Glossary

Abstract Syntax Notation (ASN.1)
A language that is used to describe structured data types formally. Within bioinformatics, it has been used by the National Center for Biotechnology Information to encode sequences, maps, taxonomic information, molecular structures, and biographical information in such a way that it can be easily accessed and exchanged by computer software.

Accession number
A unique identifier that is assigned to a single database entry for a DNA or protein sequence.

Affine gap penalty
A gap penalty score that is a linear function of gap length, consisting of a gap opening penalty and a gap extension penalty multiplied by the length of the gap. Using this penalty scheme greatly enhances the performance of dynamic programming methods for sequence alignment. See also Gap penalty.

Algorithm
A systematic procedure for solving a problem in a finite number of steps, typically involving a repetition of operations. Once specified, an algorithm can be written in a computer language and run as a program.

Alignment
Refers to the procedure of comparing two or more sequences by looking for a series of individual characters or character patterns that are in the same order in the sequences. Of the two types of alignment, local and global, a local alignment is generally the most useful. See also Local and Global alignments.

Alignment score
An algorithmically computed score based on the number of matches, substitutions, insertions, and deletions (gaps) within an alignment. Scores for matches and substitutions are derived from a scoring matrix such as the BLOSUM and PAM matrices for proteins, and affine gap penalties suitable for the matrix are chosen. Alignment scores are in log odds units, often bit units (log to the base 2). Higher scores denote better alignments. See also Similarity score, Distance in sequence analysis.

Alphabet
The total number of symbols in a sequence—4 for DNA sequences and 20 for protein sequences.

Annotation
The prediction of genes in a genome, including the location of protein-encoding genes, the sequence of the encoded proteins, any significant matches to other proteins of known function, and the location of RNA-encoding genes. Predictions are based on gene models; e.g., hidden Markov models of introns and exons in proteins encoding genes, and models of secondary structure in RNA.

Anonymous FTP
When a FTP service allows anyone to log in, it is said to provide anonymous FTP service. A user can log in to an anonymous FTP server by typing *anonymous* as the user name and his E-mail address as a password. Most Web browsers now negotiate anonymous FTP logon without asking the user for a user name and password. See also FTP.

ASCII
The American Standard Code for Information Interchange (ASCII) encodes unaccented letters a–z, A–Z, the numbers 0–9, most punctuation marks, space, and a set of control characters such as carriage return and tab. ASCII specifies 128 characters that are mapped to the values 0–127. ASCII files are commonly called "plain text," meaning that they only encode text without extra markup.

Back-propagation
When training feed-forward neural networks, a back-propagation algorithm can be used to modify the network weights. After each training input pattern is fed through the network, the network's output is compared with the desired output and the amount of error is calculated. This error is back-propagated through the network by using an error function to correct the network weights. See also Feed-forward neural network.

Baum-Welch algorithm
An expectation maximization algorithm that is used to train hidden Markov models.

Bayes' rule
Forms the basis of conditional probability by calculating the likelihood of an event occurring based on the history of the event and relevant background information. In terms of two parameters A and B, the theorem is stated in an equation: The conditional probability of A, given B, $P(A|B)$, is equal to the probability of A, $P(A)$, times the conditional probability of B, given A, $P(B|A)$, divided by the probability of B, $P(B)$. $P(A)$ is the historical or prior distribution value of A, $P(B|A)$ is a new prediction for B for a particular value of A, and $P(B)$ is the sum of the newly predicted values for B. $P(A|B)$ is a posterior probability, representing a new prediction for A given the prior knowledge of A and the newly discovered relationships between A and B.

Bayesian analysis
A statistical procedure used to estimate parameters of an underlying distribution based on an observed distribution. See also Bayes' rule.

Biochips
Miniaturized arrays of large numbers of molecular substrates, often oligonucleotides, in a defined pattern. They are also called DNA microarrays and microchips.

Bioinformatics
An interdisciplinary field involving biology, computer science, mathematics, and statistics to analyze biological sequence data, genome content, and arrangement, and to predict the function and structure of macromolecules.

Bit units
From information theory, a bit denotes the amount of information required to distinguish between two equally likely possibilities. The number of bits of information, N, required to convey a message that has M possibilities is $\log_2 M = N$ bits.

Block
Conserved ungapped patterns approximately 3–60 amino acids in length in a set of related proteins.

BLOSUM matrices
An alternative to PAM tables, BLOSUM tables were derived using local multiple alignments of more distantly related sequences than were used for the PAM matrix. These are used to assess the similarity of sequences when performing alignments.

Boltzmann distribution
Describes the number of molecules that have energies above a certain level, based on the Boltzmann gas constant and the absolute temperature.

Boltzmann probability function
See Boltzmann distribution.

Bootstrap analysis
A method for testing how well a particular data set fits a model. For example, the validity of the branch arrangement in a predicted phylogenetic tree can be tested by resampling columns in a multiple sequence alignment to create many new alignments. The appearance of a particular branch in trees generated from these resampled sequences can then be measured. Alternatively, a sequence may be left out of an analysis to determine how much the sequence influences the results of an analysis.

Branch length
In sequence analysis, the number of sequence changes along a particular branch of a phylogenetic tree.

Chebyshev's inequality
The probability that a random variable exceeds its mean is less than or equal to the square of 1 over the number of standard deviations from the mean.

Cluster analysis
A method for grouping together a set of objects that are most similar from a larger group of related objects. The relationships are based on some criterion of similarity or difference. For sequences, a similarity or distance score or a statistical evaluation of those scores is used.

Cobbler
A single sequence that represents the most conserved regions in a multiple sequence alignment. The BLOCKS server uses the cobbler sequence to perform a database similarity search as a way to reach sequences that are more divergent than would be found using the single sequences in the alignment for searches.

Coding system (neural networks)
Regarding neural networks, a coding system needs to be designed for representing input and output. The level of success found when training the model will be partially dependent on the quality of the coding system chosen.

Codon usage
Analysis of the codons used in a particular gene or organism.

COG
Clusters of orthologous groups in a set of groups of related sequences in microorganisms and yeast (*S. cerevisiae*). These groups are found by whole proteome comparisons and include orthologs and paralogs. See also Orthologs and Paralogs.

Comparative genomics
A comparison of gene numbers, gene locations, and biological functions of genes in the genomes of diverse organisms, one objective being to identify groups of genes that play a unique biological role in a particular organism.

Complexity (of an algorithm)
Describes the number of steps required by the algorithm to solve a problem as a function of the amount of data; for example, the length of sequences to be aligned.

Conditional probability
The probability of a particular result (or of a particular value of a variable) given one or more events or conditions (or values of other variables).

Consensus
A single sequence that represents, at each subsequent position, the variation found within corresponding columns of a multiple sequence alignment.

Context-free grammars
A recursive set of production rules for generating patterns of strings. These consist of a set of terminal characters that are used to create strings, a set of nonterminal symbols that correspond to rules and act as placeholders for patterns that can be generated using terminal characters, a set of rules for replacing nonterminal symbols with terminal characters, and a start symbol.

Contig
A set of clones that can be assembled into a linear order.

CORBA
The Common Object Request Broker Architecture (CORBA) is an open industry standard for working with distributed objects, developed by the Object Management Group. CORBA allows the interconnection of objects and applications regardless of computer language, machine architecture, or geographic location of the computers.

Correlation coefficient
A numerical measure, falling between -1 and 1, of the degree of the linear relationship between two variables. A positive value indicates a direct relationship, a negative value indicates an inverse relationship, and the distance of the value away from zero indicates the strength of the relationship. A value near zero indicates no relationship between the variables.

Covariation (in sequences)
Coincident change at two or more sequence positions in related sequences that may influence the secondary structures of RNA or protein molecules.

Database
A computerized storehouse of data that provides a standardized way for locating, adding, removing, and changing data. See also Object-oriented database, Relational database.

Dendogram
A form of a tree that lists the compared objects (e.g., sequences or genes in a microarray analysis) in a vertical order and joins related ones by levels of branches extending to one side of the list.

Dirichlet mixtures
Defined as the conjugational prior of a multinomial distribution. One use is for predicting the expected pattern of amino acid variation found in the match state of a hidden Markov model (representing one column of a multiple sequence alignment of proteins), based on prior distributions found in conserved protein domains (blocks).

Distance in sequence analysis
The number of observed changes in an optimal alignment of two sequences, usually not counting gaps.

Dot matrix
Dot matrix diagrams provide a graphical method for comparing two sequences. One sequence is written horizontally across the top of the graph and the other along the left-hand side. Dots are placed within the graph at the intersection of the same letter appearing in both sequences. A series of diagonal lines in the graph indicate regions of alignment. The matrix may be filtered to reveal the most-alike regions by scoring a minimal threshold number of matches within a sequence window.

Dynamic programming
A dynamic programming algorithm solves a problem by combining solutions to sub-

problems that are computed once and saved in a table or matrix. Dynamic programming is typically used when a problem has many possible solutions and an optimal one needs to be found. This algorithm is used for producing sequence alignments, given a scoring system for sequence comparisons.

Entropy

From information theory, a measure of the unpredictable nature of a set of possible elements. The higher the level of variation within the set, the higher the entropy.

Erdos and Renyi law

In a toss of a "fair" coin, the number of heads in a row that can be expected is the logarithm of the number of tosses to the base 2. The law may be generalized for more than two possible outcomes by changing the base of the logarithm to the number of outcomes. This law was used to analyze the number of matches and mismatches that can be expected between random sequences as a basis for scoring the statistical significance of a sequence alignment.

Expect value (*E*)

In a database similarity search, the probability that an alignment score as good as the one found between a query sequence and a database sequence would be found in as many comparisons between random sequences as was done to find the matching sequence. In other types of sequence analysis, *E* has a similar meaning.

Expectation maximization (sequence analysis)

An algorithm for locating similar sequence patterns in a set of sequences. A guessed alignment of the sequences is first used to generate an expected scoring matrix representing the distribution of sequence characters in each column of the alignment, this pattern is matched to each sequence, and the scoring matrix values are then updated to maximize the alignment of the matrix to the sequences. The procedure is repeated until there is no further improvement.

Extreme value distribution

Some measurements are found to follow a distribution that has a long tail which decays at high values much more slowly than that found in a normal distribution. This slow-falling type is called the extreme value distribution. The alignment scores between unrelated or random sequences are an example. These scores can reach very high values, particularly when a large number of comparisons are made, as in a database similarity search. The probability of a particular score may be accurately predicted by the extreme value distribution, which follows a double negative exponential function after Gumbel.

False negative

A negative data point collected in a data set that was incorrectly reported due to a failure of the test in avoiding negative results.

False positive

A positive data point collected in a data set that was incorrectly reported due to a failure of the test. If the test had correctly measured the data point, the data would have been recorded as negative.

Feed-forward neural network

Organizes nodes into sequence layers in which the nodes in each layer are fully connected with the nodes in the next layer, except for the final output layer. Input is fed from the input layer through the layers in sequence in a "feed-forward" direction, resulting in output at the final layer. See also Neural network.

Filtering (window size)

During pair-wise sequence alignment using the dot matrix method, random matches can be filtered out by using a sliding window to compare the two sequences. Rather than comparing a single sequence position at a time, a window of adjacent positions in the

two sequences is compared and a dot, indicating a match, is generated only if a certain minimal number of matches occur.

Fourier analysis
Studies the approximations and decomposition of functions using trigonometric polynomials.

Format (file)
Different programs require that information be specified to them in a formal manner, using particular keywords and ordering. This specification is a file format.

Forward-backward algorithm
Used to train a hidden Markov model by aligning the model with training sequences. The algorithm then refines the model to reduce the error when fitted to the given data using a gradient descent approach.

FTP (File Transfer Protocol)
Allows a person to transfer files from one computer to another across a network using an FTP-capable client program. The FTP client program can only communicate with machines that run an FTP server. The server, in turn, will make a specific portion of its file system available for FTP access, providing that the client is able to supply a recognized user name and password to the server.

Functional genomics
Assessment of the function of genes identified by between-genome comparisons. The function of a newly identified gene is tested by introducing mutations into the gene and then examining the resultant mutant organism for an altered phenotype.

Gap
Mismatch in the alignment of two sequences caused by either an insertion in one sequence or a deletion in the other.

Gap penalty
A numeric score used in sequence alignment programs to penalize the presence of gaps within an alignment. The value of a gap penalty affects how often gaps appear in alignments produced by the algorithm. Most alignment programs suggest gap penalties that are appropriate for particular scoring matrices.

Genetic algorithm
A kind of search algorithm that was inspired by the principles of evolution. A population of initial solutions is encoded and the algorithm searches through these by applying a pre-defined fitness measurement to each solution, selecting those with the highest fitness for reproduction. New solutions can be generated during this phase by crossover and mutation operations, defined in the encoded solutions.

Genome
The genetic material of an organism, contained in one haploid set of chromosomes.

Gibbs sampling method
An algorithm for finding conserved patterns within a set of related sequences. A guessed alignment of all but one sequence is made and used to generate a scoring matrix that represents the alignment. The matrix is then matched to the left-out sequence, and a probable location of the corresponding pattern is found. This prediction is then input into a new alignment and another scoring matrix is produced and tested on a new left-out sequence. The process is repeated until there is no further improvement in the matrix.

Global alignment
Attempts to match as many characters as possible, from end to end, in a set of two or more sequences.

Graph theory
A branch of mathematics which deals with problems that involve a graph or network structure. A graph is defined by a set of nodes (or points) and a set of arcs (lines or edges) joining the nodes. In sequence and genome analysis, graph theory is used for sequence alignments and clustering alike genes.

Half-bits
Some scoring matrices are in half-bit units. These units are logarithms to the base 2 of odds scores times 2.

Heuristic
A procedure that progresses along empirical lines by using rules of thumb to reach a solution. The solution is not guaranteed to be optimal.

Hexadecimal system
The base 16 counting system that uses the digits 0–9 followed by the letters A–F.

Hidden Markov Models (HMM)
In sequence analysis, a HMM is usually a probabilistic model of a multiple sequence alignment, but can also be a model of periodic patterns in a single sequence, representing, for example, patterns found in the exons of a gene. In a model of multiple sequence alignments, each column of symbols in the alignment is represented by a frequency distribution of the symbols called a state, and insertions and deletions by other states. One then moves through the model along a particular path from state to state trying to match a given sequence. The next matching symbol is chosen from each state, recording its probability (frequency) and also the probability of going to that particular state from a previous one (the transition probability). State and transition probabilities are then multiplied to obtain a probability of the given sequence. Generally speaking, a HMM is a statistical model for an ordered sequence of symbols, acting as a stochastic state machine that generates a symbol each time a transition is made from one state to the next. Transitions between states are specified by transition probabilities.

Hidden layer
An inner layer within a neural network that receives its input and sends its output to other layers within the network. One function of the hidden layer is to detect covariation within the input data, such as patterns of amino acid covariation that are associated with a particular type of secondary structure in proteins.

Hierarchical clustering
The clustering or grouping of objects based on some single criterion of similarity or difference. An example is the clustering of genes in a microarray experiment based on the correlation between their expression patterns. The distance method used in phylogenetic analysis is another example.

Hill climbing
A nonoptimal search algorithm that selects the singular best possible solution at a given state or step. The solution may result in a locally best solution that is not a globally best solution.

Homolog
A similar component in two organisms (e.g., genes with strongly similar sequences) that can be attributed to a common ancestor of the two organisms during evolution.

Horizontal transfer
The transfer of genetic material between two distinct species that do not ordinarily exchange genetic material. The transferred DNA becomes established in the recipient genome and can be detected by a novel phylogenetic history and codon content compared to the rest of the genome.

HTML
The Hyper-Text Markup Language (HTML) provides a structural description of a document using a specified tag set. HTML currently serves as the Internet lingua franca for describing hypertext Web page documents.

Hyperplane
A generalization of the two-dimensional plane to N dimensions.

Hypercube
A generalization of the three-dimensional cube to N dimensions.

Indel
An insertion or deletion in a sequence alignment.

Information content (of a scoring matrix)
A representation of the degree of sequence conservation in a column of a scoring matrix representing an alignment of related sequences. It is also the number of questions that must be asked to match the column to a position in a test sequence. For bases, the maximum possible number is 2, and for proteins, 4.32 (logarithm to the base 2 of the number of possible sequence characters).

Information theory
A branch of mathematics that measures information in terms of bits, the minimal amount of structural complexity needed to encode a given piece of information.

Input layer
The initial layer in a feed-forward neural net. This layer encodes input information that will be fed through the network model.

Interface definition language
Used to define an interface to an object model in a programming language neutral form, where an interface is an abstraction of a service defined only by the operations that can be performed on it.

Internet
The network infrastructure, consisting of cables interconnected by routers, that provides global connectivity for individual computers and private networks of computers. A second sense of the word "internet" is the collective computer resources available over this global network.

Interpolated Markov model
A type of Markov model of sequences that examines sequences for patterns of variable length in order to discriminate best between genes and non-gene sequences.

Iterative
A sequence of operations in a procedure that is performed repeatedly.

***K*-tuple**
Identical short stretches of sequences, also called words.

Likelihood
The hypothetical probability that an event which has already occurred would yield a specific outcome. Unlike probability, which refers to future events, likelihood refers to past events.

Linear discriminant analysis
An analysis in which a straight line is located on a graph between two sets of data points in a location that best separates the data points into two groups.

Local alignment
Attempts to align regions of sequences with the highest density of matches. In doing so, one or more islands of subalignments are created in the aligned sequences.

Log odds score
The logarithm of an odds score. See also Odds score.

Machine learning
The training of a computational model of a process or classification scheme to distinguish between alternative possibilities.

Markov chain
Describes a process that can be in one of a number of states at any given time. The Markov chain is defined by probabilities for each transition occurring; that is, probabilities of the occurrence of state s_j given that the current state is s_i. Substitutions in nucleic acid and protein sequences are generally assumed to follow a Markov chain in that each site changes independently of the previous history of the site. With this model, the number and types of substitutions observed over a relatively short period of evolutionary time can be extrapolated to longer periods of time. In performing sequence alignments and calculating the statistical significance of alignment scores, sequences are assumed to be Markov chains in which the choice of one sequence position is not influenced by another.

Maximum likelihood (phylogeny, alignment)
The most likely outcome (tree or alignment), given a probabilistic model of evolutionary change in DNA sequences.

Maximum parsimony
The minimum number of evolutionary steps required to generate the observed variation in a set of sequences, as found by comparison of the number of steps in all possible phylogenetic trees.

Method of moments
The mean or expected value of a variable is the first moment of the values of the variable around the mean, defined as that number from which the sum of deviations to all values is zero. The standard deviation is the second moment of the values about the mean, and so on.

Minimum spanning tree
Given a set of related objects classified by some similarity or difference score, the minimum spanning tree joins the most-alike objects on adjacent outer branches of a tree and then sequentially joins less-alike objects by more inward branches. The tree branch lengths are calculated by the same neighbor-joining algorithm that is used to build phylogenetic trees of sequences from a distance matrix. The sum of the resulting branch lengths between each pair of objects will be approximately that found by the classification scheme.

Molecular clock hypothesis
The hypothesis that sequences change at the same rate in the branches of an evolutionary tree.

Monte Carlo
A method that samples possible solutions to a complex problem as a way to estimate a more general solution.

Mutation data matrix
A scoring matrix compiled from the observation of point mutations between aligned sequences. Also refers to a Dayhoff PAM matrix in which the scores are given as log odds scores.

Nats (natural logarithm)
A number expressed in units of the natural logarithm.

Needleman-Wunsch algorithm
Uses dynamic programming to find global alignments between sequences.

Neighbor-joining method
Clusters together alike pairs within a group of related objects (e.g., genes with similar

sequences) to create a tree whose branches reflect the degrees of difference among the objects.

Neural network

From artificial intelligence algorithms, techniques that involve a set of many simple units that hold symbolic data, which are interconnected by a network of links associated with numeric weights. Units operate only on their symbolic data and on the inputs that they receive through their connections. Most neural networks use a training algorithm (see Back-propagation) to adjust connection weights, allowing the network to learn associations between various input and output patterns. See also Feed-forward neural network.

Noise

In sequence analysis, a small amount of randomly generated variation in sequences that is added to a model of the sequences; e.g., a hidden Markov model or scoring matrix, in order to avoid the model overfitting the sequences. See also Overfitting.

Normal distribution

The distribution found for many types of data such as body weight, size, and exam scores. The distribution is a bell-shaped curve that is described by a mean and standard deviation of the mean. Local sequence alignment scores between unrelated or random sequences do not follow this distribution but instead the extreme value distribution which has a much extended tail for higher scores. See also Extreme value distribution.

Object Management Group (OMG)

A not-for-profit corporation that was formed to promote component-based software by introducing standardized object software. The OMG establishes industry guidelines and detailed object management specifications in order to provide a common framework for application development. Within OMG is a Life Sciences Research group, a consortium representing pharmaceutical companies, academic institutions, software vendors, and hardware vendors who are working together to improve communication and interoperability among computational resources in life sciences research.

Object-oriented database

Unlike relational databases (see entry), which use a tabular structure, object-oriented databases attempt to model the structure of a given data set as closely as possible. In doing so, object-oriented databases tend to reduce the appearance of duplicated data and the complexity of query structure often found in relational databases.

Odds score

The ratio of the likelihoods of two events or outcomes. In sequence alignments and scoring matrices, the odds score for matching two sequence characters is the ratio of the frequency with which the characters are aligned in related sequences divided by the frequency with which those same two characters align by chance alone, given the frequency of occurrence of each in the sequences. Odds scores for a set of individually aligned positions are obtained by multiplying the odds scores for each position. Odds scores are often converted to logarithms to create log odds scores that can be added to obtain the log odds score of a sequence alignment.

Optimal alignment

The highest-scoring alignment found by an algorithm capable of producing multiple solutions. This is the best possible alignment that can be found, given any parameters supplied by the user to the sequence alignment program.

Orthologs

A pair of genes found in two species are orthologous when the encoded proteins are 60–80% identical in an alignment. The proteins almost certainly have the same three-dimensional structure, domain structure, and biological function, and the encoding

genes have originated from a common ancestor gene at an earlier evolutionary time. Two orthologs I and II in genomes A and B, respectively, may be identified when the complete genomes of two species are available: (1) in a database similarity search of all of the proteome of B using I as a query, II is the best hit found, and (2) I is the best hit when II is used as a query of the proteome of B. The best hit is the database sequence with the highest expect value (*E*). Orthology is also predicted by a very close phylogenetic relationship between sequences or by a cluster analysis. Compare to Paralogs. See also Cluster analysis.

Output layer

The final layer of a neural network in which signals from lower levels in the network are input into output states where they are weighted and summed to give an output signal. For example, the output signal might be the prediction of one type of protein secondary structure for the central amino acid in a sequence window.

Overfitting

Can occur when using a learning algorithm to train a model such as a neural net or hidden Markov model. Overfitting refers to the model becoming too highly representative of the training data and thus no longer representative of the overall range of data that is supposed to be modeled.

Pair-wise sequence alignment

An alignment performed between two sequences.

PAM scoring matrices

Percent Accepted Mutation or PAM matrices describe the probability that one base or amino acid has changed during the course of evolution. Amino acid PAM matrices are derived from families of closely related sequences and are used to assess the similarity of sequences when performing alignments.

Paralogs

Genes that are related through gene duplication events. These events may lead to the production of a family of related proteins with similar biological functions within a species. Paralogous gene families within a species are identified by using an individual protein as a query in a database similarity search of the entire proteome of an organism. The process is repeated for the entire proteome and the resulting sets of related proteins are then searched for clusters that are most likely to have a conserved domain structure and should represent a paralogous gene family.

Parametric sequence alignment

An algorithm that finds a range of possible alignments based on varying the parameters of the scoring system for matches, mismatches, and gap penalties. An example is the Bayes block aligner.

Pearson correlation coefficent

A measure of the correlation between two variables that reflects the degree to which the two variables are related. For example, the coefficient is used as a measure of similarity of gene expression in a microarray experiment. See also Correlation coefficient.

Percent identity

The percentage of the columns in an alignment of two sequences that includes identical amino acids. Columns in the alignment that include gaps are not scored in the calculation.

Percent similarity

The percentage of the columns in an alignment of two sequences that includes either identical amino acids or amino acids that are frequently found substituted for each other in sequences of related proteins (conservative substitutions). These substitutions may be found in an amino acid substitution matrix such as the Dayhoff PAM and

Henikoff BLOSUM matrices. Columns in the alignment that include gaps are not scored in the calculation.

Perceptron

A neural network in which input and output states are directly connected without intervening hidden layers.

Poisson distribution

Used to predict the occurrence of infrequent events over a long period of time or when there are a large number of trials. In sequence analysis, it is used to calculate the chance that one pair of a large number of pairs of unrelated sequences may give a high local alignment score.

Position-specific scoring matrix (PSSM)

Represents the variation found in the columns of an alignment of a set of related sequences. Each subsequent matrix column corresponds to the next column in the alignment and each row corresponds to a particular sequence character (one of four bases in DNA sequences or 20 amino acids in protein sequences). Matrix values are log odds scores obtained by dividing the counts of the residue in the alignment, dividing by the expected number of counts based on sequence composition, and converting the ratio to a log score. The matrix is moved along sequences to find similar regions by adding the matching log odds scores and looking for high values. There is no allowance for gaps. Also called a weight matrix or scoring matrix.

Posterior (Bayesian analysis)

A conditional probability based on prior knowledge and newly evaluated relationships among variables using Bayes' rule. See also Bayes' rule.

Prior (Bayesian analysis)

The expected distribution of a variable based on previous data.

Profile

A matrix representation of a conserved region in a multiple sequence alignment that allows for gaps in the alignment. The rows include scores for matching sequential columns of the alignment to a test sequence. The columns include substitution scores for amino acids and gap penalties.

Profile hidden Markov model

A hidden Markov model of a conserved region in a multiple sequence alignment that includes gaps and may be used to search new sequences for similarity to the aligned sequences.

Proteome

The entire collection of proteins that are encoded by the genome of an organism. Initially the proteome is estimated by gene prediction and annotation methods but eventually will be revised as more information on the sequence of the expressed genes is obtained.

Pseudocounts

Small number of counts that is added to the columns of a scoring matrix to increase the variability either to avoid zero counts or to add more variation than was found in the sequences used to produce the matrix.

Receiver operator characteristic

The receiver operator characteristic (ROC) curve describes the probability that a test will correctly declare the condition present against the probability that the test will declare the condition present when actually absent. This is shown through a graph of the test's sensitivity against one minus the test's specificity for different possible threshold values.

Regular expressions
This computational tool provides a method for expressing the variations found in a set of related sequences including a range of choices at one position, insertions, repeats, and so on. For example, these expressions are used to characterize variations found in protein domains in the PROSITE catalog.

Regularization
A set of techniques for reducing data overfitting when training a model. See also Overfitting.

Relational database
Organizes information into tables where each column represents the fields of information that can be stored in a single record. Each row in the table corresponds to a single record. A single database can have many tables and a query language is used to access the data. See also Object-oriented database.

Scoring matrix
See Position-specific scoring matrix.

Selectivity (in database similarity searches)
The ability of a search method to locate members of a protein family without making a false-positive classification of members of other families.

Sensitivity (in database similarity searches)
The ability of a search method to locate as many members of a protein family as possible, including distant members of limited sequence similarity.

Significance
A significant result is one that has not simply occurred by chance, and therefore is probably true. Significance levels show how likely a result is due to chance, expressed as a probability. In sequence analysis, the significance of an alignment score may be calculated as the chance that such a score would be found between random or unrelated sequences. See Expect value.

Similarity score (sequence alignment)
The sum of the number of identical matches and conservative (high scoring) substitutions in a sequence alignment divided by the total number of aligned sequence characters. Gaps are usually ignored.

Simulated annealing
A search algorithm that attempts to solve the problem of finding global extrema. The algorithm was inspired by the physical cooling process of metals and the freezing process in liquids where atoms slow down in movement and line up to form a crystal. The algorithm traverses the energy levels of a function, always accepting energy levels that are smaller than previous ones, but sometimes accepting energy levels that are greater, according to the Boltzmann probability distribution.

Single-linkage cluster analysis
An analysis of a group of related objects, e.g., similar proteins in different genomes to identify both close and more distant relationships, represented on a tree or dendogram. The method joins the most closely related pairs by the neighbor-joining algorithm by representing these pairs as outer branches on the tree. More distant objects are then progressively added to lower tree branches. The method is also used to predict phylogenetic relationships by distance methods. See also Hierarchical clustering, Neighbor-joining method.

Smith-Waterman algorithm
Uses dynamic programming to find local alignments between sequences. The key feature is that all negative scores calculated in the dynamic programming matrix are

changed to zero in order to avoid extending poorly scoring alignments and to assist in identifying local alignments starting and stopping anywhere with the matrix.

Space or time complexity
An algorithm's complexity is the maximum amount of computer memory or time required for the number of algorithmic steps to solve a problem.

Specificity (in database similarity searches)
The ability of a search method to locate members of one protein family, including distantly related members.

Stochastic context-free grammar
A formal representation of groups of symbols in different parts of a sequence; i.e., not in the same context. An example is complementary regions in RNA that will form secondary structures. The stochastic feature introduces variability into such regions.

Stringency
Refers to the minimum number of matches required within a window. See also Filtering.

Sum of pairs method
Sums the substitution scores of all possible pair-wise combinations of sequence characters in one column of a multiple sequence alignment.

Synteny
The presence of a set of homologous genes in the same order on two genomes.

Threading
In protein structure prediction, the aligning of the sequence of a protein of unknown structure with a known three-dimensional structure to determine whether the amino acid sequence is spatially and chemically compatible with that structure.

Uncertainty
From information theory, a logarithmic measure of the average number of choices that must be made for identification purposes. See also Information content.

Unified Modeling Language (UML)
A standard sanctioned by the Object Management Group that provides a formal notation for describing object-oriented design.

Viterbi algorithm
Calculates the optimal path of a sequence through a hidden Markov model of sequences using a dynamic programming algorithm.

Weight matrix
See Position-specific scoring matrix.