SUMMER TRAINING ON BIOINFORMATICS

Faculty

Guest

- **Dr. K.P. Aravindan,** Department of Pathology, Govt. Medical College, Calicut
- **Dr. R.V. Omkumar,** Rajiv Gandhi Centre for Biotechnology, Thiruvananthapuram
- **Mr. Arun Anirudhan,** Rajiv Gandhi Centre for Biotechnology, Thiruvananthapuram

Institute

- **Dr. V.A. Parthasarathy,** Director

Division of Crop Improvement & Biotechnology

- **Dr. K. Nirmal Babu**
- **Dr. B. Sasikumar**
- **Dr. Johnson. K. George**
- **Ms. D. Minoo**
- **Mr. V.N. Jayakumar**
- **Mr. Sinoj Jose**
- **Mr. K. Praveen**

Division of Crop Production and Post Harvest Technology

- **Dr. B. Chempakam**

Social Science Section

- **Dr. K.N. Kurup**
- **Dr. M.S. Madan**
- **Dr. P. Rajeev**

Division of Crop Protection

- **Dr. A.Ishwara Bhat**
- **Mr. P.S. Hareesh**

Bioinformatics Centre

- **Mr. Santhosh. J. Eapen**
- **Mr. R. Senthil Kumar**
- **Ms. K.M. Beena**
- **Ms. P K Parvathy**

Library

- **Mr. P.A. Sheriff**
- **Ms. C.K. Sushama Devi**

Biosafety Laboratory Procedures

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:

1. Avoid distracting or startling other workers.
2. Do not allow practical jokes and horseplay at any time.
3. Use laboratory equipment only for its designated purpose.
4. 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:

1. Safety glasses with side shields provide the minimum protection acceptable for regular use.
2. 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.

1. Laboratory chemicals should never be tasted.
2. A pipet bulb or aspirator should be used to pipet chemicals or to start a siphon; pipetting should never be done by mouth.
3. 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

1. Toxic chemicals or compounds of unknown toxicity should never be smelled.
2. 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.
3. 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

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

5. 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:

6. 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.

Working with biohazardous and radioactive materials

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.

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.

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.

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.

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 meet all 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.

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. Check 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.

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

Precautions while using microorganisms

Infection followed by disease will depend on the microorganisms 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 are 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.

Examples of some biohazardous chemicals – precautions

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 irritant 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.

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.

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.

Isopropanol

1. This compound contains no ingredients at concentrations of 0.1% or greater that are 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

1. *CRC Handbook of Laboratory Safety*, Third Edition. A. K. Furr (Ed.), Chemical Rubber Company. 1990. 704 pp.
2. *Prudent Practices for Handling Hazardous Chemicals in Laboratories*. Prepared by the National Research Council. 1981. 291 pp.
3. *Prudent Practices in the Laboratory*. National Research Council. 1995. 427 pp.
4. *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.

Statistical Inference

K.N. Kurup

Let us suppose that in Agriculture experiments a standard fertilizer (SF) has been traditionally used which is known to result in an yield of 100kg with an SD of 10 kg. Suppose a new fertilizer is introduced which we call a test fertilizer (TF). The manufacturer of TF claims that TF is superior to SF. How we would ascertain or reject his claim? Statistical test of inference comes to our help in solving this problem.

Let us examine the problem: We have a standard fertilizer resulting in a yield of 100kg with an SD of 10kg. What we mean by this is that if we give the SF to ***a very large number*** of plants, the resulting yield measurements will have a mean $\mu = 100$ and a SD $\sigma = 10$. That is to say***, the measurement X is distributed (say) normally with mean*** μ ***= 100 and a SD,*** σ ***=10.***

If we can think of a situation where we have experimented TF with a large number of plants, the resulting yield also will have a normal distribution with an unknown mean and an unknown SD. But, for deciding the superiority or otherwise of TF we cannot conduct an experiment on infinite number of individuals nor can we repeat the experiments infinitely on a smaller number of individuals. We may at the most conduct an experiment involving a few plants (say 100).

This experiment can be regarded as a sample from the infinite population, which TF would have generated.

Without loss of generality we may assume SD in both populations to be the same.

ie . SD (SF) = SD(TF) = σ.

If TF is superior to SF, the mean of the population generated by TF (= $\mu 1$, say) may be greater than $\mu 0$, the mean of the population generated by SF.

ie., $\mu_1 > \mu_0$

Similarly, if TF is inferior to SF,

we have, $\mu_1 < \mu_0$

Let us, to start with, assume that effects of both SF and TF are the equal.

say, SF = TF.

ie., the population generated by both SF and TF are identical.

It follows that the sample which we have considered might as well be regarded as to have come from the SF.

Now let $\bar{x}$ be the sample mean.
We know that $E(\bar{x}) = \mu_1$

And, if, TF = SF,
We have, $E(\bar{x}) = \mu_1 = \mu_0$

ie., $\bar{x}$ is an unbiased estimate of μ_0

Note. Any statistic, which is taken as an estimate of a population parameter, is called an **unbiased estimate** if the mean of all possible sample values of the statistic is equal to the population parameter.

∴ It follows that the difference between $\bar{x}$ and μ_0 is a matter of sampling error and should be comparable to the SE of $\bar{x}$.

But we know,

$\bar{x}$ Follows normal distribution (sampling distribution of $\bar{x}$) with mean μ_0 and SD $\sigma/\sqrt{n}$.

It follows that,

$$Prob\left|\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right| \geq 3 \quad \text{is} \leq .0026$$

ie., $Prob\{|\bar{x}-\mu_0|\} \geq 3 \times \sigma/\sqrt{n} \leq .0026$

In other words, if the difference d = $\bar{x}$ - μ_0 is only due to random causes, then the probability that the difference exceeds (in numerical value) 3 times its SE is very small.

Thus if the observed difference is very large (say > 3 or more multiples of SE), we can with certain confidence conclude that the difference might be not merely due to sampling (due to random causes).

ie., to say $\bar{x}$ may not be an estimate of μ_0

ie., $E(\bar{x}) = \mu_1 \neq \mu_0$ ..... (1)

And alternatively, if the difference is very small, we may say with confidence that $\bar{x}$ is an estimate of μ_0 ie., $\bar{x}$ might occur in sampling from the population with mean = μ_0 ..... (2)

(1) means that there is likely a difference in the effects of the two fertilizers and (2) would mean that they are likely to be of the same effect.

Thus the problem reduces to one of ascertaining how large the observed difference should be so that we can say with confidence that the effects are the same or different.

Let us go back and find out what happens if we take a decision to regard the difference to be significant (which means that effects are not equal) if the observed difference exceeds 3 times the SE ($\bar{x}$). If there is no significant difference, the observed value of d=$|\bar{x} - \mu_0|$ can exceed 3×SE(d) with a probability .0026. ie., only in 26 out of 10,000 experiments we expect to get the differences exceeding 3×SE(d). ie., if our sample happens to be one that may occur in 26 cases out of 10,000 trials we have committed an error in our conclusion. Alternatively if it is the one, which belongs to those cases, which can occur in 9974 cases in a series of 10,000 trials, our conclusion is right. Thus there is a chance of committing error but its probability in only .0026. In other words, we have 99.74% of confidence in the rightness of our conclusion or 0.26% of doubt in the rightness of our conclusion.

Now the statement

$$P\{|\bar{x} - \mu_0|\} \geq 3 \times SE(d) \leq .0026$$ is the same as

$$P\left\{|z| = \left|\frac{\bar{x} - \mu_0}{SE(d)}\right|\right\} \geq 3 \text{ is } \leq .0026$$

And we know z follows standard normal distribution with men 0 and SD 1. We can refer to the statistical tables (available in standard text books) and ascertain the probability that z exceeds a given value.

The value 3 taken as cut off point in the example is quite arbitrary and subjective. But it gives a confidence of 99.74%. Thus if do not require so high a confidence we may take a lower value as cut off point or if we require still more confidence we may take a higher value. In any case, in deciding so a cut off point in the range of the distribution of z, we normally run into a few other problems.

Now let us recapitulate what we have done so far.

1. We have been using a standard fertilizer which is known to result in a yield of μ_0 with a SD=σ.
2. A new fertilizer is introduced and we are required to decide whether the new fertilizer is superior to the standard fertilizer or not.
3. In order to solve the problem, we conducted an experiment involving new fertilizer (called test fertilizer, TF) on a few plants (say 'n') and observed the mean yield ($\bar{x}$).
4. We assumed that the effects of the fertilizers viz. SF and TF are same.
5. We calculated a statistic z and studied its distribution under the assumption.
6. We decided on a cut off point on the range of z (say z_0) and decided to accept the assumption if the observed value of z is < z_0 and reject the assumption if observed z is > z_0.

Thus to start with, our aim is to conceptualize the information contained in a body of data. Though we have only a sample at our disposal, our aim is to conclude about the population from which the sample is assumed to be drawn. Repeat; our aim is not to infer anything about the sample but on the population from which the sample is assumed to be drawn. This might be made clear by reference to our example. We are not studying whether $\bar{x} = \mu_0$ or $\bar{x} \neq \mu_0$ but we are attempting to find out whether $E(\bar{x}) = \mu_0$ or $E(\bar{x}) \neq \mu_0$.

Thus on the basis of sample observations, to start with we make some assumption about the population from which the sample is assumed to be drawn. Necessarily these statements will be in terms of the (or on the) parameters of the 'population distribution'. Such an assumption is called Statistical hypothesis or simply hypothesis.

Eg. $H_0 : E(\bar{x}) = \mu_0$

Usually we make such assumption as $H_0 : \mu_1 = \mu_0$ or $H_0 : E(\bar{x}_1) = E(\bar{x}_2)$ or in ordinary words, there is no difference between .....

Such hypotheses are called null hypotheses. But when we make a null hypothesis viz. $H_0 : \mu_1 = \mu_0$ we can readily see there exists alternative hypotheses

$H_1 : \mu_1 \neq \mu_0$
$H_1 : \mu_1 < \mu_0$
$H_1 : \mu_1 > \mu_0$

And we normally test H_0 against one of these possible alternatives.

Thus the test procedure involves the following steps:

1. Set up the null hypothesis (H_0).
2. Decide on the alternative hypothesis (H_1).
3. Compute a statistic 't' on the assumption that H_0 is true [under null hypothesis].
4. Divide the range of 't' into 2 regions one called critical region or rejection region usually represented by Ω and another complementary region $\Omega`$ called acceptance region.
5. Reject H_0 if the sample value of 't' falls in Ω; otherwise accept H_0.

In doing so we may commit a few errors.

Procedure is illustrated as below.

	H_0 true	H_0 false
t in Ω	Reject H_0 ✗	Reject H_0
t in $\Omega`$	Accept H_0	Accept H_0 ✗

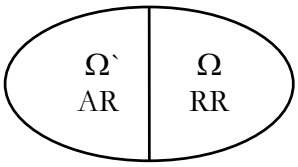
Thus if 't' falls in Ω we are rejecting the H_0 as false not knowing whether actually H_0 is true or false. And similarly we are accepting H_0 if 't' falls in $\Omega`$ without knowing actually whether H_0 is true or false.

Among the four cases enumerated we observe our procedure leads to 2 errors.

1. Rejecting H_0 when 't' falls in Ω while H_0 is true (rejecting a true hypothesis).
2. Accepting H_0 when 't' falls in $\Omega`$ while H_0 is false (accepting a false hypothesis).

Remaining two are obviously appropriate and desirable conclusions.

Sample space of 't'



The two errors cited are called Type-I and Type-II errors. Thus

> Type-I error occurs when we reject a true hypothesis (as false).
> Type-II error occurs when we accept a false hypothesis (as true).

Obviously, we need a procedure where both errors are minimised, if not eliminated. Now we observe that **type-I error can be reduced if the area of critical region is reduced**. But in doing so we observe that we are increasing the area of acceptance and therefore would lead to an increase in type-II error.

Let us define size of Type-I error as

$$\alpha = \text{Pro } \{t \text{ falls in } \Omega \text{ when } H_0 \text{ is true}\}$$
$$= P \{t \in \Omega / H_0\}$$

and size of Type-II error as

$$\beta = \text{Pro } \{ t \text{ falls in } \Omega` \text{ when } H_1 \text{ is true } \}$$
$$= P \{t \in \Omega` / H_1\}$$

when we try to reduce α, β will increase.

(Here a word of caution: by decreasing the critical region, the incidence of Type-II error does not increase proportionally as size of Type-II error depends on the distribution of 't' when H_1 is true).
This means that we may not be able to minimise both errors simultaneously and therefore, the course open to us is to run the risk of committing one of such errors fixed at certain level and then minimising other error. Usually we decide to fix the ***size of type-I error at certain predetermined level and then minimise the size of type-II error.***

We have $\beta = P \{t \in \Omega` / H_1 \}$ and $1 - \beta$ is called the **POWER OF THE TEST**. Thus when β, the size of type-II error is minimised we are really maximising the power of the test.

Note : Almost all the tests we use are those which are most powerful for a **chosen size of type-I error.** This chosen level of type-I error is called the **LEVEL OF SIGNIFICANCE OF TEST** (α). Usually in Agriculture, α is set at 5% or 1%. But with computer facilities available, now-a-days, in reporting the results of the experiment, the exact percentage (ie., the Prob { that 't' exceeds the computed value }) is given.

Critical region may be taken at right tail of the distribution or left tail of the distribution or may be split on either tails. This depends on the H_1, the alternative hypothesis.

Eg. If $H_0 : \mu_1 = \mu_0$ is tested against

H_1 : $\mu_1 \neq \mu_0$, then Ω is split on either tail (if α =0.5, we find t_1 and t_2 such that area to the right of t_1 is .025 and area to the left of t_2 is .025).

If H_1 is $\mu_1 < \mu_0$ then Ω can be taken as the left tail. If H_1 is $\mu_1 > \mu_0$, then Ω may be taken as right tail. These are accordingly called Two-tailed test or One-tailed test. One tailed test can be either left tailed or right tailed.

In actual practice, we compute the appropriate STATISTIC and the observed value is compared with the value of the statistic in the table corresponding to the chosen α. If observed value is greater than table value, we reject H_0 . If the observed value is less than the table value, we accept H_0.

Points to remember

H_0 : null hypothesis { by way of convention }
H_1 : alternative hypothesis "
Ω : critical region (or rejection area)
Ω` : complementary region (or acceptance region)

Type-I error : Error involved in rejecting H_0 : null hypothesis when it is true.
Type-II error : Error in accepting H_0, the null hypothesis, when it is false.

α = Size of type-I error (not type-I error)
= Prob {t $\in \Omega$ when H_0 is true}
= P {t $\in \Omega$ / H_0}

β = Pro {t $\in \Omega$` when H_1 is true}
= P {t $\in \Omega$` / H_1}

α = Size of type-I error is also called level of significance of test.

1 - β is called power of the test.
Critical region split on both tails - Two tailed test.

If critical region is taken completely on one tail, we have one-tailed test and this may be right tailed or left tailed.

Given a sample of 'n' observations x_1, x_2, x_3, ..., x_n. We are interested to test whether the sample has come from a population with a known μ. In other words, if $\bar{x}$ is the sample mean, we want to test the null hypothesis

$$H_0 : E(\bar{x}) = \mu$$

Case I : σ, the population SD known.

Method :
Let d= $\bar{x}$ - μ
Compute $z = \dfrac{\bar{x} - \mu}{SE(d)} = \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}}$
Compare the observed z against table z_α foe chosen α (say α = 0.05)

Reject H_0 if $z > z_\alpha$
Accept H_0 if $z < z_\alpha$

Case II : σ not known.

Method : Estimate σ^2 from the sample as follows :

Compute

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum_i \left(x_i - \bar{x}\right)^2$$

$$= \frac{1}{n-1}\left(\sum_i x_i^{\,2} - \frac{\left(\sum x_i\right)^2}{n}\right)$$

We have,

$$\hat{\sigma} = s = \sqrt{s^2}$$

Compute $t = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$ and compare 't' observed against t_α from 't' table for (n-1) d.f.

If observed $t > t_\alpha$ reject H_0
If observed $t < t_\alpha$ accept H_0

Comparison of two samples

Let us suppose we have two independent samples
$x_{11}, x_{12}, x_{13}, \ldots\ldots\ldots\ldots, x_{1n1}$
$x_{21}, x_{22}, x_{23}, \ldots\ldots\ldots\ldots, x_{2n2}$

We are interested to test whether the two samples have come from populations with same mean.

Let $E(\bar{x}_1) = \mu_1$
$E(\bar{x}_2) = \mu_2$

Our hypothesis is

$$E(\bar{x}_1) = E(\bar{x}_2)$$

ie $H_0 : \mu_1 = \mu_2$

Case I : σ_1 and σ_2 known but $\sigma_1 \neq \sigma_2$

Compute $z = \dfrac{\bar{x}_1 - \bar{x}_2}{SE(\bar{x}_1 - \bar{x}_2)}$

To compute $SE(\bar{x}_1 - \bar{x}_2)$,

variance $(\bar{x}_1 - \bar{x}_2) = V(\bar{x}_1) + V(\bar{x}_2)$

$$= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

$$\therefore SE(\bar{x}_1 - \bar{x}_2) = \sqrt{V(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

If $\sigma_1 = \sigma_2 = \sigma$, then,

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}} = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

If $\sigma_1 = \sigma_2, n_1 = n_2$,

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{n}} = \sqrt{\frac{2\sigma^2}{n}} = \sigma\sqrt{\frac{2}{n}}$$

Then compute $z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{1/n_1 + 1/n_2}}$ or

$z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{2/n}}$ as the case may be.

Conclusion as above.

Case II : If σ_1 and σ_2 not known but **equal,**

(If they are not equal, consult a statistician)

Estimate σ_1^2 and σ_2^2 from samples as follows :

$$\hat{\sigma}_1^2 = s_1^2 = \frac{1}{n_1 - 1}\left[\sum_1 x_{1i}^2 - \frac{\left(\sum x_{1i}\right)^2}{n_1}\right]$$

$$s_2^2 = \frac{1}{n_2 - 1}\left[\sum_2 x_{2i}^2 - \frac{\left(\sum x_{2i}\right)^2}{n_2}\right]$$

If σ_1 and σ_2 are assumed to be the same, we may obtain a pooled estimate of σ as follows.

Pooled estimate of σ^2 is given by

$$s^2 = \frac{(n_1 - 1)s_1^{\,2} + (n_2 - 1)s_2^{\,2}}{n_1 + n_2 - 2}$$

$$= \frac{\sum_1 (x_{1i} - \bar{x}_1)^2 + \sum_2 (x_{2i} - \bar{x}_2)^2}{n_1 + n_2 - 2}$$

$$= \frac{\sum x_{1i}^{\,2} - \frac{\left(\sum x_{1i}\right)^2}{n_1} + \sum x_{2i}^{\,2} - \frac{\left(\sum x_{2i}\right)^2}{n_2}}{n_1 + n_2 - 2}$$

$$\hat{\sigma} = s = \sqrt{s^2}$$

Compute $t = \dfrac{\bar{x}_1 - \bar{x}_2}{s\sqrt{1/n_1 + 1/n_2}}$

If $n_1 = n_2$ then,

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{2/n}}$$

Compare computed t against t_α for $(n_1 + n_2 - 2)$ d.f.

Points to remember:

i. If σ known, use 'z' test.
ii. If σ unknown, use 't' test.

In (ii) above, if n>30, compute 't' but can be referred to Z table.

Documentation and Information Retrieval

P A Sheriff

Information science is a discipline that investigates the properties and behavior of information, the forces governing the flow of information, and the means of processing information for optimum accessibility and usability. It is concerned with that body of knowledge relating to the origination, collection, organization, storage, retrieval, interpretation, transmission, transformation, and utilization of information.

What is Information Technology?

Information technology in a broad sense points out to the use of technologies in the field on information access and documentation. As such IT encompasses the

i. Computer technologies
ii. Communication technologies-satellite communication and global network
iii. Photographic techniques and

Printing and publishing technologies- which may include electronic and optical publishing – electronic mails, CD-ROMs, Optical disc, imaging Technology, etc. It facilitates automatic collection, storage, processing, retrieval and dissemination of information.

Information retrieval

It is important to distinguish between two different views of the IR problem: a computer-centered one and a human-centered one.

In the computer-centered view, the IR problem consists mainly of building up efficient indexes, processing user queries with high performance, and developing ranking algorithms, which improve the `quality' of the answer set. In the human-centered view, the IR problem consists mainly of studying the behavior of the user, of understanding his main needs, and of determining how such understanding affects the organization and operation of the retrieval system. We focus mainly on the computer-centered view of the IR problem.

Information retrieval (IR) deals with the representation, storage, organization of, and access to information items. The representation and organization of the information items should provide the user with easy access to the information in which he is interested

Information Retrieval in the Library

Libraries were among the first institutions to adopt IR systems for retrieving information. In the first generation, such systems consisted basically of an automation of previous technologies (such as card catalogs) and basically allowed searches based on author name and title. In the second generation, increased search functionality was added which allowed searching by subject headings, by keywords, and some more complex query facilities. In the third generation, which is currently being deployed, the focus is on improved advanced networking technology, Multimedia technology, CD ROM technology etc.

Sources of Information

Primary sources of Information

- Scientific journals
- Research reports and bulletins
- Proceedings of conference or congress
- Theses or dissertations
- Patents and standards

Secondary sources of Information

- Bibliographic and catalogues
- Handbooks and manuals
- Dictionaries, encyclopedias, glossaries and thesaurus
- Directories and yearbooks
- Atlases, maps and gazetteers
- Indexes and abstracting journals
- Advances and annual reviews
- Monographs, textbooks and treatises
- Statistical publications
- Biographies

Tertiary sources of Information

- Guidebooks to literature like Horticultural literature, Biological literature and Economic literature
- Bibliography of bibliographies
- Directory of directories

Primary sources of Information

A "primary" publication has been defined as that which reports the results of original research or observation.

1. *Scientific Journals*

Scientific and technological research journals or periodicals are the most important medium for the publication of the results of research or observation.

a. A periodical is published under one regular title and appears or released at a set frequency, e.g. weekly, fortnightly, monthly, quarterly, four monthly, six monthly or annually, eg. Nature (Macmillan, London)- a weekly and Indian Journal of Agricultural Sciences (ICAR, New Delhi) is a monthly journal.
b. A periodical is envisaged as continuing indefinitely until changes its title and frequency.
c. Each issue of periodical normally contains more than one article and its issues often contain other kinds of material like book reviews, announcements or reports of activities of issuing body or announcements about new books recently published or forthcoming publications.

Some conference proceedings are also published regularly like journals and share all these common characteristics of periodicals, for example ' Philosophical Transactions of the royal

society. London and Proceedings of the National Academy of Sciences of the United States of America (Fortnightly). These are published regularly as journals.

2. *Proceedings of Conferences, Congresses etc.*

Proceedings of conferences, Congresses, etc have their own kind of importance, as media for disseminating the results of scientific research. Often outstanding figures in a scientific research field will be invited to speak on the state of the art in their own areas of research.

3. Patents & Standards

Patents are descriptions of inventions of industrial, technological, chemical processes and products holding legal safeguards from governmental authority. The patent holder gets the exclusive license to produce, distribute and sell the invented item. Protection is granted for a set period of time. Patent documents thus constitute a record of the outcome of original research and development work.

Standards provide specifications and measurements in order to maintain quality of the industrial products or commodities. Bureau of Indian Standards, formerly Indian Standards institution (New Delhi0 is a unique organization in the world. It has released several thousand Indian standards and a lot many in the field of agriculture.

Secondary sources of Information

1. Bibliography

It is a list of publications. The list may contain articles/research papers or books/monographs and other macro-documents or both. It may give simple citations or along with annotations on spices.

2. Catalogue

A catalogue is a list of publications in holding. A Library catalogue, generally, available on 3'x5" size (catalogue) cards, represents the publications available in the holdings or collection of a library.

3. Handbook /Manual

Handbook, as it indicates, is a publication, which is supposed to be at hand while one is working in laboratory or office. A handbook contains descriptive information about a subject, event, or object. It contains methods, formulas, techniques, procedures, data along with figures, illustrations, etc. agricultural Engineering handbook, Edited by C.B. Richery,. New York, Mc Graw hill. 1981.

A manual is similar to a handbook. It also contains data, methods, description, but generally it gives step-by-step laboratory techniques, procedure of experiments, components and constituents of material and method of working.

4. Dictionary

A dictionary is a list of words arranged in a systematic manner-letter by letter or digit by digit. There are three types of dictionaries namely

a. General dictionaries
b. Subject dictionaries and
c. Biographical dictionaries.

A general dictionary lists words and gives the etymology or genesis of the term, its grammar (noun, pronoun, verb etc.), meaning, equivalent word, its various usages. Eg. Oxford English Dictionary.

5. *Glossary*

A glossary is a list of standard technical terms arranged alphabetically. They provide equivalent terms. They are very useful source for the persons doing translation work.

6. *Thesaurus*

Thesaurus is a list of standard terms, e.g. Webster's Collegiate Thesaurus.. Thesaurus is compiled or produced on the basis of standard scientific and technical terms used, while indexing literature in electronic and technical terms used, while indexing literature in electronic form in computer memory.

7. *Encyclopaedia*

It contains individual articles contributed by subject specialists or authoritative persons. There are two types of encyclopedias.

General: May be in single volume or multi volume, but they cover topics of general nature, which may belong to society, culture, art and science, biographies, description of objects, events, case history or history and geography etc. Such encyclopedias are available in good number. Eg. Encyclopedia Britanica, Encyclopedia Americana, World book Encyclopedia etc.

Subject Encyclopedia: Subject encyclopedias cover very specific or broad subject or general discipline, namely McGraw-Hill Encyclopedia of Science & Technology, McGraw-Hill Encyclopedia of Food, Agriculture and Nutrition.

8. *Directory and Yearbook*

Directories and yearbooks are published annually. A directory consists of addresses of individuals, name of their establishments and organization where they work, notes on the activities and interest and objectives.

CAB list of Research workers in agriculture sciences in the commonwealth, issued periodically, gives the names of scientists working 'commonwealth countries in different research organizations and their specific field of research work. ICAR, New Delhi publish a 'directory of Research workers in India; 1986. The yearbook is more descriptive and incorporates also directory information, namely ' Times of India Directory'

9. *Atlas and Maps*

Atlas contains maps. A map may be political, geographical presentation of information or data, namely for physical geography. Maps provide a vehicle to present complex of data – economic, geographic, political or agricultural in space and time through scale and cartographic-techniques.

10. *Gazetteer*

Gazetteers are, generally, governmental descriptive accounts of land and its people, soil and water, agriculture, industry, commerce and trade, education energy, communication, defence etc.

Indexing and abstracting Services

Literally an index is a pointer. An Index in a book acts as a pointer to the relevant or particular information, page or subject mater. Indexes in themselves are systematically arranged documents providing specific type or specific nature of information and indicating the same published in document, giving its full reference for quick and pin pointed location by an user. For covering a specific subject or nature of information, various kings of indexed have been seen. In general textbooks and monographs, generally, there is an author index and a subject indeed appended to the book. Eg: Author index, Subject index, Cumulative index etc.

Unlike indexing services, the abstracting services provide abstracts. Commonwealth Agricultural Bureau International (CABI), Wallingford, England is a cooperative organization to which Commonwealth countries, including India, contribute financially and helps in its administration. The basic objective of the CABI is the collection and dissemination of information in the secondary form of information.

Information retrieval in the modern context

The Web

The Web is becoming a universal repository of human knowledge and culture, which has allowed unprecedented sharing of ideas and information in a scale never seen before. Any user can create his own Web documents and make them point to any other Web documents without restrictions. This is a key aspect because it turns the Web into a new publishing medium accessible to everybody. As an immediate consequence, any Web user can push his personal agenda with little effort and almost at no cost. . Furthermore, it is causing a revolution in the way people use computers and perform their daily tasks. For instance, home shopping and home banking are becoming very popular and have generated several hundred million dollars in revenues.

Despite so much success, the Web has introduced new problems of its own. Finding useful information on the Web is frequently a tedious and difficult task. For instance, to satisfy his information need, the user might navigate the space of Web links (i.e., the hyperspace) searching for information of interest.

Agricultural Databases

Database: Data is a collection of "information in a form suitable for processing by a computer, such as digital representation of text, numbers, graphic images or sounds". A database is a collection of inter-related data of different types stored in a manner, which enables efficient storage and retrieval. Database at present is better understood with reference to electronic databases, and therefore it must be stored on a direct access storage device- a computer's Central Processing Unit (CPU), to be retrieved in different required forms. It is a collection of related objects, including tables, forms, reports, queries and scripts, created or organized by a database management system (DBMS).

CABI database: CAB International, London has built a computerized bibliographic database with abstracts since 1973. It is the largest professionally developed database covering world wide issues in agriculture, forestry, dairy, animal & veterinary science, food and nutrition, etc. CABI database is published in 47 printed abstracting journals as well as CD-ROMs.

Agris Database: Agris (FAO, Rome) has built a bibliographic database since 1975. Upto 1985 it was purely a bibliographic reference database, but from 1986 it started indexing abstracts also. Agris provides worldwide bibliographic coverage of agriculture, forestry, fishery, animal and veterinary science, food and nutrition, environment etc. The Agris database is fully computerized and is accessible on Internet. It is also published on CD-ROM as well as in print form under title "Agrindex" (monthly service).

Agricola: Agriculture on-line Access (Agricola) is a bibliographic database created by the National Agricultural Library (USA) since 1970. Today, it is the most comprehensive source of bibliographic citations covering wide range of agricultural and allied subjects - plant sciences, agriculture, animal sciences, aquaculture, veterinary science, forestry, nutrition etc. to journal articles, monographs, theses, patents, reports, AV materials etc.

Biosis database: Biosis is the World's most comprehensive bibliographic database in the field of biosciences, including life sciences, genetics, toxicology and zoological literature. Biosis database on hard disk contains over 7 million bibliographic records with abstracts since 1985 and is available online as well as on CD-ROMs.

Biotechnology database: Derwent Publications Ltd. Since 1982 have created a computerized database in the field of biotechnology. The online database is known as "Biotechnology abstracts" and its print version is also issued under title "Biotechnology abstracts". It contains abstracts in the fields of biology, biotechnology, chemicals, food and agriculture, genetics, health sciences, pharmaceuticals, plant genetics and breeding and wastes etc.

Chemical Abstracts Service (CAS): It is the largest database in the world on chemical science and technology developed by the American Chemical society (Ohio, USA).

Food & Human Nutrition Database: FAO of the United Nations since 1975 have created this bibliographic database, which covers food and agriculture, health sciences and nutrition etc.

Food science and technology database (FSTA): International Food Information Service (IFIS) since 1969 has a world leading food Science and Technology database which covers various aspects of feed science, food products and food processing including biotechnology, economics, manufacturing, legislation and packaging for any commodity. Based on information from scientific journals in the line, FSTA database contains abstracts.

CD-ROM Databases

Today a large number of companies around the world are publishing scientific data, bibliographic and full text on CD-ROMS. Silver Platter publishes the largest number of CD-ROM databases in the field of agro-biological sciences and other branches of science & technology, medicine, social sciences and humanities. Some of the important databases on CD-ROMs published in the field of agriculture and allied sciences are mentioned here.

Biosis Databases on CD-ROMs available are

-Biological abstracts on CD (1985-)
-Zoological Record on CD (1985-)
-Foods Intelligence on CD (1985-)
-Toxline Plus on CD (1985-)

CAB International CDs

CAB International has published database on CD-ROMs in 5 volumes since 1973 to date. They contain records with abstracts on different aspects of agriculture like agronomy, botany, horticulture, fertilizer, entomology, ecology and environment, economics and aquaculture. CABI also has issued specialty database CD-ROMs under its spectrum Series just like its by-product commodity based speciality abstracts, namely.

Ag Econ CD, CAB CD, Crop CD, HORT CD, Crop Pest CD, Plant gene CD, Soil CD, Tree CD, Vet CD etc.

Current contents on CD-ROM

The fastest Current Awareness Service (CAS) "Current contents" (CC) issued in different series viz. CC: Agriculture, Biology & environmental sciences etc. Apart from its Diskette edition it is available on CD-ROM. CC on CD-ROM is issued weekly. Each weekly CD-ROM presents an entire year's worth of current contents published data and thus provides the facility of browsing or searching 52 continuous weeks of information at a time. Thus, it facilitates the data/information search on a computer network.

Agrivoc- Agricultural Vocabulary

It is a multi-lingual agricultural thesaurus developed jointly by FAO and the commission for the European communities (CEC). Each citation in Agrindex included the set of descriptors (Concept/term/keyword) by which it has been indexed. With the use of these terms/descriptors document can be found in the subject index as well as document can be retrieved from the computerized system.

The Web and Digital Libraries

Three dramatic and fundamental changes have occurred due to the advances in modern computer technology and the boom of the Web. First, it became a lot cheaper to have access to various sources of information. This allows reaching a wider audience than ever possible before. Second, the advances in all kinds of digital communication provided greater access to networks. This implies that the information source is available even if distantly located and that the access can be done quickly (frequently, in a few seconds). Third, the freedom to post whatever information someone judges useful has greatly contributed to the popularity of the Web. For the first time in history, many people have free access to a large publishing medium.

Fundamentally, low cost, greater access, and publishing freedom have allowed people to use the Web (and modern digital libraries) as a highly interactive medium. Such interactivity allows people to exchange messages, photos, documents, software, videos, and to `chat' in a convenient and low cost fashion. Frequently, people are willing to exchange information as long as it does not become public. The reasons are many but the most common one is to protect oneself against misuse of private information by third parties. Thus, privacy is another issue which affects the deployment of the Web and which has not been properly addressed yet.

INSDOC

Indian National Scientific Documentation Centre (INSDOC) is a premier information organization dealing with library, documentation and information science, technology, services and systems. It is a national laboratory under the Council of Scientific and Industrial Research (CSIR) providing information and documentation services both at the national and the international level. INSDOCs activities fall under five broad categories:

1. Services, Products and Publications
2. Projects in Competency Areas
3. National Information Resource
4. Education and Training
5. International Collaboration

Services Offered

- Document copy supply
- Literature search from international/national databases.
- Foreign language translation/interpretation.
- Retrospective conversion of catalogue or library catalogue management.
- Contents, Abstracts, Photocopies Service (CAPS).
- Microfilming & slide making.
- Desk top publishing & offset printing
- Compilation of databases and directories.
- Electronic imaging of archival records.
- Standing Order Abstract Service (SOAS).
- Electronic mail service.
- Chemical Abstracts Keyword Index Service (CAKIS)

RECENT DEVELOPMENTS

Networking technology

Computer networks provide wide area communication facilities having protocols, such as-

- Public Switched Telephone network (PSTN)
- Public Data Network(PDN)
- Satellite Data Networks (SDN)
- Network of Leased Lines (NLL)
- Integrated Service digital Network (ISDN)

The computers of two different institutions/libraries are interconnected to each other by a wire path, which is dedicated to these systems for the entire duration of data communication.

CD-ROM Technology

CD-ROM a compact disc read Only memory of 4.75 inch diameter, can store 600 megabytes of data equal to 1650 floppy (each floppy can store 360 kilobytes) or 3,00,000 typed pages (1 page =2 K.B. Bytes). The first CD-ROM was manufactured in 1985. It provides instant random access to huge amount of information. Since it can store large volume of structured data- bibliographic, full-text, multi-media and image types, the CD-Rom has brought about a revolution in the information world. Today, a large number of companies around the world

are publishing scientific data, bibliographic and full text information on CD-ROMs following ISO standard 9660. the CD-ROM is read with the use of a CD-ROM Drive and a personal computer having interface card.

Multimedia Technology

Multimedia is a unique development in the field of information technology. It is a collection of multisensory presentation technologies combined through a common user interface into an information delivery system. Presentation data types can include text, images, graphics, audio, animation and video. Computer animations, video conferencing, desktop publishing, 3 dimensional exchange of information are new developments through multimedia technology.

INTERNET

The Internet connectively requires hardware (like and IBM compatible PC, a modem and a telephone line or V-SAT), operating system like Window 98 and application software like Netscape or telnet, window/NT, IRIX-commn. & WEB software.

Internet facilities make possible

- To exchange information quickly and conveniently,
- To access experts/peers in all fields,
- To receive regular updates on topics of interests,
- To build teams and enhance team work across, geographic distance.
- To access to extremely large collection of information online.
- To import and export information.

Library applications via Internet

- Acquisition- ordering publication, reviews etc.
- Cataloguing & classification- downloading of catalogue data, etc.
- Remote access to information
- Electronic books& journals- reading, down-loading , printing
- Current awareness service
- Reference service etc.

LIBRARY SERVICES

Inter Library Loan

Libraries borrow from each other on loan those publications which are not available in their holdings and thus fulfill information needs of their users.

Current Awareness Service (CAS)

CAS has been defined by B. Guha as " a device of the information system through which the users of information can be informed promptly as soon as possible after publication, but before absorption into the comprehensive secondary sources; of current literature on a broad subject field or on an area in which a group of persons are interested and presented in a manner, volume and rhythm intended to facilitate or cultivate current approach to information."

Contents Page Service
Some libraries instead of 'Journal circulation' or 'List of Periodicals Received', make photocopies of contents pages of journals received in the library and circulate among their users.

List of Periodicals Received
Some libraries inform their users about the periodicals received in the library immediately after registering them in the kardex in some other form.

List of Additions
Similar to the list of periodicals, libraries also bring out list of new additions consisting of newly acquired books, bulletins, theses, patents, special issues of advances, reviews, reports , proceedings, pamphlets, etc. Such publications are brought out at monthly, bi-monthly, quarterly frequencies.

Library Bulletins and Newsletters
A number of libraries bring out their own timely bulletins and newsletters which contain information like new acquisitions, addition of new journals, services, database searches, bibliographies compiled, information on conferences, seminars, meetings, visitors etc.

News Clipping Service
Some Libraries provide news clipping service to their user e.g. IARI Library, brings out one "Agricultural Press Bulletin" with full text every month and maintains those clippings in bound form for future reference.

SDI-Selective Dissemination of Information
It means that for an individual and is restricted exclusively to the area of his interest. The user is served only with that information which he needs. What does he not require is not served to him.

Bibliographical and Documentation Service
In research organizations, say research libraries, the documentation activities are carried on a regular basis. They compile select or topical bibliographies and documentation lists and circulates them among their own users as well as distribute to selected organizations belonging to that specific field.

Computer Based Services
With the application of Information Technology a large amount of scientific information in electronic form- electronic books, CD-ROMs, floppy-diskettes, Magnetic tapes and on computer hard disks is now available through computer networks. Libraries having these electronic facilities provide CD-ROMs database searches, computerized database searches form their own PCs and through computers on Local Area network, regional and national network and Internet service.

Resource Sharing

The libraries with their limited financial resources face serious problems in managing enormous flow of information. Libraries in the 21st century in India have to pool their resources for procuring and sharing the information sources which are cost effective. With the fast changing modern technologies for retrieval of information, the libraries should be recognized and reshaped as digital libraries for better information services.

Resource sharing through electronic media is for achieving speed in access, reducing time to search the information gaining, shorter storage space, finding simple means of selecting materials and for eliminating the need to visit a library.

Through networking

Now-a-days resource sharing computerized networking is essential and inevitable in the libraries due to tremendous changed attitude of the users and information technology. The resource sharing will depend upon the participating libraries to develop their networking and environment like resource to share, reveal the resources, response to query, reciprocity and reckon with all eventualities.

Through Internet

Internet is playing a very significant role in the resource sharing. The Internet provides a whole host of advantages to users, authors, and publishers in terms of speed of publication, speed of access, wide distribution, sharing of information resources, feedback online and others. The Internet has enabled global connectivity of computer and the development of various tools and techniques for networked information provision and access, E-mail, FTP, TELNET, USENET, World Wide Web etc are the examples.

Resource sharing at IISR

A meeting of directors and librarians of CPCRI, IISR and NRCC, Puttur was held to explore the possibility of exchange of foreign journals between these three libraries. It has been decided to cut the Journals costing more than Rs.20,000 per annum and get them on circulation from Other Libraries. The journals will be displayed in this library for a period of one month and have to be returned to the Institute in which it has been subscribed.

Similarly the procurement of CAB CDs were divided and will be getting only 2 CDs ie Hort CD and pest CD costing approximately 3.5 lakhs instead of Rs. 8.8 lakhs spent last year. The remaining CDs will be procured by CPCRI and NRCC. If the scientist require any information from the CDs subscribed by CPCRI and NRCC they have to send the keyword to CPCRI and get the information by evening of the same day.

It was decided to scan and host the content pages of journals received at each center in their web site which will be made accessible using a password to the Librarians of participating centers. The stations that do not have their won web site can depend on IISR or CPCRI web sites for hosting the data bases. The concerned libraries can send the scanned image to these institutes. It was decided to make the contents pages accessible. Necessary link may be given between websites of IISR, CPCRI and Bioinformatics centers.

Library of the future would emerge not as the mere storehouse of information but to facilitate easy access to information. What is needed may not be physically stored in one library but can be available on the Internet system.

References

1. Internet (www.google.com)

2. Lal, C. 1998: *Agricultural Libraries and Information Systems: A Handbook for User*. Published By RK Techno Science Agency. New Delhi.
3. Prakash, J.A. 2001: Resource sharing through electronic media. *Indian Journal of Information Library & Society* 14(3-4): 147-150.

IPR and Patenting

B. Sasikumar

Intellectual Property

Intellectual property right (IPR) & patenting are products of globalization & liberalization. Even common man today asks 'Why can't we go for patenting this or that?' Technology is increasingly becoming a tradable asset and a dominating factor deciding competitiveness. Therefore, high investment is being made on R&D to promote inventiveness and to develop new technology for enhancing the competitiveness. Under this circumstances protection of the results of R&D becomes important. Like the rights on the movable and immovable property, intellectual property can also be protected. This is called intellectual property rights (IPR). Knowledge in S&T is increasingly examined for its potential to earn money. Inclusion of IPR in the form of trade related intellectual property (TRIPs) in the GATT agreement is the consequence of this paradigm shift in the mindset of the developed economies.

What is Patent?

Patent is one form of intellectual property right. Patent is a legal right granted by the State to an inventor (person, persons or organization) valid within its territorial limits for a new invention employing scientific and technical knowledge. The main criteria for securing patents are novelty, inventive step and utility. Patents confer on its holder exclusive right for a limited period in consideration of the public disclosure of a new invention. In the TRIPs there are about eight forms of intellectual property rights such as Copy Rights & Related Rights, Trade Marks, Geographical Indications, Industrial designs, Lay out design of Integrated Circuits, Protection of undisclosed informatics etc in addition to Patenting

Advantages of Patent System

Patent system provides the inventor with an exclusive right to use the invention of manufacturing, marketing, renting, selling and licensing out this right to others for a specified period. This exclusiveness allows a monopoly to inventor and opportunity to make profit. This reward may encourage more creative work, which in turn bring more investment in new technology. Patenting brings the new knowledge in the public domain rather than keeping it as a personal or State secrecy. Patent documents are open and easily accessible. Such dissemination of new knowledge kindles further inventions therefore provides immense incentive to inventiveness.

Patent Legislation

Indian Patent Act 1970 is the basis for patenting in the country. This has undergone amendments. Patent (amendment) Act 1999 is now in effect.

Which Invention is Patentable?

In a general sense any invention concerning from paperclip to computer chip are patentable, provided it has novelty, involvement of inventive step and utility. Novelty means that the invention should not be known anywhere in the world as on the date of filing the patent application. Inventive step means that the invention involved must not be obvious to a person reasonably skilled in the field. Utility means the invention should be economically and technologically viable for industrial (or large scale) application. This also means social utility. An invention satisfying all these three criteria is patentable, provided such an invention is not excluded from the national patent law. Simplicity of an invention *per se* cannot be an objection for securing a patent. Thus, very simple inventions are also patentable. It is also immaterial whether the invention had happened by an accident or under an unexpected situation. It can be an incremental improvement involving an inventive step over the existing patented products of processes. In fact, majority of the patents granted across the world would fall in the latter category. Any improvement means an addition, simplification, or technology provided the improvement qualifies from the triple factor test.

What can/can't be patented

Patent grants are accorded by the patent Office, Govt. of India, for any invention on any one of the following five subjects provided the invention is 'new' in the global sense, 'non-obvious' in the sense that it involves an inventive step, and 'useful' in respect of its industrial applicability:

1. A machine/device
2. A process
3. A product/article
4. Composition of matter
5. Any improvement of any of above

Patents are not granted in India in respect of the following:

1. Claims which are frivolous, or well established natural law
2. Any invention, the primary or indented use or commercial exploitation of which could be contrary to law or morality or which causes serious prejudice to humans
3. The mere discovery of a scientific principle or the formulation of an abstract theory
4. Mere discovery of any new property of any new use for known substance or of the mere use of a known process medicine or apparatus, unless such know process results in a new product or employs at least one new variant.
5. Substance obtained by mere admixture resulting only in aggregation of prosperities of the components thereof or a process for producing such substance, and
6. Mere arrangement or re-arrangement or duplication of known devices each functioning independently of one another in a known way.

Further, it may be expressly noted that the India law also does not provide for patenting for the following subjects:

1. Method of agriculture of horticulture, or

2. Any process for medicinal, surgical, curative, prophylactic or other treatment of animals or plants to render them free of diseases or to increase their economic value or that of their products, or
3. Plant varieties, strains of animals or birds, or microorganisms per se.

Nevertheless, intellectual property protection for plant varieties is visualized under the 'Protection of Plant Varieties and Farmers' Rights Act, 2001' and some protectable subjects can be covered under Geographical Indications, Trademarks, Copyrights, Designs, etc; as per the respective Act.

Requirements and Precautions for Getting Patent

The Primary requirements are satisfaction to the triple factor test, novelty of invention, involvement of inventive step and social utility. The invention claimed should not exist in public domain under use. It also should not contain any claim containing knowledge, oral or otherwise, available within any local or indigenous community in any country. To satisfy novelty, the invention is permissible after filing provisional specifications in special situation. Normally publication is not desirable even after filing of the complete specification, in case the patenting of the same invention is anticipated in other countries.

A patent application may be provisional, complete or international. Application submitted with provisional specifications is called provisional application. This is done to gain priority in the date of filing. It is filled in Form No.2 and in normal manner except in sections "specifications" and "claim". The title should give fair indication of the art of industry in which the invention relates. This aspect should be made further clear in the description. This text may be brief, unambiguous, without fancy expressions, matter of fact and definite. In the specification part, as much information available in the invention at that point of time and at least to an extent that the given description should be adequate to discern the invention is given. Wherever adequate description is not possible without the help of drawings, such drawings should be furnished. It is not addable to include statement of claim in the provisional application.

Wherever provisional application is filled the complete application should be filled within 12 months from the date of filing of the former. Additional 4-month time may be allowed on penal charges per month. On filing complete application, the applicant can request before the acceptance of the complete specification to cancel provisional application and post-date the complete application. Complete application is also filed in the form No 2. Please note that with filing of complete application, the provisional application does not become infructuous or replaced. Both these applications are independent documents. Hence as far as the essential aspects of the invention are concerned there should be substantial similarity. If more research data to strengthen the complete specification are necessary it may be included. Non-disclosure of any vital information about the invention is liable to get the patent deemed invalid *ab initio.*

The complete specifications should disclose the best mode of carrying out the invention known to the inventor. These details should be sufficient to enable a person reasonably skilled in the art to independently work the invention without the assistance of the applicant. Whenever biological material is used for the invention its description and geographical origin should be not wrongly mentioned. In cases such patent application is filed in any other

country, such information should not be suppressed in the application. It is advisable not be disclose the patent know-how along with complete specification.

The claim or claims of a complete specification must relate to single invention. It should define the novel part of the invention disclosed for which legal protection is sought. This should define the scope and range of claims in concise, precise and accurate manner striking a comparison with the existing knowledge.

In case of a complete application is submitted after submission of provisional application, a declaration of inventorship of the invention should be included in Form No.5 along with complete application.

The last part of the text is abstract. This must provide technical information on the invention and claims, again in clear and crisp manner. This is likely to be edited at the office of the Controller.

Term of Patent and Patent Renewal

Term of plant is 20 years. It is counted from the date of filing of complete or provisional application. Patent has to be periodically renewed during the term period by the paying the specified renewal fees

Patents of Addition

Subject to the providence of IPA, 1970 and its amendments an application, if so requires, shall apply to change the specifications and claims in respect of any improvement in or modifications of the invention described in the granted patented document. This shall be allowed on the discretion of the Controller of patent. A patent of addition shall not be sealed before sealing the main patent.

Determination of Inventor(s)

To avoid intrigue, dispute and patent approval delay, it is advisable to name all those persons who contributed intellectually towards the development of patentable features of the invention as the inventors. These contributions should be discernible. This does not include mere assistance without intellectual input to the cause. Persons giving the idea which forms the germ of the invention or helping to solve any problem during invention with intellectual solutions should become inventors, even if they may not involve in carrying out the experiments and developing the invention.

Maintenance of Inventor's Diary

It is normal that all scientific studies meticulously maintain regular record of experiments. This requirement is very important for all patentable inventions in two respects. This would provide a dependable basis to determine the persons who contributed intellectually for the inventive work. In situations of later dispute on conclusion of inventors this diary may prove helpful. Another and most important use is that of the diary when maintained properly will help in establishing the date of conception of invention, if a dispute arises a later date or when another term or individual comes up simultaneously with identical invention for patent.

Furnishing Models or Samples

Whenever required by the Controller models or samples related to the invention have to be submitted by the applicants(s). Model should not exceed 12"in lengths. Sampling should be enclosed in convenient containers. In the case of biological material, if such material is not available to the public, a sample should be deposited to an authorized depository institution notified by the GOI in official Gazette.

Sequential steps involved in patenting:

1. Filling of provisional application
2. Filing of complete application. This should be within 12 months of above
3. Examination of patent.
4. Resubmission of revised application. This should be on complete rectification of all objections. Time for such resubmission to patent office is 12 months. If this is not done on time, the application may be deemed as abandoned.
5. Issue of notice to applicant. When resubmitted application is complete and taken care of all objections, it is deemed accepted and the dame is notified to the applicant.
6. Advertisement of the acceptance. All accepted applications are advertised in the official gazette of GOI (part III, section 2) for public awareness.
7. Public Inspection of Patent Document, Copies of floppies of patent specifications along with drawings are placed for free inspection by interested parties. There are 50 such inspection centers in India across all State and Union Territories and more than 10 centers in foreign countries.
8. Opposition to the grant of patent. Any opposition to the invention, specifications and claims should be filed within 4 months from the date of gazetting. Opposition should be filed in prescribed form with fee.
9. Notice of opposition to the applicant, is served by the Comptroller whenever any opposition is filed by other parties
10. Filing statement and evidence on invention. Both the applicant and the opposing parties are required to submit appropriate evidence to drive home each ones argument. Patent attorneys are involved during this and next stage.
11. Dispute resolution. May be done by the Controller with the help of Subject Matter Experts of through arguments by the concerned attorneys. The decision of the Controller can be challenged in respective High Court, if necessary.
12. Decision to grant or deny patent. Controller takes this and this is final when opposition is either absent or settled at Controller level.
13. Sealing of Patent: A process done by Comptroller on application by the applicant in prescribed Form with fees within 6 months from step (f) or within 2 months from the final proceedings under step (1).
14. Registration of Patent; again a process at Patent Comptroller level or application on prescribed Form the fee.
15. Filing of application outside India: This can be done after step (d). But could possibly be delayed or prevented in case of opposition. However, this has to be completed within 12 months from step (b).

Secrecy of the process: Normally application for patent and its specifications should be kept secret until the step 14(f) stated above. All correspondences on the subject until then are also to be treated as confidential. The staff concerned with development of invention, preparation of documents and administration at institute level and ICAR headquarter as well

as the experts engaged to examine the document are required to strictly comply with the total confidentiality of the subject.

Patent Offices

The patent system in India is administered by the patent office. The patent office has its Head Office at Calcutta and Branch Offices at Mumbai, Delhi and Chennai having territorial jurisdiction on zonal basis aa shown below;

Patent Office Branch, Todi Estates, 11th Floor, Lower Parel (West), Mumbay-400023	The States of Gujarat, Maharashtra, Madhya Pradesh, and Goa and Union Territories of Daman and Diu and Dadra and Nagar haveli Telegraphic address "PATOFFICE'
Patent Office Branch, Unit No.401 to405, 11th Floor, Municipal Market Building, Saraswati Marg, Karol Bagh, New Delhi-110005	The States of Harayana, Himachal Pradesh, Jammu and Kashmir, Punjab, Rajasthan, Uttar Pradesh and Delhi and the Union Territory of Chandigarh. Telegraphic address "PATENTS"
Patent Office Branch Rajaji Bhawan Besant Nagar Chennai-600090	The States of Andhra Pradesh, Karnataka, Kerala, Tamilnadu and Pondicheri and the Union Territories of Laccadive, Minicoy and Aminidivi Islands Telegraphic address "PATENTOFIC
Patent OFFICE (head office) 2nd M.S.O. Building 5th, 6th, 7th Floor, 234/4, Acharaya Jagadish Bose Road, Calcutta-700020	Rest of India Telegraphic address "PATENTS"

Geographical Indications

Article 22 of TRIPs under GATT agreement allows member nations to classify a good as originating from its territory, a region or a locality in that territory, where a given quality, reputation and other characteristic of the good is essentially attributable to its geographical origins. This article further permits the concerned nation to enact legal instrument to prevent:

1. Use of any means of designation or presentations of a good that indicates or suggests that the good in question originates in a geographical area other than the true place of origin in a manner, which misleads the public as to the geographical origin of the good.
2. Any use which constitutes an act of unfair competition within the meaning of Article 10 *bis* of the Paris Convention (1967)

Twist in TRIPs Tale

Contrary to the belief that patenting will promote innovation and help the developing nations, many third world economies are thinking in the opposite way. The recently set up independent Commission of Intellectual Property Right (CIPR)(with the support from the

UK secretary of state for international development) feels that IPR benefits only those who have knowledge and inventive power. In other words the Commission feels that the balance of cost and benefit is tilted in favour of the developed economies. India is a party to CIPR.

Intellectual Property Right Links

Organizations
World Intellectual Property Organization (WIPO): www.wipo.org/eng/main.htm

Worldwide patents offices
Indian Patent Facilitating Centre: http://patent.office.nic.in
Indian Patent Office: www.indianpatents.org.in
US Patent and Trademark Office: www.uspto.gov
US Copyright Office: www.loc.gov/copyright

Resources: US Intellectual Property Legal Tests
Patent Law: http://fedlaw.gsa/fedfra23.htm
Rules/Regulations: www.uspto.gov/web/offices/pac/dapp/oppd/appxr.htm

Patent search resources
Delphion Intellectual Property Network: http://www.delphion.com
International Business Machines (IBM): http://www.patents.ibm.com
Database Search Services: (Button on the Patent Office's home page): http://www.patent.gov.uk

International Patent Classification
Agriculture: http://www.mi.camcom.it/eng/patents/codes/agriculture.htm
Biotechnology Patent: http://www.nal.usda.gov/bic/Biotech Patents/
Patent Database on the Internet: http://www.sul.standford.edu/depts/swain/patent/patdbases.html
http://www.heckel.org/genera/gloss.htm

Articles and Documents Related with IPR
Value Added Patent and Scientific Information: http://www.derwent.com/news/articles/
Patent Basics: http://www.smartbiz.com/sbs/arts/piv1.htm
US Patent of Basmati Case:
http://www.american.edu/projects/mandala/TED/basmati.htm
Licensing Your Patent; An Overview:
http://www.mycounsel.com/content/intelprop/patent /license/
Like an Option of Course: http://www.cafezine.com/index article.asp? deptid=5&Id=232/

Introduction to Molecular Biology

K. Nirmal Babu, V.N. Jayakumar, Sinoj Jose, K. Praveen and D. Minoo

The term Molecular Biology was first used in 1945 by William Astbury who was referring to the study of chemical and physical structure of biological macromolecule. Molecular biology deals with molecular basis of biological and genetic specificity. It searches below the large-scale manifestations of classical biology for the underlying molecular plan. It is a 3-dimensional and structural programme that is concerned not only with fine structure but also with function and genesis of cells, organelles and macromolecules of living organisms. The three major components of molecular biology are biochemistry, genetics and cell biology. With remarkable speed many breakthroughs in genetics, molecular mechanisms of development and of cell function, structures of macromolecules, have become increasingly relevant to applied fields such as medicine and pharmaceuticals, agriculture and industrial biotechnology, leading to advances in gene therapy, drug design, genetically improved crops, synthesis of important chemicals etc. Understanding the processes of life at molecular level is perhaps the greatest achievements of twentieth century science.

The major concepts on which the edifice of modern biology stands are:

1. Living organisms are genetic systems that reproduce and occasionally mutate. They function on the basis of the genetic programmes carried on their genomes and inheritable errors (mutations) sometime occur.
2. Populations of organisms evolve though natural selection
3. All cellular organisms are made of cells that arise from preexisting cells
4. Organisms are made up of polymeric organic molecules like proteins and nucleic acids
5. Organisms use these chemical interactions to regulate themselves and maintain homeostasis
6. Organisms are mostly made of protein subunits which assemble themselves into larger functional structures
7. Membranes are very important parts of cells and create unequal distribution of ions and small molecules
8. Organisms capture energy, from light, store it in special phosphate compounds for their energy requiring processes
9. All organisms occupy some popular niche in an ecosystem, where the hallmark of their life is interaction with many other species

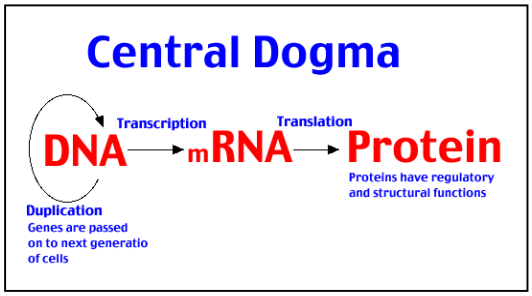
A living cell is a self contained, self-assembling, self-adjusting, self-perpetuating, constant temperature system of, molecules that extract free energy and raw materials from its environment. The cell uses this energy to maintain itself in a dynamic steady state far from equilibrium with its surroundings. The many chemical transformations within cells are organized into a network of reaction pathways promoted at each step by specific catalysts called enzymes, which the cell itself produces.

Self-replication through many generations is ensured by the self-repairing linear information coding system. Genetic information encoded as sequences or nucleotide subunits in DNA and RNA, specifies the sequence of amino acids in each distinct proteins, which ultimately determines the 3-diemnsional structures, and function of each protein.

There is a structural hierarchy in the molecular organization of cells. Cells contain organelles, such as nuclei, mitochondria, chloroplasts, which in turn contain supra molecular complexes such as membranes and ribosomes and these in turn contain cluster of macromolecules bound together by many relatively weak, non covalent forces. The monomeric components of cellular macromolecules appear to have been selected in early biological evolution. These subunit molecules are relatively few in number and evolution has combined small biomolecules to yield macromolecules of immense diversity. The first macromolecules may have been RNA-like molecules capable of catalyzing their own replication. Later on, in evolution, DNA took over the function of storing genetic information, proteins became the cellular catalysts, and RNA mediated between DNA and protein allowing the expression of genetic information.

DNA is a Requirement of Life

At the macromolecular level of life, a difference starts to emerge between living and nonliving material. DNA holds the secrets of life. Encoded in the DNA sequence is information for numerous proteins that have both structural and regulatory functions. Among the proteins found in cells are enzymes that catalyze biochemical reaction. Enzymes speed up the rate of a reaction that allows life to continue. A simple way to think of the flow of information from DNA to mRNA to protein is the molecular trinity that is central to all life and is known as the central dogma.



Energy - requirement of life

Everything in the universe has energy, which is defined as a capacity to do work. There are many different types of energy, and absolutely nothing can happen in the universe without a complete or partial transfer of energy. The ability to obtain and convert energy from its surroundings (plants and animal generally do it differently), and the use of energy to maintain itself, grow, and make more cells is metabolism

Living organism sense and respond to energy

Life forms sense and make changes in the environment. Organisms have receptors, which are molecules that detect specific stimuli. Cells adjust their metabolic activities in response to stimulus (or energy changes). Organisms respond to energy changes and their internal operating conditions remain within a narrow limit. This narrow limit of change, homeostasis, is the defining feature of life.

Biosignalling

The ability of cells to receive and act on signals from beyond the plasma membrane is fundamental to life. The receptors bind the signal molecule, amplify the signal, integrate it with input from pother receptors and transmit it in to the cell. Eukaryote cells have four signaling mechanisms viz., gated ion channels in the plasma membrane; the enzyme with a catalytic domain inside the cell and an extra cellular receptor for the signal; membrane proteins that stimulate target proteins by activating G proteins; nuclear proteins that bind steroids and act as transcription factors.

Progression through the cell cycle, is regulated by a family of protein kinases that act at specific period of the cycle, phosphorylating key proteins and modulating their activities. Oncogenes encode defective signaling proteins by continuing giving the signal for cell division they lead to tumour formation. They encode defective growth patterns receptors, G protein kinases or nuclear regulators of transcription. Apoptosis or programmed cell death involves some proteins that regulate the cell cycle and other unique to the cell death pathway. Triggered by external signal tumour necrosis factor or internal mechanisms, apoptosis result in activation of proteases that degrade cellular proteins.

The Structure of Plant Genomes

Packaging the genome into the nucleus

The genome size of corn (an average, but delightful plant) is about 10^9 base pairs. The DNA of the genome is associated with proteins, and this complex is referred to as "chromatin". The folding, coiling and looping of the 30nm chromatin fiber that further compacts the genome within the nucleus is mediated by specialized groups of proteins that play a variety of roles. Higher order chromatin structure is the least well-understood feature of the genome. It is generally agreed that the genome is organized into topologically independent super coiled loops. The anchor points, or "loop basements" that partition the genome into loops are often specialized regions of DNA/protein interactions. There appears to be two types of matrix attachment points; those that serve to package and organize the genome on a large scale (loop basement attachment sites - LBARs) and those that partition the genome into smaller, often gene-sized loops that appear to play a role in the regulation of the genome (matrix attachment regions - MARs). LBARs anchor large sections of the genome.

MARs anchor "functional" domains around genes

Matrix Attachment Regions (Scaffold Attachment Region or SAR are specialized sequences in the genome that are recognized by nuclear matrix proteins and associated matrix attachment factors. The MAR sequence is typically AT rich, several hundred bp long and contains topoisomerase II recognition sequences embedded within it. The latter is interesting as topo II is an enzyme that functions in relieving torsional stress in DNA by nicking it, then rejoining the strand after. Think about what happens when you twist up a rubber band - it gets all knotted up and *condensed* with respect to its original shape - you can return it to its original, unsupercoiled state by releasing it at its anchor (your fingers). Topo II presumably does this to super coiled loops in chromatin. MARs often flank genes, and appear play a role in "preparing" genes for transcription by creating a structure that is more accessible to the transcriptional machinery (transcriptional machinery - the suite of enzymes and other proteins required for transcription). MARs also appear to set genes apart from one another by creating independent topological domains. This means that although one section

of the genome might be somewhat de-condensed and ready for transcription, the adjacent section of the genome can remain condensed and quiescent. First order chromatin structures plays a direct role in gene regulation by creating surface features that are recognized by the transcriptional machinery, by bringing distant sequence elements.

The Nuclear Genome

Nuclear genome organization: Genome size

The size of the nuclear genome varies among organisms. The amount of DNA in a haploid cell is referred to as the C value. When the genome size of an organism is given, it is usually the haploid C value that is given. Plants have C values ranging from 107 to 1011 base pairs (bp) coding for 15 thousand to 60 thousand genes.

The genome size is roughly correlated with organism complexity. In other words,humans have larger genomes than most insects and insects have larger genomes than fungi. Although organism complexity roughly correlates with genome size, the correlation breaks down among chordates. For example, some amphibians have genomes almost 50 times larger than that of humans.

The smallest known plant genome belongs to *Arabidopsis thaliana* at $7x10^7$ (roughly the same size as yeast and the nematode *Caenorhabditis elegans*) and the one of the largest is a member of the lily family, *Fritillaria assyriaca*, with $1x10^{11}$. The lack of a direct relationship between genome size and organism complexity is called the C-value paradox. The genomes of *Arabidopsis* and *Fritillaria* code for about the same number of genes. Typically < 2% of the DNA in many genomes is transcribed and translated during normal cellular activities. Repetitive DNA contributes to the character and function of specialized structures in chromosomes and plays a role in genome organization. Non-coding, tandem repetitive DNA is referred to as satellite DNA. Dispersed repeat sequences make up a significant portion of the genome. These sequences differ from the tandem repeats of centromeres and telomeres in that copies are dispersed throughout the genome rather than lying adjacent to one another.

Not all repetitive DNAs are non-coding sequences. Large multigene families that are evolutionarily conserved are often clustered within the genome. Gene families consist of genes tandemly repeated numerous times. Even though they are arranged as tandem repeats, each gene is individually regulated. Smaller multigene family members share extended DNA sequence homology and code for functionally related proteins.

The organization and arrangement of single-copy genes is evolutionarily conserved among related plant species. Genome mapping projects have revealed that segments of chromosomes are conserved among species. Maize and sorghum, for example, contain many of the same genes and linkage groups residing at the same loci (physical locations). This co linearity of loci is called synteny.

Transposable elements

Transposable elements can make up a significant portion of the nuclear genome. They are sections of DNA that move, or transpose, from one site in the genome to another. These mobile DNA elements carry genetic information with them as they transpose, making them important features of genome organization. There are two basic types of transposable elements, Ac/Ds type elements and retrotransposons, which are almost certainly viral in origin.

Chromosomes

Each chromosome is a single molecule of DNA complexed with proteins. Chromosomes are found as homologous pairs in diploid cells, each thereby containing one allele of each gene pair in a diploid genome in either homozygous or heterozygous condition. These genes follow the principle of segregation and independent assortment.

Telomeres

The telomere is the structure that defines the end of a chromosome. This specialized chromosomal cap offsets the tendency for DNA to shorten with each round of replication. Telomeres contain repetitive sequence elements, but unlike those of centromeres, these sequences are highly conserved among eukaryotes in both sequence and arrangement.

The Chloroplast and Mitochondrial Genome

Plastids and mitochondria have their own genomes. The genome, the machinery utilized in its replication and regulation is very similar to prokaryotic systems. All plastids within a plant contain exactly the same genome. Variation exists only in the number of copies. Gene content and organization of genes within the genome is also highly conserved. The content of the mitochondrial genome is also conserved among plants, but the physical arrangement of the DNA is highly variable. The mitochondrial genome also contains smaller linear or circular DNA molecules known as Mitochondtrial plasmids and their function is mostly unknown.

The plastid genomes of evolutionarily distant plants are highly conserved. Thus the plant mitochondria provide an interesting perspective on the evolution of eukaryotic genomes.

DNA

Between 1928 and 1953, researchers (Frederick Griffith 1928, Oswald T Avery, Colin M. MacLeod and Maclyn McCarty 1944, Alfred Hershey and Martha Chase1953) confirmed DNA as the genetic material. In 1956, A. Gierer and G. Schramm suggested that the RNA carried all of the genetic information necessary for the synthesis of new viruses. Although the genetic information in cells is carried by DNA, viruses have genomes of double stranded or single stranded DNA or RNA, while viroids consist of small circular molecules of RNA.

DNA is a double helix consisting of antiparellel strands in which the nucleotide units are linked by 5' to 3' phosphodiester bonds (Watson and Crick, 1953). It replicates semi conservatively in the presence of a series of enzymes called polymerases, helicases, topoisomerases, gyrase and ligase. The DNA synthesis is semi discontinuous and primed by RNA. In which the leading strand of DNA growing 5' to 3' is extended continuously but the lagging strand that grows overall in the opposite 3' to 5' direction is made as short okazaki fragments. The concept of base pairing is central to all processes involving nucleic acids. Disruption of the base pair is a crucial aspect of the function of a double stranded molecule, while the ability to form base pairs is essential for the activity of single stranded nucleic acid. The nucleic acid sequences can be compared in terms of similarity and complementarity. The complementarity base pairing is used to transcribe RNA representing one strand of DNA. The stretch of DNA may code for protein. The genetic code describes the relationship between the sequence of DNA and the sequence of protein. The genetic code is the collection of base sequences that correspond to each amino acid and to stop signals for translation. The genetic code is a triplet, non overlapping and highly redundant. The reading frame that consists exclusively of triplets that represent amino acids is called Open Reading Frame or ORF. The ORF when translated into proteins has a reading frame that starts with initiation codon (AUG) and a termination codon. Mutations change the base sequence of

DNA and the spontaneous mutations are generally concentrated in specific sites called 'hot spots'.

The first systematic attempt to associate genes with enzymes showed that each metabolic step is catalyzed by a single enzyme whose production is the responsibility of a single gene. A mutation in the gene alters the activity of the protein for which it is responsible. This led to one gene one-enzyme hypothesis, which is more precisely expressed at present as one gene one polypeptide chain.

Genes and genomes can be mapped by the use of overlapping fragments which can be extended into a sequence. Restriction sites can be used as genetic markers. The existence of polymorphism allows linkage maps to be constricted using restriction fragments.

All types of eukaryotic genomes contain interrupted genes (exon and intron). Their proportion lower in yeast. A few genes are un interrupted in eukaryotes. The position of introns are conserved when the organization of the homologues is compared between the species. The conservation of exons can be used between the species. The intron sequences vary while the exon sequences remain well related. This can be used to isolate related genes in different species.

Most of the genes belong to the families defined by the position of relative exons of individual members. Families evolved by duplication of genes followed by divergence between the copies.

Transcription generates a single stranded RNA identical in sequence to one of the duplex DNA. Several different types of RNA (mRNA, tRNA, rRNA) are generated by transcription. Transcription converts the nucleotide sequence of RNA into amino acids sequence comprising of proteins. The ribosomes are the sites of protein synthesis. Transcription is regulated by the interaction between trans- acting factors and cis-acting sites. A trans- acting factor is the product of the regulatory gene and is usually protein. A cis-acting site is a sequence that functions by being recognized *in situ.* Bacterial genes coding for proteins whose functions are related may be organized in a cluster that is transcribed into a polycistronic m RNA from a single promoter. The control of this promoter regulates the expression of the entire pathway. This unit of regulation containing structural genes and cis-acting elements is called an operon. In phage Lambda, the genes are organized into groups and whose expression is controlled by individual genes.

Molecular biology is one of the fastest growing sciences and its implications and applications are far reaching. With increasing knowledge of molecular biology coupled with recombination technology and genetic engineering has helped in understanding the basic architecture of life at molecular level. This information lead to a better understanding of genomes of many organisms like *Arabidopsis*, yeast, rice,*C.elegans* and humans. This information has lead to newer approaches like genomics and proteomics, isolation of large number of genes better understanding of the gene function and gene therapy.

References

1. Benjamin Lewin (2000). *Genes VII*, University Oxford Press, U.K
2. David Freifelder (1998). *Molecular Biology*, Malhotra Publishing House, New Delhi.
3. P.K. Gupta (1999). *Cell and Molecular Biology*, Rastogi Publications, Merrut.

Internet Basics, Search Engines, Search Strategies

Santhosh J. Eapen

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.

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 stand out 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.

Internet Technologies

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.

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.

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.

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.

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.

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 Internet allows 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.

Search Strategies

If you are looking for a broad, common topic use a search engine that tends to give you fewer results of higher quality usually, 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.

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 them and understand the method and its results you use it.

Timeliness: Many sites on the web are out-of-date 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 then you should be careful.

References

http://www.med.nyu.edu/rcr/btr/index.html
Cynthia Gibas and Per Jambeck 2001. *Developing Bioinformatics Computer Skills*. O'Reilly, USA.

Genomics and its Applications

V.A. Parthasarathy

Introduction

Genomics is a rapidly emerging area of research, which came into existence with the initiation of various sequencing projects. This subject now promises to become a dominant theme of intellectual activity. Apart from revolutionizing the biological research, genomics provide enormous job opportunities, as the mining of sequence data will require enormous efforts and manpower. Simultaneous advances in computational methods have provided an organized and rigorous framework for biology. As it will be understood from the following, one need not be a part of mega sequencing project to benefit from genomics. Even small laboratories can take advantage of the results from genomics by combining the techniques of biology and computer.

Genomics: Biological background

The fundamental building blocks of life are proteins. Enzymes, which are the molecular machines responsible for virtually all of the chemical transformations that cells are capable of, are proteins. In addition, much of the structure of a cell is made up of proteins. That part of the structure which is *not* made up of proteins is produced by enzymes which *are* proteins. A human contains on the order of 100,000 different proteins. It is the properties of and the interactions between these 100,000 proteins that make us what we are.

Proteins are variable length linear, mixed polymers of 20 different amino acids. These topologically linear polymers fold upon themselves to generate a shape characteristic of each different protein, and this shape along with the different chemical properties of the 20 amino acids determine the function of the protein. Thus, in theory, knowing the sequence of a protein (the order with which the amino acids occurred) one could infer its function.

The order of amino acid in a protein is determined by the genetic information stored in DNA. DNA is a linear polymer of 4 nucleotides - deoxyAdenosine monophosphate (abbreviated A), deoxyThymidine monophosphate (abbreviated T), deoxyGuanosine monophosphate (abbreviated G) and deoxyCytidine monophosphate (abbreviated C).

A property of both DNA is that the linear polymers can pair one with another, such pairing being sequence specific. In such double polymers (referred to as a "double helix" due to the shape they assume) G pairs with C and A pairs with T . One strand DNA can serve as a template for the construction of a complementary strand, and this complementary strand can be used to recreate the original strand. This is the basis of DNA replication and thus all of genetics. Information on DNA is translated to proteins through RNA.

Four different nucleotides taken three at a time can result in 64 different possible triplet codes; more than enough to encode 20 amino acids. The way that these 64 codes are mapped onto 20 amino acids is first, that one amino acid may be encoded by 1 to 6 different triplet codes, and second, that 3 of the 64 codes, called stop codons, specify "end of peptide sequence".

The fact that DNA nucleotides need to be read three at a time to specify a protein sequence implies that a DNA sequence has three different reading frames determined by whether you start at nucleotide one, two, or three. (Nucleotide four will be in the same frame as nucleotide one and so on. Both strands of DNA can be copied into RNA (for translation into protein). Thus, a DNA sequence with its (inferred) complementary strand can specify six different reading frames.

It is possible to chemically determine the sequence of amino acids in a protein and of nucleotides in RNA or DNA. However, it is vastly easier at present to determine the sequence of DNA than that of RNA or protein. Since the sequence of a protein can be determined from the DNA sequence which encodes it, most protein sequences are in fact inferred from DNA sequences.

What is genomics?

In simpler terms, genomics refer to the DNA sequence analysis to identify the regions coding for the proteins. Whoever, the billions of DNA sequence do not tell us what all the genes do, how cells work, how cells form organisms, what goes wrong in disease, how we age or how to develop a drug. The purpose of genomics is to understand biology, not simply to identify the component parts but also understand how the components work together to comprise functioning cell and organism. In other terms, deciphering the sequence of bases of DNA would unlock the whole blueprint of the development of an organism.

Genomics, earlier restricted to sequencing of individual genes or short regions of the genome, received a shot in the arm with the launch of Human Genome project (HGP) in 1988, which envisage the sequencing of entire 3.2 billion bases of human genome. HGP also led to the development of additional new methodologies required for sequencing whole genome. During the different phases of HGP, the entire genomes of many prokaryotic and eukaryotic organisms have been sequenced. Bacteria *E. coli*, yeast

Saccharomyces cerevisiae, nematode *Caenorhabditis elegans*, fruitfly *Drosophila melanogaster* are some of the organisms whose genome was completely sequenced before the first draft of human genome sequence was published in early 2000. HGP aroused worldwide interest among plant biologists to undertake similar ventures in plants. As a result, two genome projects viz. AGI (*Arabidopsis* genome initiative) and RGP (Rice genome project) were initiated to sequence *Arabidopsis* and Rice genome respectively. The complete sequencing of these two genomes is already published.

Recent trends in Genomics

Once whole genome information is available for an organism, the challenge turns from identifying the parts to understanding their function. This is referred to as **functional genomics**. Functional genomics involves assignment of function of the identified genes, organisation and control of genetic pathways that come together to make up the physiology of an organism. Various techniques have been developed to understand the function of various genes, **DNA arrays or gene chips** being the most important, that allows complex mixtures of RNA and DNA to be interrogated in a parallel and quantitative fashion. Recently, **Proteomics**, the large scale analysis of proteins, is widely used to study genes and genomes. While the proteomics covers much of the functional analysis of gene products or

functional genomics including large scale identification or localization studies of proteins and interaction studies, **Structural genomics** focuses on large scale study of protein interaction.

Applications of genomics

Genomics has wide implications in all fields of biology. The study of various genomes holds the key to understanding the origin and evolution of plants and animals. Sequencing of prokaryotic genomes will aid vaccine design and allow exploration of new microbial energy sources. Genomics will impact molecular medicine the greatest, leading to improved diagnosis of disease and drug designing. In longer term, isolation of 'healthy' genes for use in 'gene therapy' will be much easier task.

Genomics will also be of vital importance for the agriculture. Only a few plant species stand between prosperity, on the one hand and hunger and starvation on the other. If we understand the genomes of crops like rice, wheat, maize, beans and potato. We can ensure a better future with the capability of more perfect and precise genetic manipulation for yield and survival under adverse conditions.

Genomics and Bioinformatics

Various genome projects are generating gigantic data bases containing sequence of billions of bases, details of when and in which tissues of the body various genes are turned on, the shapes of the proteins the genes encode, how the proteins interact with one another and the role those interactions play in disease. Unraveling the importance of such vast data is more necessary and the progress and ultimately profit relies on the ingenuity of the end user. Data capture; data warehousing and data mining have become major issues for biotechnologist and biological scientist. Advancement in information technology, particularly Internet, is being used to gather and access ever-increasing information in biology and biotechnology. Bioinformatics uses advances in the area of computer sciences, information science, computer and information technology, and communication technology to solve complex problems in life sciences.

Data generated from genome project are deposited in **Genebanks,** a term used for data banks containing DNA sequences. Now data banks for protein sequence and structures also available. One of the basic operations in bioinformatics involves searching for similarity, or homologies, between a newly sequenced piece of DNA and previously sequenced DNA segment from various organisms. Finding near match allows to predict the type of protein the new sequence encodes. Although it is not possible to *completely* predict the function or shape (structure) of a protein from its sequence *de novo*, *some* useful inferences about structure and function can be drawn, especially by comparing the sequence of a protein of unknown structure and function to sequences of proteins with known structure and function. Second, if the goal of structure/function prediction is to be reached in the future, it will be because of partial analyses done in the present. Third, by comparing the sequence of equivalent proteins from different species of animals (such equivalent proteins are called "homologues"), one can draw inferences about the evolution of these species from their common ancestors.

A popular set of software programs for comparing DNA sequence is **BLAST** (Basic Local Alignment Search Tool). BLAST is part of a suit of DNA and protein sequence search tools accessible in various customize versions from various data base providers or through NCBI (National Centre for Biotechnology Information). NCBI also offers **Entrez,** a so called

meta-search tool that covers most of NCBI's databases, including those housing three dimensional protein structures, the complete genomes of organisms such as yeast, and reference to scientific journals that back up the database entries.

Conclusion

The association of computational and molecular technological developments is opening up new and unprecedented opportunities for genetic research. The combined use of sequence information, computational tools, integrated knowledge databases and the traditional approaches of biology increase the hopes of understanding the function and regulation of all genes and proteins; and deciphering the underlying working of the cell.

Proteomics and its Applications

R.V. Omkumar

With complete genome sequences now available for several prokaryotic and eukaryotic organisms, biological researchers are faced with the unprecedented scientific challenges of assigning molecular and cellular functions to thousands of newly predicted gene products and explaining how these products cooperate in complex physiological processes. To address this problem, the field of proteomics has emerged with the goals of developing and applying methods for the global analysis of protein expression and function. It is anticipated that the creation of effective methodologies for the rapid and parallel analysis of proteins will accelerate the functionalization' of these biomolecules and, by doing so, discover, among other things, new biomarkers and therapeutic targets for the diagnosis and treatment of diseases of humans and generally increase our mechanistic understanding of biological processes. Nonetheless, for proteomic researchers to achieve these lofty goals, they must first confront and eventually overcome several technical challenges that currently limit efforts to characterize systematically proteins from highly complex samples.

Indeed, although it might be conceptually attractive to view proteomics as simply the 'protein equivalent' of genomics, such a view falls far short of depicting the myriad of methodological problems that are uniquely associated with the scientific investigation of proteins.For example, in contrast to RNA or DNA, proteins cannot be amplified by methods analogous to PCR, that is, the amount of protein present in a given sample is the amount of protein that must be analyzed. Thus, proteomic researchers are confronted with the difficult task of detecting, identifying and characterizing numerous low abundance proteins at their natural cellular concentrations, even when the levels of these biomolecules are 6-8 orders of magnitude lower than those of high-abundance proteins present in the same sample (as is often the case). Similarly, unlike RNA or DNA, proteins do not inherently possess well defined high-affinity and/or high selectivity binding partners. Thus, whereas the field of gene microarrays has been able to capitalize on the special interactions that oligonucleotides share with their antisense partners, protein microarray researchers must devise a specific capture reagent for each protein of interest, a process that promises to be both expensive and laborious, but nonetheless of crucial importance. Finally, proteinsexhibit a range of biochemical properties that far exceeds the relatively homogenous behavior of oligonucleotides and is crucially dependent on the precise 3D structure of the folded polypeptides. The diversity of traits exhibited by proteins, ranging from extreme pi values to membrane association to post-translational modification, means that methods to handle and process these molecules are rarely as generally applicable as we would like.

Considering all of the challenges that face the field of proteomics, it might be worth asking - isn't genomics good enough? Clearly, however, the answer is becoming an increasingly emphatic 'no'. Accumulating examples in the literature of the limited correlation that exists between steady-state rnRNA and protein abundance, as well as the changes in mRNA and protein abundance induced by perturbations, serve as a constant reminder of the complex manner in which cells regulate protein expression in a post-transcriptional manner. Additionally, proteins are modified by an ever-increasing number of post-translational events, many of which are dynamic in nature, highlighting the need to not only detect, but

also characterize proteins from native proteomes. Finally, it is important to recognize that the physiological and pathological events that form the basis for health and disease are, at their core, protein-driven processes. To understand these events in molecular detail, the proteome, in all of its spatially and temporally regulated forms, must be analyzed.

Mass spectrometry instrumentation for proteomics

Mass spectrometry has become a primary tool for proteomics because of its capabilities for rapid and sensitive protein identification and quantitation. It is now possible to identify thousands of proteins from microgram sample quantities in a single day and to quantify relative protein abundances. However, the need for increased capabilities for proteome measurements is immense and is now driving both new strategies and instrument advances. These developments include those based on integration with multidimensional liquid separations and high accuracy mass measurements and promise more than order of magnitude improvements in sensitivity, dynamic range and throughput for proteomic analyses in the near future.

Multidimensional peptide separation

Multidimensional peptide separation will play an increasingly important role in the drive to identify and quantitate the proteome. By increasing the peak and load capacity, multidimensional approaches increase the number and dynamic range of peptides that can be analyzed in a complex biological organism. Separation methods using different physical properties peptides have been combined with varying degrees of success. The ultimate goal is a rapid separation strategy that can be coupled with analytical methods, such as mass spectrometry, to provide comprehensive monitoring of the changing concentration, interactions, and structures of proteins in the proteome.

Protein chips

A series of exciting reports over the past two years has established the usefulness of protein chips and made important advances in preparing protein arrays. However, several technical challenges must still be addressed to make these tools available to the wider community of researchers. Here, we discusses these challenges and survey recent opportunities for creating quantitative assays, preparing and immobilizing large numbers of proteins, using detection methods to analyze the results of chip-based experiments, and using informatics tools to interpret these results.

Drug discovery

Proteomics is playing an important role in the pharmaceutical industry. The new methods for the global analysis of gene expression, which can analyse large numbers of parameters simultaneously, are ideal for accurately establishing drug action, as well as identifying possible toxic side effects at the molecular level. New techniques offer opportunities for screening gene expression at the mRNA level. However, as discussed above, mRNA analysis alone fails to provide a comprehensive view of gene expression. Moreover, many drugs act at the level of the cellular protein, and it is this level that is analysed with proteomics, making these data more relevant for drug action and possible toxicity.Consequently, proteomics is becoming an important tool in the areas of drug development and toxicology.

Essentially, the proteomes of cells exposed to the compound under investigation and non-treated cells are compared. Qualitative and quantitative differences in protein synthesis are scored and significant differences in the proteomes recorded. If required, the proteins of

interest may be further processed for identification. Alternatively, the pattern of protein changes induced by a novel drug can be compared against a database of protein changes induced by compounds of known activities. Drugs with similar modes of action are expected to affect the cellular proteome in similar ways. Thus, it may not always be necessary to know the identities of the proteins affected to establish drug action or toxicity. These experimental assays can be carried out by exposing laboratory-grown cell lines to the drugs under investigation or by exposing suitable laboratory animals. The use of laboratory animals has some advantages over *in vitro* grown cells since some drugs may be metabolised *in vivo* and the response of the metabolites should be included in the assessment of drug action and toxicity.

The future of proteomics

The methods comprising proteomics are being applied throughout the biosciences and the data derived from these ongoing and future studies will influence our thinking in biology for years to come. However, there remain a number of technical challenges to be overcome before proteomics can compete with current genomic technologies in the areas of sensitivity and sample throughput. The currently used methods are labour intensive and, therefore, not ideal for large-scale studies involving the analysis of large sample numbers. The successful application of these protocols requires a significant input of infrastructure and staff. In addition, 2DE is not optimal for analysing specific types of proteins. Although the 'tricks of the trade' may go some way to overcoming some of these limitations, it is unlikely that 2DE can be used on its own to fully achieve the complete definition of a cell's proteome. New methods are on the horizon to aid in the analysis of low abundance and difficult proteins based on mass spectrometry techniques. Work is in progress on the production of protein arrays, that are analogous to the nucleic acid micro-arrays which have played such a big role in the exploitation of genome sequence data. Once available, protein arrays will provide a means of rapidly determining the complete proteome of a cell for multiple conditions, and are certain to be widely used in the pharmaceutical industry. Only time will tell if the economics of the technology will allow them to be used in the academic environment. One thing is certain: proteomics is here to stay. Data derived from the diverse technologies making up proteomics will complement existing studies on the genome to provide a full picture of cellular metabolism. This is just the start for a new and expanding golden age for the proteins.

References

1. Banks R E, Dunn M J, Hochstrasser D F, Sanchez J-C, Blackstock W, Pappin D J and Selby P J (2000) Proteomics: new perspectives, new biomedical opportunities. *Lancet*, **356,** 1749 – 1756.
2. Blackstock W P and Weir M P (1999) Proteomics: quantitative and physical mapping of cellular proteins. *Trends in Biotechnology*,**17,** 121 – 127.
3. Cash P (2000) Proteomics in medical microbiology. *Electrophoresis*, **21,** 1187 – 1201. Pennington S R and Dunn M J (Editors) (2001) *Proteomics: From protein sequence to function*. Bios Scientific Publishers Ltd:Oxford, UK.
4. Wilkins M R, Williams K L, Appel R D and Hochstrasser D F (Editors) (1997) Proteome Research: New Frontiers inFunctional Genomics. Springer: Heidelberg, Germany
5. Phil Cash (2002) Proteomics: The protein revolution *Biologist* 49 (2), 58-62
6. A trends guide to proteomics . Trends in Biotechnology (Suppl.) 20(12). 51pp.

Human Genome Project

K.P. Aravindan

The Human Genome Project (HGP) is one of the great feats of exploration in history, an inward voyage of discovery rather than an outward exploration of the planet or the cosmos. An international research effort to sequence and map all of the genes - together known as the genome - of members of our species *Homo sapiens*. The HGP has successfully completed in April 2003, the 50th anniversary of the Watson-Crick discovery of DNA. The hereditary material of all multi-cellular organisms is the famous double helix of deoxyribonucleic acid (DNA), which contains all of our genes. DNA, in turn, is made up of four chemical bases, pairs of which form the "rungs" of the twisted, ladder-shaped DNA molecules. All genes are made up of stretches of these four bases, arranged in different ways and in different lengths.

HGP researchers are deciphering the human genome in three major ways: determining the order, or "sequence," of all the bases in our genome's DNA; making maps that show the locations of genes for major sections of all our chromosomes; and producing what are called linkage maps, complex versions of the type originated in early *Drosophila* research, through which inherited traits (such as those for genetic disease) can be tracked over generations.

Entities involved in accomplishing the HGP have included the National Institutes of Health (NIH), which created the National Human Genome Research Institute (NHGRI) in 1989; the United States Department of Energy (DOE); numerous universities and other research facilities throughout the United States; and international partners in the United Kingdom, France, Germany, Japan and China.

Human Genome Project Goals

The Human Genome Project (HGP) refers to the international 13-year effort formally begun in October 1990 to discover all the estimated 30,000 to 35,000 human genes including mapping, sequencing, and identifying genes; storing and analyzing data; and addressing the ethical, legal, and social issues (ELSI) that may arise from availability of personal genetic information. The ultimate goal of the HGP was to obtain the DNA sequence of the 3 billion DNA subunits present in human DNA and make them accessible for further biological study. As part of the HGP, parallel studies have been carried out on selected model organisms such as the bacterium *E. coli* and the mouse to help develop the technology and interpret human gene function.

Ethical, Legal, and Social Issues

- **Fairness in the use of genetic information** by insurers, employers, courts, schools, adoption agencies, and the military, among others.
- **Privacy and confidentiality** of genetic information
- **Psychological impact and stigmatization** due to an individual's genetic differences.
- **Reproductive issues** including adequate informed consent for complex and potentially controversial procedures, use of genetic information in reproductive decision making, and reproductive rights.

- **Clinical issues** including the education of doctors and other health service providers, patients, and the general public in genetic capabilities, scientific limitations, and social risks; and implementation of standards and quality-control measures in testing procedures.
- **Uncertainties** associated with gene tests for susceptibilities and complex conditions (e.g., heart disease) linked to multiple genes and gene-environment interactions.
- **Conceptual and philosophical implications** regarding human responsibility, free will vs genetic determinism, and concepts of health and disease.
- **Health and environmental issues** concerning genetically modified foods (GM) and microbes.
- **Commercialization of products** including property rights (patents, copyrights, and trade secrets) and accessibility of data and materials.

Area	Goal	Achieved	Date Achieved
Genetic Map	2- to 5-cM resolution map (600 – 1,500 markers)	1-cM resolution map (3,000 markers)	September 1994
Physical Map	30,000 STSs	52,000 STSs	October 1998
DNA Sequence	95% of gene-containing part of human sequence finished to 99.99% accuracy	99% of gene-containing part of human sequence finished to 99.99% accuracy	April 2003
Capacity and Cost of Finished Sequence	Sequence 500 Mb/year at < $0.25 per finished base	Sequence >1,400 Mb/year at <$0.09 per finished base	November 2002
Human Sequence Variation	100,000 mapped human SNPs	3.7 million mapped human SNPs	February 2003
Gene Identification	Full-length human cDNAs	15,000 full-length human cDNAs	March 2003
Model Organisms	Complete genome sequences of E. coli, S. cerevisiae, C. elegans, D. melanogaster	Finished genome sequences of E. coli, S. cerevisiae, C. elegans, D. melanogaster, plus whole-genome drafts of several others, including C. briggsae, D. pseudoobscura, mouse and rat	April 2003
Functional Analysis	Develop genomic-scale technologies	High-throughput oligonucleotide synthesis	1994
		DNA microarrays	1996
		Eukaryotic, whole-genome knockouts (yeast)	1999
		Scale-up of two-hybrid system for protein-protein interaction	2002

Potential Benefits of Human Genome Project Research

Rapid progress in genome science and a glimpse into its potential applications have spurred observers to predict that biology will be the foremost science of the 21st century. Technology and resources generated by the Human Genome Project and other genomics research are already having a major impact on research across the life sciences.

Some current and potential applications of genome research include

- Molecular medicine
- Risk assessment
- Bioarchaeology, anthropology, evolution, and human migration
- DNA forensics (identification)

Molecular Medicine

- *Improved diagnosis of disease*
- *Earlier detection of genetic predispositions to disease*
- *Rational drug design*
- *Gene therapy and control systems for drugs*
- *Pharmacogenomics "custom drugs"*

Technology and resources promoted by the Human Genome Project are starting to have profound impacts on biomedical research and promise to revolutionize the wider spectrum of biological research and clinical medicine. Increasingly detailed genome maps have aided researchers seeking genes associated with dozens of genetic conditions, including myotonic dystrophy, fragile X syndrome, neurofibromatosis types 1 and 2, inherited colon cancer, Alzheimer's disease, and familial breast cancer.

Risk Assessment

- *Assess health damage and risks caused by radiation exposure, including low-dose exposures*
- *Assess health damage and risks caused by exposure to mutagenic chemicals and cancer-causing toxins*
- *Reduce the likelihood of heritable mutations*

Understanding the human genome will have an enormous impact on the ability to assess risks posed to individuals by exposure to toxic agents. Scientists know that genetic differences make some people more susceptible and others more resistant to such agents. Far more work must be done to determine the genetic basis of such variability.

Bioarchaeology, Anthropology, Evolution, and Human Migration

- *Study evolution through germline mutations in lineages*
- *Study migration of different population groups based on female genetic inheritance*
- *Study mutations on the Y chromosome to trace lineage and migration of males*
- *Compare breakpoints in the evolution of mutations with ages of populations and historical events*

Understanding genomics will help us understand human evolution and the common biology we share with all of life. Comparative genomics between humans and other organisms such as mice already has led to similar genes associated with diseases and traits. Further

comparative studies will help determine the yet-unknown function of thousands of other genes.

Comparing the DNA sequences of entire genomes of differerent microbes will provide new insights about relationships among the three kingdoms of life: archaebacteria, eukaryotes, and prokaryotes.

DNA Forensics (Identification)

- *Identify potential suspects whose DNA may match evidence left at crime scenes*
- *Exonerate persons wrongly accused of crimes*
- *Identify crime and catastrophe victims*
- *Establish paternity and other family relationships*
- *Identify endangered and protected species as an aid to wildlife officials (could be used for prosecuting poachers)*
- *Detect bacteria and other organisms that may pollute air, water, soil, and food*
- *Match organ donors with recipients in transplant programs*
- *Determine pedigree for seed or livestock breeds*
- *Authenticate consumables such as caviar and wine*

Any type of organism can be identified by examination of DNA sequences unique to that species. Identifying individuals is less precise at this time, although when DNA sequencing technologies progress further, direct characterization of very large DNA segments, and possibly even whole genomes, will become feasible and practical and will allow precise individual identification.

To identify individuals, forensic scientists scan about 10 DNA regions that vary from person to person and use the data to create a DNA profile of that individual (sometimes called a DNA fingerprint). There is an extremely small chance that another person has the same DNA profile for a particular set of regions.

References

1. www.ornl.gov
2. www.nature.com/
3. www.sciencemag.org
4. http://www.hgmp.mrc.ac.uk/
5. http://genome.ucsc.edu/
6. www.er.doe.gov/
7. http://www.genome.gov/Education/

Polymerase Chain Reaction (PCR) Based Techniques

Johnson K. George

The PCR (Polymerase Chain Reaction) is an *in vitro* method for the enzymatic synthesis of specific DNA sequences, using two oligonucleotide primers that hybridize to opposite strands and flank the region of interest in the target DNA. A repetitive series of cycles involving template denaturation , primer annealing, and the extension of the annealed primers by DNA polymerase results in the exponential accumulation of a specific fragment whose termini are defined by the 5' ends of the primers. Because the primer extension products synthesized in one cycle can serve as a template in the next, the number of target DNA copies approximately doubles at every cycle. Thus, 20 cycles of PCR yields about a million-fold amplification. This method, which was invented by *Kary Mullis* was originally applied by a group in the Human Genetics Department at Cetus to the amplification of human β-globin DNA and to the prenatal diagnosis of sickle-cell anemia Initially, the PCR used the Klenow fragment of *E. coli* DNA polymerase I to extend the annealed primers. This enzyme was inactivated by the high temperature required to separate the two DNA strands at the outset of each PCR cycle. The introduction of the thermostable DNA polymerase (*Taq* polymerase) isolated from *Thermus aquaticus* transformed the PCR into a simple and robust reaction which could now be automated by a thermal cycling device. The use of the *Taq* polymerase not only simplified the PCR procedure but significantly increased the specifity and the overall yield of the reaction. The higher temperature optimum for the *Taq* polymerase allowed the use of higher temperatures for primer annealing and extension, thereby increasing the overall stringency of the reaction and minimizing the extension of primers that were mismatched with the template. The effect of varying the reaction parameters (e.g., enzyme, primer and Mg concentration as well as the temperature cycling protocol) is discussed below. Although, for any given pair of oligonucleotide primers, an optimal set of conditions can be established, there is no single set of conditions that will be optimal for all possible reactions.

One problem with detecting specific DNA sequences, especially those of unique genes, is that a relatively large amounts of pure DNA are needed. In addition, blotting techniques are laborious and time consuming to carry out. To circumvent these problems it is possible to enzymatically amplify a specific region of DNA using polymerase chain reaction (PCR). The ability to detect minute amounts of specific DNA sequences has resulted in the broad application of PCR for diagnosis, forensics, molecular epidemiology, etc. In addition, PCR has revolutionized many of the approaches used in molecular biology, such as gene cloning and sequencing.

Although the PCR is considered primarily a method for producing copies of a specific sequence, it is also a very powerful and precise way of altering a particular template sequence. Since the oligonucleotide primers become physically incorporated into the amplified product and mismatches between the 5' end of the primer and *initial* template are tolerated, it is possible to introduce new sequence information adjacent to the target sequence via the primers. Thus, for cloning a given sequence, one is no longer constrained by the naturally occurring restriction sites and one may add any restriction enzyme recognition sequence to

the 5' ends of the prime creating a new restriction site in the double-stranded amplification product. Furthermore, specific nucleotide substitutions, insertions, and deletions can also be introduced into the amplified product with the appropriate primers.

PCR mechanism

PCR amplifies a specific segment of DNA that lies between two known primer sequences. DNA strands are separated by heating and then annealed with a pair of **primers,** which are complementary to the opposing strands. DNA polymerase recognizes this small region of duplex DNA as a substrate and in the presence of nucleotides will synthesize the complementary strands of both template strands. If this procedure is repeated, the newly synthesized fragments also serve as templates for subsequent rounds resulting a geometric amplification. The end product of PCR is a dsDNA molecule that is defined by the 5'-ends of the primers. In Other words, the length of the DNA molecule is determined by the distance between the primers.

Typically, PCR results in a million-fold amplification of the target DNA. Therefore, sequences that only represent a small proportion of the total DNA can be detect after PCR and generally the DNA does not have to be highly purified. Furthermore, the procedure is relatively easy to carry out and does not require expensive equipment or reagents.

Practical aspects

The first step of PCR is to combine the template DNA, primers, dNTPs, Mg and Taq polymerase in a single tube. Special thin-walled tubes of uniform thickness are used to ensure rapid and equal temperature changes throughout the reaction volume. Typical concentrations are approximately 0.1 pM of the target DNA, 2 nM of primers with a Tm > 55°, 20-200 •M dNTPs (lower concentration lessen mispriming), and 2 units of enzyme. The optimal Mg2+ concentration depends upon the total dNTP concentration which includes: free dNTPs, primers, and template DNA. Since the template DNA is the most variable, it is generally necessary to titrate the optimal Mg concentration for different sources of templates. In addition, the EDTA (a typical component of buffers used for nucleic acid isolations) will affect the free Mg concentration.

The thermocycler is programmed for the desired times and temperatures of denaturation, annealing and polymerization. The optimal conditions are determined empirically. Most thermocyclers can also be programmed to vary these parameters in different cycles. For example, the denaturation step in the first cycle is sometimes carried out for 5 minutes to ensure a more complete melting of the target DNA. Likewise, after the last polymerization cycle it is possible to lower the temperature to 4o (optimal storage conditions).

PCR Applications: some of the applications of PCR are:

- Diagnosis
- Taxonomy
- Forensics
- Molecular Epidemiology
- Gene Expression
- Sequencing
- Gene Cloning

- Probe Generation
- Site-DirectedMutagenes

PCR Parameters

The oligonucleotide primers are the most critical element in terms of successful PCR (Box). Computer programs are used to examine DNA sequences for potential primers. One consideration is distance between the primers. Smaller DNA fragments are amplified more efficiently than longer DNA fragments and it is often particularly difficult to amplify fragments larger than one kb. The primers should also be unique in that they should not hybridize to other sites in the target DNA. Generally primers of 20 nucleotides or greater will provide a satisfactory level of uniqueness. The primers should not exhibit complementarity to each other or internal complementarity within a primer. Both of these will phenomenon will prevent the primer from annealing to the template DNA. A balance nucleotide composition also tends to improve the function of a primer. Designing a primer so that the last one or two bases on the 3'-end are either a G or a C will result in a stronger hybridization and ensure recognition by the polymerase.Purifying the primers by HPLC will improve their fidelity, but is not absolutely necessary.

PCR Procedure

1. Mix DNA, primers, dNTPs and Taq polymerase.
2. Set the thermocycler for desired times.

Primer Design

. template length
. uniqueness (18-28 bases)-for specific sequence amplification
. no primer dimers
. no internal complementarity
. 50% GC ratio
. HPLC purification

Amplified DNA is analyzed by gel electrophoresis. In most applications the desired result is a single band, or amplicon, detectable with ethidium bromide. Primers are usually designed so that a sample can be tested for the presence or absence of the expected band. Size heterogeneity can be determined if exhibited by that particular locus. Southern blotting of the PCR products can also be carried out for either increased sensitivity or specificity.

Site Directed Mutagenesis

PCR can also be used to manipulate DNA. For example, site-directed mutagenesis can be carried out by designing primers with single nucleotide mismatches. The PCR fragments will then be cloned into a plasmid and then transfected into *E. coli*. The host will correct the mismatches resulting in some of the clones having the introduced nucleotide. Similarly, restriction sites are easily added to the PCR products for subsequent subcloning

RT-PCR

It is also possible to use PCR for the analysis of RNA and gene expression. This method, called RNA-PCR or RT-PCR, detects specific transcripts and exhibits a greater sensitivity than Northern blotting. The first step of RT-PCR is to make a DNA copy of the mRNA (i.e., cDNA). The copy is made using reverse transcriptase (RT), an enzyme of retroviruses which exhibits RNA-dependent DNA polymerase activity. There are three basic strategies for the synthesis of cDNA: specific priming, oligo-dT priming, or random priming. The cDNA is then subjected to PCR using a specific primer pair.

Modifications of PCR

Some modifications of the PCR methods, which are used in finger printing of genomes to analyze unknown sequences, have been described.

Inverse PCR

Inverse PCR (also known as chromosome crawling) is a method to amplify the sequences flanking a knownsequence. Genomic DNA is digested with a restriction enzyme which produces appropriately sized fragments (eg., 3- 5 kb) containing the target sequence. The fragments are circularized with DNA ligase and amplified with primers that are directed away from each other. This results in the amplification of the sequences that directly flank the known sequence. The boundary between the two flanking sequences is demarcated by the restriction site used to disgest the genomic DNA.

Anchor PCR

Known sequences, or 'anchors', can be also added to unknown sequences and used as the source of PCR primers.This anchor PCR is often used to amplify the unknown 5'-end of a cDNA (i.e., mRNA). (Sometimes called RACE for rapid amplification of cDNA ends.) In this method a cDNA copy of the mRNA is made using a primer based on known sequence. A homopolymer tail is added to the 3'-end of the cDNA with terminal deoxynucleotide transferase. A primer containing sequence complementary to the homopolymer tail is used in conjuction with a primer from the known sequence to amplify the target DNA.

RAPD (Random Amplified Polymorphic DNA)

This is the simplest of all PCR based markers. Random pieces of the genome of 300 to 2000 base pairs (bp) long are amplified using a single oligonucleotide primer of arbitrary sequence. The technique requires no prior knowledge of DNA sequence to generate genetic markers and is fairly simple. A discrete PCR product is produced when the primer binds to sites on opposite strands of the DNA that are within an amplifiable distance (less than 3000 bp usually). The presence or absence of this specific product is thought to represent mutations in the primer-binding sites of the DNA . RAPD reactions normally amplify three to ten different pieces of DNA .

Primers used are about 10 bp long and because of this short length, the annealing temperature must be low (35 - 40 C). Products are realised on agarose gel and stained with ethidium bromide (polyacrylamide gel and silverstaining can also be used). There are variations on this technique, using shorter primers (5 bp) or two arbitrary primers, as well as alternative methods of staining and fragment resolution.

In RAPD analysis the total DNA from an organism is mixed with 10 bp single stranded DNA (primer) together with the four different deoxynucleotides and a heat stable DNA polymerase enzyme. The reaction mixture is placed on a thermocycler (PCR-machine) which can change the temperature of the reaction rapidly according to a predefined PCR programme.

A typical PCR programme for RAPDs consists of::
1 min. 95°C (Denaturation) followed by 45 cycles
2 min 37°C (Reannealing)
2 min 72°C (Elongation)
1 min 95°C (Denaturation)

Reannealing at low tempeature makes the primer molecules attach to complementary site on the DNA. Elongation is the period when the polymerase enzyme elongates the attached primers from their 3' end. Denaturation separates the new formed strands.

After af few PCR cycles DNA pieces with well defined length similar to the one between the two original primer sites dominate the mixture. This is because they amplify exponentially. In most RAPD reactions the result will be amplification of 5 to 10 different DNA segments whose length depend on the primer recognition sites available in the genome.

The mixture after the PCR reaction is normally loaded on an agarose gel for electrophoresis to separate amplification products according to their size. Staining of the gel with ethidium bromide, and subsequent observation under ultraviolet light is the next step. Ethidium bromide will concentrate between the bases in the DNA and give rise to yellow flourescent bands under the UV-light.

Two modifications of RAPD markers have been described as DNA Amplification fingerprinting (DAF) and Arbitrarily Primed PCR (AP-PCR). DAF uses short random primers of 5-8 bp and visualises the relatively greater number of amplification products by polyacrylamide gel electrophoresis (PAGE) and silver staining. AP-PCR uses slightly longer primers and amplification products are radiolabelled and also resolved by PAGE.

Advantages

The broad use of RAPD markers in many studies is due to several factors: (1) no prior sequence knowledge is necessary; (2) small amounts of DNA can be used (25 nanograms per reaction); (3) the technique is technically simple; (4) detection is not radioactive and (5) a large number of potential markers can be generate using readily available markers. This, in addition to its swift pace of obtaining data, makes RAPD a popular technique, having been used in studies of interspecific hybridization, introgression, identification of clones, development of markers linked to sex determination and measurements of genetic diversity, as well as many other studies. The technique also reveals relatively high levels of polymorphism and has a low cost compared to other techniques, such as allozymes and RFLP.

Problems

The low annealing temperature can cause some unspecific binding, with primers binding to sequences that are not completely complementary.
Accurate scoring of RAPD markers depends on reliable amplification of diagnostic fragments, which can be influenced by factors such as variations in DNA purity, magnesium chloride concentration, annealing temperature, primer to template ratio and different thermostable polymerases. All of these can change banding patterns. Some RAPD markers can amplify despite primer-template mismatch. Anomalous RAPD bands can occur if there is competition among binding sites, as well as contaminant DNA from parasites and symbiotic organisms.

There is, as with SSR bands, the assumption that co-migrating bands represent homologous loci. More accurate staining by PAGE and silverstaining can reveal errors, although it is more expensive and time consuming than simple agarose gel electrophoresis.
Southern hybridization can be used to test homology. In this technique, a single band is excised from a gel and used as a hybridisation probe. If the probe hybridizes to the equal length bands in other amplified samples, the assumption of length homology is correct. This verification is also time consuming, although in several studies, bands of equal length have been found not to be homologous. An alternative test of homology is via restriction enzyme digestion of gel-isolated fragments. Fragments that generate the same restriction profile for two (or two of three) tested endonucleases are considered homologous.

RAPD markers tend to underestimate genetic distances between distantly related individuals, for example in inter-specific comparisons. It is wise to be cautious when using RAPD for taxonomic studies above the species level.

AFLP (Amplification Fragment Length Polymorphism)
This technique uses restriction fragment analysis, but a PCR approach is used instead of hybridization with a probe. Extracted DNA is cut with two different restriction enzymes (a rare and a frequent cutter) to produce well-defined restriction fragments with sticky ends. Synthetic double stranded linkers (adaptors) of about 20 bp with matching sticky ends are ligated on all the restriction fragments. These ligated fragments are then amplified in PCR with about 20 nucleotide length primers recognising linkers in each end of the fragments. The primers used are usually labelled with P33 or some nonradioactive labelling system. The resulting product is visualised on an agarose or polyacrylamide gel. Sequencing machines can be used to read the sequence of the amplification products directly from the acrylamide gel if primers have been labelled with fluorescent dyes.

Smearing can result from a high number of products, which can be reduced using a two step amplification procedure. First, part of the total amount of restriction fragments are amplified in the "preamplification" step, with primers containing one extra "selective" nucleotide on their 3' end. This selective nucleotide will only allow amplification of restriction fragments with a matching nucleotide next to the linker. The next amplification (the 'selective' amplification), an additional one to two (normally three) selective nucleotides on the primers will further reduce the number of fragments down to a manageable number (50 - 100 fragments), and include a radiolabelled primer.

In AFLP analysis the genomic DNA of the organism is extracted as for RFLP. During the restriction step the purified DNA is cut usually with two different restriction enzymes to produce restriction fragments with sticky ends characteristic for each enzyme.

In the ligation step, synthetic double stranded linkers with matching sticking ends are ligated on the DNA fragments.During preamplification, a fraction of the ligated restriction fragments are amplified in a preamplification PCR reaction with primers annnealing to the ligated linkers. The preamplification uses primers with one selective nucleotide. This means the primer has got one nucleotide extra at its 3' end beyond the ligated linker. Only restriction fragments with the correct nucleotides in both ends will amplify. Only two of the four DNA fragments (framed) fit the primers used and amplify.

In the selective amplification new primers equal to the ones used for preamplification but with one or two extra selective nucleotides at the 3' end are used to selectively amplify a fraction of the fragments for analysis. In the figure only one of the preamplifyied fragments fit the primers at both ends and will amplify in the reaction.

AFLP reactions can be analysed by separation according to fragment size on a polyacrylamide gel. Visualisation is often done by autoradiography if one primer has P^{33}labeling. The simplified autoradiograph from an AFLP shows 27 different amplification products.

Advantages

The major advantage of the AFLP technique is the large number of polymorphisms that the method generates. Most applications of AFLP have been for gene mapping. Its ability to differentiate individuals in a population makes the technique useful for paternity analyses, geneflow studies and cultivar research. Other advantageous features of the AFLP technique are that no sequence information is required , the PCR technique is fast, results are reproducible, and standard kits are available.

Problems

The number of steps needed to produce results can be a limitation, instead of only one round of PCR amplification, this technique involves two. Additional costs come from the use of an automated sequencer, or sequencing gel apparatus, instead of simple agarose gel electrophoresis and ethidium bromide staining.

The choice of primer may have a large influence on the amount and quality of variation uncovered. The choice of primer may influence the number of bands amplified and the level of polymorphisms found, which in turn is linked to the taxonomic level of the investigation.

AFLPs are generally acclaimed for their reproducibility, which sets the technique apart from RAPDs. Tests in a network of laboratories have shown that it is possible to reproduce banding patterns across a range of laboratories. Duplications from the same plant were also examined, and while three duplicates returned exactly the same banding patterns, the other two were 98.9% and 97.6% similar, although some other studies have found different banding patterns when samples were re-run.

The scoring of AFLP fragments, like RAPDs, is open to a certain amount of interpretation, with many studies stating that only "distinct, reproducible, well-resolved" bands were scored (sometimes less than half of all the polymorphic bands found), without defining the criteria for these bands. This can lead different people to score the same results in different ways.

It is often assumed that co-migrating bands are homologous, although this may not always be the case as a particular sized band may consist of bands from different regions of the genome. Some researchers argue that the chance that two co-migrating AFLP fragments do not represent identical alleles of one locus is small, which they believe is due to the highly selective amplification and sharp resolution of PAGE.

Microsatellites

Also known as Simple Sequence Repeats (SSRs) short tandem repeats (STRs) or simple sequence length polymorphisms (SSLPs), microsatellites are tandem repeats of sequence units generally less than 5 bp in length, e.g. (TG)n or (AAT)n . They are thought to be produced by errors in DNA replication, when the DNA polymerase "slips" when copying in the repeats region, changing the number of repeats in the region (Robinson & Harris, 1999).

Microsatellites appear to be hypervariable, in addition to which their co-dominance and reproducibility make them ideal for genome mapping, as well as for population genetic studies and diversity studies. Inter-SSRs (ISSRs) are a variant of the RAPD technique, as are chloroplast microsatellites (cpSSRs), which are similar to nuclear microsatellites but the repeat is usually only 1 bp, i.e. (T)n.

The microsatellite protocol is simple, once primers have been designed. The first stage is a PCR, depending upon the method of detection one of the primers is fluorescently or radioactively labelled. The PCR products are separated on a high resolution polyacrylamide gels, and the products detected with a fluorescence detector (for example an automated sequencer) or an X-ray film.

Advantages

As with AFLPs, the great advantage of microsatellite analysis is the large number of polymorphisms that the method reveals, with one locus in soybean (Glycine max) reported to have 26 alleles. Unlike AFLPs, microsatellites are co-dominant markers, thus heterozygotes can be readily identified. Microsatellite co-dominance will increase the efficiency and accuracy of population genetic measures based on these markers compared with other markers, such as AFLPs and RAPDs.

Problems

Unless useful primers have been designed in previous studies, it is necessary to screen an organism for microsatellites. There are many different ways of screening, all of them are practically complex and expensive and may yield only a small number of potential microsatellite loci.

Slippage can be a significant problem when analysing mono- and di-nucleotide repeats. During the amplification process the polymerase can "slip", leading to the production of

differently sized products that differ by approximately 1-5 repeat units from the expected product. Such products are usually less intense than the desired product, and in practice can usually be discounted. However, if the products of a heterozygous individual overlap then it is sometimes difficult to differentiate the true and slippage products.

Inaccurate allele identification may also be caused by the tendency of Taq polymerase to add an adenosine nucleotide to the 3'-end of the amplified product. This addition is determined in a template- and marker-specific manner, which may not be a problem if the extra nucleotide is always, or never, added. However, errors may occur in size determination if the extra nucleotide is only occasionally added.

Homology is the greatest problem facing the use of SSRs in phylogenetic analyses. Microsatellite analyses assume that co-migrating fragments are homologous, which may not always the case. Caution should be advised in the exclusive use of band size in the interpretation of SSR length polymorphisms. Indel mutations can cause a change in the size of an allele and may lead to misidentification of an allele if only size was measured. Since no large-scale tests of SSR homology have taken place in plants, it is difficult to estimate the percentage of bands in a microsatellite survey that are non-homologous.

The most problematic homology uncertainty is within the repeat unit, whether two fragments that co-migrate are identical by descent or just identical in state. The problem of homology depends upon the mutation rate of the repeats. If it is low then the probability that a mutation is unique and similar alleles are identical by descent is high. Vice versa, if mutation rate is high then the probability increases that two co-migrating alleles are just identical in state and non-homologous.

Mutations in the binding region of one or both of the microsatellite primers may inhibit annealing that may result in the reduction or loss of the PCR product (null alleles). These null alleles may be manifested as fewer heterozygotes than expected in a randomly mating population or by the appearance of "empty" lanes.

ISSR (Inter-Simple Sequence Repeat)

This is a technique in which primers that anneal to microsatellites during PCR lead to amplification of the piece of DNA between two primers that are within a PCR-able distance. The primers for this system are constructed to anneal to microsatellites with one extra nucleotide in the 3'-end. The extra nucleotide will permit amplification only if the primer binds to a 5'-end of a microsatellite with a suitable first nucleotide in the flanking sequence. Such extra nucleotides are termed anchors and they assure that amplification will always start from the 5'-end of the microsatellite. ISSR will often amplify 25 to 50 products in one reaction. Polymorphisms are mostly of the dominant type because of changes in the anchoring nucleotide, but codominant types occur if length of the intervening space between the microsatellites has changed (Andersen, 2000c). In ISSR, primers that anneal to microsattelites are used. To assure amplification only if the primer anneals correctly to the 5'-end of the microsattelite such primers have a selective anchoring nucleotide at their 3'-end. The primer will only be elongated by a polymerase if this nucleotide fits the first nucleotide outside the microsatellite.

CAACAA...CAAG

In the above primer the last G-nucleotide is the anchor. If both primers anneal correctly, the sequence between two microsatellites are amplified ISSR markers were developed from the common SSR motifs present in eukaryotic organisms, based on the premise that SSR loci are evenly dispersed throughout the genome, and the chance of hitting two SSRs with a common motif, orientated on opposing DNA strands within amplifiable distance of one another was high enough that single-primer amplifications should yield a high degree of polymorphic bands. The resulting PCR products are anonymous SSR loci. The resulting amplified fragments are resolved on agarose or polyacrylamide gels.

ISSR markers are inherited as dominant or codominant genetic markers and are interpreted using presence or absence of bands. The absence of a band presumably signifies that divergence has occurred at either one of both of the primer sites. It could also signify that a SSR site (ISSR primer binding site) might have been lost or a chromosomal structure rearrangement may have taken place.

ISSR has been used to assess the abundance and/or inheritance of SSR loci in the genome; to identify cultivars and gather preliminary data for genome mapping projects.

Because ISSRs are a variant of the RAPD technique, many of the problems of that technique (band scoring and homology) are shared with it, as are the advantages (no prior sequence knowledge and large number of primers available); (See RAPD section above).

Conclusion

No method is without its problems, and each is useful in its own way. The choice of technique must lie with the nature of the research undertaken, and the problems of each method must be factored into both the initial choice of technique as well as the analysis of the resulting data from that research

References

1. Caetano-Anolles 1996. Scanning of nucleic acids by *invitro* amplification. Nature Biotechnology 14:1668-1674
2. Caetano-Anolles, Trigiano,R.N. 1997.Nucleic acid markers in agricultural biotechnology
3. Kozian D.H., Kirschbaum J.1999. Comparative gene- expression analysis.TIBTECH. 17:73-77.
4. Nanda S.K, Jain S.K.1994.In vitro nucleic acid amplification systems, Curr. Sci.66:421-429.

DNA Microarray Technology

A. Ishwara 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 new technology that 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 Figure 1).

2. Mechanical microspotting
It is used in the production of cDNA microarrays. In this method 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 (Figure 1).

3. In jetting

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.

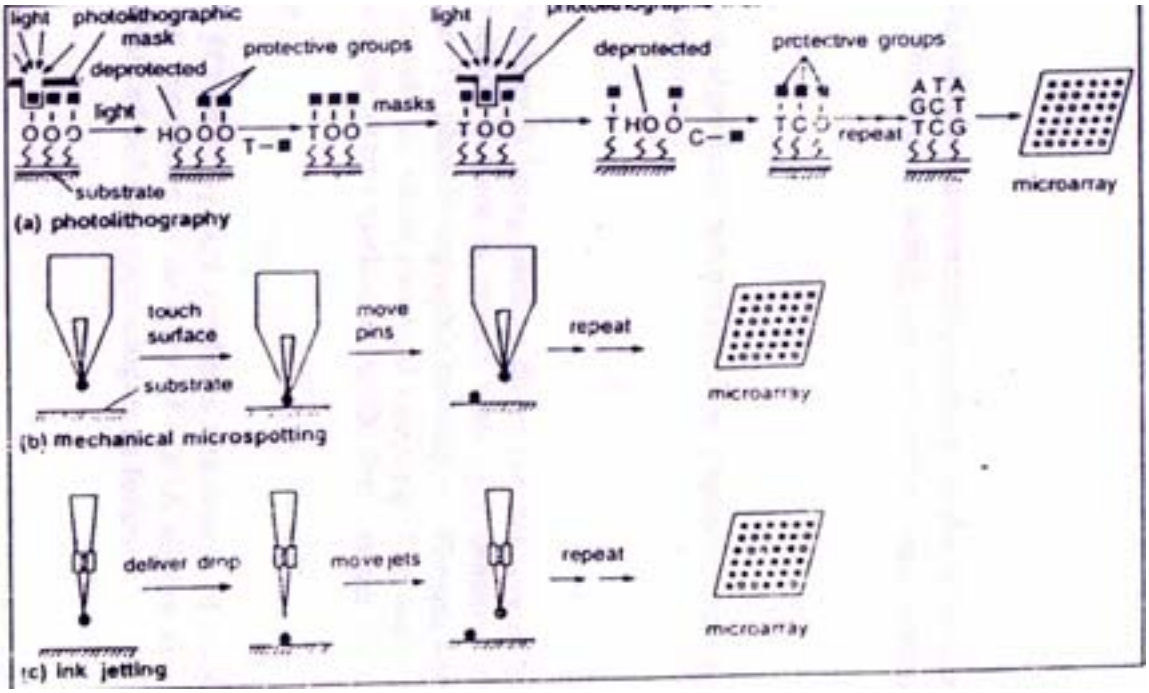


Figure 1: Line diagram depicting principal steps in the production different kinds of microarrays

Hybridization and Detection method

Hybridization of the target DNA to a microarray yields sequence information. The target DNA is labeled and incubated with array. If the target DNA has regions complementary to the probe on the array, then the target DNA will hybridize with probes under a fixed set of hybridization conditions. The fraction of probes bound to targets will vary with a base composition with the probes and the extent of the target probe match. In general for a given land probes with high GC contents will hybridize more strongly than those with high AT contents. Similarly probe matching the target will hybridize more strongly than probes with miss matches, insertion and deletions. Various detection methods are currently available for the analysis of hybridization patterns on microarrays of immobilized probes.

The target DNA is either non-radioactively labeled (biotin or digoxigenin labeling) or radioactively labeled. However there are drawbacks with the detection methods involving radioactivity (low resolution). In order to avoid this problem fluorochromes may be used which will also allow direct detection due to fluorescence. This also allows multiplexing, where more than one target DNAs labeled with different flurochromes can be used for

hybridization of microarrays and DNA chips. The hybridization patterns can be scanned either on lens based system (confocal microscope) or on CCD.

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 probe concentration that are equal to or less than the target concentration display second order hybridization kinetics.

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 excess a positive effects 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°C and 55 to 70°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

Conclusion
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

1. Granheaud, S., Bertucci, F and Jordon, B.R. 1999. Expression profiling: DNA microarrays in many guises. BioEssays, 21: 781-790.
2. Heller, R.A., Schena, M., Chai, A., Shelon, D nad Bedilion, T. 1997. Discovery and analysis of inflammatory disease related genes using cDNA microarrays. *Proc. Natl. Sci., USA* **94**: 2150-2155.
3. Lemieux, B., Aharoni, A and Schena, M. 1998. Overview of DNA chip technology. *Molecular Breeding* **4**: 277-289.
4. Ramsay, G. 1998. DNA chips: State of art. *Nature Biotech*, **16**: 40-44.
5. 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.
6. Southern M.E.D. 1996. DNA chips: Analysing sequence by hybridization to oligonucloetides on a large scale. *Trends Genetics* **12**: 110-115.

Geographic Information System (GIS) In Biological Science

M.S. Madan

Informatics is the discipline that integrates information technologies with scientific and technical disciplines. While bioinformatics deals with the vast amounts of genome data, geoinformatics works with scientific data that have spatial or geographic coordinates. Since, the subject geoinformatics is heavily dependent on the software 'Geographic Information System', it is essential to know the answer for the question: what is GIS? It is hard to define the term Geographic Information System (GIS), which represents the integration of many subject areas. A broadly accepted definition of GIS is the one provided by the National Centre of Geographic Information and Analysis:

> *GIS is a system of hardware, software and procedures to facilitate the management, manipulation, analysis, modelling, representation and display of geo-referenced data to solve complex problems regarding planning and management to resources* (NCGIA, 1990)

An understanding of what Geographic Information Systems represent may be made easy when we define the component parts of the term separately.

Geographic...
The term 'Geographic' is used because GIS tend to deal primarily with `geographic' or `spatial' features. Here the objects are referenced or related to a specific location in space. The objects may be physical, cultural or economic in nature and presented in the form of maps. Features on a map are pictorial representations of spatial objects in the real world. Symbols, colours and line styles are used to represent the different spatial features on the two-dimensional map.

Information...
This represents the large volumes of data, which are usually handled within a GIS. All real world objects have their own particular set of characteristics or descriptive attributes. This non-spatial alphanumeric data plus locational information needs to be stored and managed for all spatial features of interest. Historically maintained as paper files, computer technology has enabled much more efficient handling and management of information within automated database management systems.

Systems...
This term is used to represent the systems approach taken by GIS, whereby complex environments are broken down into their component parts for ease of understanding and handling but are considered to form an integrated whole. Computer technology has aided and even necessitated this approach so that most information systems are now computer based.

Computer systems are becoming vital for the storage and manipulation of the increasing volumes of data, the handling of complex spatial algorithms and the integration of data of different scales, projections and formats. All of which are essential to GIS.

GIS is thus a computer based information system used to digitally represent and analyse the geographic features present on the Earth' surface and the events (non-spatial attributes linked to the geography under study) that taking place on it. "Every object present on the Earth can be geo-referenced", is the fundamental key of associating any database to GIS. Here, term 'database' is a collection of information about things and their relationship to each other, and 'geo-referencing' refers to the location of a layer or coverage in space defined by the co-ordinate referencing system.

Work on GIS began in late 1950s, but first GIS software came only in late 1970s from the lab of the ESRI. Canada was the pioneer in the development of GIS as a result of innovations dating back to early 1960s. Much of the credit for the early development of GIS goes to Roger Tomilson. Evolution of GIS has transformed and revolutionized the ways in which planners, engineers, managers etc. conduct the database management and analysis.

GIS technology integrates common database operations such as query and statistical analysis with the unique visualization and geographic analysis benefits offered by maps. GIS has the ability to relate the attribute information to the spatial locality. These abilities distinguish GIS from other information systems and make it valuable to a wide range of public and private enterprises for explaining events, predicting outcomes, and planning strategies. (ESRI)

GIS and Other Related Disciplines of Science

GIS is an integrated multidisciplinary science consisting of many traditional disciplines. Table 1 summarizes how these disciplines make up GIS with respect to the functions.

Table 1. Relations of Traditional Disciplines with GIS

Discipline	Functions of GIS								
	Data Acquisition	Mapping	Pre-Processing	Data Structure	Data Base	Spatial Analysis	Modeling	Display	Appli-cation
Geography		*				*			*
Cartography	*	*						*	*
Remote Sensing	*	*						*	*
Photogrammetry	*	*						*	*
Surveying	*	*							
Geodesy		*							
Statistics			*		*	*			
Operations Research						*	*		
Computer Science			*	*	*	*	*	*	
Mathematics				*		*	*		
Civil Engineering						*	*		
Urban Planning						*	*		*

GIS is now becoming an independent discipline in the name of "Geomatic", "Geoinformatics"; or "Geospatial Information Science" that is used in many departments of government, university and research organizations world over.

Answers GIS Can Give

There is another way to describe GIS by listing the type of questions it can (or should be able to) answer. Location, Condition, Trends, patterns, Modeling, Aspatial questions, Spatial questions. There are five type of questions that a sophisticated GIS can answer:

Location (What is at…….?)
The first of these questions seeks to find out what exists at a particular location. A location can be described in many ways, using, for example place name, postal code, or geographic reference such as longitude/latitude or x/y.

Condition (Where is it…….?)
The second question is the converse of the first and requires spatial data to answer. Instead of identifying what exists at a given location, one may wish to find location(s) where certain conditions are satisfied (e.g., an uncultivated land parcel of at-least 2000 square meters in size, within 100 meters of national high way, and with soils suitable for supporting buildings)

Trends (What has changed since……..?)
The third question might involve both the first two and seeks to find the differences (e.g. reduction or increase in area under forest in the state of Kerala) over time.

Patterns (What spatial patterns exists……..?)
This question is more sophisticated. One might ask this question to determine whether landslides are mostly occurring near streams in the high ranges. It might be just as important to know how many anomalies there are that does not fit the pattern and where they are located.

Modeling (What if……..?)
"What if…" questions are posed to determine what happens to the spice industry of the state of Kerala, for example, if a new road is added to connect all the growing centers with marketing centers and port cities? or if the chemicals and fertilizers used in spice cultivation seeps into the local ground water supply? Answering this type of question requires both geographic and other information (as well as specific models). GIS permits spatial operation.

Aspatial Questions
"What's the average number of rainy days/year in pepper growing districts of Kerala ?" is an aspatial question - the answer to which does not require the stored value of latitude and longitude; nor does it describe where the places are in relation with each other.

Spatial Questions
" How many cardamom marketing (auction) centres located in the states of Kerala, Tamil Nadu and Karnataka " OR " Which auction centre lie within 50 Kms. of each other? ", OR " What is the shortest route passing through all the auction centers in Kerala state". These are spatial questions that can only be answered using latitude and longitude data and other information such as the radius of earth. Geographic Information Systems can answer such questions.

In order to meet the above requirements and answer the questions the following functions are necessary for GIS (Table 2).

Data Acquisition and Pre-processing

Data used in GIS often come from many types, and are stored in different ways. A GIS provides tools and a method for the integration of different data into a format to be compared and analysed. Data sources are mainly obtained from manual digitization and scanning of aerial photographs, paper maps, and existing digital data sets. Remote-sensing satellite imagery and GPS are promising data input sources for GIS.

Database Management and Update

After data are collected and integrated, the GIS must provide facilities, which can store and maintain data. Effective data management has many definitions but should include all of the following aspects: data security, data integrity, data storage and retrieval, and data maintenance abilities.

Table 2. Basic functions of GIS

Functions	**Sub-Functions**
Data Acquisition and preprocessing	Digitizing Editing Topology Building Projection Transformation Format Conversion Attribute Assignment
Database Management and Retrieval	Data Archival Hierarchical Modeling Network Modeling Relational Modeling Attribute Query Object-oriented Database etc.
Spatial Measurement and Analysis	Measurement Operations Buffering Overlay Operations Connectivity Operations etc.
Graphic Output and Visualization	Scale Transformation Generalization Topographic Map Statistical Map 3D Bird's Eye View etc.

Presenting Results

One of the most exciting aspects of GIS technology is the variety of different ways in which the information can be presented once it has been processed by GIS. Traditional methods of tabulating and graphing data can be supplemented by maps and three dimensional images. Visual communication is one of the most fascinating aspects of GIS technology and is available in a diverse range of output options.

Applications of GIS

The applications or uses for a GIS are endless, wherever spatial features need to be modeled and analyzed. GIS is being applied in remarkable ways in diverse range of field. Stand-alone sub-systems of GIS has been developed depending on the field of concern. The commonly available subsystems are:

- Land Information System (LIS)
- Environmental Information System (LIS)
- Resources Information System
- Spatial Data Handling System
- AM/FM
- Planning Information System

Listed below are a few descriptions of specific field oriented applications of GIS.

Agriculture: GIS and remote sensing technologies contribute substantially to the agricultural industry. Land suitability analysis, water shed management, land reclamation, precision farming are some of the fields of agriculture, where GIS technology is used. Land records in many states are already digitized and has reduced the volume of files and papers.

Ecology and conservation: Colombia is building a GIS database to prioritize which lands should be set aside to the national park service. GIS also helps in understanding seasonal migration patterns is important in managing water access for wildlife and livestock. GIS is being applied to evaluate the ecological costs and benefits because they provide access for ecological management but also interrupt the ecological landscape.

Business geographics: GIS is used to evaluate how the pending relocation of a corporate office would affect employees commute to work or to mine customer databases to identify clusters of customers, enhance productivity of mail promotions, and improve client retention.

Defense and intelligence: The U.S. Air Force uses GIS technology to manage, maintain, and visualize millions of climatological records. The Swedish armed forces have done extensive work on sophisticated symbolizing of military and civilian facilities to improve military planning. The Canadian Army has customized GIS software to integrate it with a land force command system.

Electric and gas (AM/FM): GIS is used to analyze power circuits to minimize losses and to improve voltage levels. It is modeling scenarios of device placement for optional electrical benefit. GIS also manages construction, operation, and maintenance of power transmission. A database on energy usage of every building in a country can be built and this information can be used for planning energy plants and designing distribution systems. Bangalore Development Authority has developed GIS based stand-alone software for maintenance of water supply sewage system maintenance in the Bangalore corporation area.

Emergency management and public safety: The provision of optimum locations for emergency service centres can also be aided by GIS analysis of the various parameters such as access to roads, population density and various health indicators. Network analysis can be utilized to

define optimum routes, such as shortest or fastest, for the routing of emergency service vehicles. NIIT-GIS Ltd. has prepared Emergency Response System for Delhi and is put in to operation.

Environmental resource management: GIS is being used for a complete evaluation of its effects on irrigation, hydropower, health, mining, education, tourism, and telecommunications. Ecological balance models are combined with GIS software to provide tools for environmental management. In Korea, land zoning in national parks is being analyzed with the criteria of scenic quality, elevation, slope, and natural state. It was found that some parks were not correctly zoned.

Federal and state government systems: The Tennessee Valley Authority has built a land information system to help administer land records, natural and cultural resources, land-use planning, and compliance with laws and executive orders. The U.S. National Oceanic and Atmospheric Administration is building a tool to collect metadata such as bounding coordinates, map projections, and attribution information. In Qatar, television cameras are being inserted into water and sewer networks to create video records of pipe conditions. These images are integrated with in a GIS and give operators information for maintenance. GIS is being applied to develop scenarios of land-use patterns over the next five, 10 and 15 years. In the Ukraine, a new land registration system is being developed based on high-resolution satellite imagery and innovative software techniques. Egypt government ministry is building a system to manage the Nile River channel, canals, drains, and pumps. In Florida, a hydraulic computer model is used to reduce sanitary sewer overflows. When major rainstorms come, satellite imagery is used to estimate rainfalls and assist in the operation of sewer pump stations. In Canada, a hydrodynamic / pollutant transport model has been built to simulate the effects of multiple pollution sources under different conditions

Forestry: The construction and use of roads in forested basin can contribute significantly to sediment deposition and therefore building a road sediment model can establish a maintenance plan. GIS is used to calculate colony areas and foraging zones of endangered species. In India forest departments are the pioneers in applying GIS to monitor the changing pattern of forest cover in the country.

Health care: GIS is being used to present geographic, socioeconomic, demographic, and health care utilization data. The epidemiology of rare diseases and an individual's exposure to environmental risk factors can be estimated using GIS. GIS is also being used to examine factors such as age, race, education, elevation, and access to public health programs.

Education: An educational agency is using GIS to help students discover geography and foster critical thinking and inquiry. A high school is incorporating GIS in its curriculum to teach students a "sense of place" by showing them how their personal actions have relevance on a global scale.

Mining and geosciences: GIS is used to monitor acid mine drainage on surface waters, elevations, hydrology, mined areas, and water quality data. A mining service company is using GIS to create three-dimensional databases for nuclear waste repositories, mineral exploration programs, and groundwater monitoring purpose.

Oceanography, coastal zone and marine resources: The U.S. Naval Oceanographic Office is using remotely sensed sea temperature data to study oceanic fronts and eddies. In Washington State, a GIS is mapping the current shoreline, calculating change rates, and projecting shoreline erosion hazards.

Real estate: Habitat for Humanity, an organization building houses for low-income families, uses GIS to analyze a proposed subdivision and create a plan that preserves most of the existing trees. A real estate companies are using GIS for site selection for multi-site users by considering factors like access, visibility, zoning, and entitlement process.

Remote sensing and imagery: A digital imagery company is using geo-referenced airborne sensors to create real-time spatial data. Images are sent to ground stations and are fused, reformatted and subject to automatic feature extraction.

Telecommunications: In Colombia, the fiber-optic trunk network is being captured in a GIS database with a representation of each of the network's element features. In Indonesia, GIS is employed to manage radiotelephony by studying radio station placement, the demographics of a customer area, and the maintenance of equipment. A telecommunications-consulting firm is using data on land use and land cover to predict signal attenuation for wireless communication systems.

Transportation: In Korea, GIS monitors real-time traffic conditions to mitigate traffic bottlenecks on freeways. The State of Georgia applies GIS technology to manage roadway pavement. Customized applications developed by RITES for validation of proposals for Delhi Corporation has helped in planning and executing the Delhi Metro rails. Roads and motorways need to be maintained and monitored for accident trouble spots. GIS are ideal for representing the spatial relationships between sections and storing the associated information tied to each section of road. Maintenance records can also be incorporated into the GIS and so provide up to date displays of the state of the road network and the sections which require immediate maintenance.

Facility management for the utilities: The utility industries tend to have vast numbers of facilities to manage in order to provide large customer regions with an efficient and reliable service. Gas, water, electricity and sewage utilities for instance own a lot of land, buildings, cables, pipes and other physical facilities which need monitoring, maintaining and managing in order to provide an effective service.

Market analysis: The spatial distribution of the population and particularly the different age groups and socio-economic sectors are essential information to the market analyst attempting to discover the most suitable place to launch a new product or sell a particular brand. The effectiveness of any given marketing strategy can also be modeled and evaluated.

Population analysis and prediction: The spatial distribution of the population and the predicted level of a population are essential information to planners and developers when deciding what type of facilities need to be constructed now in order to best suit the needs of the future population. Census data thus provide an important input to GIS.

GIS thus essentially enable the relationships between various spatial features to be visualised and analyzed, which in turn encourages a better understanding of the interactions between

the various features. GIS also enable the data to be manipulated and analysed quickly and flexibly in a single system, which is an extremely powerful capability.

Components of GIS

GIS constitutes of five key components: hardware, software, data, users/people and methods. Data tend to be at the center of any GIS system, while the computer components of the system support data management and analysis.

Hardware

It consists of the computer system on which the GIS software will run. The choice of hardware system range from 300MHz Personal Computers to Super Computers having capability in Tera FLOPS. The computer forms the backbone of the GIS hardware, which gets its input through the scanner or a digitizer board. Scanner converts a picture into a digital image for further processing. The output of scanner can be stored in many formats e.g. TIFF, BMP, JPG etc. A digitizer board is flat board used for vectorisation of a given map objects. Printers and plotters are the most common output devices for a GIS hardware setup.

Software

GIS software provides the functions and tools needed to store, analyze, and display geographic information. GIS softwares in use are MapInfo, ARC/Info, AutoCAD Map, etc. The software available can be said to be application specific. When the low cost GIS work is to be carried out desktop MapInfo is the suitable option. It is easy to use and supports many GIS feature. If the user intends to carry out extensive analysis on GIS, ARC/Info is the preferred option. For the people using AutoCAD and willing to step into GIS, AutoCAD Map is a good option.

People

GIS users range from technical specialists who design and maintain the system to those who use it to help them perform their everyday work. The people who use GIS can be broadly classified into two classes. The CAD/GIS operator, whose work is to vectorise the map objects. The use of this vectorised data to perform query, analysis or any other work is the responsibility of a GIS engineer/user.

Data

Geographic data and related tabular data can be collected in-house or purchased from a commercial data provider. The digital map forms the basic data input for GIS. Tabular data related to the map objects can also be attached to the digital data. The primary requirement for the source data is that the locations for the variables are known. Locations may be annotated by x, y and z coordinates of longitude, latitude and elevation or by highway mile markers. A GIS will integrate spatial data with other data resources and can even use a DBMS, used by most organization to maintain their data, to manage spatial data.

Sources of GIS Data: Various sources of data are digitized and scanned maps which are purchased, shared or obtained from Internet or created by user from toposheets and traced maps, databases, GPS – Global Positioning System, field sampling of attributes, remote sensing imagery, aerial photographs and geography networks.

Geographic data types: Although the two terms, data and information, are often used indiscriminately, they both have a specific meaning. Data can be described as different observations, which are collected and stored. Information is that data, which is useful in

answering queries or solving a problem. Digitizing a large number of maps provides a large amount of data after hours of painstaking works, but the data can only render useful information if it is used in analysis.

Spatial and non-spatial data: Geographic data are organised in a geographic database called geo-database. This database can be considered as a collection of spatially referenced data that acts as a model of reality. Geospatial data has both **spatial** and **thematic** components i.e. its geographic position and its attributes or properties. In other words, spatial data (where is it?) and attribute data (what is it?).

Attribute data: The attributes refer to the properties of spatial entities. They are often referred to as non-spatial data since they do not in themselves represent location information.

Spatial data: Geographic position refers to the fact that each feature has a location that must be specified in a unique way. To specify the position in an absolute way a coordinate system is used. For small areas, the simplest coordinate system is the regular square grid. For larger areas, certain approved cartographic projections are commonly used. Internationally there are many different coordinate systems in use. Geographic object can be shown by four types of representation viz., points, lines, areas, and continuous surfaces.

Continuous surface: Continuous surfaces are three-dimensional spatial objects with not only a position in space, a length and a width, but also a depth or height (in other words they have a volume). These spatial objects have not been discussed further because most GIS do not include real volumetric spatial data.

Thus the GIS database contains both map data (depicting location of geographical objects) and attribute data (describing physical characteristics of each object). Physical characteristics (such as timber species and tree diameter) and/or non-physical characteristics (such as estimated market value and management codes) are examples of attribute data that could be contained in a GIS used to analyze forestry problems. During a GIS analysis, site (map) data is linked with situation (attribute) data for each mapped timber stand. It is this link, which is automatically performed by the GIS software that gives GIS its analytical power. The relationship between map data and its associated attribute data are shown in Figure 4. for typical natural resource and urban maps.

GIS data can broadly be grouped into two types data models based on the type of analysis for which the data is used:

- **Raster**
  - Pixels
  - Rows, columns and value
  - Satellite images and aerial photos
  - Eg. Hydrologic Modeling, surface analysis
- **Vector**
  - Shapes (Points, lines and polygons)
  - Features (Districts, Roads etc.)
  - Attributes – size, type, length, etc.
  - Eg. Network analysis

In GIS terminology, the individual spatial phenomena or map themes are referred to as map layers. One layer can contain roads, another soils, and another can indicate land ownership. Each layer comprises all of the pertinent map and attribute data (Fig. 1).

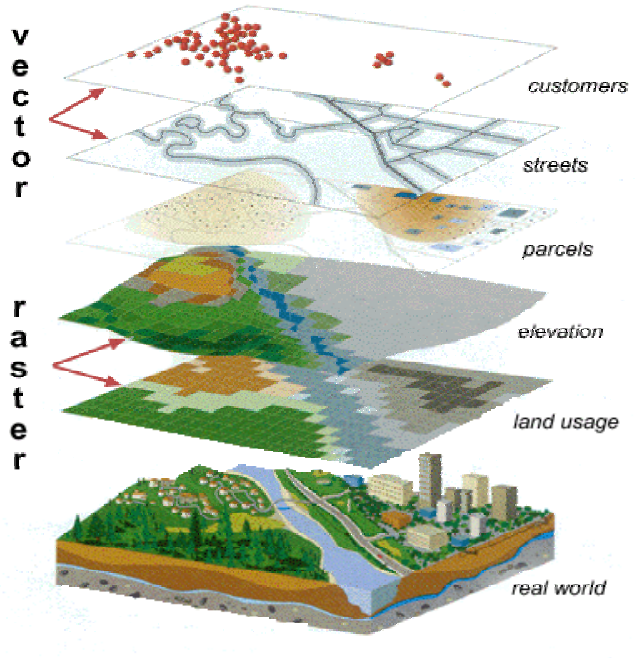


Fig. 1. The concept of layers (ESRI)

A vector model stores all spatial data as primitive features of points, lines, and polygons. Points are stored as discrete coordinate pairs, lines as linked sets of coordinates, and polygons as areas with an alternative form of a line that has the same beginning and ending coordinates. Attribute tags such as "lake" are then added to the spatial data file. The end result of converting a map to a vector data file is a GIS-compatible digital. Figure below present types of presentations on a map.

Data Analysis

The Analytical and Operational Functions, examines how GIS software links map and attribute data in order to analyze spatial problems. Geographical Information Systems are used to perform a number of fundamental spatial analysis operations. In general GIS is used in getting done operations such as a) information retrieval, b) network analysis (optimal routing, finding closest facilities, finding service areas, tracing the network etc.), c) surface analysis and d) criterion analysis.

All these operations can use any number of analytical processes. However almost all spatial analyses utilize some fundamental operations such as 1) Topological Map Overlay, 2) Buffer Generation, 3) Feature Extraction, 4) Feature Merging and 5) Relational Database Operations - Join and Relate.

Topological Map Overlay

Topological map overlay creates new features and attribute relations by overlaying the features from two input map layers. Features from each input layer are combined to create

new output features. Attributes of each input feature are combined from the two input layers to describe each new output feature, thus creating new attribute relationships.

Often, the manipulation of multiple data layers is required to achieve the objective of the overlay operation. This is done in a stepwise fashion - two input layers are combined to form an intermediate layer, this intermediate layer is then combined with a third layer to form another intermediate layer, and so forth until the desired resultant map layer is achieved.

Use of GIS is driven by the need to answer geographical or spatial questions. The ensuing data collection, database analysis, and output are in response to those questions. The purpose of collecting data for a GIS is (a) to inventory a geographically defined area (for example, to locate all state-owned parcels within a particular area) or (b) to test hypotheses and build models. Though the data is initially collected for one of these reasons, data collected to answer one set of questions is frequently used in subsequent analyses to answer questions that were not anticipated at the time the data was collected. For example, after capturing detailed soils data to determine what crops can be grown, the data may be useful for answering questions about septic tank suitability.

Spatial retrieval, delineation and classification, and measurement are major three separate operations, but are commonly used together in analysis of data. For example, a GIS user interested in legislative reapportionment may want to retrieve a district map, classify zones according to party affiliation, and determine the size of the area under study**.**

Buffering and Networking

The buffer function examines an area, which surrounds an object of interest. This function is used to create zones and to determine routes within zones. For example, a zone can be created based on specified distances from map features (e.g., the area within five miles of a road). Network functions examine the movement of objects along an interconnected pathway (e.g., traffic flow along a map of highway segments).

Overlay Analysis and Merging

The overlay function allows the user to "stack" map layers on one another, showing spatial relationships between the layers. Merging generalizes classes within map layers by combining attributes to reveal new map features.

Map algebra utilities allow the user to specify mathematical relationships between map layers. Entire maps can be added, subtracted, multiplied, and divided according to user-specified rules. For example, a new map can be generated by determining the difference in elevation between a topographic map and the corresponding map of the water table.

Conclusion

As an integrated approach to managing and analyzing map and attribute data, geographic information systems are becoming increasingly common. GIS applications begin with a spatial question and proceed to a data collection phase during which both map and attribute data are collected. The initial spatial question drives the data collection and the model-building process. In the analysis phase, which follows, both map and attribute data are manipulated via the analytical tools provided by the GIS software.

The tools presented in this paper for performing certain analyses are commonly used and represent broad categories of functionality. However, this discussion does not constitute a complete or comprehensive listing of GIS tools. The increasing influence and use of GIS can be largely attributed to its ability to support decision-making. Toward this end, a GIS needs reliable data and must be able to analyze and synthesize this data quickly. For the GIS to reach its full potential it must be soundly implemented. This implementation must consider database development and the accuracy and the long-term maintenance of data sets.

During the data acquisition stage, special attention must be paid to data accuracy, quality, and scale. Most data collected for a GIS comes from published maps. The scale of these maps relates to the relative spatial accuracy or precision of the data they contain. The degree of accuracy needed depends on the type of application. In defining minor watersheds for a county, locations specified to +/- 40 feet is well within acceptable limits because 40 feet is relatively insignificant for that particular analysis. However, in parcel mapping, where property lines need to be defined, a more accurate map at a larger scale with more accurate and precise data would probably be needed.

Collection and creation of the data to be stored in the database is typically one of the largest expenditures in implementing a GIS. For this reason, data must be as reliable and accurate as possible. Complete documentation of all databases is important to assure the integrity of the data, the reliability of subsequent analyses, and the ability to maintain the system over time. A commitment must be made to maintain these databases. Remember that the utility of GIS is a direct function of the data contained within the system. Only dedicated care of the database can ensure that database manipulations and spatial analysis will produce results of the highest order.

Internet Resources In Bioinformatics

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 littoral 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 organize the information associated with these molecules, on a large scale. In short, 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 make up. 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 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. 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 Gene Bank. Numerous biological databases and information sources are available on the web. Major databases are discussed here while other relevant ones are summarized in Table 1.

a. Nucleotide sequence databases

GenBank (http://www.ncbi.nlm.nih.gov/Genbank/GenbankSearch.html): GenBank is the nucleotide sequence database built and distributed by the National Center for Biotechnology Information (NCBI) at the National Institutes of Health. As on today GenBank contains more than 22 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 http://www.ncbi.nlm.nih.gov/Entrez/index.html, or a sequence query can be submitted through the **BLAST** family of programs (see http://www.ncbi.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** (**B**asic **L**ocal **A**lignment **S**earch **T**ool) 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 (http://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 accept data and issue 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.

b. 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 non-redundant, 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 to many other proteomics resources are available from the **ExPASy** (**E**xpert **P**rotein **A**nalysis **S**ystem) 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, SWISS-PROT, 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 (PIR-PSD) (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.

c. Genome Databases

Entrez Genome (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome): The whole genomes of over 800 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, malarial parasite) can be found on the Genomic Biology page under "organism-specifc 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://kegg.genome.ad.jp/): 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/e2k1/ath1/): 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://www.tigr.org/tigr-scripts/CMR2/CMRHomePage.spl): 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).

Computational Aspects of Pairwise Sequence Alignment

R. Senthilkumar and Arun Anirudhan

Sequence alignment is the procedure of comparing two (pair wise) or more (multiple) sequences by searching for a series of individual characters or character patterns that are in the same order in the sequences. Identical or similar characters are placed in the same column; non-identical characters can either be placed in the same column as a mismatch or opposite a gap in the other sequence. In an optimal alignment, gaps are placed to bring as many identical or similar characters as possible into vertical register. There are two types of sequence alignment, global and local. In global alignment, an attempt is made to align the entire sequence using as many characters as possible, upto both ends of each sequence. Sequences that are quite similar and approximately the same length are suitable candidates for global alignment. In local alignment, stretches of sequences with the high density of matches are aligned, thus generating one or more islands of matches or sub-alignments in the aligned sequences. Local alignments are more suitable for aligning sequences that are similar along some of their lengths but dissimilar in others. Sequences that differ in length or sequences that share a common domain are suitable candidates for this.

Fig 1: Global and Local alignment

Sequence alignments are useful for discovering functional, structural and evolutionary information in biological sequences.

Alignment of two sequences

1. Dot matrix analysis
2. Dynamic programming.
3. Word or k-tuple method (FASTA or BLAST)

Unless the sequences are known to be very much alike, the dotmatrix method is used first, because the method displays any possible sequence alignments as diagonals on the matrix. It can readily reveal any insertion/deletion. The major limitation of the method is that the most dot-matrix computer program doesn't show actual alignment.

The dynamic programming method, first used for global sequence alignment of sequences by Needleman & Wunch (1970) and for the local alignment by smith & Waterman (1981) provides one or more alignments of the sequences. An alignment is generated by starting at the ends of the two sequences and attempting to match all possible pairs of characters

between the sequences and following a scoring scheme for matches, mismatches and gaps. This procedure generates a matrix of numbers that represent all possible alignments between the two sequences. The highest set of sequential scores in the matrix defines an optimal alignment. For proteins, an amino acid substitution matrix, PAM250 or BLOSUM62 is used to score matches and mismatches.

The dynamic programming method is guaranteed in a mathematical sense to provide the optimal alignment for a given set of user defined variables, including choice of scoring matrix and gap penalties. The dynamic programming finds solutions to smaller pieces of the problem and then puts them all together to form a complete and optimal final solution. This method can be slow due to the large number of computational steps.

The word or k-tuple methods are used by FASTA and BLAST algorithms. They align two sequences quickly, first searching for identical short stretches of sequences (words) and then by joining these words into an alignment by the DP method. These methods are fast enough to be suitable for searching an entire database for the sequences that align best with and input test sequence. The FASTA and BLAST are heuristic, ie an empirical method of computer programming in which rules of thumb are used to find solutions and feedback is used to improve performances.

Dotmatrix Sequence Comparison

In the dotmatrix method of sequence comparison, one sequence A is listed across the top of the page and other sequence B is listed down the left side. Starting with the first character in B, one then moves across the page keeping in the first row and placing a dot in any column where the character in A is the same. This process is continued until the page is filled with dots representing all the possible matches of A characters with B characters. Any region of similar sequence is revealed by a diagonal row of dots. Isolated dots not on the diagonal represent random mismatches that are probably not related to any significant alignment. Detection of matching regions may be improved by filtering out random matches in dotmatrix. Dotmatrix analysis can also be used to find direct and inverted repeats within sequences. It can also reveal the presence of repeats of the same sequence character many times. The occurrence of such repeats of the same character increases the difficulty of aligning sequences because they create alignments of artificially high score.

Dynamic Programming Algorithm for Sequence Alignment

Dynamic programming is a computational method ie used to align two proteins or DNA sequences. This provides best or optimal alignments. Optimal alignments provide useful information to biologists concerning sequence relationships by giving the best information as to which character in a sequence should be in the same column in an alignment and which are insertions in one of the sequence (deletion in the other). This information is important for making functional, structural and evolutionary predictions on the basis of sequence alignments. Another feature of the dynamic programming algorithm is that the alignments obtained depend on the choice of a scoring system for comparing character pairs and penalty scores for gap.

In deciding to perform a sequence alignment, it is important the keep the goal of the analysis in mind. Is the investigator interested in trying to find out whether two proteins have similar domains or structural features or whether they share a common ancestor relationship? The

desired objective will influence the way the analysis is done. Global or local, type of scoring matrix and the value of gap penalties to be used? Dayhoff PAM matrices are based on an evolutionary model of protein change. BLOSUM matrices are designed to identify members of same family.

For sequences that are very similar, eg: - greater than 95%, the sequence alignment is quite obvious and a computer program may not even be needed to produce the alignment. As the sequences become less and less similar, the alignment becomes more difficult to produce and one is less confident of the result.

Alignment of two sequences without allowing gaps requires and algorithm that performs a number of comparisons roughly proportional to the square of the average sequence length, as in a dot matrix. If the alignment is to include gaps of any lengths at any position in their sequence, the number of comparisons that must be made becomes astronomical and it is not achievable by direct comparison methods. Dynamic programming is a method of sequence alignment that can take gaps into account but that requires a manageable number of comparisons.

Global alignment of two protein sequences by the Needleman-Wunsch algorithm with enhancements by Smith and Waterman

This example illustrates a global alignment of two hypothetical sequences, sequence 1 = MNALSDRT and sequence 2 = MGSDRTTET.

	GAP	M	N	A	L	S	D	R	T
GAP	0	-12	-16	-20	-24	-28	-32	-36	-40
M	-12								
G	-16								
S	-20								
D	-24								
R	-28								
T	-32								
T	-36								
E	-40								
T	-44								

A. Prepare a 10 x 11 matrix and place sequence 1 across the top of the matrix and sequence 2 down the left side. Leave an extra row and an extra column before each sequence labeled GAP to allow for gaps at the end of alignment. Fill in the extra row and column with the penalties for gaps of length zero to 8. The gap penalty used here is GAP = - 12- 4 (x - 1), where x is the length of the gap. -12 is the penalty for opening the gap in the alignment, and -4 is the penalty for each additional sequence character in the gap. The reason for choosing this particular penalty scheme is discussed below.

	GAP	M	N	A	L	S	D	R	T
GAP	0	-12	-16	-20	-24	-28	-32	-36	-40
M	-12	(6)	(-2)						
G	-16	(-3)	(0)						
S	-20								
D	-24								
R	-28								
T	-32								
T	-36								
E	-40								
T	-44								

B. Fill in the score for each amino acid pair in the matrix. Shown in parentheses are examples for the four possible matches between the first two amino acids. These scores are taken from the log odds form of the Dayhoff scoring matrix at 250 PAMs .

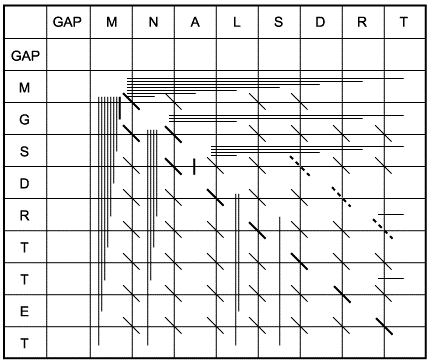
	GAP	M	N	A	L	S	D	R	T
GAP	0	-12	-16	-20	-24	-28	-32	-36	-40
M	-12	6 (6)	-6 (-2)						
G	-16	-6 (-3)	6 (0)						
S	-20								
D	-24								
R	-28								
T	-32								
T	-36								
E	-40								
T	-44								

C. Calculate the score in each of the above positions. The maximum score of the M/M position is the GAP/GAP score of 0 plus 6 for an M/M match, or 6. The arrow indicates the previous matrix position that was used to obtain a score of 6; i.e., the box labeled with a score of 0. Similarly, the maximum possible score in the N/M position is 6 - 12 (one gap penalty) = -6, that of the M/G position is 6 - 12 = -6, and that of the N/G position is 6 + 0 = 6 (no gap penalty). Note that each sequential row and column must be completed before moving to a lower row or more rightward column.

D. Complete the matrix by choosing at each position the maximum possible score (E). Keep track of all moves made to reach a maximum score at each position in a second matrix, the trace-back matrix (F).

	GAP	M	N	A	L	S	D	R	T
GAP	0	-12	-16	-20	-24	-28	-32	-36	-40
M	-12	6 (6)	-6 (-2)	-10	-14	-18	-22	-26	-30
G	-16	-6 (-3)	6 (0)	-5	-10	-13	-17	-22	-26
S	-20	-10	-5	7	-5	-8	-13	-17	-21
D	-24	-14	-8	-5	3	-5	-4	-14	-17
R	-28	-18	-14	-9	-8	3	-6	2	-10
T	-32	-22	-18	-13	-11	-7	3	-7	5
T	-36	-26	-22	-17	-15	-10	-7	2	-4
E	-40	-30	-25	-21	-20	-15	-7	-8	2
T	-44	-34	-30	-24	-23	-19	-15	-8	-5

E. The scoring matrix is completed to find the highest score at each matrix position. The process started in C has been continued to fill in the entire matrix with the maximum possible score at each matrix position. The right column and lowest row are then examined for the highest possible score because the alignment is a global one, meaning that the alignment will end only when the end of one of the sequences has been reached. Any remaining unmatched sequence will be opposite gaps. The highest-scoring box in the right-hand column and lowest row is a 5 in row 7. If end gaps were not being penalized, this would be the end of the search for the best score. However, if the alignment were to end here, there are three unmatched positions left in sequence 2, and each will be opposite a gap. Thus, an additional penalty score for three gaps (-20) corresponding to the heavy dotted line will have to be subtracted from 5, leaving an alignment score of 5 - 20 = -15. By subtracting any remaining end gap penalties from all positions in the last column and bottom row (not shown), one finds that the best score is actually -5 in the right-hand, lowest corner of the matrix obtained by a diagonal move to this position, giving a score of -8+3 = -5.



F. The trace-back matrix is used to find which characters align. This matrix shows all of the moves made by the algorithm from one matrix position to another to calculate the maximum score at each position. Because the highest-scoring position in the last row and column is the -5 in the rightmost, lower corner, this position is in the alignment. Thus, the last T in each sequence will be matched. The task is to find a path back through the matrix using the

moves made to get to that highest-scoring position and stopping at the beginning of one of the sequences. When the path turns from a previous diagonal move, a gap is placed opposite the next character in sequence 2 if the path turns upward and sequence 1 if it turns to the left. Two paths, shown as darker lines, are possible. These are also shown in part I above and correspond to the alignments 1 and 2 shown below. Note that if end gap penalties were not used, the path shown by the light dotted line in part I would be the correct alignment. This alignment, alignment 3, is also shown below. Note that wherever there are two paths leading to a matrix position, i.e., two possible ways of achieving that score, two alternative alignments will branch from that position. It is not too difficult to see that there are many possible paths through the scoring matrix that represent different alignments.

```
sequence 1  M - N A L S D R T
sequence 2  M G S D R T T E T
score        6 -12 1  0 -3 1  0 -1  3 = -5
```

Alignment 1. Although this alignment has a low and insignificant score of -5, it is the best-scoring alignment that can be made between these two short sequences with the Needleman-Wunsch algorithm with end gaps penalized. Note that the score of -5 is also found at the lowest position in the last column, corresponding to the alignment of the last characters in the sequences. Normally, it only makes sense to use a global alignment method for producing an alignment between sequences that are about the same length and that are expected to align along their entire lengths. The end gap penalty forces the ends to align. For sequences that are quite similar along their lengths, using end gap penalties will not have the dramatic effect that it does in this hypothetical example.

```
sequence 1  M N - A L S D R T
sequence 2  M G S D R T T E T
score        6 -12 1 0 -3 1 0 -1 3 = -5
```

Alignment 2. This second alignment is found by the trace-back procedure because there were two possible paths at one location in the matrix. This alignment scores slightly lower than the alignment 1 above. The difference is in the placement of a single gap opposite either a G or an S, and in the slightly higher score for the N/S versus the N/G alignment (1 vs. 0). This result illustrates that the dynamic programming alignment method may find more than one alignment having the same or almost the same score. Programs such as GCG GAP and BESTFIT can be set to provide several of these alternative alignments.

```
Sequence 1  M N A  L S D R T  -  -  -
Sequence 2   -  - M G S D R T T E T
Score        0  0 -1  -4 2 4  6  3 0  0  0 = 10
```

Alignment 3. (no end gap penalty included). On initial observation, this alignment has a great deal more appeal than the above two and has a much higher score. However, all of the gaps needed to make the alignment have been put on the ends, where they do not count. Leaving out end gaps also makes the analysis less rigorous mathematically (they appear in the rigorous proofs) and leaves doubt as to whether or not evolutionary conclusions may be drawn (Smith and Waterman 1981a,b). There is a positive effect of helping to identify the region of similarity SDRT without scoring end gaps, but using the Needleman-Wunsch

algorithm, this region could still be missed if this amino acid pattern was broken by higher-scoring neighboring alignments. The Smith-Waterman local alignment algorithm discussed below is specifically designed to find such conserved patterns of local alignment.

Local sequence alignment by the Smith-Waterman algorithm

	GAP	M	N	A	L	S	D	R	T
GAP	0	0	0	0	0	0	0	0	0
M	0	6	0	0	4	0	0	0	0
G	0	0	6	1	0	5	1	0	0
S	0	0	1	7	0	2	5	1	1
D	0	0	2	1	3	0	6	4	1
R	0	0	0	0	0	3	0	12	3
T	0	0	0	1	0	1	3	0	15
T	0	0	0	1	0	1	1	2	3
E	0	0	1	0	0	0	4	0	2
T	0	0	0	2	0	1	0	3	3

A. Scoring matrix for Smith-Waterman alignment of sequence 1, MNALSDRT, and sequence 2, MGSDRTTET. These same sequences, the PAM250 scoring matrix and gap penalty scores (-12 and -4 for gap opening and gap extension penalties, respectively) for internal and end gaps, were used. The major difference between this scoring matrix and the Needleman-Wunsch matrix is that there are no negative scores in the Smith-Waterman scoring matrix. The effect of this change is that an alignment can begin anywhere without receiving a negative penalty from a previously low- scoring alignment. Once an alignment has been built, it stops when negative alignment scores or the introduction of gaps reduces the following alignment scores to 0. Thus, only a portion of each sequence that was in this high-scoring region will be reported. Note that in this example the initial end gap penalty does not have any effect because all first row and column scores are 0, the minimum allowed by the Smith-Waterman algorithm. Because a gap penalty at the end of the alignment produces a score of zero, the end gap penalty similarly has no effect.

	GAP	M	N	A	L	S	D	R	T
GAP									
M		↖			↖				
G		↖	↖		↖	↖	↖	↖	
S		↖	↖	↖		↖		↖	
D			↖		↖		↖		
R				↖	↖	↖		↖	
T			↖	↖	↖	↖	↖	↖	
T		↖				↖		↖	
E			↖		↖		↖		
T				↖					

B. The trace-back matrix of the above Smith-Waterman scoring matrix. To find the optimal local alignment, the highest-scoring position in the scoring matrix is located (15), and the trace-back from this position is followed up to a zero in the matrix. The resulting sequence alignment is shown below. As opposed to the complex moves in the Needleman-Wunsch

matrix, which are designed to test many combinations of matches, mismatches, and gaps, only simple diagonal moves were made in the Smith-Waterman matrix. Thus, there is only one alignment starting from the highest position. However, many other lower-scoring alignments are apparent, such as the second highest-scoring alignment of MNA with MGS starting at the position that scores 7. Note that this second alignment does not include any of the same amino acids, and not even any of the same aligned amino acids, that were used in the first alignment. It is possible to have multiple local alignments that do use the same aligned amino acid pairs, as there was in the global alignment example given above, but there are no examples in this matrix. The distinctions regarding which alignments use the same aligned pairs, which use the same residues in a different alignment, and which use entirely different residues are described in Chapter 3.

Sequence 1	S D R T
Sequence 2	S D R T
Score	2 4 6 3 = 15

C. Sequence alignment determined by the above procedure. Note that the score of the alignment is the same as that shown by the highest-scoring position in the scoring matrix. The inclusion of any additional sequence would reduce the score below 15.

Scoring Matrices

Introduction

It is assumed that the sequences being sought have an evolutionary ancestral sequence in common with the query sequence. The best guess at the actual path of evolution is the path that requires the fewest evolutionary events. All substitutions are not equally likely and should be weighted to account for this. Insertions and deletions are less likely than substitutions and should be weighted to account for this. It is necessary to considerater that the choice of search algorithm influences the sensitivity and selectivity of the search. The choice of similarity matrix determines both the pattern and the extent of substitutions in the sequences the database search is most likely to discover.

There have been extensive studies looking at the frequencies in which amino acids substituted for each other during evolution. The studies involved carefully aligning all of the proteins in several families of proteins and then constructing phylogenetic trees for each family. Each phylogenetic tree can then be examined for the substitutions found on each branch. This can then be used to produce tables(scoring matrices) of the relative frequencies with which amino acids replace each other over a short evolutionary period. Thus a substitution matrix describes the likelihood that two residue types would mutate to each other in evolutionary time.

A substitution is more likely to occur between amino acids with similar biochemical properties. For example the hydrophobic amino acids Isoleucine(I) and valine(V) get a positive score on matrices adding weight to the likeliness that one will substitute for another. While the hydrophobic amino acid isoleucine has a negative score with the hydrophilic amino acid cystine(C) as the likeliness of this substitution occurring in the protein is far less. Thus matrices are used to estimate how well two residues of given types would match if they were aligned in a sequence alignment.

Importance of scoring matrices

- Scoring matrices appear in all analysis involving sequence comparison.
- The choice of matrix can strongly influence the outcome of the analysis.
- Scoring matrices implicitly represent a particular theory of evolution.
- Understanding theories underlying a given scoring matrix can aid in making proper choice.

Types of matrices

- PAM
- BLOSSUM
- DNA Identity Matrix

Differences between PAM and BLOSSUM

- PAM matrices are based on an explicit evolutionary model (that is, replacements are counted on the branches of a phylogenetic tree), whereas the Blosum matrices are based on an implicit rather than explicit model of evolution.
- The sequence variability in the alignments used to count replacements. The PAM matrices are based on mutations observed throughout a global alignment, this includes both highly conserved and highly mutable regions. The Blosum matrices are based only on highly conserved regions in series of alignments forbidden to contain gaps.
- The method used to count the replacements is different, unlike the PAM matrix, the Blosum procedure uses groups of sequences within which not all mutations are counted the same.

Equivalent PAM and Blossum matrices

The following matrices are roughly equivalent...

- PAM100 ==> Blosum90
- PAM120 ==> Blosum80
- PAM160 ==> Blosum60
- PAM200 ==> Blosum52
- PAM250 ==> Blosum45

The **Blosum matrices** are best for detecting local alignments, The **Blosum62 matrix** is the best for detecting the majority of weak protein similarities and the **Blosum45 matrix** is the best for detecting long and weak alignments.

PAM (Point Accepted Mutation) matrix

Amino acid scoring matrices are traditionally PAM (Point Accepted Mutation) matrices which refer to various degrees of sensitivity depending on the evolutionary distance between sequence pairs. In this manner PAM40 is most sensitive for sequences 40 PAMs apart. PAM250 is for more distantly related sequences and is considered a good general matrix for protein database searching. For nucleotide sequence searching a simpler approach is used which either convert a PAM40 matrix into match/mismatch values which takes into consideration that a purine may be replaced by a purine and a pyrimidine by a pyrimidine.

e.g. The PAM 250 matrix

This is appropriate for searching for alignments of sequence that have diverged by 250 PAMs, 250 mutations per 100 amino acids of sequence. Because of back mutations and silent mutations this corresponds to sequences that are about 20 percent identical.

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

In this example Isoleucine (I) is likely to be substituted by valine (V) and gets a score of 4. Isoleucine (I) is unlikely to be substituted for Cystine and gets a score of -2.

BLOSSUM (Blocks Substitution Matrix)
The BLOSUM matrices, also used for protein database search scoring (the default in blastp), are divided into statistical significance degrees which, in a way, are reminiscent of PAM distances. For example, BLOSUM64 is roughly equivalent to PAM 120. BLOSSUM Blocks Substitution Matrix). BLOSSUM matrices are most sensitive for local alignment of related sequences. The BLOSUM matrices are therefore ideal when tying to identify an unknown nucleotide sequence.

e.g. Blosum 45 Matrix This is derived from sequence blocks clustered at the 45% identity level.

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

W	-2	-3	-4	-3	-4	-3	-2	-2	-2	-4	-3	-2	-2	-3	-2	-2	1	3	15
C	-3	-4	-3	-3	-2	-3	-3	-3	-3	-1	-1	-1	-2	-1	-3	-2	-2	-3	-5	12
	G	P	D	E	N	H	Q	K	R	S	T	A	M	V	I	L	F	Y	W	C

DNA Identity Matrix (Unitary Matrix)

Here a you only get a positive score for a match, and a score of -10000 for a mismatch. As such a high penalty is given for a mismatch, no substitutuion should be allowed, although a gap may be permitted.

	A	T	G	C
A	1			
T	-10000	1		
G	-10000	-10000	1	
C	-10000	-10000	-10000	1

References

1. Mount, D.W (2002) Bioinformatics: Sequence and Genome Analysis P No: 51-138.
2. http://www.ebi.ac.uk/index.html
3. http://www.bioinformaticsonline.org

Sequence Similarity Searching using BLAST and FASTA

R. Senthilkumar and P.K. Parvathi

Sequence similarity searching is an important methodology in computational molecular biology. Initial clues to understanding the structure or function of a molecular sequence arise from homologies to other molecules that have been previously studied. Genome database searches reveal biologically significant sequence relationships and suggest future investigation strategies. Sequence alignments provide a powerful way to compare novel sequences with previously characterized genes. Both functional and evolutionary information can be inferred from well-designed queries and alignments. BLAST (Basic Local Alignment Search Tool) and FASTA provides a method for rapid searching of nucleotide and protein databases.

BLAST

The BLAST algorithm was written balancing speed and increased sensitivity for distant sequence relationships. Instead of relying on global alignments (commonly seen in multiple sequence alignment programs) BLAST emphasizes regions of local alignment to detect relationships among sequences which share only isolated regions of similarity. Therefore, BLAST is more than a tool to view sequences aligned with each other or to find homology, but a program to locate regions of sequence similarity with a view to comparing structure and function.

BLAST search can be performed:

1. Locally using local databases.
2. Remotely at NCBI using databases at NCBI (daily updated) via a Web browser. (www.ncbi.nlm.nih.gov/blast)

Selecting the BLAST Program

The BLAST search pages allow you to select from several different programs.
Below is a table of these programs.

If sequence is NUCLEOTIDE

Length	Database	Purpose	BLAST Program
20 bp or longer	Nucleotide	Identify the query sequence	MEGABLAST (accept batch queries) Standard BLAST (blastn)
		Find sequences similar to query sequence	Standard BLAST (blastn)
		Find similar proteins to translated query in a translated database	Translated BLAST (tblastx)
	Protein	Find similar proteins to translated query in a protein database	Translated BLAST (blastx)
7 - 20 bp	Nucleotide	Find primer binding sites or map short contiguous motifs	Search for short, nearly exact matches

If sequence is PROTEIN

Length	Database	Purpose	BLAST program
15 residues or longer	Protein	Identify the query sequence or find protein sequences similar to query	Standard Protein BLAST (blastp)
		Find members of a protein family or build a custom position-specific score matrix	PSI-BLAST
		Find proteins similar to the query around a given pattern	PHI-BLAST
	Conserved Domains	Find conserved domains in the query	CD-search (RPS-BLAST)
	Conserved Domains	Find conserved domains in the query and identify other proteins with similar domain architectures	Conserved Domain Architecture Retrieval Tool (CDART)
	Nucleotide	Find similar proteins in a translated nucleotide database	Translated BLAST (tblastn)
5-15 residues	Protein	Search for peptide motifs	Search for short, nearly exact matches

Selecting the BLAST Database

We can select several NCBI databases to compare with query sequences. Some databases are specific to proteins or nucleotides and cannot be used in combination with certain BLAST programs

Nucleotides

Database	Description
Nr	All non-redundant GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS, GSS, or HTGS sequences).
Month	All new or revised GenBank+EMBL+DDBJ+PDB sequences released in the last 30 days.
Dbest	Non-redundant database of GenBank+EMBL+DDBJ EST Divisions.
Dbsts	Non-redundant database of GenBank+EMBL+DDBJ STS Divisions.
Mouse ests	The non-redundant Database of GenBank+EMBL+DDBJ EST Divisions limited to the organism mouse.
human ests	The Non-redundant Database of GenBank+EMBL+DDBJ EST Divisions limited to the organism human.
other ests	The non-redundant database of GenBank+EMBL+DDBJ EST Divisions all organisms except mouse and human.
Yeast	Yeast (*Saccharomyces cerevisiae*) genomic nucleotide sequences. Not a collection of all Yeast nucelotides sequences, but the sequence fragments from the Yeast complete genome.
E. coli	*E. Coli* (*Escherichia coli*) genomic nucleotide sequences.
Pdb	Sequences derived from the 3-dimensional structure of proteins.

Kabat [kabatnuc]	Kabat's database of sequences of immunological interest. For more information http://immuno.bme.nwu.edu/
Patents	Nucleotide sequences derived from the Patent division of GenBank.
Vector	Vector subset of GenBank(R), NCBI, (ftp://ncbi.nlm.nih.gov/pub/blast/db/ directory).
Mito	Database of mitochondrial sequences
Alu	Select Alu repeats from REPBASE, suitable for masking Alu repeats from query sequences. It is available at ftp://ncbi.nlm.nih.gov/pub/jmc/alu. (Eukaryotic Promotor Database ISREC in Epalinges s/Lausanne (Switzerland).)
Gss	Genome Survey Sequence, includes single-pass genomic data, exon-trapped sequences, and Alu PCR sequences.
Htgs	High Throughput Genomic Sequences.

Proteins

Database	Description
Nr	All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF
Month	All new or revised GenBank CDS translation+PDB+SwissProt+PIR released in the last 30 days.
Swissprot	The last major release of the SWISS-PROT protein sequence database These are uploaded to our system when they are received from EMBL.
Patents	Protein sequences derived from the Patent division of GenBank.
Yeast	Yeast (Saccharomyces cerevisiae) protein sequences. This database is not to be confused with a listing of all Yeast protein sequences. It is a database of the protein translations of the Yeast complete genome.
E. coli	E. coli (Escherichia coli) genomic CDS translations.
Pdb	Sequences derived from the 3-dimensional structure Brookhaven Protein Data Bank.
kabat [kabatpro]	Kabat's database of sequences of immunological interest. For more information http://immuno.bme.nwu.edu/
Alu	Translations of select Alu repeats from REPBASE, suitable for masking Alu repeats from query sequences. It is available at ftp://ncbi.nlm.nih.gov/pub/jmc/alu.

Query Sequence: The BLAST web pages accept input sequences in three formats; FASTA sequence format, NCBI Accession numbers, or GIs.

Set Subsequence: A region of the query sequences can be searched by entering the range in nucleotides or protein residues in the "Form" and "To" boxes provided under "Set Subsequence".

Limit by Entrez Query

BLAST searches can be limited to the results of an Entrez query against the database chosen. This can be used to limit searches to subsets of the BLAST databases. Any terms can be entered that would normally be allowed in an Entrez search session.

For example:

Protease NOT hiv1 [Organism]

This will limit a BLAST search to all proteases, except those in HIV 1. This can also be used to limit searches to a particular molecule type:

biomol_mrna [PROP] AND brain

To limit to a specific organism you can either select using the pulldown menu, form a list of the most common organism in the databases. Or enter the name of the organism in the

Entrez Query field with the [Organism] qualifier. For example:

Mus musculus[Organism]

Filter (Low-complexity)

Mask off segments of the query sequence that have low compositional complexity, as determined by the SEG program of Wootton & Federhen or, for BLASTN, by the DUST program of Tatusov and Lipman. Filtering can eliminate statistically significant but biologically uninteresting reports from the blast output (e.g., hits against common acidic-, basic- or proline-rich regions), leaving the more biologically interesting regions of the query sequence available for specific matching against database sequences.

Filtering is only applied to the query sequence (or its translation products), not to database sequences. Default filtering is DUST for BLASTN, SEG for other programs.

Filter (Human repeats)

This option masks Human repeats (LINE's and SINE's) and is especially useful for human sequences that may contain these repeats. Filtering for repeats can increase the speed of a search especially with very long sequences (>100 kb) and against databases which contain large number of repeats (htgs).

Mask Lower Case

With this option selected you can cut and paste a FASTA sequence in upper case characters and denote areas you would like filtered with lower case. This allows you to customize what is filtered from the sequence during the comparison to the BLAST databases

Expect

The statistical significance threshold for reporting matches against database sequences; the default value is 10, meaning that 10 matches are expected to be found merely by chance, according to the stochastic model of Karlin and Altschul. If the statistical significance ascribed to a match is greater than the EXPECT threshold, the match will not be reported. Lower EXPECT thresholds are more stringent, leading to fewer chance matches being reported. Increasing the threshold shows less stringent matches. Fractional values are acceptable.

Gapped BLAST and PSI-BLAST

The Gapped BLAST algorithm allows gaps (deletions and insertions) to be introduced into the alignments that are returned. Allowing gaps means that similar regions are not broken into several segments. The scoring of these gapped alignments tends to reflect biological relationships more closely.

Position-Specific Iterated BLAST (PSI-BLAST) provides an automated, easy-to-use version of a "profile" search, which is a sensitive way to look for sequence homologues. The program first performs a gapped BLAST database search. The PSI-BLAST program uses the information from any significant alignments returned to construct a position-specific score matrix, which replaces the query sequence for the next round of database searching. PSI-BLAST may be iterated until no new significant alignments are found. At this time PSI-BLAST may be used only for comparing protein queries with protein databases.

PHI-BLAST

PHI-BLAST is a search program that combines matching of regular expressions with local alignments surrounding the match. The most important features of the program have been incorporated into the BLAST software framework.

If you had a protein sequence S and a regular expression pattern P occurring in S, PHI-BLAST helps answer the question: What other protein sequences both contain an occurrence of P and are homologous to S in the vicinity of the pattern occurrences?

PHI-BLAST may be preferable to just searching for pattern occurrences because it filters out those cases where the pattern occurrence is probably random and not indicative of homology.

PHI-BLAST may be preferable to other flavors of BLAST because it is faster and because it allows the user to express a rigid pattern occurrence requirement.

MegaBLAST

Megablast is optimised for aligning sequences that differ slightly as a result of sequencing or other similar errors. It is up to 10 times faster than more common sequence similarity programs and can be used to compare two large sets of sequences against each other.

FASTA

Originally called FAST-P, for Fast Protein search tool, but were modified to use DNA sequences as well, and became FASTA for FAST(ALL). Using a single sequence as input or query sequence, FASTA searches a database of sequences for the best pairwise alignment between 2 sequences. The FASTA Family does many types of searches and contains programs that will do local homology searching, assess statistical significance, and produce global alignments and others.

The FASTA family of programs

The FASTA search pages allow us to select from several different programs. Below is a table of these programs.

Program	Function
fasta3	scan a protein or DNA sequence library for similar sequences
fastx/y3	compare a DNA sequence to a protein sequence database, comparing the translated DNA sequence in forward and reverse frames.
Tfastx/y3	compares a protein to a translated DNA data bank
fasts3	compares linked peptides to a protein databank
fastf3	compares mixed peptides to a protein databank

FASTA search can be performed

1. Locally using local databases
2. Remotely at various sites using remotely maintained databses and programs via a web browser (http://www.ebi.ac.uk/fasta33/)

Databases

Below is the list of databases available for FASTA search .We can restrict our search to a particular database. Multiple databases selection is possible by clicking on them.

Abbreviation	Database Name
swall	SWALL Non-Redundant Protein sequence database Swissprot+Trembl+TremblNew
swissprot	Swiss-Prot Protein Database
Swnew	Updates to Swiss-Prot
sptrembl	SPTREMBL (TrEMBL)
remtrembl	REMTREMBL (uncurated entries in TrEMBL)
Prints	FingerPrints
IPI	International Protein Index
EURO patents	European patents database.
JAP patents	Japanese patents database. .
USPTO patents	American patents database.
SGT	Structural Genomic Targets Database
PDB	Protein Database of Brookhaven
ENSEMBL	ENSEMBL CDS PEP

EMBL	The EMBL Database (All divisions)
EFUN	EMBL Fungi
EINV	EMBL Invertebrates
EHUM	EMBL Human
EMAM	EMBL Mammalian
EORG	EMBL Organelles
EPHG	EMBL Phages
EPLN	EMBL Plants
EPRO	EMBL Prokaryote
EROD	EMBL Rodents
EMUS	EMBL Mouse
ESTS	EMBL STSs
ESYN	EMBL Synthetic
EUNA	EMBL Unclassified
EVRL	EMBL Viral
EVRT	EMBL Vertebrates
EEST	EMBL ESTs
EGSS	EMBL Genome Survey Sequences
EHTG	EMBL High Throughput Genome Sequences
EMNEW	EMBL New (Updates)
EMALL	EMBL (All divisions) + EMBL New (All Updates)
IMGT	IMGT Immunogenetics Database
HGBASE	European SNP database

Query Sequence

Query sequence can be cut and pasted or typed into the large text window. A free text (raw) sequence is simply a block of characters representing a DNA/RNA or Protein sequence. The sequence can also be in GCG, FASTA, EMBL, GenBank, PIR, NBRF, Phylip or Swiss-Prot format. Partially formatted sequences will not be accepted.

Upload A File

A file containing a valid sequence in any format can be uploaded from the computer using this option.

Results
This option allows to choose between email and interactive runs. If the search is interactive the results will be delivered to the browser window when they are ready or results will be delivered by email.

Matrix
This option is to set which comparison matrix should be used when searching the database. The default matrix for blast is blosum62

Gap Penalties

Gapopen: Penalty for the first residue in a gap (-12 by default for fasta with proteins, -16 for DNA).
Gapext: Penalty for additional residues in a gap (-2 by default for fasta with proteins, -4 for DNA).

Histogram
Setting this option to "yes" will display the search histogram of the expected frequency of chance occurrence of the database matches found. It provides you with a way of quickly checking to see if your statistical estimates are as you might expect. The histogram presents observed and expected distribution of E values.

Alignments
Setting this options to any number available in the menu allows you to set the maximum number of reported alignments in the output file.

Ktup
Change this value to limit the word-length the the search should use. A word-length of 2 is sensitive enough for most protein database searches. The thumb rule is that the larger the word-length the less sensitive, but faster the search will be. For DNA searches a ktup of 6 is the default.

Strand
This option lets you choose which DNA strand to search with when you are using a DNA sequence to compare against the DNA databanks. The 'default' is to search the 'both' strands. 'top' means the sequence will be searched as it is input into the form. 'bottom' means: reverse and complement your input sequence.

Expectation value upper limit

Here you may set the expectation value upper limit for score and alignment display. Generally, in evaluating the E() scores, the following rules of thumb can be used, sequences with E() less than 0.01 are almost always found to be homologous, sequences with E() between 1 and 10 frequently turn out to be related as well. *The defaults are 10.0 for FASTA with protein searches, 5.0 for translated DNA/protein comparisons, and 2.0 for DNA/DNA searches.*

Expectation value lower limit
Expectation value lower limit for score and alignment display. A value of 1e-6 prevents library sequences with E()- values lower than 1e-6 from being displayed. This allows the use

to focus on more distant relationships. Thus with this option if set will filter out the best matches and allow more distant relationships to be displayed.

The default setting for this is zero.

Sequence range
This options allows the user to denote which region within the query sequence should be searched.

Database sequence range to search
This option is similar to the above except that it sets the sequence range to search within the database. If the user wishes to search all entries with no more than 300 aa/nt he/she must type **1-300** in the text window. It is also possible to use ranges such as 1000-3000 which indicates sequences with at least 1000 aa/nt and up-to, but no more than 3000 aa/nt.

Similarities Between BLAST and FASTA

- Blast and Fasta both search databases of sequence information Databases must be specially formatted for Blast searching, and these can be used for searching using Fasta as well.
- They both provide statistics that allow us to assess whether we could expect a particular matching sequence to be found by random chance or not.
  In general, blast searches will be faster, but fasta may produce better alignments.

Differences between BLAST and FASTA

- Fasta and blast use different scoring matrices and gap penalties by default . (BLOSUM50/ -12 for the first gap and -2 for each additional gap for fasta, BLOSUM62, 12 for the first gap and -1 for each additional gap for blast2) Blastp will show alignments between several domains in the same sequences. Fasta shows one alignment (Smith/Waterman) for each sequence pair.
- Blast searches using a default word size of 11 for nucleotides, and 3 for peptides. If you are searching with a nucleic acid that does not code for protein, you are advised to shorten this word size to 7, or to use fasta instead.
- BLAST is not able to search with very short sequences, (e.g. primer sequences).but fasta can search it by setting a ktup value of 1.
- By default fasta searches one strand of DNA, not both.
- By default, blast "filters" input sequences for low complexity or repeat regions. Fasta does not do this.
- Blastx does 6 independent searches (one for each translation frame) while fastx3 and fasty3 do a single forward (or backward with the -i flag) search which allows frameshifts in computing the similarity score and alignments. Thus the latter are more sensitive and can produce much better alignments than blastx when using poor quality sequence. (This is similarly true for tblastn versus tfastx3 and tfasty3)
- There is an option using native fasta, to give a lower-bound expectation value cutoff that prevents very closely related sequences from being shown

Reference

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

2. Altschul, S.F., Boguski, M.S., Gish, W. & Wootton, J.C. (1994) "Issues in searching molecular sequence databases." Nature Genet. 6:119-129
3. Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.
4. http://www.ncbi.nlm.nih.gov/BLAST/tutorial/
5. http://www.ebi.ac.uk/fasta33/index.html

Multiple Sequence Alignment

R. Senthilkumar

For many genes a database search will reveal a whole number of homologous sequences. The evolution and the sequence conservation in such a group can reasonably be achieved by the sequence pair wise comparison methods. Pairwise comparisons do not readily show positions that are conserved among a whole set of sequences and tend to miss subtle similarities that become visible when observed simultaneously among many sequences. A multiple sequence alignment (MSA) arranges a set of sequences in a scheme where positions believed to be homologous are written in a common column. Like in a pairwise alignment, when a sequence does not possess an amino acid in a particular position this is again denoted by a dash.

MSA gives biologist the ability to extract biologically important but perhaps widely dispersed sequence similarities that can give biologist hints about the evolutionary history of certain sequences. In pairwise alignment, when two sequences align, it is concluded that there is probably a functional relationship between the two sequences. Whereas for MSA, if it is known that there is a functional similarity amongst a number of sequences, we can use MSA to find out where the similarity comes from.

The MSA, homology search algorithm is sometimes called a "many-against-each-other" search because the input is a small, defined set of sequences which are compared only against each other, not against an entire database. This is in contrast to the BLAST homology search algorithm, a "one-against-all" homology search, in which the input is a single sequence that is compared against all other known sequences listed in the database. Thus the starting point for an MSA is a set of sequences that are already presumed to be homologous.

Multiple Sequence Alignment algorithms:
MSA is very important for finding similar domains in a set of sequences and for doing phylogenetic analysis. Following are the algorithms commonly used for MSA.

1. The Sum of Pairs Method: The sum of pairs method provides a way to optimize the MSA by maximizing the number of matches (or by minimizing the cost or number of mismatches) summed over all columns in the MSA. The program MSA uses this method. The scores between sequences may be weighted in order to reduce the bias of having alike sequences in the alignment.

The Sum of Pairs (SP) method is as follows:

Given (1) a set of N aligned sequences each of length L in the form of a L x N MSA matrix M and (2) a substitution matrix (PAM or BLOSUM) that gives a cost c(x,y) for aligning two characters x, y .
The SP score SP (m_i) for the ith column of M denoted by m_i is calculated using the formula

$$SP(m_i) = SUM_{(I =< j < k =<)} [\, c(m_i^j, m_i^k) \,]$$

Where m_i^j is the jth entry in the ith column

Then the SP score for M is SUM$(1 =< I =< L)$ [SP (m_i)] ie the SP score for the MSA matrix M is calculated by first adding up all possible pairwise alignment scores for the ith column and then summing up the scores for all the columns in M.

Example

Consider aligning the following 4 protein sequences

S1 = AQPILLLV
S2 = ALRLL
S3 = AKILLL
S4 = CPPVLILV

Next consider the following MSA matrix M

A Q P I L L L V
A L R - L L - -
A K - I L L L -
C P P V L I L V

Assume c (match) = 1, c (mismatch) = -1, and c (gap) = -2, also assume c (-, -) = 0 to prevent the double counting of gaps.

Then the SP score for the 4th column of M would be

SP (m4) = SP (I, -, I, V)
= c(I,-) + c(I,I) + c(I,V) + c(-,I) + c(-, V) + c(I,V)
= -2 + 1 + (-1) + (-2) + (-2) +(-1)
= -7

To find SP (M) we would find the score of each m_i and then SUM all the SP (m_i) scores to get the score M.

To find the optimal score using this method we need to consider all possible MSA matrices. As before we assume that we have N sequence each of length L.
Consider the case where N = L = 2, such that S1 = AT, S2 = CG, then we have the following possible matrices

(we assume no column can contain all gaps)

A T A - T A T - A T - A - T A T - - - A T - A T
C G C G - C - G - C G - C G - - C G C G - C - G
- - A T = $(2 + 1)^2$ matrices
C G - -

In general for a MSA of N sequences each of length L, with the assumption that no column can contain all gaps, there are $(L + 1)^N$ possible MSA matrices.

Problems

Consider column 1 of our example ie A, A, A, C for this column we get SP (m4) = SP (A, A, A, C)

= 1 + 1 + (-1) + 1 + (-1) + (-1)
= 0

where as if we had A,A,A,A we get a score of SP(A,A,A,A) = 1+1+1+1+1+1 = 6 , thus we get a difference of 6 for what could be explained by a single mutation. The SP method tends to overweight the influence of mutations

For most real cases involving long sequences and/or a large number of sequences, the SP method is too time consuming

2.The STAR Alignment Method

In star alignment method first using a pairwise alignment method (DP,etc) find the sequence that is most similar to all the other sequences. Then using this "best" sequence as the center (of a star, hence the name) align the other sequences following the once a gap always a gap rule.

For example consider the following set of sequences

S1 A T T G C C A T T
S2 A T G G C C A T T
S3 A T C C A A T T T T
S4 A T C T T C T T
S5 A C T G A C C

Following is the similarity matrix for the pairwise comparing of the sequences.

	S1	S2	S3	S4	S5	SUM sim(Si, Sj) I/=J
S1	-	7	-2	0	-3	2
S2	7	-	-2	0	-4	1
S3	-2	-2	-	0	-7	-11
S4	0	0	0	-	-3	-3
S5	-3	-4	-7	-3	-	-17

For this example S1 is the center of the STAR

The best alignment between S1 and the other sequences as follows:

S1 | A T T G C C A T T S1 | A T T G C C A T T
S2 | A T G G C C A T T S5 | A C T G A C C - -

S1 | A T T G C C A T T - -

S3 | A T C - C A A T T T T

S1 | A T T G C C A T T
S4 | A T C T T C - T T

Build the MSA, start with S1 & S2 as
A T T G C C A T T
A T G G C C A T T adding S3 using once a gap always a gap

A T T G C C A T T - -
A T G G C C A T T - -
A T C - C A A T T T T continuing in this fashion we obtain for our MSA of all the sequences

A T T G C C A T T - -
A T G G C C A T T - -
A T C - C A A T T T T
A T C T T C - T T - -
A C T G A C C - - - -

The time complexity is dominated by computing the pairwise alignment which again for N sequences we have O (N^2) pairs. We consider each pairwise alignment to take L^2 time where again L is the length of each sequence.

In practice neither the SP nor the STAR method are used much. Instead a method related to the STAR method called CLUSTALW is usually used.

3.The CLUSTALW Method

CLUSTALW is a progressive method for MSA .It starts with using a pairwise method to determine the most related sequences and then progressively adding less related sequences or groups of sequences to the initial alignment.

CLUSTAL comes in 3 versions

CLUSTAL - gives equal weight to all sequences.
CLUSTALW - has the ability to give different weights to the sequences and other program parameters.
CLUSTALX - provides a GUI (Graphics User Interface) to CLUSTAL.

The CLUSTALW Algorithm

Step 1: Determine all pairwise alignment between sequences and determine degrees of similarity between each pair.
Step 2: Construct a similarity tree.
Step 3: Combine the alignments starting from the most closely related groups to the most distantly related groups, as in STAR.

Step 1

Use a pairwise alignment (implementation dependent) method to compute pairwise alignment amongst the sequences.

Using the pairwise alignments compute a "distance" between all pairs of sequences. A method commonly used is as follows:

For each pairwise alignment look at the non-gapped positions and count the number of mismatches between the two sequences ,then divide this value by the number of non-gapped pairs to calculate the distance, for example

```
N K L - O N                    distance = 1/4 = .25
- M L N O N
```

After computing the "distance" between all pairs of sequences we put them in to a matrix. For example if we consider a set of 7 sequences we could have the following matrix:

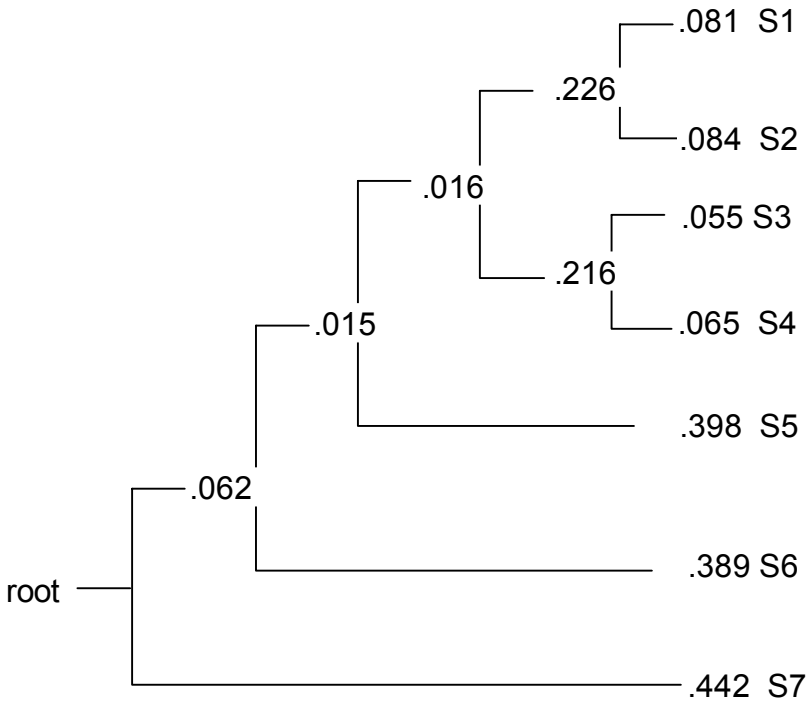
Seq.	S1	S2	S3	S4	S5	S6	S7
S1	-						
S2	.17	-					
S3	.59	.60	-	(SYMMETRIC)			
S4	.59	.59	.13	-			
S5	.77	.77	.75	.75	-		
S6	.81	.82	.73	.74	.80	-	
S7	.87	.86	.86	.88	.93	.90	-

Step 2

Construct a similarity tree. The CLUSTALW package uses a technique called the Neighbor Joining method to construct the similarity tree. For the similarity matrix given above we could construct the following similarity tree.

Combine the alignments starting from the most closely related groups to the most distantly related groups by going from tip of tree to the root of the tree. In our example we first align S1 with S2 (grp1) then S3 with S4 (grp2), then align grp1 with grp2, we continue until we reach root of the tree.

To align groups of sequences we build a similarity matrix using a somewhat complicated technique that considers the average of all the pairwise scores from the groups we want to align. We then use DP on this matrix to get an alignment between the two groups.



Step 3

To assign a weight to a sequence use the values on similarity tree from the sequence to the root of the tree. If two or more sequences share a branch, which may indicate an evolutionary relationship, its value is split amongst the sequences. This reduces the weight for related sequences. S7 would be given a weight of .442 whereas S3 weight would be

Wt (S3) = .055 + .216/2 + .016/4 + .015/5 +. 062/6 = .194
.194 would be used to adjust the value of S3 on the tree.

Multiple sequence alignment using clustal W

ClustalW is a general purpose multiple sequence alignment program for DNA or proteins available at www.ebi.ac.uk/clustalw .It produces biologically meaningful multiple sequence alignments of divergent sequences. It calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen. Evolutionary relationships can be seen via viewing Cladograms or Phylograms. Trees can also be calculated from multiple alignments. The program has some adjustable parameters with reasonable defaults.

Adjustable parameters in clustalw:

Search title**:** We can type any text as title of the alignment.
Alignment: It is possible to run a full alignment using a stringent algorithm for generating the tree guide or a fast algorithm.
Output: The result of the multiple sequence alignment can obtained in ALN, GCG, PHYLIP, PIR and GDE formats.

Outorder: To decide which order the sequences should be printed in the alignment either in input or alignment order.

Colour: To display the alignment in colours. This option only works with ALN or GCG output format. The colouring of residues takes place according to the following physiochemical criteria:

AVFPMILW	RED	Small (small+ hydrophobic (incl.aromatic -Y))
DE	BLUE	Acidic
RHK	MAGENTA	Basic
STYHCNGQ	GREEN	Hydroxyl + Amine + Basic – Q
Others	Grey	

Consensus symbols: An alignment with display by default the following symbols denoting the degree of conservation observed in each column:
"*" means that the residues or nucleotides in that column are identical in all sequences in the alignment.
":" means that conserved substitutions have been observed, according to the COLOUR table above.
"." means that semi-conserved substitutions are observed.

Fast pairwise alignment options
Ktup: This option allows to choose which 'word-length' to use when calculating fast pairwise alignments.
Window: This option is to set the window length when calculating fast pairwise alignments.
Score: This option allows to decide which score to take into account when calculating a fast pairwise alignment.
Topdiag: To Select the top diagonals should be integrated when calculating a fast pairwise alignment.
Pairgap: To set the gap penalty when generating fast pairwise alignments.

Multiple sequence alignment options
Matrix : This option is to choose which matrix series to use when generating the multiple sequence alignment. The program goes through the chosen matrix series, spanning the full range of amino acid distances.
BLOSUM (Henikoff). These matrices appear to be the best available for carrying out data base similarity (homology searches). The matrices used are: Blosum80, 62, 40 and 30.

PAM (Dayhoff). These have been extremely widely used since the late '70s. The matrices used are: PAM 120, 160, 250 and 350 matrices.

GONNET. These matrices were derived using almost the same procedure as the Dayhoff one (above) but are much more up to date and are based on a far larger data set. They appear to be more sensitive than the Dayhoff series. The matrices used are: GONNET 40, 80, 120, 160, 250 and 350 matrices.

Default values are:

dna: iub
protein blosum
Gapopen: To set the penalty for opening a gap. The default value is 10.
Endgap: To set the penalty for closing a gap.
Gapext: To set the penalty for extending a gap. The default value is 0.05.
Gapdist : To set the gap separation penalty. The default value is 8.
Phylogenetic tree
Phylogram is a branching diagram (tree) assumed to be an estimate of a phylogeny, branch lengths are proportional to the amount of inferred evolutionary change. A Cladogram is a branching diagram (tree) assumed to be an estimate of a phylogeny where the branches are of equal length, thus cladograms show common ancestry, but do not indicate the amount of evolutionary "time" separating taxa. Tree distances can be shown, just click on the diagram to get a menu of options.

Other Multiple alignment tools:
Multalign: It is also a multiple sequence alignment tool. The approach is based on the conventional dynamic-programming method of pairwise alignment. Initially, a hierarchical clustering of the sequences is performed using the matrix of the pairwise alignment scores. The closest sequences are aligned creating groups of aligned sequences. Then close groups are aligned until all sequences are aligned in one group. The pairwise alignments included in the multiple alignment form a new matrix that is used to produce a hierarchical clustering. If it is different from the first one, iteration of the process can be performed.

PileUp
PileUp creates a multiple sequence alignment using a simplification of the progressive alignment method of Feng and Doolittle.The method used is similar to the method described by Higgins and Sharp. The multiple alignment procedure begins with the pairwise alignment of the two most similar sequences, producing a *cluster* of two aligned sequences. This cluster can then be aligned to the next most related sequence or cluster of aligned sequences. Two clusters of sequences can be aligned by a simple extension of the pairwise alignment of two individual sequences. The final alignment is achieved by a series of progressive, pairwise alignments that include increasingly dissimilar sequences and clusters, until all sequences have been included in the final pairwise alignment.

Before alignment, the sequences are first clustered by similarity to produce a *dendrogram*, or tree representation of clustering relationships. It is this dendrogram that directs the order of the subsequent pairwise alignments. PileUp can plot this dendrogram so that it can display the order of the pairwise alignments that created the final alignment.

Resources for Multiple Sequence Alignment:
1. www.ebi.ac.uk/clustalw
2. http://www2.igh.cnrs.fr/bin/align-guess.cgi
3. http://barton.ebi.ac.uk/servers/amas_server.html
4. http://cbrg.inf.ethz.ch/Server/MultAlign.html

5. http://www.ch.embnet.org/software/TCoffee.html

Molecular Phylogeny

R. Senthilkumar

Introduction

The diversity of life is truly amazing, but all living organisms do share certain similarities. All living organisms can replicate, and the replicator molecule is DNA.

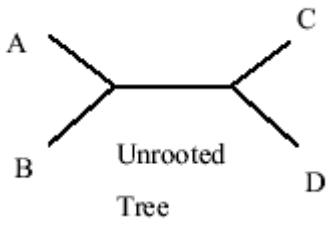
Species have evolved through speciation**,** which refers to the process through which one species becomes divided into more than one new species. The history of speciation tells how these species are related to each other; if they evolved to be separate species relatively recently, then they are close relatives; if they evolved to be separate species long ago, then they are distant relatives. So the pattern of evolutionary relationships, or history of speciation, among species is called as phylogeny. Species evolve as ancestral species split into more than one new species, it is convenient to represent phylogeny as a tree in which lines represent species and places where lines split into two represent points where ancestral species speciated into two new species.

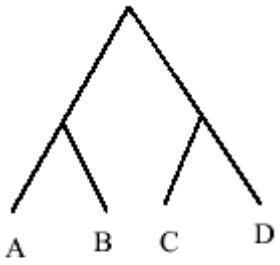
Phylogeny tree

Phylogeny trees are graphs that are composed of nodes and branches. The node represent taxonomic units and branch lines represent relationship between them. The taxonomic units represented by the nodes may be species, organisms (whole genome or ribosomal RNA), strains (closely related microbes), individual genes (or gene families), repetitive DNA sequences, metabolic pathways, secondary structures, any discrete character(s),populations, or microbial communities. In phylogeny tree there are internal and external nodes. External nodes represent present day taxonomic unit and internal nodes represents ancestral taxonomic units. External nodes are the data used to construct the tree called Operational Taxonomic Units (OTU) where internal nodes are inferred from the relationship among the OTUs.

Rooted and unrooted tree

The phylogeny tree can be represented as rooted or unrooted trees. In a rooted tree there exists one node from which a unique path leads to any other node, with the direction of the path corresponding to the evolutionary process. In an unrooted tree, only the relationship among the OTUs are specified, not the evolutionary (time) relationship. The number of tree topologies of rooted tree is much higher than that of the unrooted tree for the same number of OTUs. Therefore, the error of the unrooted tree topology is smaller than that of the rooted tree. An unrooted tree may be converted to a rooted tree by identifying an out group.





Rooted
Tree

Two means to root an unrooted tree:

- The outgroup method :include in the analysis a group of sequences known a priori to be external to the group under study; the root is by necessity on the branch joining the outgroup to other sequences.
- Make the molecular clock hypothesis: All lineages are supposed to have evolved with the same speed since divergence from their common ancestor.The root is at the equidistant point from all tree leaves.

The number of trees that are possible rises exponentially with the number of OTUs. For two OTUs there is only one tree is possible. For three OTUs there are three trees possible. For 10 OTUs 34,459,425 rooted trees are possible.

Sequences used for phylogenitic analysis:

- DNA-Very sensitive, non-uniform mutation rates
- cDNA/RNA-Useful for more remote homologies
- Protein Sequences-Useful for most remote homologies, deep phylogenies, more uniform mutation rates, more character states

Steps to construct phylogeny tree

Get DNA/RNA/Protein Sequences

↓

Construct multiple sequence alignment

↓

Compute phylogeny
(example :pairwise distances for distance methods)

↓

Build tree: topology + branch lengths

↓

Estimate reliability

↓

Visualize

Distance Matrix methods

Distance matrix methods fit a tree to a matrix of pairwise distances between the OTUs. For example if the OTUs are sequences, we may perform the pairwise or multiple sequence alignment to calculate a similarity score between each pair of sequences. The inverse of the similarity is taken to be the distance. The distances between a pair of sequences may also be obtained by straightforward one –on –one comparisons.
For example, if K is distance,

$$K=N/L$$

Where N is the total number of differences (substitutions) between the two sequences and L is the length of the sequence.

Jukes and cantor formula for distance is

$$K= -(3/4)\ln(1-4p/3) \text{ where } p=N/L$$

Unweighed Pair-Group Method with Arithmetic Mean (UPGMA)

UPGMA is the simplest method for tree reconstrctions and was originally developed for constructing taxonomic phenograms, i.e., trees that reflect the phenotypic similarities between OTUs. It can also be used to construct phylogenetic trees if the rates of evolution are approximately constant among lineages. However, it should be emphasized that this method is NOT reliable if the rates of evolution vary among lineages.

The UPGMA method employs a sequential clustering algorithm in which a phylogenetic tree is built in a stepwise manner. That is, the method first identifies the two OTUs that are most similar to each other (i.e. have the shortest distance) and treat them as a new single OTU. Such an OTU is referred to as a **composite OTU**. Subsequently, from among the new group of OTUs, the method identifies the pair with the highest similarity, and so on, until only two OTUs are left.

For example, let us consider the case of four OTUs (A, B, C, D). The pairwise evolutionary distances, such as Jukes and Cantors estimates are given by the following matrix.

	A	B	C	D
A	0	D(AB)	D(AC)	D(AD)
B	D(BA)	0	D(BC)	D(BD)
C	D(CA)	D(CB)	0	D(CD)
D	D(DA)	D(DB)	D(DC)	0

In this matrix dij stands for the distance between OTUs j and i. The first two OTUs to be clustered are the ones with the SMALLEST distance. Let us assume that dAB has the smallest value. Then, OTUs A and B are the first to be clustered.

After the first clustering, A and B are treated as a single composite OTU, and a NEW distance matrix is computed**.** In this matrix, d(AB)C = (dAC + dBC)/2 and d(AB)D = (dAD + dBD)/2. The distance between a simple OTU and a composite OTU is the average of the distances between the simple OTU (C or D) and the two OTUs in the composite OTU (A and B).

	AB	C	D
AB	0	D(C(AB))	D(D(AB))
C	D(C(AB))	0	D(CD)
D	D(D(AB))	D(DC)	0

If D(AB)C for example turns out to be the smallest distance in the new matrix, then OTU C will be joined to the composite OTU (AB).

A, B and C are treated as a single composite OTU, and a NEW distance matrix is computed.

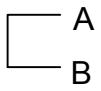
	ABC	D
ABC	0	D(D(ABC))
D	D(D(ABC))	0

In this matrix, d(ABC)D = (dAD + dBD + dCD)/3. That is to say, the distance between the simple OTU and the composite OTU is the average of the distances between the simple OTU (D) and the three OTUs in the compostie OTU (ABC).

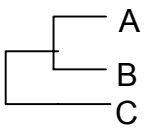
The final step consists of clustering the last OTU, D, with the composite OTU ABC.

In general, the distance between two composite OTUs is computed as the arithmetic mean of the pairwise distances between the participating simple OTUs in the two composite OTUs. For example, the distance between the composite OTUs (ij) and (mn)is : d(ij)(mn) = (dim + din + djm + djn)/4

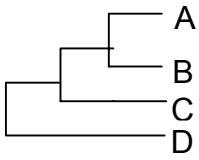
The procedure above is equivalent to the following construction of the tree.



Step 1



Step2



Step3

Neighbour joining method

The principle of the neighbour-joining method is to find neighbours sequentially that may minimize the total length of the tree. The method starts with a starlike tree in which there is no clustering of OTUs. The first step is to separate a pair of OTUs (e.g. 1 ad 2) from all the

other OTUs. In this tree there is only one internal branch which connects the nodes X and Y, where X is the common node for OTUs 1 and 2, and Y is the common node for all other OTUs.

(1,2)-**nodeX**-internal branch-**nodeY-**(3,4,5 etc).

Any pair of OTUs can take the positions 1 and 2 in the tree and there are N(N-1)/2 possible trees of that format. Among these possible trees, the one that gives the smallest sum of branch lengths is chosen. Lets for example assume that when OTUs A and B are placed at position 1,2 the sum of branch lenghs are smaller than for any other OTUs at that position. This pair of OTUs (AB) is then regarded as a composite OTU and the arithmetic mean distances between OTUs are computed to form a new distance matrix. The next pair of OTUs that gives the smallest sum of branch lengths is then chosen. This procedure is then repeated until all OTUs have been clustered and the final topology has been obtained.

By successively identifying nearest neighbors, this method takes a series of shortcuts and the number of possible trees compared for a given set of sequences is much smaller than the total number of possible trees. As a result, the method is fast and efficient. Because it also works reasonably well, as inferred from simulation studies, it has become one of the most favored distance matrix methods.

Maximum parsimony method

The principle of maximum parsimony searches for a tree that requires the smallest number of evolutionary changes to explain the differences observed among the OTUs under study. Such a tree is called a maximum parsimony tree. Often more than one tree with the same minimum number of changes are found, so that no unique tree can be inferred. A key term used in maximum parsimony is **informative sites**. A nucleotide site is informative only if it favors some trees over others. Consider the following hypothetical sequences:

1.AAGGA
2.AGGGG
3.AGACA
4. AGACG

There are three possible unrooted trees for four sequences:

Tree 1: (1-2)(3-4), Tree 2: (1-3)(2-4) and Tree 3: (1-4)(2-3)

The first nucleotide site provides no indication about which one of the three possible trees may be correct since all sequences have A at this position. This site is non informative.

The second nucleotide site is also not informative, since sequence 1 has an A at this position while all other sequences have G.

In contrast, the third nucleotide site is informative. The third site suggests that 1-2 and 3-4 should be clustered since 1 and 2 have G at this position, while 3 and 4 have A at this position. Nucleotide position 3 supports tree 1.

The fourth nucleotide site is also informative. This site also suggests that 1-2 and 3-4 should be clustered since 1 and 2 have G at this position, while 3 and 4 have C at this position. Nucleotide position 4 supports tree 1.

The fifth nucleotide site is also informative. However, this site suggests that 1-3 and 2-4 should be clustered since 1 and 3 have A at this position and 2 and 4 have G at this position. Nucleotide position 5 supports tree 2.

Thus, in the case of four OTUs, and informative site favors only one of the three possible alternative trees. It is easy to see that the tree supported by the largest number of informative sites is the maximum parsimony tree (in the example given above tree 1 would be the maximum parsimony tree). When the number of OTUs under study is larger than four, the situation becomes more complicated because there are many more possible trees to consider and because inferring the number of substitutions for each alternative tree becomes more tedious. However, the basic principle remains simple - it is to infer the minimum number of substitutions required for a given tree.

The procedure for inferring a maximum parsimony tree can be summarized as follows:

Identify all the informative sites.

↓

for each possible tree, infer the minimum number of substitutions at each informative site

↓

sum up the numbers over all informative sites,

↓

choose the tree or trees with the smallest number of substitutions at each informative site

↓

sum up the numbers over all informative sites

If there is more than one maximum parsimony tree, no unique tree is inferred. Note that the procedure neglects all substitutions at noninformative sites.

Maximum Likelihood Method

This method also uses **each position** in an alignment, evaluates all possible trees, and calculates the **likelihood for each tree** using an explicit **model of evolution**. The evolutionary model could simply mean that one assumes that changes between all nucleotides (aminoacids) are equally probable. The program will then assign all possible nucleotides to the internal nodes of the tree in turn and calculate the probability that each

such sequence would have generated data (if two sister taxa have the nucleotide A, the reconstruction that assumes derivation from a C would be assigned a low probability compared with a derivation that assumes there already was an A). The probabilities for all possible reconstruction are summed up to yield the likelihood for one particular site. The likelihood's for each aligned position are then multiplied to provide a likelihood for each tree. The tree with the maximum likelihood is the most probable tree. The parameters to be considered are the **branch lengths** for each topology, and the likelihood is maximized to estimate branch lengths. This is the slowest method of all but seems to give the best result and the most information about the tree.

Phylogeny tools

Numerous tools are now available each having their own useful features. PHYLIP, PAUP, PUZZLE and FastDNAml are common among them.

PHYLIP

PHYLIP, the Phylogeny Inference Package is, is a package of programs available as source code in C and can be run on several platforms including UNIX and Macintosh machines. Programs can be used to perform parsimony, distance and maximum likelihood analyses of both DNA and protein datasets. It can also compute consensus trees, compute distances between trees, draw trees, resample data sets by bootstrapping or jackknifing, edit trees, and compute distance matrices. It can handle data that are nucleotide sequences, protein sequences, gene frequencies, restriction sites, restriction fragments, distances, discrete characters, and continuous characters.

It is available free, from its web site: http://evolution.gs.washington.edu/phylip.html

Table :The List of Phylip programs

Program	Description
clique	Finds largest clique of mutually compatible characters.
consense	Computes majority-rule and strict consensus trees.
Contml	Phylogenies from maximum likelihood on gene frequency data.
Contrast	Computes contrasts and correlations for comparative method studies.
Dnacomp	Phylogenies from DNA by the Compatibility criterion.
dnadist	Computes four distance measures from nucleic acid sequences.
Dnainvar	Phylogenetic invariants method.
Dnaml	Phylogenies from DNA by maximum likelihood.
Dnamlk	Phylogenies from DNA by maximum likelihood with a molecular clock.
Dnamove	Interactive DNA phylogenies by parsimony or compatibility.
dniapars	Nucleotide phylogenies by parsimony methods.
dnapenny	DNA phylogenies by branch-and-bound.

dollop	Phylogenies from the Dollo or polymorphism parsimony criteria.
Dolmove	Interactive construction of phylogenies using the Dollo or polymorphism parsimony criteria.
Dolpenny	Branch-and-bound phylogenies from the Dollo or polymorphism parsimony criteria.
drawgram	Draws cladograms and phenograms.
drawtree	Draws unrooted phylogenies.
Factor	Recodes multistate data to two state data.
fitch	Phylogenies by the Fitch-Margoliash and least-squares methods.
Gendist	Computes three genetic distances from gene frequency data.
kitsch	Phylogenies by the Fitch-Margoliash and least-squares methods with a molecular clock.
mix	Wagner, Camin-Sokal and mixed parsimony methods.
Move	Interactive construction of phylogenies from two-state data.
neighbor	Neighbour-joining and UPGMA methods.
Penny	Finds all most parsimonious trees by branch-and-bound.
Protdist	Compute distance measures from protein sequences.
Protpars	Phylogenies by protein parsimony.
Restml	Phylogenies by maximum likelihood from restriction enzyme data.
Retree	Reroot, flip branches and change species names and branch lengths.
Seqboot	Bootstrap, jack-knife or permutation resampling analyses.

PAUP

PAUP stands for Phylogenetic Analysis Using Parsimony is one the most sophisticated parsimony program available. It includes as many functions in a single platform independent program with menu interface. It is used to infer evolutionary trees from sequence and other data using maximum parsimony, maximum likelihood, and distance methods.

FastDNAml

FastDNAml is a program for estimating maximum likelihood phylogenetic trees from nucleotide sequences.It is a derived version and optimized program DNAML from **package PHYLIP .** From a whole of nucleic sequences, fastDNAml builds a phylogeny according to the method of , Maximum Likelihood. An analysis **bootstrap** can be associated with the treatment.

TREE-PUZZLE

TREE-PUZZLE is a program to reconstruct phylogenetic trees from molecular sequence data (DNA or amino acid) by maximum likelihood. It implements a fast tree search algorithm called quartet puzzling that allows analysis of large data sets and automatically

assigns estimations of support to each internal branch. TREE-PUZZLE also computes pairwise maximum likelihood distances as well as branch lengths for user specified trees.

RRTree

RRTree (Relative-Rate Tests with a tree) is a program to compare substitution rates between biological sequences grouped in phylogenetically defined lineages. It presents a choice of distance methods for coding or non-coding DNA or protein sequences. The input file contains a set of aligned sequences in a standard format and output is to the screen or written to a text file.

References

1. Molecular Evolution. W-H.Li (1997),Sinauer Associates Inc.,USA
2. Mount, D.W. Bioinformatics: Sequence and Genome Analysis, Cold Spring Harbor Laboratory Press, 2000.
3. Andreas D .Baxevanis,B.F.Francis Ouellette Bioinformatics :A Practical guide to the analysis of Genes and proteins Wiley- Interscience (2001).
4. Olsen, G. J., H. Matsuda, R. Hagstrom, and R. Overbeek**.** 1994. FastDnaML: a tool for construction of phylogenetic trees of Dna sequences using maximum likelihood. Comput Appl **Biosci. 10:41-8.**

PERL In Bioinformatics

Arun Anirudhan

Perl is a Programming Language Written by Larry Wall in late 80's to process mail on Unix systems and since extended by a huge cast of characters. The name is said to stand for:

1. Pathologically Eclectic Rubbish Lister
2. Practical Extraction and Report Language

Perl Properties

1. Interpreted Language
2. "Object-Oriented"
3. Cross-platform
4. Forgiving
5. Great for text
6. Extensible, rich set of libraries
7. Popular for web pages
8. Extremely popular for bioinformatics

Mechanics of Writing Perl Scripts

Some hints to help you get going.

Creating the Script

A Perl script is just a text file. Use any text (programmer's) editor.

By convention, Perl script files end with the extension *.pl*.

The Emacs text editor has a *Perl mode* that will auto-format your Perl scripts and highlight keywords. Perl mode will be activated automatically if you end the script name with .pl. Otherwise, you can force Emacs to enter Perl mode by placing this line somewhere near the top of the file:

```
# -*- mode: perl -*-
```

The next time you open the file, Emacs will enter Perl mode.

Running the Script

Option 1

Run the perl program from the command line, giving it the name of the script file to run.

```
(~) 50% perl time.pl
The time is now Thu Sep 16 18:09:28 1999
```

Option 2

Put the magic comment *#!/usr/bin/perl* at the top of the script.

```
#!/usr/bin/perl
```

```
# file: time.pl
$time = localtime;
print "The time is now $time\n";
```

Make the script executable with *chmod +x time.pl*:

```
(~) 51% chmod +x time.pl
```

Run the script as if it were a command:

```
(~) 52% time.pl
The time is now Thu Sep 16 18:12:13 1999
```

Useful Perl Command-Line Options

You can call Perl with a few command-line options to help catch errors:

-c
: Perform a syntax check, but don't run.

-w
: Turn on verbose warnings.

-d
: Turn on the Perl debugger.

Usually you will invoke these from the command-line, as in *perl -cw time.pl* (syntax check *time.pl* with verbose warnings). You can also put them in the top line: *#!/usr/bin/perl -w*.

Perl Statements

A Perl script consists of a series of *statements* and *comments*. Each statement is a command that is recognized by the Perl interpreter and executed. Statements are terminated by the semicolon character (;). They are also usually separated by a newline character to enhance readability.

A *comment* begins with the # sign and can appear anywhere. Everything from the # to the end of the line is ignored by the Perl interpreter. Commonly used for human-readable notes.

Some Statements

```
$sum = 2 + 2; # this is a statement

$f = <STDIN>; $g = $f++;  # these are two statements

$g = $f
  /
  $sum;       # this is one statement, spread across 3 lines
```

The Perl interpreter will start at the top of the script and execute all the statements, in order from top to bottom, until it reaches the end of the script. This execution order can be modified by loops and control structures.

Blocks

It is common to group statements into *blocks* using curly braces. You can execute the entire block conditionally, or turn it into a *subroutine* that can be called from many different places. Example blocks:

```
{  # block starts
  my $EcoRI = 'GAATTC';
  my $sequence = <STDIN>;
  print "Sequence contains an EcoRI site" if $sequence=~/$EcoRI/;
}  # block ends

my $sequence2 = <STDIN>;
if (length($sequence) < 100) {  # another block starts
  print "Sequence is too small. Throw it back\n";
  exit 0;
} # and ends

foreach $sequence (@sequences) {  # another block
  print "sequence length = ",length($sequence),"\n";
}
```

Literals

Literals are constant values that you embed directly in the program code. Perl supports both *string literals* and *numeric literals*.

String Literals

String literals are enclosed by single quotes (') or double quotes (''):

```
'The quality of mercy is not strained.';  # a single-quoted string
"The quality of mercy is not strained.";  # a double-quoted string
```

The difference between single and double-quoted strings is that variables and certain special escape codes are interpolated into double quoted strings, but not in single-quoted ones. Here are some escape codes:

\n	New line
\t	Tab
\r	Carriage return
\f	Form feed
\a	Ring bell
\040	Octal character (octal 040 is the space character)
\0x2a	Hexadecimal character (hex 2A is the "*" character)
\cA	Control character (This is the ^A character)
\u	Uppercase next character
\l	Lowercase next character
\U	Uppercase everything until \E
\L	Lowercase everything until \E
\Q	Quote non-word characters until \E
\E	End \U, \L or \Q operation

```
"Here goes\n\tnothing!";
  # evaluates to:
  # Here goes
  #       nothing!
```

```
'Here goes\n\tnothing!';
  # evaluates to:
  # Here goes\n\tnothing!

"Here goes \unothing!";
  # evaluates to:
  # Here goes Nothing!

"Here \Ugoes nothing\E";
  # evaluates to:
  # Here GOES NOTHING!

"Alert! \a\a\a";
  # evaluates to:
  # Alert! (ding! ding! ding!)
```

Putting backslashes in strings is a problem because they get interpreted as escape sequences. To inclue a literal backslash in a string, double it:

```
"My file is in C:\\Program  Files\\Accessories\\wordpad.exe";

  # evaluates to: C:\Program Files\Accessories\wordpad.exe
```

Put a backslash in front of a quote character in order to make the quote character part of the string:

```
"She cried \"Oh dear! The parakeet has flown the coop!\"";

  # evaluates to: She cried "Oh dear! The parakeet has flown the coop!"
```

Numeric Literals

You can refer to numeric values using integers, floating point numbers, scientific notation, hexadecimal notation, and octal. With some help from the Math::Complex module, you can refer to complex numbers as well:

```
123;       # an integer

1.23;      # a floating point number

-1.23;     # a negative floating point number

1_000_000; # you can use _ to improve readability

1.23E45;   # scientific notation

0x7b;      # hexadecimal notation  (decimal 123)
```

```
0173;      # octal notation (decimal 123)

use Math::Complex;  # bring in the Math::Complex module

12+3*i;    # complex number 12 + 3i
```

Backtick Strings

You can also enclose a string in backtics (`). This has the unusual property of executing whatever is inside the string as a Unix system command, and returning its output:

```
`ls -l`;
# evaluates to a string containing the output of running the
# ls -l command
```

Lists

The last type of literal that Perl recognizes is the *list*, which is multiple values strung together using the comma operator (,) and enclosed by parentheses. Lists are closely related to *arrays*, which we talk about later.

```
('one', 'two', 'three', 1, 2, 3, 4.2);
  # this is 7-member list contains a mixure of strings, integers
  # and floats
```

Operators

Perl has numerous *operators* (over 50 of them!) that perform operations on string and numberic values. Some operators will be familiar from algebra (like "+", to add two numbers together), while others are more esoteric (like the "." string concatenation operator).

Numeric & String Operators

The "." operator acts on strings. The "!" operator acts on strings and numbers. The rest act on numbers.

Operator	Description	Example	Result
.	String concatenate	'Teddy' 'Bear'	'TeddyBear
=	Assignment	$a = 'Teddy'	$a variable contains 'Teddy'
+	Addition	3+2	5
-	Subtraction	3-2	1
-	Negation	-2	-2
!	Not	!1	0
*	Multiplication	3*2	6
/	Division	3/2	1.5
%	Modulus	3%2	1
**	Exponentiation	3**2	9
<FILEHANDLE>	File input	<STDIN>	Read a line of input from standard input
>>	Right bit shift	3>>2	0 (binary 11>>2=00)
<<	Left bit shift	3<<2	12 (binary 11<<2=1100)
|	Bitwise OR	3|2	3 (binary 11|10=11

&	Bitwise AND	3&2	2 (binary 11&10=10
^	Bitwise XOR	3^2	1 (binary 11^10=01

Operator Precedence

When you have an expression that contains several operators, they are evaluated in an order determined by their *precedence*. The precedence of the mathematical operators follows the rules of arithmetic. Others follow a precedence that usually does what you think they should do. If uncertain, use parentheses to force precedence:

```
2+3*4;    # evaluates to 14, multiplication has precedence over addition
(2+3)*4;  # evaluates to 20, parentheses force the precedence
```

Logical Operators

These operators compare strings or numbers, returning TRUE or FALSE:

Numeric Comparison		String Comparison	
3 == 2	equal to	'Teddy' eq 'Bear'	equal to
3 != 2	not equal to	'Teddy' ne 'Bear'	not equal to
3 < 2	less than	'Teddy' lt 'Bear'	less than
3 > 2	greater than	'Teddy' gt 'Bear'	greater than
3 <= 2	less or equal	'Teddy' le 'Bear'	less than or equal
3 >= 2	greater than or equal	'Teddy' ge 'Bear'	greater than or equal
3 <=> 2	compare	'Teddy' cmp 'Bear'	compare
		'Teddy' =~ /Bear/	pattern match

The <=> and cmp operators return:
-1 if the left side is less than the right side
0 if the left side equals the right side
+1 if the left side is greater than the right side

File Operators

Perl has special *file operators* that can be used to query the file system. These operators generally return TRUE or FALSE.

Example:

```
print "Is a directory!\n" if -d '/usr/home';
print "File exists!\n" if -e '/usr/home/lstein/test.txt';
print "File is plain text!\n" if -T '/usr/home/lstein/test.txt';
```

There are many of these operators. Here are some of the most useful ones:

-e filename	file exists
-r filename	file is readable
-w filename	file is writable
-x filename	file is executable
-z filename	file has zero size
-s filename	file has nonzero size (returns size)
-d filename	file is a directory
-T filename	file is a text file
-B filename	file is a binary file

-M filename age of file in days since script launched
-A filename same for access time

Functions

In addition to its operators, Perl has many *functions*. Functions have a human-readable name, such as print and take one or more arguments passed as a list. A function may return no value, a single value (AKA "scalar"), or a list (AKA "array"). You can enclose the argument list in parentheses, or leave the parentheses off.

A few examples:

```
  # The function is print.  Its argument is a string.
  # The effect is to print the string to the terminal.
print "The rain in Spain falls mainly on the plain.\n";

  # Same thing, with parentheses.
print("The rain in Spain falls mainly on the plain.\n");

  # You can pass a list to print.  It will print each argument.
  # This prints out "The rain in Spain falls 6 times in the plain."
print "The rain in Spain falls ",2*4-2," times in the plain.\n";

  # Same thing, but with parentheses.
print ("The rain in Spain falls ",2*4-2," times in the plain.\n");

  # The length function calculates the length of a string,
  # yielding 45.
length "The rain in Spain falls mainly on the plain.\n";

  # The split function splits a string based on a delimiter pattern
  # yielding the list ('The','rain in Spain','falls mainly','on the plain.')
split '/','The/rain in Spain/falls mainly/on the plain.';
```

Variables

A variable is a symbolic placeholder for a value, a lot like the variables in algebra. Perl has several built-in variable types:

Scalars: $variable_name
A single-valued variable, always preceded by a $ sign.
Arrays: @array_name
A multi-valued variable indexed by integer, preceded by an @ sign.
Hashes: %hash_name
A multi-valued variable indexed by string, preceded by a % sign.
Filehandle: FILEHANDLE_NAME

A file to read and/or write from. Filehandles have no special prefix, but are usually written in all uppercase.

Scalar Variables

Scalar variables have names beginning with $. The name must begin with a letter or underscore, and can contain as many letters, numbers or underscores as you like. These are all valid scalars:

$foo
$The_Big_Bad_Wolf
$R2D2
$_____A23
$Once_Upon_a_Midnight_Dreary_While_I_Pondered_Weak_and_Weary

You assign values to a scalar variable using the = operator (not to be confused with ==, which is numeric comparison). You read from scalar variables by using them wherever a value would go.

A scalar variable can contain strings, floating point numbers, integers, and more esoteric things. You don't have to predeclare scalars. A scalar that once held a string can be reused to hold a number, and vice-versa:

Processing Command Line Arguments

When a Perl script is run, its command-line arguments (if any) are stored in an automatic array called @ARGV. You'll learn how to manipulate this array later. For now, just know that you can call the shift function repeatedly from the main part of the script to retrieve the command line arguments one by one.

Printing the Command Line Argument

Code:

```
#!/usr/bin/perl
# file: echo.pl

$argument = shift;
print "The first argument was $argument.\n";
```

Output:

```
(~) 50% chmod +x echo.pl
(~) 51% echo.pl tuna
The first argument was tuna.
(~) 52% echo.pl tuna fish
The first argument was tuna.
(~) 53% echo.pl 'tuna fish'
The first argument was tuna fish.
(~) 53% echo.pl
```

Basic I/O

I/O means "Input/Output". It's how your program communicates with the world.

Output

The print() function does it all:

Code:

```
#!/usr/bin/perl
# file: print.pl
```

```
$sidekick = 100;
print "Maxwell Smart's sidekick is ",$sidekick-1,".\n";
print "If she had a twin, her twin might be called ",2*($sidekick-1),".\n";
```

Output:

```
(~) 50% chmod +x print.pl
(~) 51% print.pl
Maxwell Smart's sidekick is 99.
If she had a twin, her twin might be called 198.
```

Input

The <> operator does input. It reads a line of input from the terminal. At the point that <> appears, the script will stop and wait for the user to type of line of input. Then <> will copy the input line into a variable.

```
#!/usr/bin/perl
# file: dog_years.pl

print "Enter your age: ";
$age = <>;
print "Your age in dog years is ",$age/7,"\n";
```

Output:

```
(~) 50% dog_years.pl
Enter your age: 42
Your age in dog years is 6
```

We will learn later how to take input from a file rather than the terminal.

Problems

Modify the "add" script from yesterday so that it checks that both arguments are defined. This will allow negative numbers and 0:

```
% add 2
Please provide two arguments.

% add 2 -5
-3
```

Modify the script again so that it checks that both arguments are positive integers. Zero is allowed, but -1 is not:

```
% add 2 -5
Please provide two positive integers.
```

Rewrite the script to use -w and use strict.

Write a script to read a line of DNA from the keyboard and print "yes!" if it contains an EcoRI site (GAATTC or gaattc).

Write a script to compare two strings given on the command line arguments and print "right order" if they are in alphabetic order, and "wrong order" if they are not:

```
% order Fred Lucy
right order

% order Lucy Fred
wrong order
```

Write a script to compare two strings given on the command line and print them out in correct alphabetic order:

```
% reorder Fred Lucy
Fred Lucy

% reorder Lucy Fred
Fred Lucy
```

Write a script named "same.pl" to read two strings from the terminal. Compare them in a case-sensitive manner and print "same" if they are the same, "different" if they are different:

```
% same.pl
Enter string 1: lucy
Enter string 2: Lucy
different
```

Modify this script to compare the strings in a case-INsensitive manner (hint, use the "lc" or "uc" functions change to upper or lowercase.
Modify this script to use the \U and/or \L string escapes.
Write a script named "percent" to calculate percentages, where the percentage is $a/($a+$b) * 100. Make sure that the script does not crash when given two numbers that add up to zero:

```
% percent.pl 50 150
25%

% percent.pl 50 -50
You are trying to trick me! at line 4.
```

Modify this script to use the printf() function to produce nicely formatted floating point numbers (hint: after checking the Perl books, try "man sprintf" and "man printf" to learn about this wonderful function).

```
% percent.pl 50 150
25.00 %
```

Write a program to check the command-line argument for a valid american social security number (999-99-9999). Print "yes" if it matches:

```
% ssno.pl 058-28-1282
yes

% ssno.pl "bananas are yummy"
no
```

Modify the previous program to read the social security number from the keyboard.

Write a program named "pali.pl" to detect palindromes. It must be able to handle changes in case.

```
% pali.pl "Madam in Eden Im Adam"
yes!

% pali.pl ggaattcc
yes!

% pali.pl "cold spring harbor laboratory"
no!
```

Modify the program to work even if there is extraneous punctuation:

```
% pali.pl "A man, a plan, a canal... Panama!
yes!
```

(Hint: Look up the s/// pattern matching & substitution function in the Perl reference guide. We will cover this formally in a few days.)

Numeric Comparisons

Use == to Compare Two Numbers for Equality

```
$a =  4 == 4;      # TRUE
$a =  4 == 2 + 2;  # TRUE
$a =  4 == $b;     # depends on what $b is
```

Do not confuse == with =

== is for numeric comparison.

= is for assignment.

Use != to Compare Two numbers for Non-Equality

```
$a =  4 != 4;      # FALSE
$a =  4 != 2 + 2;  # FALSE
$a =  4 != $b;     # depends on what $b is
```

Use > and < for "Greater than", "Less than"

```
$a =  4 > 3;      # TRUE
$a =  4 < 3;      # FALSE
$a =  4 > $b;     # depends on what $b is
```

Use >= and <= for "Greater than or Equal", "Less than or Equal"

```
$a =  4 >= 3;      # TRUE
$a =  4 >= 4;      # TRUE
$a =  4 <= $b;     # depends on what $b is
```

Use <=> to Compare Two Strings

$result = $a <=> $b
$result is
-1 if the left side is less than the right side
0 if the left side equals the right side
+1 if the left side is greater than the right side
NB: <=> is really useful in conjunction with the sort() function.

String Comparisons
Use eq to Compare Two Strings for Equality

```
$a =  'fred' eq 'fred';      # TRUE
$a =  'fred and lucy' eq 'fred' . ' and ' . 'lucy';  # TRUE
$a =  'fred' eq $b;          # depends on what $b is
```

Do not confuse == with eq
== is for numeric comparison.
eq is for string comparison.

```
$a =  'fred' == 'lucy';      # WRONG WRONG WRONG!
```

Use ne to Compare Two Strings for Non-Equality

```
$a =  'fred' ne 'fred';      # FALSE
$a =  'fred' ne 'lucy';      # TRUE
$a =  'fred' eq $b;          # depends on what $b is
```

Use gt, lt, ge, ne for "Greater than", "Less than", "Greater or Equal" etc.
String comparison is in ASCII alphabetic order.

```
$a =  'fred' gt 'lucy';  # FALSE
$a =  'fred' lt 'lucy';  # TRUE
$a =  'Lucy' lt 'lucy';  # TRUE
$a =  'Lucy' lt 'fred';  # TRUE !!
```

In ASCII alphabetic order, the set of capital letters is less than the set of lowercase letters.

Use cmp to Compare Two Strings

$result = $a cmp $b
$result is
-1 if the left side is less than the right side
0 if the left side equals the right side
+1 if the left side is greater than the right side

NB: *cmp* is really useful in the sort() function.

Truth and the Comparison Operations
If a comparison operation is true, it returns 1.
If a comparison operation is false, it returns undefined.

```
$a =  4 == 1+3;
print "The answer is $a","\n";
```
The answer is 1.

Conditional Statements
The operators if and unless can be used to execute a statement conditionally.

```
print "a equals b\n"   if $a eq $b;
print "a is twice b\n" if $a*2 == $b;

print "a does not equal b\n"  unless $a eq $b;
```

If-Else Statements
Use else blocks for either/or constructions.

```
if ($a == $b) {
  print "a equals b\n";
  $a += $b;
} else {
  print "a does not equal b\n";
  die "Operation aborted!";
}
```

Logical Operators
To combine comparisons, use the and, or and not logical operators. In some scripts, you might see their cryptic cousins, &&, || and !:

Lower precedence	Higher precedence	Description
$a and $b	$a && $b	TRUE if $a AND $b are TRUE
$a or $b	$a || $b	TRUE if either $a OR $b are TRUE
not $a	!$b	TRUE if $a is FALSE

```
if ($a < 100 and $a > 0) {
  print "a is the right size\n";
} else {
  die "out of bounds error, operation aborted!";
}

if ($a < 100 && $a > 0) {
  print "a is the right size\n";
} else {
  die "out of bounds error, operation aborted!";
}
```

```
if ($a >= 100 or $a <= 0) {
  die "out of bounds error, operation aborted!";
}

if ($a >= 100 || $a <= 0) {
  die "out of bounds error, operation aborted!";
}
```

To Reverse Truth, use not or !

```
 $ok = ($a < 100 and $a > 0);
 print "a is too small\n" if not $ok;

 # same as this:
 print "a is too small\n" unless $ok;

 # and this:
 print "a is too small\n" if !$ok;
```

and vs &&, or vs ||
&& has higher precedence than and. || has higher precedence than or. This is an issue in assignments:

Low precedence operation:

```
 $ok = $a < 100 and $a > 0;

 # This doesn't mean:
 $ok = ($a < 100 and $a > 0);

 # but:
 ($ok = $a < 100) and $a > 0;
```

High precedence operation:

```
 $ok = $a < 100 && $a > 0;

 # This does mean
 $ok = ($a < 100 && $a > 0);
```

When in doubt, use parentheses.

The "or die" Idiom
The or, and || operators short circuit. If what is on the left is true, then what is on the right is never evaluated, because it doesn't need to be.

```
 $a = 10;
```

```
$b = 99
$a < 100 or $b < 100;  # $b comparison never evaluated
```

The die() Function Aborts Execution with an Error Message

```
die "\$a is the wrong size" unless ($a < 100 and $a > 0);
```

You Combine them Idiomatically Like This

```
($a < 100 and $a > 0) or die "\$a is the wrong size";
```

You can use "and" in the Same Way
If what is on the left of the "and" is FALSE, then Perl doesn't evaluate what's on the right, because it doesn't need to.

```
$a < 0 and die "\$a is negative";
```

File Tests
A bunch of operators are used to check whether files exist, directories exist, files are readable, etc.
-e <filename>
file exists
-r <filename>
file is readable
-x <filename>
file is executable
-w <filename>
file is writable
-d <filename>

filename is a directory

```
 -w "./fasta.out" or die "Can't write to file";

 print "This file is executable\n" if -x "/usr/bin/perl";
```

What is False?
The number 0, the string "0", the empty string, the empty list and undefined are all False.

Arrays
An *Array* Is a List of Values
For example a list with the number 3.14 as the first *element*, the string 'abA' as the second element, and the number 65065 as the third element.

"Literal Representation"
We write the list as above as

(3.14, 'abA', 65065)
If $pi = 3.14 and $s = 'abA' we can also write
($pi, $s, 65065)

We can also do integer ranges:
(-1..5)
shorthand for
(-1, 0, 1, 2, 3, 4, 5)

Counting down not allowed!
Array Variables and Assignment
my $x = 65065;
my @x = ($pi, 'abA', $x);
my @y = (-1..5);
my @z = ($x, $pi, @x, @y);
my ($first, @rest) = @z;
Getting at Array Elements
$z[0] # 65065
$z[0] = 2;
$z[0] # 2
$z[$#z]; # 5
Skip "slices" for now.
Push, Pop, Shift, Unshift
Add 9 to the *end* of @z;
push @z, 9;
Take the 9 off the end of @z, and then take the 5 off the end:
my $end1 = pop @z;
my $end2 = pop @z;
Add 9 to the *beginning* of @z;
unshift @z, 9;
Take the 9 off the beginning of @z, and then take the 3.14 off the beginning:
my $b1 = shift @z;
my $b2 = shift @z;

Hashes
A hash is a lookup table. We use a *key* to find an associated *value*.
my %translate;
$translate{'atg'} = 'M';
$translate{'taa'} = '*';
$translate{'ctt'} = 'K'; # oops
$translate{'ctt'} = 'L'; # fixed
print $translate{'atg'};

Problems
These problems are to be done over the course of several workshops, depending on time. The lecturer will tell you which problems to attempt during the workshop!

Write a script that reads a series of lines with one word per line until the end of the file, then prints a summary of how many times each word appears in the file. You can find a test file named *word_list.txt* in */net/share/perl_problems*. Hint: try using a hash.

```
(~) wordcount.pl word_list.txt
   word appears 14 times
   the appears 10 times
   ...
```

Your results will be substantially different from this. This is just an example.
Modify the above script to print out the words and word counts in alphabetical order.

Write a script to read a text file one line at a time. Determine the length of each line, not counting the newline. When the text file is completely written, print out the total number of lines and the total number of characters in each line.

You'll find a test file named *example1.fasta* in */net/share/perl_problems*

```
% count_lines example1.fasta
TOTAL_LINES = 1392
TOTAL_CHARACTERS = 97441
```

Note, your results will differ from this -- this is just an example for output format.

Modify the previous script to compute the distribution of line lengths. At the end of the script, print out a sorted two-column list of line lengths and the number times each length was seen (hint: use a hash and the sort function).

```
% line_distribution example1.fasta
TOTAL_LINES = 1392
Length     Count
12         1
20         2
28         1
36         1
40         3
60         89
```

Note, your results will differ from this -- this is just an example for output format.

Modify the previous script to present the distribution data sorted by frequency rather than length:

```
% line_distribution2 example1.fasta
TOTAL_LINES = 1392
Length     Count
60         89
40         3
20         2
12         1
28         1
36         1
```

"Unwrap" the contents of a FASTA file, so that each sequence is printed out as one long line. Print the identifier, followed by a tab, followed by the sequence (hint: read up on the input record separator, $/, and on the pattern matching operator):

```
% unwrap example1.fasta
M43911
GATTCCGATCCCCCCCCCAGTTTGACCAAAGTTCAGAGGAAATCCCAGACCAAC
....
L54931
GGGTGGTGGTGAGAGAGAGCGATTGAAAGCTATATATATATGACCGATTCACAG
GT....
L54932
TAGTTGATTCAGTCCGATTTCAATTGATTTCCCGTATATCCTTAAGGGTTTAAA..
..
```

In *Zea maize*, the oligonucleotides Pu-C-G and Pu-C-X-G are two potentially methylated mcrBC recogition sites. Write a program to search for such sites in a FASTA file. The output should be the name of the sequence and the count of such sites found. (Hint: pipe the output of the unwrap program to this script in order to avoid reparsing the FASTA file.)

From the same sequence file, search for sequences of (TA)n repeats, where n >= 5 and "mask" them by replacing them with N's.
From the same sequence file, perform a reverse complimentation on each sequence and print them out in tab-delimited format:

```
% unwrap example1.fasta | reverse_complement
   M43911
GTTGGTCTGGGATTTCCTCTGAACTTTGGTCAAACTGGGGGGGGGGATCGGAA
TC...
   L54931
ACCTGTGAATCGGTCATATATATAGCTTTCAATCGCTCTCTCTCACCACCACCC...
   L54932
TTTAAACCCTTAAGGATATACGGGAAATCAATTGAAATCGGACTGAATCAACT
A...
```

From the same sequence file, split each sequence into three-letter "codons" separated by spaces:

```
% unwrap example1.fasta | codons
M43911 GAT TCC GAT CCC CCC CCC AGT TTG ACC AAA GTT CAG AGG
AAA...
L54931 GGG TGG TGG TGA GAG AGA GCG ATT GAA AGC TAT ATA TAT
GAC...
L54932 TAG TTG ATT CAG TCC GAT TTC AAT TGA TTT CCC GTA TAT CCT...
```

Modify the program to print out codons in three different reading frames. Call it "codons_threeframe":

```
% unwrap example1.fasta | codons_threeframe
```

M43911.1 GAT TCC GAT CCC CCC CCC AGT TTG ACC AAA GTT CAG AGG AAA...
M43911.2 ATT CCG ATC CCC CCC CCA GTT TGA CCA AAG TTC AGA GGA AAC...
M43911.3 TTC CGA TCC CCC CCC CAG TTT GAC CAA AGT TCA GAG GAA ACC...
L54931.1 GGG TGG TGG TGA GAG AGA GCG ATT GAA AGC TAT ATA TAT GAC...
L54931.2 GGT GGT GGT GAG AGA GAG CGA TTG AAA GCT ATA TAT ATG ACT...
...

Pipe the output of this program to a codon translation program named "ribosome":

```
% unwrap example1.fasta | codons_threeframe | ribosome
M43911.1 P  G  G  *  U  L  L  M  X  X  X  X  X  X
M43911.2 ....
M43911.3
L54931.1
L54931.2
```

To help you, cut and paste this translation table:

```
%CODON_TABLE = (
TCA => 'S',TCG => 'S',TCC => 'S',TCT => 'S',
TTT => 'F',TTC => 'F',TTA => 'L',TTG => 'L',
TAT => 'Y',TAC => 'Y',TAA => '*',TAG => '*',
TGT => 'C',TGC => 'C',TGA => '*',TGG => 'W',
CTA => 'L',CTG => 'L',CTC => 'L',CTT => 'L',
CCA => 'P',CCG => 'P',CCC => 'P',CCT => 'P',
CAT => 'H',CAC => 'H',CAA => 'Q',CAG => 'Q',
CGA => 'R',CGG => 'R',CGC => 'R',CGT => 'R',
ATT => 'I',ATC => 'I',ATA => 'I',ATG => 'M',
ACA => 'T',ACG => 'T',ACC => 'T',ACT => 'T',
AAT => 'N',AAC => 'N',AAA => 'K',AAG => 'K',
AGT => 'S',AGC => 'S',AGA => 'R',AGG => 'R',
GTA => 'V',GTG => 'V',GTC => 'V',GTT => 'V',
GCA => 'A',GCG => 'A',GCC => 'A',GCT => 'A',
GAT => 'D',GAC => 'D',GAA => 'E',GAG => 'E',
GGA => 'G',GGG => 'G',GGC => 'G',GGT => 'G');
```

Pipe the output of these programs to a program named "longest_orf", to identify the longest ORF in the three possible reading frames:

```
% unwrap example1.fasta | codons_threeframe | ribosome | longest_orf
ID       Frame       Length (aa)
M43911   2-480       1650
L54931   31-638      202
L54932   1-1032      344
```

Write a program called "gc_content" to determine the GC content of sequences in a FASTA file. The program should use a "sliding window" to compute the %GC, and allow the user to

specify the size of the window on the command line (hint: use the standard Perl Getopt::Long module). The output should look like this:

```
% unwrap example1.fasta | gc_content --window 50
1     48.1
2     48.2
3       48.1
...
1000  55.8
```

Basic Loops

Loops let you execute the same statements over and over again.

***while* Loops**

A *while* loop has a condition at the top. The code within the body will execute until the code becomes false.

```
while ( TEST ) {
   Code to execute
} continue {
   Optional code to execute at the end of each loop
}
```

The *continue* Block

while loops can have an optional *continue* block containing code that is executed at the end of each loop, just before jumping back to the test at the top:

```
#!/usr/local/bin/perl
# file: count_up.pl

$count = 1;
while ( $count <= 5 ) {
   print "count: $count\n";
} continue {
    $count++;
}
```

continue blocks will make more sense after we consider loop control variables.

The *until* Loop

Sometimes you want to loop until some condition becomes true, rather than until some condition becomes false. The *until* loop is easier to read than the equivalent *while (!TEST)*.

```
my $counter = 5;
until ( $counter < 0 ) {
   print $counter--,"\n";  }
```

***foreach* Loops**

foreach will process each element of an array or list:

```
foreach $loop_variable ('item1','item2','item3') {
  print $loop_variable,"\n";
}

@array = ('item1','item2','item3');
foreach $loop_variable (@array) { # same thing, but with an array
  print $loop_variable,"\n";
}

@array = ('item1','item2','item3');
foreach  (@array) { # same difference
  print $_,"\n";
}
```

The last example is interesting. It shows that if you don't explicitly give *foreach* a loop variable, the special scalar variable *$_* is used.

Advanced Loops

The *for* Loop

Consider the standard while loop:

```
initialization code
while ( Test code ) {
  Code to execute in body
} continue {
  Update code
}
```

Loop Control

The next, last, and redo statements allow you to change the flow of control in the loop mid-stream, as it were. You can use these three statements in while loops, until loops, and for and foreach loops, but not in the do-until and do-while variants.

next

The next statement causes the rest of the loop to be skipped and control to pass back to the conditional test at the top. If there's a continue block, it is executed before control returns to the top of the loop.

```
$done = 0;
while (!$done) {
  $line = <STDIN>;
  chomp $line;
  next if $line eq 'SKIP';
  print $line,"\n";
} continue {
  $done++ if $line eq 'END';
}
```

last
The last statement causes the loop to terminate prematurely, even if the loop conditional is still true:

```
while ( $line = <STDIN> ) {
  chomp $line;
  last if $line eq 'END';
  print $line,"\n";
}
```

redo
The redo statement is rarely used. It causes flow of control to jump to the top of the loop, like next. However, the continue block, if any, is not executed. In a for loop, the update expression is not executed.

```
for (my $i=0; $i<10; $i++ {
  chomp ($line = <STDIN>);
  redo if $line eq 'SKIP'; # $i won't get incremented in this case
  print "Read line $i\n";
}
```

Basic I/O
I/O means input/output, and is necessary to get computer programs to talk to the rest of the world.
The STDIN, STDOUT and STDERR Filehandles

Every Perl scripts starts out with three connections to the outside world:

STDIN
Standard input, used to read input. Initially connected to the keyboard, but can be changed from shell using redirection (<) or pipe (|).

STDOUT
Standard output, used to write data out. Initially connected to the terminal, but can be redirected to a file or other program from the shell using redirection or pipes.
STDERR
Standard error, used for diagnostic messages. Initially connected to the terminal, etc.

In addition to these three filehandles, you can create your own.

Filehandles
You can create your own filehandles using the open function, read and/or write to them, and then clean up using close.

open
open opens a file for reading and/or writing, and associates a filehandle with it. You can choose any name for the filehandle, but the convention is to make it all caps. In the examples, we use FILEHANDLE.

open a file for reading
open FILEHANDLE,"cosmids.fasta"
alternative form:
open FILEHANDLE,"<cosmids.fasta"
open a file for writing
open FILEHANDLE,">cosmids.fasta"
open a file for appending
open FILEHANDLE,">>cosmids.fasta'
open a file for reading and writing
open FILEHANDLE,"+<cosmids.fasta"

The Magic of <>
The bare <> function when used without any explicit filehandle is magical. It reads from each of the files on the command line as if they were one single large file. If no file is given on the command line, then <> reads from standard input.

String Substitution
String substitution allows you to replace a pattern or character range with another one using the s/// and tr/// functions.

The s/// Function
s/// has two parts: the regular expression and the string to replace it with: s/*expression*/*replacement*/.
$h = "Who's afraid of the big bad wolf?";
$i = "He had a wife.";

$h =~ s/w.+f/goat/; # yields "Who's afraid of the big bad goat?"
$i =~ s/w.+f/goat/; # yields "He had a goate."
If you extract pattern matches, you can use them in the replacement part of the substitution:
$h = "Who's afraid of the big bad wolf?";

$h =~ s/(\w+) (\w+) wolf/$2 $1 wolf/;
yields "Who's afraid of the bad big wolf?"

What is a Subroutine?
We have been using a form of subroutines all along. Perl *functions* are basically built in subroutines. You call them (or "invoke") a function by typing its name, and giving it one or more arguments.

Example: Length

```
my $seq = 'ATGCAAATGCCA';

my $seq_length = length $seq;  ## OR
my $seq_length = length($seq);

# $seq_length now contains 12
```

Perl gives you the opportunity to define your own functions, called "subroutines". In the simplest sense, subroutines are *named blocks of code* that can be reused as many times as you wish.

Using a Module

A module is a package of useful subroutines and variables that someone has put together. Modules extend the ability of Perl.

Example 1: The File::Basename Module

The File::Basename module is a standard module that is distributed with Perl. When you load the File::Basename module, you get two new functions, *basename* and *dirname*.

basename takes a long UNIX path name and returns the file name at the end. *dirname* takes a long UNIX path name and returns the directory part.

```
#!/usr/bin/perl
# file: basename.pl

use strict;
use File::Basename;

my $path = '/bush_home/bush1/lstein/C1829.fa';
my $base = basename($path);
my $dir  = dirname($path);

print "The base is $base and the directory is $dir.\n";
```

The output of this program is:

The base is C1829.fa and the directory is /bush_home/bush1/lstein.

The use function loads up the module named *File::Basename* and imports the two functions. If you didn't use use, then the program would print an error:

Undefined subroutine &main::basename called at basename.pl line 8.

Problems

Create a script called "add" script to sum two arguments:

```
% add 2 3
5
```

Modify this script so that it checks that both arguments are present:

```
% add 2
Please provide two numeric arguments.
```

Create a script called "now" to print the current time of day:

```
% now
It is now Sun Jun 6 16:35:40 1999
```

Create a script to produce the reverse complement of a sequence (hint, use the reverse and tr/// functions:

```
% reversec GAGAGAGAGAGTTTTTTTTT
AAAAAAAAACTCTCTCTCTC
```

EMBOSS

Arun Anirudhan

EMBOSS is a new, free Open Source software analysis package specially developed for the needs of the molecular biology (e.g. EMBnet) user community. The software automatically copes with data in a variety of formats and even allows transparent retrieval of sequence data from the web. Also, as extensive libraries are provided with the package, it is a platform to allow other scientists to develop and release software in true open source spirit. EMBOSS also integrates a range of currently available packages and tools for sequence analysis into a seamless whole. EMBOSS breaks the historical trend towards commercial software packages.

The EMBOSS suite:

- Provides a comprehensive set of sequence analysis programs (approximately 100)
- Provides a set of core software libraries (AJAX and NUCLEUS)
- Integrates other publicly available packages
- Encourages the use of EMBOSS in sequence analysis training.
- Encourages developers elsewhere to use the EMBOSS libraries.

Applications in EMBOSS

EMBOSS consists over 100 programs (applications). These are just some of the areas covered:

- Sequence alignment
- Rapid database searching with sequence patterns
- Protein motif identification, including domain analysis
- EST analysis
- Nucleotide sequence pattern analysis, for example to identify CpG islands.
- Simple and species-specific repeat identification
- Codon usage analysis for small genomes
- Rapid identification of sequence patterns in large scale sequence sets.
- Presentation tools for publication
- And much more.

These programs are grouped by the following functions

- ALIGNMENT
  - CONSENSUS
  - DIFFERENCES
  - DOT PLOTS
  - GLOBAL
  - LOCAL
  - MULTIPLE
- DISPLAY
- EDIT
- ENZYME KINETICS

- FEATURE TABLES
- INFORMATION
- NUCLEIC
    - 2D STRUCTURE
    - CODON USAGE
    - COMPOSITION
    - CPG ISLANDS
    - GENE FINDING
    - MOTIFS
    - MUTATION
    - PRIMERS
    - PROFILES
    - REPEATS
    - RESTRICTION
    - TRANSCRIPTION
    - TRANSLATION
- PHYLOGENY
- PROTEIN
    - 2D STRUCTURE
    - 3D STRUCTURE
    - COMPOSITION
    - MOTIFS
    - MUTATION
    - PROFILES
- UTILS
    - DATABASE CREATION
    - DATABASE INDEXING
    - MISC

Useful Themes in EMBOSS

Many EMBOSS programs have functionality in common. They all understand the same sorts of sequence formats, output formats and Feature formats. The following are descriptions of some of the common themes in EMBOSS.

Uniform Sequence Addresses

The Uniform Sequence Address, or USA, is a standard sequence naming used by all EMBOSS applications.

The USA syntax is basically one of:

- "format::file"
- "format::file:entry"
- "dbname:entry"
- "@listfile" (a file of file-names)

The "::" and ":" syntax is to allow, for example, "embl" and "pir" to be both database names and sequence formats.

Sequence Formats

The format to use on input can be specified by giving the format name with two colons before the file holding your sequences. For example:

embl::myfile.seq

The format is not required. When reading in a sequence, EMBOSS will guess the **sequence format** by trying all known formats until one succeeds.

When writing out a sequence, EMBOSS will use **fasta** format by default. We can specify another format to use, for example:

gcg::myresults.seq

Alignment Formats

When an alignment of two or more sequences is done by programs in EMBOSS, then the resulting output is written to a file.

There are many different programs in EMBOSS that do many different types of alignments. Some of these programs have been incorporated into EMBOSS from pre-existing programs and some were specially written for it.

The resulting assortments of programs were starting to produce alignment output in a variety of different formats.

Feature Formats

When reading or writing features associated with a sequence, there are a standard set of formats that are used.

The feature files can either be a standard sequence format with a feature table as part of the sequence format, or the features can be held in a file without the associated sequence.

Report Formats

When a report is produced of an analysis by an EMBOSS program, then the resulting output is written to a file.

There are many different programs in EMBOSS that create many different types of reports. Some of these programs have been incorporated into EMBOSS from pre-existing programs and some were specially written for it.

The resulting assortment of programs were starting to produce report output in a variety of different formats

Running EMBOSS Programs

EMBOSS programs are run by typing them at the UNIX prompt, or by using an interface.

Interfaces try to help you by making access to files easier and typically by presenting the programs parameters to you in a form to be filled in. There are many available interfaces.

GUI interfaces

Jemboss- A graphical user interface for EMBOSS, written in Java and called Jemboss.

Staden- The Staden package provides sequence analysis free to academic users. External applications can be included through tcl/tk scripts. Interactive graphics is included and would be a useful display tool for EMBOSS

Web interfaces

Pise

PISE (french pronunciation like the english word 'peas') was developed by Catherine Letondal at the Institute Pasteur. Using an XML definition for each application, PISE generates a Web interface and other common interfaces automatically.

W2H

W2H (german pronunciation "way 2 har") was developed by Martin Senger (see AppLab) at the German EMBnet node as a common web interface for their sequence analysis users.

WEMBOSS

WEMBOSS is being developed by Marc Colet at the Belgian EMBnet node as an EMBOSS web interface for their sequence analysis users.

AppLab

Applab was developed by Martin Senger at EBI.

SRSWWW

SRS is the most widely used biological database indexing and query system worldwide. Originally developed by Thure Etzold at EMBL and EBI, it is now developed by him at Lion Bioscience in Cambridge in collaboration with the EBI.

Bionavigator

BioNavigator is a web interface developed by eBioinformatics Inc, a commercial bioinformatics service provider based in Sydney Australia, in collaboration with the Australian EMBnet node.

Other interfaces

CoLiMate
Kaptain
DISguiSE
SeqPup
CINEMA
Acedb

Reference

1. Rice,P. Longden,I. and Bleasby,A. "EMBOSS: The European Molecular Biology Open SoftwareSuite" Trends in Genetics June 2000, vol 16, No 6. pp.276-277

EXOME and SEQtools

R. Senthilkumar

Many commercial bioinformatics software are available with different programs/ applications for sequence analysis, database search, gene identification, pattern analysis and many other tasks. In this chapter we will discuss about the softwares SEQtools and EXOME.

EXOME

Exome is the Advanced Automated Gene Annotation And Sequence Analysis Software Tool. Exome assists biologists analyze biological data like DNA/RNA on factors like sequence statistics, ORF locations, restriction enzyme sites and coding regions. Exome can also conduct gene prediction and oligomer searches. The friendly graphical user interfaces, assist users analyze sequence data on multiple parameters using a comprehensive multi-tasking window. This concurrent analysis of multiple sequences can even be set up in a scheduler for automatic processing.

Some of the important features of EXOME are

- Import and export nucleotide and protein sequences
- Edit sequences through user-friendly interfaces
- Analyze nucleotide sequence characteristics: like the sequence composition, restriction mapping, ORF finding, predicting genes and coding regions, user defined pattern searches
- Find gene coding regions Dual search outputs through Fourier Transform and Window analysis for sensitive and accurate prediction.
- Translate nucleotide sequences into protein sequences.
- Analyze protein sequence characteristics Sequence composition (amino acid composition with grouping on the basis of physico-chemical characteristics), molecular weight, extinction coefficient and aliphatic index of protein.
- Predict isoelectric point and hydropathy plot: for proteins from their amino acid sequences and detect conserved patterns in protein sequences through a link to the PROSITE dictionary
- Compare target sequences with other sequences:On the NCBI server (using the BLAST algorithm) or multiple sequence alignment with the Clustal W
- Design Primers for PCR: with selection of optimum primers based on a number of critical parameters
- Phylogenetic analysis through tree plot of related sequences
- Schedule analysis of multiple sequences simultaneously in a batch process

Genescan in Exome

GeneScan algorithm predicts the coding regions based on Fourier Analysis of sequence data. Studies have shown that the GeneScan algorithm is one of the most accurate and specific techniques for prediction of probable genes. GeneScan makes use of a universal property of coding sequences and is thus, independent of a training data set of genes. Unlike neural net-based methods, it requires no prior knowledge of the nature and the character of the sequences that constitute the gene in a specific organism

SEQtools

SEQtools is a comprehensive program package for batch handling and analysis of nucleotide and protein sequences. The program includes a series of trivial functions to carry out common operations. In addition SEQtools will assist more demanding operations like unattended blast search and result parsing with hundreds of sequences. Special functions are included for design of microarray gene expression analysis experiments, for expression analyses with the SAGE procedure and for managing small EST projects. Utilities are also included for primer design and ordering, renaming files, creating codon usage tables, building local searchable databases, aligning nucleotide and protein sequences, comparing sequences and a lot more.

Some of the important features of SEQtools are:

- **Back-translate protein sequence** This function *Protein/Back-Translate Protein* performs a back-translation of a peptide sequence into a degenerate DNA sequence and calculates the degree of degeneration. If the degeneration exceeds 10,000 times, the actual value is not displayed.
- **Base composition** Displays the base composition as well as the AT, GC and N content of the currently displayed DNA sequence.
- **Blast client search** This function allows to submit all - or a selected range - of the sequences in a project for blast search connected as a client to the blast server at NCBI (National Center for Biotechnological Information).
- **Blast in-house server** This function allows perform blast searches on a local blast server on local copies of GenBank databases.
- **Codon calculator** It is used to translate codons into amino acids or amino acids into their respective codons.
- **Compare sequences (dot matrix method)** This function allows to compare two sequences, or a sequence with itself, using a dot matrix approach for finding identical regions in the two sequences.
- **Hydrophobicity plot** The function progressively evaluates the hydrophilicity and hydrophobicity of a protein along its amino acid sequence.
- **Restriction map** This function is used to search of restriction sites .
- **Multiple sequence alignment** This function uses the Clustal programs to perform the actual alignment. The ClustalW and Clustal X programs are integrated with SEQtools.
- **Plasmid editor** this function also includes a number of facilities designed to assist in constructing plasmids.
- **Primer design** *Primer Design* is used to construct primers for PCR amplification and DNA sequencing.
- **Primer properties** This form Analysis/Primer Properties shows several characteristics of the current primer sequence including the number of bases, the GC percentage, the molecular weight, the A260 value and the Tm.
- **Chromatograms** SEQtools either perform a simple extraction the sequence information from trace files or perform basecalling using the LifeTrace program, a high quality basecaller written by Dirk Walters. Sequences loaded as trace files maintain a link to the chromatogram, which can be retrieved and viewed with Chromas.

Proteomics Tools

R. Senthilkumar

Proteomics refers to the science and the process of analyzing and cataloging all the proteins encoded by a genome (a proteome). Since the majority of all known and predicted proteins have no known cellular function, the hope is that proteomics will bridge the chasm separating what raw DNA and protein primary sequence reveals about a protein and its cellular function. Determining protein function on a genomewide scale can provide critical pieces to the metabolic puzzle of cells. Because proteins are involved in one measure or another in disease states (whether induced by bacterial or viral infection, stress, or genetic anomaly), complete descriptions of proteins, including sequence structure and function, will substantially aid the current pharmaceutical approach to therapeutics development. This process, known as rational drug design, involves the use of specific structural and functional aspects of a protein to design better proteins or small molecule ligands that can serve as activators or inhibitors of protein function.

The multidimensional nature of proteomics data (for example, 2D-PAGE gel images) presents novel collection, normalization, and analysis challenges. Data collection issues are being overcome by sophisticated proteomic systems that semiautomate and integrate the experimental process with data collection. Improvements in the experimental technology have increased the number of proteins that can be identified, with consistency, within a single gel; however, making comparisons and looking for patterns and relationships between proteins and/or particular environmental, disease, or developmental states requires data mining and knowledge discovery tools.

Proteomics tools

Modern experimental technologies are providing seemingly endless opportunities to generate massive amounts of sequence, expression, and functional data. The drive to capitalize on this enormous pool of information in order to understand fundamental biological phenomena and develop novel therapeutics is pushing the development of new computational tools to capture, organize, categorize, analyze, mine, retrieve, and share data and results. Most current computational applications will suffice for analyses of specific questions using relatively small data sets. But to expand scientific horizons, to accommodate the larger and larger data sets, and to find patterns and see relationships that span temporal and spatial scales, new tools that broaden the scope and complexity of the analyses are needed. Many of these data mining tools available are highlighted below.

Protein identification & characterization

- **AACompIdent(http://us.expasy.org/tools/aacomp/)** - Identify a protein by its amino acid composition
- **AACompSim(http://us.expasy.org/tools/aacsim/)** - Compare the amino acid composition of a SWISS-PROT entry with all other entries
- **MultiIdent(http://us.expasy.org/tools/multiident/)** - Identify proteins with *pI*, *Mw*, amino acid composition, sequence tag and peptide mass fingerprinting data

- **PeptIdent(http://us.expasy.org/tools/peptident.html)** - Identify proteins with peptide mass fingerprinting data, *pI* and *Mw* Experimentally measured, user-specified peptide masses are compared with the theoretical peptides calculated for all proteins in SWISS-PROT, making extensive use of database annotations
- **TagIdent(http://us.expasy.org/tools/tagident.html)** - Identify proteins with *pI*, *Mw* and sequence tag, or generate a list of proteins close to a given *pI* and *Mw*
- **FindMod(http://us.expasy.org/tools/findmod/**)- Predict potential protein post-translational modifications and potential single amino acid substitutions in peptides. Experimentally measured peptide masses are compared with the theoretical peptides calculated from a specified SWISS-PROT entry or from a user-entered sequence, and mass differences are used to better characterize the protein of interest.
- **GlycoMod(http://us.expasy.org/tools/glycomod/)**-Predict possible oligosaccharide structures that occur on proteins from their experimentally determined masses (can be used for free or derivatized oligosaccharides and for glycopeptides)
- **GlycanMass(http://us.expasy.org/tools/glycomod/glycanmass.html)** - Calculate the mass of an oligosaccharide structure
- **FindPept(http://kr.expasy.org/tools/findpept.html)** - Identify peptides that result from unspecific cleavage of proteins from their experimental masses, taking into account artefactual chemical modifications, post-translational modifications (PTM) and protease autolytic cleavage
- **PeptideMass(http://us.expasy.org/tools/peptide-mass.html)** - Calculate masses of peptides and their post-translational modifications for a SWISS-PROT or TrEMBL entry or for a user sequence
- **PepMAPPER (http://wolf.bms.umist.ac.uk/mapper/)-** Peptide mass fingerprinting tool from UMIST, UK
- **Mascot (http://www.matrixscience.com/search_form_select.html)-** Peptide mass fingerprint, sequence query and MS/MS ion search from Matrix Science Ltd., London
- **PepSea (http://195.41.108.38/PepSeaIntro.html)** - Protein identification by peptide mapping or peptide sequencing from Protana, Denmark
- **PeptideSearch(http://www.mann.embl-heidelberg.de/GroupPages/PageLink/peptidesearchpage.html)** - Peptide mass fingerprint tool from EMBL Heidelberg
- **ProteinProspector(http://prospector.ucsf.edu/)** - A variety of tools from UCSF (MS-Fit, MS-Tag, MS-Digest, etc.) for mining sequence databases in conjunction with mass spectrometry experiments [Mirrors at **UCL-Ludwig**, UK / **Ludwig Institute Melbourne** (Australia)]
- **PROWL(http://prowl.rockefeller.edu/)** - Protein chemistry and mass spectrometry resource from Rockefeller and NY Universities
- **CombSearch(http://us.expasy.org/tools/CombSearch/**)- An experimental unified interface to query several protein identification tools accessible on the web

DNA -> Protein

- **Translate (http://us.expasy.org/tools/dna.html)**- Translates a nucleotide sequence to a protein sequence

- **Protein machine(http://www2.ebi.ac.uk/translate/)** - Nucleotide to protein translation at EBI
- **MBS translator (http://mbshortcuts.com/translator/)**- Nucleotide to protein translation at MBS
- **Backtranslation(http://www.entelechon.com/eng/backtranslation.html)** - Translates a protein sequence back to a nucleotide sequence
- **Genewise (http://www.sanger.ac.uk/Software/Wise2/genewiseform.shtml)** - Compares a protein sequence to a genomic DNA sequence, allowing for introns and frameshifting errors
- **FSED(http://ir2lcb.cnrs-mrs.fr/d_fsed/fsed.html)** - Frameshift error detection
- **LabOnWeb(http://www.labonweb.com/)** - Elongation, expression profiles and sequence analysis of ESTs using Compugen LEADS clusters

Similarity searches

- BLAST and WU-BLAST - Interfaces to various versions of the **B**asic **L**ocal **A**lignment **S**earch **T**ool
  - **BLAST** at EMBnet-CH/SIB/CSCS (Switzerland) **(http://www.ch.embnet.org/software/BottomBLAST.html?)**
  - **ExPASy Interface to EMBnet-CH/SIB/CSCS BLAST (http://kr.expasy.org/cgi-bin/BLASTEMBnet-CH.pl)**
  - **BLAST** at NCBI **(http://www.ncbi.nlm.nih.gov/BLAST/)**
  - **WU-BLAST** at Bork's group in EMBL (Heidelberg) **(http://www.bork.embl-heidelberg.de:8080/Blast2/)**
  - **WU-BLAST** and **BLAST** at the EBI (Hinxton) **(http://www2.ebi.ac.uk/blast2/)**
  - **BLAST** at PBIL (Lyon) **(http://pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_blast.html)**
- **Bic** ultra-fast rigorous (Smith/Waterman) similarity searches using the Bioccelerator [At **EBI**, **DKFZ** or at **Weizmann**] **(http://genome.dkfz-heidelberg.de/genweb/)**
- **DeCypher II** (**http://timelogic.com/esdemo.html**)- Smith/Waterman or FrameSearch search using the DeCypher hardware accelerator
- **Fasta3(http://www2.ebi.ac.uk/fasta3/)** - FASTA version 3 at the EBI
- **FDF(http://www.ch.embnet.org/software/FDF_form.html)** - Smith/Waterman type searches on Paracel's Fast Data Finder (FDF) at EMBnet-CH
- **PropSearch(http://www.infobiosud.univ-montp1.fr/SERVEUR/PROPSEARCH/propsearch.html)**; searches for structural homologs using a 'properties' approach [At **EMBL** or at **Montpellier**]
- **SAMBA (http://www.irisa.fr/SAMBA/)**- Systolic Accelerator for Molecular Biological Applications
- **SAWTED(http://bonsai.lif.icnet.uk/bmm/sawted/intro.html)** - Structure Assignment With Text Description
- **Scanps (http://www2.ebi.ac.uk/scanps/)**- Similarity searches using Barton's algorithm

Pattern and profile searches

- **InterPro Scan(http://www.ebi.ac.uk/interpro/scan.html)** - Integrated search in PROSITE, Pfam, PRINTS and other family and domain databases
- **ScanProsite (http://kr.expasy.org/tools/scnpsite.html)**- Scans a sequence against PROSITE or a pattern against SWISS-PROT and TrEMBL
- **ProfileScan(http://hits.isb-sib.ch/cgi-bin/hits_motifscan)** - Scans a sequence against protein profile databases (including PROSITE)
- **Frame-ProfileScan(http://www.isrec.isb-sib.ch/software/PFRAMESCAN_form.html)** - Scans a short DNA sequence against protein profile databases (including PROSITE)
- **Pfam HMM search(http://pfam.wustl.edu/hmmsearch.shtml)**; scans a sequence against the Pfam protein families db [At **Washington University** or at **Sanger Centre**]
- **FPAT(http://alfredo.wustl.edu/fpat/)** - Regular expression searches in protein databases
- **PRATT(http://www2.ebi.ac.uk/pratt/)** - Interactively generates conserved patterns from a series of unaligned proteins
- **PPSEARCH (http://www2.ebi.ac.uk/ppsearch/)-** Scans a sequence against PROSITE (allows a graphical output); at EBI
- **PROSITE scan(http://pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_prosite.html)** - Scans a sequence against PROSITE (allows mismatches); at PBIL
- **PATTINPROT (http://pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_pattinprot.html)**- Scans a protein sequence or a protein database for one or several pattern(s); at PBIL
- **SMART(http://smart.embl-heidelberg.de/)** - Simple Modular Architecture Research Tool; at EMBL
- **TEIRESIAS(http://www.research.ibm.com/bioinformatics/teiresias.html)** - Generate patterns from a collection of unaligned protein or DNA sequences; at IBM
- **Hits (http://hits.isb-sib.ch/)-** Relationships between protein sequences and motifs

Post-translational modification prediction

- **Sulfinator(http://kr.expasy.org/tools/sulfinator/)** - Prediction of tyrosine sulfation sites
- **PSORT(http://psort.nibb.ac.jp/)** - Prediction of protein sorting signals and localization sites
- **SignalP (http://www.cbs.dtu.dk/services/SignalP/)**- Prediction of signal peptide cleavage sites
- **ChloroP(http://www.cbs.dtu.dk/services/ChloroP/)** - Prediction of chloroplast transit peptides
- **MITOPROT(http://www.mips.biochem.mpg.de/cgi-bin/proj/medgen/mitofilter)** - Prediction of mitochondrial targeting sequences
- **Predotar(http://www.inra.fr/Internet/Produits/Predotar/)** - Prediction of mitochondrial and plastid targeting sequences
- **NetOGlyc (http://www.cbs.dtu.dk/services/NetOGlyc/)-** Prediction of type O-glycosylation sites in mammalian proteins

- **DictyOGlyc(http://www.cbs.dtu.dk/services/DictyOGlyc/)** - Prediction of GlcNAc O-glycosylation sites in Dictyostelium
- **YinOYang(http://www.cbs.dtu.dk/services/YinOYang/)** - O-beta-GlcNAc attachment sites in eukaryotic protein sequences
- **big-PI Predictor(http://mendel.imp.univie.ac.at/gpi/)** - GPI Modification Site Prediction
- **DGPI(http://dgpi1.pathbot.com/)** - Prediction of GPI-anchor and cleavage sites (Mirror site)
- **NetPhos(http://www.cbs.dtu.dk/services/NetPhos/)** - Prediction of Ser, Thr and Tyr phosphorylation sites in eukaryotic proteins
- **NetPicoRNA(http://www.cbs.dtu.dk/services/NetPicoRNA/)** - Prediction of protease cleavage sites in picornaviral proteins

Primary structure analysis

- **BLITZ (Bic_SW):(http://www.ebi.ac.uk/searches/blitz_input.html)** Smith & Waterman algorithm implementation for protein database searches.
- **ProSite Database:(http://www.expasy.ch/tools/scnpsite.html)** Scan prosite for patterns which match your sequence.
- **ProtScale:(http://www.expasy.ch/cgi-bin/protscale.pl)** Calculate amino acid scales or local concentrations of particular residues.
- **Multiple sequence alignments at Baylor:(http://dot.imgen.bcm.tmc.edu:9331/multi-align/multi-align.html)** Align multipule protein (or nucleic acid) sequences using ClustalW.
- **BOXSHADE server:(http://ludwig-sun1.unil.ch:8080/software/BOX_form.html)** Take a multiple sequence alignment file (i.e. an alignment file from ClustalW analysis) and produce graphical or color output for slides and prints.
- **PSORT:(http://psort.nibb.ac.jp/)** Search for various sorting signals in proteins.
- **ProtParam(http://kr.expasy.org/tools/protparam.html)** - Physico-chemical parameters of a protein sequence (amino-acid and atomic compositions, *pI*, extinction coefficient, etc.)
- **Compute pI/Mw(http://kr.expasy.org/tools/pi_tool.html)** - Compute the theoretical *pI* and *Mw* from a SWISS-PROT or TrEMBL entry or for a user sequence
- **MW, pI, Titration curve(http://www.up.univ-mrs.fr/~wabim/d_abim/compo-p.html)** - Computes *pI*, composition and allows to see a titration curve
- **REP (http://www.embl-heidelberg.de/~andrade/papers/rep/search.html)**- Searches a protein sequence for a repeats
- **SAPS (http://www.isrec.isb-sib.ch/software/SAPS_form.html)**- Statistical analysis of protein sequences at EMBnet-CH [Also available at **EBI**]
- **Coils(http://www.ch.embnet.org/software/COILS_form.html)** - Prediction of coiled coil regions in proteins (Lupas's method) at EMBnet-CH [Also available at **PBIL**]
- **Paircoil(http://nightingale.lcs.mit.edu/cgi-bin/score**) - Prediction of coiled coil regions in proteins (Berger's method)
- **Multicoil(http://nightingale.lcs.mit.edu/cgi-bin/multicoil)** - Prediction of two- and three-stranded coiled coils

- **PEST(http://www.lif.icnet.uk/LRITu/projects/pest)** - Identification of PEST regions
- **PESTfind (http://www.at.embnet.org/embnet/tools/bio/PESTfind/)**- Identification of PEST regions at EMBnet Austria
- **HLA_Bind(http://bimas.dcrt.nih.gov/molbio/hla_bind/)** - Prediction of MHC type I (HLA) peptide binding
- **SYFPEITHI(http://syfpeithi.bmi-heidelberg.com/scripts/MHCServer.dll/home.htm)** - Prediction of MHC type I and II peptide binding
- **ProtScale(http://kr.expasy.org/cgi-bin/protscale.pl)** - Amino acid scale representation (Hydrophobicity, other conformational parameters, etc.)
- **Drawhca(http://www.lmcp.jussieu.fr/~soyer/www-hca/hca-form.html)** - Draw an HCA (Hydrophobic Cluster Analysis) plot of a protein sequence
- **Protein Colourer(http://www2.ebi.ac.uk/cgi-bin/translate/visprot.pl)** - Tool for coloring your amino acid sequence
- **Colorseq(http://pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_color.html)** - Tool to highlight (in red) a selected set of residues in a protein sequence
- **HelixWheel / HelixDraw(http://www.site.uottawa.ca/~turcotte/resources/HelixWheel/)** - Representations of a protein fragment as a helical wheel
- **RandSeq(http://kr.expasy.org/tools/randseq.html)** - Random protein sequence generator

Secondary structure prediction

- **AGADIR(http://www.embl-heidelberg.de/Services/serrano/agadir/agadir-start.html)** - An algorithm to predict the helical content of peptides
- **BCM PSSP(http://dot.imgen.bcm.tmc.edu:9331/pssprediction/pssp.html)** - Baylor College of Medicine
- **Prof(http://www.aber.ac.uk/~phiwww/prof/)** - Cascaded Mutiple Classifiers for Secondary Structure Prediction
- **GOR I** (Garnier et al, 1978) [At **PBIL** or at **SBDS**] **(http://pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_gor.html)**
- **GOR II** (Gibrat et al, 1987) **(http://pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_gib.html)**
- **GOR IV** (Garnier et al, 1996) **(http://pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_gor4.htm)**
- **HNN (http://pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_nn.html**)- Hierarchical Neural Network method (Guermeur, 1997)
- **Jpred(http://barton.ebi.ac.uk/servers/jpred.html)** - A consensus method for protein secondary structure prediction at EBI
- **nnPredict(http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html)** - University of California at San Francisco (UCSF)
- **PredictProtein(http://cubic.bioc.columbia.edu/predictprotein/)** - PHDsec, PHDacc, PHDhtm, PHDtopology, PHDthreader, MaxHom, EvalSec from Columbia University

- **PREDATOR(http://www.embl-heidelberg.de/cgi/predator_serv.pl)** - Protein secondary structure prediction from single or multiple sequences at EMBL (Argos' group)
- **PSA(http://bmerc-www.bu.edu/psa/request.htm)** - BioMolecular Engineering Research Center (BMERC) / Boston
- **PSIpred(http://insulin.brunel.ac.uk/psipred/)** - Various protein structure prediction methods at Brunel University
- **SOPM(http://pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_sopm.html)** (Geourjon and Del?ge, 1994)
- **SOPMA(http://pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_sopma.html)** (Geourjon and Del?ge, 1995)

Tertiary structure

- **SWISS-MODEL(http://kr.expasy.org/swissmod/SWISS-MODEL.html)** - An automated knowledge-based protein modelling server
- **Geno3d(http://geno3d-pbil.ibcp.fr/)** - Automatic modelling of protein three-dimensional structure
- **CPHmodels(http://www.cbs.dtu.dk/services/CPHmodels/)** - Automated neural-network based protein modelling server
- **3D-PSSM(http://bonsai.lif.icnet.uk/foldfitnew/)** - Protein fold recognition using 1D and 3D sequence profiles coupled with secondary structure information (Foldfit)
- **SWEET (http://www.dkfz-heidelberg.de/spec/sweet2/doc/index.html)-** Constructing 3D models of saccharids from their sequences
- **Swiss-PdbViewer (http://kr.expasy.org/spdbv/)**- A program to display, analyse and superimpose protein 3D structures
- **The Protein Data Bank**:(**http://www.rcsb.org/pdb/)** Brookhaven Protein Data Bank. Retrive coordinates of various solved structures. View in 3D with appropriate add-in viewers.
- **DALI: (http://www2.ebi.ac.uk/dali/)**3D structure alignment server.
- **Molecular Modeling: (http://www.ncbi.nlm.nih.gov/Structure/)** Molecular modeling at NIH.
- **PROSPECT:**(**http://compbio.ornl.gov/structure/prospect/) PRO**tein **S**tructure **P**rediction and **E**valuation **C**omputer **T**oolkit is a threading-based protein structure prediction system.
- **3D Structure Viewers**
  - **RasMol, Chime, & Protein Explorer:**(**http://www.umass.edu/microbio/rasmol/)** 3D Molecular Visualization Freeware.
  - **Cn3D:**(**http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml)** Allows you to view 3D structures from NCBI's Entrez retrieval service. Simultaneously displays structure, sequence, and alignment.
  - **MolMol:(http://www.mol.biol.ethz.ch/wuthrich/software/molmol/)** A free molecular graphics program for displaying, analyzing, and manipulating the three-dimensional structure of biological macromolecules. Runs on UNIX and Windows NT/95/98/2000.

- **WebLab Viewer Lite:(http://www.msi.com/life/products/weblab/viewer/register/lite/download_lite.html)** Free molecular visualization software from MSI.

Transmembrane regions detection

- **DAS(http://www.sbc.su.se/~miklos/DAS/)** - Prediction of transmembrane regions in prokaryotes using the Dense Alignment Surface method (Stockholm University)
- **HMMTOP(http://www.enzim.hu/hmmtop/)** - Prediction of transmembrane helices and topology of proteins (Hungarian Academy of Sciences)
- **PredictProtein (http://dodo.cpmc.columbia.edu/predictprotein/)-** Prediction of transmembrane helix location and topology (Columbia University)
- **SOSUI(http://sosui.proteome.bio.tuat.ac.jp/sosuiframe0.html)** - Prediction of transmembrane regions (TUAT; Tokyo Univ. of Agriculture & Technology)
- **TMAP(http://www.mbb.ki.se/tmap/)** - Transmembrane detection based on multiple sequence alignment (Karolinska Institut; Sweden)
- **TMHMM(http://www.cbs.dtu.dk/services/TMHMM-2.0/)** - Prediction of transmembrane helices in proteins (CBS; Denmark)
- **TMpred(http://www.ch.embnet.org/software/TMPRED_form.html)** - Prediction of transmembrane regions and protein orientation (EMBnet-CH)
- **TopPred 2(http://www.sbc.su.se/~erikw/toppred2/)** - Topology prediction of membrane proteins (Stockholm University)

Sequence alignment

- **SIM + LALNVIEW(http://kr.expasy.org/tools/sim-prot.html)** - Alignment of two protein sequences with SIM, results can be viewed with **LALNVIEW**
- **LALIGN(http://www.ch.embnet.org/software/LALIGN_form.html)** - Finds multiple matching subsegments in two sequences
- **Dotlet (http://www.isrec.isb-sib.ch/java/dotlet/Dotlet.html)**- A Java applet for sequence comparisons using the dot matrix method

References

1. www.expasy.org
2. www.mtrc.net

Comparative Molecular Modeling

R. SenthilKumar

Functional characterization of a protein sequence is one of the most frequent problems in biology. This task is usually facilitated by accurate three-dimensional (3D) structure of the studied protein. The experimental elucidation of the 3D-structure of proteins is however often hampered by difficulties in obtaining sufficient protein, diffracting crystals and many other technical aspects. Therefore the number of solved 3D-structures increases only slowly compared to the rate of sequencing of novel cDNAs, and no structural information is available for the vast majority of the protein sequences registered in the SWISS-PROT database. In the absence of an experimentally determined structure, comparative or homology modeling can sometimes provide a useful 3D model for a protein that is related to at least one known protein structure.

Proteins from different sources and sometimes diverse biological functions can have similar sequences, and it is generally accepted that high sequence similarity is reflected by distinct structure similarity. Indeed, the relative mean square deviation (rmsd) of the alpha-carbon co-ordinates for protein cores sharing 50% residue identity is expected to be around 1Å. This fact served as the premise for the development of comparative protein modeling (also often called modeling by homology or knowledge-based modeling), which is presently the most reliable method. Comparative modeling predicts the 3D structure of a given protein sequence (target) based primarily on its alignment to one or more proteins of known structure (templates). The prediction process consists of fold assignment, target-template alignment, model building, and model evaluation. The number of protein sequences that can be modeled and the accuracy of the predictions are increasing steadily because of the growth in the number of known protein structures and because of the improvements in the modeling software. It is currently possible to model with useful accuracy significant parts of approximately one half of all known protein sequences. Even models with errors may be useful, because some aspects of function can be predicted from only coarse structural features of a model

STEPS IN COMPARATIVE MODELING

1. Identification of modelling templates

The starting point in comparative modeling is to identify all protein structures related to the target sequence, and then select those structures that will be used as templates. This step is facilitated by numerous protein sequence and structure databases, and database scanning software available on the web. Templates can be found using the target sequence as a query for searching structure databases such as the Protein Data Bank, SCOP, DALI , CATH and NCBI .

There are three main classes of protein comparison methods that are useful in fold identification. The first class includes the methods that compare the target sequence with each of the database sequences independently, using pairwise sequence-sequence comparison. The performance of these methods in searching for related protein sequences and structures has been evaluated exhaustively. Frequently used programs in this class include FASTA and BLAST.

The second set of methods relies on multiple sequence comparisons to improve the sensitivity of the search. A widely used program in this class is PSI-BLAST, which iteratively expands the set of homologs of the target sequence. For a given sequence, an initial set of homologs from a sequence database is collected, a weighted multiple alignment is made from the query sequence and its homologs, a position specific scoring matrix is constructed from the alignment, and the matrix is used to search the database for additional homologs. These steps are repeated until no additional homologs are found. In comparison to BLAST, PSI-BLAST finds homologs of known structure for approximately twice as many sequences.

The third class of methods is the so-called threading or 3D template matching methods. These methods rely on pairwise comparison of a protein sequence and a protein of known structure. Whether or not a given target sequence adopts any one of the many known 3D folds is predicted by an optimization of the alignment with respect to a structure dependent scoring function, independently for each sequence-structure pair. That is, the target sequence is threaded through a library of 3D folds. These methods are especially useful when there are no sequences clearly related to the modeling target, and thus the search cannot benefit from the increased sensitivity of the sequence profile methods. A useful fold assignment approach is to accept an uncertain assignment provided by any of the methods, build a full-atom comparative model of the target sequence based on this match, and make the final decision about whether or not the match is real by evaluating the resulting comparative model.

Once a list of all related protein structures has been obtained, it is necessary to select those templates, several factors should be taken into account when selecting the templates.

- The family of proteins, which includes the target and the templates, can frequently be organized in sub-families. The construction of a multiple alignment and a phylogenetic tree can help in selecting the template from the sub-family that is closest to the target sequence.
- The template properties that determine protein structure (*e.g. solvent*, pH, ligands, and quaternary interactions) should be compared to the required environment for the model.
- The priority of the criteria for template selection depends on the purpose of the comparative model. For instance, if a protein-ligand model is to be constructed, the choice of the template that contains a similar ligand is probably more important than the resolution of the template. On the other hand, if the model is to be used to analyze the geometry of the active site of an enzyme, it is preferable to use a high-resolution template. It is not necessary to select only one template. In fact, the use of several templates approximately equidistant from the target sequence generally increases the model accuracy

2. Aligning the target sequence with the template sequence

The target sequence now needs to be aligned with the template sequence. Most fold assignment methods produce an alignment between the target sequence and template structures. However, this is often not the optimal target-template alignment for comparative modeling. Searching methods are usually tuned for detection of remote relationships, not for optimal alignments. Therefore, once templates have been selected, a specialized method should be used to align the target sequence with the template structures. For closely related protein sequences with identity higher than 40%, the alignment is almost always correct. Regions of low local sequence similarity become common when the overall sequence identity

is below 40%. The alignment becomes difficult in the "twilight zone" of less than 30% sequence identity. As the sequence similarity decreases, alignments contain an increasingly large number of gaps and alignment errors, regardless of whether they are prepared automatically or manually. For example, only 20% of the residues are likely to be correctly aligned when two proteins share 30% sequence identity. Maximal effort to obtain the most accurate alignment possible is needed because no current comparative modeling method can recover from an incorrect alignment. There is a great variety of protein sequence alignment methods, many of which are based on dynamic programming techniques. A frequently used program for multiple sequence alignment is CLUSTAL. In the more difficult alignment cases, it is frequently beneficial to rely on multiple structure and sequence information first, the alignment of the potential templates is prepared by superposing their structures. Next, the sequences that are clearly related to the templates and are easily aligned with them are added to the alignment. The same is done for the target sequence. Finally, the two profiles are aligned with each other, taking structural information into account as much as possible.

Building the model

Once an initial target-template alignment has been built, a variety of methods can be used to construct a 3D model for the target protein. The original and still widely used methods are modeling by rigid-body assembly, modeling by segment matching, relies on the approximate positions of conserved atoms in the templates, modeling by satisfaction of spatial restraints, uses either distance geometry or optimization techniques to satisfy spatial restraints obtained from the alignment. Accuracies of the various model building methods are relatively similar when used optimally. Other factors, such as template selection and alignment accuracy usually, have a larger impact on the model accuracy, especially for models based on less than 40% sequence identity to the templates.

a. Framework construction

The next step is the construction of a framework, which is computed by averaging the position of each atom in the target sequence, based on the location of the corresponding atoms in the template. When more than one template is available, the relative contribution, or weight, of each structure is determined by its local degree of sequence identity with the target sequence.

b. Building non-conserved loops

Following framework generation, loops for which no structural information was available in the template structures are not defined and therefore must be constructed. Although most of the known 3D-structures available share no overall similarity with the template, there may be similarities in the loop regions, and these can be inserted as loop structure in the new protein model. Using a "spare part" algorithm, one searches for fragments, which could be accommodated onto the framework among the Brookhaven Protein Data Bank (PDB) entries determined with a resolution better than 2.5 Å. Each loop is defined by its length and its "stems", namely the alpha carbon (C□) atom co-ordinates of the four residues preceding and following the loop. The fragments which correspond to the loop definition are extracted from the PDB entries and rejected if the relative mean square deviation (rmsd) computed for their "stems" is greater than a specified cut-off value. Furthermore, only fragments which do not overlap with neighbouring segments should be retained. The accepted "spare parts" are sorted according to their rmsd, and a C□ framework based on the five best fragments can be

added to the model. In order to ensure that the best possible fragments are used for loop rebuilding, the rmsd cut-off can be incremented from 0.2 onwards until all loops are rebuilt.

c. Completing the backbone

Since the loop building only adds C□ atoms□ the backbone carbonyl and nitrogens must be completed in these regions. This step can be performed by using a library of pentapeptide backbone fragments derived from the PDB entries determined with a resolution better than 2.0 Å. These fragments are then fitted to overlapping runs of five C□ atoms of the target model. The co-ordinates of each central tripeptide are then averaged for each target backbone atom (N, C, O) and added to the model. This process yields modelled backbones that differ from experimental co-ordinates by approx. 0.2 Å rms.

d. Adding side chains

For many of the protein side chains there is no structural information available in the templates. These cannot therefore be built during the framework generation and must be added later. The number of side chains that need to be built is dictated by the degree of sequence identity between target and template sequences. To this end one uses a table of the most probable rotamers for each amino acid side chain depending on their backbone conformation. All the allowed rotamers of the residues missing from the structure are analysed to see if they are acceptable by a van der Waals exclusion test. The most favoured rotamer is added to the model. The atoms defining the □ 1 and □ 2 angles of incomplete side chains can be used to restrict the choice of rotamers to those fitting these angles. If some side chains cannot be rebuilt in a first attempt, they will be assigned initially in a second pass. This allows some side chains to be rebuilt even if the most probable allowed rotamer of a neighbouring residue already occupies some of this portion of space. The latter may then switch to a less probable but allowed rotamer. In case that not all of the side chains can be added, an additional tolerance of 0.15 Å can be introduced in the van der Waals exclusion test and the procedure repeated.

e. Model refinement

Idealisation of bond geometry and removal of unfavorable non-bonded contacts can be performed by energy minimisation with force fields such as CHARMM, AMBER or GROMOS. The refinement of a primary model should be performed by no more than 100 steps of steepest descent, followed by 200-300 steps of conjugate gradient energy minimisation. Experience has shown models optimised that energy minimisation (or molecular dynamics) usually move away from a control structure. It is thus necessary to keep the number of minimisation steps to a minimum. Constraining the positions of selected atoms (such as C□, or using a B-factor based function□ in each residue generally helps avoiding excessive structural drift during force field computations.

4. Model evaluation

The quality of the predicted model determines the information that can be extracted from it. Thus, estimating the accuracy of 3D protein models is essential for interpreting them. The model can be evaluated as a whole as well as in the individual regions. There are many model evaluation programs and servers. The quality of the predicted model determines the information that can be extracted from it. Thus, estimating the accuracy of 3D protein models is essential for interpreting them. The model can be evaluated as a whole as well as in the individual regions. There are many model evaluation programs and servers.

The first step in model evaluation is to determine if the model has the correct fold A model will have the correct fold if the correct template is picked and if that template is aligned at least approximately correctly with the target sequence. The confidence in the fold of a model is generally increased by a high sequence similarity with the closest template, an energy based Z-score, or by conservation of the key functional or structural residues in the target sequence.

Once the fold of a model is accepted, a more detailed evaluation of the overall model accuracy can be obtained based on the similarity between the target and template sequences. Sequence identity above 30% is a relatively good predictor of the expected accuracy. The reasons are the well known relationship between structural and sequence similarities of two proteins, the "geometrical" nature of modeling that forces the model to be as close to the template as possible, and the inability of any current modeling procedure to recover from an incorrect alignment. The dispersion of the model target structural overlap increases with the decrease in sequence identity. If the target template sequence identity falls below 30%, the sequence identity becomes unreliable as a measure of expected accuracy of a single model. Models that deviate significantly from the average accuracy are frequent. It is in such cases that model evaluation methods are particularly useful.

In addition to the target-template sequence similarity, the environment can strongly influence the accuracy of a model. For instance, some calcium-binding proteins undergo large conformational changes when bound to calcium. If a calcium-free template is used to model the calcium-bound state of the target, it is likely that the model will be incorrect irrespective of the target-template similarity or accuracy of the template structure. This also applies to the experimental determination of protein structure; a structure must be determined in the functionally meaningful environment.

A basic requirement for a model is to have good stereochemistry. Some useful programs for evaluating stereochemistry are PROCHECK, ROCHECKNMR, AQUA, SQUID and WHATCHECK. The features of a model that are checked by these programs include bond lengths, bond angles, peptide bond and sidechain ring planarities, chirality, mainchain and sidechain torsion angles, and clashes between non-bonded pairs of atoms.

There are also methods for testing 3D models that implicitly take into account many spatial features compiled from high resolution protein structures. These methods are based on 3D profiles and statistical potentials of mean force. Programs implementing this approach include VERIFY3D, PROSAII, HARMONY and ANOLEA. The programs evaluate the environment of each residue in a model with respect to the expected environment as found in the high-resolution X-ray structures.

ERRORS IN COMPARATIVE MODELS

As the similarity between the target and the templates decreases, the possibilities of error can increase.

Errors in sidechain packing: As the sequences diverge, the packing of sidechains in the protein core changes. Sometimes even the conformation of identical sidechains is not conserved, a pitfall for many comparative modeling methods. Sidechain errors are critical if

they occur in regions that are involved in protein function, such as active sites and ligand-binding sites.

Distortions and shifts in correctly aligned regions: As a consequence of sequence divergence, the mainchain conformation changes, even if the overall fold remains the same. Therefore, it is possible that in some correctly aligned segments of a model, the template is locally different (< 3Å) from the target, resulting in errors in that region. The structural differences are sometimes not due to differences in sequence, but are a consequence of artifacts in structure determination or structure determination in different environments (*e.g.*, packing of subunits in a crystal). The simultaneous use of several templates can minimize this kind of an error.

Errors in regions without a template: Segments of the target sequence that have no equivalent region in the template structure (*i.e.* insertions or loops) are the most difficult regions to model. If the insertion is relatively short, less than 9 residues long, some methods can correctly predict the conformation of the backbone. Conditions for successful prediction are the correct alignment and an accurately modeled environment surrounding the insertion.

Errors due to misalignments: The largest source of errors in comparative modeling are misalignments, especially when the target-template sequence identity decreases below 30%. However, alignment errors can be minimized in two ways. First, it is usually possible to use a large number of sequences to construct a multiple alignment, even if most of these sequences do not have known structures. Multiple alignments are generally more reliable than pairwise alignments. The second way of improving the alignment is to iteratively modify those regions in the alignment that correspond to predicted errors in the model .

Incorrect templates: This is a potential problem when distantly related proteins are used as templates (*i.e.*, less than 25% sequence identity). Distinguishing between a model based on an incorrect template and a model based on an incorrect alignment with a correct template is difficult. In both cases, the evaluation methods will predict an unreliable model. The conservation of the key functional or structural residues in the target sequence increases the confidence in a given fold assignment.

APPLICATIONS OF COMPARATIVE MODELING

Comparative modeling is an increasingly efficient way to obtain useful information about the proteins of interest. For example, comparative models can be helpful in designing mutants to test hypotheses about a protein's function, identifying, designing and improving ligands for a given binding site, modeling substrate specificity, predicting antigenic epitopes, simulating protein-protein docking, inferring function from a calculated electrostatic potential around the protein, facilitating molecular replacement in X-ray structure determination, refining models based on NMR constraints, testing and improving a sequence-structure alignment, confirming a remote structural relationship, and rationalizing known experimental observations.

Programs and web servers useful for comparative modeling

Database
CATH - www.biochem.ucl.ac.uk/bsm/cath/
GenBank – www.ncbi.nlm.nih.gov/Genbank/
GeneCensus- bioinfo.mbb.yale.edu/genome/
MODBASE- guitar.rockefeller.edu/modbase/
PDB - www.rcsb.org/pdb/
PRESAGE presage.berkeley.edu
SCOP - scop.mrc-lmb.cam.ac.uk/scop
TrEMBL - srs.ebi.ac.uk

Template Search
123D - 123d.ncifcrf.gov/123D+.html
BLAST - www.ncbi.nlm.nih.gov/BLAST/
DALI- www2.ebi.ac.uk/dali
FastA - www.ebi.ac.uk/fasta33
MATCHMAKER - bioinformatics.burnham-inst.org
PHD, TOPITS - S cubic.bioc.columbia.edu/predictprotein/
PROFIT - www.came.sbg.ac.at
THREADER - insulin.brunel.ac.uk/~jones/threader.html
FRSVR - S fold.doe-mbi.ucla.edu

Sequence Alignment
BCM SERVER - searchlauncher.bcm.tmc.edu
BLAST2 - www.ncbi.nlm.nih.gov/blast
CLUSTAL -www2.ebi.ac.uk/clustalw/
FASTA3 - www2.ebi.ac.uk/fasta3/
MULTALIN - pbil.ibcp.fr

Modelling
COMPOSER - www.tripos.com/software/composer.html
CONGEN - www.congenomics.com/congen/congen.html
MODELLER -guitar.rockefeller.edu/modeler

InsightII - www.accelrys.com
ICM -www.molsoft.com
QUANTA - www.accelrys.com
SYBYL - www.tripos.com
SCWRL - www.fccc.edu/research/labs/dunbrack/scwrl/
SWISS-MOD – www.expasy.org/swissmod/SWISS-MODEL.html
WHAT IF - www.cmbi.kun.nl/whatif/

Model evaluation
ANOLEA - www.fundp.ac.be/sciences/biologie/bms/CGI/anolea.html
AQUA - urchin.bmrb.wisc.edu/~jurgen/aqua/
BIOTECH - biotech.embl-heidelberg.de:8400
ERRAT - www.doe-mbi.ucla.edu/Services/ERRAT/
PROCHECK - www.biochem.ucl.ac.uk/~roman/procheck/procheck.html
ProsaII - www.came.sbg.ac.at
PROVE - www.ucmb.ulb.ac.be/UCMB/PROVE
SQUID - www.ysbl.york.ac.uk/~oldfield/squid/
VERIFY3D -www.doe-mbi.ucla.edu/Services/Verify_3D/
WHATCHECK -www.sander.embl-heidelberg.de/whatcheck/
Methods evaluation
CASP -predictioncenter.llnl.gov
CAFASP - cafasp.bioinfo.pl
EVA - cubic.bioc.columbia.edu/eva/
LiveBench - bioinfo.pl/LiveBench**/**

Bioinformatics in India

Santhosh J. Eapen

Bioinformatics has been in the news of late. It is becoming increasingly important due to the interest of the pharmaceutical industry in genome sequencing projects. There is a vital need to harness this information for medical diagnostic and therapeutic uses, and there are opportunities for other industrial applications. This field is evolving rapidly, which makes it challenging for biotechnology professionals to keep up with recent advancements. With its strengths in science and software, India proved to be a natural choice for the emerging field of Bioinformatics—using information technology to make sense of biological data.

Biotechnology Information System of India

Recognizing the importance of information technology for pursuing advanced research in modern biology and biotechnology, India launched a Bioinformatics programme, envisaged as a distributed database and network organisation, during 1986-87. The establishment of a National Bioinformatics Network known as Biotechnology Information System of India (BTIS) by the Department of Biotechnology (DBT), Government of India in the year 1987 initiated the Bioinformatics activities in India. The network programme consists of 10 Distributed Information Centres (DICs), 50 Sub-Distributed Information Centres (Sub-DICs) and an apex Biotechnology Information Centre (BTIC) that coordinates the activities of the entire network. BTISnet is unique and one of the well recognized major scientific networks in the world, dedicated to providing the-state-of-the-art infrastructure, technologies, education, manpower and tools in Bioinformatics.

The entire network has emerged as a very sophisticated scientific infrastructure for Bioinformatics involving state-of-the-art computational and communication facilities. The computer communication network, linking all the Bioinformatics centres, is playing a vital role in the success of the Bioinformatics programme. Database development, R&D activities in Bioinformatics, human resource development and a variety of services in support of biotechnology R&D programmes and projects, has made this programme very popular and useful to the scientific community.

A national node of the European Molecular Biology Network (EMBnet), has been established at the Centre of DNA Fingerprinting and Diagnostics (CDFD), Hyderabad. Four mirror sites for mirroring important biological databases are being established at IISc, JNU, Poona University, and IMTECH to promote and support R&D activities in Genomics and Proteomics, the two emerging fields of biotechnology requiring critical support of genomic databases. With these resources now available on the BTISnet, it has now become a single largest information resource for all references to biotechnology related literature, scientific data, patent information, policy matters and related issues.

Human resource development has been recognized as an important area for effective sustenance of the Bioinformatics programme. A number of workshops and training programmes were conducted under the BTIS programme on the use of computers and databanks in modern biology and biotechnology. Considering the importance of the subject, some institutions and university departments have introduced a formal course of training in Bioinformatics as a 3- credit one-semester course in their existing post-graduate programmes in Biotechnology. DBT has initiated a long-term academic course in Bioinformatics leading to the award of an advanced diploma in Bioinformatics. The course is being conducted by Madurai Kamaraj University, Madurai, Pune University, Pune, Jawaharlala University and Calcutta University.

Six biocomputing facilities under the umbrella of "National Facilities on Interactive Graphics and Molecular Modeling" have been established with the task of providing discipline wise facilities to the scientists working in the area of molecular structure modeling, 3D structures, active site modeling, crystal structures, conformational anlysis, protein and DNA structures and interactions, homology studies and like. These biocomputing centres are located at 1) Indian Institute of Science, Bangalore, 2) Centre for Cellular and Molecular Biology, Hyderabad, 3) Madurai Kamaraj University, Madurai, 4) University of Poona, Pune, 5) Bose Institute, Calcutta and 6) Jawaharlal Nehru University, New Delhi.

The BTIS centers have developed more than 100 databases on various aspects of biotechnology. Several major International databases for application to genomics and proteomics have been established in the form of mirrors as part of the network. A BIOGRID INDIA in the form Virtual Private Network (VPN) has been established to link these databases through high speed and large bandwidth network to promote faster sharing of information. The network was established through HCL Infosystems and is being coordinated by National Brain Research Center (NBRC). Eleven institutions viz. JNU, DBT, Delhi University, South Campus (Rice Genome Project), CBT, NII, NBRC (Gurgaon), MKU Madurai, IISc Bangalore, CDFD, Hyderabad, University of Pune and IMTECH, Chandigarh have been networked under this project in the first phase. The mirror sites of internationally recognized genomic databases such as GDB, Protein Data Bank (PDB), Plant

Genome Data Banks, Databases of European Bioinformatics institute (EBI) and Public domain Bioinformatics software packages have been mirrored as part of VPN. The advantage of mirroring these databases in India is to provide unhindered access to large amount of data bank for analysis of primary and secondary information sources. The network will act as knowledge pathways for discoveries.

The DBT has identified Bioinformatics as an area of high priority during the tenth plan period. The draft National Bioinformatics Policy 2003 of the DBT has called for a total revamping of the existing programme by allowing 51 per cent equity by foreign companies and institutions in government funded research centres. The policy has also recommended the setting up of an independent autonomous institution, a National Virtual Centre for Bioinformatics, with control over the existing network.

The Corporate Sector

Besides, there are several private organizations where Bioinformatics research is progressing. Companies like Bigtec conducted work on systems for data management of clinical trials that are compatible with US FDA requirements for electronic records systems. It is also working currently on clustering analysis of micro array data with detection of expression patterns. Strand Genomics was involved in developing IP rich solutions using PCs running Red Hat and Debian Linux for the problems faced by pharma and biotech companies. The company is also developing an integrated software suite christened Oyster for accelerating drug discovery process, which will include features like macro molecular structure modelling and prediction and 3D structure visualisation. On its part, Kshema Technologies has developed some solutions in the areas of ultra-trace chemical detection based on mass spectrometry techniques. Tamil Nadu Industrial Development Corporation (TIDCO) plans to invest close to Rs 20 crore to develop and commercialise advanced laboratory and computational technologies for high throughput DNA sequencing and for accelerating the identification and interpretation of disease-causing genetic mutations. Companies like BIOCON, ASTRA-ZENICA, Dr. Reddy's Lab, TATA-ELXSI, Jubilant, Ocimum, BIOSIS etc. are some of the leading private firms in India, active in this emerging field.

Because of the hype, several IT companies show a keen interest in this field. IBM Global Services India is currently in the process of scanning the entire spectrum of life sciences and has already identified some areas of work related to Bioinformatics. Satyam too threw in the hat when it entered into a five-year agreement with the Centre for Cellular and Molecular Biology (CCMB) to develop tools to sift through masses of genetic material in search of vital DNA fragments. The CCMB also looked at possible collaboration with three other Indian IT companies.

Bioinformatics Courses

In India, there are three types of courses available. Most notable among these are the post-graduate advanced diploma courses offered by universities like University of Pune, Jawaharlal Nehru University, New Delhi, Madurai Kamaraj University, Calcutta University and University of Hyderabad. Recently several universities like Bharthiar University, Coimbatore is offering M. Sc Bioinformatics courses. The Government of Karnataka in association with ICICI has established an Institute of Bioinformatics and Applied Biotechnology (www.ibab.ac.in), which offers post-graduate course in Bioinformatics and short-term training programmes. Bioinformatics Institute of India at Noida

(www.bioinformaticscentre.org) is a non-profit making educational, R&D and devlopment centre of Bioinformatics, offering various diploma courses.

In some of the IITs one can pursue Bioinformatics research leading to Ph. D. degree. Several short-term courses are conducted by various DICs and DISCs mentioned above. The latest trend is local computer firms offering various courses in Bioinformatics. One should be extremely careful in choosing the right centre, which offer quality courses. Companies are still looking for professionals with backgrounds in Bioinformatics, IT, molecular biology, and protein separation chemistry. There is also a demand for software engineers who can craft the instrument control and data acquisition instrumentation as well as data analysis tools, including database-searching tools. Finally, to sort out all the data being generated, biostatisticians are very much in need. Though Indian institutions have been providing quality education both in biosciences and IT, much more still needs to be done for India to excel in this new area.

References

1. http://www.btisnet.ac.in
2. http://www.ibab.ac.in
3. http://www.bioinformaticscentre.org
4. Biopolicy. Biospectrum April 2003. pp. 48-51

Annexure I

General Conversion Tables and Formulae

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

Annexure II

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 :

(a) Circular DNA molecule : 2 x (moles of DNA) x (number of sites)
(b) 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	1.6×10^{-12}	9.8×10^{11}	1.6nM	3.2nM
1.0 A_{260} pBR 322 DNA =	50µg/ml	1.8×10^{-11}	1.1×10^{13}	18.0nM	36.0nM
1.0 A_{260} linker (8 MER) =	50µg/ml	9.8×10^{-9}	5.9×10^{15}	10.0µM	20.0µM

Annexure III

Codon Usage Table

	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.

Annexure IV

Amino Acid Abbreviations and Molecular Weights

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

$(2 \times 10^6) / (660 \times \text{Number of Bases}) = \text{pmole ends}/\mu g$

Exact molecular weight of an oligonucleotide

$[(A \times 312.2) + (G \times 328.2) + (C \times 288.2) + (T \times 303.2) - 61.0] = \text{MW (g/mol) of specific oligonucleotide}$

Annexure V

BOOKS ON BIOINFORMATICS

Adams M.D., Fields C., Venter J.C. (Eds.) **Automated DNA sequencing and analysis,** Academic Press, London (1994).

Baxevanis A., Ouellette F.B.F. (Eds.) Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins John Wiley and Sons, New York (1998).

Bishop M.J., Rawlings C.J. (Eds.) Nucleic acid and protein sequence analysis. A Practical Approach,IRL Press, Oxford (1987)

Bishop M.J., Rawlings C.J. (Eds.) DNA and protein sequence analysis. A Practical approach IRL Press, Oxford (1997)

Bishop M.J. (Ed.) **Guide to human genome computing.** Second edition Academic Press, London (1998).

Colwell R.R., Swartz D.G., McDonald M.T. (Eds.) Biomolecular data: A resource in transition. Oxford University Press, Oxford (1989).

Doolittle R.F. Of URFs and ORFs: a primer on how to analyze derived amino acid sequences.University Science Books, Mill Valley California (1986).

Doolittle R.F. (Ed.) Molecular Evolution: computer analysis of protein and nucleic acid sequences (Methods in Enzymology, Vol. 183). Academic Press, San Diego (1990).

Doolittle R.F. (Ed.) Computer methods for macromolecular sequence analysis (Methods in Enzymology, Vol. 266). Academic Press, San Diego (1996).

Durbin R., Eddy S.R., Krogh A., Mitchison G. Biological sequence analysis. Probabilistic models of proteins and nucleic acids Cambridge University Press, Cambridge (1998).

Gribskov M., Devereux J. (Eds.) **Sequence analysis primer** Stockton Press, New York (1991).

Griffin A.M., Griffin H.G. (Eds.) Methods in molecular biology Vol. 24: Computer analysis of sequence data, part 1. Humana Press, Totowa (1994).

Griffin A.M., Griffin H.G. (Eds.) Methods in molecular biology Vol. 25: Computer analysis of sequence data, part II. Humana Press, Totowa (1994).

Gusfield D. Algorithms on strings, trees, and sequences. Computer science and computational biology. Cambridge University Press, Cambridge (1997).

Lesk A.M. (Ed.) Computational molecular biology. Sources and methods for sequence analysis. Oxford University Press, Oxford (1988).

Peruski L.F. Jr,, Harwood Peruski A. The Internet and the new biology: tools for genomic and molecular research. American Society for Microbiology, Washington DC (1997).

SchuIze-Kremer S. **Molecular Bioinformatics: algorithms and applications.** Walter de Gruyter,Berlin - New-York (1995).
Setubal J., Meidanis J. **Introduction to computational molecular biology.** PWS Publishing Co.,Boston (1996).

Sillince J., Sillince M. **Molecular databases for protein sequences and structure studies: an introduction.** Springer Verlag, Berlin (1992).

Smith D.W. (Ed.) **Biocomputing. Informatics and genome projects.** Academic Press, New York (1994).

Suhai S. (Ed.) Computational methods in genome research. Plenum Press, New York (1994).

Suhai S. (Ed.) **Theoretical and computational methods in genome research.** Plenum Press, New York (1997).

Swindell S.R., Miller R.R., Myers G.S.A. (Eds.) Internet for the molecular biologist Horizon Scientific Press, Norfolk (1996).

Swindell S.R. (Ed.) Methods in molecular biology Vol. 70: Sequence data analysis guidebook.Humana Press, Totowa (1997).

Trifonov E.N., Brendel V. **GNOMIC, a dictionary of genetic codes.** Balaban Publishers, Philadelphia(1986).

Von Heijne G.; **Sequence analysis in molecular biology. Treasure trove or trivial pursuit.** Academic Press, London (1987).

Waterman M.S. (Ed.) **Mathematical methods for** DNA **sequences.** CRC Press, Boca Raton (1989).

Waterman M.S. **Introduction to computational biology: maps, sequences, and genomes**, Chapman and Hall, London (1995).

Wilkins M.R., Williams K.L, Appel R.D., Hochstrasser D.H. (Eds.) Proteome research: new frontiers in functional genomics. Sringer Verlag, Berlin (1997).

Yap T.K, Frieder 0., Martino R.L. High performance computational methods for biological sequence analysis. Kluwer Academic Publisher, Dordrecht (1996).

Annexure VI

Major Biological Datbases & Information Resources

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
MagnaportheDB (http://www.cals.ncsu.edu/fungal_genomics/mgdatabase/int.htm) - Integrated physical and genetic maps for the rice blast fungus Magnaporthe grisea
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
MitBASE (http://www3.ebi.ac.uk/Research/Mitbase/mitbase.pl) - Mitochondrial genomes, intra-species variants and mutants
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
BRENDA (http://www.brenda.uni-koeln.de/) - Extensive functional data on enzymes
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
FUNPEP (http://picsou.cmbi.kun.nl:8080/) - Low-complexity or compositionally-biased protein sequences
GenProtEC (http://genprotec.mbl.edu/) - E.coli K-12 genome, gene products and homologs
InBase (http://www.neb.com/neb/inteins.html) - All known inteins (protein splicing elements): properties, sequences, bibliography
Kabat Database (http://immuno.bme.nwu.edu/) - Sequences of proteins of immunological interest
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
Aaindex (http://www.genome.ad.jp/dbget/) - Physicochemical properties of peptides
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
BioImage (http://www-embl.bioimage.org/) - Searchable database of multidimensional biological images
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

Annexure VII

Bioinformatics Software

Archives and Indices

- Bio Catalog (http://www.ebi.ac.uk/biocat/biocat.html) links to biotechnology-related non-commercial software available via the Internet
- Indiana University Repository for Biological Software - IUBIO ARCHIVE FOR BIOLOGY (ftp://iubio.bio.indiana.edu)- Indiana University (USA)

DNA Sequence Analysis Software

- AAT- Analysis and Annotation Tool (http://genome.cs.mtu.edu/aat/aat.html) - Identify genes by comparing cDNA and protein sequence databases
- AcaClone pDRAW32 (http://www.acaclone.com) - DNA cloning, analysis and visualisation software. freeware
- ACD/Protein Manager (http://www.acdlabs.com/products/peptide/prot_mgr.html) - Sequence Analysis tools
- BioinformatiX (http://www.xpogen.com) - Analysis of microarray data
- Biopendium (http://www.inpharmatica.co.uk) - Sequence, structure and function relationships for all gene products in the public domain
- BioTools (http://www.biotools.com) - DNA, protein sequence analysis, and chromatogram analysis
- BLAST (http://www.ncbi.nlm.nih.gov/BLAST/) - Pair wise sequence alignment
- CLUSTALW (www.ebi.ac.uk/clustalw/) - Multiple sequence alignment
- CURVES (http://www.ibpc.fr/UPR9080/Curindex.html) - Nucleic acid helical analysis program
- DNASIS (http://www.oligo.net/dnasis.htm) - MacOS and Windows-based program for gene analysis
- Double Digester: (ftp.cs.yale.edu/pub/double_digester) - Yale University (USA)
- Gene Modeller (Gm) (ftp.tigr.org/pub)
- FASTA (http://www.ebi.ac.uk/fasta33/) - Pair wise sequence alignment
- Gene Construction Kit(tm) 2 (http://www.sciquest.com) - A DNA manipulation, cloning design, and illustration tool
- Gene Inspector(tm) 1.5 (http://www.sciquest.com.) - Sequence analysis package
- GeneJockey (http://www.biosoft.com/biosoft) - Editing, manipulation, and analysis of nucleic acid and protein sequences
- GENSCAN (http://genes.mit.edu/GENSCAN.html) - Predicts complete gene structures in genomic sequences
- GLIMMER (www.tigr.org/software/glimmer/) - uses interpolated Markov models to find genes in microbial DNA
- Grail(tm) Toolkit. (http://www.apocom.com) - homology-based gene assembly
- ISYS(TM) (http://www.ncgr.org/isys) - capable of integrating data sources and analysis tools from an investigator's own laboratory
- Lasergene (http://www.dnastar.com) - sequencing, primer design, sequence alignment, databases and database searching, protein analysis, and restriction map analysis
- MZEF (http://sciclio.cshl.org/genefinder/) - predicts internal protein coding exons in genomic DNA sequences
- NIH-Repositories of Sequence Analysis Software (http://molbio.info.nih.gov/molbio/software.htm) - NIH (USA)
- Onto-Express (www.openchannelfoundation.org) - automates the process of annotating a set of expressed genes with genomic functions

- OMIGA (http://www.accelrys.com/) - multiple sequence alignments and contains over 50 DNA and protein analysis functions
- pDRAW32 (www.crosswinds.net/_acaclone) - DNA analysis software by AcaClone software
- 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)
- Plasmid (http://www.redasoft.com) - vector map drawing and sequence analysis
- SeqVerter (www.genestudio.com) - nucleotide sequence converter/viewer
- SIGNALSCAN (http://bimas.dcrt.nih.gov/molbio/signal/) - finds homologies in published signal sequences
- SRS (http://www.lionbioscience.com/) - database integration and biological information search system
- TESS (http://www.cbil.upenn.edu/tess/) - locating and displaying transcription factor binding sites in DNA sequences
- The Rosetta Resolver System (www.rii.com) - high-capacity data storage, retrieval and analysis of gene expression data
- The Wisconsin Package™ (http://www.accelrys.com) - Sequence analysis contains over 100 interrelated software programs.
- TurboBLAST (http://www.turbogenomics.com/) - Implementation of BLAST, prvides critical insight into the structure and function of genes and proteins
- Visual Cloning 2000 (http://www.redasoft.com) - vector map drawing and sequence analysis program

Molecular Modeling Software

- 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.
- Chime (http://www.mdlchime.com/chime/) - Interactively displays 2D and 3D molecules directly in Web pages.
- 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
- Modeler (http://guitar.rockefeller.edu/modeller/modeller.html) - Homology / comparative modeling of protein three-dimensional structures
- NAB (http://www.scripps.edu/case) - construct models of helical and non-helical nucleic acids
- Peptool (http://www.biotools.com/) - alignment, homology building, structure prediction
- 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
- 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
- Swiss PDB Viewer (http://us.expasy.org/spdbv/) - It is tightly linked to Swiss model, an automated homology modeling server
- Web lab viewer (http://molsim.vei.co.uk/weblab/) - Create and visualize 3D molecular structures
- WHAT IF (http://www.cmbi.kun.nl/gv/servers/WIWWWI/) - protein modeling, protein structure analysis, database handler, molecular graphics package
- Display and analysis of molecular model data (MSC)

Primer Designing Software

- Amplify (http://engels.genetics.wisc.edu/amplify) - Software for PCR - Ver.2.52b
- Array Designer 2 (http://www.premierbiosoft.com/) - Design hundreds of primers for DNA or oligonucleotide microarrays
- CODEHOP (http://blocks.fhcrc.org/blocks/codehop.html) - PCR primers designed from protein multiple sequence alignments

- Fast PCR (http://www.biocenter.helsinki.fi/bi/bare-1_html/oligos.htm) - PCR primer design, DNA and protein tools program
- GenomePRIDE (http://pride.molgen.mpg.de/genomepride.html) - Primer design program that designs PCR primers or long oligos on an annotated sequence
- MethPrimer (http://itsa.ucsf.edu/~urolab/methprimer/) - A program for designing bisulfite-conversion-based Methylation PCR Primer.
- NetPrimer (http://www.PremierBiosoft.com) - Most comprehensive free primer analysis program on the web
- 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.
- Primer3 (http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi) - Pick primers from a DNA sequence
- Primerfinder (http://eatworms.swmed.edu/~tim/primerfinder/) - Tool to design oligonucleotides suitable for PCR within any sequence
- Primer Premier 5 (http://www.PremierBiosoft.com/) - Primer design program
- X-Primer (http://alces.umn.edu/pv/pub/ComBin) - for designing PCR primers - the Virtual Genome Center (USA)

RNA Analysis

- CARD (rrna.uia.ac.be/card.html) - A computer program for drawing RNA secondary structure models - The rRNA WWW Server - University of Antwerp (Belgium)
- The RNAdraw Home Page (http://www.yk.rim.or.jp/~aisoai/index.html) - by Ole Matzura
- RNA Secondary Structure Prediction Software (*ftp.nrcbsa.bio.nrc.ca/pub/*) - NRC (Canada)
- STRUCTURELAB (http://www-lmmb.ncifcrf.gov/%7Ebshapiro/RNAstructure.html) - Computational Approaches to RNA Structure Analysis
- 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

- An Alphabetic List of Genetic Analysis Software- This list mainly contains computer software on genetic linkage analysis, marker mapping, and pedigree drawing.
- DELTA: *ftp* - DEscription Language for TAxonomy (Australia)
- FastDNAml (http://bioweb.pasteur.fr/seqanal/interfaces/fastdnaml.html) - Construction of phylogenetic trees of DNA sequences using maximum likelihood
- 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
- PHYLIP phylogeny package (http://evolution.genetics.washington.edu/phylip.html) - It is a *free* package of programs for inferring phylogenies
- 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.
- 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
- TreeTop (http://www.genebee.msu.su/services/phtree_reduced.html) - Phylogenetic Tree Prediction
- TreeView (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html) - Program for displaying and printing phylogenies

WEBPHYLIP (http://sdmc.krdl.org.sg:8080/~lxzhang/phylip/) - It is a web version of the original PHYLIP package

Annexure VIII

Bioinformatics Institutions

- Agricultural Biotechnology Center (EMBnet - Hungary) - http//www.abc.hu/
- Animal Genome Database (NIAI, Japan) - http//ws4.niai.affrc.go.jp/
- Australian National Genomic Information Service - http//www.angis.org.au
- Baylor College of Medicine - www.mouse-genome.bcm.tmc.edu
- Baylor College of Medicine Genome Center - http//www.hgsc.bcm.tmc.edu/
- Beatson Institute - http//www.vet.gla.ac.uk/beatson/
- Belgian EMBnet Node (EMBnet - Belgium) - http//www.be.embnet.org/
- BioBase (EMBnet - Denmark) - http//biobase.dk/
- Bioinformatics and Metabolic Engineering Laboratory at MIT - http//bioinformatics.mit.edu
- Bioinformatics Research Center - brc.mcw.edu
- Biomedical Center (EMBnet - Sweden) - http//www.bmc.uu.se/
- Biomolecular Engineering Research Center - http//bmerc-www.bu.edu/
- Biophysics & Bioinformatics Laboratory (University of Athens, Greece) - http//biophysics.biol.uoa.gr
- Biotechnology Dept. - National Cancer Institute - http//www.biotech.ist.unige.it/
- Brutlag Bioinformatics Group - http//motif.stanford.edu/
- Caltech Genome Research Laboratory - http//informa.bio.caltech.edu/
- Canadian Bioinformatics Resource - http//www.cbr.nrc.ca
- Cavalli Lab - http//lotka.stanford.edu/
- Cedars-Sinai Medical Center Molecular Genetics Labs - http//www.csmc.edu/genetics/korenberg/korenberg.html
- Center for Bioinformatics and Computational Genomics - http//www.nersc.gov/cbcg/cbcg.html
- Center for Genomics Research - http//www.cgr.ki.se
- Center for Molecular and Biomolecular Informatics (EMBnet - the Netherlands) - http//www.caos.kun.nl/
- Center for Scientific Computing (EMBnet - Finland) - http//www.csc.fi/molbio
- Centre for DNA Fingerprinting and Diagnostics (EMBnet - India) - http//www.in.embnet.org/
- Centre for Molecular and Biomolecular Informatics (EMBnet - the Netherlands) - http//www.cmbi.kun.nl/
- Centre of Bioinformatics (EMBnet - China) - http//www.cbi.pku.edu.cn/
- Centro de Ingenieria Genetica y Biotecnologia (EMBnet - Cuba) - http//www.cu.embnet.org/
- Centro Nacional de Biotecnologia (EMBnet - Spain) - http//www.cnb.uam.es/
- CEPH (Fondation Jean Dausset) - http//www.cephb.fr/
- Cold Spring Harbor Labs - http//www.cshl.org/
- Colorado State University - klab.agsci.colostate.edu
- Danish Centre for Human Genome Research - http//www.biokemi.au.dk/febs/index.html
- Division of Mathematical Biology - http//mathbio.nimr.mrc.ac.uk/
- DKFZ The Biocomputing Service Group, German Cancer Research Center - http//www.dkfz-heidelberg.de/index_e.htm
- DOE Joint Genome Institute - http//www.jgi-psf.org
- DOE Life Sciences Division - http//www.er.doe.gov/production/ober/HELSRD_top.html
- Drosophila Genome Center - Erasmus University MGC Dept. of Clinical Genetics - http//www.eur.nl/fgg/kgen/
- European Bioinformatics Institute - http//www.ebi.ac.uk/ebi_home.html
- European Center for Biotechnology Information - http//www.ecbi.org/
- European Molecular Biology Laboratory - http//www.embl-heidelberg.de/
- Fred Hutchinson Cancer Research Centre - http//www.fhcrc.org
- GeneBee (EMBnet - Russia) - http//www.genebee.msu.su
- Genestream - http//www2.igh.cnrs.fr/home.eng.html

- GENETHON Human Genome Research Centre - http//www.genethon.fr/genethon_en.html/
- GENIUSnet (EMBnet - Germany) - http//genome.dkfz-heidelberg.de/
- George M. Church Laboratory - http//twod.med.harvard.edu/
- Gesellschaft für Biotechnologische Forschung mbH - http//www.gbf.de
- Harvard University - Dept. of Molecular & Cellular Biology - http//golgi.harvard.edu/
- Hoffmann-La Roche Ltd. - http//www.roche.com
- Human Gene Nomenclature Committee - http//www.gene.ucl.ac.uk/nomenclature/
- IEN-AdR - CNR Area di Ricerca (EMBnet - Italy) - http//area.ba.cnr.it/
- Imperial Cancer Research Fund - http//www.lif.icnet.uk/
- INFOBIOGEN (EMBnet - France) - http//www.infobiogen.fr/
- Institute of Advanced Biomedical Technologies - http//www.itba.mi.cnr.it/
- Institute of Biochemistry and Biophysics (EMBnet Poland) - http//www.ibb.waw.pl
- Institute of Cancer Research - http//www.icr.ac.uk/
- Institute of Forest Genetics - http//dendrome.ucdavis.edu/ifg/index.html
- Institute of Molecular Biology and Biotechnology (EMBnet - Greece) - http//www.imbb.forth.gr/
- Institute of Molecular Pathology - Bioinformatics - http//mendel.imp.univie.ac.at
- Institute of Virology and Envirenmental Microbiology - http//www.nerc-oxford.ac.uk/
- Instituto de Bioquimica y Biologia Molecular - IBBM (EMBnet - Argentina) - http//sol.biol.unlp.edu.ar/embnet
- Instituto de Bioquimica y Biologia Molecular - IBBM (EMBnet - Argentina) - http//sol.biol.unlp.edu.ar/embnet
- Instituto de Investigaciones Biotecnologicas, Universidad Nacional de General San Martin - http//www.iib.unsam.edu.ar/genomelab
- Instituto Gulbenkian de Ciencia (EMBnet - Portugal) - http//www.igc.gulbenkian.pt
- International Center for Cooperation in Bioinformatics Network) - http//www.iccbnet.org
- Irish National Centre for BioInformatics (EMBnet - Ireland) - http//acer.gen.tcd.ie/
- Israeli National Node (EMBnet - Israel) - http//dapsas1.weizmann.ac.il/
- Japan Science and Technology Corporation - http//www.jst.go.jp/EN/
- JENA Genome Sequencing Centre - http//genome.imb-jena.de/
- John Innes Centre - http//www.jic.bbsrc.ac.uk/
- LANL Center for Human Genome Studies - http//www-ls.lanl.gov/
- LaTrobe University Department of Genetics and Human Variation - http//www.gen.latrobe.edu.au/
- Lawrence Berkeley National Laboratory Human Genome Sequencing Department - http//www-hgc.lbl.gov/GenomeHome.html
- Lawrence Livermore Human Genome Center - http//www.bio.llnl.gov/bbrp/genome/genome.html
- LLNL Biology and Biotechnology Research Program - http//www.bio.llnl.gov/bbrp/bbrp.homepage.html
- Mammalian Genetics Unit - http//www.mgu.har.mrc.ac.uk/
- Mascon Discovery Informatics - http//wwww.masconlifesciences.com
- Minnesota University Institute of Human Genetics - http//www.ihg.med.umn.edu/index.html
- Minnesota University Medical School, Computational Biology Centers - http//www.cbc.med.umn.edu/
- Missouri University - Maize Genome Database MaizeDB - http//www.agron.missouri.edu/
- Molecular Design Institute - http//mdi.ucsf.edu/
- Mouse Genome Centre - http//www.mgc.har.mrc.ac.uk/
- Mouse Genome Informatics - http//www.informatics.jax.org/
- MRC Human Genetics Unit - http//www.hgu.mrc.ac.uk/
- Munich Information Centre for Protein Sequences. - http//www.mips.biochem.mpg.de
- National Bioinformatics Node (EMBnet - Turkey) - http//www.tr.embnet.org/
- National Center for Biotechnology Infoformation - c

- National Center for Genome Resources - http//www.ncgr.org/
- National Institute for Medical Research - http//www.nimr.mrc.ac.uk/
- National Institute of Health - http//www.nih.gov/
- NCSA Illinois Computational Biology - http//www.ncsa.uiuc.edu/Apps/CB/
- Pasteur Institute - http//www.pasteur.fr/index-en.iphtml
- Resource Center/Primary Database of the German Human Genome Project - http//www.rzpd.de
- Rockefeller University Laboratory of Statistical Genetics - http//linkage.rockefeller.edu/
- Roslin Institute Bioinformatics - http//www.ri.bbsrc.ac.uk/bioinformatics/
- Rowett Research Institute - http//www.rri.sari.ac.uk/
- Salk Institute for Biological Studies - http//www.salk.edu/
- Sanger Centre - http//www.sanger.ac.uk/
- Scottish Crop Research Institute - http//www.scri.sari.ac.uk/
- South African National Bioinformatics Institute - http//www.sanbi.ac.za/
- South African Natitonal Bioinformatics Institute - http//www.sanbi.ac.za
- Stanford Human Genome Center - http//www-shgc.stanford.edu/
- Stanford University, Genetics Department - http//genome-www.stanford.edu/
- Swiss Institute for Experimental Cancer Research) - http//www.ch.embnet.org
- Swiss Institute of Bioinformatics (EMBnet - Switzerland) - http//www.ch.embnet.org/ SWISS-PROT - Switzerland - http//www.no.embnet.org/brochure/html/swiss.html Telethon Institute of Genetics and Medicine - http//www.tigem.it/TIGEM/tigem.html
- The Babraham Institute - http//www.bi.bbsrc.ac.uk/
- The Center for Biological Sequence Analysis - http//www.cbs.dtu.dk/
- The Galton Laboratory - http//www.gene.ucl.ac.uk/
- The Human Genome Organisation (HUGO) - http//www.hugo-international.org/hugo/
- The Institute of Molecular Biology and Biotechnology - http//www.imbb.forth.gr/about.html
- The International centre for genetic engineering and biotechnology - http//www.icgeb.trieste.it/
- The Merck Genome Research Institute - http//www.mgri.org/
- The MRC Centre for Protein Engineering - http//www.mrc-cpe.cam.ac.uk/
- The MRC Laboratory of Molecular Biology - http//www.mrc-lmb.cam.ac.uk/
- The Norwegian EMBnet Node (EMBnet - Norway) - http//www.no.embnet.org/
- The Sanger Centre - http//www.sanger.ac.uk/
- The Walter and Eliza Hall Institute of Medical Research - http//www.wehi.edu.au
- The Wellcome Trust Centre for Human Genetics - http//www.well.ox.ac.uk/
- UBC Bioinformatics Centre - http//bioinformatics.ubc.ca/
- UCL - http//www.biochem.ucl.ac.uk/bsm/dbbrowser/embnet.html
- UK Human Genome Mapping Project Resource Centre (EMBnet - United Kingdom) - http//www.hgmp.mrc.ac.uk/
- University of Kansas Medical Center - http//www.kumc.edu/gec
- University of Oklahoma, Advanced Center for Genome Technology - http//www.genome.ou.edu
- University of Pennsylvania Center for Bioinformatics - http//www.pcbi.upenn.edu/
- University of Washington, Dept of Pathology - http//www.pathology.washington.edu/
- University of Washingyon Genome Center - http//www.genome.washington.edu/UWGC/
- Universität Bielefeld - http//www.techfak.uni-bielefeld.de/GK635/forschung_engl.html
- Utah Genome Center - http//www.genome.utah.edu/
- Vienna Biocenter (EMBnet - Austria) - http//www.at.embnet.org/
- Washington University Genome Sequencing Center - http//genome.wustl.edu/gsc/
- Washington University Institute for Biomedical Computing - http//www.ibc.wustl.edu/
- Weizmann Institute of Science - Bioinformatics and Biological Computing - http//bioinfo.weizmann.ac.il
- Whitehead Institute for Biomedical Research - http//www.wi.mit.edu/
- Whitehead Institute for Biomedical, Research/MIT Center for Genome Research - http//www-genome.wi.mit.edu/

- XPERT GLOBAL TECH LIMITED - www.xpertglobaltech.com

Glossary

A

Accession number An identifier supplied by the curators of the major biological databases upon submission of a novel entry that uniquely identifies that sequence (or other) entry.

Adenine A purine base found in DNA and RNA

Agricultural biotechnology (AgBio)

The application of rDNA technology to agriculturally important plants and organisms.

Algorithm

A series of steps defining a procedure or formula for solving a problem, that can be coded into a programming language and executed. Bioinformatics algorithms typically are used to process, store, analyze, visualize and make predictions from biological data.

Alignment

Explicit mapping of characters of a sequence to characters of one or more other sequence(s)

Allele

A given form of a gene that occupies a specific position or locus on a chromosome. Variant forms of genes occurring at the same locus are said to be alleles of one another.

Alternative splicing

One of the alternate combinations of a folded protein that are possible due to by recombination of multiple gene segments during mRNA splicing that occurs in higher organisms.

Alternative splice-form

One of the possible alternate combinations of exons into a folded protein that are possible by recombining multiple gene segments during mRNA splicing in higher organisms.

Alu family

A common set of dispersed DNA sequences found throughout the human genome; each is about 300 bases long and they are repeated at least 500,000 times. Alu sequences are speculated to have originated from viral RNA sequences that integrated into human DNA thousands of years ago.

Analogy

Reasoning by which the function of a novel gene or protein sequence may be deduced from comparisons with other gene or protein sequences of known function. Identifying analogous or homologous genes via similarity searching and alignment is one of the chief uses of Bioinformatics. (See also alignment, similarity search.)

Annotation

A combination of comments, notations, references, and citations, either in free format or utilizing a controlled vocabulary, that together describe all the experimental and inferred information about a gene or protein. Annotations can also be applied to the description of other biological systems. Batch, automated annotation of bulk biological sequence is one of the key uses of Bioinformatics tools.

Anticodon

The triplet of contiguous bases on tRNA that binds to the codon sequence of nucleotides on mRNA. Example: GGG codes for Glycine.

Antisense

DNA or RNA composed of the complementary sequence to the target DNA/RNA. Also used to describe a therapeutic strategy that uses antisense DNA or RNA sequences to target specific gene DNA sequences or mRNA implicated in disease, in order to bind and physically inhibit their expression by physically blocking them.

Assembly

Compilation of overlapping sequences from one or more related genes that have been clustered together based on their degree of sequence identity or similarity. Sequence assembly may be used to piece together "shotgun" sequencing fragments (see shotgun sequencing) based upon overlapping restriction enzyme digests, or may be used to identify and index novel genes from "single-pass" cDNA sequencing efforts.

B

Base pair

A pair of nitrogenous bases (a purine and a pyrimidine), held together by hydrogen bonds, that form the core of DNA and RNA i.e the A:T, G:C and A:U interactions. adenine (A), cytosine (C), guanine (G), thymine (T) and uracil (U) or their post-replicationally or post-transcriptionally modified derivatives; short for ìnucleobase.î

Beta sheet
A three dimensional arrangement taken up by polypeptide chains that consists of alternating strands linked by hydrogen bonds. The alternating strands together form a sheet that is frequently twisted. One of the secondary structural elements characteristic of proteins.
Bit score
The value S' is derived from the raw alignment score S in which the statistical properties of the scoring system used have been taken into account. Because bit scores have been normalized with respect to the scoring system, they can be used to compare alignment scores from different searches.
BLOSUM
Blocks **Su**bstitution **M**atrix. A substitution matrix in which scores for each position are derived from *observations* of the frequencies of substitutions in blocks of local alignments in related proteins. Each matrix is tailored to a particular evolutionary distance. In the BLOSUM62 matrix, for example, the alignment from which scores were derived was created using sequences sharing no more than 62% identity. Sequences more identical than 62% are represented by a single sequence in the alignment so as to avoid over-weighting closely related family members.
Blunt-end (ligation)
The joining of DNA fragments that contain no overhang at either end and consequently no DNA bases available for hybridization (cf. sticky-end ligation).

C
Character
A fundamental unit of string
cDNA (complementary DNA)
A DNA strand copied from mRNA using reverse transcriptase. A cDNA library represents all of the expressed DNA in a cell.
cDNA library
A set of DNA fragments prepared from the total mRNA obtained from a selected cell, tissue or organism.
Chimeric clone
A cloning artifact created by a foreign gene being inserted into a vector in an incorrect orientation resulting in theexpression of a protein consisting of a fusion of two different gene products.
Chromat
Data file output from most popular DNA sequencers. Chromat files consist of the fluorescent traces generated by the sequencer for each of the four chemical bases, A, C, G, and T, together with the sequence and measures of the error in the traces at each sequence position.
Chromatin
The chromosome as it appears in its condensed state, composed of DNA and associated proteins (mainly histones).
Chromosome
The structure in the cell nucleus that contains all of the cellular DNA together with a number of proteins that compact and package the DNA.
Clone
A population of genetically identical cells or DNA molecules.
Cloning
The formation of clones or exact genetic replicas.
Cluster
The grouping of similar objects in a multidimensional space. Clustering is used for constructing new features which are abstractions of the existing features of those objects.
Coding regions (CDS)
The portion of a genomic sequence bounded by start and stop codons that identifies the sequence of the protein being coded for by a particular gene.
Codon
The three letter grouping in coding RNA sequence that selects the specific amino acid for incorporation into the protein.
Complementary determining region (CDR)
The hypervariable regions of an antibody molecule, consisting of three loops from the heavy chain and three from the light chain, that together form the antigen-binding site.
Complexity (of gene sequence)
The term "low complexity sequence" may be thought of as synonymous with regions of locally biased amino acid composition.

Computational Biology
The application of computational techniques to model or describe biological systems and test theories concerning their functions.
Configuration
(in software) The complete ordering and description of all parts of a software or database system. Configuration management is the use of software to identify, inventory and maintain the component modules that together comprise one or more systems or products.
Conformation
The precise three-dimensional arrangement of atoms and bonds in a molecule describing its geometry and hence its molecular function.
Consensus sequence
A single sequence delineated from an alignment of multiple constituent sequences that represents a "best fit" for all those sequences. A "voting" or other selection procedure is used to determine which residue (nucleotide or amino acid) is placed at a given position in the event that not all of the constituent sequences have the identical residue at that position.
Constitutive synthesis (expression)
Synthesis of mRNA and protein at an unchanging or constant rate regardless of a cellís requirements (see housekeeping genes).
Contig
A length of contiguous sequence assembled from partial, overlapping sequences, generated from a "shotgun" sequencing project. Contigs are typically created computationally, by comparing the overlapping ends of several sequencing reads generated by restriction enzyme digestion of a segment of genomic DNA.
Convergence
The end-point of any algorithm that uses iteration or recursion to guide a series of data processing steps. An algorithm is usually said to have reached convergence when the difference between the computed and observed steps falls below a pre-defined threshold.
Crystal structure
Term used to describe the high resolution molecular structure derived by x- ray crytallographic analysis of protein or other biomolecular crystals.

D
Data Cleaning
A process whereby automated or semi-automated algorithms are used to process experimental data, including noise, experimental errors and other artifacts, in order to generate and store high-quality data for use in subsequent analysis. Data cleaning is typically required in high-throughput sequencing where compression or other experimental artifacts limit the amount of sequence data generated from each sequencing run or "read."
Data Mining
The ability to query very large databases in order to satisfy a hypothesis ("top-down" data mining); or to interrogate a database in order to generate new hypotheses based on rigorous statistical correlations ("bottom-up" data mining).
Data Processing
Data processing is defined as the systematic performance of operations upon data such as handling, merging, sorting, and computing. The semantic content of the original data should not be changed, but the semantic content of the processed data may be changed.
Data Warehouses
Vast arrays of heterogeneous (biological) data, stored within a single logical data repository, that are accessible to different querying and manipulation methods.
Database
Any file system by which data gets stored following a logical process. (see also relational database)
ddNTP (dideoxy nucleoside triphosphate)
An individual DNA or RNA base that cannot be extended in the 3' direction
Deconvolution
Mathematical procedure to separate out the overlapping effects of molecules such as mixtures of compounds in a high-throughput screen, or mixtures of cDNAs in a high density array.
Deletion
A chromosomal alteration in which a portion of the chromosome or the underlying DNA is lost.

Deletion mapping
Process in which different deletions in a region of DNA are created and used to map the functionally critical areas of that DNA. e.g the minimal region of DNA required for a test promoter can be ascertained by systematic deletions in the region of interest.
Dendrogram
A graphical procedure for representing the output of a hierarchical clustering method. A dendrogram is strictly defined as a binary tree with a distinguished root, that has all the data items at its leaves. Conventionally, all the leaves are shown at the same level of the drawing. The ordering of the leaves is arbitrary, as is their horizontal position. The heights of the internal nodes may be arbitrary, or may be related to the metric information used to form the clustering.
Dimer
A composite molecule formed by the binding of two molecules (see homo and heterodimers).
DNA (deoxyribonucleic acid)
The chemical that forms the basis of the genetic material in virtually all organisms. DNA is composed of the four nitrogenous bases Adenine, Cytosine, Guanine, and Thymine, which are covalently bonded to a backbone of deoxyribose-phosphate to form a DNA strand. Two complementary strands (where all Gs pair with Cs and As with Ts) form a double helical structure which is held together by hydrogen bonding between the cognate bases.
DNA fingerprinting
A technique for identifying human individuals based on a restriction enzyme digest of tandemly repeated DNA sequences that are scattered throughout the human genome, but are unique to each individual.
DNA sequencing
The technique in which the specific sequence of bases forming a particular DNA region is deciphered.
DNase (Deoxyribonuclease)
One of a series of enzymes that can digest DNA.
Domain (protein)
A region of special biological interest within a single protein sequence. However, a domain may also be defined as a region within the three-dimensional structure of a protein that may encompass regions of several distinct protein sequences that accomplishes a specific function. A domain class is a group of domains that share a common set of well-defined properties or characteristics.
Drug discovery cycle
The cycle of events required to develop a new drug. Typically this involves research, preclinical testing and clinical development, and can take from 5 to 12 years.
DUST
A program for filtering low complexity regions from nucleic acid sequences.

E
Electronic Northerns
The use of an electronic database of cDNA sequences (or probes derived from them) in order to measure the relative levels of mRNAs expressed in different cells or tissues. An example of the use of an electronic Northern might be to identify the differences in the genes expressed in prostate cancer and those in benign prostate hyperplasia, by subtracting the database of one from the other and seeing which cDNAs remain.
Epigenomics
The study of complex expression networks or linkages both spatially (within the body) and temporally (at different times in development).
Exon
The region of DNA within a gene that codes for a polypeptide chain or domain. Typically a mature protein is composed of several domains coded by different exons within a single gene.
Expressed Sequence Tags (ESTs)
A small sequence from an expressed gene that can be amplified by PCR. ESTs act as physical markers for cloning and full length sequencing of the cDNAs of expressed genes. Typically identified by purifying mRNAs, converting to cDNAs, and then sequencing a portion of the cDNAs.
Expression (gene or protein)
A measure of the presence, amount, and time-course of one or more gene products in a particular cell or tissue. Expression studies are typically performed at the RNA (mRNA) or protein level in order to determine the number, type, and level of genes that may be up-regulated or down-regulated during a cellular process, in response to an external stimulus, or in sickness or disease. Gene chips and proteomics now allow the study of expression profiles of sets of genes or even entire genomes.

E value
Expectation value. The number of different alignents with scores equivalent to or better than S that are expected to occur in a database search by chance. The lower the E value, the more significant the score.
Expression profile
The level and duration of expression of one or more genes, selected from a particular cell or tissue type, generally obtained by a variety of high-throughput methods, such as sample sequencing, serial analysis, or microarray-based detection.
Expression vector
A cloning vector that is engineered to allow the expression of protein from a cDNA. The expression vector provides an appropriate promoter and restriction sites that allow insertion of cDNA.

F
FASTA
The first widely used algorithm for database similarity searching. The program looks for optimal local alignments by scanning the sequence for small matches called "words". Initially, the scores of segments in which there are multiple word hits are calculated ("init1"). Later the scores of several segments may be summed to generate an "initn" score. An optimized alignment that includes gaps is shown in the output as "opt". The sensitivity and speed of the search are inversely related and controlled by the "k-tup" variable which specifies the size of a "word". (Pearson and Lipman)
Filtering
Also known as Masking. The process of hiding regions of (nucleic acid or amino acid) sequence having characteristics that frequently lead to spurious high scores. See SEG and DUST.
Fingerprint
A fingerprint is a set of motifs used to predict the occurrence of similar motifs, in either an individual sequence or in a database. Fingerprints are refined by iterative scanning of a composite protein sequence database. A composite or multiple-motif fingerprint contains a number of aligned motifs taken from different parts of a multiple alignment. True family members are then easy to identify by virtue of possessing all elements of the fingerprint, while subfamily members may be identified by possessing only part of it.
Frameshift
A deletion, substitution, or duplication of one or more bases that causes the reading-frame of a structural gene to shift from the normal series of triplets.
Functional genomics
The use of genomic information to delineate protein structure, function, pathways and networks. Function may be determined by "knocking out" or "knocking in" expressed genes in model organisms such as worm, fruitfly, yeast or mouse.

G
Gaps (affine gaps)
A gap is defined as any maximal, consecutive run of spaces in a single string of a given alignment. Gaps help create alignments that better conform to underlying biological models and more closely fit patterns that one expects to find in meaningful alignment.
Gap penalties
The penalty applied to a similarity score for the introduction of an insertion or deletion gap, the extension of a gap, or both. Gap penalties are usually subtracted from a cumulative score being determined for the comparison of two or more sequences via an optimization algorithm that attempts to maximize that score.
Gene Index
A listing of the number, type, label and sequence of all the genes identified within the genome of a given organism. Gene indices are usually created by assembling overlapping EST sequences into clusters, and then determining if each cluster corresponds to a unique gene. Methods by which a cluster can be identified as representing a unique gene include identification of long open reading frames (ORFs), comparison to genomic sequence, and detection of SNPs or other features in the cluster that are known to exist in the gene.
GenBank
Data bank of genetic sequences operated by a division of the National Institutes of Health.
Gene
Classically, a unit of inheritance. In practice, a gene is a segment of DNA on a chromosome that encodes a protein and all the regulatory sequences (promoter) required to control expression of that protein.

Gene chips (also Gene arrays)
The covalent attachment of oligonucleotides or cDNA directly onto a small glass or silicon chip in organized arrays. Over 50,000 different DNA fragments can be presented on a single chip providing a high throughput parallel method of probing gene expression, genotype or gene function.
Gene expression
The conversion of information from gene to protein via transcription and translation.
Gene families
Subsets of genes containing homologous sequences which usually correlate with a common function.
Gene library
A collection of cloned DNA fragments created by restriction endonuclease digestion that represent part or all of an organismís genome.
Gene product
The product, either RNA or protein, that results from expression of a gene. The amount of gene product reflects the activity of the gene.
Gene therapy
The use of genetic material for therapeutic purposes. The therapeutic gene is typically delivered using recombinant virus or liposome based delivery systems.
Genetic code
The mapping of all possible codons into the 20 amino acids including the start and stop codons.
Genetic engineering (Recombinant DNA technology)
The procedures used to isolate, splice and manipulate DNA outside the cell. Genetic Engineering allows a recombinantly engineered DNA segment to be introduced into a foreign cell or organism, and be able to replicate and function normally.
Genetic marker
Any gene that can be readily recognized by its phenotypic effect, and which can be used as a marker for a cell, chromosome, or individual carrying that gene. Also, any detectable polymorphism used to identify a specific gene.
Genome
The complete genetic content of an organism.
Genomic DNA (sequence)
DNA sequence typically obtained from mammalian or other higher-order species, which includes both intron and exon sequence (coding sequence), as well as non-coding regulatory sequences such as promoter, and enhancer sequences.
Genomics
Use f complete genomes to explain and interpret biological phenomena.
Genotype
Strictly, all of the genes possessed by an individual. In practice, the particular alleles present in a specific genetic locus.
Global Alignment
The alignment of two nucleic acid or protein sequences over their entire length.

H
H is the relative entropy of the target and background residue frequencies. H can be thought of as a measure of the average information (in bits) available per position that distinguishes an alignment from chance. At high values of H, short alignments can be distinguished by chance, whereas at lower H values, a longer alignment may be necessary.
Hairpin
A double-helical region in a single DNA or RNA strand formed by the hydrogen-bonding between adjacent inverse complementary sequences to form a hairpin shaped structure.
Haploid
A cell or organism containing only one set of chromsomes without the homologous pairs. (cf. diploid)
Heterodimer
Protein composed of 2 different chains or subunits**.**
Heteroduplex
Hybrid structure formed by the annealing of two DNA strands (or an RNA and DNA) that have sufficient complementarity in their sequence to allow hydrogen bonding.

Hidden Markov model (HMM)
A joint statistical model for an ordered sequence of variables. The result of stochastically perturbing the variables in a Markov chain (the original variables are thus "hidden"), where the Markov chain has discrete variables which select the "state" of the HMM at each step..
High-throughput screening
The method by which very large numbers of compounds are screened against a putative drug target in either cell-free or whole-cell assays. Typically, these screenings are carried out in 96 well plates using automated, robotic station based technologies or in higher- density array ("chip") formats.
Homeobox
A highly conserved region in a homeotic gene composed of 180 bases (60 amino acids) that specifies a protein domain (the homeodomain) that serves as a master genetic regulatory element in cell differentiation during development in species as diverse as worms, fruitflies, and humans.
Homeotic gene
A gene that controls the activity of other genes involved in the development of a body plan. Homeotic genes have been found in organisms ranging from plants to humans.
Homology
Rlationship by evolutionary descent from a common ancestral precursor.
Housekeeping genes
Genes that are always expressed (ie. they are said to be constitutively expressed) due to their constant requirement by the cell.
HSP
High-scoring segment pair. Local alignments with no gaps that achieve one of the top alignment scores in a given search.
Hydrogen bond
A weak chemical interaction between an electronegative atom (e.g. nitrogen or oxygen) and a hydrogen atom that is covalently attached to another atom. This bond maintains the two-helices of DNA together and is also the primary interaction between water molecules.
Hydrophilicity
(lit. water-loving) The degree to which a molecule is soluble in water. Hydrophilicity depends to a large degree on the charge and polarizability of the molecule and its ability to form transient hydrogen-bonds with (polar) water molecules.
Hydrophobicity
(lit. water-hating) The degree to which a molecule is insoluble in water, and hence is soluble in lipids. If a molecule lacking polar groups is placed in water, it will be entropically driven to finding a hyrdophobic environment (such as the interior of a protein or a membrane).

I
Identity
The extent to which two (nucleotide or amino acid) sequences are invariant.
Introns
Nucleotide sequences found in the structural genes of eukaryotes that are non-coding and interrupt the sequences containing information that codes for polypeptide chains
Isoschizomers
Two different restriction enzymes which recognize and cut DNA at the same recognition site. e.g Sma I and Xma I both recognize and cut the sequence CCCGGG.
Isozymes
Two or more enzymes capable of catalyzing the same reaction but varying in their specificity due to differences in their structures and hence their efficiencies under different environmental conditions.

J
Junk DNA
Term used to describe the excess DNA that is present in the genome beyond that required to encode proteins.

K
A statistical parameter used in calculating BLAST scores that can be thought of as a natural scale for search space size. The value K is used in converting a raw score (S) to a bit score (S').
Karyotype

The constitution (typically number and size) of chromosomes in a cell or individual.

L

"Lab on a chip"

Term describing microdevices that allow rapid, microanalytical analysis of DNA or protein in a single, fully integrated system. Typically, these devices are miniature surfaces, made of silicon, glass or plastic, which carry the necessary microdevices (pumps, valves, microfluidic controllers, and detectors) that allow sample separation and analysis. These devices are used in drug discovery, genetic testing and separation science.

lambda

A statistical parameter used in calculating BLAST scores that can be thought of as a natural scale for scoring system. The value lambda is used in converting a raw score (S) to a bit score (S').

Leucine zipper

Protein motif which binds DNA in which 4-5 Leucines are found at 7 amino acid intervals. This motif is present typically in transcription factors and other proteins that bind DNA.

Lexicon

In Bioinformatics, a lexicon refers to a pre-defined list of terms that together completely define the contents of a particular database.(strict.) The component in the grammar which is in bare form a list of words or lexical entries.

Ligand

Any small molecule that binds to a protein or receptor; the cognate partner of many cellular proteins, enzymes, and receptors.

Linkage

The association of genes (or genetic loci) on the same chromosome. Genes that are linked together tend to be transmitted together.

Linkage map

A genetic map of a chromosome or genome delineated by mapping the positions of genes to their chromosomes by their linkage to readily identifiable genetic loci.

Locus

The specific position occupied by a gene on a chromosome. At a given locus, any one of the variant forms of a gene may be present. The variants are said to be alleles of that gene.

Local Alignment

The alignment of some portion of two nucleic acid or protein sequences

Low Complexity Region (LCR)

Regions of biased composition including homopolymeric runs, short-period repeats, and more subtle overrepresentation of one or a few residues. The SEG program is used to mask or filter LCRs in amino acid queries. The DUST program is used to mask or filter LCRs in nucleic acid queries.

M

Markov chain

Any multivariate probability density whose independence diagram is a chain.

Masking

Also known as Filtering. The removal of repeated or low complexity regions from a sequence in order to improve the sensitivity of sequence similarity searches performed with that sequence.

Matrix

a 2-dimensional array, typically of numbers

Meiosis

A process within the cell nucleus that results in the reduction of the chromosome number from diploid (two copies of each chromosome) to haploid (a single copy) through two reductive divisions in germ cells.

Melting (of DNA)

The denaturation of double-stranded DNA into two single strands by the application of heat. (Denaturation breaks the hydrogen bonds holding the double-stranded DNA together).

Messenger RNA (mRNA)

The complementary RNA copy of DNA formed from a single-stranded DNA template during transcription that migrates from the nucleus to the cytoplasm where it is processed into a sequence carrying the information to code for a polypeptide domain.

Methylation

The addition of -CH3 (methyl) groups to a target site. Typically such addition occurs on to the cytosine bases of DNA. (see maternal imprinting).

Microarray
A 2D array, typically on a glass, filter, or silicon wafer, upon which genes or gene fragments are deposited or synthesized in a predetermined spatial order allowing them to be made available as probes in a high-throughput, parallel manner.
Motif
A short conserved region in a protein sequence. Motifs are frequently highly conserved parts of domains.
Modeling
In bioinformatics, modeling usually refers to molecular modeling, a process whereby the three-dimensional architecture of biological molecules is interpreted (or predicted), visually represented, and manipulated in order to determine their molecular properties. (general)
Multiple (sequence) alignment
A Multiple Alignment of k sequences is a rectangular array, consisting of characters taken from the alphabet A, that satisfies the following conditions: There are exactly k rows; ignoring the gap character, row number i is exactly the sequence s_i; and each column contains at least one character different from "-". In practice multiple sequence alignments include a cost/weight function, that defines the penalty for the insertion of gaps (the "-" character) and weights identities and conservative substitutions accordingly. Multiple alignment algorithms attempt to create the optimal alignment defined as the one with the lowest cost/weight score.
Multiplex sequencing
Approach to high-throughput sequencing that uses several pooled DNA samples run through gels simultaneously and then separated and analyzed.

N
Naked DNA
Pure, isolated DNA devoid of any proteins that may bind to it.
Nested PCR
The second round amplification of an already PCR-amplified sequence using a new pair of primers which are internal to the original primers. Typically done when a single PCR reaction generates insufficient amounts of product.
Neural net
A neural net is an interconnected assembly of simple processing elements, units or nodes, whose functionality is loosely based on the animal brain. The processing ability of the network is stored in the inter-unit connection strengths, or weights, obtained by a process of adaptation to, or learning from, a set of training patterns.
Northern blotting
A technique to identify RNA molecules by hybridization that is analogous to Southern blotting (see Southern blotting).
Nucleoside
A five-carbon sugar covalently attached to a nitrogen base.
Nucleotide
A nucleic acid unit composed of a five carbon sugar joined to a phosphate group and a nitrogen base.

O
Object-Relational Database
Object databases combine the elements of object orientation and object-oriented programming languages with database capabilities. They provide more than persistent storage of programming language objects.
Open reading frame (ORF)
Any stretch of DNA that potentially encodes a protein. Open reading frames start with a start codon, and end with a termination codon. No termination codons may be present internally. The identification of an ORF is the first indication that a segment of DNA may be part of a functional gene.
Operon
A unit of transcription consisting of one or more structural genes, an operator, and a promoter.
Optimal Alignment
An alignment of two sequences with the highest possible score.
Ortholog
Orthologs are genes in different species that evolved from a common ancestral gene by speciation. Normally, orthologs retain the same function in the course of evolution.
Overlapping clones

Collection of cloned sequences made by generating randomly overlapping DNA fragments with infrequently cutting restriction enzymes.

P

P value

The probability of an alignment occurring with the score in question or better. The p value is calculated by relating the observed alignment score, S, to the expected distribution of HSP scores from comparisons of random sequences of the same length and composition as the query to the database. The most highly significant P values will be those close to 0. P values and E values are different ways of representing the significance of the alignment.

PAM

Percent Accepted Mutation. A unit introduced by Dayhoff et al. to quantify the amount of evolutionary change in a protein sequence. 1.0 PAM unit, is the amount of evolution which will change, on average, 1% of amino acids in a protein sequence. A PAM(x) substitution matrix is a look-up table in which scores for each amino acid substitution have been calculated based on the frequency of that substitution in closely related proteins that have experienced a certain amount (x) of evolutionary divergence.

Palindrome

A region of DNA with a symmetrical arrangement of bases occuring about a single point such that the base sequences on either side of that point are identical (if the strands are both read in the same direction) e.g 5í GAATTC 3í whose complementary sequence is 3í CTTAAG 5í.

Pattern

Molecular biological patterns usually occur at the level of the characters making up the gene or protein sequence. A pattern language must be defined in order to apply different criteria to different positions of a sequence. In order to have position-specific comparison done by a computer, a pattern-matching algorithm must allow alternative residues at a given position, repetitions of a residue, exclusion of alternative residues, weighting, and ideally, combinatorial representation.

Pathways

Bioinformatics strives to define representations of key biological datatypes, algorithms and inference procedures, including sequences, structures, biological pathways and reactions.

Paralog

Paralogs are genes related by duplication within a genome. Orthologs retain the same function in the course of evolution, whereas paralogs evolve new functions, even if these are related to the original one.

Pharmacogenomics

The use of (DNA-based) genotyping in order to target pharmaceutical agents to specific patient populations. Genetic differences are known to affect responses to many types of drug therapy, and pharmacogenomics analysis serves to customize the use of pharmaceuticals for specific subgroups of patients.

Pharmacophore

The three dimensional spatial arrangment of atoms, substituents, functional groups, or chemical features that together are sufficient to describe the pharmacologically active components of a drug molecule or molecule series.

Polygenic inheritance

Inheritance involving alleles at many genetic loci.

Polymerase chain reaction (PCR)

Technique used to amplify or generate large amounts of replica DNA of a segment of any DNA whose "flanking" sequences are known.

Positional cloning

Method used to define the location of a gene on a chromosome and use this information to identify and clone the gene.

Primary sequence (protein)

The linear sequence of a polypeptide or protein.

Primer

A short oligonucleotide that provides a free 3í hydroxyl for DNA or RNA synthesis by the appropriate polymerase (DNA polymerase or RNA polymerase).

Profile

A table that lists the frequencies of each amino acid in each position of protein sequence. Frequencies are calculated from multiple alignments of sequences containing a domain of interest. See also PSSM.

Promoter (site)

A promoter site is defined by its recognition by eukaryotic RNA polymerase II; its activity in a higher eukaryote; by experimentally evidence, or homology and sufficient similarity to an experimentally defined promoter; and by observed biological function.

Protein families
Sets of proteins that share a common evolutionary origin reflected by their relatedness in function which is usually reflected by similarities in sequence, or in primary, secondary or tertiary structure. Subsets of proteins with related structure and function.
Proteome
The entire protein complement of a given organism.
Proteomics
The study of the proteome. Typically, the cataloging of all the expressed proteins in a particular cell or tissue type, obtained by identifying the proteins from cell extracts using a combination of 2D gel electrophoresis and mass spectrometry. The large scale analysis of the protein composition and function. (cf genomics)
PSI-BLAST
Position-**S**pecific **I**terative BLAST. An iterative search using the BLAST algorithm. A profile is built after the initial search, which is then used in subsequent searches. The process may be repeated, if desired with new sequences found in each cycle used to refine the profile.
PSSM
Position-specific scoring matrix; see profile. The PSSM gives the log-odds score for finding a particular matching amino acid in a target sequence.

Q
Query (sequence)
A DNA, RNA of protein sequence used to search a sequence database in order to identify close or remote family members (homologs) of known function, or sequences with similar active sites or regions (analogs), from whom the function of the query may be deduced.

R
Rational drug design (Structure based drug design)
The development of drugs based on the 3-dimensional molecular structure of a particular target.
Raw Score
The score of an alignment, S, calculated as the sum of substitution and gap scores. Substitution scores are given by a look-up table (see PAM, BLOSUM). Gap scores are typically calculated as the sum of G, the gap opening penalty and L, the gap extension penalty. For a gap of length n, the gap cost would be G+Ln. The choice of gap costs, G and L is empirical, but it is customary to choose a high value for G (10-15)and a low value for L (1-2).
Reading frame
A sequence of codons beginning with an intiation codon and ending with a termination codon, typically of at least 150 bases (50 amino acids) coding for a polypeptide or protein chain (see ORF and URF).
Recombinant DNA (rDNA)
DNA molecules resulting from the fusion of DNA from different sources. The technology employed for splicing DNA from different sources and for amplifying the resultant heterogenous DNA.
Recursion
An algorithmic procedure whereby an algorithm calls on itself to perform a calculation until the result exceeds a threshold, in which case the algorithm exits. Recursion is a powerful procedure with which to process data and is computationally quite efficient.
Relational Database
A database that follows E. F. Coddís 11 rules, a series of mathematical and logical steps for the organization and systemization of data into a software system that allows easy retrieval, updating, and expansion. An RDBMS stores data in a database consisting of one or more tables of rows and columns.
Relational Database Management Systems (RDBMS)
A software system that includes a database architecture, query language, and data loading and updating tools and other ancillary software that together allow the creation of a relational database application.
Repeats (repeat sequences)
Repeat sequences and approximate repeats occur throughout the DNA of higher organisms (mammals).
Repetitive elements

Repetitive elements provide important clues about chromosome dynamics, evolutionary forces, and mechanisms for exchange of genetic information between organisms

Restriction enzyme (restriction endonuclease)
A type of enzyme that recognizes specific DNA sequences (usually palindromic sequences 4, 6, 8 or 16 base pairs in length) and produces cuts on both strands of DNA containing those sequences only. The "molecular scissors" of rDNA technology.
Restriction fragment length polymorphisms (RFLPs)
Variation within the DNA sequences of organisms of a given species that can be identified by fragmenting the sequences using restriction enzymes, since the variation lies within the restriction site. RFLPs can be used to measure the diversity of a gene in a population.
Reverse Genetics
The use of protein information to elucidate the genetic sequence encoding that protein. Used to describe the process of gene isolation starting with a panel of afflicted patients (see positional cloning)**.**
Reverse transcriptase-PCR (RT-PCR)
Procedure in which PCR amplification is carried out on DNA that is first generated by the conversion of mRNA to cDNA using reverse transcriptase.
Ribonucleic acid (RNA)
A category of nucleic acids in which the component sugar is ribose and consisting of the four nucleotides Thymidine, Uracil, Guanine, and Adenine. The three types of RNA are messenger RNA (mRNA), transfer RNA (tRNA) and ribosomal RNA (rRNA).

S
Secondary structure (protein)
The organization of the peptide backbone of a protein that occurs as a result of hydrogen bonds e.g alpha helix, Beta pleated sheet.
SEG
A program for filtering low complexity regions in amino acid sequences. Residues that have been masked are represented as "X" in an alignment. SEG filtering is performed by default in the blastp subroutine of BLAST 2.0.
Selectivity
Selectivity of bioinformatics similarity search algorithms is defined as the significance
Sense strand
The strand of double-stranded DNA that acts as the template strand for RNA synthesis. Typically only one gene product is produced per gene, reading from the sense strand only. (Some viruses have open reading frames in both the sense and the antisense strands).
Sensitivity
Sensitivity of bioinformatics similarity search algorithms centers around two areas:
Sequence
An ordered succession of characters, units, or symbols, *i.e.* a string.
Similarity (homology) search
Given a newly sequenced gene, there are two main approaches to the prediction of structure and function from the amino acid sequence. Homology methods are the most powerful and are based on the detection of significant extended sequence similarity to a protein of known structure, or of a sequence pattern characteristic of a protein family.
Signal sequence (leader sequence)
A short sequence added to the amino-terminal end of a polypeptide chain that forms an amphipathic helix allowing the nascent polypeptide to migrate through membranes such as the endoplasmic reticulum or the cell membrane. It is cleaved from the polypeptide after the protein has crossed the membrane.
Single nucleotide polymorphisms (SNPs)
Variations of single base pairs scattered throughout the human genome that serve as measures of the genetic diversity in humans.
Single-pass sequencing
Rapid sequencing of large segments of the genome of an organism by isolating as many expressed (cDNA) sequences as possible and performing single sequencer runs on their 5í or 3í ends.
Site

Sites in sequences can be located either in DNA (e.g. binding sites, cleavage sites) or in proteins. In order to identify a site in DNA, ambiguity symbols are used to allow several different symbols at one position. Proteins, however, need a different mechanism (see Pattern). Restriction enzyme cleavage sites, for instance, have the following properties: limited length (typically, less than 20 base pairs); definition of the cleavage site and its appearance (3', 5' overhang or blunt); definition of the binding site.

Structural gene

Gene which encodes a structural protein (cf. Regulatory gene).

Structure prediction

Algorithms that predict the secondary, tertiary and sometimes even quarternary structure of proteins from their sequences.

Substitution

The presence of a non-identical amino acid at a given position in an alignment. If the aligned residues have similar physico-chemical properties the substitution is said to be "conservative".

Substitution matrix

A model of protein evolution at the sequence level resulting in the development of a set of widely used substitution matrices. These are frequently called Dayhoff, MDM (Mutation Data Matrix), BLOSUM or PAM (Percent Accepted Mutation) matrices. They are derived from global alignments of closely related sequences. Matrices for greater evolutionary distances are extrapolated from those for lesser ones.

T

Tentative Consensus (TC)

The identification of a sequence from an EST cluster that represents part or all of a complete gene.

Tentative Human Consensus sequences (THCs)

A consensus sequence generated from human EST fragments. THCs may be validated by comparison against databases of known human gene sequences, human genomic sequences, or by identification of the ORFs or other sequence features contained within the consensus as belonging to a known human gene product.

Tertiary structure

Folding of a protein chain via interactions of its sideschain molecules including formation of disulphide bonds between cysteine residues.

Transcription

The assembly of complementary single-stranded RNA on a DNA template.

Transcription factors

A group of regulatory proteins that are required for transcription in eukaryotes. Transcription factors bind to the promoter region of a gene and facilitate transcription by RNA polymerase.

Transfer RNA (tRNA)

A small RNA molecule that recognizes a specific amino acid, transports it to a specific codon in the mRNA, and positions it properly in the nascent polypeptide chain.

Translation

The process of converting RNA to protein by the assembly of a polypeptide chain from an mRNA molecule at the ribosome.

U

Unidentified reading frame (URF)

An open reading frame encoding a protein of undefined function**.**

Unitary Matrix

Also known as Identity Matrix. A scoring system in which only identical characters receive a positive score.

Uracil

Nitrogenous pyrimidine base found in RNA but not DNA.

V

Variable numbers of tandem repeats (VNTRs)

DNA sequence blocks of 2-60 base pairs which are repeated from two to more than 20 times in different individuals.

Variation (genetic)

Variation in genetic sequences and the detection of DNA sequence variants genome-wide allow studies relating the distribution of sequence variation to a population history. This in turn allows one to determine the density of SNPS or other markers needed for gene mapping studies.

Vector

Any agent that transfers material (typically DNA) from one host to another. Typically DNA vectors are autonomous DNA elements (such as plasmids) that can be manipulated and integrated into a hostís DNA or recombinant viruses.

Virtual libraries

The creation and storage of vast collections of molecular structures in an electronic database.

Visualization

Visualization is the process of representing abstract scientific data as images that can aid in understanding the meaning of the data.

W

Weight matrix

The density of binding sites in a gene or sequence can be used to derive a ratio of density for each element in a pattern of interest.

Wild type

Form of a gene or allele that is considered the "standard" or most common.

X

X chromosome

In mammals, the sex chromosome that is found in two copies in the homogametic sex (female in humans) and one copy in the hererogametic sex (male in humans).

Y

Yeast 2-hybrid system

A yeast-based method used to simultaneously identify, and clone the gene for, proteins interacting with a known protein.

Z

Z-DNA

A conformation of DNA existing as a left-handed double helix (the phosphate-sugar backbone forms a left-handed zig-zag course), which may play a role in gene regulation.

Zinc fingers

A protein motif formed by the interaction of repeated cysteine and histidine residues with a zinc ion. The spacing of the repeats results in finger like arrangements of the protein loops formed from the interaction which interact with DNA. These motifs are typically found in transcription factors.