20 to estimate y when x is 200.
- The regression equation should not be used with different populations. That is, if x is the height of a male, and y is the weight of a male, then you shouldn't use the regression equation to estimate the weight of a female.
- The regression equation shouldn't be used to forecast values not from that time frame. If data is from the 1960's, it probably isn't valid in the 1990's.

Assuming that you've decided that you can have a regression equation because there is significant linear correlation between the two variables, the equation becomes: $y' = ax + b$ or $y' = a + bx$ (some books use y-hat instead of y-prime). The Bluman text uses the second formula, however, more people are familiar with the notion of $y = mx + b$, so I will use the first.

a is the slope of the regression line: $$a = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

b is the y-intercept of the regression line: $$b = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$$

The regression line is sometimes called the "line of best fit" or the "best fit line".

Since it "best fits" the data, it makes sense that the line passes through the means.

The regression equation is the line with slope *a* passing through the $(\bar{x}, \bar{y})$ point

Another way to write the equation would $y' - \bar{y} = a(x - \bar{x})$ be

$$y' - \bar{y} = ax - a\bar{x}$$
$$y' = ax + \bar{y} - a\bar{x}$$

apply just a little algebra, and we have the formulas for a and b that we would

$$\boxed{a = \frac{SS(xy)}{SS(x)} \quad \text{and} \quad b = \bar{y} - a\bar{x}}$$

It also turns out that the slope of the regression line can be written as.

$$a = r \cdot \frac{s_y}{s_x}$$

Since the standard deviations can't be negative, the sign of the slope is determined by the sign of the correlation coefficient. This agrees with the statement made earlier that the slope of the regression line will have the same slope as the correlation coefficient.

F-TEST

Definitions:

F-distribution: The ratio of two independent chi-square variables divided by their respective degrees of freedom. If the population variances are equal, this simplifies to be the ratio of the sample variances.

Analysis of Variance (ANOVA): A technique used to test a hypothesis concerning the means of three or more populations.

One-Way Analysis of Variance:

Analysis of Variance when there is only one independent variable. The null hypothesis will be that all population means are equal, the alternative hypothesis is that at least one mean is different.

Between Group Variation: The variation due to the interaction between the samples, denoted SS(B) for Sum of Squares Between groups. If the sample means are close to each other (and therefore the Grand Mean) this will be small. There are k samples involved with one data value for each sample (the sample mean), so there are k-1 degrees of freedom.

Between Group Variance: The variance due to the interaction between the samples, denoted MS(B) for Mean Square Between groups. This is the between group variation divided by its degrees of freedom.

Within Group Variation: The variation due to differences within individual samples, denoted SS (W) for Sum of Squares Within groups. Each sample is considered independently, no interaction between samples is involved. The degrees of freedom is equal to the sum of the individual degrees of freedom for each sample. Since each sample has degrees of freedom equal to one less than their sample sizes, and there are k samples, the total degrees of freedom is k less than the total sample size: df = N - k.

Within Group Variance: The variance due to the differences within individual samples, denoted MS (W) for Mean Square Within groups. This is the within group variation divided by its degrees of freedom.

Scheffe' Test: A test used to find where the differences between means lie when the Analysis of Variance indicates the means are not all equal. The Scheffe' test is generally used when the sample sizes are different.

Tukey Test: A test used to find where the differences between the means lie when the Analysis of Variance indicates the means are not all equal. The Tukey test is generally used when the sample sizes are all the same.

Two-Way Analysis of Variance:

An extension to the one-way analysis of variance. There are two independent variables. There are three sets of hypothesis with the two-way ANOVA. The first null hypothesis is that there is no interaction between the two factors. The second null hypothesis is that the population means of the first factor are equal. The third null hypothesis is that the population means of the second factor are equal.

Factors: The two independent variables in a two-way ANOVA.

Treatment Groups: Groups formed by making all possible combinations of the two factors. For example, if the first factor has 3 levels and the second factor has 2 levels, then there will be 3x2=6 different treatment groups.

Main Effect: The effects of the independent variables

Interaction Effect: The effect one factor has on the other factor

One-way Analysis of Variance (ANOVA):

A One-Way Analysis of Variance is a way to test the equality of three or more means at one time by using variances.

Assumptions:

- The populations from which the samples were obtained must be normally or approximately normally distributed.
- The samples must be independent.
- The variances of the populations must be equal.

Hypotheses: The null hypothesis will be that all population means are equal, the alternative hypothesis is that at least one mean is different.

In the following, lower case letters apply to the individual samples and capital letters apply to the entire set collectively. That is, n is one of many sample sizes, but N is the total sample size.

Grand Mean:

The grand mean of a set of samples is the total of all the data values divided by the total sample size. This requires that you have all of the sample data available to you, which is usually the case, but not always. It turns out that all that is necessary to find perform a one-way analysis of variance are the number of samples, the sample means, the sample variances, and the sample sizes.

$$\bar{X}_{GM} = \frac{\sum x}{N}$$

Another way to find the grand mean is to find the weighted average of the sample means. The weight applied is the sample size.

$$\bar{X}_{GM} = \frac{\sum n\bar{x}}{\sum n}$$

***Total Variation*:**

The total variation (not variance) is comprised the sum of the squares of the differences of each mean with the grand mean.

$$SS(T) = \sum (x - \bar{X}_{GM})^2$$

There is the between group variation and the within group variation. The whole idea behind the analysis of variance is to compare the ratio of between group variance to within group variance. If the variance caused by the interaction between the samples is much larger when compared to the variance that appears within each group, then it is because the means aren't the same.

***Between Group Variation*:**

$$SS(B) = \sum n(\bar{x} - \bar{X}_{GM})^2$$

The variation due to the interaction between the samples is denoted SS(B) for Sum of Squares Between groups. If the sample means are close to each other (and therefore the Grand Mean) this will be small. There are k samples involved with one data value for each sample (the sample mean), so there are k-1 degrees of freedom.

The variance due to the interaction between the samples is denoted MS (B) for Mean Square Between groups. This is the between group variation divided by its degrees of freedom. It is also denoted by s_b^2 .

***Within Group Variation*:**

$$SS(W) = \sum df \cdot s^2$$

The variation due to differences within individual samples, denoted SS (W) for Sum of Squares Within groups. Each sample is considered independently, no interaction between samples is involved. The degrees of freedom is equal to the sum of the individual degrees of freedom for each sample. Since each sample has degrees of freedom equal to one less than their sample sizes, and there are k samples, the total degrees of freedom is k less than the total sample size: df = N - k.

The variance due to the differences within individual samples is denoted MS(W) for Mean Square Within groups. This is the within group variation divided by its degrees of freedom. It is also denoted by s_w^2. It is the weighted average of the variances (weighted with the degrees of freedom).

F- test statistic:

Recall that a 'F' variable is the ratio of two independent chi-square variables divided by their respective degrees of freedom. Also recall that the F test statistic is the ratio of two sample variances, well, it turns out that's exactly what we have here. The F test statistic is found by dividing the between group variance by the within group variance. The degrees of freedom for the numerator are the degrees of freedom for the between group (k-1) and the degrees of freedom for the denominator are the degrees of freedom for the within group (N-k).

$$\boxed{F = \frac{s_b^2}{s_w^2}}$$

***Summary Table*:**

All of this sounds like a lot to remember, and it is. However, there is a table which makes things really nice.

	SS	df	MS	F

Between	SS (B)	k-1	SS(B)
			----------- 	
			k-1
				MS(B)
				-------------- 	
				MS(W)

Within	SS (W)	N-k	SS(W) ----------- N-k	.

Total	SS (W) + SS (B)	N-1	.	.

Notice that each Mean Square is just the Sum of Squares divided by its degrees of freedom, and the F value is the ratio of the mean squares. Do not put the largest variance in the numerator, always divide the between variance by the within variance. If the between variance is smaller than the within variance, then the means are really close to each other and you will fail to reject the claim that they are all equal. The degrees of freedom of the F-test are in the same order they appear in the table

***Decision Rule*:**

The decision will be to reject the null hypothesis if the test statistic from the table is greater than the F critical value with k-1 numerator and N-k denominator degrees of freedom.

If the decision is to reject the null, then at least one of the means is different. However, the ANOVA does not tell you where the difference lies. For this, you need another test, either the Scheffe' or Tukey test.

Experimentation

An *experiment* deliberately imposes a *treatment* on a group of objects or subjects in the interest of observing the response. This differs from an *observational study*, which involves collecting and analyzing data without changing existing conditions. Because the validity of a experiment is directly affected by its construction and execution, attention to *experimental design* is extremely important. Treatment: In experiments, a *treatment* is something that researchers administer to experimental units. For example, a corn field is divided into four, each part is 'treated' with a different fertiliser to see which produces the most corn; a teacher practices different teaching methods on different groups in her class to see which yields the best results; a doctor treats a patient with a skin condition with different creams to see which is most effective. Treatments are administered to experimental units by 'level', where level implies amount or magnitude. For example, if the experimental units were given 5mg, 10mg, 15mg of a medication, those amounts would be three levels of the treatment. Factor: A *factor* of an experiment is a controlled independent variable; a variable whose levels are set by the experimenter**.** A factor is a general type or category of treatments. Different treatments constitute different levels of a factor. For example, three different groups of runners are subjected to different training methods. The runners are the experimental units, the training methods, the treatments, where the three types of training methods constitute three levels of the factor 'type of training'.

Experimental Designs

When we design an experiment we must have the statistical analysis in mind. The experimenter who has no basic knowledge of statistical technique generally performs the experiment such that not possible to draw any valid conclusion from the data or from the data these is very little extract any further information. Hence it is necessary that the experiment should be planned in advance.

***Some Definitions*:**

i) **Treatments**: The object of comparison which the experimenter or researcher has to try out in the field for assessing the value are known as treatments e.g.: varieties of seed treatment, insecticide, cultivation practices, different methods of seed treatments are the examples of field treatment.

ii) **Experimental material:** This is the name given to denote the material on which the experiment is performed. For example, field, soil, pot culture, patients in the hospital, herds of cow, bulls, varieties of seed etc.

iii) **Experimental error:** In field experiment we may come across variation due to the treatments and due to extraneous or inherent variation which can't altogether be eliminated but can be reduced to certain extent. This inherent variation is called experimental error.

iv) **Treatment effects:** In research study when we measure the impact of certain treatments or a thing or a known object, we get the information about that particular treatment. This is due to the typical technique of measurement must we adopted and originally is known as Treatment effect.

The choice of the experimental design depends upon the number and nature of the experiment under study. It also depends on the objects of the experiment. Another consideration is the available resources. Particular designs are to be selected depending on the situations.

Types of designs:

i) **Completely Randomized Design:** It is appropriate when the Experimental material is limited and homogeneous. Soil in the plot experiment.

ii) **Randomized Block Design:** This can be of use when the fertility gradient of the field is of one direction. It may be adopted up to 20 treatments without an appreciable loss of efficiency.

ii) **Latin Square Design:** Can be used when the fertility gradient is in two directions instead of one. It may be adopted fir number of treatments ranging from 5 to 8 or from 5-12 at the most.

iv) **Factorial Experiment:** When there are several factors with different levels to be studied simultaneously with the same precision, factorial scheme may be adopted.

i) **Split Plot Design:** When the factors are like that some of them requires large plots, like irrigation, depth of ploughing, sowing dates etc. and may be studied with different precision, Split Plot Design may be used.

ii) **Incomplete Block Design:** When the number of treatments to be tried are more it is difficult to keep variation between each treatment and between replication, in that case error variance will be more. To overcome this difficulty incomplete block design can be adopted.

Principles of designs:

The three basic principles of experimental design are Randomization, Replication and Local control. It is prime important for the investigation to understand the logic of these principles in order to design effective equipments.

i) *Replication*: The repetition of treatments under investigation is known as Replication. Since the variation in fertility cannot be allowed for directly owing to its unpredictable nature, the experimenter seeks to average out its influence over the different treatments by replication.

ii) *Randomization*: The principle of randomization is unique in modern experiment. The allocation treatment to the different plots by a random process is known as randomization of the treatments. The statistical procedure employed in making comparisons between treatments will hold good only provided the treatments are allocated randomly to various plots.

iii) *Local control*: Local control involves the principle of homogeneity. Since, a lower experiment error helps in detecting the smaller real difference between the treatments; it is desirable that it should be reduced as far as possible. The reduction of experimental error can be achieved by making use of the fact observed earlier that adjacent areas in field are relatively homogeneous than those separated widely.

Analysis of Variance (ANOVA)

A. CRD (with equal replications)

Source of variation	***Degrees of freedom (d.f)***	***Mean Square***
Treatments	(t-1)	
Error	t (r-1)	s^2
Total	tr-1	

B. CRD (If treatments have different replications, r1, r2, r3,….rn)

Source of variation	***Degrees of freedom (d.f)***	***Mean Square***
Treatments	(t-1)	
Error	(n-t) [(n-1) – (t-1)]	s^2
Total	(n-1)	

C. RBD

Source of variation	***Degrees of freedom***	***Mean square***

	(d.f)	*(m.s)*
Blocks	b-1	
Treatments	t-1	
Error	(b-1)(t-1)	s^2
Total	n-1	

D. LSD

Source of variation	*Degrees of freedom (d.f)*	*Mean square (m.s)*
Rows	t-1	
Columns	t-1	
Treatments	t-1	
Error	(t-1) (t-2)	s^2
Total	$(t^2 -1)$	

D. Factorial Experiment (Two factors)

Source of variation	D.F.	M.S.
Blocks	**(r-1)**	
Main effect (A)	**(a-1)**	
Main effect (B)	**(b-1)**	
Interaction (AB)	**(a-1)(b-1)**	
Error	**(r-1)(ab-1)**	s^2
Total	**(abr-1)**	

E. Split-plot Experiment

Source of variation	D.F.	M.S.
Main plot		
Blocks	**(r-1)**	
Main plot (A)	**(a-1)**	
Main plot Error (a)	**(a-1) (r-1)**	s^2_a
Main plot total	**ra – 1**	
Sub-plot		
Sub-plot (B)	**(b-1)**	

Interaction (AB)	**(a-1)(b-1)**	
Sub-plot Error (b)	**(r-1) a(b-1)**	s^2_b
Total	**(rab -1)**	

<u>Terminologies:</u>

$$\text{Standard error of Mean (S E mean)} = \sqrt{\frac{\text{Error mean square}}{\text{r}}}$$

$$\text{Standard error of difference (S Ed)} = \text{SEm} \ \ \text{x} \ \sqrt{2}$$

$$\text{Standard deviation (SD)} = \sqrt{\text{EMS}}$$

Coefficient of variation (CV %) = $\dfrac{S.D}{Grandmean} X100$

Grand Mean = Grand total / Total no. of observations

Critical Difference (C.D.) or Least significant difference (LSD)

= t (0.05, error df) x SEd

<u>Note</u>:

IF > 10 treatments – DMRT

IF < 10 treatments – LSD/CD

Minimum degrees of freedom (Error) for experiment: 12

<u>C.D.</u>: Agriculture (field) = 0.5% ; Agriculture (Lab.) = 0.1%

Standard Error of Difference (SEd)	**Design**
$\sqrt{\dfrac{2\ EMS}{r}}$ replication	CRD with equal
$\sqrt{EMS \quad (\dfrac{1}{r1} + \dfrac{1}{r\,2})}$ replication	CRD with unequal

RBD & LSD

$$\sqrt{\frac{2\ EMS}{r}}$$

Factorial Experiment:

$$\sqrt{\frac{2EMS}{rxlevel.of.other.factor(q)}}$$ Factor - A

$$\sqrt{\frac{2EMS}{rxlevel.of.other.factor(p)}}$$ Factor-B

$$\sqrt{\frac{2\ EMS}{r}}$$ Interaction (A x B)

Where,
Factor-A = ‘p’ levels
Factor-B = ‘q’ levels
No. of replications = ‘r’

Split-plot Experiment:

$$\sqrt{\frac{2\ E_a MS}{rs}}$$ Two main-plot treatments

$$\sqrt{\frac{2\ E_b MS}{rm}}$$ Two Sub-plot treatments

$$\sqrt{\frac{2\ E_b MS}{r}}$$ Main x two sub-plot treatments

$$\sqrt{\frac{2[E_a MS + (s-1)E_b MS}{rs}}$$ Sub x two main-plot treatments

Where,

Main plot (A)	= ‘m’ levels
Sub-plot (B)	= ‘s’ levels
Main plot & Replication	= E (a)
Interaction (A x B)	= m x s
Sub-plot & Interaction (A x B)	= E (b)

No. of replications = ‘r’

Statistical Softwares

MSTAC – C	Michigan State University (DOS Version)
SPSS	Statistical Software for Social Sciences
SAS	Statistical Analytical System
GENSTAT	Reading University, U.K.
IRRI – STAT	IRRI, Philippines
SYS STAT	Modelling
SPAR	Statistical Package for Agricultural Research, IASRI, New Delhi
INDOSTAT-C	Hyderabad
MS – Excel	MS Office
Microstat	DOS Version

Some of the popular Softwares useful in descriptive statistics are **AXUM, EXCEL** and **Harvard Graphics (HG)**. These softwares help us to present the data in a very impressive and informative way and interpretation of the large mass of data become easier for the Research worker.

References:

1. Panse, V.G. and Sukhatme, P.V. (1991). Statistical Methods for Agricultural Workers. ICAR, New Delhi.
2. Chandel, S.R.S. (1999). A HandBook of Agricultural Statistics. Achal Prakashan Mandir, Kanpur, Uttar Pradesh

Horticultural Crops as Source of Nutrients, Flavours, Natural Food Colors and Pigments

K.N. Shiva

Introduction

INDIA enjoys an enviable position in the world horticulture. India has now emerged as a major producer of horticultural crops and is presently producing about 60 million tonnes of fruits and about 120 million tonnes of vegetables. It is the largest producer of spices, coconut, arecanut and cashewnut and the second largest producer of fruits and vegetables in the world. Similarly, India is the highest producer of mango, banana, Sapota and acid lime and has recorded highest productivity in grape. Diversification and value- addition will be the key words in the Indian horticulture in the 21st century.

HORTICULTURE consisting of fruits, vegetables, spices, floriculture, medicinal and aromatic plants and plantation crops, as a pleasantry before independence, has emerged as a fast moving sector in the Indian Agricultural economy. Horticulture has emerged as the best option for diversification in agriculture for nutritional security, employment generation, and enhancing farm income. National goal of achieving 4.0% growth in agriculture, can be achieved through the major contribution in growth form horticulture, i.e. 6.0 %. It provides avenues for diversification, enhanced returns per unit area, better land and water-use and opportunities for employment generation. The wide range of horticultural crops provide ample opportunities to farmers to adopt multi-layer cropping for minimizing risk of crop failure, maximizing their farm income (Singh, 2007).

Due to poverty, malnutrition, is posing a threat to vulnerable section in Asia and the Pacific regions. This is manifest in the form of vitamin A deficiency (VAD), iron deficiency anemia (IDA) and iodine deficiency disorders. The first two could be minimized through horticulture intervention and awareness drive.

Horticultural Crops as Source of Nutrients:

In human nutrition, fruits and vegetables play an important role towards making a balanced diet. To some extent, they provide energy rich food. Fruits and vegetables are composed of a number of substances, besides nutritive components. Fruits and vegetables are rich source of vitamins and minerals. Hence, they are known as ‘protective foods’. The chief components are water, nutritive components (carbohydrates, proteins, fats, minerals and vitamins), pigments, flavour compounds, polyphenols, free organic acids, enzymes, etc. Most of the fruits contain malic and citric acids. Tartaric acid is rich in grapes and tamarind, while oxalic acid in carambola and bilimbi. Horticultural crops are good source of natural food colors and pigments too.

Banana, jackfruit, anona, sapota and fig contain carbohydrates in the range 10 to 24 per cent and are good sources of energy comparable t potato, colacasia, tender maize, yam and green peas (15.9 to 24.6 per cent carbohydrates and 79 to 125 kcal energy). Closely following this group of fruits as good sources of energy are mango, litchi, grapes, ber, pomergranate, phalsa and jamun. Fresh avocado is the only highest energy yielding fresh fruit yielding 161 to 215 kcl per 100 g of edible portion due to its high fat content (15-26%). But, fruits and vegetable furnish 90% of the vitamin C and 60% vitamin A in the world. Mango and papaya are rich in pro-vitamin A, while aonla and guava are rich in vitamin C. Banana is a good source of carbohydrate, among fruits. Fruits yield larger quantities of food per hectare compared to cereals. For example paddy yield (max.) is 3 tonnes/ha whereas it is 22 tonnes/ha in case of banana; 45 tonnes/ha in case of pineapple and 90 tonnes/ha in case of grapes. Much less area is required to obtain the calorific requirement per adult per year (11,00,000 kcal) from growing banana (0.03 hectare) or mango (0.16 hectare) than from growing wheat (0.44 hectare). Horticultural crops in general are poor sources of protein as they contain less than 2% protein. However, mushrooms and vegetable meat are excellent source of proteins. Fruits are rich sources of organic acids like citric acid in fruits and tartaric acid in grapes, which stimulate appetite and help in digestion. Papaya contains protein digesting enzymes. Manu fruits and vegetables possess laxative property due to the presence of dietary fiber and pectin, which stimulate intestinal activity.

Vegetables and fruits appear to be playing a prominent role in prevention of several chronic diseases such as heart diseases, cancer, cataract, osteoporosis, diabetes, etc. The active constituents responsible for this property have been shown to be a number of nutrients, phytochemicals and fiber. Apart from micronutrient related functions, the bioactive phytochemicals prevent degenerative processes by antioxidant activity. In order to have protective effect, it is necessary to consume 400-600 g of fruits and vegetables every day. But, the consumption level f fruits is low and widely variable from region to region in India. Fruit consumption level is as low as 1 g/day/person in the states of Manipal and Nagaland to 70 g/day/person in Union territory of Chandigarh. An increase in the intake of fruits along with vegetables will meet the RDA's of many nutrients.

Bioactive components such as carotienoids from tomatoes, glucosinolates from Brassica vegetables, phytoestrogens from soybean and phenolics and antioxidants from various plants protect and present numerous disease such as cancer and cardiovascular disease. These antioxidants are also associated with slowing the aging process and improving overall health. Phytoestrogens from *Dioscorea* species such as wild yam (*Diascrorea villosa)* also possess numerous physiological benefits. Some sweet potato varieties have anti-diabetic potential. Most of these bioactive components are plant secondary metabolites.

Type of vegetables	Examples
Roots	Carrot, beet root, radish, turnip, colocasia Tubers- potatoes, corm, topica, yam
Bulbs	Onions, garlic, leeks
Leaves	Spinach, cabbage, lettuce, amaranth, colocasia leaves
Flowers	Plantain(raw banana) flowers, broccoli
Fruits	Tomatoes, brinjal , pumpkin, cucumber, gourds, capsicum, plantain
Legumes (beans)	Peans, French beans, broad beans, green gram, drumstick
Stems	Ginger, celery stem, drumstick stem
Sprouts	Pulse sprouts, soyabean
Fungi	Mushrooms
Perennial Vegetables	Asparagus, artichokem rhubarb
Herb vegetables	Tulsi, coriander leaves, mint leaves

Nutrient	Green leafy Vegetables	Roots and Tubers	Other Vegetables
Calorie	Not significant	Better than others as they contain more starches	Poor
Carbohydrates	Not in significant amount	High in complex carbohydrates	Fair sources of fibre
Protein	Poor (exceptions agathi and drumstick)	Poor	Leguminous vegetables high in protein
Fats	Poor	Poor	Poor
Vitamins	Beta- carotene	Beta-carotene	Fair Sources of

	(Specially in coriander leaves, darker leaves, outer cabbage leaves) Vitamins C, folic acid	(Carrots, although less than green leaf vegetables), fair in vitamin C. poor in Beta- vitamins	vitamin C (specially capsicum)
Minerals	Calcium and iron-rich but not properly utilized since present in bound form (drumstick leaves and beet greens excellent sources of iron)	Poor mineral	Plantain greens have high amounts of iron, otherwise not very good sources of minerals
Miscellaneous	Withered greens are low in nutrients, carotene content is influenced by seasons (courtesy NIN, Hyderabad), agathi most nutritious, amaranth high in lignin and cellulose	Contain more starch and hence contribute more calories	Highly perishables since they have a high moisture content, small bitter ground more nutritious than other ones.

Genetically modified vegetables with quality characters

Vegetable	Characters
Broccoli	**Prolonged shelf life**
Carrot	**Improved nutrition**
Cassava	**Improved nutrition**
Lettuce	**Brown spot resistance, delayed senescence**
Melon	**Delayed ripening**
Peppers	**Delayed ripening, increased shelf life**
Potato	**Increased solids, increased carbohydrates,, improved nutritional quality, increased tyrosine, altered amino acids, black spot bruise resistance, decreased glyuco-alkaloids**
Sweet potato	**Altered amino acid composition, protein quality improved**
Tomato	**Delayed ripening, increased starch and solids, altered sugar profile, pigments altered, altered carotenoid content, protein quality altered, increased antioxidant enzymes**

Horticultural Crops as Source of Flavours:

Flavour is a sensory phenomenon having a combination of the sensations of taste, odour or aroma, heat and cold and texture or mouthfeel. Natural flavouring materials such as spices, essential oils and fruit juices have been used in food preparations.

Fruits	Volatile flavour components
Apple	Esters, alcohols, aldehydes, ketones, acids, hexenal, ethyl-2-methyl butyrate
Banana	Alcohols, esters, amyl acetate, butyl butyrate
Orange	Acetaldehyde, ethanol, limonene, esters, linalool
Peach	Benzaldehyde, benzyl alcohol
Pear	Esters of 2, 4-decadienoic acid, esters of ethyl, n-propyl and n-butyl
Pineapple	p-allylphenol, methyl esters of bita-hydroxyl butyric acids
Strawberries	Methyl and ethyl esters, propionates and butyrates

Vegetables	Flavour component
Beans	Methanol, acetone, hydrogen sulphide
Corn	Methanol, acetone, hydrogen sulphide
Peas	Methanol, acetone, dimethyl sulphide
Brassicaea group (Cabbage, cauliflower, broccoli, etc.)	s-methyl l-cysteine sulphoxide
Allium group (Onion, garlic, leek)	Diallyl di-sulphide

Bitterness in bitter gourd is due to momordicians, while it is due to cucurbitacians oxygenated tetracyclic terpenes in cucumber and pumpkin. These may occur either in free form or as glycosides.

Horticultural Crops as Source of Natural Food Colors and Pigments:

Pigments of plant origin are gaining importance globally as a potential source of natural food colors for its versatile uses and to combat a variety of health hazards caused by synthetic colors. Color has been added to our foods in one or another form for centuries. Color is an important element in enhancing delicacy of the food stuffs, constituting one of the major dietary additives. The color is contributed by various kinds of components in the food; some of them occurring naturally, while others are produced by heating or processing and the rest are added to alter the colors. The food colors can be primarily classified into two types *viz.,* (i) Natural colors and (ii) Synthetic colors (include permitted and non-permitted colors) (Indu Rani *et al.*, 2004).

Natural food color can be defined as any dye, pigment or any offer substance obtained from plant, animal, insect, algae, mineral or source capable of coloring food drug, cosmetic or any part of human body (www.foodaddtivesworld.com/natural-food-colors html). The colors are extracted from a variety of sources such as seeds, fruits, bark, leaf, root, stem, wood flower, rhizome, whole plant, etc. by conventional methods. Being biological origin, they are often called as 'Bio Colors' (Indu Rani *et al.*, 2004). The natural color of foods may be done to the presence of natural pigments such as carotenoids, chlorophylls, myoglobins and anthocyanins and chemical modification during the processing of natural constituents of foods, e.g. caramelization and color additives (Parkinson and Brown, 1981). A suitable natural color can be developed by manipulation of certain factors such as pH, heat, light, storage and other ingredients. Food elements are used in the range of 10-500 ppm in food and beverages (Coulson, 1980). According to the United States Food and Drug Administration (FDA), color pigments having natural biotical origin are exempted from certification (in 21

CFR, part 73 for food, during and cosmetics). There are 26 colors permitted to be used in food and 28 to be used in cosmetics and pharmaceuticals.

Some of the well-known dyes used in ancient times were: a red dye obtained from the roots of *Rubia tinctorium* L.; blue indigo dye from the leaves of *Indigofera tinctoria* L.; yellow dye from the stigma of *Crocus sativus* L. as well as from the rhizomes of *Curcuma longa* L. (Siva, 2007).

Pigment	Colour	Examples
Chlorophyll	Green	Guava, Country apples, amla, unripe fruits
Carotenoids	Yellow-orange	Mango, Jackfruit, papaya, peaches, pineapple
Lycopene	Red orange	Tomatoes
Anthocyanins	Purple	Grapes, blue berries, plums
Anthoxanthins	Purplish brown	Apple, gooseberry, pears, custard, apple

Carotenoids	Anthocyanins
Yellow-orange in colour	Purple blue in colour
Insoluble in water	Soluble in water
Not affected by acid or alkali	Change colour on reacting with alkali
Mangoes, oranges, papaya	Grapes, blue berries

All green coloured vegetables contain chlorophyll:

Eg. Capsicum, beans, Green leaf vegetables, chilles. Peas, amaranth, coriander, mint, etc.

Carotenoids are present in various forms in the vegetables:

Yellow corn – cryptoxanthin
Tomatoes - lycopene, beta-carotene
Red capsicum - cryptoxanthin, violaxanthin, capsorubin,
Capxanthin
Green capsicum - lutein, beta-carotene, violaxanthin
Carrots - alpha and beta carotene, lycopene, xanthophylls

Water soluble pigments: Pigments that can be easily dissolved in water are also termed as flavonoids. Anthocyanins are highly water soluble with a colour range from red to purple, anthoxanthins, on the other hand, are colourless or white.

Anthocyanins	Anthoxanthins
Water soluble	Water soluble
Red to purple in colour	White or colourless to pale yellow, depending on the pH
Disperse felly in the cell	Disperse feely in the cell
Contained in the vacuole of the cell	Contained in the vacuole of the cell
Cherries, grapes, red apples, pomegranates, Currants have these pigments	Cauliflower, onions, and some green leafy vegetables.
Anthocyanidins (without sugars) *Pelargonidin – red* *Cyanidin – reddish blue* *Delphinidin - blue*	Present as complex mixture and widely distributed in all plants

Plants as source of natural dye/color

Among the biological sources, many natural dyes/colors are obtained mainly from plants, offering various colors like red, yellow, blue, black, brown and a combination of these. Colors obtained from plant origin are eco-friendly and non-toxic, while adding to food products, hence preferred over synthetic colors. Over 2000 pigments have been reported to be produced by various parts of plant, of which only above 150 have been exploited commercially (Siva, 2007). In India alone about 450 are known to yield colors (Chandramouli, 1995), of which 50 (ten from roots, four from barks, five from leaves, seven from flowers, seven from fruits, three from seeds, eight from wood and three from gums and resins) are considered as important (Siva, 2003). Recently, plants yielding food colors have been reviewed by Shamina *et al.* (2007). Plant pigments that provide natural colors to food can be grouped into four primary classes based on their chemistry *viz.*, chlorophylls, carotenoids, flavonoids and betalaines.

Plant pigments providing natural food colors

Color	Responsible pigment	Source of presence	
		Cell component	Plant part
Green	Chlorophyll (a, b, c & d forms)	Chloroplast	Leaf
Red, orange and yellow to brown	Carotenoids	Chloroplast & Chromoplast	Leaf, bark/wood, flower, fruits, root, seed, tuber
Yellow, red, blue & orange	Flavonoids	Cytosol & vacuole	Flower, fruits
Yellow, orange, red & violet	Betalaines	Cytosol & vacuole	Flower, fruits & other parts (as occurring in beet root)

Among the pigments, green color is the dominant color in the plants, while the carotenoids are a large group of pigments associated with chlorophyll and responsible for autumn leaf pigmentation. Many of the intense colors in flowers and fruits are contributed by the flavonoid pigments and closely related compounds with a diverse range of colors, which are due to structural differences between compound and the relative concentration of specific pigments within the cells. Betalains, the restricted group of pigments, gets its name from the red-violet pigment isolated for the first time in crystalline form from the root of the beet, *Beta vulgaris*, L. (Kays, 1998).

In this context, Horticultural plants are the major sources for providing various natural colorants and pigments. These colorants and pigments are having wide range of applications/uses such as coloring food, beverages, soft drinks, confectionery, bakery products, etc. Some of the important natural food color/pigments yielding horticultural plants along with its origin and distribution are listed in the following Table.

Horticultural crops as sources of natural food colors, pigments and its applications/uses

Color	Source	Applications/uses
Chlorophylls/chlorophyllins		
Natural green & green	Nettles, grass, Celery, Green beans and peas, Green olives, Parsley, Spinach, Green turnips, Asparagus, Bell peppers, Broccoli, Brussels sprouts, Green cabbage, herbs	Pastas, confectionery, medicines, processed food, vegetable oils, delicatessen, spice preparations, ice cream, coloring materials,
Carotenoids and Xanthophylls		
Natural orange	Mushrooms	Tanning pills, fruit-spreads, candies, syrups, sauces, carbonated drinks
Yellow and red	Gardenia	Confectionery, agricultural products, and processed marine products
Yellowish orange	Turmeric, Curcumin/CU-Chloro	Beverages, sauces and confectionery, desserts, ice cream
Yellow-orange to red-orange	Paprika Extracts / Capsanthin (paprika pod) and paprika leaf extract	Beverages, processed foods, tomato products
Yellow-orange to red-orange	Paprika Oleoresin (paprika pod)	Beverages, processed foods such as sausage, dressings, dry soluble seasonings, food coatings, and snack food seasonings, tomato products
Yellowish orange	Carrot	Sorbets, beverages, confectionery, sauces
Yellowish orange	Saffron	Confectionery, agricultural products, and processed marine products
Yellow – orange	Vegetables	Baby foods, cereals, sauces, processed cheese, fruit drinks
Yellow	Ginger	Food products, beverages
Red, yellow, and orange	Carrots, sweet potatoes, spinach, and tomatoes	Food products, sauces, chutneys
Yellow	Paprika	Beverages, processed food, tomato products
Bright yellow	Marigold	Food colorants
Orange	Parsley, sweet pepper	Food products
Orange yellow	Annatto	Ice-cream and dairy products, bakery and snack foods, desserts, custard powders, cereal products, confectionery, coloring butter, citrus juices, concentrates, drinks, candies, etc.
Orange-yellow	Turmeric, saffron, parsley	Food products, cosmetic and pharmaceutical products
Orange to red	Citrus fruit skin, vegetable pulp	Food products, cosmetic and pharmaceutical products
Red to reddish orange	Paprika, red pepper, saffron, mustard	Beverages, processed food, tomato products
Red, dark red, purple red	Paprika, red pepper	Beverages, processed food, tomato products

Red to dark red	Red palm oil	Food products
Dark red	Paprika, parsley	Beverages, processed food, tomato products
Pink	Pink edible mushroom	Food products
Greenish yellow to yellow	Stinging Nettle	Ice cream, delicatessen, baked goods, fruit preparations, foods such as nettle soup and nettle cheese
Brown to reddish brown	Onion	Confectionery, stock farm and processed marine products, coloring of seasonings and sauce
Brown	Cacao	Coloring incorporated in candy or cake, chocolate, biscuits, and other products
Brown	Cinnamon	Preserving the food articles
Betalains		
Yellow to orange	Beet root	Beverages, frozen foods, fruit fillings, candies, baked goods
Bluish-red	Beet juice colors (Beet root)	Fruit preparations, condiments sauces, fillings and candies, condiments, gelatin products, certain powdered beverage products, dairy products
Red-violet and yellow	Red beet-root	Beverages, frozen foods, food stuffs, fruit fillings, chewing gums, candies, baked goods, medicinal products, coloring red to soybean products
Flavonoids / Anthocyanins		
Red, purple and blue	Strawberries, grape skin, black grapes blueberries, raspberries, red perilla	Confectionery, food products and dessert products
Red	Red cabbage	Coloring chewing gum and vegetable juice, making drinks free-from nasty smell, and sedimentation,
Red	Red-fleshed potato	Colorant used as an additive to foodstuffs, beverages, pharmaceuticals, toiletries etc.
Deep red	Elderberry	Beverages, fruit confectionery, sorbets and sauces, food colorant, desserts, soft drinks
Bright red-orange to strawberry red	Chokeberry or Red Fruit (Aronia)	Jelly making, candies, pie and cookie fillings, yogurt, sorbet, flavored milk
Bright red & purple	Hibiscus	Soft drinks and alcoholic beverages
Reddish-purple	Grape	All beverages, fruit base, sorbets and sauces
Yellow-orange	Tea (flavone)	Beverages, medicines, health-care product, food products
Purple –black	Black Currant	Beverages, confectionery, fruit preparations, soft drinks, preserves
Dark violet -blue	Indigo plant	Fruits, dairy products, cosmetics, medicines

Bluish red	Hibiscus	Bakery products and tea based beverages to enhance the brown tint, coloring food
Purplish red	Amaranthus	Coloring beverage of grape fruit juice, jelly and powder juice
Blue	Clerodendron	Coloring food

Source: Shamina *et al.* (2007)

CONCLUSION:

Malnutrition is posing a threat to vulnerable section in Asia and the Pacific regions due to poverty, which manifests in the form of vitamin A deficiency (VAD), iron deficiency anemia (IDA) and iodine deficiency disorders. Horticultural crops are potential sources to eliminate malnutrition and alleviate poverty. They are also sources of ‘protective foods’. Diversification and value-addition are keys for development of Horticulture (Scope for crop/product diversification and escape from risk of crop failure). There is ample scope for ‘Value-addition’ and export potential for Horticultural Crops.

References

Chandramouli, K.V. (1995). *Sources of Natural Dyes in India – A Compendium with Regional Names*, PPST Foundation, Chennai.

Coulson, J. (1980). Synthetic organic colors for food. In: (ed. J. Walford) *Development s in Food Colors* – I, Applied Sci., London, **3**: 47-94, 259 pp.

Kays, S.J. (1998). Postharvest physiology of perishable plant products. pp. 187-207. CBS Publishers and Distributors, New Delhi, India.

http://www.foodadditivesworld.com/natural-food-colors.html (Assessed on 23-03-2007)

Indu Rani, C., Arumuganatha, T. and Muthuvel, I. (2004). Food colors. *Agro India*, **7**(8): 15.

Parkinson, T.M. and Brown, J.P. (1981). Metabolic fate of food colorants. *Annual Review of Nutrition*, **1**: 175-205.

Shamina, A., Shiva, K.N. and Parthasarathy, V.A. (2007). Food colors of plant origin. In: *CAB Reviews - Perspectives in Agriculture, Veterinary Science, Nutrition and Natural Resources,* CAB International, **87**(2): 1-25.

Shiva, K.N., Shamina, A. and Parthasarathy, V.A. (2009). Horticultural crops as source of natural food colors and pigments. In: Training Manual on *“Flavours, nutraceuticals and food colors from Horticultural Crops”* (ICAR sponsored Winter School, 07.01.2009 to 28.01.2009) A. Shamina, K.S. Krishnamurthy, K.N. Shiva, N.K. Leela and B. Chempakam (Eds.). Published by the Director, IISR, Calicut, Kerala, pp. 36-44.

Singh, H.P. (2007). Technological interventions in Indian Horticulture – a way ahead. *Indian Horticulture*, **52**(4): 3-9 & 19.

Siva, M. (2003). Assessment of genetic variation in some dye-yielding plants using isozyme data. *Ph.D. thesis*, Bharathidasan University, Tiruchirapalli, TamilNadu, India.

Siva, M. (2007). Status of natural dyes and dye-yielding plants in India. *Current Science*, **92**(7): 916-925.

Concept of Models and Modling Biological System

K. Kandiannan

Introduction

System analysis has been used by engineers for many years and their success with this approach has inspired biologists to apply similar techniques in their discipline. The approach can be characterized by the terms: systems, models and simulation. A system is a limited part of reality that contains interrelated elements. A model is a simplified representation of a system. Simulation can be defined as the art of building mathematical models and the study of their properties with reference to those of the systems they represent. The word *system* may have different meanings for different people. In the scientific community, a system is usually said to be an arrangement, set, or collection of things connected in such a manner as to form an entirety or as to act as an integral unit. In a more mathematical sense, a system is defined as physical or abstract object that, over the time scale, receives inputs from outside its boundaries, responding with changes of state and with outputs. An important implication of this definition is the awareness of a dynamic condition, the awareness of changes and evolution of the system over time. ***A system is a collection of components and their interrelationships that have been grouped together for the purpose of studying some part of the real world***. The selection of the components to include in a system depends on the objectives of the study and actually represents our simplified view of reality. Systems can be described as collection of mutually interacting objects that are affected by outside forces.

Biological systems are made up of interacting chemical and physical processes. Living systems are composed of many subsystems and components, each having its own unique characteristics and behavior while contributing to the overall function of an entire system. These systems are highly complex; many components interact simultaneously, and their interactions are highly nonlinear or chaotic in nature. These interactions and nonlinearities must be taken into account when attempts are made to understand or predict system behavior. Our understanding of these interactions is incomplete and often guided by empirical evidence of overall system behavior instead of empirical data on processes that lead to overall system behavior.

System research comprises an analytical and a synthetic phase. System analysis a term applied to the logical organisation of data and theories concerning the behaviour of various system into model and the rigorous testing of such models for the purpose of validating and improving them ultimately and using them to predict the future behaviour of the systems which they represent. System analysis helps to identify systems, subsystems and key processes and often guides us to optimise in using the system.

MODELS

Models are quite essential in the various disciplines of the Arts and Sciences, Since it is only through the use of models that the diversity and complexity of a phenomenon can be comprehended. They are operational tools with the help of which one can understand reality, not the whole reality, but nevertheless, a useful and comprehensible segment of it. At the same time, if the models are not used properly, they might prove dangerous. The term 'model' has been interpreted by Scholars in different ways depending upon the need of the specialization. Model is an idealized representation of reality in order to demonstrate some of the properties of its segment in a simplified and an accurate form. Model is any abstraction or simplification of a system. Models can do considerably more on understanding assessing and optimizing the system. Model is a hypothesis which we can test experimentally. If the models survives considerable testing, it can becomes a theory, and if found to hold without exception, it becomes Law. It is understood that model is a simplified representation of a system. Modeling is the process of developing that representation.

CHARACTERISTICS OF MODELS

Making a choice between competing models requires an understanding of the many dimensions of model power. The four-element combination of a good and powerful model are

1. Comprehensiveness, i.e. a model should be applicable to a large number of situations.
2. Predictive-ness, i.e a model should be in a position to predict individual phenomenon providing large bits of information.
3. Efficiency, i.e. a model should be concise and to the point so that it can make the most of predictions using the fewest statements and refined mathematical techniques.
4. Accuracy, i.e. a model should have the quality of matching model predictions with observations.

whatever form a model may take and whatever might be its name, the essential characteristics of a complete model are.

- A representation of the structure, i.e that elements of the systems.
- A description of how the system works and how the elements in a system interact, i.e cause - effect relationships.
- A lucid explanation of the system and its functioning.

TYPES OF MODELS

Models are of several kinds. One particular model might be suitable under one set of circumstances, while another might be suitable in another situation. Each model may be characterised by a set of assumptions, parameters, and computational properties.

Exact model : The system or its components will be exhibited as it is. Example - Insects, Symptoms of diseases, Soil sample etc.,

Iconic model : Physical representation of a real object, usually three-dimensional and reduced in size. Clay models, toys, miniature dam, mini-farming system models are examples for this kind.

Analogue model : Representing one for other i.e. the properties of reality are shown by different materials or by other properties. Example, the elevation of the landforms is shown by contours in a map.

Symbolic model : The model represent the properties / components of objects / systems by symbols (Mathematical symbols). These can be further classified as

Qualitative model: This also called as conceptual model. It is one's impression of how system behaves depending upon his / her education, research and experience.

Quantitative models: Darwin's theory of evolution is a qualitative explanation of the process of speciation. Once, one begins to speculate on the rate of such evolutionary processes or attempts to measure genetic chances over time, then theory becomes quantitative.

Quantitative models can be empirical, e.g. regression models or explanatory which explain behaviour of the modeled system through its components and their interrelationships by the physical / chemical / biological / ecological processes. (Mechanistic models). The quantitative models can be further divided into sub-types-static or dynamic. Some typical example of static models is respiration and its dependence on biochemical processes, so on. Dynamic models are obtained when the time dimension is introduced. Static models are often part of dynamic models. If the models yields definite quantity it is deterministic model and if the result is attached with probability the model become stochastic.

HIERARCHY AND CAPABILITIES OF MODEL

Generally, models low in the hierarchy use a small time frame and require data from detailed experiments performed in closely defined conditions. Models low in the hierarchy also tend to be models of individual processes. These can be aggregated to form models at higher levels in the hierarchy, but there are limits on the extent to which this is feasible. The current conventional wisdom is that it is not desirable to build a model at more than one, or at most two, levels of hierarchy below the one at which predictions are to be made.

The capabilities of the various models are for summarizing data (SM) and may be even for interpolative prediction (IP), completely empirical (e.g., multiple regression) models may suffice. For extrapolative models must be mechanistic, and for interpretation (IN) they must be mechanistic and comprehensive. As the amount of mechanism in a model increases so does the requirement for detailed experimental data.

MATHEMATICAL MODEL

Mathematical model is a functional relationship between a dependent observable plant response such as growth, weight change, photosynthate change etc., and pertinent variables influencing the plant. The first crop models were conceptual models. These existed in the form of words only. Then considerable resources were diverted to building physical models, which could predict the advent of real world situations. The development and sophistication of advanced techniques of mathematics made it possible to use mathematical models in the form of symbols and set of equations about the system. These models are much easier to manipulate and more comfortable than the physical models.

ADVANTAGES OF MATHEMATICAL MODEL

i) They are precise and abstract.
ii) They transfer information in a logical way.
iii) They act as unambiguous medium of communication.

DISADVANTAGES OF MATHEMATICAL MODEL

i) Complexity of the symbolic logic.
ii) The solution of the equation should be interpretated carefully. (Jeffers,1988).

Initial period, field data and multiple regression were used to construct models. The vast number of variables involved and the complex interactions of these variables made this approach very difficult. Also inability faced to precisely measure many of the pertinent variables.

SIMULATION MODELS

Simulation is art of creating real situation. A system is an isolated part of the real world composed of one or more components and process selected for the study. The nature of the subject and the objectives of the study define system. System behaviour is the sum of all processes or changes in the system. Biological systems are continuous dynamic systems because they exhibit gradual behavior change.

Simulation models composed of dozens, even hundreds of algorithms, each containing a set of empirically determined constants. Simulation includes the processes necessary for operationalizing the model or solving the model to mimic real system behavior. Developing computer logic and flow diagrams, writing the computer code and implementing the code on a computer to produce desired outputs from analyzing the system are necessary tasks in the simulation process.

THE BASIC STEPS OF SYSTEMS SIMULATION

- define goals
- define systems and boundaries
- define key variables in system
- quantify relationships
- calibrations
- validation
- sensitivity analysis
- simplifications and use of models

Modeling Biological Systems

The gap in complexity between limited data and complex behavior of biological systems can be compared to the behavior complexity emerging in dynamical systems studies (e.g. chaotic systems). The latter is a well-known example of complexity emerging from apparently simple systems. As for biological processes, the large number of interacting components produces a particular difficulty of predicting the emerging behavior. In general, numerical modeling biological systems follows the process shown below.

1. Starting from experimental data, sort possible molecular processes and retain the most plausible.
2. Build a cartoon depicting the overall model and refine it until it is composed of elementary steps.
3. Translate the elementary steps into mathematical expressions using the laws of physics and chemistry.
4. Translate these expressions into time-dependent differential equations quantifying the changes in the model.
5. Analyze the differential system to assess the model.
6. Elaborate predictions based on a more detailed study of the differential system.
7. Test some selected predictions *in vitro* or *in vivo*.

This approach has proven substantial properties of various biological processes, as for example in the case of cell cycle

Many biochemical processes can be modeled using continuous domains by employing various kinetics based on the mass action law. However quite a number of biological processes involve small scale units and their dynamics can not be approximated using a global approach and needs to be considered unit-wise.

Some of the biological systems are now known to have a switch-like behavior and can only be specified in a continuous realm by using zero-order ultra-sensitive parametric functions converging to a sharply sigmoid function, which artificially complexifies the system.

The lack of formalized translations between each step makes the whole modeling process error-prone, since immersing the high-level comprehensible cartoon into a low-level differential formalism is completely dependent on the knowledge of the modeler and his/her mathematical skills. Maybe even worse, it blurs the explanatory power of the schema.

As an illustration of the last point it is well-known that the same high level process of the lysis/lysogeny decision in lambda bacteriophage infecting an *E. coli* cell can be specified using different low-level formalisms, each producing unique results contradicting the others.

The assessment step of the modeling process is usually conducted by slow and painful *parameter tinkering*, upon which some artificial integrators and rate constants are added to fit the model to the experimental data without any clue as to what meanings these integrators could have biologically speaking.

Two complementary approaches are necessary for model validation. The first is the validation from the computer science point of view, and is mainly based on intrinsic criteria. The second is the external validation, and in our case requires confirmation of model predictions by biological experiments.

In addition to classic measures such as indexes of cluster validity, our use of instrinsic criteria in comparative genomics depends on treatment of the organism as a system. We define coherency rules for predictions that take into account essential genes, requirements for connectivity in biochemical pathways, and, in the case of genome rearrangements, biological rules for genome construction. These rules are defined at appropriate levels in each application. Experimental validation is made possible by collaboration with partner laboratories in the biological sciences.

Agarose Gel Electrophoresis *Annexure I*

This is one of the routinely used techniques in molecular biology. This is used to separate DNA fragments and to assess the quality and quantity of DNA.

Principle

The gel is made from agarose, a highly purified form of the polysaccharide that is used to make agar plates on which bacteria is grown. The gel is immersed in buffer and the DNA fragments are loaded onto a well at one end of the gel and made to move through the gel by the application of electric current. DNA is negatively charged and so will move towards the positive anode. However, the polysaccharide mix of the gel retards the DNA by a process of sieving, so that small fragments move through faster and these fragments separate according to size.

The DNA is visualised by adding ethidium bromide (EtBr), a fluorescent molecule which intercalate with the DNA bases, extending the length of linear and nicked circular DNA molecules and making them more rigid. When EtBr is added, UV radiation at 254 nm is absorbed by the DNA and transmitted to the bound dye. The energy is re-emitted at 590 nm in the red-orange region of the spectrum. Ethidium bromide is a powerful mutagen and hence the gel should be handled carefully with the gloves. The DNA bands can be visualised under UV and gel documentation appliances can record the data.

Characteristic features of gel electrophoresis are:

1. The molecular weight of the DNA: The migration rate is inversely proportional to the molecular weight
2. Agarose concentration: The migration rate is inversely proportional to the agarose concentration
3. Conformation of the DNA: Linear form travels slowest and the supercoiled form travels fastest
4. Applied voltage: Typical value - 5 volts per cm. The heat generated during electrophoresis is dissipated by the buffer.
5. DNA being polyanionic at neutral pH, it migrates towards the anode.
6. The loading dye for DNA contains glycerol, which gives density to help the sample sink to the bottom of the well and marker dyes Xylene Cyanol and bromophenol blue. Bromophenol blue moves on par with 300-400 bp DNA and Xylene cyanol with 2-3 kb DNA.
7. The DNA is visualised by adding EtBr a fluorescent molecule that intercalates with the DNA bases. To 0.8% agarose gel add EtBr to give 0.5 pg/ml concentration. UV radiation at 254 nm is absorbed by the DNA and transmitted to the bound dye. The energy is re-emitted at 590 nm in the red-orange region of the spectrum.
8. EtBr is a powerful mutagen. The dye is usually incorporated into the gel or conversely the gel is stained after running by soaking in a solution of Et. Br.
9. The usual sensitivity of detection is 0.1 pg of DNA
10. The gel will be run along with a molecular weight marker, a wide range of which is commercially available.

Protocol

1. Prepare 1% agarose gel in Tris-acetate EDTA buffer (IX TAE) containing EtBr
2. To 1 gm of agarose, add 100 ml of IX TAE. Heat until dissolved. Cool the gel to 50°C and add EtBr (0.5 pg/ml) before pouring into the gel apparatus.
3. Wash the gel casting tray and comb with water to remove dirt.
4. Place the apparatus on a level surface and check with the spirit level and adjust the level.
5. Choose appropriate comb (commonly 12 slots) and fix into position.
6. Pour the gel onto the apparatus and allow it to cool and set.

7. After the gel has set firmly, pour little amount of buffer and remove the comb gently. Take care not to drag the comb and break the gel.
8. Immerse the gel slowly into the gel tank. Add sufficient amount of IX TAE buffer. Connect the electrode and check the current.
9. *Note*: Always check the electrical connections before loading the sample.
10. Load the samples into wells carefully.
11. Always load an aliquot of standard molecular weight marker along with the samples. It will help in assessing the size of the DNA fragment by comparing with the electrophoretic mobility.

Anticipated Results

Depending upon its confirmation and size, the DNA fragments will move as discrete bands. The EtBr intercalated with the DNA fragments will make it emit pink fluorescence under UV lamps, which can be recorded in any gel documentation system.

Quantitation of DNA

Reliable measurement of DNA concentration is important for many applications in molecular biology including complete digestion of DNA by restriction enzymes and amplification of target DNA by polymerase chain reaction. DNA quantification is generally carried out by spectrophotometric measurements or by agarose gel analysis.

Principle

The purines and pyrimidines in nucleic acid show absorption maxima around 260 nm (dATP 259, dCTP 272; dGDP 253, dTTP 247) of the DNA samples is pure without any contamination of protein or organic solvents vs. absorption at 260 nm can be taken to correctly assess the quantity. If the sample amount is less it can be visually estimated by gel electrophoresis.

Protocol

(a) Spectrophotometric measurements

1. Take 1 ml TE buffer in a cuvette and calibrate the spectrophotometer at 260 nm as well as 280 nm wavelength.
2. Add 2 to 5 pi of DNA, mix properly and record the optical density (OD) at both 260 and 280 nm.
3. Estimate the DNA concentration employing the following formula
4. Amount of DNA (µg/µl) = $\frac{(OD)_{260} \text{ x } 50 \text{ x dilution factor}}{1000}$
5. Judge the quality of DNA from the ration of the OD values recorded at 260 and 280nm. The A_{260}/A_{280} ration around 1.9 (1.85 – 1.95) indicates best quality of DNA.

Preparation of buffers and reagents *Annexure II*

Given below are the methods for preparing buffers and reagents, which are generally used in a molecular biology laboratory. The solutions, which are specific for any experiment, are given under the respective headings.

30% Acrylamide

Acrylamide	30.0 g
N-N'methylene bisacrylamide	0.8 g

Dissolve in distilled water and make up the volume to 100 ml. Store in a dark bottle at - 4°C for up to 2 weeks. Acrylamide is neurotoxic. Wear gloves while handling.

10% Ammonium persulphate (APS): Dissolve 1 g of APS in sterile distilled water and make up to 10 ml. Store at 4°C. It is preferable to prepare freshly.

Antibiotics: Dissolve 5 mg of ampicillin/kanamycin in 10 ml of sterile double distilled water. Filters sterilize and dispense into 200 µl aliquots and store at -20°C.

Antibody diluent/buffer (for ELISA)

Add 2g polyvinyl pyrrolidone (PVP) and 0.2g ovalbumin to 100 ml of PBS-T

Blocking solution (for ELISA)

Add 5.0 g Bovine Serum Albumin (BSA)/Spray dried milk (SDM) to 1 litre PBS-T.

Stop solution (for ELISA)

NaOH	12.0 g
Distilled H_2O to make	1 litre

All buffers contain 0.02% sodium azide as a preservative.

100 mM $CaCl_2$

Add 1.47 g of $CaCl_2$. $2H_2O$ in H_2O. Make upto 100 ml. Sterilise by autoclaving.

Catalase (CAT)

Staining:

Prepare the following solutions:

0.01%H_2O_2	75 ml
Stain solution:	
Potassium ferricyanide	0.375g
Ferric chloride	0.375g
Volume is made up to	75 ml.
7% acetic acid	

Chloroform:Isoamyl alcohol (24:1), 100ml

Mix 96ml of chloroform and 4ml of isoamyl alcohol. Mix well and store in reagent bottles at room temperature. Chloroform evaporates, so the cap should be tightly closed.

Coating buffer (Carbonate buffer, pH 9.6)

Na_2CO_3	1.59 g
$NaHCO_3$	2.93 g

Distilled H_2O to make 1litre

CPG (Cassamino Peptone Glucose) medium:

Tryptone	10g
Glucose	5g
Cassamino acid	1g
Yeast extract	1g
Agar	18g

Tetrazolium chloride (10mg/ml) 500ul/100ml (pH 7.2)

Use CPG for mass culturing *Ralstonia solanacearum*

DNA extraction buffer for soil DNA

Tris-Cl (pH 8.0)	100mM
EDTA (pH 8.0)	100mM
Na2H Po4 (pH 8.0)	100mM
NaCl	1.5M
CTAB	1%
Proteinase K	20µg
Lysozyme	100µg

DNA extraction buffer for plants

Tris Cl pH 8	100 mM
EDTA pH 8	20 mM
NaCl	1.4 M
2- mercaptoethanol	0.1%
CTAB	2%

DNA Molecular Weight Marker

Commercial molecular weight markers are normally prepared from digestion of different plasmids to produce discrete fragments.

DNA (ladder)	100 µl
50 x TAE	5 µl
6 x Gel loading dye	42 µl
SDW	103 µl
Total volume	250 µl

The marker is ready to be loaded.

Denaturing PAGE

Urea (AG)	453 g
10X TBE	50 ml
40%(19:1 ratio)	
Acrylamide : bisacrylamide	125 ml

Distilled water (make up to final vol of 1 litre)

Destaining solution

Methanol	50.0 ml
Glacial acetic acid	75 ml
Distilled water	1000 ml

Diaphorase (DIA)

Stain solution :	75ml
NADH	37.5mg
MTT	9mg
DCPIP	1.5mg
Tris-Cl, pH 8.0	7.5ml

Volume is made up with distilled water.

0.5 M EDTA (pH 8.0) : Add 186.1 g of disodium ethylene diamine tetraacetate. $2H_2O$ to 800 ml of H_2O. Stir vigorously on a magnetic stirrer. Adjust the pH to 8.0 with NaOH. Make upto 1 litre. Dispense into aliquots and sterilize by autoclaving.

Ethidium Bromide (10 mg/ml) : Add 1 g of ethidium bromide to 100 ml of H_2O. Stir on a magnetic stirrer for several hours to ensure that the dye has dissolved. Wrap the container in aluminium foil or transfer to a dark bottle and store at 4°C. **Caution**: Ethidium bromide is a mutagen and toxic. Wear gloves when working with ethidium bromide solutions and a mask when weighing it out.

70% ethanol, 100ml

Mix 70ml of ethanol and 30ml of distilled water. Dispense to reagent bottles and store at 4^0C or -20^0C.

Equilibriation of Phenol: Melt 500 g phenol in a water bath, kept at 65°C. Add 100 ml of

100 mM Tris Cl (pH 8.0) and shake to emulsify. Let phases separate. Repeat this step until the aqueous layer shows a pH between 7.0 to 7.5. Store the phenol under 50 ml of 10 mM Tris HCl pH 7.6 in an amber coloured bottle.

Enzyme conjugate diluent/buffer (PBS-T polyvinyl-pyrrolidone ovalbumin, PBS TPO) (for ELISA): Add 20.0 g polyvinylpyrrotidone (PVP, MW 44,000) and 2.0 g egg ovalbumin to 1 litre PBS-T.

Fixative solution

Methanol	400 ml
Glacial acid	70 ml
Distilled water	1000 ml

GYP Medium (1L) for *Phytophthora capsici*

L.Asparagine	20gm
$FeSO_4$7H2o	20mg
CaCl2.2H_2o	10mg
MgSO4.7H_2O	0.1gm
KH_2PO_4	0.47gm
$K_2H_2PO_4$	0.26gm
Thiamine hydrochloride	10mg
$ZnSO_4.7H_20$	1 µg/ml
$CuSO_4.5H_2O$	0.02µ/ ml
$Mncl_2$	2mg

pH adjusted to 6

Isocitrate dehydrogenase

Stain solution : 75ml

Isocitric acid	41.6mg
$MgCl_2$	166.6mg
NADP	33.3mg
NBT	8.3mg
PMS	1.7mg
1.0M Tris-Cl, pH8.0	7.5ml

Volume was made up with distilled water

Malate dehydrogenase (MDH)

Stain solution : 75ml

2.0 M Malic acid	2011.35mg (2.0M) dissolved in a small qty. of water.
(pH adjusted to pH 8.0 with NaOH)	
NAD	33.3mg
NBT	16.65mg
PMS	3.33mg
1.0M Tris-Cl, pH8.0	7.5ml

Volume was made up with distilled water

PVPH incorporated Carrot Agar (1L)

Carrot	200gm
Agar	20gm

Volume made upto 1000ml using sterile distilled water

For every 100 ml of carrot agar 10ml of PVPH was added.

LA medium (Luria agar medium): 950 ml of deionized water, add 10 g of bactotryptone, 5 g of yeast extract and 10 g of NaCl and 15 g of agar. Boil to dissolve content, adjust pH to 6.8 with 5 N NaOH and make up volume up to 1 litre with deionized water. Sterilize by autoclaving.

Luria-Bertani broth: It is same as above but prepared without agar.

Loading dye (6X)

Tris-HCl, pH 8.0	50 mM
EDTA	5 mM
Glycerol	25%
Bromophenol blue	0.2%
Xylene Cyanol FF0.	2%

This loading dye is used for Non-denaturing ds DNA

6X Loading dye

1% Bromophenol blue	200 µl
Glycerol	200 µl
10% SDS	60 µl
0.5 M EDTA	50 µl
10 X TAE	60 µl
Water	30 µl

100 mM $MgCl_2$: Add 2.03 g of $MgCl_2.6H_2O$ in H_2O and make upto 100 ml. Sterilise by autoclaving.

5M NaCl : Dissolve 292.2 g of NaCl in H_2O and make upto 1 litre. Dispense into aliquots. Sterlize by autoclaving.

Nutrient agar (NA)

Beef extract	5 g
Peptone	10 g
NaCl	5 g
Agar	20 g
Distilled water	1000 ml (pH 7.0)

Nutrient broth (NB) : Same as NA, but without agar.

Plasmid isolation buffers (Modified alkaline lysis method)

Solution I (resuspension buffer)	:	25 mM Tris HCl (pH 8.0) 50 mM Glucose 10 mM EDTA
Solution II (lysis buffer)	:	0.2 N NaOH 1 % SDS
Solution III (neutralization buffer)	:	3 M Sodium acetate (pH 4.8)

Phenol: Chloroform(1:1)

Mix equal amounts of phenol and chloroform. Equilibrate the mixture by extracting several times with 0.1M Tris.Cl (pH 7.6). Store the equilibrated mixture under an equal volume of 0.01M Tris.Cl (pH 7.6) at 4^0C in dark glass bottle.

Phosphate buffer-saline, PBS, pH 7.4

NaCl	8.0 g
$Na_2HPO_4.2H_2O$	1.44 g or Na_2 $HPO_4.12H_2O$ 2.90 g or Na_2HPO_4 1.50 g
KH_2PO_4	0.20 g
KCl	0.20 g

Distilled water to make 1 litre

PVS 2 solution pH 5.8

Glycerol	30% v/v
Ethylene glycol	15% v/v
DMSO	15% v/v
Sucrose	0.4 M

Resolving gel solution

Acrylamide stock solution	19.8 ml
Tris-SDS stock solution (pH 8.8)	30 ml
Distilled water	to 60 ml

TEMED	15 µl
10 % APS	1 ml

RNAse (10 mg/ml)

Dissolve RNAse A at a concentration of 10 mg/ml in 0.01 M sodium acetate(pH 5.2). Heat to°C for 15 minutes. Allow to cool slowly to room temperature. Adjust the pH by adding 0.1 volumes of 1 M tris-Cl(pH 7.4). Dispence into aliquots and store at -20°C.

10% Sodium dodecyl Sulfate (SDS) : Dissolve 100 g of SDS in K_2O and make upto 1 litre. Dispense into aliquots.

3M Sodium acetate (pH 4.8 and 5.2) : Dissolve 408.1 g of sodium acetate $3H_2O$ in 800 ml of distilled water. Adjust the pH to 4.8 to 5.2 with glacial acetic acid. Make upto 1 litre. Dispense into aliquots. Sterilize by autoclaving.

Staining solution

Coomassie Brilliant Blue R 250 (dissolved in 227 ml methanoi)	1.15 g
Glacial acetic acid	46 ml
Distilled water	500 ml

Sample buffer

Tris-SDS stock pH 6.8	25 ml
SDS	2.0 g
Glycerol	10 ml
2-mercaptoethanol	5.0 ml
1% bromophenol blue	0.1 ml
Distilled water	100 ml

Stacking gel solution (for two 1.5 mm gel, 3%)

Acrylamide stock solution	2 ml
Tris-SDS stock sol (pH 6.8)	10.0 ml
Distilled Water	60 ml
TEMED	70 µl
10 % APS	1 ml

1M Sucrose

Dissolve 34.23g of sucrose in 80ml of water. Make up the volume to 100ml. Filter sterilize.

Superoxide dismutase (SOD)

Stain solution : 75ml

Na_2EDTA	5.625mg
Riboflavin	3.0mg
NBT	7.5mg
1M Tris-Cl, pH8.0	3.75ml (0.5M final concentration)

Volume is made up with distilled water.

1M Tris-HCl : Dissolve 121.1 g of Tris hydroxylmethane aminomethane in 800 ml distilled H_2O. Adjust the pH to 8.0 with concentrated hydrochloric acid. Make upto 1 litre. Dispense into aliquots. Sterilise by autoclaving.

500ml, Tris-Cl, 1M pH-8.0

Dissolve 60.55gm of base in 400ml H_20. Adjust pH to 8.0 using conc. HCl (approx.21ml). Allow to cool the solution to room temperature and do final adjustments with pH. Make up to 500ml. Filter and sterilize by autoclaving. Store at 4^0C. pH of this solution is temperature dependent. Tris will completely dissolve only when pH becomes 8.0. Initial pH will be around 11

Tris-buffer saline, TBS, pH 7.5

0.02 M Tris	4.84 g
0.5 M NaCl	58.48 g

Adjust the pH to 7.5 with 1 N HCl and make upto 2 litre. This is used as Wash solution.

Tris-Glycine-SDS electrode buffer, pH 8.3

Tris	15.15g
Glycine	72.0 g
SDS	5.0 g

Dissolve in water and make up the volume to 5000 ml.

Tris-SDS stock solution pH 6.8

Tris	39.4 g
SDS	2 g

Dissolve in 900 ml distilled water. Adjust pH to 6.8 with HCL and make up the volume to 1000 ml.

Tris-SDS stock solution, pH 8.8

Tris	118.2 g
SDS	2 g

Dissolve in 900 ml distilled water. Adjust pH to 8.8 with HCI and make up the volume to 1litre.

10X TBE (500 ml)

Tris	54 g
Boric acid	27.5 g
EDTA (sodium salt)	4.65 g

TE saturated phenol : chloroform : isoamyl alcohol : To the equilibriated phenol add equal amount of chloroform : isoamyl alcohol (24:1). Store at 4°C in an amber coloured bottle under a layer of 10 mM Tris-CI (pH 7.6).

50 X TAE : Dissolve 242 g of Tris in H_2O. Add 57.1 ml of glacial acetic acid and add 100 ml of 0.5 M EDTA (pH 8.0). Make upto 1 litre. Dispense into aliquots. Sterilise by autoclaving.

Wash buffer (PBS-Tween, PBS-T) (For ELISA) : Add 0.5 ml Tween - 20 to 1 litre PBS.

List of suppliers of molecular biology grade chemicals **Annexure III**

Promega
www.promega.com
Hysel India Pvt Ltd,
41, DDA Shopping Centre,
Sukhdev Vihar,
New Delhi – 110025

Roche Applied Science
www.roche.com
Nicholas Piramal India Ltd
Wellspring 5th foor,
Mumbai – 400013

Aldrich
www.sigma-aldrich.com
Survey # 31/1, Sitharampalaya,
Mahadevpura P O
Bangalore – 560048
New England Biolabs
www.neb.com
Labmate Pvt Ltd.
183, Mount Road,
Madras-600015

Merck
www.merck.co.in
Shiv Sagar Estate A
Dr Annie Besent Road,
worli – Mumbai-400018

Hi-Media
www.himedialabs.com
A-406, Bhaveshwar Plaza,
LBS Marg
Mumbai-400086

BioRad Laboratories
www.bio-rad.com
Flat A-24,2nd Floor
15-C South Boag Road,
Gulmohar Appartments,
Chennai - 600017

Labmate
www.labmateasia.com
Baid Mehta Complex
C-Block, I Floor. 183, Mount Road
Chennai-600015

Stratagene
www.stratagen.com
Genetix, C-88, Kriti Nagar
Lower Ground Floor,
New Delhi-110015

Invitrogen
www.invitrogen.com
Genetix, C-88, Kriti Nagar
Lower Ground Floor,
New Delhi-110015

Bangalore Genei
www.bangalore.com
No-6 6th Main
BDA Industrial Suburb
Near SRS road
Peenya
Bangalore-560058

Amesham Biosciences Ltd
www.amershambiosciences.com
www.amershambiosciences.co.in
India Branch Office
No 18, II Floor
Cunningham Road
Bangalore-560052

Micro Devices Metrohm Ltd
www.mdml.com
N0-13/1, 1st Avenue, Indira Nagar,
Adyar, Chennai-600020

Millipore
www.millipore.com
Analytical Product Division
Millipore India Pvt. Ltd.
50 A, 2nd Phase, Ring Road
Peenya, Bangalore-560058

Genetix
www.qiagen.com
C-88, Kirti Nagar
Lower Ground Floor
New Delhi-110 015

General Conversion Tables and Formulae **Annexure IV**

Common Conversions of Nucleic Acids and Proteins

Weight conversion

1μg	=	10^{-6} g
1 ng	=	10^{-9} g
1 pg	=	10^{-12} g
1 fg	=	10^{-15} g

Spectrophotometric conversion

1 A_{260} unit of double-stranded DNA = 50 μg/ml
1 A_{260} unit of single-stranded DNA = 33 μg/ml
1 A_{260} unit of single-stranded RNA = 40 μg/ml

DNA molar conversions

1 μg of 1,000 bp DNA – 1.52 pmole (3.03 pmoles of ends)
1 pmole of 1000 bp DNA = 0.66 μg

Protein molar conversion

100 pmoles of 100,000 dalton protein = 10 μg
100 pmoles of 50,000 dalton protein = 5 μg
100 pmoles of 10,000 dalton protein = 1 μg

Protein/DNA conversion

1 kb of DNA = 330 amino acids of coding capacity – 3.7 x 10^4 dalton protein
10,000 dalton protein = 270 bp DNA
50,000 dalton protein = 1.35 kb DNA
100,000 dalton protein = 2.7 kb DNA

DNA Data

Average weight of DNA base pair (sodium salt) = 650 daltons

MW of a double-stranded DNA molecule = (# of base pair)’ (650 daltons/base pair)
Moles of ends of a double-stranded DNA molecule = 2’ (grams of DNA)/(MW in daltons)
Moles of ends generated by restriction endonuclease cleavage:

1. Circular DNA molecule: 2 x (moles of DNA) x (number of sites)
2. Linear DNA molecule: 2 x (moles of DNA) x (number of sites) + 2 x (moles of DNA)

1.0 kb DNA = coding capacity for 33 amino acids = 37,000 dalton protein

10,000 dalton protein = 270 bp DNA
50,000 dalton proteins = 1.35 kb DNA

Linear DNA	µg/ml	Moles/ml	Molecules/ml	Conc.	Conc.5’ ends
1.0 A_{260} DNA =	50µg/ml	1.6 x 10^{-12}	9.8 x 10^{11}	1.6nM	3.2nM
1.0 A_{260} pBR 322 DNA =	50µg/ml	1.8 x 10^{-11}	1.1 x 10^{13}	18.0nM	36.0nM
1.0 A_{260} linker (8 MER) =	50µg/ml	9.8 x 10^{-9}	5.9 x 10^{15}	10.0µM	20.0µM

	U	C	A	G	
	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C
U					
	UUA Leu	UCA Ser	UAA Stop (och)	UGA Stop (opal)	A
	UUG Leu	UCG Ser	UAG (stop (amb)	UGA Trp	G
	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC Leu	CCC Pro	CAC His	CGC Arg	C
C					
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G
	AUU lle	ACU Thr	AAU Asn	AGU Ser	U
	AUC lle	ACC Thr	AAC Asn	AGC Ser	C
A					
	AUA lle	ACA Thr	AAA Lys	AGA Arg	A
	AUG Met	ACG Thr	AAG Lys	AGG Arg	G
	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC Val	GCC Ala	GAC Asp	GGC Gly	C
G					
	GUA Val	GCA Ala	GAA Glu	GGA Gly	A
	GUG Val (Met)	GCG Ala	GAG Glu	GGG Gly	G

Bases are given as ribonucleotides. GUG usually codes for valine, but it can code for methionine to initiate an mRNA chain. Stop (och) refers to the ochre termination triplet and Stop (amb) refers to the amber.

Amino acid	**Three-letter Abbrev.**	**One-letter Symbol**	**MW**
Alanine	Ala	A	80
Arginine	Arg	R	174
Asparagine	Asn	N	132
Aspartic acid	Asp	D	133
Asparagine/aspartic acid	Asx	B	-
Cysteine	Cys	C	121
Glutamine	Gln	Q	146
Glutamic Acid	Glu	E	147
Glutamine/glutamic acid	Glx	Z	-
Glycine	Gly	G	75
Histidine	His	H	155
Isoleuicne	lle	I	131
Leucine	Leu	L	131
Lysine	Lys	K	146
Methionine	Met	M	149
Phenylalanine	Phe	F	165
Proline	Pro	P	115
Serine	Ser	S	105
Threonine	Thr	T	119
Tryptophan	Trp	W	204
Tyrosine	Tyr	Y	181
Valine	Val	V	117

Formulae

Picomole ends per microgram of double-stranded linear DNA
(2 x 10^6) /(660 x Number of Bases) = pmole ends/µg

Exact molecular weight of an oligonucleotide

[(A x 312.2) + (G x 328.2) + (C x 288.2) + (T x 303.2) – 61.0]
= MW (g/mol) of specific oligonucleotide

Nucleic acid sequence

GenBank (http://www.ncbi.nlm.nih.gov: 80/entrez/query.fcgi?db=Nucleotide)
SRS at EMBL/EBI (http://srs.ebi.ac.uk)
DNA Data Bank of (http://www.ddbj.nig.ac.jp/)
Japan (DDBJ)

Comparative Genomics

Comparative Genometrics (http://www.unil.ch/igbm/genomics/genometrics.html)- Biometric comparisons of whole genomes
EuGenes (http://iubio.bio.indiana.edu:89/)- Common summary of gene and genomic information from eukaryotic databases
Genome Information Broker (http://gib.genes.nig.ac.jp/)- Comparative analysis of completed microbial genomes
Gramene (http://www.gramene.org/)- Comparative genome analysis in the grasses

Gene Expression

ASDB (http://cbcg.lbl.gov/asdb)- Protein products and expression patterns of alternatively spliced genes
Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo)- Gene expression and hybridization array data repository
MethDB (http://www.methdb.de/)- DNA methylation data, patterns, and profiles
RECODE (http://recode.genetics.utah.edu/)- Genes using programmed translational recoding in their expression
Stanford Microarray Database (http://genome-www.stanford.edu/microarray)- Raw and normalized data from microarray experiments
Gene Expression Links (http://industry.ebi.ac.uk/~alan/MicroArray/)
TRIPLES (http://ygac.med.yale.edu/)- Transposon-insertion phenotypes, localization and expression in Saccharomyces
yMGV (http://www.transcriptome.ens.fr/ymgv/)- Yeast microarray data and mining tools

Gene Identification and Structure

Ares Lab Intron Site (http://www.cse.ucsc.edu/research/compbio/yeast_introns.html)- Yeast spliceosomal introns
COMPEL (http://compel.bionet.nsc.ru/)- Composite regulatory elements
CUTG (http://www.kazusa.or.jp/codon/)- Codon usage tables
DBTBS (http://elmo.ims.u-tokyo.ac.jp/dbtbs/)- Bacillus subtilis binding factors and promoters
DBTSS (http://elmo.ims.u-tokyo.ac.jp/dbtss/)- Transcriptional start sites
EID (http://mcb.harvard.edu/gilbert/EID/)- Protein-coding, intron-containing genes
EPD (http://www.epd.isb-sib.ch/)- Eukaryotic POL II promoters with experimentally determined transcription start sites
ExInt (http://intron.bic.nus.edu.sg/exint/exint.html)- Exon–intron structure of eukaryotic genes
FUGOID (http://wnt.cc.utexas.edu/~ifmr530/introndata/main.htm)- Functional and structural information on organellar introns
IDB/IEDB (http://nutmeg.bio.indiana.edu/intron/index.html)- Intron sequence and evolution
PALSdb (http://palsdb.ym.edu.tw/)- Putative alternative splice sites
PLACE (http://www.dna.affrc.go.jp/htdocs/PLACE)- Plant cis-acting regulatory elements
PlantCARE (http://sphinx.rug.ac.be:8080/PlantCARE/)- Plant cis-acting regulatory elements

PromEC (http://bioinfo.md.huji.ac.il/marg/promec)- Escherichia coli mRNA promoters with experimentally-identified transcriptional start sites
RRNDB (http://rrndb.cme.msu.edu/)- Variation in prokaryotic ribosomal RNA operons
RSDB (http://rsdb.csie.ncu.edu.tw/)- Repetitive elements from completed genomes
rSNP Guide (http://wwwmgs.bionet.nsc.ru/mgs/systems/rsnp/)- Single nucleotide polymorphisms in regulatory gene regions
STRBase (http://www.cstl.nist.gov/div831/strbase/)- Short tandem DNA repeats
TransCOMPEL(http://compel.bionet.nsc.ru/FunSite/CompelPatternSearch.html)- Transcriptional regulatory elements in eukaryotic genes
Transterm (http://uther.otago.ac.nz/Transterm.html)- Codon usage, start and stop signals
TRRD (http://wwwmgs.bionet.nsc.ru/mgs/dbases/trrd4/)- Transcription regulatory regions of eukaryotic genes
VIDA (http://www.biochem.ucl.ac.uk/bsm/virus_database/VIDA.html)- Virus genome open reading frames
WormBase (http://www.wormbase.org/) - Guide to C.elegans biology
YIDB (http://www.embl-heidelberg.de/ExternalInfo/seraphin/yidb.html)- Yeast nuclear and mitochondrial intron sequences

Genomic Databases

Arabidopsis Information Resource (TAIR) (http://www.arabidopsis.org/)- Arabidopsis thaliana genome
ArkDB (http://www.thearkdb.org/)- Genome databases for farm and other animals
Comprehensive Microbial Resource (http://www.tigr.org/tigr-scripts/CMR2/CMRHomePage.spl)- Completed microbial genomes
CropNet (http://ukcrop.net/) - Genome mapping in crop plants
EcoGene (http://bmb.med.miami.edu/EcoGene/EcoWeb/)- E.coli K-12 sequences
EMGlib (http://pbil.univ-lyon1.fr/emglib/emglib.html)- Completely sequenced prokaryotic genomes
GOBASE (http://megasun.bch.umontreal.ca/gobase)- Organelle genome database
GOLD (http://igweb.integratedgenomics.com/GOLD/)- Information regarding complete and ongoing genome projects
ICB (http://www.mbio.co.jp/icb)- Identification and classification of bacterial protein-coding regions
INE (http://rgp.dna.affrc.go.jp/giot/INE.html)- Rice genome analysis and sequencing
MatDB (http://mips.gsf.de/proj/thal/db/)- Arabidopsis Genome Initiative data
Medicago Genome Initiative (MGI) (https://xgi.ncgr.org/mgi)- Model legume Medicago ESTs, gene expression and proteomic data
Mendel Database (http://www.mendel.ac.uk/)- Database of plant EST and STS sequences annotated with gene family information
MitoNuc/MitoAln (http://bighost.area.ba.cnr.it/srs6bin/wgetz?-page+Liblnfo+-lib+MITONUC)- Nuclear genes coding for mitochondrial proteins
MITOP (http://www.mips.biochem.mpg.de/proj/medgen/mitop/)- Mitochondrial proteins, genes and diseases
MIPS (http://www.mips.biochem.mpg.de/)- Protein and genomic sequences
NRSub (http://pbil.univ-lyon1.fr/nrsub/nrsub.html)- B.subtilis genome
Oryzabase (http://www.shigen.nig.ac.jp/rice/oryzabase/)- Rice genetics and genomics
Phytophthora Genome Consortium Database (https://xgi.ncgr.org/pgc)- ESTs from Phytophthora infestans and Phytophthora sojae
RiceGAAS (http://RiceGaas.dna.affrc.go.jp/)- Rice genome sequence and predicted gene structure
Saccharomyces Genome Database (SGD)(http://genome-www.stanford.edu/Saccharomyces)- Saccharomyces cerevisiae genome
SubtiList (http://genolist.pasteur.fr/SubtiList/)- B.subtilis 168 genome

TIGR Microbial Database (http://www.tigr.org/tdb/mdb/mdbcomplete.html)- Microbial genomes and chromosomes
WILMA (http://www.came.sbg.ac.at/wilma/)- C.elegans annotation
ZmDB (http://zmdb.iastate.edu/)- Maize genome database

Intermolecular Interactions

BIND (http://bind.ca/) - Molecular interactions, complexes and pathways
Database of Interacting Proteins (http://dip.doe-mbi.ucla.edu/)- Experimentally-determined protein–protein interactions
Database of Ribosomal Crosslinks (DRC) (http://www.mpimg-berlin-dahlem.mpg.de/~ag_ribo/ag_brimacombe/drc/) - Ribosomal crosslinking data
DPInteract (http://arep.med.harvard.edu/dpinteract/)- Binding sites for E.coli DNA-binding proteins

Metabolic Pathways and Cellular Regulation

EcoCyc (http://ecocyc.org/)- E.coli K-12 genome, metabolic pathways, transporters and gene regulation
ENZYME (http://www.expasy.ch/enzyme/)- Enzyme nomenclature
GeneNet (http://wwwmgs.bionet.nsc.ru/mgs/systems/genenet/)- Formalized descriptions of the structure and functional organization of gene networks
Klotho (http://www.ibc.wustl.edu/klotho/)- Collection and categorization of biological compounds
Kyoto Encyclopedia of Genes and Genomes (KEGG) (http://www.genome.ad.jp/kegg)- Metabolic and regulatory pathways
LIGAND (http://www.genome.ad.jp/ligand/)- Chemical compounds and reactions in biological pathways
MetaCyc (http://ecocyc.org/)- Metabolic pathways and enzymes from various organisms
PathDB (http://www.ncgr.org/pathdb)- Biochemical pathways, compounds and metabolism
RegulonDB (http://www.cifn.unam.mx/regulondb/)- E.coli transcriptional regulation and operon organization
UM-BBD (http://umbbd.ahc.umn.edu/)- Microbial biocatalytic reactions and biodegradation pathways
WIT2 (http://wit.mcs.anl.gov/WIT2/)- Integrated system for functional curation and development of metabolic models

Protein Databases

AARSDB (http://rose.man.poznan.pl/aars/index.html)- Aminoacyl-tRNA synthetase sequences
ABCdb (http://ir2lcb.cnrs-mrs.fr/ABCdb/)- ABC transporters
AraC/XylS database (http://www.arac-xyls.org/)- AraC/XylS family of positive regulators in bacteria
ASPD (http://wwwmgs.bionet.nsc.ru/mgs/gnw/aspd)- Artificial proteins and peptides
CSDBase (http://www.chemie.uni-marburg.de/~csdbase)- Cold shock domain-containing proteins
DatA (http://luggagefast.Stanford.EDU/group/arabprotein/)- Annotated coding sequences from Arabidopsis
ESTHER (http://www.ensam.inra.fr/cholinesterase/)- Esterases and alpha/beta hydrolase enzymes and relatives
EXProt (http://www.cmbi.nl/exprot)- Proteins with experimentally-verified function
InBase (http://www.neb.com/neb/inteins.html)- All known inteins (protein splicing elements): properties, sequences, bibliography
MEROPS (http://www.merops.ac.uk/)- Proteolytic enzymes (proteases/peptidases)
MetaFam (http://metafam.ahc.umn.edu/)- Integrated protein family information

ooTFD (http://www.ifti.org/)- Transcription factors and gene expression
PhosphoBase (http://www.cbs.dtu.dk/databases/PhosphoBase/)- Protein phosphorylation sites
PLANT-Pis (http://bighost.area.ba.cnr.it/PLANT-PIs/)- Plant protease inhibitors
PlantsP (http://plantsp.sdsc.edu/) - Plant protein kinases and phosphatases
Prolysis (http://delphi.phys.univ-tours.fr/Prolysis/)- Proteases and natural and synthetic protease inhibitors
Protein Information Resource (PIR) (http://pir.georgetown.edu/)- Comprehensive, annotated, non-redundant protein sequence database
TIGRFAMs (http://www.tigr.org/TIGRFAMs)- Protein family resource for the functional identification of proteins
TRANSFAC (http://transfac.gbf.de/TRANSFAC/)- Transcription factors and binding sites
VIDA (http://www.biochem.ucl.ac.uk/bsm/virus_database/VIDA.html)- Homologous viral protein families

Protein Sequence Motifs

BLOCKS (http://blocks.fhcrc.org/)- Multiple alignments of conserved regions of protein families
CDD (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml)- Alignment models for conserved protein domains
eMOTIF (http://motif.stanford.edu/emotif)- Protein sequence motif determination and searches
InterPro (http://www.ebi.ac.uk/interpro/)- Integrated documentation resource for protein families, domains and sites
iPROCLASS (http://pir.georgetown.edu/iproclass/)- Annotated protein classification database with structure and function information
PIR-ALN (http://pir.georgetown.edu/pirwww/dbinfo/piraln.html)- Protein sequence alignments
PRINTS (http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/)- Hierarchical gene family fingerprints
ProClass (http://pir.georgetown.edu/gfserver/proclass.html)- Protein families defined by PIR superfamilies and PROSITE patterns
ProDom (http://www.toulouse.inra.fr/prodom.html)- Protein domain families
PROSITE (http://www.expasy.org/prosite)- Biologically-significant protein patterns and profiles
SBASE (http://www.icgeb.trieste.it/sbase)- Annotated protein domain sequences
SUPFAM (http://pauling.mbu.iisc.ernet.in/~supfam)- Sequence families correlated to structure

Proteome Resources

GELBANK (http://gelbank.anl.gov/) - 2D-gel electrophoresis patterns from completed genomes
Predictome (http://predictome.bu.edu/) - Putative functional links between proteins
REBASE (http://rebase.neb.com/rebase/rebase.html)- Restriction enzymes and associated methylases
SWISS-2DPAGE (http://www.expasy.ch/ch2d/) - Annotated two-dimensional polyacrylamide gel electrophoresis database

RNA Sequences

ACTIVITY (http://wwwmgs.bionet.nsc.ru/mgs/systems/activity/)- Functional DNA/RNA site activity
ARED (http://rc.kfshrc.edu.sa/ared)- AU-rich element-containing mRNAs
European Large Subunit rRNA Database

Guide RNA Database (http://biosun.bio.tu-darmstadt.de/goringer/gRNA/gRNA.html)- Guide to RNA sequences
HyPaLib (http://bibiserv.techfak.uni-bielefeld.de/HyPa/)- Structural elements characteristic for classes of RNA
PLANTncRNAs (http://www.prl.msu.edu/PLANTncRNAs/)- Plant non-protein coding RNAs with relevant gene expression information
PLMItRNA (http://bigarea.area.ba.cnr.it:8000/PLMItRNA/)- Mitochondrial tRNA genes and molecules in photosynthetic eukaryotes
Ribosomal Database Project (RDP-II) (http://rdp.cme.msu.edu/)- rRNA sequence data, alignments and phylogenies
tRNA Sequences (http://www.uni-bayreuth.de/departments/biochemie/trna/)- tRNA and tRNA gene sequences
Viroids and viroid-like RNAs (http://nt.ars-grin.gov/subviral/)- Viroids and viroid-like RNAs

Structure

ASTRAL (http://astral.stanford.edu/)- Sequences of domains of known structure, selected subsets and sequence–structure correspondences
BioMagResBank (http://www.bmrb.wisc.edu/)- NMR spectroscopic data from proteins, peptides and nucleic acids
CSD (http://www.ccdc.cam.ac.uk/prods/csd/csd.html)- Crystal structure information for organic and metal organic compounds
GTOP (http://spock.genes.nig.ac.jp/~genome/gtop-j.html)- Protein structures predicted from genome sequences
IMB Jena Image Library of Biological Macromolecules (http://www.imb-jena.de/IMAGE.html) - Visualization and analysis of three-dimensional biopolymer structures
MMDB (http://www.ncbi.nlm.nih.gov/Structure/)All experimentally-determined three-dimensional structures, linked to NCBI Entrez
PDB (http://www.rcsb.org/pdb/)- Structure data determined by X-ray crystallography and NMR
PDB-REPRDB (http://www.cbrc.jp/papia/) - Representative protein chains, based on PDB entries
PDBsum (http://www.biochem.ucl.ac.uk/bsm/pdbsum)- Summaries and analyses of PDB structures
RESID (http://www-nbrf.georgetown.edu/pirwww/dbinfo/resid.html)- Protein structure modifications
SCOP (http://scop.mrc-lmb.cam.ac.uk/scop)- Familial and structural protein relationships

Archives and Indices

1. Bio Catalog (http://www.ebi.ac.uk/biocat/biocat.html) links to biotechnology-related non-commercial software available via the Internet
2. Indiana University Repository for Biological Software- IUBIO ARCHIVE FOR BIOLOGY (ftp://iubio.bio.indiana.edu)- Indiana University (USA)

DNA Sequence Analysis Software

1. AAT- Analysis and Annotation Tool (http://genome.cs.mtu.edu/aat/aat.html)- Identify genes by comparing cDNA and protein sequence databases
2. AcaClone pDRAW32 (http://www.acaclone.com) - DNA cloning, analysis and visualisation software. Freeware
3. ACD/Protein Manager (http://www.acdlabs.com/products/peptide/prot_mgr.html)- Sequence Analysis tools
4. BioinformatiX (http://www.xpogen.com)- Analysis of microarray data
5. Biopendium (http://www.inpharmatica.co.uk)- Sequence, structure and function relationships for all gene products in the public domain
6. BioTools (http://www.biotools.com)- DNA, protein sequence analysis, and chromatogram analysis
7. BLAST (http://www.ncbi.nlm.nih.gov/BLAST/) - Pair wise sequence alignment
8. CLUSTALW (www.ebi.ac.uk/clustalw/) - Multiple sequence alignment
9. CURVES (http://www.ibpc.fr/UPR9080/Curindex.html)- Nucleic acid helical analysis program
10. DNASIS (http://www.oligo.net/dnasis.htm)- MacOS and Windows-based program for gene analysis
11. Double Digester: (ftp.cs.yale.edu/pub/double_digester)- Yale University (USA)
12. Gene Modeller (Gm) (ftp.tigr.org/pub)
13. FASTA (http://www.ebi.ac.uk/fasta33/)- Pair wise sequence alignment
14. Gene Construction Kit(tm) 2 (http://www.sciquest.com)- A DNA manipulation, cloning design, and illustration tool
15. Gene Inspector(tm) 1.5 (http://www.sciquest.com.)- Sequence analysis package
16. GeneJockey (http://www.biosoft.com/biosoft)- Editing, manipulation, and analysis of nucleic acid and protein sequences
17. GENSCAN (http://genes.mit.edu/GENSCAN.html)- Predicts complete gene structures in genomic sequences
18. GLIMMER (www.tigr.org/software/glimmer/)- uses interpolated Markov models to find genes in microbial DNA
19. Grail(tm) Toolkit. (http://www.apocom.com) - homology-based gene assembly
20. ISYS(TM) (http://www.ncgr.org/isys)- capable of integrating data sources and analysis tools from an investigator's own laboratory
21. Lasergene (http://www.dnastar.com)- sequencing, primer design, sequence alignment, databases and database searching, protein analysis, and restriction map analysis
22. MZEF (http://sciclio.cshl.org/genefinder/)- predicts internal protein coding exons in genomic DNA sequences
23. NIH-Repositories of Sequence Analysis Software (http://molbio.info.nih.gov/molbio/software.htm) - NIH (USA)
24. Onto-Express (www.openchannelfoundation.org)- automates the process of annotating a set of expressed genes with genomic functions
25. OMIGA (http://www.accelrys.com/)- multiple sequence alignments and contains over 50 DNA and protein analysis functions

26. pDRAW32 (www.crosswinds.net/_acaclone)- DNA analysis software by AcaClone software
27. Plasmid Processor (www.hytti.uku.fi/plasmid.html)- You can download full executable version of Plasmid Processor 1.02 for Windows (tm) at University of Kuopio (Finland)
28. Plasmid (http://www.redasoft.com) - vector map drawing and sequence analysis
29. SeqVerter (www.genestudio.com) - nucleotide sequence converter/viewer
30. SIGNALSCAN (http://bimas.dcrt.nih.gov/molbio/signal/)- finds homologies in published signal sequences
31. SRS (http://www.lionbioscience.com/)- database integration and biological information search system
32. TESS (http://www.cbil.upenn.edu/tess/)- locating and displaying transcription factor binding sites in DNA sequences
33. The Rosetta Resolver System (www.rii.com)- high-capacity data storage, retrieval and analysis of gene expression data
34. The Wisconsin Package™ (http://www.accelrys.com)- Sequence analysis contains over 100 interrelated software programs.
35. TurboBLAST (http://www.turbogenomics.com/)- Implementation of BLAST, prvides critical insight into the structure and function of genes and proteins
36. Visual Cloning 2000 (http://www.redasoft.com)- vector map drawing and sequence analysis program

Molecular Modeling Software

1. Antheprot (http://antheprot-pbil.ibcp.fr/Documentation_antheprot.html) - Protein sequence alignment, secondary structure predictions, sites & function detection, physico-chemical profiles, homology search and 3D display of protein structures freely available.
2. Chime (http://www.mdlchime.com/chime/) - Interactively displays 2D and 3D molecules directly in Web pages.
3. Cn3D viewer (http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml) - Visualization tool for bio molecular structures in MMDB format, sequences, and sequence alignments
4. Modeler (http://guitar.rockefeller.edu/modeller/modeller.html) - Homology / comparative modeling of protein three-dimensional structures
5. NAB (http://www.scripps.edu/case) - construct models of helical and non-helical nucleic acids
6. Peptool (http://www.biotools.com/) - alignment, homology building, structure prediction
7. PROCHECK (http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html) - geometric analysis of proteins and DNA including phi-psi angles, hbonding, secondary structure evaluation
8. Rasmol (http://www.umass.edu/microbio/rasmol/) - RasMol is a program for molecular graphics visualization tool for showing the structure of DNA, proteins and smaller molecules
9. Swiss PDB Viewer (http://us.expasy.org/spdbv/) - It is tightly linked to Swiss model, an automated homology modeling server
10. Web lab viewer (http://molsim.vei.co.uk/weblab/) - Create and visualize 3D molecular structures
11. WHAT IF (http://www.cmbi.kun.nl/gv/servers/WIWWWI/) - protein modeling, protein structure analysis, database handler, molecular graphics package
12. Display and analysis of molecular model data (MSC)

Primer Designing Softwares

1. Amplify (http://engels.genetics.wisc.edu/amplify) - Software for PCR - Ver.2.52b

2. Array Designer 2 (http://www.premierbiosoft.com/) - Design hundreds of primers for DNA or oligonucleotide microarrays
3. CODEHOP (http://blocks.fhcrc.org/blocks/codehop.html) - PCR primers designed from protein multiple sequence alignments
4. Fast PCR (http://www.biocenter.helsinki.fi/bi/bare-1_html/oligos.htm) - PCR primer design, DNA and protein tools program
5. GenomePRIDE (http://pride.molgen.mpg.de/genomepride.html) - Primer design program that designs PCR primers or long oligos on an annotated sequence
6. MethPrimer (http://itsa.ucsf.edu/~urolab/methprimer/) - A program for designing bisulfite-conversion-based Methylation PCR Primer.
7. NetPrimer (http://www.PremierBiosoft.com) - Most comprehensive free primer analysis program on the web
8. PCR Rare (http://bioinformatics.weizmann.ac.it/software/PC-Rare) - PC-Rare is a very powerful (but user friendly) software that allows the choice of specific PCR primers.
9. Primer3 (http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi) - Pick primers from a DNA sequence
10. Primerfinder (http://eatworms.swmed.edu/~tim/primerfinder/)-Tool to design oligonucleotides suitable for PCR within any sequence
11. Primer Premier 5 (http://www.PremierBiosoft.com/) - Primer design program
12. X-Primer (http://alces.umn.edu/pv/pub/ComBin) - for designing PCR primers - the Virtual Genome Center (USA)

RNA Analysis

1. CARD (rrna.uia.ac.be/card.html) - A computer program for drawing RNA secondary structure models - The rRNA WWW Server - University of Antwerp (Belgium)
2. The RNAdraw Home Page (http://www.yk.rim.or.jp/~aisoai/index.html) - by Ole Matzura
3. RNA Secondary Structure Prediction Software (*ftp.nrcbsa.bio.nrc.ca/pub/)* - NRC (Canada)
4. STRUCTURELAB (http://www-lmmb.ncifcrf.gov/%7Ebshapiro/RNAstructure.html) - Computational Approaches to RNA Structure Analysis
5. Vienna RNA Package (http://www.tbi.univie.ac.at/%7Eivo/RNA/) - The Vienna RNA Package consists of a library and some standalone programs for the prediction and comparison of RNA secondary structures *by* Ivo Hofacker - Wine (Austria)

Systematics and Taxonomy

1. An Alphabetic List of Genetic Analysis Software- This list mainly contains computer software on genetic linkage analysis, marker mapping, and pedigree drawing.
2. DELTA: *ftp* - DEscription Language for TAxonomy (Australia)
3. FastDNAml (http://bioweb.pasteur.fr/seqanal/interfaces/fastdnaml.html) - Construction of phylogenetic trees of DNA sequences using maximum likelihood
4. GeneStudio Pro (http://www.genestudio.com) - Sequence alignments, Phylogenetic analysis and display of trees and Full integration of Internet resources such as Entrez (GenBank), PubMed, and BLAST
5. PHYLIP phylogeny package (http://evolution.genetics.washington.edu/phylip.html) - It is a *free* package of programs for inferring phylogenies
6. Phylodendron (http://iubio.bio.indiana.edu/soft/molbio/java/apps/trees/) - An application for drawing phylogenetic trees, used in evolutionary biology. Options allow you to save in various formats, print, edit, modify, and adorn the tree.
7. TreeGen (http://cbrg.inf.ethz.ch/Server/subsection3_1_6.html) - Compute phylogenetic trees and 2D optimal placements based on distance matrices and optional variances
8. TreeTop (http://www.genebee.msu.su/services/phtree_reduced.html) - Phylogenetic Tree Prediction

9. TreeView (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html) - Program for displaying and printing phylogenies
10. WEBPHYLIP (http://sdmc.krdl.org.sg:8080/~lxzhang/phylip/) - It is a web version of the original PHYLIP package.

GLOSSARY *Annexure X*

Agrobacterium tumefaciens A species of gram-negative, rod-shaped soil bacteria; often contains copies of a large plasmid, designated Ti. A Ti-containing strain can infect many plants and form crown gall tumors.
Aliquot An exact fractional sample or portion of a whole (used especially of solutions).
Allele is one of several alternative forms of a gene occupying a given locus on a chromosome.
Amber codon is the nucleotide triplet UAG, one of three "nonsense" codons that cause termination of protein synthesis.
Amplified Fragment Length Polymorphism (AFLP) PCR based marker. Not based on arbitrairy priming of oligo's, but amplification of specificly selected restriction fragments (see SRFA).
Annealing is the pairing of complementary single strands of DNA to form a double helix.
Antibody is a protein (immunoglobulin) produced by B lymphocyte cells that recognizes a particular foreign "antigen", and thus triggers the immune response.
Anticoding strand of duplex DNA is used as a template to direct the synthesis of RNA that is complementary to it.
Antigen is any molecule whose entry into an organism provokes synthesis of an antibody (immunoglobulin).
Antiparallel strands of the double helix are organized in opposite orientation, so that the 5' end of one strand is aligned with the 3' end of the other strand.
Antisense gene An engineered gene placed in inverted orientation relative to a promoter which when transcribed, produces a transcript complementary to the mRNA transcribed from the normal orientation of the gene.
Arbitrairy Primed Polymerase Chain Reaction (AP-PCR) DNA amplification using a single random primer of 16-20 bases. Products are often analysed on a polyacrylamide gel and detected by autoradiography.
Autoradiography detects radioactively labelled molecules by their effect in creating an image on photographic film.
Base pair (bp) is a partnership of A with T or of C with G in a DNA double helix; other pairs can be formed in RNA under certain circumstances.
Blunt-end ligation is a reaction that joins two DNA duplex molecules directly at their ends.
Bulked Segregant Analysis (BSA) A rapid mapping strategy suitable for monogenic qualitative traits. When DNA of ten plants are bulked into one pool, all alleles must be present. Two bulked pools of segregants differing for one trait will differ only at the locus harbouring that trait.
cDNA is a single-stranded DNA complementary to an RNA, synthesized from it by reverse transcription *in vitro.*
cDNA clone is a duplex DNA sequence representing an RNA, carried in a cloning vector.
cDNA library it represents copies of the transcripts expressed in a particular tissue at a particular time. A cDNA clone thus represents the coding sequence of a gene.
Chromosome walking describes the sequential isolation of clones carrying overlapping sequences of DNA, allowing large regions of the chromosome to be spanned. Walking is often performed in order to reach a particular locus of interest.
Clone describes a large number of cells or molecules identical with a single ancestral cell or molecule.
Cloning vector is a plasmid or phage that is used to "carry" inserted foreign DNA for the purposes of producing more material or a protein product.
Coding strand of DNA has the same sequence as mRNA.
Codominant alleles both contribute to the phenotype; neither is dominant over the other.
Codon is a triplet of nucleotides that represents an amino acid or a termination signal.
Colony hybridization is a technique for using *in situ* hybridization to identify bacteria carrying chimeric vectors whose inserted DNA is homologous with some particular sequence.
Concatemer of DNA consists of a series of unit genomes repeated in tandem.
Conjugation describes, "mating" between two bacterial cells, when (part of) the chromosome is transferred from one to the other.
Consensus sequence is an idealized sequence in which each position represents the base most often found

when many actual sequences are compared.

Cot is the product of DNA concentration and time of incubation in a reassociation reaction.

Cot$_{1/2}$ is the Cot required to proceed to half completion of the reaction; it is directly proportional to the unit length of reassociating DNA.

Cryoprotectant: Compound preventing cell damage during freezing and thawing processes. Cryoprotectants are agents with high water solubility and low toxicity. Two types of cryoprotectant agent are commonly used: permeating (glycerol and DMSO (q.v.)) and non-permeating (sugars, dextran, ethylene glycol, polyvinyl pyrolidone and hydroxyethyl starch).

Cryopreservation: The preservation of germplasm resources in a dormant state by cryogenic techniques, as currently applied to storage of plant seeds and pollen, microorganisms, animal sperm, and tissue culture cell lines

Degeneracy in the genetic code refers to the lack of an effect of any change in the third base of the codon on the amino acid that is represented.

Denaturation of DNA or RNA describes its conversion from the double-stranded to the single stranded state; separation of the strands is most often accomplished by heating.

Denaturation of protein describes its conversion from the physiological conformation to some other (inactive) conformation.

Differential Display Reverse Transcriptase-PCR (DDRT-PCR) This technique enables the amplification of differentially expressed mRNA deriving from identical genotypes exposed to different treatments.

Diploid set of chromosomes contains two copies of each autosome and two sex chromosomes.

DNA Amplification Fingerprinting (DAF) refers to DNA amplification using a single random primer of 8-10 bases. DAF products are analysed on polyacrylamide gels and detected by silver staining

DNA polymerase One of several enzymes that synthesize a new DNA strand complementary to a template strand by adding nucleotides one at a time to a 3'-OH end.

DNAase is an enzyme that attacks bonds in DNA.

Doubled Haploids (DH) A progeny of doubled haploids derived from a heterozygous or F1 individual can serve as a mapping population. Making a DH is faster than RIL; DHs have a better resolution than a F2 progeny (no heterozygousity); DHs can be maintained infinitely.

Downstream identifies sequences proceeding farther in the direction of expression, for example, the coding region is downstream of the initiation codon.

Electrophoresis A technique for separating different types of molecules based on their patterns of movement in an electrical field.

End labeling describes the addition of a radioactively labelled group to one end (5' or 3') of a DNA strand.

Endonucleases cleave bonds within a nucleic acid chain: they may be specific for RNA or for single-stranded or double-stranded DNA.

Episome is a plasmid integrated into bacterial DNA.

Exon is any segment of an interrupted gene that is represented in the mature RNA product.

Exonucleases cleave nucleotides one at a time from the end of a polynucleotide chain; they may be specific for either the 5' or 3' end of DNA or RNA.

Expressed Sequence Tag (EST) PCR based marker. Highly specific oligos (16-20-mers) are designed by using sequence information of a cDNA. The locus represents a functional gene and is located in an actively transcribed region of the genome.

Expression vector is a cloning vector designed so that a coding sequence inserted at a particular site will be transcribed and translated into protein.

Filter hybridization is performed by incubating a denatured DNA preparation immobilized on a nitrocellulose filter with a solution of radioactively labeled RNA or DNA.

Fingerprint of a DNA is a pattern of fragments (usually resolved on an electrophoretic gel) generated by cleavage with an enzyme.

Gene banks or libraries These are collections of individual genes, maintained separately in a host microbe. Their chief use is as the source material in which genes are found.

Genetic code is the correspondence between triplets in DNA (or RNA) and amino acids in protein.

Genomic library it represents the entire genome: genes and intergenic regions. The genes contain both the coding sequences and the adjacent regulatory sequences (i.e. the **promoter** which are responsible for determining both **when** (in relation to development or in relation to external stimuli and **where** (tissue or organ) the gene is expressed. Genomic libraries can be made in various types of vectors, depending on the purpose.

Homology The extent to which two nucleic acid molecules have the same nucleotide sequence or two proteins have the same amino acid sequence. Homology can be determined by direct comparisons of sequence data, or estimated by DNA-DNA or DNA-RNA hybridization. The degree of identity between chromosomes or chromosome segments also determine homology.

Hybridization The pairing of complementary DNA or RNA strands to give stable DNA-DNA or DNA-RNA duplexes. The efficiency of hybridization is a test of sequence homology.

Insert The piece of foreign DNA introduced into a vector molecule.

Inter-Simple Sequence Repeat amplification (ISSR) PCR-based multi-locus marker system using oligonucleotide primers homologous to SSR sequences (such as (GATA) n). To avoid stutterings these primers can be anchored to unique genomic sequences flanking the repeat. Amplification products are only obtained in case SSRs in opposite orientation are found within a PCR-able distance, with a flanking sequence matching the oligo's. 3'-anchoring gives better results than 5'-anchoring. Repeat polymorphisms within the SSR do not influence the chance for ISSR polymorphisms.

Inverse Polymerase Chain Reaction (IPCR) A technique to amplify genomic DNA flanking the insertion site of a transposon or T-DNA construct. The obtained flanking genomic DNA can be used as RFLP-probe to determine the map position of the insertion site of the contruct.

Kb is an abbreviation for 1000 base pairs of DNA or 1000 bases of RNA.

Ligase Chain Reaction (LCR) Technique to detect a single-base substitution in a known sequence. The ligase enzyme links two adjacent oligo's, which hybridize agains the target sequence. In case of a mismatch the ligation of the adjacent oligo's will fail. Ligated oligo's resemble the target sequence and are available as target sequence for unligated oligo's in subsequent cycles resulting in an exponential amplification of ligated oligos.

Leader is the nontranstated sequence at the 5' end of mRNA that precedes the initiation codon.

Leading strand of DNA is synthesized continuously in the 5'-3' direction.

Ligation is the formation of a phosphodiester bond to link two adjacent bases separated by a nick in one strand of a double helix of DNA (The term can also be applied to blunt-end ligation and to joining of RNA).

Map distance is measured as cM (centiMorgans) = per cent recombination (sometimes subject to adjustments).

Melting of DNA means its denaturation.

Melting temperature (T_m) is the midpoint of the temperature range over which DNA is denatured.

Miniprep A small-scale preparation of plasmid or phage DNA commonly used after cloning to analyze the DNA inserted into a cloning vector.

Near Isogenic Lines (NIL) Homozygous line, produced by recurrent selection, which only heterozygous and different from the recurrent parent with respect to a locus of interest.

Nick translation describes the ability of *E. coli* DNA polymerase 1 to use a nick as a starting point from which one strand of a duplex DNA can be degraded and replaced by resynthesis of new strand; is used to introduce radio-activity labelled nucleotides into DNA *in vitro.*

Northern blotting is a technique for transferring RNA from an agarose gel to a nitrocellulose filter on which it can be hybridized to a complementary DNA.

Open reading frame (ORP) contains a series of triplets coding for amino acids without any termination codons; sequence is (potentially) translatable into proteins.

Palindrome is **a** sequence of DNA that is the same when one strand is read left to right or the other is read right to left; consists of adjacent inverted repeats.

PCR (polymerase chain reaction) describes a technique in which cycles of denaturation, annealing with primer, and extension with DNA polymerase, are used to amplify the number of copies of a target DNA sequence by $>10^6$ times.

Phenotype is the appearance or other characteristics of an organism, resulting from the interaction of its genetic constitution with the environment.

Plasmid is an autonomous self-replicating extrachromosomal circular DNA.

Primer is a short sequence (often of RNA) that is paired with one strand of DNA and provides a free 3'-OH end at which a DNA polymerase starts synthesis of a deoxyribonucleotide chain.

Promoter is a region of DNA involved in binding of RNA polymerase to initiate transcription.

Proofreading refers to any mechanism for correcting errors in protein or nucleic acid synthesis that involves scrutiny of individual units after they have been added to the chain.

Quantitative Trait Locus (QTL) Single locus from a series of polygenes, which are involved in a quantitative trait.

Random Amplified Polymorphic DNA (RAPD) A PCR product that is obtained from genomic DNA using a single or a combination of typically 10-mer oligonucleotides. Alleles are visualized by the fragments that are amplified, separated on agarose gels and stained with EtBr. RAPDs show dominant inheritance. Variation is based on the position and orientation of primer-annealing sites and the interval they span

Recombinant DNA DNA molecules in which sequences that are not naturally contiguous have been placed next to each other by *in vitro* manipulations. The different sequences placed within a recombinant DNA molecule frequently are obtained from entirely different organisms.

Reporter gene is a coding unit whose product is easily assayed (such as chloramphenicol acetyl transferase, GUS); it may be connected to any promoter of interest so that expression of the gene can be used to assay promoter function.

Restriction enzymes recognize specific short sequences of (usually) unmethylated DNA and cleave the duplex (sometimes at target site, sometimes elsewhere, depending on type).

Restriction fragment length polymorphism (RFLP) refers to inherited differences in sites for restriction enzymes for example, caused by base changes in the target sites that result in differences in the lengths of the fragments produced by cleavage with the relevant restriction enzyme. RFLPs are used for genetic mapping to link the genome directly to a conventional genetic marker.

Recombinant Inbred Lines (RIL) An F4 population derived from a F2 by single seed descent (SSD). The consequence of infinite selfing is a doubling of the recombination frequency; a immortalization of the population and the exclusion of dominance variance due to heterozygousity.

Restriction Fragment Length Polymorphism (RFLP) A DNA fragment used to probe Southern blots of restricted genomic DNA from different strains of the same species. This results in the visualisation of variation in the size and/or number of detected restriction fragments generated from the different strains. The detected length variation is based on DNA sequence variation caused by insertions, deletions or changes in restriction sites.

Restriction map is a linear array of sites on DNA cleaved by various restriction enzymes.

Reverse transcription is synthesis of DNA on a template of RNA; accomplished by reverse transcriptase enzyme.

RNA polymerase An enzyme which catalyzes the synthesis of RNA from a DNA template in the absence of a 3' primer molecule. Transcription in prokaryotes is accomplished by a single RNA polymerase that synthesizes all classes of RNA molecules. Eukaryotes have three RNA polymerases, with different transcriptional specificities; RNA polymerase I (or A) synthesizes the large rRNA precursor; II (or B) synthesizes mRNA; III (or C) synthesizes tRNA and 5S rRNA species.

RNase is an enzyme whose substrate is RNA.

SI nuclease is an enzyme that specifically degrades unpaired (single-stranded) sequences of DNA.

Screening libraries A library has to be screened in order to find a clone. Where amino acid sequence of a specific protein is known, an oligonucleotide (short piece of DNA) can be designed and either used directly as a hybridization probe or for PCR. If no probe is available, a shotgun screening approach has to be taken, of which differential or subtractive hybridization are frequently used (see Old and Primrose for details).

Shotgun experiment is cloning of an entire genotype in the form of randomly generated fragments.

Selective Restriction Fragment Amplification (SRFA) A procedure taking oligo's those are elongated with selective bases. Because of these bases not all restiction fragments can be amplified in a PCR reaction.

Sequence Characterized Amplified Regions (SCAR) PCR based marker. A genomic DNA fragment at a single genetically defined locus that is identified by PCR amplification using a pair of specific (16-24-mer) oligonucleotice primers.

Sequence Tagged Microsatelite Site (STMS) A PCR-based marker. The DNA sequence of unique DNA flanking microsatellites (GATA) n, (CT) n etc. is used to construct 20-mer oligos. Amplification products spanning the microsatellite show rich allele divergency.

Sequence Tagged Site (STS) A PCR-based marker derived from a RFLP-probe, by constructing primers according to the end-sequence of the RFLP-probe

Silent mutations do not change the product of a gene.

Simple Sequence Repeats (SSR) Synonymous to STR or micro satellite repeats, in particular the dinucleotide repeats $(AC)_n$ $(AG)_n$ $(AT)_n$

Single Nucleotide Polymorphisms e (SNP) Polymorphism based on a nucleotide substitution. Used as a marker diagnostic for a specific trait. Often mentioned in connection with a technique which allows the specific recognition of the SNP

Single Stranded Confirmational Polymorphism (SSCP) Relies on secondary and tertiary structure differences between denatured and rapidly cooled amplified DNA fragments that differ slightly in their DNA sequence. Different SSCP alleles are resolved on non-denaturing acrylamide gels, usually at low temperatures. The ability to resolve alleles depends on the conditions of electrophoresis.

Southern blotting describes the procedure for transferring denatured DNA from an agarose gel to a nitrocellulose filter where it can be hybridized with a complementary nucleic acid.

Sticky ends are complementary single strands of DNA that protrude from opposite ends of a duplex or from ends of different duplex molecules, can be generated by staggered cuts in duplex DNA.

Structural gene code for any RNA or protein product other than a regulator.

Syntenic genetic loci lie on the same chromosome.

Termination codon is one of three triplet sequences. DAG (amber), UAA (ochre), or UGA that cause termination of protein synthesis; they are also called nonsense codons.

Ti plasmid A class of large, conjugative plasmids found in the soil bacterium *Agrobacterium tumefaciens* and responsible for the crown gall disease of broad-leaved (dicotyledonous) plants.

T_m is the abbreviation for melting temperature.

Transcription is synthesis of RNA on a DNA template.

Transformation of bacteria describes the acquisition of new genetic markers by incorporation of added DNA.

Transgenic plants are created by introducing new DNA sequences into the host DNA either through biolistic or through *Agrobacterium* mediated delivery.

Transgenic of cells, cell cultures, plants, or progeny; having received a foreign or modified gene by one of the various methods of transformation as described in this manual.

Transient expression The short-term detectable expression of the product of a gene which has been transferred into protoplasts, cells, or plants. Marker genes on the DNA construct (such as GUS or GFP β-glucuronidase) that are readily assayed in plant tissues are often used to determine whether DNA has been transferred into the cells. Transient expression indicates that the marker gene is transcribed and translated in the cells assayed, and does not imply that the gene or DNA construct has been heritably incorporated into the plant genome.

Translation is synthesis of protein on the mRNA template.

Transposon is a DNA sequence able to insert itself at a new location in the genome (without any sequence relationship with the target locus).

Upstream identifies sequences proceeding in the opposite direction from expression; for example, the bacterial promoter is upstream from the transcription unit, the initiation codon is upstream of the coding region.

Variable Number Tandem Repeat (VNTR) A genetic locus studied from a Southern blot, probed with a labelled minisatellite repeat. It's used in fingerprinting and forensic studies. It is not similar to a STMS. The core units vary in lenght from 11 to 60 bp.

Western blots A procedure for transfer of proteins after separation on a polyacrylamide gel to a suitable immobilizing matrix such as a nitrocellulose sheet. The proteins attached to the support matrix can then be probed with a specific antibody to identify a particular protein species.

Yeast Artificial Chromosome (YAC) A plasmid that contains all the sequences mecessary for stable maintenance in yeast (a centromere, DNA replication origin and telomeres), as well as a yeast selectable marker.

β-glucoronidase (GUS) Reporter gene from *E. coli; a* hydrolase that catalyzes the cleavage of a wide variety of β-glucuronides. One of the most commonly used is 5-bromo-4-chloro-3-indolyl glucuronide (X-GLUC), which shows dark blue colour in transformed cells and tissues.