Glossary of Bioinformatics

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

3' flanking region: A region of DNA which is NOT copied into the mature mRNA, but which is present adjacent to 3' end of the gene (see Figure 4). It was originally thought that the 3' flanking DNA was not transcribed at all, but it was discovered to be transcribed into RNA, but quickly removed during processing of the primary transcript to form the mature mRNA. The 3' flanking region often contains sequences which affect the formation of the 3' end of the message. It may also contain enhancers or other sites to which proteins may bind.

3' untranslated region: A region of the DNA which IS transcribed into mRNA and becomes the 3' end or the message, but which does not contain protein coding sequence. Everything between the stop codon and the polyA tail is considered to be 3' untranslated (see Figure 4). The 3' untranslated region may affect the translation efficiency of the mRNA or the stability of the mRNA. It also has sequences which are required for the addition of the poly(A) tail to the message (including one known as the "hexanucleotide", AAUAAA).

3Dseq Database: Providing annotation of the PDB sequences to a uniform standard, and to provide cross-references to the **SWISS-PROT** database

5' flanking region: A region of DNA which is NOT transcribed into RNA, but rather is adjacent to 5' end of the gene (see Figure 4). The 5'-flanking region contains the promoter, and may also contain enhancers or other protein binding sites.

5' untranslated region: A region of a gene which IS transcribed into mRNA, becoming the 5' end of the message, but which does not contain protein coding sequence. The 5'-untranslated region is the portion of the DNA starting from the cap site and extending to the base just before the ATG translation initiation codon (see Figure 4). While not itself translated, this region may have sequences which alter the translation efficiency of the mRNA, or which affect the stability of the mRNA.

A

Adaptation “AD-ap-TAY-shun”: a feature produced by natural selection for its current function.

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.

Accession Number line (EMBL): The AC (Accession Number) line lists the accession numbers associated with this entry.

accession number line (SWISS-PROT): The AC (ACcession number) line lists the accession number(s) associated with an entry.

Acquired mutations: Gene changes that arise within individual cells and accumulate throughout a person's lifetime; also called somatic mutations

Acrylamide gels: A polymer gel used for electrophoresis of DNA or protein to measure their sizes (in daltons for proteins, or in base pairs for DNA). See "Gel Electrophoresis".

Acrylamide gels are especially useful for high resolution separations of DNA in the range of tens to hundreds of nucleotides in length

Adenine: A purine base found in DNA and RNA

Agarose gels: A polysaccharide gel used to measure the size of nucleic acids (in bases or base pairs). See "Gel Electrophoresis". This is the gel of choice for DNA or RNA in the range of thousands of bases in length, or even up to 1 megabase if you are using pulsed field gel electrophoresis.

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, analyse, visualise and make predictions from biological data.

Alignment: The result of a comparison of two or more gene or protein sequences in order to determine their degree of base or amino acid similarity. Sequence alignments are used to determine the similarity, homology, function or other degree of relatedness between two or more genes or gene products

Alignment score: The alignment score, represents the likelihood that the described alignment is not random, providing an indication of its validity. They are calculated by totaling the scores for each matched pair of residues at each position in the alignment, plus unmatched residues are given the gap open penalty, (the gap penalty for non-affine searches), or the gap extension penalty, if appropriate in the alignment, and if the affine search is running.

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.

Alleles: Variant forms of the same gene. Different alleles produce variations in inherited characteristics such as eye colour or blood type.

Alternative splicing: One of the alternate combinations of a folded protein that are possible due to recombination of multiple gene segments during mRNA splicing that occurs 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

Alzheimer's disease: A disease that causes memory loss, personality changes, dementia and, ultimately, death. Not all cases are inherited, but genes have been found for familial forms of Alzheimer's disease.

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

Amino acids with acidic side chains: These have a carboxylic acid group in their side chain and are very hydrophilic

Amino acids with aliphatic hydrophobic side chains: The hydrophobic side chains of these amino acids will not form hydrogen bonds or ionic bonds with other groups. These hydrophobic amino acids tend to be buried in the centre of proteins away from the surrounding aqueous environment.

Amino acids with basic side chains: The positive charge on these side chains makes them hydrophilic and they are likely to be found at the protein surface

Amino acids with Neutral side chains: The single hydrogen atom side chain has no strong hydrophobic or hydrophilic properties

Amino acids with uncharged but polar side chains: The side chains of these amino acids are uncharged at physiological pH

Annotation: A combination of comments, notations, references, and citations, either in free format or utilising 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.

Antibiotic resistance: Plasmids generally contain genes which confer on the host bacterium the ability to survive a given antibiotic. If the plasmid pBR322 is present in a host, that host will not be killed by (moderate levels of) ampicillin or tetracycline. By using plasmids containing antibiotic resistance genes, the researcher can kill off all the bacteria which have not taken up this plasmid, thus ensuring that the plasmid will be propagated as the surviving cells divide.

Antigenic: An EMBOSS application. Antigenic predicts potentially antigenic regions of a protein sequence, using the method of Kolaskar and Tongaonkar. Analysis of data from experimentally determined antigenic sites on proteins has revealed that the hydrophobic residues Cys, Leu and Val, if they occur on the surface of a protein, are more likely to be a part of antigenic sites. A semi-empirical method which makes use of physicochemical properties of amino acid residues and their frequencies of occurrence in experimentally known segmental epitopes was developed by Kolaskar and Tongaonkar to predict antigenic determinants on proteins. Application of this method to a large number of proteins has shown that their method can predict antigenic determinants with about 75% accuracy which is better than most of the known methods. This method is based on a single parameter and thus very simple to use.

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.

ArrayExpress: **ArrayExpress** is a public repository for microarray based gene expression data.

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.

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.

Autosome: Any of the non-sex-determining chromosomes. Human cells have 22 pairs of autosomes.

B

Background extinction "ek-STINK-shun": the extinction that lineages have "normally" experienced throughout life's history. Even when the Earth is not experiencing a major catastrophe, lineages are constantly going extinct—and this is background extinction. Brief periods (often associated with major geologic or climactic changes) in which extinction rates have been elevated across many lineages are called mass extinctions and are not a part of background extinction.

Backtranseq: An EMBOSS application. backtranseq takes a protein sequence and makes a best estimate of the likely nucleic acid sequence it could have come from. It does this by using a codon frequency table. For each amino acid, the corresponding most frequently occuring codon is used in the construction of the nucleic acid sequence.

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

Banana: An **EMBOSS** application. banana predicts bending of a normal (B) DNA double helix, using the method of Goodsell & Dickerson, NAR 1994 11;22(24):5497-5503

This program calculates the magnitude of local bending and macroscopic curvature at each point along an arbitrary B-DNA sequence, using any desired bending model that specifies values of twist, roll and tilt as a function of sequence. The data, based on the nucleosome positioning data of Satchwell et al 1986 (J. Mol. Biol. 191, 659-675), correctly predicts experimental A-tract curvature as measured by gel retardation and cyclization kinetics and successfully predicts curvature in regions containing phased GGGCCC sequences. (This is the model 'a' described in the Goodsell & Dickerson paper). This model - showing local bending at mixed sequence DNA, strong bends at the sequence GGC, and straight, rigid A-tracts - is the only model, out of six models investigated in Goodsell & Dickerson paper, that is consistent with both solution data from gel retardation and cyclization kinetics and structural data from x-ray crystallography. The consensus sequence for DNA bending is 5 As and 5 non-As alternating. "N" is an ambiguity code for any base, and "B" is the ambiguity code for "not A" so "BANANA" is itself a bent sequence - hence the name of this program. The program outputs both a graphical display and a text file of the results.

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.

Binding site: A place on cellular DNA to which a protein (such as a transcription factor) can bind. Typically, binding sites might be found in the vicinity of genes, and would be involved in activating transcription of that gene (promoter elements), in enhancing the transcription of that gene (enhancer elements), or in reducing the transcription of that gene (silencers). NOTE that whether the protein in fact performs these functions may depend on some condition, such as the presence of a hormone, or the tissue in which the gene is being examined. Binding sites could also be involved in the regulation of chromosome structure or of DNA replication.

Bioinformatics: The field of endeavour that relates to the collection, organisation and analysis of large amounts of biological data using networks of computers and databases (usually with reference to the genome project and DNA sequence information)

Biosed: An **EMBOSS** application. Description: biosed is a simple sequence editing utility that searches for a target sub-sequence in one or more input sequences and replaces it with a specified second sub-sequence (or optionally just deletes the found target sub-sequence).

Blastn program: Blastn will search a DNA sequence against a DNA databank.

Blastp program: Blastp will compare a protein sequence against the protein database of your choice

BlastProDom.pl: A tool that scans the families in the ProDom database. These families are built by an automated process based on recursive use of PSI-BLAST homology searches

Blastx program: Blastx will translate a nucleic acid sequence in all six reading frames and compare all these against the protein database of your choice

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

Blotting: A technique for detecting one RNA within a mixture of RNAs (a Northern blot) or one type of DNA within of a mixture of DNAs (a Southern blot). A blot can prove whether that one species of RNA or DNA is present, how much is there, and its approximate size. Basically, blotting involves gel electrophoresis, transfer to a blotting membrane (typically nitrocellulose or activated nylon), and incubating with a radioactive probe. Exposing the membrane to X-ray film produces darkening at a spot correlating with the position of the DNA or RNA of interest. The darker the spot, the more nucleic acid was present there.

Btwisted: An **EMBOSS** application. btwisted takes a region of a pure DNA sequence and calculates by simple arithmetic the probable overall twist of the sequence and the stacking energy.

C

Cai: An **EMBOSS** application. cai calculates the Codon Adaptation Index. This is a simple, effective measure of synonymous codon usage bias. The index uses a reference set of highly expressed genes from a species to assess the relative merits of each codon, and a score for a gene is calculated from the frequency of use of all codons in that gene. The index assesses the extent to which selection has been effective in moulding the pattern of codon usage. In that respect it is useful for predicting the level of expression of a gene, for assessing the adaptation of viral genes to their hosts, and for making comparisons of codon usage in different organisms. The index may also give an approximate indication of the likely success of heterologous gene expression

Candy: Candy is a small-size project providing access to a set of controlled vocabularies.

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

Carrier: A person who has a recessive mutated gene, together with its normal allele. Carriers do not usually develop disease but can pass the mutated gene on to their children.

CAT assay: An enzyme assay. CAT stands for chloramphenicol acetyl transferase, a bacterial enzyme which inactivates chloramphenicol by acetylating it. CAT assays are often performed to test the function of a promoter. The gene coding for CAT is linked onto a promoter (transcription control region) from another gene, and the construct is "transfected" into cultured cells. The amount of CAT enzyme produced is taken to indicate the transcriptional activity of the promoter (relative to other promoters which must be tested in parallel). It is easier to perform a CAT assay than it is to do a Northern blot, so CAT assays were a common method for testing the effects of sequence changes on promoter function. Largely supplanted by the reporter gene luciferase.

CC line (EMBL): The CC lines are free text comments about the entry, and may be used to convey any sort of information thought to be useful

CC line (SWISS-PROT): The CC lines are free text comments on the entry, and are used to convey any useful information.

CCAAT box: (CAT box, CAAT box, other variants) A sequence found in the 5' flanking region of certain genes which is necessary for efficient expression. A transcription factor (CCAAT-binding protein, CBP) binds to this site.

cDNA clone: cDNA clone: "complementary DNA"; a piece of DNA copied from an mRNA. The term "clone" indicates that this cDNA has been spliced into a plasmid or other vector in order to propagate it. A cDNA clone may contain DNA copies of such typical mRNA regions as coding sequence, 5'-untranslated region, 3' untranslated region or poly(A) tail. No introns will be present, nor any promoter sequences (or other 5' or 3' flanking regions). A "full-length" cDNA clone is one which contains all of the mRNA sequence from nucleotide #1 through to the poly(A) tail.

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.

Chaos: An **EMBOSS** application. Create a chaos game representation plot for a sequence. A box is drawn and an AGCT is drawn at each corner. Starting from the middle, move half way to the corner of the box representing the first base in the sequence and draw a dot. Then for each subsequent base move half way to the corresponding box corner and draw a dot. Finally display the number and percentage values of AGCT bases. The result is an image of a square sprinkled with dots. Areas which are devoid of dots (or heavily covered with dots) indicate short sequence motifs that are unusually infrequent (or frequent). The sequence of such motifs can be deduced by looking to see which quarter of the square the region is in - the letter that this quarter belongs to is the first base of the motif. The quarter is then quartered again and the appropriate base letters are assigned to the corners of the quarter - the eigth part that the region is in gives the second base of the motif. The process continues until you have identified the 1/16th or 1/32nd, etc. of the original square containing the unusual region and you now have the sequence of the motif.

Charge: An **EMBOSS** application. charge reads a protein sequence and writes a file (or plots a graph) of the charges of the amino acids within a window of specified length as the window is moved along the sequence

Checktrans: An **EMBOSS** application. Reports STOP codons and ORF statistics of a protein

chemical base: An essential building block. DNA contains four complementary bases: adenine, which pairs with thymine, and cytosine, which pairs with guanine. In RNA, thymine is replaced by uracil.

Chips: An **EMBOSS** application. Codon usage statistics

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.

Cirdna: An **EMBOSS** application. Draws circular maps of DNA constructs.

Cladogram: A Cladogram is a branching diagram (**tree**) assumed to be an estimate of a phylogeny where the branches are of equal length, thus cladograms show common ancestry, but do not indicate the amount of evolutionary "time" separating taxa.

Clone: The term "clone" can refer either to a bacterium carrying a cloned DNA, or to the cloned DNA itself. If you receive a clone from a collaborator, you should first figure out if they send you DNA or bacteria. If it is DNA, your first job is to introduce it ("transform" it) into bacteria [see "Transformation (with respect to bacteria)"]. Occasionally, someone might send just the "insert", rather than the whole plasmid. Firstly it is necessary to splice that DNA into a convenient vector, and only then can you transform it into bacteria. To "clone" something is to produce copies of it. To clone a piece of DNA, one would insert it into some type of vector (say, a plasmid) and put the resultant construct into a host (usually a bacterium) so that the plasmid and insert replicate with the host. An individual bacterium is isolated and grown and the plasmid containing the "cloned" DNA is re-isolated from the bacteria, at which point there will be many millions of copies of the DNA - essentially an unlimited supply. An investigator wishing to clone some gene or cDNA rarely has that DNA in a purified form, so practically speaking, to "clone" something involves screening a cDNA or genomic library for the desired clone. See also "Probe" for a description of how one might start a cloning project, and "Screening" for how the probe in used.

Cloning: The formation of clones or exact genetic replicas

ClustalW: **ClustalW** produces multiple alignments of protein sequences, such tools are important tools in studying sequences. The basic information they provide is identification of conserved sequence regions. This is very useful in designing experiments to test and modify the function of specific proteins, in predicting the function and structure of proteins, and in identifying new members of protein families. Sequences can be aligned across their entire length (global alignment) or only in certain regions (local alignment). This is true for pairwise and multiple alignments. Global alignments need to use gaps (representing insertions/deletions) while local alignments can avoid them, aligning regions between gaps. ClustalW is a fully automatic program for global multiple alignment of DNA and protein sequences. The alignment is progressive and considers the sequence redundancy. Trees can also be calculated from multiple alignments. The program has some adjustable parameters with reasonable defaults.

ClustalW Format: The first non-blank line must contain the word "CLUSTAL". Sequences are interleaved on separate lines with gaps represented by dashes. Each sequence line starts with the sequence name which is separated from the aligned sequence residues by spaces or tabs. Each set of interleaved sequence segments is separated by one or more blank lines.

CluSTr: The **CluSTr** (Clusters of SWISS-PROT+ **TrEMBL** proteins) database offers an automatic classification of **SWISS-PROT** + **TrEMBL** proteins into groups of related proteins.

CluSTr Database: The **CluSTr** database offers an automatic classification of **SWISS-PROT** + **TrEMBL** proteins into groups of related proteins.

Codata Format: The first line starts with the text ENTRY". The end of a sequence is delineated by "///". The "SEQUENCE" line specifies the beginning of the sequence lines (starting on the next line), and no sequence is assumed to appear in the entry if the "SEQUENCE" line is missing.

Codcmp: An **EMBOSS** application. Codon usage table comparison

Coderet: An EMBOSS application. Extract CDS, mRNA and translations from feature tables.

Coding sequence: The portion of a gene or an mRNA which actually codes for a protein. Introns are not coding sequences; nor are the 5' or 3' untranslated regions (or the flanking regions, for that matter - they are not even transcribed into mRNA). The coding sequence in a cDNA or mature mRNA includes everything from the AUG (or ATG) initiation codon through to the stop codon, inclusive.

Coding strand: An ambiguous term intended to refer to one specific **strand** in a double-stranded gene.

Codon: In an mRNA, a codon is a sequence of three nucleotides which codes for the incorporation of a specific amino acid into the growing protein. The sequence of codons in the mRNA unambiguously defines the primary structure of the final protein. Of course, the codons in the mRNA were also present in the genomic DNA, but the sequence may be interrupted by introns.

Compseq: An **EMBOSS** application. Counts the composition of dimer/trimer/etc words in a sequence Description This takes a specified length of sequence and counts the number of distinct subsequences of that length that there are in the input sequence(s). It can read in the result of a previous compseq analysis and use this to set the expected frequencies of the subsequences. Unless you tell 'compseq' otherwise, it expects each word to be equally likely. The 'Expected' frequency therefore of any dimer is 1/16 - this is simply the inverse of the number of possible dimers (AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT). Similarly, the 'Expected' frequency of any trimer is 1/64, etc. Obviously this is not the case in real sequences - there will be bias in favour of some words. Compseq cannot otherwise guess what the 'Expected' frequency is. You can, however, tell it what the Expected frequencies are by giving compseq the output of the analysis of another set of sequences, produced by a previous compseq run. So you take a set of sequences that are representative of the type of sequence you expect and you run compseq on it to get your expected sequence frequencies. You then take the sequences you wish to investigate, run compseq on them giving compseq the expected frequencies that you have established, above. You tell compseq what the file of expected frequencies is by specifying it with '-infile filename' on the command-line.

Computational Genomics Group: The Computational Genomics Group develops research in the fields of genome sequence annotation, classification of protein function, protein sequence motif discovery, data mining, ontologies for molecular biology, metabolic pathways, deep phylogeny, knowledge representation in molecular biology databases, pattern discovery in sequences and transcription

Cons: An **EMBOSS** application. Description: cons calculates a consensus sequence from a multiple sequence alignment. To obtain the consensus, the sequence weights and a scoring matrix are used to calculate a score at each position in the alignment. The residue (or nucleotide) i in an alignment column, is compared to all other residues (j) in the column. The score for i is the sum over all residues j (not i=j) of the score(ij)*weight(j) . Where score(ij) is taken from a nucleotide or protein scoring matrix (see -datafile qualifier) and the "weight(j)" is the weighting given to the sequence j, which is given in the alignment file. The highest scoring type of residue is then found in the column. If the number of positive matches for this residue is greater than the "plurality value" then this residue is the consensus. The positive matches for a residue i are calculated as being the sum of weights of all the residues that increase the score of residue i (i.e. positive). Where no consensus is found at a position i, an 'n' or an 'x' character is output; (depending on it being a DNA or protein sequence). The "plurality" qualifier allows the user to set a cut-off for the number of positive matches below which there is no consensus. The "identity" qualifier provides the facility of setting the required number of identities at a site for it to give a consensus at that position. Therefore, if this is set to the number of sequences in the alignment only columns of identities contribute to the consensus. The "setcase" qualifier sets the threshold for the positive matches above which the consensus is is upper-case and below which the consensus is in lower-case.

Consensus sequence: A ‘nominal’ sequence inferred from multiple, imperfect examples. Multiple lanes of shotgun sequence can me merged to show a consensus sequence. The optimal sequence of nucleotides recognized by some factor. A DNA binding site for a protein may vary substantially, but one can infer the consensus sequence for the binding site by comparing numerous examples. For example, the (fictitious) transcription factor ZQ1 usually binds to the sequences AAAGTT, AAGGTT or AAGATT. The consensus sequence for that factor is said to be AARRTT (where R is any purine, i.e. A or G). ZQ1 may also be able to weakly bind to ACAGTT (which differs by one base from the consensus).

Cosmid: A type of vector used for cloning 35-45 kb of DNA. These are plasmids carrying a phage l cos site (which allows packaging into l capsids), an origin of replication and an antibiotic resistance gene. A plasmid of 40 kb is very difficult to put into bacteria, but can replicate once there. Cosmids, however, have a cos site, and thus can be packaged into l phage heads (a reaction which can be performed in vitro) to allow efficient introduction into bacteria (you'll have to look up the cos site elsewhere).

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

Cpgplot: An **EMBOSS** application. **Plot CpG rich areas** Description: CpG refers to a C nucleotide immediately followed by a G. The 'p' in 'CpG' refers to the phosphate group linking the two bases. Detection of regions of genomic sequences that are rich in the CpG pattern is important because such regions are resistant to methylation and tend to be associated with genes which are frequently switched on. Regions rich in the CpG pattern are known as CpG islands. It has been estimated that about half of all mammalian genes have a CpG-rich region around their 5' end. It is said that all mammalian house-keeping genes have a CpG island! Non-mammalian vertebrates have some CpG islands that are associated with genes, but the association gets equivocal in the farther taxonomic groups. Finding a CpG island upstream of predicted exons or genes is good contributory evidence. By default, this program defines a CpG island as a region where, over an average of 10 windows, the calculated % composition is over 50% and the calculated Obs/Exp (i.e. Observed/Expected) ratio is over 0.6 and the conditions hold for a minimum of 200 bases. These conditions can be modified by setting the values of the appropriate parameters. The Observed number of CpG patterns in a window is simply the count of the number of times a 'C' is found followed immediately by a 'G'. The Expected number of CpG patterns is calculated for each window as the number of CpG dinucleotides you would expect to see in that window based on the frequency of C's and G's in that window. Thus, the Expected frequency of CpG's in a window is calculated as the number of 'C's in the window multiplied by the number of 'G's in the window, divided by the window length. Expected = (number of C's * number of G's) / window length This program reads in one or more sequences and calculates the Obs/Exp ratio, the percentage CpG over a window which is moved along the sequence. These calculated values can be plotted, together with the regions which match this program's definition of a CpG island.

CpGReport: **CpGReport** will produce an **EMBL** formated report with a Feature Table that contains a Key for each island found along with location/qualifiers that depict the position of the island, its size, the total sum of C+G's in the island, the %CG and the observed/expected value max for it. At the bottom of each report the total number of islands found is printed or a 'No islands found' text.

Cusp: An **EMBOSS** application to Create a codon usage table . It reads one or more coding sequences (CDS sequence only) and calculates a codon frequency table. The output file can be used as a codon usage table in other applications.

Cutseq: An EMBOSS application. Removes a specified section from a sequence Description: This simple editing program allows you to cut out a region from your sequence by specifying the begin and end positions of the sequence to remove. It removes the sequence from the specified start to the end positions (inclusive) and returns the rest of the sequence in the output file.

Cystic fibrosis: An inherited disease in which a thick mucus clogs the lungs and blocks the ducts of the pancreas

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

Cytosine: A pyrimidine base found in DNA and RNA

D

DNA sequencing: The technique in which the specific sequence of bases forming a particular DNA region is deciphered.

DNA Strands: 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 the reverse and complement sequence to your input sequence will be searched against the database entry. A gene is composed of DNA, which is located in the nucleus. It is a double helix consisting of 2 strands. Many tools will have options where you can search against the top, bottom or both strands of DNA. This bonding between strands is known as a base pair. A base pair is simply a pair of bases which form bonds with each other. There are only two base pairs found in DNA: adenine(A) and thymine(T) form one base pair, and cytosine(C) and guanine(G) form the other. This piece of hypothetical DNA could produce 2 RNA sequences based upon which strand is used as the template. They are similar to the reverse strand of DNA except Uradine(U) replaces thymine(T), found in DNA.

DNase: Deoxyribonuclease, a class of enzymes which digest DNA. The most common is DNase I, an endonuclease which digests both single and double-stranded DNA.

Domain: 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.

Dominant allele: A gene that is expressed, regardless of whether its counterpart allele on the other chromosome is dominant or recessive. Autosomal dominant disorders are produced by a single mutated dominant allele, even though its corresponding allele is normal.

Dot blot: A technique for measuring the amount of one specific DNA or RNA in a complex mixture. The samples are spotted onto a hybridization membrane (such as nitrocellulose or activated nylon, etc.), fixed and hybridized with a radioactive probe. The extent of labeling (as determined by autoradiography and densitometry) is proportional to the concentration of the target molecule in the sample. Standards provide a means of calibrating the results.

Dotmatcher: An **EMBOSS** application. A dotplot is a graphical representation of the regions of similarity between two sequences. The two sequences are placed on the axes of a rectangular image and (subject to threshold conditions) wherever there is a similarity between the sequences a dot is placed on the image. Where the two sequences have substantial regions of similarity, many dots align to form diagonal lines. It is therefore possible to see at a glance where there are local regions of similarity as these will have long diagonal lines. It is also easy to see other features such as repeats (which form parallel diagonal lines), and insertions or deletions (which form breaks or discontinuities in the diagonal lines). dotmatcher uses a threshold to define whether a match is plotted (calculated from the substitution matrix). A window of specified length is moved up all possible diagonals and a score is calculated within each window for each position along the diagonals. The score is the sum of the comparisons of the two sequences using the given similarity matrix along the window. If the score is above the threshold, then a line is plotted on the image over the position of the window.

Dotpath: An **EMBOSS** application. A dotplot is a graphical representation of the regions of similarity between two sequences. The two sequences are placed on the axes of a rectangular image and wherever there is a similarity between the sequences a dot is placed on the image. Where the two sequences have substantial regions of similarity, many dots align to form diagonal lines. It is therefore possible to see at a glance where there are local regions of similarity. dotpath is very similar to the program dottup which looks for places where words (tuples) of a specified length have an exact match in both sequences and draws a diagonal line over the position of these words. Using a longer word size thus displays less random noise, runs extremely quickly, but is less sensitive. dotpath finds all matches of size -wordsize or greater between two sequences. It then reduces the matches found to the minimal set of long matches that do not overlap. This is a way of finding the (nearly) optimal path aligning two sequences. It is not the true optimal path as produced by the algorithms used in water or needle, but for very closely related sequences it will produce the same result and will work well with very long sequences. If you wish to compare the path found by dotpath to the set of all matches found then the qualifier -overlaps will show all matches in red except for the matches in the minimal path which are shown in black, as normal.

Dottup: An **EMBOSS** application. A dotplot is a graphical representation of the regions of similarity between two sequences. The two sequences are placed on the axes of a rectangular image and (in the simplest forms of dotplot) wherever there is a similarity between the sequences a dot is placed on the image. Where the two sequences have substantial regions of similarity, many dots align to form diagonal lines. It is therefore possible to see at a glance where there are local regions of similarity as these will have long diagonal lines. It is also easy to see other features such as repeats (which form parallel diagonal lines), and insertions or deletions (which form breaks or discontinuities in the diagonal lines). dottup looks for places where words (tuples) of a specified length have an exact match in both sequences and draws a diagonal line over the position of these words. This is a fast, but not especially sensitive way of creating dotplots. It is an acceptable method for displaying regions of substantial similarity between two sequences. Using a longer word (tuple) size displays less random noise, runs extremely quickly, but is less sensitive. Shorter word sizes are more sensitive to shorter or fragmentary regions of similarity, but also display more random points of similarity (noise) and runs slower.

Dreg: An **EMBOSS** application. This searches for matches of a regular expression to a nucleic acid sequence. A regular expression is a way of specifying an ambiguous pattern to search for. Regular expressions are commonly used in some computer programming languages and may be more familiar to some users than to others. The following is a short guide to regular expressions in EMBOSS: ^ use this at the start of a pattern to insist that the pattern can only match at the start of a sequence. (eg. '^AUG' matches a start codon at the start of the sequence) $ use this at the end of a pattern to insist that the pattern can only match at the end of a sequence (eg. 'A+$' matches a poly-A sequence at the end of the sequence) () groups a pattern. This is commonly used with '|' (eg. '(AUG)|(ATG)' matches either the DNA or RNA form of the initiation codon) | This is the OR operator to enable a match to be made to either one pattern OR another. There is no AND operator in this version of regular expressions. The following quantifier characters specify the number of time that the character before (in this case 'x') matches: x? matches 0 or 1 times (ie, '' or 'x') x* matches 0 or more times (ie, '' or 'x' or 'xx' or 'xxx', etc) x+ matches 1 or more times (ie, 'x' or 'xx' or 'xxx', etc) Quantifiers can follow any of the following types of character specification: x any character (ie 'A') x the character after the backslash is used instead of its normal regular expression meaning. This is commonly used to turn off the special meaning of the characters '^$()|?*+[]-.'. It may be especially useful when searching for gap characters in a sequence (eg '.' matches only a dot character '.') [xy] match one of the characters 'x' or 'y'. You may have one or more characters in this set. [x-z] match any one of the set of characters starting with 'x' and ending in 'y' in ASCII order (eg '[A-G]' matches any one of: 'A', 'B', 'C', 'D', 'E', 'F', 'G') [^x-z] matches anything except any one of the group of characters in ASCII order (eg '[^A-G]' matches anything EXCEPT any one of: 'A', 'B', 'C', 'D', 'E', 'F', 'G') . the dot character matches any other character (eg: 'A.G' matches 'AAG', 'AaG', 'AZG', 'A-G' 'A G', etc.) Combining some of these features gives the example: '([AGC]+GGG)|(TTTGGG)' which matches one or more of any one of 'A' or 'G' or 'C' followed by three 'G's or it matches just 'TTTGGG'. Regular expressions are case-sensitive. The pattern 'AAAA' will not match the sequence 'aaaa'.

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

DSSP: Dictionary Secondary Structure of Protein. Definition of secondary structure of proteins given a set of 3D coordinates.

E

E-BioSci: E-BioSci aims

F

Familial cancer: Cancer, or a predisposition toward cancer, that runs in families.

FASTA: **FASTA** (pronounced FAST-Aye) stands for FAST-All, reflecting the fact that it can be used for a fast protein comparison or a fast nucleotide comparison. This program achieves a high level of sensitivity for similarity searching at high speed. This is achieved by performing optimised searches for local alignments using a substitution matrix. The high speed of this program is achieved by using the observed pattern of word hits to identify potential matches before attempting the more time consuming optimised search. The trade-off between speed and sensitivity is controlled by the ktup parameter, which specifies the size of the word. Increasing the ktup decreases the number of background hits. Not every word hit is investigated but instead initially looks for segment's containing several nearby hits.

fasta sequence format: This **format** contains a one line header followed by lines of sequence data. Sequences in fasta formatted files are preceded by a line starting with a" >" symbol. The first word on this line is the name of the sequence. The rest of the line is a description of the sequence. The remaining lines contain the sequence itself. Blank lines in a FASTA file are ignored, and so are spaces or other gap symbols (dashes, underscores, periods) in a sequence. Fasta files containing multiple sequences are just the same, with one sequence listed right after another. This format is accepted for many multiple sequence alignment programs.

```
>FOSB_MOUSE Protein fosB. 338 bp
MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGS
PPTAAASQECAGLGEMPGSFVPTVTAITTSQDLQWL
VQPTLISSMAQSQGQPLASQPPAVDPYDMPGTSYS
TPGLSAYSTGGASGSGGPSTSTTTSGPVSARPARA
RPRRPREETLTPEEEEKRRVRRERNKLAAAKCRNR
RRELTDRLQAETDQLEEEKAELESEIAELQKEKERLEF
VLVAHKPGCKIPYEEGPGPGPLAEVRDLPGSTSAKE
DGFGWLLPPPPPPPLPFQSSRDAPPNLTASLFTHSE
VQVLGDPFPVVSPSYTSSFVLTCPEVSAFAGAQRTS
GSEQPSDPLNSPSLLAL
```

fasta3: A program that scans a protein or DNA sequence library for similar sequences

fastf3: A program that compares mixed peptides to a protein databank

fasts3: A program that compares linked peptides to a protein databank

fastx3: A program that compares a DNA sequence to a protein sequence database, comparing the translated DNA sequence in forward and reverse frames.

fasty3: A program that compares a DNA sequence to a protein sequence database, comparing the translated DNA sequence in forward and reverse frames.

Feature Header line (EMBL): The FH (Feature Header) lines are present only to improve readability of an entry when it is printed or displayed on a terminal screen. The lines contain no data and may be ignored by computer programs.

Feature Table line (EMBL): The FT (Feature Table) lines provide a mechanism for the annotation of the sequence data. Regions or sites in the sequence which are of interest are listed in the table. A complete and definitive description of the feature table is given here.

Feature Table line (SWISS-PROT): The FT (Feature Table) lines provide a precise but simple means for the annotation of the sequence data. The table describes regions or sites of interest in the sequence. In general the feature table lists post-translational modifications, binding sites, enzyme active sites, local secondary structure or other characteristics reported in the cited references.

Filter: The filter option, if set to true, when running a blast, will allow you to mask out various segments of the query sequence for regions which are non-specific for sequence similarity searches. Filtering can eliminate statistically significant but biologically uninteresting reports from the output, for example hits against common acidic-, basic- or proline-rich regions, leaving the more biologically interesting regions of the query sequence available for specific matching against database sequences. Filtering is only applied to the query sequence, not to database sequences. The program used for this, with nucleotide query sequences is known as DUST written by Tatusov, R. L., and Lipman, D.J. The SEG program is used for filtering low complexity regions in amino acid sequences from your protein query sequence and was written by Wootton, J.C., and Federhen, S.

Findkm: An **EMBOSS** application. Takes a file of enzymatic data and plots Michaelis Menten and Hanes Woolf plots of the data. From these it calculates the Michaelis Menten constant (Km) and the maximum velocity (Vmax) of the reaction.

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.

FingerPRINTScan: A tool that scans against the fingerprints in the PRINTS database. These fingerprints are groups of motifs that together are more potent than single motifs by making use of the biological context inherent in a multiple motif method.

Flat Query-anchored with identities alignment: The 'flat' display shows inserts as deletions on the query. Identities are displayed as dots. Mismatches displayed as single letter nucleotide abbreviations (c,t,a or g). Gaps are introduced with a "-" symbol.

Flat Query-anchored without identities alignment: The 'flat' display shows inserts as deletions on the query. Identities are displayed as as single letter nucleotide abbreviations (c,t,a or g). Mismatches displayed as single letter nucleotide abbreviations (c,t,a or g). Gaps are introduced with a "-" symbol.

Footprinting: A technique by which one identifies a protein binding site on cellular DNA. The presence of a bound protein prevents DNase from "nicking" that region, which can be detected by an appropriately designed gel.

Freak: An **EMBOSS** application. freak takes one or more sequences as input and a set of bases or residues to search for. It then calculates the frequency of these bases/residues in a window as it moves along the sequence. The frequency is output to a data file or (optionally) plotted. The default set of bases is 'cg' which will calculate the frequency of 'G' + 'C' bases within the default moving window of 30 bases.

FSSP: Fold classification based on Structure-Structure alignment of Proteins.

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

Fuzznuc: An **EMBOSS** application. fuzznuc uses PROSITE style patterns to search nucleotide sequences. Patterns are specifications of a (typically short) length of sequence to be found. They can specify a search for an exact sequence or they can allow various ambiguities, matches to variable lengths of sequence and repeated subsections of the sequence. fuzznuc intelligently selects the optimum searching algorithm to use, depending on the complexity of the search pattern specified.

Fuzzpro: An **EMBOSS** application. fuzzpro uses PROSITE style patterns to search protein sequences. Patterns are specifications of a (typically short) length of sequence to be found. They can specify a search for an exact sequence or they can allow various ambiguities, matches to variable lengths of sequence and repeated subsections of the sequence. fuzzpro intelligently selects the optimum searching algorithm to use, depending on the complexity of the search pattern specified.

Fuzztran: An **EMBOSS** application. fuzztran uses PROSITE style protein patterns to search nucleic acid sequences translated in the specified frame(s). Patterns are specifications of a (typically short) length of sequence to be found. They can specify a search for an exact sequence or they can allow various ambiguities, matches to variable lengths of sequence and repeated subsections of the sequence. fuzztran intelligently selects the optimum searching algorithm to use, depending on the complexity of the search pattern specified.

G

Gap: A **gap** is a maximal consecutive run of spaces in a single string of a given alignment. It corresponds to an atomic insertion or deletion of a substring. Causes of gaps: A single mutation can create a gap (very common). Unequal crossover in meiosis can lead to insertion or deletion of strings of bases. DNA slippage in the replication procedure can result in the repetition of a string. Retrovirus insertions. Translocations of DNA between chromosomes

gap extension penalty: The gap extension penalty is added to the standard gap open penalty for each base or residue in the gap. This is how long gaps are penalised. If you don't like long gaps, just increase the extension gap penalty. Usually you will expect a few long gaps rather than many short gaps, so the gap extension penalty should be lower than the gap penalty. An exception is where one or both sequences are single reads with possible sequencing errors in which case you would expect many single base gaps. You can get this result by setting the gap open penalty to zero (or very low) and using the gap extension penalty to control gap scoring.

gap open: The gap open penalty is the score taken away for the initiation of the gap in sequence or in structure. To make the match more significant you can try to make the gap penalty larger. It will decrease the number of gaps and if you have good alignment without many gaps, its Z-score will be higher.

Gap Penalties: Introduction of gaps into sequence alignments allows the alignment to be extended into regions where one sequence may have lost or gained sequence characters not found in the other. If the gap penalty is too low, then a high sequence alignment score is achievable even between unrelated or random sequences. A penalty is subtracted for each gap introduced into an alignment because the gap increases uncertainty into an alignment. If gaps are introduced without a penalty than they can be introduced at random and eventually all characters will be aligned in even random sequences.The gap penalty is used to help decide whether on not to accept a gap or insertion in an alignment when it is possible to achieve a good alignment residue-to-residue at some other neighbouring point in the sequence. One cannot let gaps/insertion occur without penalty, because an unreasonable 'gappy' alignment would result. Biologically, it should in general be easier for a protein to accept a different residue in a position, rather than having parts of the sequence chopped away or inserted. Gaps/insertions should therefore be more rare than point mutations (substitutions). Thus, 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. In general, the lower the gapping penalties, the more gaps and more identities are detected but this should be considered in relation to biological significance.

gap penalties – adjusting: Fasta, Blast, Blitz and Clustalw use slightly different terms to refer to gap initiation and gap extension penalties. In general, gapopen and opengap are the former while gapext and extendgap the latter. Some of the later improvements to these programs include the possibility to penalise gaps separately on the database sequences and then query sequences separately. Such is the case of blitz. In clustalw, a gap penalty exists which penalises separately the length of a gap, closing a gap and the introduction of a pairwise gap in both sequences. Gap penalty values are designed to reduce the score when an alignment has been broken by an insertion in one of the sequences. The value should be small enough to allow a previously accumulated alignment to continue with an insertion in one of the sequences but should not be so large that this previous alignment score is removed completely. You could tweak gap open and gap extension penalties (which combined produce the overall gap penalty) in 2 ways: Keep the score similar regardlass of gap length. Allow a constant overall gap penalty regardless of gap length, in other words have a zero gap extension penalty and just penalise when you open a gap. These types of penalty schemes assume that sequences are just as likely to change by large as by small insertions and deletions. This will penalise a large gap by the same extent as a small gap. The score becomes larger as a linear function of gap length: Have a larger gap opening penalty followed by a gap extension penalty that is smaller than the gap open penalty. This will penalise several small gaps by the same extent as 1 large gap.

Garnier: An **EMBOSS** application. This is an implementation of the original Garnier Osguthorpe Robson algorithm (GOR I) for predicting protein secondary structure. Secondary structure prediction is notoriously difficult to do accurately. The GOR I algorithm is one of the first semi-successful methods. The Garnier method is not regarded as the most accurate prediction, but is simple to calculate on most workstations. The accuracy of any secondary structure prediction program is not much better than 70% to 80% at best. This is an early algorithm and will probably not predict with much better than about 65% accuracy.

GCG: On June 1, 2001, the **Genetics Computer Group (GCG)**, Oxford Molecular, MSI, and Synopsys joined together to become a single company. Accelrys is the new leader in simulation and informatics software for the pharmaceutical, biotechnology, and chemicals process industries.

Founded in 1982 as a service of the Department of Genetics at the University of Wisconsin, GCG became a private company in 1990 and was acquired by Oxford Molecular Group in 1997. The company was one of the pioneers of bioinformatics and its Wisconsin Package sequence analysis tools are widely used and well regarded throughout the pharmaceutical and biotechnology industries and in academia. To support enterprise bioinformatics efforts, GCG developed SeqStore, its Oracle-based data management system. Desktop solutions are delivered to bench scientists through products such as MacVector and OMIGA. Following the September 2000 acquisition by Pharmacopeia, GCG and Oxford Molecular were combined with MSI and Synopsys Scientific Systems, with the goal of creating a single provider of simulation and informatics tools capable of building an integrated technology platform for research and development. That sole provider is Accelrys. Accelrys customers will benefit from the overlap of GCG bioinformatics expertise with MSI protein modelling capabilities -- a unique combination in biocomputing and structural proteomics.

Geecee: An **EMBOSS** application. This calculates the fraction of G+C bases of the input nucleic acid sequence(s). It reads in nucleic acid sequences, sums the number of 'G' and 'C' bases and writes out the result as the fraction (in the interval 0.0 to 1.0) of the length of the whole sequence.

Gel electrophoresis: A method to analyse the size of DNA (or RNA) fragments. In the presence of an electric field, larger fragments of DNA move through a gel slower than smaller ones. If a sample contains fragments at four different discrete sizes, those four size classes will, when subjected to electrophoresis, all migrate in groups, producing four migrating "bands". Usually, these are visualised by soaking the gel in a dye (ethidium bromide) which makes the DNA fluoresce under UV light.

Gel shift assay: (aka gel mobility shift assay (GMSA), band shift assay (BSA), electrophoretic mobility shift assay (EMSA)) A method by which one can determine whether a particular protein preparation contains factors which bind to a particular DNA fragment. When a radiolabelled DNA fragment is run on a gel, it shows a characteristic mobility. If it is first incubated with a cellular extract of proteins (or with purified protein), any protein-DNA complexes will migrate slower than the naked DNA - a shifted band.

Genbank Format: GenBank is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences. Although there is daily exchange of information with the **EMBL** Nucleotide Sequence Database, it has it's own sequence format shown below. Each **GenBank** entry includes a concise description of the sequence, the scientific name and taxonomy of the source organism, and a table of features that identifies coding regions and other sites of biological significance, such as transcription units, sites of mutations or modifications, and repeats. Protein translations for coding regions are included in the feature table. Bibliographic references are included along with a link to the Medline unique identifier for all published sequences. Each sequence entry is composed of lines. Different types of lines, each with their own format, are used to record the various data that make up the entry.

Gene: A unit of DNA which performs one function. Usually, this is equated with the production of one RNA or one protein. A gene contains coding regions, introns, untranslated regions and control regions.

Gene deletion: The total loss or absence of a gene

Gene markers: Landmarks for a target gene, either detectable traits that are inherited along with the gene, or distinctive segments of DNA

Gene Name line (SWISS-PROT): The GN (Gene Name) line contains the name(s) of the gene(s) that code for the stored protein sequence.

Gene therapy: The use of genetic material for therapeutic purposes. The therapeutic gene is typically delivered using recombinant virus or liposome based delivery systems.

Genemark: Genemark. The problem of predicting gene locations in newly sequenced DNA is well known but still far from being successfully resolved. A novel approach to the problem based on the frame dependent (non-homogeneous) Markov chain models of protein-coding regions was previously suggested. This approach is, apparently, one of the most powerful "search by content" methods. The initial idea of the method combines the specific Markov models of coding and non-coding region together with Bayes' decision making function and allows easy generalisation for employing of higher order Markov chain models. Another generalisation allows the analysis of both DNA strands simultaneously. Currently known gene searching methods perform the analysis of the two DNA strands in turn, one after another. In doing this all the known methods fail in the sense that they generate false (artifactual) prediction signals for the given strand when the real coding region is located on the complementary DNA strand. This common drawback is avoided by employing the Bayesian algorithm which uses an additional non-homogeneous Markov chain model of the "shadow" of the coding region -- the sequence which is complementary to the protein-coding sequence.

GeneQuiz: **GeneQuiz** is an integrated system for large-scale biological sequence analysis, that goes from a protein sequence to a biochemical function, using a variety of search and analysis methods and up-to-date protein and DNA databases. Applying an "expert system" module to the results of the different methods, GeneQuiz creates a compact summary of findings. It focuses on deriving a predicted protein function, based on the available evidence, including the evaluation of the similarity to the closest homologue in the database (identical, clear, tentative, or marginal). The analysis yields everything that can possibly be extracted from the current databases, including three-dimensional models by homology, when the structure can be reliably calculated.

Genetic code: The mapping of all possible codons into the 20 amino acids including the start and stop codons.

Genetic Code Viewer: Genetic Code Viewer is a simple tool for showing different versions of genetic code used by various taxonomic groups.

Genetic marker: Any gene that can be readily recognised 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 total DNA contained in each cell of an organism. Mammalian genomic DNA (including that of humans) contains 6x109 base pairs of DNA per diploid cell. There are somewhere in the order of a hundred thousand genes, including coding regions, 5' and 3' untranslated regions, introns, 5' and 3' flanking DNA. Also present in the genome are structural segments such as telomeric and centromeric DNAs and replication origins, and intergenic DNA.

Genome & Proteome Fasta: **This tool** provides sequence similarity and homology searching against complete proteome or genome databases using the Fasta3 programs to sequences similar to your query.

Genomes Database: A **database** of complete genome sequences

Genomic blot: A type of Southern blot specifically used to analyse a mixture of DNA fragments derived from total genomic DNA. Because genomic DNA is very complicated, when it has been digested with restriction enzymes, it produces a complex set of fragments ranging from tens of bp to tens of thousands of bp. However, any specific gene will be reproducibly found on only one or a few specific fragments. A million identical cells will produce a million identical restriction fragments for any given gene, so probing a genomic Southern with a gene-specific probe will produce a pattern of perhaps one or just a few bands.

H

Haploid: A cell or organism containing only one set of chromsomes without the homologous pairs.

Helix-loop-helix: A protein structural motif characteristic of certain DNA-binding proteins

Helixturnhelix: An **EMBOSS** application. helixturnhelix uses the method of Dodd and Egan and finds helix-turn-helix nucleic acid binding motifs in proteins. The helix-turn-helix motif was originally identified as the DNA-binding domain of phage repressors. One alpha-helix lies in the wide groove of DNA; the other lies at an angle across DNA.

Hereditary mutation: A gene change in the body's reproductive cells (egg or sperm) that becomes incorporated in the DNA of every cell in the body; also called germline mutation.

Heterodimer: Protein composed of 2 different chains or subunits.

HGVbase: HGVbase (Human Genome Variation database) consists of all known sequence variations in the human genome.

Hidden Markov Model: 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. The perturbed values can be continuous and are the "outputs" of the HMM. A Hidden Markov Model is equivalently a coupled mixture model where the joint distribution over states is a Markov chain. Hidden Markov models are valuable in bioinformatics because they allow a search or alignment algorithm to be trained using unaligned or unweighted input sequences; and because they allow position-dependent scoring parameters such as gap penalties, thus more accurately modelling the consequences of evolutionary events on sequence families.

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

HMM: Hidden Markov model. 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. The perturbed values can be continuous and are the "outputs" of the HMM. A Hidden Markov Model is equivalently a coupled mixture model where the joint distribution over states is a Markov chain. Hidden Markov models are valuable in bioinformatics because they allow a search or alignment algorithm to be trained using unaligned or unweighted input sequences; and because they allow position-dependent scoring parameters such as gap penalties, thus more accurately modeling the consequences of evolutionary events on sequence families.

Hmmpfam: A tool that scans the hidden markov models (HMMs) that are present in the protein domain databases Pfam, TIGRFAMMs and SMART.

Hmoment: An **EMBOSS** application. hmoment plots or writes out the hydrophobic moment. Hydrophic moment is the hydrophobicity of a peptide measured for a specified angle of rotation per residue. Periodicities in the polar/apolar character of the amino acid sequence of a protein can be examined by assigning to each residue a numerical hydrophobicity and searching for periodicity in the resulting one-dimensional function. The strength of each periodic component is the quantity that has been termed the hydrophobic moment. When proteins of known three-dimensional structure are examined, it is found that sequences that form alpha helices tend to have, on average, a strong periodicity in the hydrophobicity of af 3.6 residues, the period of the alpha helix. The angle of rotation per residue in alpha helices is 100 degrees. Similarly, many sequences that form strands of beta sheets tend to have a periodicity in their hydrophobicity of about 2.3 residues, the period typical of beta structure. The angle of rotation per residue in beta sheets is 160 degrees. This means that many protein sequences tend to form the periodic structure that maximises their amphiphilicity. The hydrophobic moment is measured within a moving window using the method of Eisenberg et al. The default angle of 100 degrees is used for the alpha-helix results and the default of 160 degrees is used for the beta-sheet results. These angles can be changed if required using the appropriate options. hmoment can plot two graphs when the option '-double' is given, one for the alpha helix moment and one for the beta sheet moment. Otherwise it just plots the alpha helix moment.

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.

Homology: (strict) Two or more biological species, systems or molecules that share a common evolutionary ancestor. (general) Two or more gene or protein sequences that share a significant degree of similarity, typically measured by the amount of identity (in the case of DNA), or conservative replacements (in the case of protein), that they register along their lengths. Sequence "homology" searches are typically performed with a query DNA or protein sequence to identify known genes or gene products that share significant similarity and hence might inform on the ancestry, heritage and possible function of the query gene.

Homology and Similarity Searching: 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. Statistical methods are less successful but more general and are based on the derivation of structural preference values for single residues, pairs of residues, short oligopeptides or short sequence patterns. The transfer of structure/function information to a potentially homologous protein is straightforward when the sequence similarity is high and extended in length, but the assessment of the structural significance of sequence similarity can be difficult when sequence similarity is weak or restricted to a short region.

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. Statistical methods are less successful but more general and are based on the derivation of structural preference values for single residues, pairs of residues, short oligopeptides or short sequence patterns. The transfer of structure/function information to a potentially homologous protein is straightforward when the sequence similarity is high and extended in length, but the assessment of the structural significance of sequence similarity can be difficult when sequence similarity is weak or restricted to a short region.

Housekeeping genes: Genes that are always expressed (ie. they are said to be constitutively expressed) due to their constant requirement by the cell.

HPI: Human Proteomics Initiative, by the Swiss Institute of Bioinformatics (SIB) and the European Bioinformatics Institute (EBI), to annotate all known human sequences according to the quality standards of SWISS-PROT

Human Proteomics Initiative: Human Proteomics Initiative, by the Swiss Institute of Bioinformatics (SIB) and the European Bioinformatics Institute (EBI), to annotate all known human sequences according to the quality standards of SWISS-PROT.

Huntingtons disease: An adult-onset disease characterised by progressive mental and physical deterioration; it is caused by an inherited dominant gene mutation

Hybridization: The reaction by which the pairing of complementary strands of nucleic acid occurs. DNA is usually double-stranded, and when the strands are separated they will re-hybridize under the appropriate conditions. Hybrids can form between DNA-DNA, DNA-RNA or RNA-RNA. They can form between a short strand and a long strand containing a region complementary to the short one. Imperfect hybrids can also form, but the more imperfect they are, the less stable they will be (and the less likely to form). To "anneal" two strands is the same as to "hybridize" them.

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.

I

ID (IDentification) line (EMBL): The ID (IDentification line) line is always the first line of an entry. The general form of the ID line is: Term ID, entryname, dataclass, molecule, division, sequence length (Base Pairs).

ID (IDentification) line (SWISS-PROT): The ID (IDentification) line is always the first line of an entry. The general form of the ID line is: Term, ID ENTRY_NAME, DATA_CLASS, MOLECULE_TYPE, SEQUENCE_LENGTH.

Iep: An **EMBOSS** application. This calculates the isoelectric point of a protein from its amino acid composition assuming that no electrostatic interactions change the propensity for ionisation. Adjusting the pH of an aqueous protein solution to the point where the numbers of positive and negative charges on the protein are equal brings the protein to its isoelectric point. This is often the point of lowest solubility, presumably because it is the point at which there are fewest intermolecular repulsions, so that the molecules tend to form aggregates. The application can make a plot of the ionisation curve with respect to pH and can write an output file of the data.

IMGT: ImMunoGeneTics database, compromising **IMGT/LIGM-DB** database of immunoglobulins and T-cell receptors, **IMGT/HLA** database of the human MHC complex and **IMGT/MHC** covering the MHC complex of non-human species.

IMGT/HLA Database: The **International ImMunoGeneTics database**. Provides a specialist database for sequences of the human major histocompatibility complex (HLA).

IMGT/LIGM Database: The **International ImMunoGeneTics database** at the Laboratoire d'ImmunoGénétique Moléculaire, a comprehensive database of IG and TR from human and other vertebrates, with translation for fully annotated sequences.

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

Imprinting: A biochemical phenomenon that determines, for certain genes, which one of the pair of alleles, the mother's or the father's, will be active in that individual

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)

Industry Programme: The **Industry Programme** is a forum through which the **EBI** can provide training and research of benefit to the European pharmaceutical, biotechnology, consumer-goods, chemical and agricultural industries. The Industry Programme enables industry to quickly adapt to, and maximise the benefit from, innovations in the fast-growing field of bioinformatics. Offering training, research, and the development and adaptation of bioinformatics resources that are particularly relevant to industry, the programme content remains at the cutting edge of bioinformatics through joint development by the **EBI** and its industry partners.

Infoalign: An **EMBOSS** application. infoalign is small utility to list some simple properties of sequences in an alignment. It will write a table containing one line per sequence. The information is written out in columns separated by space or TAB characters. The columns of data are: the sequences' USA, name, two measures of length, counts of gaps, and numbers of identical, similar and different residues or bases in this sequence when compared to a reference sequence, together with a simple statistic of the % change between the reference sequence and this sequence. The reference sequence can be either the calculated consensus sequence (the default) or it can be one of the set of aligned sequences, specified by either the ordinal number of that sequence in the input file, or by its name. Any combination of these types of information can be easily selected or unselected. By default, the output file starts each line with the USA of the sequence being described, so the output file is a list file that can be manually edited and read in by any other **EMBOSS** program that can read in one or more sequence to be analysed.

Infoseq: An **EMBOSS** application. This is a small utility to list the sequences' USA, name, accession number, type (nucleic or protein), length, percentage C+G, and/or description. Any combination of these types of information can be easily selected or unselected. By default, the output file starts each line with the USA of the sequence being described, so the output file is a list file that can be manually edited and read in by any other **EMBOSS** program that can read in one or more sequence to be analysed.

Insert: In a complete plasmid clone, there are two types of DNA - the "vector" sequences and the "insert". The vector sequences are those regions necessary for propagation, antibiotic resistance, and all those mundane functions necessary for useful cloning. In contrast, however, the insert is the piece of DNA in which you are really interested.

Intelligenetics alignment format: Intelligenetics alignment format uses `|' to show identities and `:' to show conservative replacements and places these indicators between the two aligned sequences.

IntEnz: The Integrated relational Enzyme database (**IntEnz**) is supported by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB) and will contain enzyme data approved by the Nomenclature Committee. The goal is to create a single relational enzyme database, using the resources already available from the Trinity College in Dublin, ENZYME database (SIB) and BRENDA database (University of Cologne). IntEnz, implemented and supported by the EBI, will be the master copy of the Enzyme database.

InterPro: An integrated documentation resource for protein families, domains and sites. **InterPro** combines a number of databases that use different methodologies and a varying degree of biological information on well-characterised proteins to derive protein signatures

InterProScan: InterProScan is a tool that combines different protein signature recognition methods native to the **InterPro** member databases into one resource with look up of corresponding **InterPro** and GO annotation

Intron: Introns are portions of genomic DNA which ARE transcribed (and thus present in the primary transcript) but which are later spliced out. They are not present in the mature mRNA. Note that although the 3' flanking region is often transcribed, it is removed by endonucleolytic cleavage and not by splicing. It is not an intron.

IPI: A non-redundant human proteome set constructed from SWISS-PROT, TrEMBL, **Ensembl** and RefSeq.

Isochore: Isochore plots GC content over a sequence. It is intended for large sequences such as complete chromosomes or large genomic contigs, although interesting results can also be obtained from shorter sequences

isochore-emboss: An **EMBOSS** application. The nuclear genomes of vertebrates are mosaics of isochores, very long stretches (>300kb) of DNA that are homogeneous in base composition and are compositionally correlated with the coding sequences that they embed. Isochores can be partitioned in a small number of families that cover a range of GC levels (GC is the molar ratio of guanine+cytosine in DNA), which is narrow in cold-blooded vertebrates, but broad in warm-blooded vertebrates. This application plots GC content over a sequence. It is intended for large sequences such as complete chromosomes or large genomic contigs, although interesting results can also be obtained from shorter sequences.

J

Jembossctl: An **EMBOSS** application. jembossctl should not be run by typical users. It is a slave program for the Jemboss server. If this means nothing to you, then you do not need to know anything more, just ignore this program. If you are setting up a Jemboss server (This is not the Jemboss interface that anyone can download, it is the software that listens to the Jemboss interfaces and runs the **EMBOSS** programs for you), then this program should be 'chmod 4755' to get root. Details are to be found on the Jemboss web pages.

K

Karyotype: The constitution (typically number and size) of chromosomes in a cell or individual.

KeyWord line (EMBL): The KW (KeyWord) lines provide information which can be used to generate cross-reference indexes of the sequence entries based on functional, structural, or other categories deemed important. The keywords chosen for each entry serve as a subject reference for the sequence, and will be expanded as work with the database continues. Often several KW lines are necessary for a single entry.

KeyWord line (SWISS-PROT): The KW (KeyWord) lines provide information that can be used to generate indexes of the sequence entries based on functional, structural, or other categories.

Kimura Correction of distances: This options, when using ClustalW allows you to set on distances correction (correction for multiple substitutions). This is because, as sequences diverge, more than one substitution will happen at many sites. However, you only see one difference when you look at the present day sequences. Therefore, this option has the effect of stretching branch lengths in trees (especially long branches). The corrections used here (for DNA or proteins) are both due to Motoo Kimura.

Kinase: A kinase is in general an enzyme that catalyses the transfer of a phosphate group from ATP to something else. In molecular biology, it has acquired the more specific verbal usage for the transfer onto DNA of a radiolabelled phosphate group. This would be done in order to use the resultant "hot" DNA as a probe.

Knock-out experiment: A technique for deleting, mutating or otherwise inactivating a gene in a mouse. This laborious method involves transfecting a crippled gene into cultured embryonic stem cells, searching through the thousands of resulting clones for one in which the crippled gene exactly replaced the normal one (by homologous recombination), and inserting that cell back into a mouse blastocyst. The resulting mouse will be chimeric but, if you are lucky, its germ cells will carry the deleted gene. A few rounds of careful breeding can then produce progeny in which both copies of the gene are inactivated.

KTUP: Change this value to limit the word-length the a Fasta search should use.

L

Leucine zipper: A motif found in certain proteins in which Leu residues are evenly spaced through an a-helical region, such that they would end up on the same face of the helix. Dimers can form between two such proteins. The Leu zipper is important in the function of transcription factors such as Fos and Jun and related proteins.

Leukaemia: Cancer that begins in developing blood cells in the bone marrow

Library: A library might be either a genomic library, or a cDNA library. In either case, the library is just a tube carrying a mixture of thousands of different clones - bacteria or l phages. Each clone carries an "insert" - the cloned DNA.

Ligand: Any small molecule that binds to a protein or receptor; the cognate partner of many cellular proteins, enzymes, and receptors.

Ligase: An enzyme, T4 DNA ligase, which can link pieces of DNA together. The pieces must have compatible ends (both of them blunt, or else mutually compatible sticky ends), and the ligation reaction requires ATP.

Ligation: The process of splicing two pieces of DNA together. In practice, a pool of DNA fragments are treated with ligase (see "Ligase") in the presence of ATP, and all possible splicing products are produced, including circularised forms and end-to-end ligation of 2, 3 or more pieces. Usually, only some of these products are useful, and the investigator must have some way of selecting the desirable one

Lindna: An **EMBOSS** application. The program 'lindna' draws linear maps of DNA constructs. It uses the graphical shapes: ticks, ranges, and blocks to represent genetic markers (e.g, genes and ESTs) and places them according to their position in a DNA fragment. The markers can be organised in different groups. The program reads in one input file in which the user specifies the names and positions of the genetic markers. In this file the user also enters data for controlling the appearance of the markers. A refined customisation of the drawing can be achieved by running the program with '-options' on the command line and changing the values of the desired parameters.

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.

Listor: An EMBOSS application. listor reads in two sets of sequences and writes out a list file (file of file names) that result from the logical union of these two sets of sequences. It is a simple way of manipulating and editing lists or sets of sequences to produce a list file. When comparing sequences to see if they are the same between two sets of sequences, no use is made of the ID name or accession number of the sequences. Only the sequences themselves are compared. The comparison of the sequences is case-independent. The logical union is an OR operation by default. Other available operations are: AND, XOR and NOT. The (default) logical OR of the two sets of sequences is simply the result of merging the two sets of sequences, (without listing any shared sequences twice). A logical AND simply lists those sequences that occur in both sets of sequences. A logical XOR lists those sequences that ONLY occur in the first set or only occur in the second set - sequences occuring in both sets are ignored (the opposite of an AND). A logical NOT lists all those sequences in the first set except for those that also occur in the second set.

Local alignment: An alignment that searches for segments of the two sequences that match well. There is no attempt to force entire sequences into an alignment, just those parts that appear to have good similarity, according to some criterion.

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

M/S with identities Alignment: The databases alignments are anchored (shown in relation to) to your query sequence. Identities are displayed as dots (.). Mismatches are displayed as single letter nucleotide abbreviations(c,t,a or g). Gaps are introduced with a "-" symbol.

M/S without identities alignment: The databases alignments are anchored (shown in relation to) to your query sequence. Identities are shown as single letter nucleotide abbreviations. Mismatches displayed as single letter nucleotide abbreviations(c,t,a or g). Gaps are introduced with a "-" symbol

Macromolecular Structure Database: Macromolecular Structure Database. The European project for the collection, management and distribution of data about macromolecular structures.

Markov chain: Any multivariate probability density whose independence diagram is a chain.The variables are ordered, and each variable "depends" only on its neighbours in the sense of being conditionally independent of the others. Markov chains are an integral component of hidden Markov models.

Marscan: An **EMBOSS** application. Description: Matrix/scaffold attachment regions (MARs/SARs) are genomic elements thought to delineate the structural and functional organisation of the eukaryotic genome. Originally, MARs and SARs were identified through their ability to bind to the nuclear matrix or scaffold. Binding cannot be assigned to a unique sequence element, but is dispersed over a region of several hundred base pairs. These elements are found flanking a gene or a small cluster of genes and are located often in the vicinity of cis-regulatory sequences. This has led to the suggestion that they contribute to higher order regulation of transcription by defining boundaries of independently controlled chromatin domains. There is indirect evidence to support this notion. In transgenic experiments MARs/SARs dampen position effects by shielding the transgene from the effects of the chromatin structure at the site of integration. Furthermore, they may act as boundary elements for enhancers, restricting their long range effect to only the promoters that are located in the same chromatin domain. marscan finds a bipartite sequence element that is unique for a large group of eukaryotic MARs/SARs. This MAR/SAR recognition signature (MRS) comprises two individual sequence elements that are <200 bp apart and may be aligned on positioned nucleosomes in MARs. The MRS can be used to correctly predict the position of MARs/SARs in plants and animals, based on genomic DNA sequence information alone. Experimental evidence from the analysis of >300 kb of sequence data from several eukaryotic organisms show that wherever a MRS is observed in the DNA sequence, the corresponding genomic fragment is a biochemically identifiable SAR. The MRS is a bipartite sequence element that consists of two individual sequences of 8 (AATAAYAA) and 16 bp (AWWRTAANNWWGNNNC) within a 200 bp distance from each other. One mismatch is allowed in the 16 bp pattern. The patterns can occur on either strand of the DNA with respect to each other. The 8 bp and the 16 bp sites can overlap. Where there are many possible MRS sites caused by many 8 bp and/or 16 bp pattern sites located within 200 bp of each other, then only the 8 bp site and the 16 bp site that occur closest to each other are reported. Once a MRS has been reported, no more sites will be looked for within 200 bp of that site. This reduces (but maybe will not totally eliminate) over-reporting of the clusters of MRS's that tend to occur within a MAR/SAR. Not all SARs contain a MRS. Analysis of >300 kb of genomic sequence from a variety of eukaryotic organisms shows that the MRS faithfully predicts 80% of MARs and SARs, suggesting that at least one other type of MAR/SAR may exist which does not contain a MRS. The problem of how to define and find MARs is still being actively invetsigated.

Maskfeat: An EMBOSS application. maskfeat reads in a sequence with its associated features. The features can be found in the annotation of the sequence if it is in a format such as EMBL or SWISSPROT which includes features in the annotation, or they may be supplied explicitly in a GFF file by using the command-lin option '-gff filename'. The feature table is then searched for features whose type matches the specified feature type to be masked. By default, the type is 'repeat*' (i.e. any type whose name starts with 'repeat'). You can specify the name of any other type of feature, or features that you wish to mask. If you wish to specify more than one type of feature, separate their names with spaces or commas. The names of the types of feature to be found may be wild-carded with asterisks '*' to find gruops of feature types sharing a common part of their names. If you are unsure of the names of feature types in use, please consult http://www3.ebi.ac.uk/Services/WebFeat/ for a list of the **EMBL** feature types and see Appendix A of the Swissprot user manual in http://www.expasy.ch/txt/userman.txt for a list of the Swissprot feature types. If any features matching the specified names of feature types are found, then those regions of the sequence will be masked out by replacing that part of the sequence by masking characters. The default masking characters are 'X' for a protein sequence and 'N' for a nucleic acid sequence, although you can specify your own masking character, if required.

Maskseq: An **EMBOSS** application. This simple editing program allows you to mask off regions of a sequence with a specified letter. Why would you wish to do this? It is common for database searches to mask out low-complexity or biased composition regions of a sequence so that spurious matches do not occur. It is just possible that you have a program that has reported such biased regions but which has not masked the sequence itself. In that case, you can use this program to do the masking. You may find other uses for it.

Matcher: An **EMBOSS** application. Description: matcher compares two sequences looking for local sequence similarities using a rigorous algorithm. matcher is based on Bill Pearson's 'lalign' application, version 2.0u4 Feb. 1996 Lalign uses code developed by X. Huang and W. Miller (Adv. Appl. Math. (1991) 12:337-357) for the "sim" program, which is a linear-space version of an algorithm described by M. S. Waterman and M. Eggert (J. Mol. Biol. 197:723-728). Like water, matcher is rigorous, but also very slow. The advantage of matcher is that it uses far less memory than water, so you are much less likely to run out of memory when aligning large sequences. matcher will also report a specified number of alignments between the two sequences showing the actual local alignments. (water will only report the single best match.) The default number of alignments output is 1, but can be increased to (for example) the 10 best alignments by using the '-alternatives 10' command-line qualifier. In some cases, for example multidomain proteins or cDNA and genomic DNA comparisons, there may be many interesting and significant alignments.

Matrix: It is assummed that the sequences being sought have an evolutionary ancestral sequence in common with the query sequence. The best guess at the actual path of evolution is the path that requires the fewest evolutionary events. All substitutions are not equally likely and should be weighted to account for this. Insertions and deletions are less likely than substitutions and should be weighted to account for this. It is necessary to consider that the choice of search algorithm influences the sensitivity and selectivity of the search. The choice of similarity **matrix** determines both the pattern and the extent of substitutions in the sequences the database search is most likely to discover. There have been extensive studies looking at the frequencies in which amino acids substituted for each other during evolution. The studies involved carefully aligning all of the proteins in several families of proteins and then constructing phylogenetic trees for each family. Each phylogenetic tree can then be examined for the substitutions found on each branch. This can then be used to produce tables(scoring matrices) of the relative frequencies with which amino acids replace each other over a short evolutionary period. Thus a substitution matrix describes the likelihood that two residue types would mutate to each other in evolutionary time. A substitution is more likely to occur between amino acids with similar biochemical properties. For example the hydrophobic amino acids Isoleucine(I) and valine(V) get a positive score on matrices adding weight to the likeliness that one will substitute for another. While the hydrophobic amino acid isoleucine has a negative score with the hydrophilic amino acid cystine(C) as the likeliness of this substitution occurring in the protein is far less. Thus matrices are used to estimate how well two residues of given types would match if they were aligned in a sequence alignment.

MaxSprout: **MaxSprout** is a fast database algorithm for generating protein backbone and side chain co-ordinates from a C(alpha) trace. The backbone is assembled from fragments taken from known structures. Side chain conformations are optimised in rotamer space using a rough potential energy function to avoid clashes.

MEDLINE: MEDLINE is a bibliographic database covering the fields of medicine, nursing, dentistry, veterinary medicine, the health care system, and the pre-clinical sciences. MEDLINE searches are available using the EBI´s **SRS** server.

Megamerger: An **EMBOSS** application. megamerger takes two overlapping sequences and merges them into one sequence. It could thus be regarded as the opposite of what splitter does. The sequences can be very long. The program does a match of all sequence words of size 20 (by default). It then reduces this to the minimum set of overlapping matches by sorting the matches in order of size (largest size first) and then for each such match it removes any smaller matches that overlap. The result is a set of the longest ungapped alignments between the two sequences that do not overlap with each other. If the two sequences are identical in their region of overlap then there will be one region of match and no mismatches. It should be possible to merge sequences that are Mega bytes long. Compare this with the program merger which does a more accurate alignment of more divergent sequences using the Needle and Wunsch algorithm but which uses much more memory. The sequences should ideally be identical in their region of overlap. If there are any mismatches between the two sequences then megamerger will still attempt to create a merged sequence, but you should check that this is what you required. A report of the actions of megamerger is written out. Any actions that require a choice between using regions of the two sequences where they have a mismatch is marked with the word WARNING!. The sequence in these regions is written out in uppercase. All other regions of the output sequence are written in lowercase. Where there is a mismatch then the sequence that is chosen to supply the region of the mismatch in the final merged sequence is that sequence whose mismatch region is furthest from the start or end of the sequence

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

Merger: An **EMBOSS** application. Description: This joins two overlapping nucleic acid sequences into one merged sequence. It uses a global alignment algorithm (Needleman & Wunsch) to optimally align the sequences and then it creates the merged sequence from the alignment. When there is a mismatch in the alignment between the two sequences, the correct base to include in the resulting sequence is chosen by using the base from the sequence which has the best local sequence quality score. The following heuristic is used to find the sequence quality score: If one of the bases is a 'N', then the other sequence's base is used, otherwise: A window size around the disputed base is used to find the local quality score. This window size is increased from 5, to 10 to 20 bases or until there is a clear decision on the best choice. If there is no best choice after using a window of 20, then the base in the first sequence is used.

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).

MGED: The **Microarray Gene Expression Data** (MGED) Society is an international organisation of biologists, computer scientists, and data analysts that aims to facilitate the sharing of microarray data generated by functional genomics and proteomics experiments.

MIAMExpress: Microarray data submissions to **ArrayExpress** database.

Microarray: Microarrays allow snapshots to be made of expression levels for thousands of genes in a single experiment.

Microarray Group: Microarrays allow snapshots to be made of expression levels for thousands of genes in a single experiment. They are already generating massive amounts of valuable functional genomics data. The **Microarray Informatics Team** at the **EBI** was established in May 2000 to address this problem of managing and analysing this data.

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\

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

Motif: A conserved element of a protein sequence alignment that usually correlates with a particular function. Motifs are generated from a local multiple protein sequence alignment corresponding to a region whose function or structure is known. It is sufficient that it is conserved, and is hence likely to be predictive of any subsequent occurrence of such a structural/functional region in any other novel protein sequence.

MPsrch: **MPsrch** is a biological protein sequence comparison tool that implements the true Smith and Waterman algorithm. It runs a search on a HP/COMPAQ cluster, a family of massively parallel computers. It allows an exhaustive search in a reasonable computational time. MPsrch utilises an exhaustive algorithm, which is recognised as the most sensitive sequence comparison method available, whereas Blast utilises an heuristic one, which speeds up searches by reducing the complexity of the problem. As a consequence, MPsrch is capable of identifying hits in cases where Blast fails with fewer false hits.

MPsrch_pp: Uses a protein query to search a protein sequence database, using amino-acid match scoring derived from a specified table. Given a novel protein sequence, this is the standard choice to detect related proteins already known in the databases. Only a single gap penalty is used during the search; in most cases, the best alignments between related sequences do not involve long gaps or regions with multiple gaps.

MPsrch_ppa: Uses a protein query to search a protein sequence database, using amino-acid match scoring derived from a specified table. In addition, the inclusion of gaps is controlled by the two penalties known as gapopen and gapextend. Given a novel protein sequence, this may be the choice to detect distantly related proteins in the databases. These proteins may align with long gapped regions, possibly in loops on protein surfaces that may not contain critical functional residues. If you want to detect weak candidate alignments it may be necessary to repeat the searches with a variety of gap penalties.

mRNA: This is messenger RNA, it is a copy of the information carried by a gene on the DNA. The role of mRNA is to move the information contained in DNA to the translation machinery (ribosomes).

Msbar: An **EMBOSS** application. Description: This program changes a sequence a lot or a little, attempting to emulate various forms of mutation. You can set the number and types of mutations. It can act on the following sizes of sequence: Point (single base or residue change) Codon (not applicable in proteins) Block of sequence (of a specified minimum and maximum random size) If the sequence is nucleic, the codon and block-sized operations can optionally be done in-frame. This causes the minimum block size to be set to 3 and the randomly chosen positions to be multiples of 3. For each of the above size of sequence it can produce the effects of any of the following types of mutation at a randomly chosen position: Insertion of a randomly generated sequence Deletion Change (deletion then insertion of a random sequence of the same size) Duplication at an adjacent position Move region from one position to another (without deletion of the original) Any of the above, chosen at random. None of the above The input and output sequences may not differ if only a few changes are chosen as (for example) one in four nucleic acid point substitutions will not change the sequence. N.B. There is no selection of the types of mutation to produce viable sequence as there would be in a real organism. In particular, there is no attempt to bias mutations of nucleic acid sequences to conform to the C+G ratio in the sequence or to bias the codons in the direction of the frequencies used in the organism. This program emulates mutation, not selection. This program was named from the acronym of "Mutate Sequence Beyond All Recognition", by analogy with the acronym "fubar" commonly used in the US and UK armed forces.

MSD: Macromolecular Structure Database. The European project for the collection, management and distribution of data about macromolecular structures.

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 A, that satisfies the following conditions: There are exactly k rows; ignoring the gap character, row number i is exactly the sequence sI; 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.

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.

Mutations Database: **Sequence Variation Database**, aims to record changes in genomes and map their effects at higher levels of cellular and organismic information processing.

Mwcontam: An **EMBOSS** application. Description: mwcontam finds molecular weights that are common between a set of mass spectrometry result files. Such molecular weights are usually a form of contamination resulting from autolysis of a protease, degradation of the matrix or presence of keratin from the sample preparer. The output of mwcontam, with minimal editing, can be added to the data file for the mwfilter program ('Emwfilter.dat').

Mwfilter: An **EMBOSS** application. Description: mwfilter is designed to remove unwanted (noisy) data from mass spectrometry output in proteomics. Given a list of molecular weights this program removes those which are: Contaminating trypsin or keratin Modified oxy-methionine or oxy-threonine Peaks associated with sodium ions. The last two operations can be done as most peaks are reported in both modified and unmodified forms. Removal of modified peaks aids in database searching for protein identification.

myGrid: **myGrid**. Developing the infrastructural middleware necessary for an "e-Biologist's" workbench.

N

NCBI-Blast2: BLAST stands for Basic Local Alignment Search Tool.The emphasis of this tool is to find regions of sequence similarity, which will yield functional and evolutionary clues about the structure and function of your novel sequence. **WU-BLAST 2.0** and **NCBI BLAST2** are distinctly different software packages, although they have a common lineage for some portions of their code, so the two packages do their work differently and obtain different results and offer different features.

Needle: An **EMBOSS** application. This program uses the Needleman-Wunsch global alignment algorithm to find the optimum alignment (including gaps) of two sequences when considering their entire length. The Needleman-Wunsch algorithm is a member of the class of algorithms that can calculate the best score and alignment in the order of mn steps, (where 'n' and 'm' are the lengths of the two sequences). These dynamic programming algorithms were first developed for protein sequence comparison by Needleman and Wunsch, though similar methods were independently devised during the late 1960's and early 1970's for use in the fields of speech processing and computer science. What is the optimal alignment? Dynamic programming methods ensure the optimal global alignment by exploring all possible alignments and choosing the best. It does this by reading in a scoring matrix that contains values for every possible residue or nucleotide match. Needle finds an alignment with the maximum possible score where the score of an alignment is equal to the sum of the matches taken from the scoring matrix. An important problem is the treatment of gaps, i.e., spaces inserted to optimise the alignment score. A penalty is subtracted from the score for each gap opened (the 'gap open' penalty) and a penalty is subtracted from the score for the total number of gap spaces multiplied by a cost (the 'gap extension' penalty). Typically, the cost of extending a gap is set to be 5-10 times lower than the cost for opening a gap

Needle Program: This is a true implementation of the Needleman-Wunsch algorithm and so produces a full path matrix. It therefore cannot be used with genome sized sequences unless you have a lot of memory and a lot of time. Needle is for aligning two sequences over their entire length. This works best with closely related sequences. If you use needle to align very distantly-related sequences, it will produce a result but much of the alignment may have little or no biological significance. A true Needleman Wunsch implementation like needle needs memory proportional to the product of the sequence lengths. For two sequences of length 10,000,000 and 1,000 it therefore needs memory proportional to 10,000,000,000 characters. Two arrays of this size are produced, one of integers and one of floats so multiply that figure by 8 to get the memory usage in bytes. That doesn't include other overheads. Therefore only use water and needle for accurate alignment of reasonably short sequences.

New EBI Web Taxonomy browser: A taxonomy database, integrating taxonomy data compiled at NCBI and data specific to the **Swiss-Prot** protein knowledgebase

Newcpgreport: An **EMBOSS** application. This application is used in the production of the CpG Island database 'CPGISLE'. It produces CPGISLE database entry format reports for a potential CpG island. See the FTP site: ftp://ftp.ebi.ac.uk/pub/databases/cpgisle/ for the finished database. CpG refers to a C nucleotide immediately followed by a G. The 'p' in 'CpG' refers to the phosphate group linking the two bases. Detection of regions of genomic sequences that are rich in the CpG pattern is important because such regions are resistant to methylation and tend to be associated with genes which are frequently switched on. Regions rich in the CpG pattern are known as CpG islands. It has been estimated that about half of all mammalian genes have a CpG-rich region around their 5' end. It is said that all mammalian house-keeping genes have a CpG island! Non-mammalian vertebrates have some CpG islands that are associated with genes, but the association gets equivocal in the farther taxonomic groups. Finding a CpG island upstream of predicted exons or genes is good contributory evidence for that gene's existance. By default, this program defines a CpG island as a region where, over an average of 10 windows, the calculated % composition is over 50% and the calculated Obs/Exp ratio is over 0.6 and the conditions hold for a minimum of 200 bases. These conditions can be modified by setting the values of the appropriate parameters. The Expected number of CpG patterns in a window is calculated as the number of 'C's in the window multiplied by the number of 'G's in the window, divided by the window length. This program reads in one or more sequences and finds regions where there is a high absolute frequency of CpG dimers as well as a high proportion of CpG compared to GpC.

Newcpgseek: An **EMBOSS** application. newcpgseek reports CpG rich regions of a sequence as candidate CpG islands. CpG refers to a C nucleotide immediately followed by a G. The 'p' in 'CpG' refers to the phosphate group linking the two bases. Detection of regions of genomic sequences that are rich in the CpG pattern is important because such regions are resistant to methylation and tend to be associated with genes which are frequently switched on. Regions rich in the CpG pattern are known as CpG islands. It has been estimated that about half of all mammalian genes have a CpG-rich region around their 5' end. It is said that all mammalian house-keeping genes have a CpG island! Non-mammalian vertebrates have some CpG islands that are associated with genes, but the association gets equivocal in the farther taxonomic groups. Finding a CpG island upstream of predicted exons or genes is good contributory evidence. CpG islands are usually defined as >200bp with %GC > 50% and obs/exp CpG > 0.6". However this program uses a running sum rather than a window to produce a score: if there is not a CpG at position i, then decrement runSum counter, but if CpG then runSum += CPGSCORE. Spans > threshold are searched for recursively. If the score is higher than a threshold (17 at the moment) then a putative island is declared. This program reads in one or more sequences and finds regions where there is a high absolute frequency of CpG dimers as well as a high proportion of CpG compared to GpC.

Newseq: An **EMBOSS** application. This allows you to type a sequence into a file in a quick and easy manner. The length of the sequence you can type in is restricted to a fairly short length (typically less than 255 characters). This length restriction is not a property of the **EMBOSS** package, but of the computer system you are using. This is because, as you type in response to a prompt from this program, what you type is stored in the computer operating system before being handed over to the program. There is often a limit of less than 255 characters on the length of a response that a computer system will allow you to give. Despite this restriction, it is expected that this program will be a useful and easy way of constructing new sequence files. (You wouldn't want to type a long sequence in by hand, anyway, would you?)

NEWT: New **EBI** Web Taxonomy browser. A taxonomy database, integrating taxonomy data compiled at NCBI and data specific to the **Swiss-Prot** protein knowledgebase.

Noreturn: An **EMBOSS** application. The way that Unix and PC operating systems store simple text files, (including sequence files), differs slightly. Unix files have a hidden character called 'new line' at the end of every line. PC files have two hidden characters called 'carriage return' and then 'new line' at the end of every line. When files are transferred from PC machines to Unix machines, it is often useful to convert the file from the PC format to the Unix format, otherwise commands like 'more', to display the file, and text editors can become confused. This simple utility removes 'carriage return' characters from such files, converting them from PC format to Unix format text files. **EMBOSS** programs can read in both PC and Unix text file formats, so it is not necessary for you to use this utility all of the time.

Northern blot: A technique for analyzing mixtures of RNA, whereby the presence and rough size of one particular type of RNA (usually an mRNA) can be ascertained. See "Blotting" for more information. After Dr. E. M. Southern invented the Southern blot, it was adapted to RNA and named the "Northern" blot.

Notseq: An **EMBOSS** application. When you have a set of sequences (a file of multiple sequences) and you wish to remove one or more of them from the set, then use notseq. This program was written for the case where a file containing several sequences is being used as a small database, but some of the sequences are no longer required and must be deleted from the file. notseq splits the input sequences into those that you wish to keep and those you wish to exclude. notseq takes a set of sequences as input together with a list of sequence names or accession numbers. It also takes the name of a new file to write the files that you want to keep into, and optionally the name of a file that will contain the files that you want excluded from the set. notseq then reads in the input sequences. It outputs the ones that match one of the sequence names or accession numbers to the file of excluded sequences, and those that don't match are output to the file of sequences to be kept. Note that the names of the sequences to be excluded are not standard **EMBOSS** USAs. Only the name or accession number should be specified, not the database or file that these entries may occur in. These excluded sequence names will be matched against the names of the input sequences to see if there is a match. Wildcarded names may be specified by using '*'s. Any specified names of sequences to be excluded that are not found are simply ignored.

Nthseq: An **EMBOSS** application. In EMBOSS, when an application has to write out many sequences, the normal style is to write them all into one file containing multiple sequences. This default behaviour can be changed by using the qualifier '-ossingle' which writes many sequences into many files, each containing one sequence. The program seqretsplit will take a file containing many sequences and will output many files, each containing one sequence. However you have no choice over the naming of the files - they are named after the ID name of the sequence they contain. If, however you have a situation where you have a file containing multiple sequences and you wish to extract one of them, then this application may be useful. nthseq allows you to specify the name of the output file, so you may find that it is useful to include this program in scripts where you need to be able to specify the name of the resulting sequence files you create. This application extracts the indicated sequence from a multiple set of sequences and writes it out.

Nuclear run-on: A method used to estimate the relative rate of transcription of a given gene, as opposed to the steady-state level of the mRNA transcript (which is influenced not just by transcription rates, but by the stability of the RNA). This technique is based on the assumption that a highly-transcribed gene should have more molecules of RNA polymerase bound to it than will the same gene in a less-active state. If properly prepared, isolated nuclei will continue to transcribe genes and incorporate 32P into RNA, but only in those transcripts that were in progress at the time the nuclei were isolated. Once the polymerase molecules complete the transcript they have in progress, they should not be able to re-initiate transcription. If that is true, then the amount of radiolabel incorporated into a specific type of mRNA is theoretically proportional to the number of RNA polymerase complexes present on that gene at the time of isolation. A very difficult technique, rarely applied appropriately.

Nuclease: An enzyme which degrades nucleic acids. A nuclease can be DNA-specific (a DNase), RNA-specific (RNase) or non-specific. It may act only on single stranded nucleic acids, or only on double-stranded nucleic acids, or it may be non-specific with respect to strandedness. A nuclease may degrade only from an end (an exonuclease), or may be able to start in the middle of a strand (an endonuclease). To further complicate matters, many enzymes have multiple functions; for example, Bal31 has a 3'-exonuclease activity on double-stranded DNA, and an endonuclease activity specific for single-stranded DNA or RNA

Nucleotide: A nucleic acid unit composed of a five carbon sugar joined to a phosphate group and a nitrogen base.

Nucleotide bases: Nucleotide bases fall into two categories depending on the ring structure of the base. Purines (Adenine and Guanine) are two ring bases, pyrimidines (Cytosine and Thymine) are single ring bases. Mutations in DNA are changes in which one base is replaced by another. A mutation that conserves the ring number is called a transition (e.g., A -> G or C -> T) a mutation that changes the ring number are called transversions. (e.g. A -> C or A -> T and so on).

O

Octanol: An **EMBOSS** application. Protein sequences that form transmembrane regions are assumed to have a thermodynamic preference for a hydrophobic environment (inside the membrane lipid bilayer), rather than an aqueous environment in water. The free energy change for each amino acid residue between a lipid and a water environment can be measured experimentally, and the values for peptides can be shown to be additive (White and Wimley 1999). The octanol program calculates two free energy differences. The first is the free energy difference between solution in water and association with the interface (glycerol group) of a POPC (palmitoyloleoylphosphocholine) bilayer. The second is the free energy difference between water and octanol, equivalent to the environment inside a lipid bilayer. Residues which can be buried inside a lipid bilayer must be in a region of the peptide where most residues show a free energy difference in favour of being in an octanol environment or at least being in the lipid/water interface region. White and Wimley (1999) showed that a sliding window of either free energy difference will indicate the location of probably transmembrane regions, but that the best indicator is the difference between the two values, which is the free energy difference between the interface and octanol environments. The free energies are calculated over a sliding window of 19 residues, about the size of a membrane spanning alphahelix. The energy values for each residue are added over the window.

Oddcomp: An **EMBOSS** application. oddcomp searches a series of protein files, reporting the identifier for those that exceed a certain amino acid composition threshold in a portion of the sequence. oddcomp was written to answer the question 'which proteins contain at least n X and m Y in p residues'. One could search for serine rich or polyglutamine rich, collagen helix, or similar proteins using this program. oddcomp takes as input an amino acid composition data file in the same format as the output from compseq. It can tolerate any word length within reason for the memory capacity of the machine in question. Only the first two fields in the composition data file are used: the word and the number of occurrences. Any word not mentioned is initialised to a threshold of zero. oddcomp measures the amino acid composition (this can be dimers etc as well as monomers) in a sliding window. If and when composition meets or exceeds all the specified thresholds, the sequence is reported and oddcomp moves to the next sequence. It does not report where in the sequence it found the matching region, merely the sequence ID. oddcomp was originally written to identify SR/RS containing proteins. eg. specifying a window of forty amino acids containing at least 3 SR and 4 RS words. To search for a specific set of words in a sequence, edit the input composition data file to delete any words from the input file in which you are not interested. The search is a boolean AND so there must be (from the short example above) at least 3 SR AND at least 4 RS for the sequence to be reported. If your total words specified exceeds window-wordlength+1 you will never get any hits. Only one word size can be used and is specified at the top of the file in compseq-style output.

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

Oncogene: A gene in a tumor virus or in cancerous cells which, when transferred into other cells, can cause transformation (note that only certain cells are susceptible to transformation by any one oncogene). Functional oncogenes are not present in normal cells. A normal cell has many "proto-oncogenes" which serve normal functions, and which under the right circumstances can be activated to become oncogenes. The prefix "v-" indicates that a gene is derived from a virus, and is generally an oncogene (like v-src , v-ras, v-myb , etc). See also "Transformation (with respect to cultured cells)".

Open reading frame: Any region of DNA or RNA where a protein could be encoded. In other words, there must be a string of nucleotides (possibly starting with a Met codon) in which one of the three reading frames has no stop codons. See "Reading frame" for a simple example.

Organism Classification line (EMBL): The OC (Organism Classification) lines contain the taxonomic classification of the source organism.

Organism Classification lines (SWISS-PROT): The OC (Organism Classification) lines contain the taxonomic classification of the source organism.

Organism Species line (EMBL): The OS (Organism Species) line specifies the preferred scientific name of the organism which was the source of the stored sequence.

Organism Species line (SWISS-PROT): The OS (Organism Species) line specifies the organism(s) which was (were) the source of the stored sequence.

Organism taxonomy Cross-Reference line: The OX (Organism taxonomy Cross-Reference) line is used to indicate the identifier to a specific organism in a taxonomic database.

ORIEL: (Online Research Information Environment for the Life Sciences) ORIEL aims to explore and develop methods and technologies for integration, exploitation and dissemination of large disparate information resources. It is a companion project to E-BioSci and seeks EU funds under IST. Co-ordinated by EMBO, it involves other partners: EBI, University of Oxford, ingenta UK Ltd, CINES, CNR-ITB, ICGEB, and CSIC.

P

p-value: probability value. i.e. sorts blast output from most statistically significant (lowest P-value) to least statistically (highest P-value). The P-values are a function of N, as used in Karlin-Altschul Sump (sum probability) statistics or Poisson statistics.

pairwise Alignment: Aligns your query sequence and database matches in pairs. Matches are connected with a "|" symbol. Mismatches are opposed with a spce. Gaps are introduced with a "-" symbol. e.g.

palindrome: An **EMBOSS** application. palindrome looks for inverted repeats (stem loops) in a nucleotide sequence. It will find inverted repeats that include a proportion of mismatches and gaps (bulges in the stem loop). It works by finding all possible inverted matches satisfying the specified conditions of minimum and maximum length of palindrome, maximum gap between repeated regions and number of mismatches allowed. Secondary structures like inverted repeats in genomic sequences may be implicated in initiation of DNA replication. Some genomic sequence entries in the databases are composed of unfinished, draft sequence with gaps of unknown size between contigs. The positions of these gaps are often indicated by runs of 200 N characters. To prevent palindrome producing large, uninformative outputs, any palindromes found that are composed only of N's will not be reported.

Parasite Genomes WU-Blast2: This application allows you to enter a sequence and perform **BLAST** searches with it against the different parasite genome databases available.

Pasteseq: An **EMBOSS** application. This simple editing program allows you to insert one sequence into another sequence after a specified position and to then write out the results to a sequence file.

Patmatdb: An **EMBOSS** application. Takes a protein motif and compares it to a set of protein sequences. It returns the number of matches there were between the motif and each matched sequence, length of match, start and end positions of match, and writes out an alignment.

patmatdb-emboss: An **EMBOSS** application. Takes a protein motif and compares it to a set of protein sequences. It returns the number of matches there were between the motif and each matched sequence, length of match, start and end positions of match, and writes out an alignment.

Patmatmotifs: An **EMBOSS** application. patmatmotifs takes a protein sequence and compares it to the PROSITE database of motifs. For a description of PROSITE, we can do no better than to quote the PROSITE user's documentation: PROSITE is a method of determining what is the function of uncharacterized proteins translated from genomic or cDNA sequences. It consists of a database of biologically significant sites and patterns formulated in such a way that with appropriate computational tools it can rapidly and reliably identify to which known family of protein (if any) the new sequence belongs. In some cases the sequence of an unknown protein is too distantly related to any protein of known structure to detect its resemblance by overall sequence alignment, but it can be identified by the occurrence in its sequence of a particular cluster of residue types which is variously known as a pattern, motif, signature, or fingerprint. These motifs arise because of particular requirements on the structure of specific region(s) of a protein which may be important, for example, for their binding properties or for their enzymatic activity. These requirements impose very tight constraints on the evolution of those limited (in size) but important portion(s) of a protein sequence. To paraphrase Orwell, in Animal Farm, we can say that "some regions of a protein sequence are more equal than others" ! The use of protein sequence patterns (or motifs) to determine the function(s) of proteins is becoming very rapidly one of the essential tools of sequence analysis. This reality has been recognized by many authors, as it can be illustrated from the following citations from two of the most well known experts of protein sequence analysis, R.F. Doolittle and A.M. Lesk: "There are many short sequences that are often (but not always) diagnostics of certain binding properties or active sites. These can be set into a small subcollection and searched against your sequence (1)". "In some cases, the structure and function of an unknown protein which is too distantly related to any protein of known structure to detect its affinity by overall sequence alignment may be identified by its possession of a particular cluster of residues types classified as a motifs. The motifs, or templates, or fingerprints, arise because of particular requirements of binding sites that impose very tight constraint on the evolution of portions of a protein sequence (2)." The home web page of PROSITE is: http://www.expasy.ch/prosite/ It is common to find that a search of the PROSITE database against a protein sequence will report many matches to the short motifs that are indicative of the post-translational modification sites, such as glycolsylation, myristylation and phosphorylation sites. These reports are often unwanted and are not normally reported. You can turn reporting of these short motifs on by giving the '-noprune' option on the command-line. Your **EMBOSS** administrator must have set up the local **EMBOSS** PROSITE database using the utility 'prosextract' before this program will run.

PCR: A technique for replicating a specific piece of DNA in-vitro , even in the presence of excess non-specific DNA. Primers are added (which initiate the copying of each strand) along with nucleotides and Taq polymerase. By cycling the temperature, the target DNA is repetitively denatured and copied. A single copy of the target DNA, even if mixed in with other undesirable DNA, can be amplified to obtain billions of replicates. PCR can be used to amplify RNA sequences if they are first converted to DNA via reverse transcriptase. This two-phase procedure is known as ‘RT-PCR’. Polymerase Chain Reaction (PCR) is the basis for a number of extremely important methods in molecular biology. It can be used to detect and measure vanishingly small amounts of DNA and to create customised pieces of DNA. It has been applied to clinical diagnosis and therapy, to forensics and to vast numbers of research applications. It would be difficult to overstate the importance of PCR to science.

PDB: Brookhaven Protein Sequence Database, the single worldwide repository for the processing and distribution of 3-D biological macromolecular structure data.

Penetrance: A term indicating the likelihood that a given gene will actually result in disease.

Pepcoil: An **EMBOSS** application. Coiled coils are formed by two or three alpha helices in parallel and in register that cross at an angle of approximately 20 degrees, are strongly amphipathic and display a pattern of hydrophilic and hydrophobic residues that is repeated every seven residues. The seven positions of the heptad repeat are designated a through g, a and d being generally hydrophobic, while the others are hydrophilic. The parallel two-stranded alpha-helical coiled coil is the most frequently encountered subunit-oligomerization motif in proteins. pepcoil calculates the probability of a coiled-coil structure for windows of 28 residues through a protein sequence using the method of Lupas A, van Dyke M & Stock J (1991); Science 252:1162-4

Pepinfo: Pepinfo detects and displays various useful metrics about a protein sequence. It can plot and display the following properties: plots of hydrophobicity (using the method of Kyte & Doolittle), of OHM (Sweet & Eisenberg), or of consensus parameters (Eisenberg et al)) a histogram of the presence of residues with the physicochemical properties: Tiny, Small, Aliphatic, Aromatic, Non-polar, Polar, Charged, Positive, Negative.

pepinfo-emboss: An **EMBOSS** application. pepinfo detects and displays various useful metrics about a protein sequence. It can plot and display the following properties: plots of hydrophobicity (using the method of Kyte & Doolittle), of OHM (Sweet & Eisenberg), or of consensus parameters (Eisenberg et al)) a histogram of the presence of residues with the physico-chemical properties: Tiny, Small, Aliphatic, Aromatic, Non-polar, Polar, Charged, Positive, Negative. The data are also written out to a data file.

Pepnet: An **EMBOSS** application. This is a method of displaying the residues of a protein in a simple 3,4,3,4 repeating pattern that emulates at a simple level the arrangement of residues around an alpha helix. It is therefore easy to see patterns of amphipathicity that you may wish to investigate in more detail by using displays such as pepwheel. You can specify which residues to mark up in squares, diamonds and octagons.

Pepstats: Pepstats Outputs a report of simple protein sequence information including: molecular weight, number of residues, average residue weight charge, iso electric point, for each type of amino acid: number, molar percent, DayhoffStat, for each physicochemical class of amino acid: number, molar percent, DayhoffStat is the amino acid's Dayhoff statistic divided by the molar percent. The Dayhoff statistic is the amino acid's relative occurrence per 1000 aa normalised to 100 by rls@ebi.ac.uk (original work from 1993)

Pepwheel: An **EMBOSS** application. pepwheel displays peptide sequences in a helical representation. This gives a view of a helix from a protein sequence looking down the axis of the helix. It is useful for highlighting amphipathicity and other properties of residues around a helix.

Pepwindow: Reads in a protein sequence and displays a graph of the classic Kyte & Doolittle hydropathy plot of that protein.

pepwindow-emboss: An **EMBOSS** application. pepwindow reads in a protein sequence and displays a graph of the classic Kyte & Doolittle hydropathy plot of that protein.

Pepwindowall: An **EMBOSS** application. pepwindowall produces a set of superimposed Kyte & Doolittle hydropathy plots from an aligned set of protein sequences. The result is the same as running pepwindow on a set of proteins with aligning gaps and superimposing the plots. It is useful for visualising the average hydropathy and its variability along the alignment.

Pestfind: An **EMBOSS** application. pestfind allows rapid and objective identification of PEST motifs in protein target sequences. Briefly, the PEST hypothesis was based on a literature survey that combined both, information on protein stability as well as protein primary sequence information. Initially, the study relied on 12 short-lived proteins with well-known properties, but was continually extended later. The initial group of proteins included E1A, c-myc, p53, c-fos, v-myb, P730 phytochrome, heat shock protein 70 (HSP 70), HMG-CoA reductase, tyrosine aminotransferase (TAT), ornithine decarboxylase (ODC), alpha-Casein and beta-Casein. Although all these proteins exerted various different cellular functions it became apparent that they shared high local concentrations of amino acids proline (P), glutamic acid (E), serine (S), threonine (T) and to a lesser extent aspartic acid (D). From that it was concluded that PEST motifs reduce the half-lives of proteins dramatically and hence, that they target proteins for proteolytic degradation. PEST means Black Death in German, so that the name of this programme sounds a bit strange, at least in our ears.

PfScan: Tool that scans against PROSITE profiles. These profiles are based on weight matrices and are more sensitive for the detection of divergent protein families.

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

Phage display: A technique in which phage are engineered to fuse a foreign peptide or protein with their capsid (surface) proteins and hence display it on their cell surfaces. The immobilized phage may then be used as a screen to see what ligands bind to the expressed fusion protein exhibited (displayed) on the phage surface.

Phagemid: A type of plasmid which carries within its sequence a bacteriophage replication origin. When the host bacterium is infected with "helper" phage, the phagemid is replicated along with the phage DNA and packaged into phage capsids.

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

Phylip Format: The first line of the input file contains the number of species, the number of sequences and their length (in characters) separated by blanks. The next line contains the sequence name, followed by the sequence in blocks of 10 characters.

Phylogram: Phylogram is a branching diagram (**tree**) assumed to be an estimate of a phylogeny, branch lengths are proportional to the amount of inferred evolutionary change

Phylum: The segmentation of the animal kingdom into about 30 major groups collectively known as phyla. The members of each phylum share the same basic structure and organisation. For instance, fish, birds, and human beings belong to one phylum - the Chordata - because all have spinal cords.

Plasmid: A circular piece of DNA present in bacteria or isolated from bacteria. Escherichia coli, the usual bacteria in molecular genetics experiments, has a large circular genome, but it will also replicate smaller circular DNAs as long as they have an "origin of replication". Plasmids may also have other DNA inserted by the investigator. A bacterium carrying a plasmid and replicating a million-fold will produce a million identical copies of that plasmid. Common plasmids are pBR322, pGEM, pUC18.

Pleitropy: The multiple effects on an organism's 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.

Plotcon: An **EMBOSS** application. Displays a graphical representation of the similarity along a set of aligned sequences. The similarity is calculated by moving a window of a specified length along the aligned sequences. Within the window, the similarity of any one position is taken to be the average of all the possible pairwise scores of the bases or residues at that position. The pairwise scores are taken from the specified similarity matrix. The average of the position similarities within the window is plotted. The program is useful for determining where the quality of alignments is good or bad.

Plotorf: An **EMBOSS** application. Plot potential open reading frames. A graphical representation of where the open reading frames are in all 6 reading frames is shown. The ORFs are displayed as blue boxes. ORFs in this program are defined as being regions between START and STOP codons. Note that this definition of an ORF would miss those exons in eukaryotic genomic sequences which do not contain a START codon. plotorf is only really useful when dealing with prokaryotic or mRNA eukaryotic sequences. The default START codon is: "ATG". The default STOP codons are: "TAA,TAG,TGA". You can specify your own set of start and stop codons using the -start and -stop qualifiers.

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

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

poisson statistics: The occurrence of two or more HSPs involving the query sequence and the same database sequence can be modelled as a Poisson process by selecting this option. An important result of applying Poisson statistics is that an HSP having a low score and high Expect value (low statistical significance) may be ascribed a statistically significant Poisson P-value when the HSP appears in the context of additional match(es) of equal or greater score with the same database sequence. The Poisson P-value for any given HSP is a function of its expected frequency of occurrence and the number of HSPs observed against the same database sequence with scores at least as high. The Poisson Pvalue for a group of HSP events is the probability that at least as many HSPs would occur by chance alone, each with a score at least as high as the lowest-scoring member of the group. HSPs which appear on opposite strands of a nucleotide query or database sequence are considered to be independent, distinguishable events, and are counted separately.

PolyA tail: After an mRNA is transcribed from a gene, the cell adds a stretch of A residues (typically 50-200) to its 3' end. It is thought that the presence of this "polyA tail" increases the stability of the mRNA (possibly by protecting it from nucleases). Note that not all mRNAs have a polyA tail; the histone mRNAs in particular do not.

Polyadenylation site: A site on the 3'-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. These so-called poly(A) "tails" increase mRNA stability and allow one to isolate mRNA from cells by PCR-amplification using poly(T) primers.

Polydot: An **EMBOSS** application. A dotplot is a graphical representation of the regions of similarity between two sequences. The two sequences are placed on the axes of a rectangular image and (subject to threshold conditions) wherever there is a similarity between the sequences a dot is placed on the image. Where the two sequences have substantial regions of similarity, many dots align to form diagonal lines. It is therefore possible to see at a glance where there are local regions of similarity. polydot compares all sequences in a set of sequences, draws a dotplot for each pair of sequences by marking where words (tuples) of a specified length have an exact match in both sequences and optionally reports all identical matches to feature files.

Polygenic inheritance: Inheritance involving alleles at many genetic loci.

Polymerase: An enzyme which links individual nucleotides together into a long strand, using another strand as a template. There are two general types of polymerase — DNA polymerases (which synthesize DNA) and RNA polymerase (which makes RNA). Within these two classes, there are numerous sub-types of polymerase, depending on what type of nucleic acid can function as template and what type of nucleic acid is formed. A DNA-dependant DNA polymerase will copy one DNA strand starting from a primer, and the product will be the complementary DNA strand. A DNA-dependant RNA polymerase will use DNA as a template to synthesise an RNA strand.

Polymerase chain reaction: A technique for replicating a specific piece of DNA in-vitro , even in the presence of excess non-specific DNA. Primers are added (which initiate the copying of each strand) along with nucleotides and Taq polymerase. By cycling the temperature, the target DNA is repetitively denatured and copied. A single copy of the target DNA, even if mixed in with other undesirable DNA, can be amplified to obtain billions of replicates. PCR can be used to amplify RNA sequences if they are first converted to DNA via reverse transcriptase. This two-phase procedure is known as 'RT-PCR'. Polymerase Chain Reaction (PCR) is the basis for a number of extremely important methods in molecular biology. It can be used to detect and measure vanishingly small amounts of DNA and to create customized pieces of DNA. It has been applied to clinical diagnosis and therapy, to forensics and to vast numbers of research applications. It would be difficult to overstate the importance of PCR to science.

Polymorphism: (lit. many forms) The existence of a gene in a population in at least two different forms at a frequency far higher than that attributable to recurrent mutation alone. Variations in a population may be measured by determining the rate of mutation in polymorphic genes (see SNPs).

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. The location of the gene is determined by linkage analysis of DNA from a large family containing afflicted and normal members to identify linkages between the transmission of the disease gene and observable genetic markers. This information is then used to screen (by chromosomal jumping and walking) the location for putative genes. The disease gene must be compared between the afflicted and normal family members and be shown to be different in the two groups. The full sequencing of the gene will then provide information regarding the characteristics and function of the gene product, and a potential explanation for the cause of the disease.

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

Post-transcriptional regulation: Any process occurring after transcription which affects the amount of protein a gene produces. Includes RNA processing efficiency, RNA stability, translation efficiency, protein stability. For example, the rapid degradation of an mRNA will reduce the amount of protein arising from it. Increasing the rate at which an mRNA is translated will increase the amount of protein product.

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.

Post-translational regulation: Any process which affects the amount of protein produced from a gene, and which occurs AFTER translation in the grand scheme of genetic expression. Actually, this is often just a buzz-word for regulation of the stability of the protein. The more stable a protein is, the more it will accumulate.

PPSearch: This tool allows you to search your query sequence for protein motifs. You can rapidly compare your query protein sequence against all patterns stored in the PROSITE pattern database. PROSITE is a database of protein families and domains. It is based on the observation that, while there is a huge number of different proteins, most of them can be grouped, on the basis of similarities in their sequences, into a limited number of families. Proteins or protein domains belonging to a particular family generally share functional attributes and are derived from a common ancestor. It is apparent, when studying protein sequence families, that some regions have been better conserved than others during evolution. These regions are generally important for the function of a protein and/or for the maintenance of its three- dimensional structure. By analysing the constant and variable properties of such groups of similar sequences, it is possible to derive a signature for a protein family or domain, which distinguishes its members from all other unrelated proteins. A pertinent analogy is the use of fingerprints by the police for identification purposes. A fingerprint is generally sufficient to identify a given individual. Similarly, a protein signature can be used to assign a newly sequenced protein to a specific family of proteins and thus to formulate hypotheses about its function. PROSITE currently contains patterns and profiles specific for more than a thousand protein families or domains. Each of these signatures comes with documentation providing background information on the structure and function of these proteins.

PQS server: The PQS server allows for searching of the list of likely quaternary structures generated at the EBI. The system is a SQL front end to a database on characteristics of the quaternary structure files.

Pratt: Pratt is a program that allows the user to efficiently search for patterns conserved in a set of protein sequences. It allows the user to define the class of patterns to be searched for, and then finds conserved patterns in this class. The time used by the program depends on the set of sequences, the class of patterns defined the minimum number of sequences a pattern is to match if an alignment or a query sequence is given, the greediness of the search.

Pred No: Pred No: The interest in any alignment depends on the strength of the sequence relationship it discloses. Any two random sequences are likely to contain some short region of similar residues, and the predicted number is calculated to indicate how often random or unrelated real sequences, like most of those in the database, are likely to show a score greater or equal to the alignment score.

Preg: An **EMBOSS** application. This searches for matches of a regular expression to a protein sequence. A regular expression is a way of specifying an ambiguous pattern to search for. Regular expressions are commonly used in some computer programming languages and may be more familiar to some users than to others.

preRNA: This is precursor RNA, an RNA transcript before it is processed into mRNA, rRNA, tRNA, or other cellular RNA species, any RNA species that is not yet the mature RNA product.

Prettyplot: An **EMBOSS** application. prettyplot reads in a set of aligned DNA or protein sequences. It displays them graphically, with conserved regions highlighted in various ways.

Prettyseq: An **EMBOSS** application. This writes out a nicely formatted display of the sequence with the translation (within specified ranges) displayed beneath it. The translated nucleic acid region will be shown in lower-case letters while the rest of the input sequence will be left in the input case. The base and residue numbers of the sequences are shown beside the sequences in the output. Slightly unusually, this application uses the codon usage tables to translate the codons.

Primary transcript: When a gene is transcribed in the nucleus, the initial product is the primary transcript, an RNA containing copies of all exons and introns. This primary transcript is then processed by the cell to remove the introns, to cleave off unwanted 3' sequence, and to polyadenylate the 5' end. The mature message thus formed is then exported to the cytoplasm for translation.

Primer: A small oligonucleotide (anywhere from 6 to 50 nt long) used to prime DNA synthesis. The DNA polymerases are only able to extend a pre-existing strand along a template; they are not able to take a naked single strand and produce a complementary copy of it de-novo. A primer which sticks to the template is therefore used to initiate the replication. Primers are necessary for DNA sequencing and PCR.

Primer extension: This is a method used to figure out how far upstream from a fixed site the start of an mRNA is. For example, perhaps you have isolated a cDNA clone, but you don't think that the clone has all of the 5' untranslated region. To find out how much is missing, you would first sequence the part you have, and figure out which strand is coding strand (usually the coding strand will have a large open reading frame). Next, you ask the

DNA Synthesis Facility to make an oligonucleotide complementary to the 5'-most region of the coding strand (and thus complementary to the mRNA). This "primer" is hybridised to mRNA (say, a mixture of mRNA containing the one in which you are interested), and reverse transcriptase is added to copy the mRNA from the primer out to the 5' end. The size of the resulting DNA fragment shows how far away from the 5' end your primer is.

Primersearch: An **EMBOSS** application. primersearch reads in primer pairs from an input file and searches them against sequence(s) specified by the user. Each of the primers in a pair is searched against the sequence and potential amplimers are reported. The user can specify a maximum percent mismatch level; for example, 10% mismatch on a primer of length 20bp means that the program will classify a primer as matching a sequence if 18 of the 20 base pairs matches. It will only report matches if both primers in the pair have a match in opposite orientations.

Prints: A fingerprint is a group of conserved motifs used to characterise a protein family. Prints is a compendium of such protein fingerprints. Fingerprints can encode protein folds and functionalities more flexibly and powerfully than can single motifs, their full diagnostic potency deriving from the mutual context afforded by motif neighbours.

Printsextract: An **EMBOSS** application. printsextract preprocesses the PRINTS database for use with the program PSCAN. This program derives matrix information from the final motif sets of the PRINTS data file (prints.dat). It creates files in the **EMBOSS** data subdirectory PRINTS these being a matrix file and files containing text information for each fingerprint. Running this program may be the job of your system manager.

Probe: A fragment of DNA or RNA which is labelled in some way (often incorporating 32P or 35S), and which is used to hybridise with the nucleic acid in which you are interested. For example, if you want to quantitate the levels of alpha subunit mRNA in a preparation of pituitary RNA, you might make a radiolabelled RNA in-vitro which is complementary to the mRNA, and then use it to probe a Northern blot of the pit RNA. A probe can be radiolabelled, or tagged with another functional group such as biotin. A probe can be cloned DNA, or might be a synthetic DNA strand. As an example of the latter, perhaps you have isolated a protein for which you wish to obtain a cDNA or genomic clone. You might (pay to) microsequence a portion of the protein, deduce the nucleic acid sequence, (pay to) synthesize an oligonucleotide carrying that sequence, radiolabel it and use it as a probe to screen a cDNA library or genomic library.

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. Each position in the profile contains scores for all of the possible amino acids, as well as one penalty score for opening and one for continuing a gap at the specified position. Attempts have been made to further improve the sensitivity of the profile by refining the procedures to construct a profile starting from a given multiple alignment. Other representations for sequence domains or motifs do not necessarily require the presence of a correct and complete multiple alignment, such as hidden Markov models.

Profit: An **EMBOSS** application. profit takes a simple frequency matrix produced by prophecy and searches with this to find matches in the input sequence(s) you are searching. Scores for the matches are calculated from the simple frequency matrix. It is the sum of scores at each position of the matrix. A 'simple frequency matrix' is simply a count of the number of times any particular amino acid occurs at each position in the alignment used to create it. Simple frequency matrices are created using the program prophecy with the option '-type F' to create the correct type of matrix. The alignment should not have gaps in it. The resulting matrix is moved to each position in the sequence(s) you are searching. At each position in the sequence, the frequencies of the amino acids or bases covered by the length of the matrix is read from the matrix. The sum of these frequencies at each position of the matrix is the score for that position of the sequence. If this score is above the threshold percentage of the maximum possible score for that matrix, then a hit is reported.

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

Promoter: The first few hundred nucleotides of DNA "upstream" (on the 5' side) of a gene, which control the transcription of that gene. The promoter is part of the 5' flanking DNA, i.e. it is not transcribed into RNA, but without the promoter, the gene is not functional. Note that the definition is a bit hazy as far as the size of the region encompassed, but the "promoter" of a gene starts with the nucleotide immediately upstream from the cap site, and includes binding sites for one or more transcription factors which can not work if moved farther away from the gene.

Prophecy: An **EMBOSS** application. This creates a profile matrix file from a nucleic acid or a protein sequence alignment. The profile matrix file can then be used by profit or prophet.

Prophet: An **EMBOSS** application. prophet finds matches between a GRIBSKOV or HENIKOFF profile produced by prophecy and one or more sequences. Note: prophet does NOT use the 'simple frequency matrices' produced by prophecy. If you have a 'simple frequency matrix'you should use the program profit to scan sequences.

Prosextract: An **EMBOSS** application. Takes the IDentity, ACcession number and motif PAttern line contents from prosite entries. Also converts the PAttern into a regular expression and writes these four pieces to an output file - defaulted to be called 'prosite.lines'

PROSITE Database: **PROSITE** is a database of protein families and domains. It consists of biologically significant sites, patterns and profiles that help to reliably identify to which known protein family (if any) a new sequence belongs.

Protein: Proteins are macromolecules made up from 20 different amino acids, also referred to as residues.

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.

protein product: The protein molecule assembled under the direction of a gene

Protein Quaternary Structure Query: The PQS server allows for searching of the list of likely quaternary structures generated at the EBI. The system is a SQL front end to a database on characteristics of the quaternary structure files.

Proteome analysis: Proteome analysis provides comprehensive statistical and comparative proteome analyses of the predicted proteomes of all fully sequenced organisms present in the **SWISS-PROT** and **TrEMBL** databases

Proteomes Database: A comprehensive statistical and comparative analyses of the predicted proteomes of fully sequenced organisms.

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)

Proto-oncogene: A gene present in a normal cell which carries out a normal cellular function, but which can become an oncogene under certain circumstances. The prefix "c-" indicates a cellular gene, and is generally used for proto-oncogenes (examples: c-myb , c-myc , c-fos , c-jun , etc).

Pscan: An **EMBOSS** application. PRINTS is a database of diagnostic protein signatures, or fingerprints. Fingerprints are groups of conserved motifs or elements that together form a diagnostic signature for particular protein families. An uncharacterised sequence matching all motifs or elements can then be readily diagnosed as a true match to a particular family fingerprint. They can be used to diagnose family relationships in newly-determined sequences (especially from genome projects). Usually the motifs or elements do not overlap, but are separated along a sequence, though they may be contiguous in 3D-space. Fingerprints can encode protein folds and functionalities more flexibly and powerfully than can single motifs, full diagnostic potency deriving from the mutual context provided by motif neighbours. Diagnostically, this is more powerful than using single motifs by virtue of the biological context afforded by matching motif neighbours.

Pulsed field gel electrophoresis: A gel technique which allows size-separation of very large fragments of DNA, in the range of hundreds of kb to thousands of kb. As in other gel electrophoresis techniques, populations of molecules migrate through the gel at a speed related to their size, producing discrete bands. In normal electrophoresis, DNA fragments greater than a certain size limit all migrate at the same rate through the gel. In PFGE, the electrophoretic voltage is applied alternately along two perpendicular axes, which forces even the larger DNA fragments to separate by size.

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

R

Radar: Many large proteins have evolved by internal duplication and many internal sequence repeats correspond to functional and structural units. We have developed an automatic algorithm, **RADAR**, for segmenting a query sequence into repeats. The segmentation procedure has three steps: (i) repeat length is determined by the spacing between suboptimal self-alignment traces. (ii) repeat borders are optimised to yield a maximal integer number of repeats. (iii) distant repeats are validated by iterative profile alignment. The method identifies short composition biased as well as gapped approximate repeats and complex repeat architectures involving many different types of repeats in the query sequence. No manual intervention and no prior assumptions on the number and length of repeats are required. Comparison to the Pfam-A database indicates good coverage, accurate alignments, and reasonable repeat borders. Screening the Swissprot database revealed 3,000 repeats not annotated in existing domain databases. A number of these repeats had been described in the literature but most were novel. This illustrates how in times when curated databases grapple with ever increasing backlogs, automatic (re)analysis of sequences provides an efficient way to capture this important information.

Random primed synthesis: If you have a DNA clone and you want to produce radioactive copies of it, one way is to denature it (separate the strands), then hybridise to that template a mixture of all possible 6-mer oligonucleotides. Those oligos will act as primers for the synthesis of labelled strands by DNA polymerase (in the presence of radiolabelled precursors).

Rank Function: Rank Function: A problem with the standard scoring of alignments is that the score alone does not give a measure of the improbability of the alignment. Intuitively, it is the most improbable alignments that may be the most helpful in pointing to possible structural or functional properties of these sequence domains.The ranking function is obtained by modelling the behaviour of the shorter sequence (in a comparison) in matching properties, and taking the length ratio into account. It is expressed as the number of results expected or above the observed score for each alignment.

Raw Format: Like text/plain format except that it removes any white space or digits, accepts only alphabetic characters and rejects anything else. This means that it is safer to use this format than plain format. If you have digits and spaces or TAB characters, these are removed and ignored. If you have other non-alphabetic characters (for example, punctuation characters), then the sequence will be rejected as erroneous.

Reading Frames: Once the RNA has been transcribed, it travels from the DNA template to the ribosome on the endoplasmic reticulum to be translated for protein synthesis. Each

3 bases in the RNA sequence codes for 1 amino acid. As you may not be sure what position to start at when predicting what protein sequence may be produced by this code, you could start with one of 3 positions from either end of the RNA sequence. Thus there are 6 possible predicted protein sequences resulting from such a peice of code. These are known as the 6 possible **reading frames**. There are 3 forward frames and 3 reverse sense frames.

Rebaseextract: An **EMBOSS** application. The Restriction Enzyme database (REBASE) is a collection of information about restriction enzymes and related proteins. It contains published and unpublished references, recognition and cleavage sites, isoschizomers, commercial availability, methylation sensitivity, crystal and sequence data. DNA methyltransferases, homing endonucleases, nicking enzymes, specificity subunits and control proteins are also included. Most recently, putative DNA methyltransferases and restriction enzymes, as predicted from analysis of genomic sequences, are also listed. The home page of REBASE is: http://rebase.neb.com/ This program derives recognition site and cleavage information from the "withrefm" file of an REBASE distribution. It creates three files in the **EMBOSS** data subdirectory REBASE. A pattern file, a reference file and a supplier file. The **EMBOSS** programs that find restriction cutting sites use the data files produced by this program and will not work without them.

Recessive: Any trait that is expressed phenotypically only when present on both alleles of a gene

Recessive allele: A gene that is expressed only when its counterpart allele on the matching chromosome is also recessive (not dominant). Autosomal recessive disorders develop in persons who receive two copies of the mutant gene, one from each parent who is a carrier

Recoder: An **EMBOSS** application. recoder scans a given nucleotide sequence for restriction sites. It reports single base positions in the restriction pattern which when mutated remove the restriction site whilst maintaining the same translation (in frame 1 of the input sequence). Several restriction enzymes can be specified or alternatively all the enzymes in the REBASE database can be investigated. To find out whether the single point mutations found by 'recoder', introduce new restriction sites, 'silent' should be run on the original sequence. ('Silent' searches for silent point mutation sites which maintain the same translation. The output for 'recoder' is similar to the format used by 'silent'.

Recombinant DNA: 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).

Redata: An **EMBOSS** application. The Restriction Enzyme database (REBASE) is a collection of information about restriction enzymes and related proteins. It contains published and unpublished references, recognition and cleavage sites, isoschizomers, commercial availability, methylation sensitivity, crystal and sequence data. DNA methyltransferases, homing endonucleases, nicking enzymes, specificity subunits and control proteins are also included. Most recently, putative DNA methyltransferases and restriction enzymes, as predicted from analysis of genomic sequences, are also listed. The home page of REBASE is: http://rebase.neb.com/ This program searches the REBASE database for information on a specified restriction enzyme. It outputs a report including the cut site, isoschizomers, references and commercial suppliers of the enzyme.

Reference Author line (EMBL): The RA (Reference Author) lines list the authors of the paper (or other work) cited.

Reference Author line (Swissprot): The RA (Reference Author) lines list the authors of the paper (or other work) cited.

Reference Comment line (EMBL): The RC (Reference Comment) line type is an optional line type which appears if the reference has a comment.

Reference Comment line (SWISS-PROT): The RC (Reference Comment) lines are optional lines which are used to store comments relevant to the reference cited.

Reference Cross-reference line (EMBL): The RX (Reference Cross-reference) line type is an optional line type which contains a cross-reference to an external citation or abstract database.

Reference Cross-Reference line (SWISS-PROT): The RX (Reference Cross-Reference) line is an optional line which is used to indicate the identifier assigned to a specific reference in a bibliographic database.

Reference Location line: The RL (Reference Location) lines contain the conventional citation information for the reference.

Reference Location line (EMBL): The RL (Reference Location) line contains the conventional citation information for the reference.

Reference Number line: The RN (Reference Number) line gives a sequential number to each reference citation in an entry.

Reference Number line (EMBL): The RN (Reference Number) line gives a unique number to each reference citation within an entry.

Reference Position line (EMBL): The RP (Reference Position) line type is an optional line type which appears if one or more contiguous base spans of the presented sequence can be attributed to the reference in question.

Reference Position line (SWISS-PROT): The RP (Reference Position) line describes the extent of the work carried out by the authors of the reference cited.

Reference Title line (EMBL): The RT (Reference Title) lines give the title of the paper (or other work).

Reference Title lines (SWISS-PROT): The RT (Reference Title) lines give the title of the paper (or other work) cited

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).

Remap: An **EMBOSS** application. The Restriction Enzyme database (REBASE) is a collection of information about restriction enzymes and related proteins. It contains published and unpublished references, recognition and cleavage sites, isoschizomers, commercial availability, methylation sensitivity, crystal and sequence data. DNA methyltransferases, homing endonucleases, nicking enzymes, specificity subunits and control proteins are also included. Most recently, putative DNA methyltransferases and restriction enzymes, as predicted from analysis of genomic sequences, are also listed. The home page of REBASE is: http://rebase.neb.com/ This program uses REBASE data to find the recognition sites and/or cut sites of restriction enzymes in a nucleic acid sequence. This program displays the cut sites on both strands by default. It will optionally also display the translation of the sequence. There are many options to change the style of display to aid in making clear presentations. One potentially very useful option is '-flatreformat' that displays not only the cut sites which many other restriction cut-site programs will show, but also shows the recognition site.

Repeats: Repeat sequences and approximate repeats occur throughout the DNA of higher organisms (mammals). For example, the Alu sequences of length about 300 characters, appear hundreds of thousands of times in Human DNA with about 87% homology to a consensus Alu string. Some short substrings such as TATA-boxes, poly-A and (TG)* also appear more often than by chance. Repeat sequences may also occur within genes, as mutations or alterations to those genes. Repetitive sequences, especially mobile elements, have many applications in genetic research. DNA transposons and retroposons are routinely used for insertional mutagenesis, gene mapping, gene tagging, and gene transfer in several model systems.

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.

Restover: An **EMBOSS** application. The Restriction Enzyme database (REBASE) is a collection of information about restriction enzymes and related proteins. It contains published and unpublished references, recognition and cleavage sites, isoschizomers, commercial availability, methylation sensitivity, crystal and sequence data. DNA methyltransferases, homing endonucleases, nicking enzymes, specificity subunits and control proteins are also included. Most recently, putative DNA methyltransferases and restriction enzymes, as predicted from analysis of genomic sequences, are also listed. The home page of REBASE is: http://rebase.neb.com/ restover takes a specified sequence and a short sequence of a cut-site overhang and searches the REBASE database for matching enzymes that create the desired overhang sequence when they cut the input sequence.

Restriction enzyme: A class of enzymes ("restriction endonucleases") generally isolated from bacteria, which are able to recognise and cut specific sequences ("restriction sites") in DNA.

Restriction fragment: The piece of DNA released after restriction digestion of plasmids or genomic DNA. See "Restriction enzyme". One can digest a plasmid and isolate one particular restriction fragment (actually a set of identical fragments). The term also describes the fragments detected on a genomic blot which carry the gene of interest.

Restriction map: A "cartoon" depiction of the locations within a stretch of known DNA where restriction enzymes will cut.

Reverse transcriptase: An enzyme which will make a DNA copy of an RNA template - a DNA-dependant RNA polymerase. RT is used to make cDNA; one begins by isolating polyadenylated mRNA, providing oligo-dT as a primer, and adding nucleotide triphosphates and RT to copy the RNA into cDNA.

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

Reverse Translator: Reverse Translator has been created to help the calculation of the probable DNA level point mutations underlying the reported amino acid substitution.

Revseq: An **EMBOSS** application. This takes a sequence and outputs the reverse complement (also known as the anti-sense or reverse sense) sequence. It can also output just the reversed sequence or just the complement of the sequence.

Ribonucleic acid: 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).

Riboprobe: A strand of RNA synthesised in-vitro (usually radiolabelled) and used as a probe for hybridisation reactions. An RNA probe can be synthesised at very high specific activity, is single stranded (and therefore will not self anneal), and can be used for very sensitive detection of DNA or RNA.

Ribosome: A cellular particle which is involved in the translation of mRNAs to make proteins. Ribosomes are a complex consisting of ribosomal RNAs (rRNA) and several proteins.

RNA: Ribonucleic acid. 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).

RNase: Ribonuclease; an enzyme which degrades RNA. It is ubiquitous in living organisms and is exceptionally stable. The prevention of RNase activity is the primary problem in handling RNA

RNase protection assay: This is a sensitive method to determine (1) the amount of a specific mRNA present in a complex mixture of mRNA and/or (2) the sizes of exons which comprise the mRNA of interest. A radioactive DNA or RNA probe (in excess) is allowed to hybridise with a sample of mRNA (for example, total mRNA isolated from tissue), after which the mixture is digested with single-strand specific nuclease. Only the probe which is hybridised to the specific mRNA will escape the nuclease treatment, and can be detected on a gel. The amount of radioactivity which was protected from nuclease is proportional to the amount of mRNA to which it hybridised. If the probe included both intron and exons, only the exons will be protected from nuclease and their sizes can be ascertained on the gel.

Robinson group: The **Robinson group** is investigating & advising on the e-Science & Grid technology requirements of the EMBL-EBI, through application development plus participation in standards development.

rRNA: This is ribosomal RNA, it is a component of the ribosomes, the protein synthetic factories in the cell.

RSF Format: **RSF** means rich sequence format and it is created by the Editor in SeqLab. The format is recognised by the word !!RICH_SEQUENCE at the beginning of the file. It contains one or more sequences that may or may not be related. In addition to the sequence data, each sequence can be annotated with descriptive sequence information such as: Creator/author of the sequence, Sequence weight Creation date, One-line description of the sequence, Offset, or the number of leading gaps in a sequence that is part of an alignment or fragment assembly project Known sequence features.

S

SAPS: **SAPS** evaluates by statistical criteria a wide variety of protein sequence properties. Properties considered include compositional biases; clusters and runs of charge and other amino acid types; different kinds and extents of repetitive structures; locally periodic motifs; and anomalous spacings between identical residue types. The statistics are computed for any single (or appropriately concatenated) protein sequence input. Statistically significant sequence features highlighted by SAPS in the input sequence may suggest promising regions for experimental investigation. The program also finds application in the description of conserved features of families of proteins as well as in the inverse problem of deriving protein groupings based upon sequence features. Short sequences are subject to larger statistical fluctuations than longer sequences. The statistical evaluations of SAPS are reliable only for sequences of at least about 200 residues. Shorter sequences may in some cases be appropriately concatenated and analysed as a representative combined sequence (e.g., histones, or Ras family proteins).

SARS: Severe acute respiratory syndrome (SARS) is a respiratory illness that has recently been reported in Asia, North America, and Europe. This fact sheet provides basic information about the disease and what is being done to combat its spread. To find out more about SARS, go to www.cdc.gov/ncidod/sars/ and www.who.int/csr/sars/en/ .

ScanPS: **SCANPS** (SCAN Protein Sequence) is a program written to perform algorithms related to the Smith-Waterman local similarity search. It runs on variety of conventional hardware, and can be used to perform sequence database searches using full dynamic programming. Features are: Full Smith-Waterman style searching with a single sequence. Multiple domain matches found against each database sequence. Iterative profile searching similar in concept to PSI-BLAST, but with full dynamic programming on each cycle for additional sensitivity. Significance of matches calculated ``on the fly'' for each search. Efficient implementation on Intel CPUs by using MMX and SSE instructions. Output of each search as pairwise alignments and multiple alignments.

ScanRegExp: Tool that scans against the regular expressions in PROSITE

Screening: To screen a library is to select and isolate individual clones out of the mixture of clones. For example, if you needed a cDNA clone of the pituitary glycoprotein hormone alpha subunit, you would need to make (or buy) a pituitary cDNA library, then screen that library in order to detect and isolate those few bacteria carrying alpha subunit cDNA.

Secondary structure: The organisation of the peptide backbone of a protein that occurs as a result of hydrogen bonds e.g alpha helix, Beta pleated sheet.

Seealso: An **EMBOSS** application. This program takes the name of an existing program in **EMBOSS** (or a program in one the associated EMBASSY packages) and gives a list of the programs which share some functionality with it. It does this by noting the functional groups that the program belongs to and reporting any programs which share those functional groups. The functional groups of a program are set in the ACD file (this is the part of the program in the **EMBOSS** system which specifies the required parameters, some help on the parameters, the one-line description of the program, etc.) Normally the names of the groups are fairly specific, such as: "NUCLEIC COMPOSITION", resulting in the reporting of only a tightly restricted set of other programs also dealing with the analysis of "NUCLEIC COMPOSITION". The '-explode' qualifier will increase the number of groups that the program belongs to by splitting the group name at selected points to produce such groups as: "NUCLEIC", "NUCLEIC COMPOSITION" and "COMPOSITION". All programs with the exploded group names which also include "NUCLEIC" and "COMPOSITION" will now be reported. The result of this qualifier is thus a report of a larger number of programs with a more tenuous link to the specified program. In other words, use the '-explode' qualifier to decrease the specificity of the search. The groups that the program belongs to can be output by using the '-groups' qualifier.

Selectivity: Selectivity of bioinformatics similarity search algorithms is defined as the significance threshold for reporting database sequence matches. As an example, for **BLAST** searches, the parameter E is interpreted as the upper bound on the expected frequency of chance occurrence of a match within the context of the entire database search. E may be thought of as the number of matches one expects to observe by chance alone during the database search.

Sense strand: A gene has two strands: the sense strand and the anti-sense strand. The Sense strand is, by definition, the same 'sense' as the mRNA; that is it can be translated exactly as the mRNA sequence can.

Sensitivity: Sensitivity of bioinformatics similarity search algorithms centres around two areas: First, how well can the method detect biologically meaningful relationships between two related sequences in the presence of mutations and sequencing errors; secondly how does the heuristic nature of the algorithm affect the probability that a matching sequence will not be detected. At the user's discretion, the speed of most similarity search programs can be sacrificed in exchange for greater sensitivity - with an emphasis on detecting lower scoring matches.

SeqDB: The **Sequence Database Group** is an amalgamation of activities related to the production of protein sequence, protein family and nucleotide sequence databases at the EBI.

Seqmatchall: An **EMBOSS** application. This takes a set of sequences and does an all-against-all pairwise comparison of words (fragments of the sequences of a specified fixed size) in the sequences, finding regions of identity between any two sequences. The larger the specified word size, the faster the comparison will proceed. Regions whose stretches of identity are shorter than the word size will be missed. You should therefore choose a word size that is small enough to find those regions of similarity you are interested in within a reasonable time-frame.

Seqret: An **EMBOSS** application. The simplicity of the above description of this program greatly understates the rich functionality of this program. Because **EMBOSS** programs can take a wide range of qualifiers that slightly change the behaviour of the program when reading or writing a sequence, this program can do many more things than simply "read and write a sequence". seqret can read a sequence or many sequences from databases, files, files of sequence names, the command-line or the output of other programs and then can write them to files, the screen or pass them to other programs. Because it can read in a sequence from a database and write it to a file, seqret is a program for extracting sequences from databases. Because it can write the sequence to the screen, seqret is a program for displaying sequences. seqret can read sequences in any of a wide range of standard sequence formats. You can specify the input and output formats being used. If you don't specify the input format, seqret will try a set of possible formats until it reads it in successfully. Because you can specify the output sequence format, seqret is a program to reformat a sequence. seqret can read in the reverse complement of a nucleic acid sequence. It therefore is a program for producing the reverse complement of a sequence. seqret can read in a sequence whose begin and end positions you have specified and write out that fragment. It is therefore a utility for doing simple extraction of a region of a sequence. seqret can change the case of the sequence being read in to upper or to lower case. It is therefore a simple sequence beautification utility. seqret can do any combination of the above functions. The sequence input and output specification of this (and many other **EMBOSS** programs) is described as being a Uniform Sequence Address. The Uniform Sequence Address, or USA, is a somewhat tongue-in-cheek reference to a URL-style sequence naming used by all **EMBOSS** applications. The USA is a very flexible way of specifying one or more sequences from a variety of sources and includes sequence files, database queries and external applications.

Seqretsplit: An **EMBOSS** application. seqretsplit is exactly the same as the program seqret except that when it writes out more than one sequence, it writes each sequence to an individual file. Its main use is therefore to split a file containing multiple sequences into many files, each containing one sequence.

Sequence: As a noun, the sequence of a DNA is a buzz word for the structure of a DNA molecule, in terms of the sequence of bases it contains. As a verb, "to sequence" is to determine the structure of a piece of DNA; i.e. the sequence of nucleotides it contains.

sequence data line (SWISS-PROT): The sequence data line has a line code consisting of two blanks rather than the two-letter codes used until now. The sequence counts 60 amino acids per line, in groups of 10 amino acids, beginning in position 6 of the line.

sequence data lines (EMBL): The sequence data lines have lines of code starting with two blanks. The sequence is written 60 bases per line, in groups of 10 bases separated by a blank character, beginning in position 6 of the line. The direction listed is always 5' to 3'

Sequence Database Group: The Sequence Database Group is an amalgamation of activities related to the production of protein sequence, protein family and nucleotide sequence databases at the EBI.

SeQuence header line (EMBL): The SQ (SeQuence header) line marks the beginning of the sequence data and gives a summary of its content.

SeQuence header line (SWISS-PROT): The SQ (SeQuence header) line marks the beginning of the sequence data and gives a quick summary of its content.

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

Sequence Version line (EMBL): The SV (Sequence Version) line contains the new format of the nucleotide sequence identifier.

Sex chromosomes: The chromosomes that determine the sex of an organism. Human females have two X chromosomes; males have one X and one Y.

Shotgun cloning: The practice of randomly clipping a larger DNA fragment into various smaller pieces, cloning everything, and then studying the resulting individual clones to figure out what happened. For example, if one was studying a 50 kb gene, it "may" be a bit difficult to figure out the restriction map. By randomly breaking it into smaller fragments and mapping those, a master restriction map could be deduced. See also Shotgun sequencing.

Shotgun sequencing: A way of determining the sequence of a large DNA fragment which requires little brainpower but lots of late nights. The large fragment is shotgun cloned (see above), and then each of the resulting smaller clones ("subclones") is sequenced. By finding out where the subclones overlap, the sequence of the larger piece becomes apparent. Note that some of the regions will get sequenced several times just by chance.

Showalign: An **EMBOSS** application. showalign displays an aligned set of protein or a nucleic acid sequences in a style suitable for publication. The output is sent to the screen by default for the user to view, but it can write the results to a file. The output highlights various differences or similarities between each of the sequences and a reference sequence by setting selected types of matches to a reference sequence to be '.' characters. The reference sequence can be either the calculated consensus sequence (the default) or it can be one of the set of aligned sequences, specified by either the ordinal number of that sequence in the input file, or by its name. The output sequences can be displayed in either the input order (the default) or they can be sorted in order of their similarity to the reference sequence or sorted alphabetically by their names.

Showdb: An **EMBOSS** application. This writes out a simple table displaying the names, contents and available ways of accessing the sequence databases. The available ways of accessing the databases are 'ID', 'Query' and 'All'. These refer to the way that you can search the databases to get entries from them, which is governed by the ways the database has been set up and the way it is organised and indexed. Different databases may have different access capabilities, depending on how your local site is organised. **EMBOSS** has been designed to be extremely flexible in its use of sequence databases formats, so that it is easy to set **EMBOSS** up to use your site's existing databases. Sometimes this means that it is hard to extract entries from some databases in particular ways. For example, a flat file database with no index is only useful for reading all entries, while a database located in another site that is available via the WWW may only provide single entries.

Showfeat: An **EMBOSS** application. Showfeat reads a protein or nucleic sequence and its feature table, and writes a text representation of the features to standard output.

Showorf: An **EMBOSS** application. Showorf displays a nucleic acid sequence with its protein translation in a style suitable for publication. The translation can be done in any frame or combination of frames. It uses codon frequency files to do the translation. You can specify the codon frequency file that you use with the '-cfile' option. The default table is 'Ehum.cut'.

Showseq: An **EMBOSS** application. showseq displays a protein or a nucleic acid sequence in a style suitable for publication. The output is sent to the screen by default for the user to view, but it can write the results to a file. The display style of the output is very flexible. The user can select a style from the pre-set choice of formats, or can design a style to suit their purposes and aesthetic tastes.

Shuffleseq: An **EMBOSS** application. This takes a sequence as input and outputs one or more sequences whose order has been randomly shuffled. No bases or residues are changed, only their order. The number of shuffled sequences output can be set by the '-shuffle' qualifier.

Sickle-cell anaemia: An inherited, potentially lethal disease in which a defect in hemoglobin, the oxygen-carrying pigment in the blood, causes distortion (sickling) and loss of red blood cells, producing damage to organs throughout the body.

Sigcleave: An **EMBOSS** application. sigcleave predicts the site of cleavage between a signal sequence and the mature exported protein. The predictive accuracy is estimated to be around 75-80% for both prokaryotic and eukaryotic proteins.

Silent: An **EMBOSS** application. silent does a scan of a nucleic acid sequence for silent mutation restriction enzyme sites. silent finds positions in a sequence where a point mutation could be made to introduce a specified restriction enzyme recognition site without changing the translation.

Similarity 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. Statistical methods are less successful but more general and are based on the derivation of structural preference values for single residues, pairs of residues, short oligopeptides or short sequence patterns. The transfer of structure/function information to a potentially homologous protein is straightforward when the sequence similarity is high and extended in length, but the assessment of the structural significance of sequence similarity can be difficult when sequence similarity is weak or restricted to a short region.

Single nucleotide polymorphisms: Variations of single base pairs scattered throughout the human genome that serve as measures of the genetic diversity in humans. About 1 million SNPs are estimated to be present in the human genome, and SNPs are useful markers for gene mapping studies.

Sirna: An **EMBOSS** application. Description: RNA interference, or RNAi, is a phenomenon in which double stranded RNA (dsRNA) effects silencing of the expression of genes that are highly homologous to either of the RNA strands in the duplex. Gene silencing in RNAi results from the degradation of mRNA sequences, and the effect has been used to determine the function of many genes in Drosophilia, C. elegans, and many plant species. The duration of knockdown by siRNA can typically last for 7-10 days, and has been shown to transfer to daughter cells. Of further note, siRNAs are effective at quantities much lower than alternative gene silencing methodologies, including antisense and ribozyme based strategies.

Sixpack: An **EMBOSS** application. sixpack takes a nucleic acid sequence and writes out the forward and reverse senses of the sequence with the 3 forward and three reverse translations in a pretty display format. It also writes a file containing the open reading frames that are larger than the specified minimum size (default 1 base, showing all possible open reading frames). These open reading frames are written as protein sequences in the default output sequence format. An open reading frame is defined in this program as any possible translation between two STOP codons.

Skipseq: An **EMBOSS** application. skipseq skips the first few sequences in a multiple set of sequences, and writes out the rest of them. skipseq is a variant of the standard program for reading and writing sequences, seqret. seqret has an option to allow it to only read the first sequence from a multiple set of sequences (-firstonly). seqret cannot, however, skip the first few sequences from a multiple set of sequence, writing out the rest; this is what skipseq is for. In all other respects, skipseq is the same as seqret.

Slot blot: Similar to a dot blot, but the analyte is put onto the membrane using a slot-shaped template. The template produces a consistently shaped spot, thus decreasing errors and improving the accuracy of the analysis. See Dot blot

SNP-Fasta33: **SNP-Fasta33**. There are several types of DNA sequence variation, including insertions and deletions, copy number differences of repeated sequences, and single base pair differences. Single base pair differences are the most frequent case of sequence variation and when the variant sequence type has a frequency of at least 1% in the population, they are termed single nucleotide polymorphisms (SNPs). SNPs have many properties that make them useful for the study of sequence variation, they have high frequency and are stable, having much lower mutation rates than do repeat sequences. SNPs are useful polymorphic markers for mapping and discovering the genes associated with common diseases, or genes related to drug responsiveness.

snRNA: This is small nuclear RNA and refers to a number of small RNA molecules found in the nucleus. These RNA molecules are important for several processes including RNA splicing and maintenance of the telomeres, or chromosome ends. They are always found associated with specific proteins and the complexes are referred to as small nuclear ribonucleoproteins (SNRNP).

Solution hybridization: A method closely related to RNase protection (see "RNase protection assay"). Solution hybridisation is designed to measure the levels of a specific mRNA species in a complex population of RNA. An excess of radioactive probe is allowed to hybridise to the RNA, then single-strand specific nuclease is used to destroy the remaining unhybridised probe and RNA. The "protected" probe is separated from the degraded fragments, and the amount of radioactivity in it is proportional to the amount of mRNA in the sample which was capable of hybridisation. This can be a very sensitive detection method.

Somatic cells: All body cells except the reproductive cells

Somatic mutation: Gene changes that arise within individual cells and accumulate throughout a person's lifetime; also called acquired mutations.

Southern blot: A technique for analysing mixtures of DNA, whereby the presence and rough size of one particular fragment of DNA can be ascertained. See "Blotting". Named after its inventor, Dr E. M. Southern.

SP-ML: The **SWISS-PROT** and **TrEMBL** protein sequence databases in **XML format**. It provides the users with an easily parsable view on the rich data in these two databases.

Splice form: By using alternative splicing, a single message precursor from DNA can generate an entire family of mRNAs and proteins. This can be utilised to create specificity in cell-cell or cell-ligand interactions. A cell may produce a given protein, but it will be a different splice-form of the protein than that produced by an adjacent cell. In this manner, the two cells have the potential to interact differently with other cells or molecules. Two places where this has been extremely important is in the production of cell-surface specificity proteins in the immune and nervous systems.

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.

Splitter: An **EMBOSS** application. This simple editing program allows you to split a long sequence into smaller, optionally overlapping, subsequences. There should be little requirement to split sequences into smaller sub-sequences in EMBOSS, but there may be circumstances where memory usage becomes restrictive when dealing with truly large sequences. In this case, memory usage may be reduced by repeating the analysis several times on split sub-sequences. If you need to split a large sequence into smaller subsequences so that a non- **EMBOSS** program can analyse the smaller sequence, it may also be useful to write the sub-sequences into separate files instead of the default

EMBOSS behaviour of concatenating them together into one file. To write the output sequences to separate files, use the command-line switch '-ossingle'.

SPTR: A comprehensive non-redundant protein sequence database that combines the high quality of annotation in **SWISS-PROT** with the completeness of the weekly updated translation of all protein coding sequences from the **EMBL** Nucleotide Sequence Database.

SRS: Sequence Retrieval System (SRS). A datawarehouse developed at the EMBL/EBI by Dr. Thure Etzold. IIt is now owned and developed by Lion Bioscience AG. This system for querying and retrieving data from a wide variety of bioinformatic databases can be found at the EBI: **http://srs.ebi.ac.uk/**

Stable transfection: A form of transfection experiment designed to produce permanent lines of cultured cells with a new gene inserted into their genome. Usually this is done by linking the desired gene with a "selectable" gene, i.e. a gene which confers resistance to a toxin (like G418, aka Geneticin). Upon putting the toxin into the culture medium, only those cells which incorporate the resistance gene will survive, and essentially all of those will also have incorporated the experimenter's gene.

Sticky ends: After digestion of a DNA with certain restriction enzymes, the ends left have one strand overhanging the other to form a short (typically 4 nt) single-stranded segment. This overhang will easily re-attach to other ends like it, and are thus known as "sticky ends". For example, the enzyme BamHI recognizes the sequence GGATCC, and clips after the first G in each strand: The overhangs thus produced can still hybridise ("anneal") with each other, even if they came from different parent DNA molecules, and the enzyme ligase will then covalently link the strands. Sticky ends therefore facilitate the ligation of diverse segments of DNA, and allow the formation of novel DNA constructs.

Stretcher: An **EMBOSS** application. stretcher calculates a global alignment of two sequences using a modification of the classic dynamic programming algorithm which uses linear space. A global pairwise alignment is one where it is assumed that the two sequences have diverged from a common ancestor and that the program should try to stretch the two sequences, introducing gaps where necessary, in order to show the alignment over the whole length of the two sequences that best illustrates their similarities. In contrast, a local alignment program like matcher simply finds local, small parts of the two sequences where there is some similarity and makes no assumption about the whole length of the sequence needing to be similar. The standard sequence global alignment program using the Needleman & Wunsch algorithm, as implemented in the program needle, requires O(MN) space and O(N) time. This is standard computer-science language for it needing an amount of computer memory that is proportional to the product of the two sequences being aligned and taking an amount of time that is proportional to the shorter of the two sequences. So if a 1 kb and a 10 kb sequence take 10 Mega-words of memory and 10 minutes to align, you should expect that in order to align a 10 kb sequence and a 1 Mb sequence you will need appoximately 10 Giga-words of memory and 100 minutes. Computer memory will rapidly be exhausted as the size of the aligned sequences increases. This program implements the Myers and Miller algorithm for finding an optimal global alignment in an amount of computer memory that is only proportional to the size of the smaller sequence - O(N). In computing, a benefit is seldom gained without a cost elsewhere. The cost of gaining a memory-efficient alignment is that it takes about twice the amount of time to do the alignment as the Needleman & Wunsch algorithm. In computer-science language the time is approximately O(2N).

Stringency: A term used to describe the conditions of hybridisation. By varying the conditions (especially salt concentration and temperature) a given probe sequence may be allowed to hybridise only with its exact complement (high stringency), or with any somewhat related sequences (relaxed or low stringency). Increasing the temperature or decreasing the salt concentration will tend to increase the selectivity of a hybridisation reaction, and thus will raise the stringency.

Structure prediction: Algorithms that predict the secondary, tertiary and sometimes even quarternary structure of proteins from their sequences. Determining protein structure from sequence has been dubbed "the second half of the Genetic Code" since it is the folded tertiary structure of a protein that governs how it functions as a gene product. As yet most structure prediction methods are only partially successful, and typically work best for certain well-defined classes of proteins.

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

Stssearch: An **EMBOSS** application. stssearch searches a DNA sequence database with a set of STS primers and reports expected matches. stssearchs reads in one or more sequences to be searched. For each pair of primers, it looks for matches between the primers and the query sequence in either orientation. Any matches found will be reported. Only one primer need match for it to be reported.

Sub-cloning: If you have a cloned piece of DNA (say, inserted into a plasmid) and you need unlimited copies of only a part of it, you might "sub-clone" it. This involves starting with several million copies of the original plasmid, cutting with restriction enzymes, and purifying the desired fragment out of the mixture. That fragment can then be inserted into a new plasmid for replication. It has now been subcloned.

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.

Sump statistics: Tends to rank database matches in a more intuitive order than Poisson statistics. Here the statistical significance ascribed to a set of HSPs may be higher than that ascribed to any individual member of the set. Only when the ascribed significance satisfies the user-selectable (or default) expected threshold will the match be reported to the user.

Supermatcher: An **EMBOSS** application. This is a rough and ready local alignment program for large sequences. The reason it is rough and ready is that wordmatch is used to find all the word matches between the first sequence and another sequence. Then by calculating the highest score for a diagonal we can then use this as the centre point for a Smith-Waterman type calculation of a width given by the user. So a narrow diagonal Smith-Waterman is calculated hence the results will be rough but due to the space saving much larger sequences can be aligned.

SVA: Sequence Version Archive at the EBI. This provides access to all public sequence records that ever existed in the **EMBL** Nucleotide Sequence Database. The SVA can be accessed at **http://www.ebi.ac.uk/embl/sva/**.

SWALL: A comprehensive non-redundant protein sequence database that combines the high quality of annotation in **SWISS-PROT** with the completeness of the weekly updated translation of all protein coding sequences from the **EMBL** Nucleotide Sequence Database.

SWISS-NEW: **Swiss-Prot** Updates

SWISS-PROT Database: Swiss-Prot Protein Database, a curated protein sequence database that provides a high level of annotation, a minimal level of redundancy and high level of integration with other databases.

Swiss-Prot Format: **SWISS-PROT** is an annotated protein sequence database. The **SWISS-PROT** protein knowledgebase consists of sequence entries. Sequence entries are composed of different line types, each with their own format. For standardisation purposes the format of **SWISS-PROT** follows as closely as possible that of the **EMBL** Nucleotide Sequence Database. The **Swiss-Prot** user manual is available here. The entries in the **SWISS-PROT** database are structured so as to be usable by human readers as well as by computer programs. The explanations, descriptions, classifications and other comments are in ordinary English. Wherever possible, symbols familiar to biochemists, protein chemists and molecular biologists are used. Each sequence entry is composed of lines.

Different types of lines, each with their own format, are used to record the various data that make up the entry.

Syco: An **EMBOSS** application. syco is a frame-specific gene finder that tries to recognize protein coding sequences by virtue of the similarity of their codon usage to a codon frequency table. syco finds regions of each forward reading frame of a nucleic acid sequence that show strong codon preference. syco is useful for locating protein coding regions, determining their reading frames, estimating the level of expression of a gene, and locating nucleic acid sequencing errors. It is essential to use the correct codon usage file for the species.

T

Taq polymerase: A DNA polymerase isolated from the bacterium Thermophilis aquaticus and which is very stable to high temperatures. It is used in PCR procedures and high temperature sequencing.

TATA box: A sequence found in the promoter (part of the 5' flanking region) of many genes. Deletion of this site (the binding site of transcription factor TFIID) causes a marked reduction in transcription, and gives rise to heterogeneous transcription initiation sites.

Taxy: Taxy is a graphical application for navigating, querying and searching taxonomic data with a user friendly interface and a plug-in system for accessing different sources of data. The main plug-in is for accessing the data of NCBI taxonomy database releases. Also there is a plug-in for accessing in-house databases available for internal use only.

Terminator line: The // (terminator) line contains no data or comments and designates the end of an entry.

Terminator line (EMBL): The // (terminator) line also contains no data or comments. It designates the end of an entry.

Tertiary structure: Folding of a protein chain via interactions of its sideschain molecules including formation of disulphide bonds between cysteine residues.

Textsearch: An **EMBOSS** application. This is a small utility search for words in the description text of a sequence and for each match list the sequence's name and/or description. NB. It only searches the description line of the annotation, not the full annotation.

tfastx3: A program that compares a protein to a translated DNA data bank

tfastx3: A program that compares a protein to a translated DNA data bank

tfextract: An **EMBOSS** application. The TRANSFAC Database is a database of eukaryotic cis-acting regulatory DNA elements and trans-acting factors. It covers the whole range from yeast to human. TRANSFAC started in 1988 with a printed compilation (Nucleic Acids Res. 16: 1879-1902, 1988) and was transferred into computer-readable format in 1990 (BioTechForum - Advances in Molecular Genetics (J. Collins, A.J. Driesel, eds.) 4:95-108, 1991).

Tfm: An **EMBOSS** application. This program displays the help documentation for an **EMBOSS** program. The contributors of the **EMBOSS** programs do attempt to provide an adequate description of the programs. This documentation is primarily held as HTML pages at **http://www.uk.embnet.org/Software/EMBOSS/Apps/**

Tfscan: An **EMBOSS** application. The TRANSFAC Database is a database of eukaryotic cis-acting regulatory DNA elements and trans-acting factors. It covers the whole range from yeast to human.

Thornton's Group: As can be seen below, this group has a wide range of research interests, with some emphasis on obtaining an understanding of the following from biomolecular structure, (although sequence-based methods play a crucial role as well). Research Interests : Enzyme active sites, Protein-Protein interactions, Protein-ligand interactions, Protein-DNA Interactions, Structure and Modelling.

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. Since many organs are multifunctional, they have developed highly specialised cell types to perform different functions. Identifying the section of an organ that is homogenous for a particular cell type ensures that the gene expression profiles extracted from those cells will accurately resemble the class of cells that make up the tissue.

Tissue-specific expression: Gene function which is restricted to a particular tissue or cell type. For example, the glycoprotein hormone alpha subunit is produced only in certain cell types of the anterior pituitary and placenta, not in lungs or skin; thus expression of the glycoprotein hormone alpha-chain gene is said to be tissue-specific. Tissue specific expression is usually the result of an enhancer which is activated only in the proper cell type.

Tmap: An **EMBOSS** application. This program predicts transmembrane segments in proteins, utilising the algorithm described in: "Persson, B. & Argos, P. (1994) Prediction of transmembrane segments in proteins utilising multiple sequence alignments J. Mol. Biol. 237, 182-192." tmap reads in one or more aligned protein sequences.

Tranalign: An **EMBOSS** application. tranalign is a re-implementation in **EMBOSS** of the program mrtrans by Bill Pearson.

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

Transcription: The process of copying DNA to produce an RNA transcript. This is the first step in the expression of any gene. The resulting RNA, if it codes for a protein, will be spliced, polyadenylated, transported to the cytoplasm, and by the process of translation will produce the desired protein molecule.

Transcription factor: A protein which is involved in the transcription of genes. These usually bind to DNA as part of their function (but not necessarily). A transcription factor may be general (i.e. acting on many or all genes in all tissues), or tissue-specific (i.e. present only in a particular cell type, and activating the genes restricted to that cell type). Its activity may be constitutive, or may depend on the presence of some stimulus; for example, the glucocorticoid receptor is a transcription factor which is active only when glucocorticoids are present.

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.

Transeq: An **EMBOSS** application. This translates nucleic acid sequences to the corresponding peptide sequence.

Transfection: A method by which experimental DNA may be put into a cultured mammalian cell. Such experiments are usually performed using cloned DNA containing coding sequences and control regions (promoters, etc) in order to test whether the DNA will be expressed. Since the cloned DNA may have been extensively modified (for example, protein binding sites on the promoter may have been altered or removed), this procedure is often used to test whether a particular modification affects the function of a gene.

Transformation: With respect to cultured cells, a change in cell morphology and behavior which is generally related to carcinogenesis. Transformed cells tend to exhibit characteristics known collectively as the "transformed phenotype" (rounded cell bodies, reduced attachment dependence, increased growth rate, loss of contact inhibition, etc). There are different "degrees" of transformation, and cells may exhibit only a subset of these characteristics. Not well understood, the process of transformation is the subject of intense research. With respect to bacteria, the process by which a bacteria acquires a plasmid and becomes antibiotic resistant. This term most commonly refers to a bench procedure performed by the investigator which introduces experimental plasmids into bacteria.

Transgenic mouse: A mouse which carries experimentally introduced DNA. The procedure by which one makes a transgenic mouse involves the injection of DNA into a fertilised embryo at the pro-nuclear stage. The DNA is generally cloned, and may be experimentally altered. It will become incorporated into the genome of the embryo. That embryo is implanted into a foster mother, who gives birth to an animal carrying the new gene. Various experiments are then carried out to test the functionality of the inserted DNA.

Transient transfection: When DNA is transfected into cultured cells, it is able to stay in those cells for about 2-3 days, but then will be lost (unless steps are taken to ensure that it is retained - see Stable transfection). During those 2-3 days, the DNA is functional, and any functional genes it contains will be expressed. Investigators take advantage of this transient expression period to test gene function.

Translation: The process of decoding a strand of mRNA, thereby producing a protein based on the code. This process requires ribosomes (which are composed of rRNA along with various proteins) to perform the synthesis, and tRNA to bring in the amino acids. Sometimes, however, people speak of "translating" the DNA or RNA when they are merely reading the nucleotide sequence and predicting from it the sequence of the encoded protein. This might be more accurately termed "conceptual translation".

Transmembrane region: The region of a transmembrane protein that actually spans the membrane. Transmembrane regions are usually hydrophobic in order to be thermodynamically compatible with the lipid bilayer portion of the membrane. They may consist of either alpha-helical or beta-strand secondary structure elements, but in either case the external residues (the ones facing the membrane) are invariably hydrophobic while the internal residues may be hydrophilic (as in the case of a pore or channel) or polar. One common transmembrane structural domain is the seven-helix bundle seen in numerous channel proteins.

TrEMBL Database: **TrEMBL** is a computer-annotated supplement to SWISS-PROT. **TrEMBL** contains the translations of all coding sequences (CDS) present in the **EMBL** Nucleotide Sequence Database, which are not yet integrated into **SWISS-PROT**.

TrEMBL New : Translated **EMBL** Updates

Trimmest: An **EMBOSS** application. EST and mRNA sequences often have poly-A tails at the end of them. This utility removes those poly-A tails.

Trimseq: An **EMBOSS** application. This program is used to tidy up the ends of sequences, removing all the bits that you would really rather were not published.

tRNA: This is transfer RNA, it is the information adapter molecule. It is the direct interface between amino-acid sequence of a protein and the information in DNA. Therefore it decodes the information in DNA.

Tumor suppressor: A gene that inhibits progression towards neoplastic transformation. The best-known examples of tumor suppressors are the proteins p53 and Rb.

Twofeat: An **EMBOSS** application. twofeat reads in the feature tables of sequences and reports occurances of pairs of specified features.

U

Union: An **EMBOSS** application. union reads in several sequences, concatenates them and writes them out as a single sequence

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

Upstream activator sequence: A binding site for transcription factors, generally part of a promoter region. A UAS may be found upstream of the TATA sequence (if there is one), and its function is (like an enhancer) to increase transcription. Unlike an enhancer, it can not be positioned just anywhere or in any orientation.

Upstream/Downstream: In an RNA, anything towards the 5' end of a reference point is "upstream" of that point. This orientation reflects the direction of both the synthesis of mRNA, and its translation - from the 5' end to the 3' end. In DNA, the situation is a bit more complicated. In the vicinity of a gene (or in a cDNA), the DNA has two strands, but one strand is virtually a duplicate of the RNA, so it's 5' and 3' ends determine upstream and downstream, respectively. NOTE that in genomic DNA, two adjacent genes may be on different strands and thus oriented in opposite directions. Upstream or downstream is only used in conjunction with a given gene.

Uracil: Nitrogenous pyrimidine base found in RNA but not DNA.

V

Vector: The DNA "vehicle" used to carry experimental DNA and to clone it. The vector provides all sequences essential for replicating the test DNA. Typical vectors include plasmids, cosmids, phages and YACs.

Vectorstrip: An **EMBOSS** application. vectorstrip is intended to be useful for stripping vector sequence from the ends of sequences of interest. For example, if a fragment has been cloned into a vector and then sequenced, the sequence may contain vector data eg from the cloning polylinker at the 5' and 3' ends of the sequence. vectorstrip will remove these contaminating regions and output trimmed sequence ready for input into another application.

W

Water: An **EMBOSS** application. water uses the Smith-Waterman algorithm (modified for speed enhancments) to calculate the local alignment.

Water Program: Water uses the Smith-Waterman algorithm (modified for speed enhancements) to calculate the local alignment.

WEBIN: Interactive system for **submitting** DNA sequences to EMBL/GenBank/DDBJ.

Webin-Align: **Webin-Align**. Interactive submission tool for sequence alignments

Western blot: A technique for analysing mixtures of proteins to show the presence, size and abundance of one particular type of protein. Similar to Southern or Northern blotting (see "Blotting"), except that (1) a protein mixture is electrophoresed in an acrylamide gel, and (2) the "probe" is an antibody which recognises the protein of interest, followed by a radioactive secondary probe (such as 125I-protein A).

Whichdb: An **EMBOSS** application. whichdb searches all available **EMBOSS** databases for sequences with a specified ID name or accession number.

Wild type: Form of a gene or allele that is considered the "standard" or most common.

Wobble: An **EMBOSS** application. Wobble plots the third position variability as an indicator of a potential coding region.

Wordcount: An **EMBOSS** application. Displays all the words of the specified length with the number of times it occurs.

Wordmatch: An **EMBOSS** application. Finds all exact matches of a given minimum size between 2 sequences displaying the start points in each sequence and the match length.

Wossname: An **EMBOSS** application. This allows a user to search for keywords or parts of words in the brief documentation (as displayed by a program when it first starts). The program name and the brief description is output. If no words to search for are specified, then details of all the **EMBOSS** programs are output.

WU-Blast2: WU-Blast2 stands for Washington University Basic Local Alignment Search Tool Version 2.0. It is used to compare a novel sequence with those contained in nucleotide and protein databases by aligning the novel sequence with previously characterised genes. The emphasis of this tool is to find regions of sequence similarity, which will yield functional and evolutionary clues about the structure and function of this novel sequence. Regions of similarity detected via this type of alignment tool can be either local, where the region of similarity is based in 1 location, or global, where regions of similarity can be detected across otherwise unrelated genetic code. The fundamental unit of **BLAST** algorithm output is the High-scoring Segment Pair (HSP). An HSP consists of two sequence fragments of arbitrary but equal length whose alignment is locally maximal and for which the alignment score meets or exceeds a threshold or cutoff score. A set of HSPs is thus defined by two sequences, a scoring system, and a cutoff score; this set may be empty if the cutoff score is sufficiently high. In the programmatic implementations of the **BLAST** algorithm described here, each HSP consists of a segment from the query sequence and one from a database sequence. The sensitivity and speed of the programs can be adjusted via the standard **BLAST** algorithm parameters W, T, and X (Altschul et al., 1990); selectivity of the programs can be adjusted via the cutoff score. The approach to similarity searching taken by the **BLAST** programs is first to look for similar segments (HSPs) between the query sequence and a database sequence, then to evaluate the statistical significance of any matches that were found, and finally to report only those matches that satisfy a user-selectable threshold of significance. Findings of multiple HSPs involving the query sequence and a single database sequence may be treated statistically in a variety of ways. By default the programs use ''Sum'' statistics (Karlin and Altschul, 1993). As such, the statistical significance ascribed to a set of HSPs may be higher than that ascribed to any individual member of the set. Only when the ascribed significance satisfies the user-selectable threshold (EXP THR parameter) will the match be reported to the user.

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).

XX line: The XX line in database entries contains no data or comments. It is used instead of blank lines to avoid confusion with the sequence data lines.

Y

YAC: Yeast artificial chromosome. This is a method for cloning very large fragments of DNA. Genomic DNA in fragments of 200-500 kb are linked to sequences which allow them to propagate in yeast as a mini-chromosome (including telomeres, a centromere and an ARS - an autonomous replication sequence). This technique is used to clone large genes and intergenic regions, and for chromosome walking

Yank: An **EMBOSS** application. yank is a simple utility to add a specified sequence name to a list file. In fact, it writes out not just the name of the sequence, but also the start and end position of a region within that sequence and, if the sequence is nucleic, it can specify whether the sequence is the reverse complement. Without the program yank you would need to use a text editor such as pico to create the appropriate list files. yank makes this process easy.

Yeast 2-hybrid system: A yeast-based method used to simultaneously identify, and clone the gene for, proteins interacting with a known protein. The basis of this method is a "transcriptional reporter assay" (see definition) in which reporter gene expression is dependent on two domains. The first domain is linked to the known protein. The second domain is genetically linked to a library. If the library is screened against the known protein the two domains will interact only if a protein from the library binds the known protein, resulting in transcription activation of the reporter gene, and a blue color. The "blue yeast clone" will contain the gene encoding the newly identified protein.

Z

Zinc finger: A protein structural motif common in DNA binding proteins. Four Cys residues are found for each "finger" and one finger can bind a molecule of zinc.

Reference:

www.ebi.ac.uk/2can/glossary/