附录: 生物信息学主要英文术语及释义

Abstract Syntax Notation (ASN.l)（**NCBI**发展的许多程序，如显示蛋白质三维结构的**Cn3D**等所使用的内部格式）
A language that is used to describe structured data types formally, Within bioinformatits,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（记录号）
A unique identifier that is assigned to a single database entry for a DNA or protein sequence.
Affine gap penalty（一种设置空位罚分策略）
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（算法）
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（联配/比对/联配）
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（联配/比对/联配值）
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 aftine 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（字母表）
The total number of symbols in a sequence-4 for DNA sequences and 20 for protein sequences.
Annotation（注释）
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（匿名FTP）

When a FTP service allows anyone to log in, it is said to provide anonymous FTP ser-vice. 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 O-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 O-127. ASCII tiles are commonly called plain text, meaning that they only encode text without extra markup.

BAC clone（细菌人工染色体克隆）

Bacterial artificial chromosome vector carrying a genomic DNA insert, typically 100–200 kb. Most of the large-insert clones sequenced in the project were BAC clones.

Back-propagation（反向传输）

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（Baum-Welch算法）

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

Baye's rule（贝叶斯法则）

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 condition-al probability of *A,* given B, P(AIB), is equal to the probability of A, P(A), times the conditional probability of B, given A, P(BIA), divided by the probability of B, *P(B). P(A)* is the historical or prior distribution value of *A, P(BIA)* 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(AIB)* 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（贝叶斯分析）

A statistical procedure used to estimate parameters of an underlying

distribution based on an observed distribution. See also Baye's rule.

Biochips（生物芯片）

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

Bioinformatics （生物信息学）

The merger of biotechnology and information technology with the goal of revealing new insights and principles in biology. /The discipline of obtaining information about genomic or protein sequence data. This may involve similarity searches of databases, comparing your unidentified sequence to the sequences in a database, or making predictions about the sequence based on current knowledge of similar sequences. Databases are frequently made publically available through the Internet, or locally at your institution.

Bit score （二进制值/ **Bit** 值）

The value S' is derived from the raw alignment score S in which the statistical properties of the scoring system used have been taken into account. Because bit scores have been normalized with respect to the scoring system, they can be used to compare alignment scores from different searches.

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, AJ, required to convey a message that has A4 possibilities is log2 *M* = N bits.

BLAST （基本局部联配搜索工具，一种主要数据库搜索程序）

Basic Local Alignment Search Tool. A set of programs, used to perform fast similarity searches. Nucleotide sequences can be compared with nucleotide sequences in a database using BLASTN, for example. Complex statistics are applied to judge the significance of each match. Reported sequences may be homologous to, or related to the query sequence. The BLASTP program is used to search a protein database for a match against a query protein sequence. There are several other flavours of BLAST. BLAST2 is a newer release of BLAST. Allows for insertions or deletions in the sequences being aligned. Gapped alignments may be more biologically significant.

Block（蛋白质家族中保守区域的组块）

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

BLOSUM matrices（模块替换矩阵，一种主要替换矩阵）

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（**Boltzmann** 分布）

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(Boltzmann概率函数)
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 deter-mine how much the sequence influences the results of an analysis.
Branch length（分支长度）
In sequence analysis, the number of sequence changes along a particular branch of a phylogenetic tree.
CDS or cds （编码序列）
Coding sequence.
Chebyshe, d 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.
Clone （克隆）
Population of identical cells or molecules (e.g. DNA), derived from a single ancestor.
Cloning Vector （克隆载体）
A molecule that carries a foreign gene into a host, and allows/facilitates the multiplication of that gene in a host. When sequencing a gene that has been cloned using a cloning vector (rather than by PCR), care should be taken not to include the cloning vector sequence when performing similarity searches. Plasmids, cosmids, phagemids, YACs and PACs are example types of cloning vectors.
Cluster analysis（聚类分析）
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（直系同源簇）

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

Comparative genomics（比较基因组学）

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)（算法的复杂性）

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（条件概率）

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).

Conservation （保守）

Changes at a specific position of an amino acid or (less commonly, DNA) sequence that preserve the physico-chemical properties of the original residue.

Consensus（一致序列）

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 （序列重叠群/拼接序列）

A set of clones that can be assembled into a linear order. A DNA sequence that overlaps with another contig. The full set of overlapping sequences (contigs) can be put together to obtain the sequence for a long region of DNA that cannot be sequenced in one run in a sequencing assay. Important in genetic mapping at the molecular level.

CORBA（国际对象管理协作组制定的使**OOP**对象与网络接口统一起来的一套跨计算机、操作系统、程序语言和网络的共同标准）

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（相关系数）

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)（共变）

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

Coverage (or depth) （覆盖率/厚度）

The average number of times a nucleotide is represented by a high-quality base in a collection of random raw sequence. Operationally, a 'high-quality base' is defined as one with an accuracy of at least 99% (corresponding to a PHRED score of at least 20).

Database（数据库）

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.

Depth （厚度）

See coverage

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 hid-den 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（序列距离）

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

DNA Sequencing （**DNA** 测序）

The experimental process of determining the nucleotide sequence of a region of DNA. This is done by labelling each nucleotide (A, C, G or T) with either a radioactive or fluorescent marker which identifies it. There are several methods of applying this technology, each with their advantages and disadvantages. For more information, refer to a current text book. High throughput laboratories frequently use automated sequencers, which are capable of rapidly reading large numbers of templates. Sometimes, the sequences may be generated more quickly than they can be characterised.

Domain （功能域）

A discrete portion of a protein assumed to fold independently of the rest of the protein and possessing its own function.

Dot matrix（点标矩阵图）
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.
Draft genome sequence （基因组序列草图）
The sequence produced by combining the information from the individual sequenced clones (by creating merged sequence contigs and then employing linking information to create scaffolds) and positioning the sequence along the physical map of the chromosomes.
DUST （一种低复杂性区段过滤程序）
A program for filtering low complexity regions from nucleic acid sequences.
Dynamic programming（动态规划法）
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.
EMBL （欧洲分子生物学实验室，**EMBL** 数据库是主要公共核酸序列数据库之一）
European Molecular Biology Laboratories. Maintain the EMBL database, one of the major public sequence databases.
EMBnet （欧洲分子生物学网络）
European Molecular Biology Network: http://www.embnet.org/ was established in 1988, and provides services including local molecular databases and software for molecular biologists in Europe. There are several large outposts of EMBnet, including EXPASY.
Entropy（熵）
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 out-comes. 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.
EST （表达序列标签的缩写）

See Expressed Sequence Tag

Expect value (E)（E值）

E value. The number of different alignents with scores equivalent to or better than S that are expected to occur in a database search by chance. The lower the E value, the more significant the score. 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.

Exon （外显子）

Coding region of DNA. See CDS.

Expressed Sequence Tag (EST) （表达序列标签）

Randomly selected, partial cDNA sequence; represents it's corresponding mRNA. dbEST is a large database of ESTs at GenBank, NCBI.

FASTA （一种主要数据库搜索程序）

The first widely used algorithm for database similarity searching. The program looks for optimal local alignments by scanning the sequence for small matches called "words". Initially, the scores of segments in which there are multiple word hits are calculated ("init1"). Later the scores of several segments may be summed to generate an "initn" score. An optimized alignment that includes gaps is shown in the output as "opt". The sensitivity and speed of the search are inversely related and controlled by the "k-tup" variable which specifies the size of a "word". (Pearson and Lipman)

Extreme value distribution（极值分布）

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（假阴性）

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 （假阳性）
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 （反向传输神经网络）
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.

Filtering （过滤）
Also known as Masking. The process of hiding regions of (nucleic acid or amino acid) sequence having characteristics that frequently lead to spurious high scores. See SEG and DUST.

Finished sequence（完成序列）
Complete sequence of a clone or genome, with an accuracy of at least 99.99% and no gaps.

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

Format (file)（格式）
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)（文件传输协议）
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 tile system available for FTP access, providing that the client is able to supply a recognized user name and password to the server.

Full shotgun clone （鸟枪法克隆）
A large-insert clone for which full shotgun sequence has been produced.

Functional genomics（功能基因组学）
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 （空位/间隙/缺口）
A space introduced into an alignment to compensate for insertions and deletions in one sequence relative to another. To prevent the accumulation of too many gaps in an alignment, introduction of a gap causes the deduction of a fixed amount (the gap score) from the alignment score. Extension of the gap to encompass additional nucleotides or amino acid is also penalized in the scoring of an alignment.
Gap penalty（空位罚分）
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（遗传算法）
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.
Genetic map （遗传图谱）
A genome map in which polymorphic loci are positioned relative to one another on the basis of the frequency with which they recombine during meiosis. The unit of distance is centimorgans (cM), denoting a 1% chance of recombination.
Genome（基因组）
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（整体联配）
Attempts to match as many characters as possible, from end to end, in a set of two or

more sequences.

Gopher (一个文档发布系统，允许检索和显示文本文件**)**

Graph theory（图论）

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.

GSS（基因综述序列）

Genome survey sequence.

GUI（图形用户界面）

Graphical user interface.

H （相对熵值）

H is the relative entropy of the target and background residue frequencies. (Karlin and Altschul, 1990). H can be thought of as a measure of the average information (in bits) available per position that distinguishes an alignment from chance. At high values of H, short alignments can be distinguished by chance, whereas at lower H values, a longer alignment may be necessary. (Altschul, 1991)

Half-bits

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

Heuristic（启发式方法）

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（**16**制系统）

The base 16 counting system that uses the digits O-9 followed by the letters A-F.

HGMP （人类基因组图谱计划）

Human Genome Mapping Project.

Hidden Markov Model (HMM)（隐马尔可夫模型）

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（隐藏层）

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（分级聚类）

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.

Homology（同源性）

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（水平转移）

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 com-pared to the rest of the genome.

HSP （高比值片段对）

High-scoring segment pair. Local alignments with no gaps that achieve one of the top alignment scores in a given search.

HTGS/HGT（高通量基因组序列）

High-throughout genome sequences

HTML（超文本标识语言）

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.

Identity （相同性/相同率）

The extent to which two (nucleotide or amino acid) sequences are invariant.

Indel（插入或删除的缩略语）

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 max-imum possible number is 2, and for proteins, 4.32 (logarithm to the base 2 of the number of possible sequence characters).

Information theory（信息理论）

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（输入层）

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（因特网）

The network infrastructure, consisting of cables interconnected by routers, that pro-vides 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.

Intranet（内部网）

Intron （内含子）

Non-coding region of DNA.

Iterative（反复的/迭代的）

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

Java（一种由 **SUN Microsystem** 开发的编程语言）

K （**BLAST** 程序的一个统计参数）

A statistical parameter used in calculating BLAST scores that can be thought of as a natural scale for search space size. The value K is used in converting a raw score (S) to a bit score (S').

K-tuple（字/字长）

Identical short stretches of sequences, also called words.

lambda （λ，**BLAST** 程序的一个统计参数）

A statistical parameter used in calculating BLAST scores that can be thought of as a natural scale for scoring system. The value lambda is used in converting a raw score (S) to a bit score (S').

LAN（局域网）

Local area network.

Likelihood（似然性）

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 pointsin a location that best separates the data points into two groups.

Local alignment（局部联配）

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（概率对数值）

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

Low Complexity Region (LCR) （低复杂性区段）

Regions of biased composition including homopolymeric runs, short-period repeats, and more subtle overrepresentation of one or a few residues. The SEG program is used to mask or filter LCRs in amino acid queries. The DUST program is used to mask or filter LCRs in nucleic acid queries.

Machine learning（机器学习）

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

Markov chain（马尔可夫链）

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 sj given that the current state is sp 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.

Masking （过滤）

Also known as Filtering. The removal of repeated or low complexity regions from a sequence in order to improve the sensitivity of sequence similarity searches performed with that sequence.

Maximum likelihood (phylogeny, alignment)（最大似然法）

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

Maximum parsimony（最大简约法）

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 mini-mum 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.

MMDB （分子建模数据库）

Molecular Modelling Database. A taxonomy assigned database of PDB (see PDB) files, and related information.

Molecular clock hypothesis（分子钟假设）

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

Monte Carlo（蒙特卡罗法）

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

Motif （模序）

A short conserved region in a protein sequence. Motifs are frequently highly conserved parts of domains.

Multiple Sequence Alignment （多序列联配）

An alignment of three or more sequences with gaps inserted in the sequences such that residues with common structural positions and/or ancestral residues are aligned in the same column. Clustal W is one of the most widely used multiple sequence alignment programs

Mutation data matrix（突变数据矩阵，即**PAM**矩阵）

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.

N50 length （**N50** 长度，即覆盖 **50%**所有核苷酸的最大序列重叠群长度）

A measure of the contig length (or scaffold length) containing a 'typical' nucleotide. Specifically, it is the maximum length L such that 50% of all nucleotides lie in contigs (or scaffolds) of size at least L.

Nats (natural logarithm)

A number expressed in units of the natural logarithm.

NCBI （美国国家生物技术信息中心）

National Center for Biotechnology Information (USA). Created by the United States Congress in 1988, to develop information systems to support the

biological research community.

Needleman-Wunsch algorithm（Needleman-Wunsch算法）

Uses dynamic programming to find global alignments between sequences.

Neighbor-joining method（邻接法）

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（神经网络）

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.

NIH （美国国家卫生研究院）

National Institutes of Health (USA).

Noise（噪音）

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（正态分布）

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)（国际对象管理协作组）

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 inter-operability among computational resources in life sciences research. See CORBA.

Object-oriented database（面向对象数据库）

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（概率/几率值）
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.

OMIM （一种人类遗传疾病数据库）
Online Mendelian Inheritance in Man. Database of genetic diseases with references to molecular medicine, cell biology, biochemistry and clinical details of the diseases.

Optimal alignment（最佳联配）
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.

ORF （开放阅读框）
Open Reading Frame. A series of codons (base triplets) which can be translated into a protein. There are six potential reading frames of an unidentifed sequence; TBLASTN (see BLAST) transalates a nucleotide sequence in all six reading frames, into a protein, then attempts to align the results to sequeneces in a protein database, returning the results as a nucleotide sequence. The most likely reading frame can be identified using on-line software (e.g. ORF Finder).

Orthologous（直系同源）
Homologous sequences in different species that arose from a common ancestral gene during speciation; may or may not be responsible for a similar function. 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 1 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 11 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（输出层）
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 outpu t 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 hid-den 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.

P value （P 值/概率值）

The probability of an alignment occurring with the score in question or better. The p value is calculated by relating the observed alignment score, S, to the expected distribution of HSP scores from comparisons of random sequences of the same length and composition as the query to the database. The most highly significant P values will be those close to 0. P values and E values are different ways of representing the significance of the alignment.

Pair-wise sequence alignment（双序列联配）

An alignment performed between two sequences.

PAM（可接受突变百分率/可以观察到的突变百分率，它可作为一种进化时间单位）

Percent Accepted Mutation. A unit introduced by Dayhoff et al. to quantify the amount of evolutionary change in a protein sequence. 1.0 PAM unit, is the amount of evolution which will change, on average, 1% of amino acids in a protein sequence. A PAM(x) substitution matrix is a look-up table in which scores for each amino acid substitution have been calculated based on the frequency of that substitution in closely related proteins that have experienced a certain amount (x) of evolutionary divergence.

Paralogous（旁系同源）

Homologous sequences within a single species that arose by gene duplication. 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 entireproteome 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.

PDB（主要蛋白质结构数据库之一）

Brookhaven Protein Data Bank. A database and format of files which describe the 3D structure of a protein or nucleic acid, as determined by X-ray crystallography or nuclear magnetic resonance (NMR) imaging. The

molecules described by the files are usually viewed locally by dedicated software, but can sometimes be visualised on the world wide web.

Pearson correlation coefficent（Pearson相关系数）

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（相似百分率）

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（感知器，模拟人类视神经控制系统的图形识别机)

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

PHRED（一种广泛应用的原始序列分析程序，可以对序列的各个碱基进行识别和质量评价）

A widely used computer program that analyses raw sequence to produce a 'base call' with an associated 'quality score' for each position in the sequence. A PHRED quality score of X corresponds to an error probability of approximately 10-X/10. Thus, a PHRED quality score of 30 corresponds to 99.9% accuracy for the base call in the raw read.

PHRAP （一种广泛应用的原始序列组装程序）

A widely used computer program that assembles raw sequence into sequence contigs and assigns to each position in the sequence an associated 'quality score', on the basis of the PHRED scores of the raw sequence reads. A PHRAP quality score of X corresponds to an error probability of approximately 10-X/10. Thus, a PHRAP quality score of 30 corresponds to 99.9% accuracy for a base in the assembled sequence.

Phylogenetic studies（系统发育研究）

PIR （主要蛋白质序列数据库之一，翻译自 GenBank）

A database of translated GenBank nucleotide sequences. PIR is a redundant (see Redundancy) protein sequence database. The database is divided into four categories:

PIR1 - Classified and annotated.

PIR2 - Annotated.

PIR3 - Unverified.

PIR4 - Unencoded or untranslated.

Poisson distribution（帕松分布）

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)（特定位点记分矩阵，**PSI-BLAST** 等搜索程序使用）

The PSSM gives the log-odds score for finding a particular matching amino acid in a target sequence. 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（分布型）

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. See also PSSM.

Profile hidden Markov model（分布型隐马尔可夫模型）

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（蛋白质组）

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.

Proteomics （蛋白质组学）

Systematic analysis of protein expression of normal and diseased tissues that involves the separation, identification and characterization of all of the proteins in an organism.

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.

PSI-BLAST （BLAST 系列程序之一）
Position-Specific Iterative BLAST. An iterative search using the BLAST algorithm. A profile is built after the initial search, which is then used in subsequent searches. The process may be repeated, if desired with new sequences found in each cycle used to refine the profile. Details can be found in this discussion of PSI-BLAST. (Altschul et al.)
PSSM （特定位点记分矩阵）
See position-specific scoring matrix and profile.
Public sequence databases （公共序列数据库，指 GenBank、EMBL 和 DDBJ）
The three coordinated international sequence databases: GenBank, the EMBL data library and DDBJ.
Q20 (Quality score 20)
A quality score of > or = 20 indicates that there is less than a 1 in 100 chance that the base call is incorrect. These are consequently high-quality bases. Specifically, the quality value "q" assigned to a basecall is defined as:
q = -10 x log10(p)
where p is the estimated error probability for that basecall. Note that high quality values correspond to low error probabilities, and conversely.
Quality trimming
This is an algorithm which uses a sliding window of 50 bases and trims from the 5' end of the read followed by the 3' end. With each window, the number of low quality (10 or less) bases is determined. If more than 5 bases are below the threshold quality, the window is incremented by one base and the process is repeated. When the low quality test fails, the position where it stopped is recorded. The parameters for window length low quality threshold and number of low quality bases tolerated are fixed. The positions of the 5' and 3' boundaries of the quality region are noted in the plot of quality values presented in the" Chromatogram Details" report.
Query （待查序列/搜索序列）
The input sequence (or other type of search term) with which all of the entries in a database are to be compared.
Radiation hybrid (RH) map （辐射杂交图谱）
A genome map in which STSs are positioned relative to one another on the basis of the frequency with which they are separated by radiation-induced breaks. The frequency is assayed by analysing a panel of human–hamster hybrid cell lines, each produced by lethally irradiating human cells and fusing them with recipient hamster cells such that each carries a collection of human chromosomal fragments. The unit of distance is centirays (cR), denoting a 1% chance of a break occuring between two loci
Raw Score （初值，指最初得到的联配值 S）
The score of an alignment, S, calculated as the sum of substitution and gap scores. Substitution scores are given by a look-up table (see PAM, BLOSUM). Gap scores are typically calculated as the sum of G, the gap opening penalty

and L, the gap extension penalty. For a gap of length n, the gap cost would be G+Ln. The choice of gap costs, G and L is empirical, but it is customary to choose a high value for G (10-15)and a low value for L (1-2).

Raw sequence （原始序列/读胶序列）

Individual unassembled sequence reads, produced by sequencing of clones containing DNA inserts.

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 tesls sensitivity against one minus the test specificity for different possible threshold values.

Redundancy （冗余）

The presence of more than one identical item represents redundancy. In bioinformatics, the term is used with reference to the sequences in a sequence database. If a database is described as being redundant, more than one identical (redundant) sequence may be found. If the database is said to be non-redundant (nr), the database managers have attempted to reduce the redundancy. The term is ambiguous with reference to genetics, and as such, the degree of non-redundancy varies according to the database manager's interpretation of the term. One can argue whether or not two alleles of a locus defines the limit of redundancy, or whether the same locus in different, closely related organisms constitutes redundency. Non-redundant databases are, in some ways, superior, but are less complete. These factors should be taken into consideration when selecting a database to search.

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（关系数据库）

Organizes information into tables where each column represents the fields of informa-tion 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.

Scaffold （支架，由序列重叠群拼接而成）

The result of connecting contigs by linking information from paired-end reads from plasmids, paired-end reads from BACs, known messenger RNAs or other sources. The contigs in a scaffold are ordered and oriented with respect to one another.

Scoring matrix（记分矩阵）
See Position-specific scoring matrix.
SEG （一种蛋白质程序低复杂性区段过滤程序）
A program for filtering low complexity regions in amino acid sequences. Residues that have been masked are represented as "X" in an alignment. SEG filtering is performed by default in the blastp subroutine of BLAST 2.0. (Wootton and Federhen)
Selectivity (in database similarity searches)（数据库相似性搜索的选择准确性）
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)（数据库相似性搜索的灵敏性）
The ability of a search method to locate as many members of a protein family as possi-ble, including distant members of limited sequence similarity.
Sequence Tagged Site （序列标签位点）
Short cDNA sequences of regions that have been physically mapped. STSs provide unique landmarks, or identifiers, throughout the genome. Useful as a framework for further sequencing.
Significance（显著水平）
A significant result is one that has not simply occurred by chance, and therefore is prob-ably 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 calcu-lated as the chance that such a score would be found between random or unrelated sequences. See Expect value.
Similarity score (sequence alignment) （相似性值）
Similarity means the extent to which nucleotide or protein sequences are related. The extent of similarity between two sequences can be based on percent sequence identity and/or conservation. In BLAST similarity refers to a positive matrix score. The sum of the number of identical matches and conservative (high scoring) substitu-tions in a sequence alignment divided by the total number of aligned sequence charac-ters. 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 pro-gressively added to lower tree branches. The method is also used to predict phylogenet-ic relationships by distance methods. See also Hierarchical clustering, Neighbor-joining method.

Smith-Waterman algorithm（Smith-Waterman算法）

Uses dynamic programming to find local alignments between sequences. The key fea-ture 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.

SNP （单核苷酸多态性）

Single nucleotide polymorphism, or a single nucleotide position in the genome sequence for which two or more alternative alleles are present at appreciable frequency (traditionally, at least 1%) in the human population.

Space or time complexity（时间或空间复杂性）

An algorithms 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)（数据库相似性搜索的特异性）

The ability of a search method to locate members of one protein family, including dis-tantly related members.

SSR （简单序列重复）

Simple sequence repeat, a sequence consisting largely of a tandem repeat of a specific k-mer (such as (CA)15). Many SSRs are polymorphic and have been widely used in genetic mapping.

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 sec-ondary
structures. The stochastic feature introduces variability into such regions.

Stringency

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

STS （序列标签位点的缩写）

See Sequence Tagged Site

Substitution （替换）

The presence of a non-identical amino acid at a given position in an alignment. If the aligned residues have similar physico-chemical properties the substitution is said to be "conservative".

Substitution Matrix （替换矩阵）

A substitution matrix containing values proportional to the probability that amino acid i mutates into amino acid j for all pairs of amino acids. such matrices are constructed by assembling a large and diverse sample of verified pairwise alignments of amino acids. If the sample is large enough to be statistically significant, the resulting matrices should reflect the true probabilities of mutations occuring through a period of evolution.

Sum of pairs method
Sums the substitution scores of all possible pair-wise combinations of sequence charac-ters in one column of a multiple sequence alignment.
SWISS-PROT （主要蛋白质序列数据库之一）
A non-redundant (See Redundancy) protein sequence database. Thoroughly annotated and cross referenced. A subdivision is TrEMBL.
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.
TrEMBL （蛋白质数据库之一，翻译自 **EMBL**）
A protein sequence database of Translated EMBL nucleotide sequences.
Uncertainty（不确定性）
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 nota-tion for describing object-oriented design.
UniGene （人类基因数据库之一）
Database of unique human genes, at NCBI. Entries are selected by near identical presence in GenBank and dbEST databases. The clusters of sequences produced are considered to represent a single gene.
Unitary Matrix （一元矩阵）
Also known as Identity Matrix. A scoring system in which only identical characters receive a positive score.
URL（统一资源定位符）
Uniform resource locator.
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.