Introduction To Bioinformatics

20-22 November, 2002

RECENT TRENDS IN BIOINFORMATICS

Dr. T. MADHAN MOHAN,
Director (Bioinformatics) Department of Biotechnology,
Govt. of India, New Delhi

The astronomical growth of biological information through large scale sequencing of genomes has resulted in an unprecedented synergism between information technology and molecular biology. Constraints of cost, time and uncertainty that are associated with 'wet' molecular biology studies and data-intensive techniques like x-ray crystallography has made in-silicon approaches more and more popular among biologists today. Present scope of Bioinformatics encompass the wide area of data acquisition, storage, retrieval, analysis and interpretation along with parallel development of application tools in terms of computational infrastructure, software, databases and so on.

Advancing Beyond the Genome Project:

With the completion of the draft sequence of the human genome and the completion of the sequencing of a number of genomes from a variety of organisms, the main challenge before the bioinformatician is to analyse and interpret the information. In this regard, the following issues are of particular importance: -

- High Throughput Genome Assembly
- Annotation
- Comparative Modeling and Assignment of Protein Folds
- Prediction of Structure, Dynamics and Thermodynamics
- Advanced biotechnological applications e.g. drug design, gene therapy etc.
- System modeling

Applications of Bioinformatics:

Present day application of Bioinformatics is diverse. It ranges from studies of evolution of life on Earth to generation of designer drugs. Sequence analysis focus upon the finding of new genes, analyse structure of the gene to determine its function and correlate how an altered structure of a gene can be linked with diseases. Molecular modeling studies attempts to understand how the three-dimensional topography of a protein is related to its function. Other complex applications include modeling of cell signaling and metabolic pathways, studying protein-protein interaction, understand mechanisms how protein families evolve and map the expression pattern of a plethora of genes in different cells and tissues.

The Genome Databases:

In the international vista, Bioinformatics is progressing at an astonishing rate. The major thrust areas seem to be the acquisition of sequence data, incorporating them in form of classified databases, integrating sequence information with structure data, developing tools for effective data-mining and developing a common platform for resource sharing and integration. In this last area, the outcome had been the International Nucleotide Sequence Database Collaboration, involving the three major genome databases of the world, GenBank, EMBL and DDBJ. The INSDC comprise of:

1. The taxonomy project – for using a uniform taxonomy for all databases
2. The Feature table – giving shared rules to allow the database to be exchanged among the three organizations
3. db_xref qualifier – explicit referencing of specific sequence within the database
4. Country qualifier – country of origin of the sequence.

In order to achieve the multitude of objectives on the global scale, growth of public domain databases have been particularly instrumental. Today we can retrieve an unimaginable amount of information on virtually any aspect of cellular biology from the Internet, whether it is bibliographic or genomic or structural or even functional. This information acts as the seed data for valuable downstream research. The public domain, apart from providing the basic information, also fortunately provides the utilities required for analysing and interpreting such data. It ranges from multiple sequence alignment to virtual gene expression studies, electronic PCR and so on. Some of the major molecular biology servers operating worldwide include:

- National Centre for Biotechnology Information: http://www.ncbi.nlm.nih.gov
- European Bioinformatics Institute: http://www.ebi.ac.uk
- DNA Databank of Japan: http://www.ddbj.nig.ac.jp
- ExPAsy Molecular Biology Server: http://www.expasy.ch
- Australian National Genome Information Service: http://morgan.angis.su.oz.au
- The Protein Data Bank: http://www.rcsb.org/pdb

An exhaustive list of public domain resources has been recently compiled by the Bioinformatics Centre of the Institute of Microbial Technology, Chandigarh and is available in its server at http://imtech.res.in/pdsb/

The Many Faces of Genome Analysis:

The availability of genome information provides the bioinformatician with a new set of challenges. This is to analyse the seemingly fragmentary body of knowledge. Currently, the predominant areas of Bioinformatics data analysis include:

- Sequence Alignment Studies
- Prediction of Protein Structure

The two areas have within them the commonalties and differences that one might expect.

Sequence Analysis studies*:* Sequence analysis studies are of two major types namely pair wise alignment e.g. one that is experienced through BLAST searches and multiple alignment studies e.g. one experienced with programmes like Clustal. In all cases the idea is to find the similarity or differences between a set of sequences and attempting to infer how they arose and how they are changing. Sequence analysis is an extraordinary tool for studying evolutionary relationships among genomes, gene duplication, splicing and so on.

***Protein Structure Prediction**: -* Predicting the structure of a protein from the sequence information is one of the most exciting areas of Bioinformatics. The crux of the process lies in the assignment of folds and domains in the primary structure, thereby developing an acceptable model of tertiary and quaternary level structures. Lately, several methods are available e.g. comparative modeling, threading methods, *ab initio* methods and use of genetic algorithms.

Mining from the Genome – In Trail of the Treasure

A well-equipped Bioinformatics laboratory in place, the fortunes to follow are stupendous. Euphemistically called genome data mining, the method uses traditional methods like sequence similarity studies, multiple alignment etc. coupled with more complex developments such as Serial Analysis of Gene Expression [SAGE], electronic PCR, microarray informatics and so on. Some of the areas of application include:

- Gene identification
- Drug discovery
- Phylogenomics
- Detection of genomic markers and polymorphism
- Understanding gene expression profiles
- Exploring new metabolic and regulatory pathways
- Assigning functions to unknown ESTs
- Understanding protein-protein interactions.

High Performance Computing – Adding Tooth to a Blunted Knife:

With the exponential growth of genome information and the need to accomplish quick analysis, the once extraordinary power of serial computation is beginning to falter. Given the present sizes of the known genomes, an operation such as whole genome comparison is likely to take unacceptably long times. This constraint has resulted in the evolution of the concept of parallel computation. Parallelisation efficiently scales up the process by cutting down computation time.

The Indian Side of Bioinformatics:

India embarked upon major national initiative for Bioinformatics studies. Leading the bandwagon is the Department of Biotechnology's Biotechnology Information System of India. Spread out as a distributed resource across the country with about 61 centres, the network is expected to bolster the Indian attempt to harness the deluge of biological information. Other programmes include the High Performance Computing Initiatives at the Centre for Development of Advanced Computing, Pune, where parallel computing is used to address problems of evolutionary biology, large-scale genome comparison and biological system modeling. Super computing facility has been established at IIT, Delhi for the promotion of in-silico drug development.

The Biotechnology Information System Network:

Structure: - The network comprises of 10 Distributed Information Centres [DICs] and 50 Distributed Information Sub-Centres [DISCs]. The entire system is headed and coordinated by an Apex Biotechnology Information Centre at the DBT Headquarters in New Delhi.

Activities: - The major activities of the Bioinformatics Centres are

- To provide a national bioinformation network to cover the diverse areas of the multidisciplinary areas of biotechnology
- To develop information resources, develop databases, information handling tools and techniques
- To establish information linkages with the international organizations
- To evolve programmes of education and implement human resource development in Bioinformatics
- To undertake research and development activities in the field of Bioinformatics.

State-of-the Art: Aided by a highly sophisticated communication backbone and six Interactive Graphics Facilities for Molecular modeling the network has been instrumental in the development of important databases and software. Four long term Advanced Diploma Courses in Bioinformatics at the Post-M.Sc. level is currently operational at Madurai, Pune, Calcutta and New Delhi for increasing the production of trained personnel in the area. The network also maintains an array of mirror sites of some of the major public domain databases for the benefit of practicing scientists. M.Sc, M.Tech and Ph.D. programmes in Bioinformatics have also been introduced this year.

As we gear up for the post-genomic era, the BTISnet is poised to play an increasingly crucial role.

Biological databases

N. RAVI KUMAR,
SCIENTIST (COMPUTER APPLICATIONS)
CPCRI, KASARAGOD

Introduction

1.1 What is Bioinformatics?

Bioinformatics has been defined in many ways. The most precise, in our opinion is "*Bioinformatics is conceptualizing biology in terms of molecules (in the sense physical-chemistry) and then applying "informatics" techniques (derived from disciplines such as applied Math, CS and Statistics) to understand and organize the information associated with these molecules, on a large-scale*."

-by Mark Gerstein

Bioinformatics is the field of science in which biology, computer science, and information technology merge into a single discipline. 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. There are three important sub-disciplines within bioinformatics:

i. Development of new algorithms and statistics with which to assess relationships among members of large data sets.

ii. Analysis and interpretation of various types of data including nucleotide and amino acid sequences, protein domains and protein structures.

iii. Development and implementation of tools that enable efficient access and management of different types of information.

1.2. Database implementations

Database is developed in relational database management system (RDBMS), and the database is distributed in large "flat-files" (ascii text files). This applies to for example the EMBL nucleotide sequence database. Currently people are considering using object oriented data models and object oriented database management systems (OODBMS).

2. Database Access

Genetics reached a major milestone as an international research team announced it has completed the first plant genome sequence. The species Arabidopsis thaliana has emerged as the plant counterpart of the laboratory mouse, offering clues to how all sorts of living organisms behave genetically, with potentially widespread applications for agriculture, medicine and energy. To analyse genome and protein sequences tools are available. The popular tools are FASTA and BLAST.

2. 1 The FASTA program package

Fasta is sequence comparison software that uses the method of Pearson and Lipman The program compares a DNA sequence to a DNA database or a protein sequence to a protein database. Practically, Fasta is a family of programs. Although there are a large number of programs in this package, they belong to three groups:

(i) "Conventional" Library Search programs:
FASTA3, FASTX3, FASTY3, TFASTA3, TFASTX3, TFASTY3, SSEARCH3.

(ii) Programs for searching with short Fragments:
FASTS3, FASTF3, TFASTS3, TFASTF3.

(iii) Statistical Significance: PRSS3.

Programs that start with fast search protein databases, while TFAST programs search translated DNA databases. *Table*-I gives a brief description of the programs

.

The FASTA program package does not include any protein or DNA sequence libraries. Protein databases are available on CD-ROM from the PIR and EMBL (see below), or via anonymous FTP from many different sources Genbank at the National for Biotechnology Information (www.ncbi.nlm.nih.gov and ftp://ncbi.nlm.nih.gov) and EMBL at the European Bioinformatics Institute (www.ebi.ac.uk).

2.1 Table I. Comparison programs in the FASTA3 package

fasta3 Compare a protein sequence to a protein sequence database or a DNA sequence to a DNA sequence database using the FASTA algorithm (Pearson and Lipman, 1988, Pearson, 1996). Search speed and selectivity are controlled with the ktup(wordsize) parameter. For protein comparisons, ktup = 2 by default; ktup =1 is more sensitive but slower. For DNA comparisons, ktup=6 by default; ktup=3 or ktup=4 provides higher sensitivity; ktup=1 should be used for oligonucleotides (DNA query lengths < 20).

ssearch3 Compare a protein sequence to a protein sequence database or a DNA sequence to a DNA sequence database using the Smith Waterman algorithm (Smith and Waterman, 1981). ssearch3 is about 10-times slower than FASTA3, but is more sensitive for full-length protein sequence comparison.

fastx3/fasty3 Compare a DNA sequence to a protein sequence database, by comparing the translated DNA sequence in three frames and allowing gaps and frame shifts. fastx3 uses a simpler, faster algorithm for alignments that allows frame shifts only between codons; fasty3 is slower but produces better alignments with poor quality sequences because frame shifts are allowed within codons.

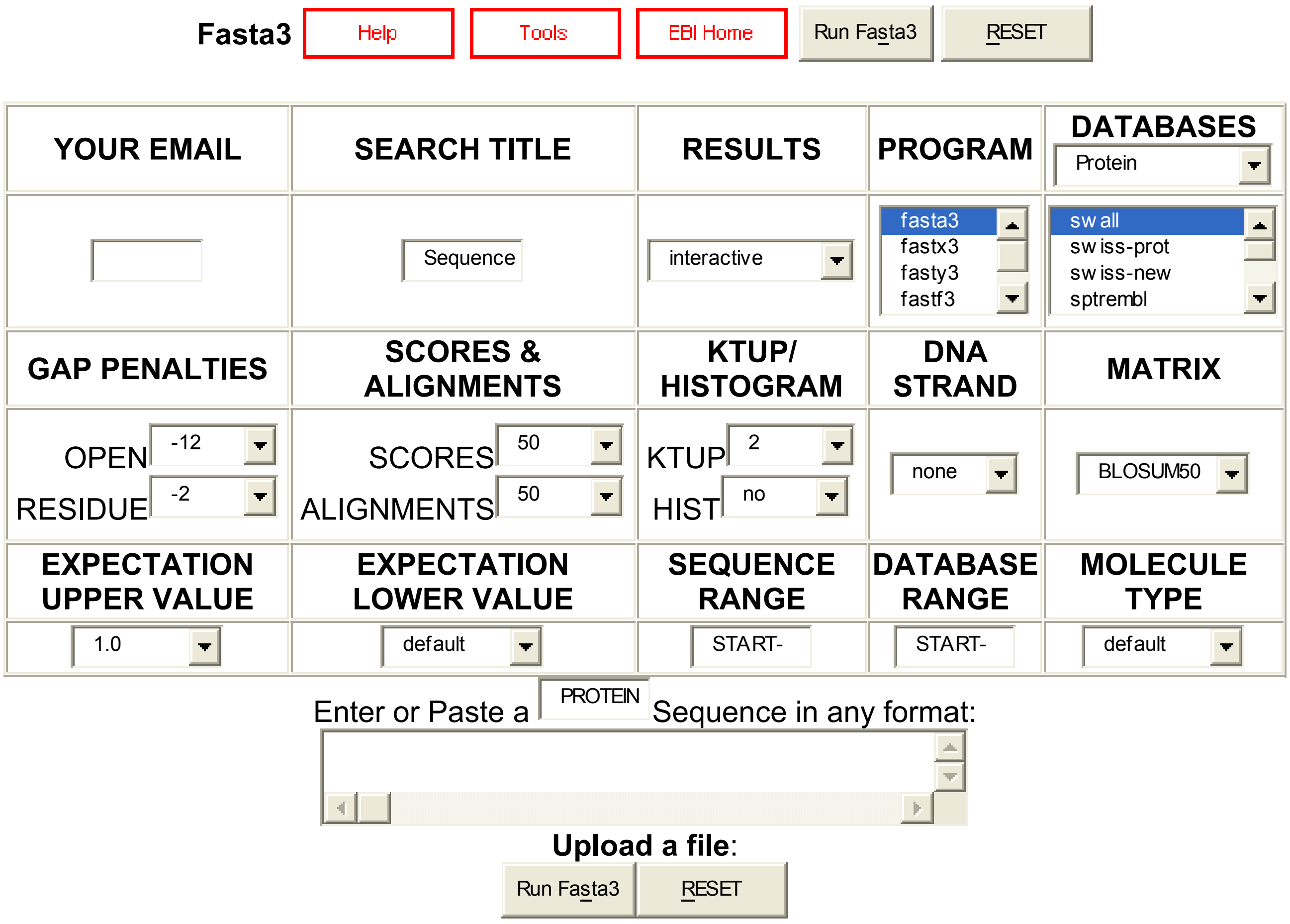
tfastx3/tfasty3 Compare a protein sequence to a DNA sequence database, calculating similarities with frame shifts to the forward and reverse orientations.

tfasta3 Compare a protein sequence to a DNA sequence database, calculating similarities (without frame shifts) to the 3 forward and three reverse reading frames. tfastx3 and tfasty3 are preferred because they calculate similarity over frame shifts.

fastf3/tfastf3 Compares an ordered peptide mixture, as would be obtained by Edman degradation of a CNBr cleavage of a protein, against a protein (fastf) or DNA (tfastf) database.

fasts3/tfasts3 Compares set of short peptide fragments, as would be obtained from mass-spec. Analysis of a protein, against a protein (fasts) or DNA (tfasts) database.

2.3. FASTA3 Software



YOUR EMAIL	SEARCH TITLE	RESULTS	PROGRAM	DATABASES Protein
	Sequence	interactive	fasta3 fastx3 fasty3 fastf3	swall swiss-prot swiss-new sptrembl
GAP PENALTIES	SCORES & ALIGNMENTS	KTUP/ HISTOGRAM	DNA STRAND	MATRIX
OPEN -12 RESIDUE -2	SCORES 50 ALIGNMENTS 50	KTUP 2 HIST no	none	BLOSUM50
EXPECTATION UPPER VALUE	EXPECTATION LOWER VALUE	SEQUENCE RANGE	DATABASE RANGE	MOLECULE TYPE
1.0	default	START-	START-	default

Enter or Paste a PROTEIN Sequence in any format:

Upload a file:

2.3.1 About Sequence Formats

Sequence formats are simply the way in which the amino acid or DNA sequence is recorded in a computer file. Different programs expect different formats, so if you are to submit a job successfully, it is important to understand what the various formats look like.

i. Your sequences

You can cut & paste or type a sequence into the large text window. The server can accept most common sequence formats like fasta, GCG, EMBL, Genbank, NBRF and Phylip. Free text sequences which are simply a block of characters representing a DNA or Protein sequence are also accepted. Please note that partially formatted sequences will not be accepted. The server will return an error.

Copying and Pasting directly from word processors may yield unpredictable results, as hidden/control characters may be present. IF your sequence is DNA and contains more than 50% ambiguity codes it will be rejected by the system (both email and interactive submissions). It is best to use the part(s) of the sequence that contain real DNA (good consensus regions) as the likelihood if these regions giving positive scores is higher than if they contained many 'N's which the programs will attempt to match to the entire database(s). If successful, these database scans will contain false positive hits, which are of limited use.

ii. Short sequences

For very short nucleic acid sequences, the strategy recommended is to decrease the word length (ktup) from 6 to 1. In this way a significant increase in sensitivity is achieved without loss of biological significance. Fasta may not return as many hits as blast does in this type of searches but the relevance of fasta results is much higher than with blast. For Fastf3 and Fasts3 you must use the following format to enter the fragments:

```
>mgstm1
MGCEN,
MIDYP,
MLLAY,
MLLGY
```

Examples of Sequence Formats:

i. EMBL Format:

```
ID AB00001 standard; DNA; PRO; 155 BP.
XX
AC AB00001;
XX
NI d1137212
XX
DT 10-SEP-1997 (Rel. 52, Created)
DT 09-APR-1998 (Rel. 55, Last updated, Version 2)
XX
DE Gluconobacter oxydans 16S rRNA, partial sequence.
XX
SQ Sequence 155 BP; 39 A; 34 C; 47 G; 35 T; 0 other;
GGGCTACACA  CGTGCTACAA   TGGCGGTGAC  AGTGGGAAGC
TATGTGGTGA   CACAGTGCTG 60
ATCTCTAAAA   GCCGTCTCAG   TTCGGATTGT   ACTCTGCAAC
TCGAGTACAT   GAAGGTGGAA 120
TCGCTAGTAA   TCGCGGATCA   GCATGCCGCG  GTGAA 155
//
```

ii. Fasta Format:

A sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The description line is distinguished from the sequence data by a greater-than (">") symbol in the first column. It is recommended that all lines of text be shorter than 80 characters in length. An example sequence in FASTA format is:

```
>gi|532319|pir|TVFV2E|TVFV2E envelope protein
ELRLRYCAPAGFALLKCNDADYDGFKTNCSNVSVVHCTNLMNTTVTTGLLL
NGSYSENRT
QIWQKHRTSNDSALILLNKHYNLTVTCKRPGNKTVLPVTIMAGLVFHSQKYN
LRLRQAWC
HFPSNWKGAWKEVKEEIVNLPKERYRGTNDPKRIFFQRQWGDPETANLWF
NCHGEFFYCK
MDWFLNYLNNLTVDADHNECKNTSGTKSGNKRAPGPCVQRTYVACHIRSVI
IWLETISKK
TYAPPREGHLECTSTVTGMTVELNYIPKNRTNVTLSPQIESIWAAELDRYKLV
EITPIGF
APTEVRRYTGGHERQKRVPFVXXXXXXXXXXXXXXXXXXXXXXVQSQHLLA
GILQQQKNL
LAAVEAQQQMLKLTIWGVK
```

iii. PIR Format:

```
>P1;FOSB_HUMAN
sw:fosb_human => FOSB_HUMAN
MFQAFPGDYD   SGSRCSSSPS    AESQYLSSVD    SFGSPPTAAA
SQECAGLGEM   PGSFVPTVTA   ITTSQDLQWL VQPTLISSMA
QSQGQPLASQ   PPVVDPYDMP   GTSYSTPGMS    GYSSGGASGS
GGPSTSGTTS   GPGPARPARA RPRRPREETL      TPEEEEKRRV
RRERNKLAAA  KCRNRRRELT   DRLQAETDQL   EEEKAELESE
IAELQKEKER LEFVLVAHKP     GCKIPYEEGP    GPGPLAEVRD
LPGSAPAKED   GFSWLLPPPP     PPPLPFQTSQ     DAPPNLTASL
FTHSEVQVLG     DPFPVVNPSY    TSSFVLTCPE     VSAFAGAQRT
SGSDQPSDPL    NSPSLLAL*
C;ID FOSB_HUMAN STANDARD; PRT; 338 AA.
C;AC P53539;
C;DT 01-OCT-1996 (REL. 34, CREATED)
```

```
C;DT 01-OCT-1996 (REL. 34, LAST SEQUENCE UPDATE)
C;DT 01-OCT-1996 (REL. 34, LAST ANNOTATION UPDATE)
C;DE FOSB PROTEIN (G0/G1 SWITCH REGULATORY PROTEIN 3). . . .
```

iv. Phylip Format:

```
1 380
FOS_HUMAN MMFSGFNADY EASSSRCSSA SPAGDSLSYY
HSPADSFSSM GSPVNAQDFC

TDLAVSSANF IPTVTAISTS PDLQWLVQPA LVSSVAPSQT RAPHPFGVPA

PSAGAYSRAG VVKTMTGGRA QSIGRRGKVE QLSPEEEEKR
RIRRERNKMA

AAKCRNRRRE LTDTLQAETD QLEDEKSALQ TEIANLLKEK EKLEFILAAH

RPACKIPDDL GFPEEMSVAS LDLTGGLPEV ATPESEEAFT LPLLNDPEPK

PSVEPVKSIS SMELKTEPFD DFLFPASSRP SGSETARSVP DMDLSGSFYA

ADWEPLHSGS LGMGPMATEL EPLCTPVVTC TPSCTAYTSS
FVFTYPEADS

FPSCAAAHRK GSSSNEPSSD SLSSPTLLAL
```

2.3.2 Your Email

You must type your valid email address in this text box. It is not necessary to fill in the box if you are running your search interactively (RESULT=interactive).

2.3.3 Search title

You may type any text you want to help you identify the database search results.

2.3.4 Results

This option lets you choose between email and interactive runs. The email run sends result to given email address in the EMAIL text box. The default value is email.

2.3.5 Databases

Choose databases you which to run your protein sequence against. You can choose multiple databases by clicking on them. The choices will appear highlighted. Please note that Netscape and Internet Explorer behave differently when doing this. To choose multiple databases in Netscape simply click on the database names. To deselect a database simply click on it again. You have to press SHIFT and click on the databases in order to make a multiple choice.

i. Database names references:

Database	Reference
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
SGT	Structural Genomic Targets Database
PDB	Protein Database of Brookhaven
ENSEMBL	ENSEMBL CDS PEP
EMBL	The EMBL Database (All divisions) Non Interactive
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
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) Non Interactive
IMGT	IMGT Immunogenetics Database
HGBASE	European SNP database

ii. Databases with related Program

Database	Description	Program
EMBL	EMBL Database	fasta,blastn,blastx
EMBLNEW	EMBL Updates	fasta,blastn,blastx
EMALL	EMBL + Updates	fasta,blastn,blastx
EMBL Divisions (i.e. EFUN)	EMBL by Division	fasta
EEST	EMBL ESTs	fasta,blastn.blastx
EVEC	EMBL Vectors	fasta,blastn.blastx
SWISS-PROT	Swiss-Prot protein sequences	fasta,blastp,blitz
TrEMBL	Translated EMBL	fasta,blastp,blitz
SWISS-NEW	Swiss-Prot Updates	fasta,blastp,blitz
TrEMBL New	Translate EMBL Updates	fasta,blastp,blitz
SWALL	Non-Redundant protein sequence database	fasta,blastp,scanps,blitz, ssearch3
PDB	Brookhaven protein sequence database	blastp, scanps
Gpcrdb	Gene Protein Coupled receptors protein sequence database	blastp
Prints	Protein Fingerprints database	blastp

iii. The Genbank DNA sequence library

Because of the large size of DNA databases, you will probably want to keep DNA databases in only one or possibly two formats. The FASTA3 programs that search DNA databases - fasta3, tfastx/y3, and tfasta3 can read DNA databases in Genbank flat file (not ASN.1), FASTA, GCG/compressed-binary, BLAST1.4 (pressdb), and BLAST2.0 (formatdb) formats, as well as EMBL format. If you are also running the GCG suite of sequence analysis programs, you should use GCG/compressed-binary format or BLAST2.0 format for your fasta3 searches. If not, BLAST2.0 is a good choice. These files are considerably more compact than Genbank flat files, and are preferred. The NCBI does not provide software for converting from Genbank flat files to Blast2.0 DNA databases, but you can use the Blast formatdb program to convert ASN.1 formatted Genbank files, which are available from the NCBI ftp site.

The NCBI also provides the nr, swissprot, and several EST databases that are used by BLAST in FASTA format from: ftp://ncbi.nlm.nih.gov/blast/db. These databases are updated nightly.

iv. The NBRF protein sequence library

You can obtain the PIR protein sequence database (Barker et al., 1998) from National Biomedical Research Foundation; Georgetown University Medical Center; 3900 Reservoir Rd, N.W.; Washington, D.C. 20007or via ftp from nbrf.georgetown.edu or from the NCBI (ncbi.nlm.nih.gov/repository/PIR). The data in the ASCII directory is in PIR Codata format, which is not widely used. Recommend format is the PIR/VMS format data (libtype=5) in the vms directory.

2.3.6 Matrix

There are many scoring matrices available that help in discerning biological significance. Their efficiency depends on the type of experiment and the data used during the alignment phase of the programs. 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.

The BLOSUM matrices, also used to protein database search scoring (the default in blastp), are divided into statistical significance degrees which, in a way, are reminisent of PAM distances. For example, BLOSUM64 is roughly equivalent to PAM 120

Use this option to set which comparison matrix should be used when searching the database. The default matrix for blast is blosum62. You may choose from a complete list of matrices, which should cover various evolutionary constraints.

2.3.7 Gap penalties

When aligning two sequences together it is often required to insert gaps in them in order to optimise the alignment. This can be done on the basis of identities alone, inserting gaps in the sequences as required where there are no matches. However, this is not recommended for biological sequence comparisons because similarities are then not taken into consideration. A scoring scheme, often referred to as a comparison matrix, is used which gives a high positive score when the identical residues or bases are properly aligned. Slightly less if a similarity or homology is possible (i.e. a conservative substitution) and even negative scores for alignment pairs, which are not biologically significant.

When two sequences are aligned together a diagonal is created which depicts the best alignment path for these. This diagonal may be broken in places due to mismatches. If there are too many of these the diagonal is subdivided into several smaller ones. In order to make the alignment better gap initiation and gap extension penalties are introduced which penalise the total alignment score.

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).

2.3.8 Histogram

Setting this option to "yes" will display the search histogram of the expected frequency of chance occurrence of the database matches found.

2.3.9 Scores

Setting this option to any number available in the menu allows you to set to maximum number of reported scores in the output file.

2.3.10 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.

2.3.11 KTUP

Change this value to limit the word-length for searching. A word-length of *two* 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 six is the default. Please note that if you do not specify a ktup larger than three when doing a nucleic database search ktup will be set automatically to six.

2.3.12 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.

2.3.13 Expectation value upper limit

Here you may set the expectation value upper limit for score and alignment display. The defaults are 10.0 for FASTA with protein searches, 5.0 for translated DNA/protein comparisons, and 2.0 for DNA/DNA searches.

2.3.14 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 relationships that are distant.

2.3.15 Sequence range

This option allows the user to denote which region within the query sequence should be searched. For example, if the user submits a sequence of 380 aa/nt he/she may wish the search the database using the region comprising positions 50 to 200. In this case, the user should type into the text field the numbers: **50-200**.

2.3.16 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 that indicates sequences with at least 1000 aa/nt and up-to, but no more than 3000 aa/nt.

2.3.17 Molecule type

This option is used to choose or enforce the molecule type of the query in use for a search. This is useful when using programs such as tfastx and tfasty. Please note that this option needs not to be changed when using the standard fasta3 program.

2.3.18 Sequence input window

You can cut & paste or type a sequence into the large text window. A free text sequence is simply a block of characters representing a DNA or Protein sequence. You may also paste a sequence in GCG, FASTA, EMBL, GenBank, PIR, NBRF or Phylip format. Partially formatted sequences will not be accepted.

2.3.19 Upload a file

You may upload a file from your computer which containing a valid protein sequence in any format (GCG, FASTA,PIR,etc.) using this option. Please note that this option only works with Netscape Browsers!

3. Basic Local Alignment Search Tool

3.1 Introduction

The Basic Local Alignment Search Tool (BLAST), for comparing gene and protein sequences against others in public databases, now comes in several flavours including Position specific iterative BLAST(PSI-BLAST), PHI-BLAST, and BLAST 2 sequences. Specialized BLASTs are also available for human, microbial, and malaria genomes, as well as for vector contamination, immunoglubulins, and tentative human consensus sequences.

The core of NCBI 's BLAST services is BLAST 2.0 otherwise known as "Gapped BLAST". This service is designed to take protein and nucleic acid sequences and compare them against a selection of NCBI databases.

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 (Altschul et al., 1990). 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.

3.2 Selecting the BLAST Program

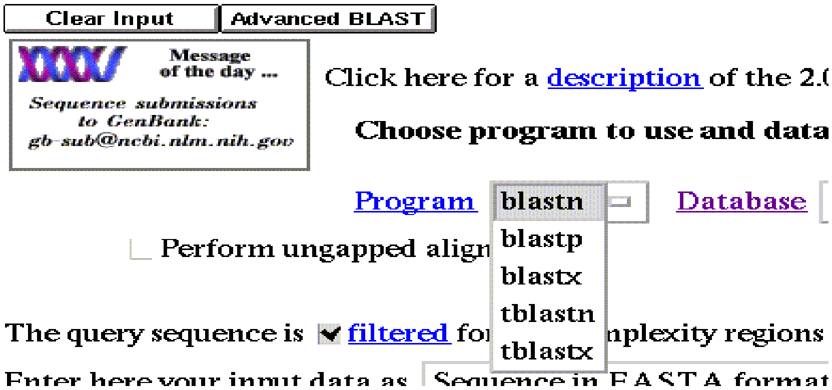
The BLAST search pages allow you to select from several different programs. Below is a table of these programs.

Program	Description
blastp	Compares an amino acid query sequence against a protein sequence database.
blastn	Compares a nucleotide query sequence against a nucleotide sequence database.
blastx	Compares a nucleotide query sequence translated in all reading frames against a protein sequence database. You could use this option to find potential translation products of an unknown nucleotide sequence.
tblastn	Compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames.
tblastx	Compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database. Please note that the tblastx program cannot be used with the nr database on the BLAST Web page because it is computationally intensive.

To select a BLAST program for your search:

i. Open the Basic BLAST search page.

ii. From the "Program" Pull Down Menu select the appropriate program.



Click here for a description of the 2.(

Choose program to use and data

Program blastn

□ Perform ungapped align

The query sequence is ☑ filtered fo nplexity regions

Enter here your input data as Sequence in FASTA format

Figure 1. Using the pull down menu to select a BLAST program.

3.3 Selecting the BLAST Database

You can select several NCBI databases to compare your query sequences against. Note that some databases are specific to proteins or nucleotides and cannot be used in combination with certain BLAST programs (for example a blastn search against swissprot).

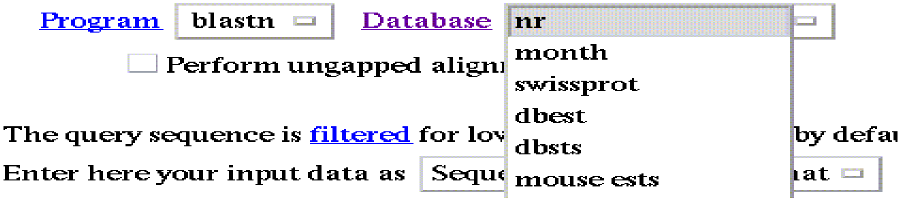
3.3.1 Proteins

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 (no updates). 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. See "Alu alert" by Claverie and Makalowski, Nature vol. 371, page 752 (1994).

3.3.2 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	The non-redundant Database of

ests	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 (Rel. 1.0, July 1995).	
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. See "Alu alert" by epd	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.	



Program blastn Database nr

□ Perform ungapped align

The query sequence is filtered for lov by defa

Enter here your input data as Seque at

Figure 2. Using the Pull Down Menu to select the BLAST database.

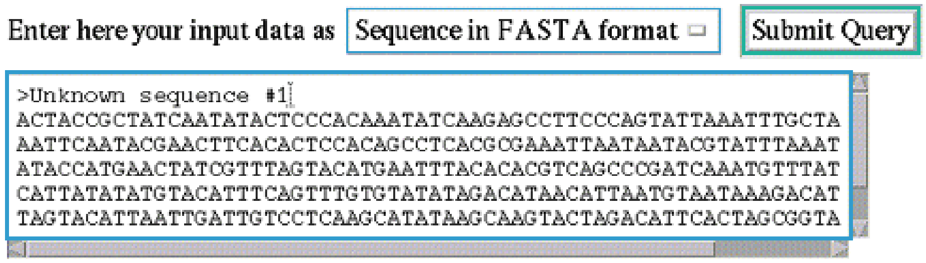
3.4 Entering your Sequence

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

3.4.1 FASTA Format

A description of the FASTA format is located on the Basic BLAST search pages.

1. Open your FASTA formatted sequence in a text editor as plain text.
2. Use your mouse to highlight the entire sequence.
3. Select Edit/Copy from the menu in your text editor.
4. Go to the BLAST search page in your web browser.
5. Use your mouse to select the main input field titled "Enter your input data here", by clicking it once.
6. Select Edit/Paste from the browser's menu.
7. You should now see your FASTA sequence in this field.
8. Set the pull down menu to "Sequence in FASTA format".



Enter here your input data as Sequence in FASTA format

```
>Unknown sequence #1
ACTACCGCTATCAATATACTCCCACAAATATCAAGAGCCTTCCCAGTATTAAATTTGCTA
AATTCAATACGAACTTCACACTCCACAGCCTCACGCGAAATTAATAATACGTATTTAAAT
ATACCATGAACTATCGTTTAGTACATGAATTTACACACGTCAGCCCGATCAAATGTTTAT
CATTATATATGTACATTTCAGTTTGTGTATATAGACATAACATTAATGTAATAAAGACAT
TAGTACATTAATTGATTGTCCTCAAGCATATAAGCAAGTACTAGACATTCACTAGCGGTA
```

Figure 3. Example of a FASTA sequence in the input field.

3.5 Accession or GI number

If you know the Accession number or the GI of a sequence in GenBank, you can use this as the query sequence in a BLAST search.

1. Go to the BLAST search page in your web browser.
2. Use your mouse to select the main input field titled "Enter your input data here", by clicking it once.
3. Using the keyboard enter the GenBank Accession number or the GI number.
4. Set the Pull Down Menu to "Accession or GI".

3.6 Submitting your Search

1. Make sure you have selected the correct BLAST program and BLAST database.
2. If you have entered your FASTA sequence or an Accession or GI number, click the "Submit Query Button".
3. BLAST will now open a new window and tell you it is working on your search.
4. Once your results are computed they will be presented in the window.

4 Database Searching, Browsing and Analysis Tools

4.1 Database Searching

SRS6	Database searching using the Sequence Retrieval System version 6.x
InterPro	Integrated Resource of Protein Domains and Functional Sites
CORBA Tools	EBI CORBA Tools
EnsEMBL	Automatic annotation of eukaryotic genomes
	Index of Protein Profiles Search the Index of Protein Profiles

4.2 Homology Searches

4.2.1 DNA and Protein Searches

Fasta3 (new)	A new version of the fasta3 interface with more user options (currently undergoing testing - please report problems to support@ebi.ac.uk)
Fasta3	Sequence similarity and homology searching against nucleotide and protein database using Fasta3. Please note that this service will be phased out shortly. Use the new version above.
WU-Blast2	Washington University blast2 (blast 2.0 with gaps)
NCBI-Blast2	NCBI blast2 (blastall) program

4.2.2 Protein Searches

MPsrch	Edinburgh Biocomputing Systems´ very fast implementation of the true Smith and Waterman algorithm
bic_sw	Compugen´s Bic2´s Smith & Waterman algorithm implementation for protein database searches
Scanps2.3	NEW! Version 2.3 of Scanps Fast implementation of the true Smith & Waterman algorithm for protein database searches
InterProScan	Search protein sequences againts InterPro member databases
FingerPRINTScan	PRINTS Protein fingerprint searches
ppsearch	Protein motifs searches
GeneQuiz	Highly automated analysis of biological sequences

4.3 Specialised Servers

Genome and Proteome

Fasta3 server	Completed Genomes and Proteomes fasta3 server
Parasites blast	Parasite Genomes blast server
EGI blast	Blast server of EST clusters and alignments based on the EuroGeneIndexes
SNP-Fasta3 server	European SNP database (HGBASE) Fasta3 searches

4.3 3D Structure Analysis

DALI	Comparison of protein structures in 3D
FSSP	Fold classification and structural alignments
MaxSprout	Reconstruction of 3D coordinates from C(alpha) trace
TOPS	Protein structural topology
3Dee	Domain search using PDB codes
PQS	Protein quaternary structure query

4.4 Analysis Tools

ClustalW_mp	Multiple sequence alignments
AMAS	Analyze multiply aligned sequences
GeneMark	Gene prediction service
Pratt	Protein motif discovery
Radar	Protein repeat detection
Mutation Checker	Sequence validation
Jpred	Consensus secondary structure prediction
GenomeBuilder	GenomeBuilder is for the analysis and visualisation of EST sequences
PredictProtein	Secondary and Tertiary protein structure prediction (PHDacc, PHDsec, PHDhtm, TOPITS)
W2H v3.1	A WWW interface to the GCG sequence analysis software

4.5 Utilities

CpG Islands	CpG Islands finder
CpG Plot/CpGreport	CpG Island finder and plotting tool (EMBOSS)
Align	Pairwise global and local alignment tool (EMBOSS)
Transeq	DNA sequence translation tool (EMBOSS)
Prima	Primer detection tool (EMBOSS)
Pepstats/ (EMBOSS)	
Pepwindow/Pepinfo	EMBOSS programs for basic protein sequence analysis
SAPS	Statistics on protein sequences
Genetic Code Viewer	Review of genetic code differences

Protein Engine	Translate DNA sequences
Protein Colourer	The Protein Colourer
Reverse Translator	Reverse complement checker
Protein Annotation Assistant	Tool that help with Protein Annotation
AppLab v1.0	A CORBA-Java based application wrapper
Webin	New sequence submissions to EMBL database
Webin-Align	Submission of alignments to EMBL
XEMBL	EMBL records in BSML or AGAVE XML format
readseq	Sequence format conversion utility

5 Plant Genome Mapping Projects

5.1 African Projects

Bean Genome Mapping Projects - Africa
Corn Genome Mapping Projects -Africa
Fungi Genome Mapping Projects -Africa

5.2 Australian Projects

Barley Genome Mapping Projects	Bean Genome Mapping Projects
Cereal Genome Mapping Projects	Corn Genome Mapping Projects
Cotton Genome Mapping Projects	Fungi Genome Mapping Projects
Lettuce Genome Mapping Projects	Mushroom Genome Mapping Projects
Pine Genome Mapping Projects	Rye Genome Mapping Projects
Sorghum Genome Mapping Projects	Sugarcane Genome Mapping Projects
Wheat Genome Mapping Projects	

5.3 European Projects

Apple Genome Mapping Projects	Arabidopsis Genome Mapping Projects
Barley Genome Mapping Projects	Bean Genome Mapping Projects
Brassicas Genome Mapping Projects	Cabbage Genome Mapping Projects
Cereal Genome Mapping Projects	Clover Genome Mapping Projects
Corn Genome Mapping Projects	Cucumber Genome Mapping Projects
Fungi Genome Mapping Projects	Grass Genome Mapping Projects
Lilium Genome Mapping Projects	Mushroom Genome Mapping Projects
Pea Genome Mapping Projects	Peach Genome Mapping Projects
Plum Genome Mapping Projects	Potato Genome Mapping Projects
Rye Genome Mapping Projects	Snapdragon Genome Mapping Projects
Spruce Genome Mapping Projects	Tobacco Genome Mapping Projects
Tomato Genome Mapping Projects	Wheat Genome Mapping Projects

5.4 North American Projects

Alfalfa Genome Mapping Projects	Almond Genome Mapping Projects
Apple Genome Mapping Projects	Arabidopsis Genome Mapping Projects
Asparagus Genome Mapping Projects	Barley Genome Mapping Projects
Bean Genome Mapping Projects	Berry Genome Mapping Projects
Brassicas Genome Mapping Projects	Cabbage Genome Mapping Projects
Carrot Genome Mapping Projects	Celery Genome Mapping Projects
Chrysanthemum Genome Mapping Projects	Citrus Genome Mapping Projects
Citrus and Other Perennial Fruit Crops	Clover Genome Mapping Projects
Cocoa Genome Mapping Projects	Corn Genome Mapping Projects

Cotton Genome Mapping Projects	Cucumber Genome Mapping Projects
Cuphea Genome Mapping Projects	Fungi Genome Mapping Projects
Grass Genome Mapping Projects	Lentil Genome Mapping Projects
Lettuce Genome Mapping Projects	Melon Genome Mapping Projects
Oat Genome Mapping Projects	Onion Genome Mapping Projects
Papaya Genome Mapping Projects	Pea Genome Mapping Projects
Peach Genome Mapping Projects	Peanut Genome Mapping Projects
Pear Genome Mapping Projects	Pepper Genome Mapping Projects
Pine Genome Mapping Projects	Plum Genome Mapping Projects
Poplar Genome Mapping Projects	Potato Genome Mapping Projects
Rice Genome Mapping Projects	Rose Genome Mapping Projects
Rye Genome Mapping Projects	Sorghum Genome Mapping Projects
Soybean Genome Mapping Projects	Spruce Genome Mapping Projects
Squash Genome Mapping Projects	Sugarcane Genome Mapping Projects
Sunflower Genome Mapping Projects	Tobacco Genome Mapping Projects
Tomato Genome Mapping Projects	Turf Grass Genome Mapping Projects
Wheat Genome Mapping Projects	

6. Bioinformatics and Molecular Biology Links

Bioinformatics Institute	
NCBI	http://www.ncbi.nlm.nih.gov
EBI	http://www.ebi.ac.uk
DDBJ	http://www.ddbj.nig.ac.jp/
ExPASy	http://www.expasy.ch
Alces	http://alces.med.umn.edu/
NCGR	http://www.ncgr.org/
EMBL	http://www.embl.org/
Sequence Search	
fasta3	http://www.ebi.ac.uk/fasta3/
blast2	http://www.ebi.ac.uk/blast2/
UTR home page	http://bigarea.area.ba.cnr.it:8000/EmblT/UTRHome/
Codon usage	http://bioweb.pasteur.fr/seqanal/interfaces/codonw.html
ESTblast	http://www.hgmp.mrc.ac.uk/ESTBlast/
Promoter scan	
Promoter scan	http://bimas.dcrt.nih.gov/molbio/proscan/
Promoter Finder	http://www.fruitfly.org/seq_tools/promoter.html
Place	http://www.dna.affrc.go.jp/htdocs/PLACE/
Improbizer(dynamic)	http://www.cse.ucsc.edu/~kent/improbizer/improbizer.html
Translation of DNA to protein	
ExPASy-tranlate tool	
Alces-translate tool	http://alces.med.umn.edu/rawtrans.html

MBS translator	http://mbshortcuts.com/translator/
Transmembrane domains	
DAS-domain prediction	
TMHMM - Helices	http://www.cbs.dtu.dk/services/TMHMM-2.0/
Topology	http://www.sbc.su.se/~erikw/toppred2/
TmPRED	http://www.ch.embnet.org/software/TMPRED_form.html
Stock centres	
NASC(Arabidopsis)	http://nasc.nott.ac.uk/
ABRC (Arabidopsis)	http://aims.cps.msu.edu/aims/
Maize Genetics Cooperation	http://w3.ag.uiuc.edu/maize-coop
Plant Germplasm	http://www.ars-grin.gov/npgs/
Physico-chemical properties of proteins	
Isoelectric point	http://www.embl-heidelberg.de/~toldo/JaMBW/3/1/6/
RNA secondary structure	http://www-camis.stanford.edu/projects/helix/sstruct/
Peptide mass calculator	http://ca.expasy.org/tools/peptide-mass.html
Mass fingerprint search 'Mascot'	http://www.matrixscience.com/search_form_select.html
Protein Prospector	http://prospector.ucsf.edu/
TagIdent	http://ca.expasy.org/tools/tagident.html
Literature	
PubMed	http://www3.ncbi.nlm.nih.gov/PubMed/medline.html
Highwire	http://highwire.stanford.edu/lists/freeart.dtl
J.Biol.Chem.	http://www-jbc.stanford.edu/jbc/
Web of Science(UK)	http://wos.mimas.ac.uk/
BIOME	http://biome.ac.uk/biome.html
Public Library of Science	http://www.publiclibraryofscience.org/
Reprint Directory	http://www.umanitoba.ca/faculties/medicine/radiology/search/universities.html
Sequence manipulation tools	
Restriction analysis	http://www.ccsi.com/firstmarket/cutter/cut2.html
BCM	http://dot.imgen.bcm.tmc.edu:9331/multi-align/multi-align.html
Reverse complement	http://www.ualberta.ca/~stothard/javascript/rev_comp.html
OUBC	http://analysis.molbiol.ox.ac.uk/pise_html/index.html
Backtranslate	http://www.entelechon.com/eng/backtranslation.html
Rebase	http://rebase.neb.com
Phylogeny tools	
Phylip(Singapore)	http://sdmc.krdl.org.sg:8080/~lxzhang/phylip/
Phylip (Pasteur)	http://bioweb.pasteur.fr/seqanal/phylogeny/phylip-uk.html
FastDNAml	http://bioweb.pasteur.fr/seqanal/interfaces/fastdnaml.html
TreeGen	http://cbrg.inf.ethz.ch/subsection3_1_6.html#SECTION0001600000000000000

Phylogenetic tree prediction	http://www.genebee.msu.su/services/phtree_reduced.html
Gene identification	
WebGene	http://www.itba.mi.cnr.it/webgene/
GenScan	http://genes.mit.edu/GENSCAN.html
NetPlantGene2	http://www.cbs.dtu.dk/services/NetGene2/
Sequence retrival	
SRS session	http://srs.ebi.ac.uk/
HMM Genie	http://www.fruitfly.org/seq_tools/genie.html
Plant Sequence Projects	
Rice	http://www.tigr.org/tdb/e2k1/osa1/index.shtml
MaizeDB	http://www.agron.missouri.edu/
Brassica napus	http://www.cbc.umn.edu/ResearchProjects/Brassica/index.html
Ongoing eukaryotic projects	http://wit.integratedgenomics.com/GOLD/eukaryagenomes.html
ATCC	http://www.atcc.org
E.coli genetic stock center	http://cgsc.biology.yale.edu/
Fungal genetics stock center	http://www.fgsc.net/
Bacillus genetic stock center	http://bacillus.biosci.ohio-state.edu
Free software	
pDRAW32	http://www.crosswinds.net/~acaclone/
DNA Tools	http://www.crc.dk/phys/dnatools.htm
ABI Sequence trace viewer	http://www.genomex.com/sharedwebsite/freeware.html
AlignACE	http://arep.med.harvard.edu/mrnadata/mrnasoft.html
AnnHYB	http://annhyb.free.fr/
DNAforWindows	http://website.lineone.net/~molbio/
Hidden markov models in bioinformatics	
UCSC HMM Applications	http://www.cse.ucsc.edu/research/compbio/HMM-apps/HMM-applications.html
Pfam	http://www.sanger.ac.uk/Software/Pfam/search.shtml
Decypher2	http://decypher.stanford.edu/index_by_algo.htm
123D Sanger	http://genomic.sanger.ac.uk/123D/123D.html
FRSVR	http://www.doe-mbi.ucla.edu/people/fischer/TEST/getsequence.html
SAMt98	http://www.cse.ucsc.edu/research/compbio/HMM-apps/model-library-search.html
Arabidopsis databases	
AGI(US)	AGI(US)
AGI(JP)	AGI(JP)
AGI(EU)	AGI(EU)
Plant microarray databases	Plant microarray databases
Stanford microarray database	Stanford microarray database

MIPS	MIPS
CropNet	CropNet
Alignment tools	
ClustalW	ClustalW
Boxshade	Boxshade
DiAlign	DiAlign
Stretcher	Stretcher
Multiple alignment resource	Multiple alignment resource

ELECTRONIC INFORMATION RESOURCES

M.P RAMESH KUMAR
Technical Officer (T6), CPCRI, KASARAGOD

1.0 Introduction

The challenges put forward by information explosion and the rising cost of makes access to information difficult. The transition from conventional information sources to the latest electronic media has made information available quickly. Internet has made communication international and instantaneous. The advent of Internet technology and its adoption by the organizations has resulted in the evolution of the intranet. Many important services like e-mail, discussion forums have become handy solutions for the dissemination of information. An exposure to the current developments in IT and familiarization of the methods and techniques of information access through Internet and Intranet will be of immense help to the scientific community.

2.0 Internet

The Internet provides access to a wealth of information on countless topics contributed by people throughout the world. It is a self-publishing medium. Anyone with a small amount of technical skill and access to a host computer can publish on the Internet. The resources on the Internet are not catalogued as in the case of library, which makes the retrieval of information a difficult task. The number of sites is increasing rapidly and the addresses of sites frequently change and the sites disappear often. These facts are to be kept in mind when you search for information on the web. As in the case of any other information resource the sites are to be evaluated for its content.

A variety of information sources and services are available on the Internet. A brief account is provided here on the information resources on the Internet with special reference to biotechnology, Bioinformatics, agriculture and biology.

2.1 Bibliographic databases

Bibliographic databases contain descriptive information (citation and subject headings) for publications, such as books, periodical articles, videotapes or government documents. Different bibliographic services help users to locate information sources or the information itself.

2.1.1 PubMed Central - http://www.ncbi.nlm.nih.gov/Literature/index.html

It is a digital archive of life sciences journal literature managed by the National Center for Biotechnology Information (NCBI) at the U.S. National Library of Medicine (NLM). It is not a journal publisher. Access to PubMed Central (PMC) is free and unrestricted. It is a part of **ENTREZ (**http://www.ncbi.nlm.nih.gov/ Entrez/)is a retrieval system for searching several linked databases.

2.1.2 BIOSIS Previews http://www.biosis.org/products_services/previews.html

Site for plant biology and other non-medical biology articles and conference papers from 1969 to the present (only 34% of the journals in BIOSIS overlap with MEDLINE). Since 1993 a "sequence data" field has been available but rather than containing actual nucleotide or amino acid sequences this field contains the accession number for the sequence from databases such as GenBank, EMBL and SwissProt, if the author included this number in the article text. A very small percentage of records actually use this field. BIOSIS is available from a number of database vendors, most of which will also provide links from the citation to your library's online journals.

2.1.3 ResearchIndex (formerly CiteSeer) – http://researchindex.com/

This is a free, full-text index to the freely available research articles on the web. "Although availability varies greatly by discipline, over a million research articles are freely available on the web. Some journals and conferences provide free access online, others allow authors to post articles on the web, and others allow authors to purchase the right to post their articles on the web." (Lawrence 2001) this index is popular with computer scientists because a great deal of their literature is available this way, and because ResearchIndex also provides citation analysis. While Web of Science doesn't index conference papers (which are a mainstay in computer science), ResearchIndex does (if the proceedings are on the web for free). It also offers reference linking, extraction of citation context, related document detection and the BibTeX entry for each article.

2. 2 Non Bibliographic Databases

Sequence and other non-bibliographic databases are the central, most important type of information resource. But there are hundreds of databases that might be considered relevant to bioinformatics. There are specialized databases for each species, and separate databases for different types of information (nucleic acid sequences, protein sequences, protein structures, biochemical and biophysical information, etc.). The multiplicity of databases makes selection confusing, and the databases themselves can be challenging to understand and navigate. Nomenclature is not standard. Data formats/metadata schemes are not standard. Databases struggle with data redundancy and charges that they contain a lot of "junk." There are a lot of interrelated pieces of information surrounding a gene (genome location, structure, sequence, expression information, chemistry, etc.) or a protein, which lead to somewhat complicated database structures and links to related databases which may or may not be intuitive. The following list of databases is intended to orient the reader to the major databases. To become a proficient searcher in each database takes considerable training

2. 2.1 EMBL Nucleotide Database - http://www.ebi.ac.uk/swissprot/index.html

Europe’s primary collection of nucleotide sequences managed by the European Bioinformatics Institute (EBI)

2. 2.2 SWISS-PROT - http://www.ebi.ac.uk/swissprot/index.html

A complete annotated protein sequence database managed by the European Bioinformatics Institute (EBI)

2. 2.3 Protein Data Bank (PDB) - http://www.rcsb.org/pdb/

The PDB was established at Brookhaven National Laboratories in 1971, making it the first public bioinformatics database. The PDB is now operated by the Research Collaboratory for Structural Bioinformatics (RCSB), which is a collaborative effort of the San Diego Supercomputing Center, Rutgers University, and the National Institute of Standards and Technology (NIST). The PDB is a repository of experimentally determined three-dimensional structures of biological macromolecules (proteins, enzymes, nucleic acids, protein-nucleic acid complexes, and viruses) derived from x-ray crystallography and NMR experiments (see http://www.rcsb.org/pdb/experimental_methods.html for a helpful overview of these methods).

2. 2.4 Entrez Genome - http://www.ncbi.nlm.nih.gov/entrez/

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

2.3 Training Materials/Tutorials

The Internet offers various free publications and training materials intended for Internet applications and different categories with clear explanation of recommendations.

2.3.1 The Bioinformatics Training Resource
http://www.med.nyu.edu/rcr/rcr/btr/

The BTR is an organized collection of links to online tutorials, online courses, essays, book chapters, course syllabi, glossaries, bibliographies of key papers, etc. In short everything that interested scientists need in order to train themselves in the emerging discipline of bioinformatics.

2.3.2 IETF/TERENA Training Materials Catalogue
http://www.trainmat.ietf.org/

It is a catalogue of quality Network Training Materials for use by Internet trainers in training their users.

2.4 Softwares

A large number of freeware and open source softwares related to biocomputing is available for download at different sites. Free utilities for enhancing computer operations are also available.

2.4.1 Mathtools.net - http://www.mathtools.net

Various software tools are available for scientific and engineering needs.

2.4.2 Bioinformatics.Org - http://bioinformatics.org/

An international non-profit, academe-based organization committed to opening access to bioinformatics research projects, providing Open Source software for bioinformatics by hosting its development, and keeping biological information freely available

2.4.3 ExPASy Proteomics tools – http://www.expasy.ch/

Large number of softwares for proteomics analysis is available

2.4.4 Program Files - www.programfiles.com

Provides utilities for general computing

2.4.5 Tools in Specific Programming Languages

Bioinformatics makes use of a number of programming languages, including C++, Perl, Java, Python, XML, Ruby and Lisp. Worth noting here is the development of the various Bio*.org projects that now cluster under the umbrella group called the Open Bioinformatics Foundation (http://www.open-bio.org/), which was incorporated in October 2001. Each of these projects is an international association of developers of open source tools (software programs or program modules) for bioinformatics, genomics and life science research written in their particular language. Each association attempts to archive, mirror or provide pointers to any and all biology-related code in their specific language that is freely available for download. BioPerl.org was the first of these projects.

2.5 E-Mail Discussion Group Or Usenet Newsgroup

Thousands of e-mail discussion groups or Usenet newsgroups, which cover a wealth of topics are available over internet. By registering in this forum one can discuss with others counterparts and observe discussions made by others. Belonging to these groups is somewhat like receiving a daily newspaper on topics of interest. These groups provide a chance to know the latest developments in ones subject area. In addition, they help to find out how to locate required information.

E-mail discussion groups are managed by software programs. There are three in common use: Listserv, Majordomo, and Listproc. The commands for using these programs are similar.

2.5.1 Biotechnology in Food and Agriculture http://www.fao.org/biotech/forum.asp

The FAO Electronic forum on Biotechnology in Food and Agriculture was established in March 2000 to provide an open forum that will allow a wide range of parties, including governmental and non-governmental organizations, policy makers and the general public, to discuss and exchange views and experiences about specific issues concerning biotechnology in food and agriculture for developing countries. This is done through a series of moderated e-mail conferences, each running for a limited time period only, on specific topics, for which background (before the conferences) and summary documents (after) are produced. A glossary of biotechnology terminology and contact information is also provided

2.5. 2 Molecular Biology Techniques Forum http://www.nwfsc.noaa.gov/protocols/methods/methods.html

This is a professional forum for reader-submitted techniques, tips and questions concerning molecular biology. User can post questions or methods to the forum. It has four categories:

1) **DNA/PCR methods**: for questions directly related to DNA manipulation (isolation, purification, sequencing, etc) and questions regarding various PCR methodologies
2) **RNA Methods**: Use this category for the exchange of ideas, methodologies and references regarding the isolation, manipulation and analysis of RNA. (i.e.: Extraction protocols, Nothern Blot analysis, RNase Protection, Differential Display, In Vitro Transcription, etc.) 3) Protein methods: Use this category for questions regarding problems manipulating proteins in molecular biology applications (expression, detection, etc)
3) **General Methods**: Use this category for general molecular biology questions that don't fit specifically into any of the categories above, including software questions

2.6 Web sites of important organizations

2.6.1 Food and Agriculture Organization of the United Nations (FAO) (http://www.fao.org)

Information about FAO and a link to the World Agricultural Information Centre is included. This Web site is multilingual: French, Spanish, Arabic, and Chinese are available. Through the "Search our site" button, access to databases like AGRIS, Current Agricultural Research Information System (CARIS), AGROVOC (multilingual thesaurus), FAO publications catalog, and more are available and free to all (see also: World Agricultural Information Centre (WAICENT)

2.6.2 Consultative Group on International Agricultural Research (CGIAR) (http://www.cgiar.org)

A total of 16 international agricultural research centers make up this network, covering such subjects as: tropical agriculture, forestry, maize, potatoes, dry land agriculture, aquatic resources, agroforestry, semi-arid tropics, food policy, tropical agriculture, livestock, plant genetic resources, rice and water research, etc.

Links are provided to each center, the center's publications, ongoing research projects, and a collection of photographs.

2.6.3 United States Department of Agriculture (USDA) (USA) http://www.usda.gov

In-depth information such as national agricultural statistics and a Natural Resources (Soil) Conservation Service directory are available on this large and complex portal via "Agencies, Services, and Programs". There are also links to many USDA subdivisions with their own Web sites and a partial subject index.

2.6.4 National Centre for Biotechnology Information http://www.ncbi.nlm.nih.gov/

Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information

2.7 Electronic Journals, Newsletters and Alerting Services

Reputed publishers of journals like Springer-Verlag and the Academic Press provide online access to their journals besides the print edition on registration or subscription. They offer contents and abstracts of their journals online generally free of cost. Full text articles of some of these journals are accessible online free of cost. Certain journals published only on the web and hence do not have a print version.

E-journals can offer value added features over printed publications. For instance, the “ Public Access Computer Systems Review” enables authors to ‘revisit’ their articles in order to update them. The articles, which are very often visited in an archive, can be monitored. The advent of electronic journals facilitates correspondence of readers with the original author. They can be easily searched and they also provide e-mail alerting services contents to registered users.

2.7.1 URLs of some freely available journals:

Biotechnology and Development Monitor http://www.biotech-monitor.nl/
Electronic Journal of Biotechnology http://www.ejb.org/
Issues in Science and Technology Online http://www.nap.edu/issues/
Journal of Agricultural Economics http://www.ncgr.org/jag/
Agricultural outlook
http://www.ers.usda.gov/publications/agoutlook/archives
Current Science http://www.ias.ac.in/currsci
The Scientist www.the-scientist.com

2.8 Online Public Access Catalogues (OPAC) on the Internet.

Most of the world’s best libraries have put their library catalogues on the Internet, which provide access to a large volume of bibliographic information. For example: LOCIS (Library of Congress Information System) (http://www.loc.gov/ catalog/), The British Library Public Catalog (http://blpc.bl.uk/)

2.9 Virtual Libraries

The Internet resulted in the birth of a new type of libraries called the “virtual libraries” providing information a click away. The Online Books Page (http://digital.library.upenn.edu/books/index.html) is a directory of books that can be freely read right on the Internet . Internet Public Library (www.ipl.org) is a public service organization and a learning/teaching environment at the

University of Michigan School of Information, which provides access to about 20,000 titles.

2.10 General Reference Works

A variety of general and special dictionaries, encyclopedias, almanacs, maps, atlases etc. are available online.

Examples:

BioTech Life Science Dictionary
http://biotech.icmb.utexas.edu/search/dict-search.html
A useful online dictionary covering terms from biochemistry, biotechnology, botany, cell biology and genetics)
Access Excellence Graphics Gallery -
http://www.accessexcellence.org/AB/GG/
Graphics Gallery is a series of labeled diagrams with explanations representing the important processes of biotechnology.
Genomics Glossary - http://www.genomicglossaries.com/
Actually a collection of several glossaries and taxonomies, including a Bioinformatics Glossary.
US Patent Information http://www.sciweb.com/features/patents.cfm
Encyclopedia Britannica http://www.britannica.com
Information Please (Almanacs) http://www.infoplease.com/
Merriam-Webster Online www.m-w.com/

3.0 CD-ROM Databases

The CDROMs which emerged as a better way to play music has now became a powerful tool for the dissemination of information. Some databases of agricultural interest are mentioned below.

3.1 AGRIS

It is the international information system for the agricultural sciences and technology. The system identifies worldwide literature (both conventional and non-conventional; the so-called "gray" literature), dealing with all aspects of agriculture and publishes in electronic format.

3.2 AGRICOLA

This is a machine-readable database of bibliographic records created by the National Agricultural Library and its cooperators. The records describe publications and resources encompassing all aspects of agriculture and allied disciplines, including plant and animal sciences, forestry, entomology, soil and water resources, agricultural economics, agricultural engineering, agricultural products, alternative farming practices, and food and nutrition. Auxiliary subjects that support NAL's Information Center activities, such as agricultural trade and marketing, rural information, and animal welfare, are also included.

3.3 TROPAG and RURAL

The Royal Tropical Institute of the Netherlands in the framework of the European Consortium for Agricultural Research in the Tropics: The TROPAG database corresponds to the monthly publication Agriculture and Environment for Developing Regions, and the RURAL database covers social and economic development in rural areas of developing countries during 1985 through 1994.

3.4 NISC discover

Biotechnology abstracts for the period 1983 to 1999 published by National Information Services Corporation, Baltimore

3.5 CAB Abstracts

CAB ABSTRACTS is a bibliographic database compiled by CAB International. It covers the significant research and development literature in the fields of agriculture, forestry, aspects of human health, human nutrition, animal health and the management and conservation of natural resources. Over three million records have been added to the database since its computerization in 1973. These records are made available through a wide range of products and services both printed and electronic.

3.6 Current contents

Published by Institute of Scientific Information, which brings out contents pages of latest issues of important journals published in the field of agriculture, biology and environmental sciences. This is published in print, in floppies and on CD.

4.0 Conclusion

The information resources described above are not exclusive. It only provides an indication about the modern information sources

INFORMATION RETRIEVAL

1.0 Introduction

There are a number of features that are common to all information retrieval systems including CD-ROMS, Internet or OPAC (Online Public Access Catalogue). Indexing techniques help us to access information from these information retrieval systems. Controlled or free indexing languages are used to retrieve information depending on the interface design. Different users will have different information needs and it is necessary to employ different search strategies to get relevant output. One should specify the correct combination of terms for successful retrieval of information.

2.0 Formulation of Queries

Three steps are involved in any computer database search:

2.1 Identifying the concepts

While conducting any database search, the topic may be broken down into its component concepts. For example, if you want to find information on 'Coconut – Cacao Intercropping' the component concepts will be '**coconut', 'cacao'** and '**intercropping'**

2.2 Listing the keywords for each concept

List the keywords to describe each concept after identifying the concepts. Some concepts may have single keyword, while others may have many. Here one should consider synonyms, plurals and alternative spellings etc.
For example:

The concept ***Coconut*** will have keywords **cocos nucifera**, **coconuts** in addition to the key word **coconut**

The concept ***Cocoa*** will have key words **Cacao**, **theobroma cacao** in addition to cocoa

The concept ***Intercropping*** will have key words **Intercropping, multiple cropping**

Depending on the focus of search, there may be other keywords.

2.3 Specifying the logical relationships among your keywords

Once the keywords are known, establish the logical relationships among them. The formal name for this technique is Boolean logic. Boolean logic allows you to specify the relationships among search terms by using any of three logical operators: **AND, OR, NOT.**

Coconut AND Cacao	Files containing both these terms
Coconut OR Cacao	Files containing at least one of these terms
Coconut NOT Cacao	Files containing the term coconut but it should not contain the term cacao

In most CD-ROM database search softwares, combining terms with **AND, OR** and **NOT** is possible, whereas web search engines differ greatly in their support to Boolean logic. Some search engines offer Boolean searching without mentioning the logical operators by name. If we list different search terms and choose **ALL** those terms for searching it denotes **AND** logic. Specifying **ANY** of these terms denotes **OR** logic. Most search engines use a type of implied Boolean logic, in which **symbols** or **spaces** are used to denote logical relationships. For example, **+coconut +breeding** denotes **AND** logic. **+coconut –cacao** denotes **NOT** logic.

Certain search interfaces allow you to use a **proximity operator**. This is a type of **AND** logic which specifies the distance between words in a ource file. The Internet search engines **AltaVista** and **Lycos** allows the **NEAR** operator. For example, consider a search: **physiology NEAR morphology.** In AltaVista, the two terms must appear within 10 words apart from each other in the source file whereas in Lycos user can specify the distances. Use of this option helps to gain relevance in search results.

3.0 Field Searching

Electronic records, like print records, are organized into separate fields. A typical web page or CD database is composed of different fields . The following major fields may be available in web page: **Title, Domain, Host (or site), URL, and links to other pages**. A CD-ROM database may contain different fields like **Title, Author, Source, Year of publication** etc. Some search engines/softwares allow retrieving the records by specifying the field label in combination with your search term(s). Field searching on the Web/CD-ROM is a powerful tool, which allows specifying exactly where the search engine/software has to look in the document for the required information.

- In the case of a Web Site, **title** is the word or words that appears in the blue banner at the very top of your browser's window. If one knows the subject of a page, it's a good bet that important words describing that subject will appear within its title. Searching for a keyword in the title field, rather than as a keyword alone, will very likely produce more relevant responses.

  Example: **title:"coconut breeding"** (Altavista)
  allintitle:biocomputing (Google)
  returns pages that have these words in the title. (Title searches may not work very well with one-word title entries in the case of Web.)
  Coconut in TI (words in article title in WinSPIRS)
- To seek information from a particular kind of site, or a particular site one can limit field search to one of the current top level domains or the site itself.

  Example:
  domain:edu AND "plantation crops" AND intercropping (Altavista)
  site:cgiar.org coconut (Google)
- You can also limit your search to a particular language or year of publication (in CD Search) or updated date (Internet search)
  Example:
  Coconut AND (LA="ENGLISH") AND (PY > "1990") (WinSPIRS)
  In Search Engines you can use the **Advanced Search** to limit your search.

4.0 Tips on Conducting Searches

- Read the directions at each search site or CD-ROM search software since the technique for formulating a search depends on the search engine/software, which vary in available options.
- For a multi-term search, determine the Boolean logic as they may differ with respect to the search engine/software. For example, a search about the relationship between coconut and cacao can be formulated as: **+coconut +cacao** or as **coconut AND cacao**
- Most engines require the operators (AND, OR, AND NOT/NOT) capitalized. Other engines will accept either CAPS or lower case, so it is safe to stick to CAPS.
- Use synonyms or alternate spellings in search statements and connect these terms with **OR** logic.
- Group portions of Boolean queries together using brackets "( )". Example **(coconut or coconuts) and (intercropping or multicropping)**
- Use double quotation marks (" ") to search for a phrase
- Use truncation and wildcards (e.g., *, ?) to look for variations in spelling and word form.
- Spell the terms correctly.

- Include most important keywords first in the string.
- Type keywords and phrases in lower case to find both lower and upper case versions. Usually upper case letters will return only an exact match.
- Repeat the search using alternative terms in case results are not satisfactory.
- In the case Web searching, experiment with different search engines. No two search engines work from with the same index.
- During web search, try search engines, which allow searching multiple search engines simultaneously.
- In case of many results, or irrelevant results
  - Make use of field search
  - Add concept words or repeat the search within the result (Some search engines like Hotbot or Lycos or Google extend this facility)
  - Link appropriate terms with the Boolean AND (+) so that each term is required to appear in the record.
  - Use term proximity operators if they are available to locate documents in which terms are close together. (Lycos offers a number of term proximity options.)
  - If one of your search terms is a phrase, be sure to enclose it within quotations, e.g., "Environmental Biotechnology"
  - Use the Boolean NOT to keep out irrelevant records.
- If you have too few results:
  - Drop off the least important concept(s) to broaden your subject
  - Use more general vocabulary
  - Add alternate terms or spellings for individual concepts and connect with the Boolean OR
  - Try the option available on some engines to find related documents to one or more of your relevant hits. (Google, MetaMission and ProFusion are among the engines that offer this type of feature.)

5.0 Web Search Engines

An Internet search engine allows the user to enter keywords relating to topic and retrieve information about Internet sites containing those keywords. A search engine is a searchable database of Internet files collected by a computer program (called a wanderer, crawler, robot, worm, spider). Index is created from the title, full text, size, URL, etc of the collected files. There are no selection criteria for the collection of files, though evaluation can be applied to ranking schemes that return the results of a query.

There are two types of search engines viz. Individual search engines and Meta search engines.

5.1 Individual Search Engines

Individual search engines compile their own searchable databases on the web. They may be general or specialized search engines based on the types of files/sites they index.

Examples:

General

Google (http://www.google.com)
Teoma (http://www.teoma.com)
Altavista (http://www.altavista.com)
AskJeeves (http://www.askjeeves.com)

Specialised

FTP Search: These search engines search for FTP sites on thenet.
FTP Find (http://www.ftpfind.com)
FTP Search (http://www.ftpsearch.com)

Multimedia & Images: They can be used for searching for multimedia and image file.Lycos Multimedia Search **(http://richmedia.lycos.com/)**
AltaVista Photo Finder (http://image.altavista.com/cgi-bin/avncgi)
FindSounds.com (http://www.findsounds.com/)

Scirus (http://www.scirus.com/)

Scirus is a comprehensive science specific search engine. Elsevier Science, the leading international publisher of scientific information, launched it. Both web sites and journal sources are indexed by Scirus
AgriSurf (http://www.agrisurf.com/)
This Web site is very extensive, searchable agriculture focused search engine and directory and includes links in 95 countries. The sites are selected by agriculture experts.

5.2 Meta Search Engines

Metasearch engines do not crawl the web compiling their own searchable databases. Instead, they search the databases of multiple sets of individual search engines simultaneously, from a single site and using the same interface. Metasearchers provide a quick way of finding out which engines are retrieving the best results for you in your search.

Examples:
Ixquick (http://www.ixquick.com/)
Metor (http://www.metor.com/)
Metacrawler (http://www.metacrawler.com)

6.0 Subject directories

An increasing number of universities, libraries, companies, organizations, and even volunteers are creating subject directories to catalog portions of the Internet. These directories are organized by subject and consist of links to Internet resources relating to these subjects. The major subject directories available on the Web tend to have overlapping but different databases. Most directories provide a search capability that allows you to query the database on your topic of interest.

Directories are useful for general topics, for topics that need exploring, and for browsing. Subject directories come in assorted flavors. There are general directories, academic directories, commercial directories, portals and now, vortals. INFOMINE (http://infomine.ucr.edu/) is a large directory of Web sites of scholarly interest compiled by the University of California. The WWW Virtual Library (http://www.vlib.org/) is one of the oldest and most respected subject directories on the Web. This directory consists of individual subject collections, many of which are maintained at universities throughout the world.

Portals are directories that have been created or taken over by commercial interests and then reconfigured to act as gateways to the web. These portal sites not only link to popular subject categories, they also offer additional services such as email, current news, stock quotes, travel information and maps. Yahoo! (www.yahoo.com) is the most famous example of a commercial portal.

Vortals, or vertical portals, (eg. xrefer (http://www.xrefer.com/ - science specific portal) are subject-specific directories, as opposed to the broader, more generalized and varied collection of subjects and sites.
Subject directories differ significantly in selectivity. For example, the famous Yahoo! site does not carefully evaluate user-submitted content when adding Web pages to its database. *It is therefore NOT a reliable research source and should not be used for this purpose.* In contrast, INFOMINE selects only those sources considered useful to the academic and research community. Consider the policies of any directory that you visit.

7.0 Deep Web

The concept of the "deep" or "invisible" Web has emerged recently. This refers to content that is stored in databases accessible on the Web but not available via search engines. In other words, this content is "invisible" to search engines. This is because spiders cannot or will not enter into databases and extract content from them as they can from static Web pages. In the past, these databases were fewer in number and referred to as specialty databases, subject specific databases, and so on.
The only way to access information on the invisible Web is to search the databases themselves. As the number of Web-accessible databases grows, it will become essential that they be used to conduct successful information finding on the Web.

Other content not gathered by spiders includes non-textual files such as multimedia files, graphical files, and documents in non-standard formats such as Portable Document Format (PDF).

Keep in mind that many search engine sites and commercial portals feature searchable databases as part of their package of services. This phenomenon falls under the heading of converging content. For example, you can visit AltaVista and look up news, maps, jobs, auctions, items for purchase, etc., all things outside the purview of a spider- gathered index. As another example, Google integrates searches of PDF and Microsoft Office files into its general search service.

Examples of sites that collect content from the deep Web:

The Invisible Web http://www.invisibleweb.com/
Invisible-web.net http://www.invisible-web.net/
Search.Com http://www.search.com/
WebData http://www.webdata.com/

8.0 How Search Engines Work

Search engines compile their databases by employing "spiders" or "robots" (specially created software programs) to crawl through web space from link to link, identifying and perusing pages. Sites with no links to other pages may be missed by “spiders” altogether. Once the spiders get to a web site, they typically index most of the words on the publicly available pages at the site. Web page owners may submit their URLs to search engines for "crawling" and eventual inclusion in their databases.

It is important to remember that when you are using a search engine, you are NOT searching the entire web, as it exists at this moment. You are actually searching a portion of the web, captured in a fixed index created at an earlier date.

Search engines are the best means devised yet for searching the web. Stranded in the middle of this global electronic library of information without either a card catalog or any recognizable structure, there is no other way to find information.

On the down side, the sheer number of words indexed by search engines increases the likelihood that they will return hundreds of thousands of irrelevant responses to simple search requests. Remember, they will return lengthy documents in which your keyword appears only once.

8.1 Points to remember while using search engines

- Search engines do not index all the documents available on the Web.
- The content of databases generally will not show up in a search engine result.
- Search engine features are proliferating and are in a constant state of flux.
- Search techniques may differ according to the search engine
- Most major search engine indexes consist of the full text of source files..
- Many search engines have an interface for basic searches as well as a separate interface for advanced or more full-featured queries
- Search engines don't always pay attention to everything included in your search statements.
- Because only partially relevant files are often returned as a result of your search, good relevancy ranking is important.
- Don't be impressed by a large number of hits in response to a search
- Most search engines employ the principles of Boolean logic in the formulation of search queries.
- Search engines tend to have a default Boolean logic

9.0 Conducting CD-ROM Search

A large number of publishers are producing their databases in CD-ROMs. CD-ROM products have been specifically designed for ease of use and facilitate end-user searching. Retrieval facilities are similar to those that might be expected in any information retrieval product. They support Boolean logic, truncation using wild cards, field specific searching and phrase searching. They provide retrieval software for searching the databases, which have a user-friendly interface. Most of these interfaces are menu based. Unfortunately there is little standardization among the products from different database suppliers and even among the products from a specific supplier. Users are likely to encounter a range of different softwares if they search more than one CD-ROM. This will have different search facilities and different interfaces.

10.0 Conclusion

Search strategy for information retrieval is basically the same in the case of Internet and CD-ROM database searching. The main difference is in the search interface and on the type of information retrieved. Most of the CD-ROM databases contain bibliographic information while a variety of information accessed over the Web. For retrieval of information from a CD-ROM database specific retrieval softwares are required and search engines take up this job in the case of Internet.

AN INTRODUCTION TO DNA ANALYSIS SOFTWARES

DEVAKUMAR, K
BIOTECHNOLOGY SECTION, CPCRI, KASARAGOD

1.0 Introduction

DNA markers are increasingly being used for the analysis of plant genome organisation, genetic relationships, diversity estimation, genome characterisation etc.

An array of molecular markers is available for plant genome analysis. RFLP (Restriction Fragment Length Polymorphism), RAPD (Randomly Amplified Polymorphic DNA), AFLP (Amplified Fragment Length Polymorphism), SSR (Simple Sequence Repeats), DAF (DNA amplification fingerprinting) are some of the widely used markers. RFLP and SSR are co-dominant markers whereas other markers are dominant markers.
The enormous amount of data generated by molecular markers makes data analysis and interpretation a stupendous task. A number of computer programmes are available, which can be used to analyse large data sets.

2.0 RAP DISTANCE

This programme is mainly for the analysis of patterns of RAPD fragments. It calculates pairwise distances between the samples and output is stored in several formats for other programmes. There are also programmes for PTP (Permutation Tail Probability) and assessment of the significance of trees generated by the neighbour joining method.

Following are the important commands for this software.

The program is started by the command

```
C:\rapd\rapd <enter>
```

Main menu will appear, which has four submenus

M1: For the RAPDistance input/editing programmes
M2: For distance calculation programme
M3: For analysis programs
M4: for help
Q: Quit

For selecting the option, type letters and number and enter. E.g. type M1 for evoking input program.

Option 11 in M1 submenu stores data in a RAPdistance datafile.

The program will ask for

1. a name for the datafile that is to hold the data.
2. the number of samples;
3. a name for each sample;
4. the number of populations into which the samples are to be grouped.
5. the number of samples in each population

6. the number of the primers used for generating the data
7. the name of each primer
8. the length of the primer
9. the total number of bands generated by each primer

The program then prompts you to record whether each of the defined bands is present '1' or not '0'.

The entered data is held in file 'filename.dat'.

This file can be edited or converted to an ASCII file for printing using different options.

Option 21 in M2 submenu is used to calculate pairwise distance and similarity matrix.
It calculates pairwise distances between the DNA samples using the band data from a RAPDisatnce datafile, based on any of the 18 metrices and provides the results as triangular matrix in different file format.

*.NJT required by some Njtree and UPGMA program.
*.DIP required by the DIPLOMO matrix comparison programs
*.DIM for use with DIPLOMO
*.PHY required by PHYLIP package
*.NTS require by the NTSYS package
*.DIS required by WINOMA

Option 31 in submenu M3 generates a neighbor joining tree using a pair distance file in *.NJT format. The output file NJTREE.OUT contains the inter-out/node branching order and length.

NJTREE.OUT is then used to draw the dendrogram by using TDRAW command. The output of TDRAW is an ASCII file and a postscript file.
The postscript file is printed by copying the file to printer (with postscript menu) with the command "copy /B 'filename' Prn: ".

3.0 PHYLIP

It is a package of programmes for inferring phylogenies and carrying out certain related task. It contains programmes for molecular sequence data, distance matrix data, gene frequencies and continuous characters, discrete date and plotting trees and consensus trees.

http://evolution.genetics.washington.edu/phylip.html

4.0 NTSYS-PC

It is a system of programmes that is used to find and display structure in multivariate data. It can be used to compute various measures of similarity and dissimilarity between all pairs of objects and then summarize this information either in terms of nested sets of similar objects (cluster analysis) or in terms of a spatial arrangement along one or more coordinate axes (Ordination analysis or various types of multidimensional scaling analysis).

INPUT file: It contains four kinds of records.

Comments: These optional lines are used to identify the data.
Matrix parameter line: This line contains 4 integer numbers and possible a real number

A. The number is a code for the type of matrix.
1 rectangular data matrix
2 symmetric dissimilarity matrix
3 symmetric similarity matrix
4 diagonal matrixes

B. Second and third numbers are the numbers of rows and columns in the matrix
C. The fourth number is to indicate whether missing data is present.

Row and column labels: Labels must be furnished if a B, E or L is placed after the numbers of rows or and L after the number of columns in the previous line.

Matrix dateline: The elements of the matrix are entered with rows in the input matrix corresponding to the one or more lines in the input file Values must be separated by one or more blanks or commas in some applications.

Cluster analysis: The most common use of NTSYS-PC is for performing various types of agglomerative cluster analysis of some type of similarity or dissimilarity matrix.

The following are the commands for cluster analysis compute a distance matrix using **QUALITATIVE** option

Do a UPGMA analysis of the distance matrix using **SAHN** option
Display phenogram using **TREE** option

Principal coordinate analysis (PCOORDA): It is one of the most important methods of ordination analysis. It constructs a new set of orthogonal coordinate axes such that the projection of points onto them have maximum variance. Different steps are:

Standardise the data using **STANDARDISATION** option
Compute distance matrix among objects using **SIMINT** option
Double center the distance matrix using **DOUBLE CENTRE** option
Extract eigen vectors using **EIGENVECTOR** option
Display the projection using **MOD3D** option

5.0 BIOSYS 2

It is a computer programme for the analysis of electrophoretically detectable allelic variation. This performs most types of electrophoretic data analysis commonly employed in biochemical population genetics and systematics. It can be used to compute allele frequencies and genetic variability measures, to test for deviation of genotype frequencies from Hardy-Weinberg expectations, to test for linkage disequilibrium between alleles at pairs of loci, to calculate F-statistics, to perform heterogeneity chi-square analyses, to calculate a variety of similarity and distance coefficients, to construct phenograms using cluster analysis, and to estimate phylogenies by the distance Wagner procedure.

6.0 ARELEQUIN

It provides with a large set of methods and statistical tests, in order to extract information on genetic and demographic features of a collection of population samples. Arlequin can handle several types of data either in haplotypic or genotypic form. The basic data types are / DNA sequences, RFLP data, Micro-satellite data, standard data and allele frequency data.

This software is available at http://anthro.unige.ch/arlequin

7.0 GENEPOP

Genepop performs three major tasks (1) It computes exact tests for hardy weinberg equilibrium for population differentiation and for genotypic disequilibrium among pairs of loci, (2) It computes estimates of classical population parameters, such as FST and other correlation, allele frequencies etc. and (3) It converts the input GENEPOP file to formats used by other programmes.

The software is available at http://www.cefe.cnrs-mop.fr/.

8.0 Conclusion

These softwares have been widely used for molecular marker data analysis. The use of these softwares makes analysis of the data much easier and allow the researcher to interpret their results from several angels.

GENOMICS AND ITS APPLICATION

V.A PARTHASARATHY
DIRECTOR,IISR CALICUT

1.0 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.

2.0 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), deoxyThymidinemonophosphate (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 64different 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 soon). 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.

2.1 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 blue print 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 3.2billion 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.

2.2 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.

2.3 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.

3.0 Genomics and Bioinformatics

Various genome projects are generating gigantic databases 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 *denovo*, *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.

4.0 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.

Bioinformatics 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.

Active site

The amino acid residues at the catalytic site of an enzyme. These residues provide the binding and activation energy needed to place the substrate into its transition state and bridge the energy barrier of the reaction undergoing catalysis

Adenine

A purine base found in DNA and RNA

Agents

Independent, autonomous, software modules that can search the Internet for data or content pertinent to a particular application, such as a gene, protein, or biological system.

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.

Amino acid
One of the 20 chemical building blocks that are joined by amide (peptide) linkages to form a polypeptide chain of a protein

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.

Antigen
Any foreign molecule that stimulates an immune response in a vertebrate organism. Many antigens are proteins such as the surface proteins of foreign organisms.

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.

Assay
A method for measuring a biological activity. This may be enzyme activity, binding affinity, or protein turnover. Most assays utilize a measurable parameter such as color, fluorescence or radioactivity to correlate with the biological activity.

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.

Autoradiography
A method used to locate radioisotope-labeled materials which have been separated in gels or are present in blots. The location of the radiolabeled material is determined by overlaying the test material with a photographic film that is sensitive to the radioisotope.

B

Bacterial artificial chromosome (BAC)
Cloning vector that can incorporate large fragments of DNA. (see YACS)

Bacteriophage
A virus that infects bacteria. The bacteriophage DNA has served as a basis for cloning vectors, and is also utilized to create phage libraries containing human or other genes.

Baculovirus
An insect virus which forms the basis of a protein expression system

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.

Bioinformatics
The application of computer and information sciences to (complex) biological questions. Generally the term bioinformatics applies to molecular bioinformatics, but logic demands that analysis of biological data on the organismal or population scale also qualify as bioinformatics. Thus the study of T-cell fluctuations in AIDs patients or population dynamics of meadow voles amount to bioinformatics; and similarly depend upon computational techniques. Probably the most significant non-molecular bioinformatics concerns the nervous systems of higher animals, and ìneuroinformaticsî has already appeared as a term in the literature (considering that a human brain has on the order of 100 billion neurons, itís easy to see why). [*Cf.* computational biology

Bivalent
Having two binding sites; having 2 free electrons available for binding.

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
Carboxyl group
The -COOH functional group, acidic in nature, found in all amino acids

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.

Cell
The basic unit of any living organism.

Cell Cycle
The life cycle of a cell which is marked by cell division which is separated into four phases: G1, S, G2, and M. DNA replication is confined to the S(synthesis) phase, and chromosomal separation in the M (mitotic) phase**.**

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.

Clinical trials
Research studies that involve patients. Biotechnology companies typically use clinical trials to assess the efficacy and safety of new therapies and to answer scientific questions. Typically, there are 3 phases during a clinical trial. Phase I is designed to evaluate the safety of the product in humans; phase II analyses the effects of dose escalation, and phase III definitively evaluates the clinical efficacy of the product.

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.

Combinatorial chemistry
The use of chemical methods to generate all possible combinations of chemicals starting with a subset of compounds. The building blocks may be peptides, nucleic acids or small molecules. The libraries of compounds

formed by this methodology are used to probe for new pharmaceutical reagents (see high-throughput screening).

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.

Cosmids
DNA vectors that allow the insertion of long fragments of DNA (up to 50 kbases).

Crystal structure
Term used to describe the high resolution molecular structure derived by x-ray crytallographic analysis of protein or other biomolecular crystals.

Cytoplasm
The medium of the cell between the nucleus and the cell membrane.

Cytosine
A pyrimidine base found in DNA and RNA.

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).

Disulphide bond
Covalent link formed between the sulphur atoms of two different cysteine residues in a protein. Important in maintaining the folded structure of a protein, and also for linking different proteins in a complex.

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 microarrays
The deposition of oligonucleotides or cDNAs onto an inert substrate such as glass or silicon. Thousands of molecules may be organized spatially into a high-density matrix. These DNA chips may be probed to allow expression monitoring of many thousands of genes simultaneously. Uses include study of polymorphisms in genes, de novo sequencing or molecular diagnosis of disease.

DNA polymerase
An enzyme that catalyzes the synthesis of DNA from a DNA template given the deoxyribonucleotide precursors.

DNA probes
Short single stranded DNA molecules of specific base sequence, labeled either radioactively or immunologically, that are used to detect and identify the complementary base sequence in a gene or genome by hybridizing specifically to that gene or sequence.

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
An agent that affects a biological process. Specifically, a molecule whose molecular structure can be correlated with its pharmacological activity.

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.

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.

Electrophoresis
The use of an external electric field to separate large biomolecules on the basis of their charge by running them through acrylamide or agarose gels.

Enhancers
DNA sequences that can greatly increase the transcription rates of genes even though they may be far upstream or downstream from the promoter they stimulate.

Enzyme
A class of proteins that are capable of catalyzing chemical reactions (the making or breaking of chemical bonds). They do so by orienting their substrates into a suitable geometry in a particular location (the active site) where electrophilic or nucleophilic amino acid residues can participate in the reaction. Enzymes are protein catalyst that speeds up chemical reactions that would otherwise be prohibitively slow under physiological conditions.

Epigenomics
The study of complex expression networks or linkages both spatially (within the body) and temporally (at different times in development).

Equilibrium constant
Value that describes the equilibrium state of the reversible reaction between two molecular species.

Eukaryote
A cell or organism with a distinct membrane-bound nucleus as well as specialized membrane-based organelles (see also prokaryote).

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.

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

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.

Fusion protein

The protein resulting from the genetic joining and expression of 2 different genes (see chimeric)

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.

Gel electrophoresis
A technique by which molecules are separated by size or charge by passing them through a gel under the influence of an external electric field.

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.

Glycosylation
The addition of carbohydrate groups (sugars) e.g. to polypeptide chains
Guanine (G)
One of the nitrogenous purine bases found in DNA and RNA

H

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.

HLA complex
Another name for the MHC in humans; refers to the "Human Leukocyte Antigen" complex located on chromosome 6.

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.

Human Anti-Murine Antibody Response (HAMA)
An immune response generated in humans to antibodies raised in murine (e.g. mouse or rat) cells.

Hybridization
The interaction of complementary nucleic acid strands. This can occur between two DNA strands or between DNA and RNA strands, and is the basis of many techniques such as Southern and northern blots.

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

Idiotype
Antibody variants localized to the variable portion of an immunoglobulin that are recognised by their antigenic determinants.

Immunoglobulin
A member of the globulin protein family consisting of two light and two heavy chains linked by disulfide bonds. All antibodies are immunoglobulins.

in situ hybridization

A variation of the DNA/RNA hybridization procedure in which the denatured DNA is in place in the cell and is then challenged with RNA or DNA extracted from another source. (See also fluorescence in situ hybridization).

Integration
The physical insertion of DNA into the host cell genome. The process is used by retroviruses where a specific enzyme catalyses the process or can occur at random sites with other DNA (eg. transposons).

Intracellular signalling
The communication of a molecular message from the surface of the cell to the nucleus via the participation of a series of molecules, including receptors, enzymes, proteins, and small-molecules.

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.

Iteration
A series of steps in an algorithm whereby the processing of data is performed repetitively until the result exceeds a particular threshold. Iteration is often used in multiple sequence alignments whereby each set of pairwise alignments are compared with every other, starting with the most similar pairs and progressing to the least similar, until there are no longer any sequence-pairs remaining to be aligned.

J

Junk DNA
Term used to describe the excess DNA that is present in the genome beyond that required to encode proteins.

K

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.

Lead compound
A candidate compound identified as the best "hit" (tight binder) after screening of a combinatorial (or other) compound library, that is then taken into further rounds of screening to determine its suitability as a drug.

Lead optimization
The process of converting a putative lead compound ("hit") into a therapeutic drug with maximal activity and minimal side affects, typically using a combination of computer-based drug design, medicinal chemistry and pharmacology.

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.

M

Map unit
A measure of genetic distance between two linked genes that corresponds to a recombination frequency of 1%.

Markov chain
Any multivariate probability density whose independence diagram is a chain.

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.

Microfluidics
The miniaturization of chemical reactions or pharmacalogical assays into microscopic tubes or vessels in order to greatly increase their throughput, by placing many of them side-by-side in an array.

Missense mutation

A point mutation in which one codon (triplet of bases) is changed into another designating a different amino acid.

Mitosis
The nuclear division that results in the replication of the genetic material and its redistribution into each of the daughter cells during cell division.

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)

Monomer
A single unit of any biological molecule or macromolecule, such as an amino acid, nucleic acid, polypeptide domain, or protein.

Monovalent
Having one binding site; strictly, an atom with only one free electron available for binding in its highest energy shell.

Multigene family
A set of genes derived by duplication of an ancestral gene, followed by independent mutational events resulting in a series of independent genes either clustered together on a chromosome or dispersed throughout the genome.

Multiple (sequence) alignment
A Multiple Alignment of *k* sequences is a rectangular array, consisting of characters taken from the alphabet $\underline{A}$, 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.

Mutagen
Any agent that can cause an increase in the rate of mutations in an organism.

Mutation
An inheritable alteration to the genome that includes genetic (point or single base) changes, or larger scale alterations such as chromosomal deletions or rearrangements.

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.

Nonsense mutation
A point mutation in which a codon specific for an amino-acid is converted into a nonsense codon.

Northern blotting
A technique to identify RNA molecules by hybridization that is analogous to Southern blotting (see Southern blotting).

Nuclease
Any enzyme that can cleave the phosphodiester bonds of nucleic acid backbones.

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.

Oligonucleotide
A short molecule consisting of several linked nucleotides (typically between 10 and 60) covalently attached by phosphodiester bonds.

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.

Operator
A segment of DNA that interacts with the products of regulatory genes and facilitates the transcription of one or more structural genes.

Operon
A unit of transcription consisting of one or more structural genes, an operator, and a promoter.

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

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.

Peptide

A short stretch of amino acids each covalently coupled by a peptide (amide) bond.

Peptide bond (amide bond)
A covalent bond formed between two amino acids when the amino group of one is linked to the carboxy group of another (resulting in the elimination of one water molecule).

Phage (Bacteriophage)
A virus that infects bacterial cells and serves as a useful vector for introducing genes into bacteria for a number of purposes.

.

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.

Phenotype
Any observable feature of an organism that is the result of one or more genes.

Phylum
The segmentation of the animal kingdom into about 30 major groups collectively known as phyla.

Physical map
A physical map consists of a linearly ordered set of DNA fragments encompassing the genome or region of interest. Physical maps are of two types, macro-restriction maps and ordered clone maps.

Plasmid
Any replicating DNA element that can exist in the cell independently of the chromosomes. Synthetic plasmids are used for DNA cloning. Most commonly found in bacterial cells.

Pleitropy
The multiple effects on an organismís phenotype due to a single gene or allele e.g the cytokines which can bind to multiple cellular receptors and effect growth and multiple immune pathways.

Point mutation
A mutation in which a single nucleotide in a DNA sequence is substituted by another nucleotide.

Poly(A) tail
The stretch of Adenine (A) residues at the 3í end of eukaryotic mRNA that is added to the pre-mRNA as it is processed, before its transport from the nucleus to the cytoplasm and subsequent translation at the ribosome.

Polyadenylation site
A site on the 3í-end of messenger RNA (mRNA) that signals the addition of a series of Adenines during the RNA processing step and before the mRNA migrates to the cytoplasm.

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.

Polypeptide
A single chain of covalently attached amino acids joined by peptide bonds. Polypeptide chains usually fold into a compact, stable form (a domain) that is part (or all) of the final protein.

Positional cloning
Method used to define the location of a gene on a chromosome and use this information to identify and clone the gene.

Post-transcriptional modification
Alterations made to pre-mRNA before it leaves the nucleus and becomes mature mRNA.

Post-translational modification
Alterations made to a protein after its synthesis at the ribosome. These modifications, such as the addition of carbohydrate or fatty acid chains, may be critical to the function of the protein.

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).

Probe
Any biochemical that is labelled or tagged in some way so that it can be used to identify or isolate a gene, RNA, or protein.
Profile

Sequence profiles are usually derived from multiple alignments of sequences with a known relationship, and consist of tables of position-specific scores and gap-penalties.

Prokaryote
An organism or cell that lacks a membrane-bounded nucleus. Bacteria and blue-green algae are the only surviving prokaryotes (cf. Eukaryote).

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)

Purine
A nitrogen-containing compound with a double-ring structure. The parent compound of Adenine and Guanine.

Pyrimidine
A nitrogen-containing compound with a single six-membered ring structure.
The parent compound of Thymidine and Cytosine.

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.

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).

Reagents
Sources of biological or chemical material that can be used as the starting blocks in laboratory experiments.

Recessive
Any trait that is expressed phenotypically only when present on both alleles of a gene (cf dominant).

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.

Recombination
A new combination of alleles resulting from the rearrangement occuring by crossing-over or by independent assortment (see crossing over).

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.

Regulatory gene
A DNA sequence that functions to control the expression of other genes by producing a protein that modulates the synthesis of their products (typically by binding to the gene promoter). (cf. Structural gene).

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

Replication
The synthesis of an informationally identical macromolecule (e.g. DNA) from a template molecule.

Repressor
The protein product of a regulatory gene that combines with a specific operator (regulatory DNA sequence) and hence blocks the transcription of genes in an operon.

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.

Restriction map
A physical map or depiction of a gene (or genome) derived by ordering overlapping restriction fragments produced by digestion of the DNA with a number of restriction enzymes.

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
A DNA polymerase that can synthesise a complementary DNA (cDNA) strand using RNA as a template - a so-called RNA-dependent DNA polymerase.

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.

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.

Sequence Tagged Site (STS)
A unique sequence from a known chromosomal location that can be amplified by PCR. STSs act as physical markers for genomic mapping and cloning.

Sexual PCR (Molecular Diversity)
Sexual PCR is a form of PCR in which similar, but not identical, DNA sequences are reassembled to obtain novel juxtapositions, simulating the result of genetic recombination.

Shotgun cloning
The cloning of an entire gene segment or genome by generating a random set of fragments using restriction endonucleases to create a gene library that can be subsequently mapped and sequenced to reconstruct the entire genome.

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.

Southern blotting
A procedure for the identification of DNA by transmitting a fragment isolated on an agarose gel to a nitrocellulose filter where it can be hybridized with a complementary "probe" sequence.

Splice site
The sequence found at the 5í and 3í region of exon/intron boundaries, usually defined by a consensus sequence:
Intron
5í CAGGTAAGT---------TNCAGG 3í
A G C T
N represents any nucleotide; the bottom line represents alternative nucleotides at the indicated positions.

Splice form
By using alternative splicing, a single message precursor from DNA can generate an entire family of mRNAs and proteins.

Splicing
The joining together of separate DNA or RNA component parts. For example, RNA splicing in eukaryotes involves the removal of introns and the stitching together of the exons from the pre-mRNA transcript before maturation.

Solvent accessibility
The surface area (typically measured in square angstroms) of a biological molecule, usually a protein, that is exposed to solvent in its native, folded form. Determining the solvent accessibility of a protein helps define which amino acids in its molecular sequence are on the exterior of the molecule, and thus available to participate in interactions with other molecules.

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 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.

Thymine
A pyrimidine base found in DNA but not in RNA.
Tissue
Section of an organ that consists of a largely homogenous population of cell types.

Transcript
The single-stranded mRNA chain that is assembled from a gene template.

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.

Transformation
A genetic alteration to a cell as a result of the incorporation of DNA from a genetically diferent cell or virus; can also refer to the introduction of DNA into bacterial cells for genetic manipulation.

Transgene
A foreign gene that is introduced into a cell or whole organism (eg.transgenic mice) for therapeutic or experimental purposes.

Translation
The process of converting RNA to protein by the assembly of a polypeptide chain from an mRNA molecule at the ribosome.

Transmembrane region
The region of a transmembrane protein that actually spans the membrane.

Tissue
Section of an organ that consists of a largely homogenous population of cell types.

U

Unidentified reading frame (URF)
An open reading frame encoding a protein of undefined function**.**

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.

Western blot
Technique in which specific antibodies are used to identify their antigens from a mixture of proteins.

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.