



Glossary...

- Affine gap costs** – A scoring system for gaps within alignments that charges a penalty for the existence of a gap and an additional per-residue penalty proportional to the gap's length.
- Algorithm** – A fixed procedure, embodied in a computer program.
- Alignment score** – A numerical value that describes the overall quality of an alignment. Higher numbers correspond to higher similarity.
- Batch Entrez** – A feature in Entrez that allows the retrieval of many sequences at once and saves the sequences to a file on a local computer. This is particularly useful if you want to download a set of sequences to analyze locally (for example, to do a multiple alignment on a local computer).
- bfind mode** – One of two modes used by DBGET. Keywords can be entered in a search box. See 'bget' mode.
- bget mode** – One of two modes used by DBGET. Entry name or the accession number of an entry can be entered in a search box. See 'bfind mode'.
- Biochemical pathway** – A network of interacting molecules that is responsible for a specific biochemical function, such as a metabolic pathway or a signal transduction pathway.
- Bit score** – A scaled version of an alignment's raw score that accounts for the statistical properties of the scoring system used.
- BLAST** – Basic Local Alignment Search Tool. A heuristic sequence comparison algorithm, developed by researchers at the National Center for Biotechnology Information (NCBI) and others, that is used to search sequence databases for optimal local alignments to a query.
- Bootstrapping** – A statistical method that is often used to estimate the reproducibility of specific features of phylogenetic trees.
- Cluster analysis** – A process of assigning data points (sequences) into groups (clusters), starting from pairwise distances. Useful for identifying outliers and weak links between groups. Fairly easy to do by hand for small datasets.
- Command line** – Interacting with software by typing specific commands. Generally considered less 'user friendly' than a 'graphical user interface'.
- Comparative genomics** – The study of comparing complete genome sequences, often by computational methods, to understand general principles of genome structure and function.
- Content** – An extended or variable-length region of genomic DNA with a particular function, such as an exon.
- Controlled vocabulary** – A vocabulary that contains specific words that are consistently applied to all entries in a database. The MeSH system is an example of a controlled vocabulary.
- Deductive database** – A database that contains both facts (often in the form of a relational database) and rules for reasoning (often in logic programming) so that new facts can be dynamically generated from stored facts.
- DNA chip technology** – New technology for parallel processing thousands of DNA segments, such as for detecting mutation patterns in genomic DNAs or expression patterns of mRNAs.
- Domain** – A portion of a protein that folds independently of the rest of the protein, or is at least assumed to do so.
- DUST** – Program for filtering low-complexity regions of DNA structure.
- Dynamic programming** – A type of algorithm widely used for constructing sequence alignments and for evaluating all possible candidate gene structures.
- E value** – Expectation value. The number of distinct alignments, with score equivalent to or better than the one of interest, that are expected to occur in a database search purely by chance. The lower the E value, the more significant the score is.
- EST** – Expressed Sequence Tag. A short cDNA (complementary DNA) sequence.
- Extreme value distribution** – The probability distribution applicable to the scores of optimal local alignments.
- Family, subfamily, superfamily** – Family groups, sequences or domains that are clearly related and usually have a similar function. A superfamily groups several families that are related by (divergent) evolution and usually still share some functional elements. A subfamily groups sequences within a family that are particularly closely related. There is no truly accepted consensus on the use of these terms.
- FASTA** – A popular heuristic sequence comparison algorithm, devised by W. Pearson and D. Lipman, that is used to search sequence databases for optimal local alignments to a query.
- Filtering** – See 'masking'.
- Fold** – The overall folding pattern of a 3-D protein structure.
- Free-form text** – The opposite of a 'controlled vocabulary'. Free text has no structured set of words, such that two related entries might not be identified in a search because different words are used to describe each entry.
- ftp** – File Transfer Protocol. A mechanism for transferring files across a network.
- Functional genomics** – The study of obtaining an overall picture of genome functions, including the expression profiles at the mRNA level (transcriptome) and the protein level (proteome).
- Gap** – Within an alignment of two sequences, several adjacent null characters in one sequence aligned with adjacent letters in the other.
- Gap score** – The score assigned to a gap.
- Gapped alignment** – An alignment in which gaps are permitted.
- Gene family** – Two or more genes that are related by divergent evolution from a common ancestor, either by speciation or gene duplication.
- Global alignment** – The alignment of two complete nucleic acid or protein sequences.
- Graphical user interface** – Software that allows a user to interact via 'user-friendly' menu and mouse-driven commands, as is typical of Macintosh and Windows applications and less common for UNIX applications; as opposed to a 'command line' interface of typed or scripted commands.
- GSS** – Genome Survey Sequence. Includes single-pass genomic data, exon-trapped sequences and Alu PCR sequences.
- Heuristics** – A term in computer science that refers to 'guesses' made by a program to obtain approximately accurate results. Typically, these are used to increase the speed of a program greatly at the cost of potentially yielding suboptimal results. BLAST and FASTA use heuristics based on knowledge of how sequences evolve.
- High-throughput DNA sequencing** – Experimental procedures for determining massive amounts of genomic DNA or cDNA sequence data using highly automated sequencing machines.
- HMM** – Hidden Markov Model. The extension of a Markov model. A pattern recognition method that can be used to represent the alignment of multiple sequences or sequence segments by attempting to capture common patterns of residue conservation.
- HMNER** – A software package for profile hidden Markov model analysis.
- Homology domain** – A region in a protein sequence with similarity to an otherwise unrelated protein. This term should be used only if the region is of a size sufficient to form a domain. A homology domain can contain several motifs.
- HTGS** – High-throughput Genomic Sequence. A short segment of 'unfinished' genomic sequence.
- Iterative search** – After performing an initial search against the database, the high scoring matches are used to search the database again. In some cases (intermediate sequence path), these sequences are used on their own; in others, the sequences are joined together in an alignment or profile.
- ktup** – Parameter to FASTA that affects speed and sensitivity. The ktup = 1 setting is slower and more sensitive than ktup = 2.
- Links** – A feature used by Entrez to identify associated entries in other databases. For example, for a given sequence, there are links to the literature.
- Linux** – A freely available but commercial-strength clone of the UNIX operating system. A godsend for starting bioinformatics on a budget. It is easily installed alongside Windows on a PC, so the same machine can be booted into either Linux or Windows.
- Local alignment** – The alignment of segments from two nucleic acid or protein sequences.
- Long branch attraction** – The artifactual placement of rapidly evolving sequences with other rapidly evolving sequences and slowly evolving sequences with other slowly evolving sequences in phylogenetic trees. This placement is independent of the true phylogenetic relationships.



- Low-complexity region** – A region of a nucleic acid or protein sequence with highly biased residue composition, or consisting of many short near-perfect repeats.
- Markov model** – A statistical model for sequences in which the probability of each letter depends on the letters that precede it.
- Masking** – Some regions of sequences have particular characteristics (such as repeated patterns) that lead to spurious high scores. Masking replaces these regions of sequence with an 'X' (for proteins) or 'N' (for nucleic acids).
- MEDLINE** – A free on-line literature database of papers in biomedical sciences (see <http://www.ncbi.nlm.nih.gov/Entrez/medline.html>).
- MeSH** – Medical Subject Headings. A controlled vocabulary of medical and scientific terms assigned in a consistent manner to documents in the MEDLINE database. Using these terms can make searching more efficient than searching through free text.
- Meta-analysis** – Mathematical and statistical methods to combine findings from independent studies to obtain more robust or reliable conclusions.
- Motif** – A short conserved region in a protein sequence. Motifs frequently form a recognition sequence or are highly conserved parts of domains. Motif is sometimes used in a broader sense for all localized homology regions, independent of their size.
- Motif descriptor** – A data structure that stores information about a sequence family, motif or domain family. Typical examples are consensus sequences, patterns, profiles and HMMs.
- mRNA expression profile** – The identities and absolute or relative expression levels of mRNAs that characterize a particular cell type or physiological, developmental or pathological state.
- Multiple alignment** – An alignment of three or more sequences, with gaps (spaces) inserted in the sequences such that residues with common structural positions and/or ancestral residues are aligned in the same column of the multiple alignment.
- Needleman–Wunsch algorithm** – The standard dynamic programming algorithm for finding optimal global alignments.
- Neighbors** – A feature used by Entrez to identify related entries within a single database. For example, for a given sequence, neighbors are similar sequences.
- Neural net** – A statistical pattern recognition method; a type of nonlinear regression.
- nr database** – Nonredundant database of protein or DNA sequences available from the NCBI BLAST Web site.
- Null character** – A character, usually represented by '-', inserted into one sequence to allow it to be aligned with another.
- Oligonucleotide chips** – High-density arrays of oligonucleotide probes that are synthesized by manufacturing techniques similar to those for silicon chips and that are used to detect, for example, variations of DNA sequence patterns by hybridization studies.
- Optimal alignment** – A global or local alignment of two sequences with the highest possible score.
- Orthologs** – Homologous sequences in different species that arose from a common ancestor gene during speciation; may or may not be responsible for a similar function. See 'paralogs'.
- P value** – Similar to an *E* value, the probability of a hit occurring by chance with this score or better, as opposed to the expected number of hits. A *P* value has a maximum of 1.0, whereas an *E* value has, as a maximum, the number of sequences in the database that was searched. For small (significant) *P* values, *P* and *E* are approximately equal, so the choice of one or the other in a software package is arbitrary. NCBI BLAST 2.0, FASTA and HMMER report *E* values. WU-BLAST 2.0 reports *P* values.
- Paralinear (logdet) distances** – An algorithm, based on a general mathematical model of evolution, that can accurately estimate the distance between two aligned sequences. It is unaffected by one of the three principal artefacts of phylogenetic reconstruction (unequal rate effects).
- Paralogs** – Homologous sequences (that is, sequences that share a common evolutionary ancestor) that diverged by gene duplication. See 'orthologs'.
- Pattern** – A descriptor for short sequence motifs, consisting of amino acid characters and meta-characters that can represent ambiguities or variable length insertions.
- pdb database** – Protein Data Bank. The repository of solved protein structures.
- Profile** – A linear model representing characteristics of a sequence group, the simplest form of which is the tabulation of the frequency of amino acids (or nucleotides) in each position of the multiple sequence alignment. Also called 'position-specific scoring matrices'. Profiles that do not allow insertions and deletions are also called 'weight matrices'.
- Proteomics** – Technically and conceptually similar to functional genomics, but with the aim of studying biological aspects of all proteins at once in a systematic manner.
- PSI-BLAST** – Position-specific Iterated BLAST. An iterative search that uses the BLAST algorithm to provide fast searches, and builds a profile at every iteration.
- Raw score** – The score of an alignment, usually defined as the sum of substitution and gap scores.
- Regular expression** – A text pattern that conforms to regular grammar and that is used for text pattern matching in the UNIX system, as well as for representing consensus sequence patterns in biology.
- Rooted tree** – A phylogenetic tree in which the last common ancestor of all genes, or organisms, displayed on the tree is specified by the initial bifurcation of the tree.
- Score** – In the context of computational genefinding, a function used to evaluate different candidate gene predictions; dynamic programming finds the candidate gene structure with the best score.
- SEG** – Program for filtering low-complexity regions of DNA structure.
- Sequence motif** – A short conserved amino acid (or nucleotide) sequence pattern that represents a specific functional site of a protein (or nucleic acid) molecule.
- Signal** – A local functional site in genomic DNA, such as a splice site.
- Smith–Waterman algorithm** – The standard dynamic programming algorithm for finding optimal local alignments.
- SSEARCH** – Implementation of the Smith–Waterman algorithm, provided in the FASTA package with empirical statistical estimates.
- STS** – Sequence-tagged Site. A short genomic landmark sequence.
- Subfamily** – see 'family'.
- Substitution matrix** – The collection of all substitution scores.
- Substitution score** – The score for aligning a particular pair of letters.
- Superfamily** – see 'family'.
- SWISS-PROT** – A curated protein sequence database that provides a high level of annotations, a minimal level of redundancy and a high level of integration with other databases. It is maintained collaboratively by the Department of Medical Biochemistry at the University of Geneva and the European Bioinformatics Institute (EBI).
- Target frequencies** – Among alignments representing true biological relationships, the frequencies with which the various letter pairs are aligned. Target frequencies vary with the degree of evolutionary divergence separating homologous sequences.
- Taxon (pl. taxa)** – A group of one or more organisms. The term is often applied to the organisms represented by the terminal branches of a tree.
- Ungapped alignment** – An alignment in which gaps are not permitted.
- UNIX** – A computer operating system.
- Unrooted tree** – A phylogenetic tree in which the last common ancestor of all genes, or organisms, on the tree is not specified.
- Weight matrix** – A statistical model in which each position in a sequence is modeled with a separate, independent probability distribution.
- Wormpep** – A periodically updated compilation of the protein sequences for predicted *Caenorhabditis elegans* genes from the nematode genome project.
- WU-BLASTP** – The Washington University version of gapped BLASTP.
- Z value** – The number of standard deviations from the mean.