Glossary of Bioinformatics Terms

This collection of terms and definitions commonly encountered in the bioinformatics literature will be updated periodically as Current Protocols in Bioinformatics grows. In addition, an extensive glossary of genetic terms can be found on the Web site of the National Human Genome Research Institute (http://www.genome.gov/glossary.cfm). The entries in that online glossary provide a brief written definition of the term; the user can also listen to an informative explanation of the term using RealAudio or the Windows Media Player.

TERMS

2D gel electrophoresis: Two-dimensional polyacrylamide electrophoresis. A technique for separating large numbers of proteins by loading them onto a gel, applying an electric current to separate by pI (see isoelectric point), then separating them by molecular weight in the perpendicular direction.

Accessible surface area: The surface area of a protein or macromolecule that can be contacted by water molecules or other solvent or solute molecules. Accessible surface area is measured in square Angstroms (Å²) and is often used to assess the quality of protein structures and the strength of hydrophobic interactions.

Algorithm: Any sequence of actions (e.g., computational steps) that perform a particular task.

Alignment: Two or more sequences that have been lined up, based on some algorithm. Sequence-based alignment algorithms attempt to match as many identical residues or conservatively substituted positions as possible. Sequences can also be aligned based on their three-dimensional structures.

Allele: Any of the forms of a gene that may occur at a given locus.

Alternative splicing: The process through which a cell can generate many different protein products from a single gene. A gene is transcribed into a primary RNA transcript, containing both exons and introns; these exons (and sometimes introns) can then be combined into one or more different mRNA molecules, each encoding for a different protein.

Analogous: In phylogenetics, characters that have the same role but are not related by ancestry.

ASN.1: Abstract syntax notation one. ASN.1 is a formal language for abstractly describing messages or information to be exchanged. ASN.1 is used extensively by the National Center for Biotechnology Information in representing sequence, structure, interaction, mapping, and bibliographic records.

Assembler: A program designed to deduce the sequence of original DNA used to make a shotgun library from a set of shotgun reads. Also see shotgun sequencing.

Average mass: The mass of an ion calculated by averaging all common isotope variations. This quantity typically is used when the resolution of the instrumentation is not adequate to distinguish individual isotopes.

BAC: Bacterial artificial chromosome. A vector used to clone segments of DNA roughly 100 to 200 kb in length. Also see YAC.
**Base caller:** A program used to convert raw sequencer output to an ordered list of base identities and quality scores. Also see chromatogram.

**Bayesian network:** A method in machine learning used to predict a feature in a dataset given some known but incomplete information. Based on Bayes’ rule for conditional probability.

**BLOSUM matrix:** Blocks substitution matrix. See PAM matrix.

**Bonferroni correction:** A conservative statistical multiple comparison correction used to correct the significance threshold (α value) of multiple independent statistical tests, which together may have a greater chance of generating false positives than the individual test.

**Boolean:** Refers to an expression or variable that can have only a true or false value. Named after George Boole, a British mathematician who developed the theory of algebraic logic or Boolean algebra, which is now used in almost all electronic computation.

**Bottom-up proteomics:** Another term for shotgun proteomics (see below), referring to the fact that the analysis begins with the peptide constituents of a sample.

**Browser:** Program used to access sites on the Internet. Using hypertext markup language (HTML), browsers are capable of representing a Web page the same way, regardless of computer platform.

**Candidate gene:** A gene that is implicated in the causation of a disease or phenotype. Candidate genes lie in a region that has been identified through genetic mapping. The protein product of a candidate gene may implicate the candidate gene as being the actual disease gene being sought.

**Captured gap:** In sequencing, a gap that is contained within one or more subclones.

**cDNA:** Complementary DNA. Single-stranded DNA that has been synthesized from an mRNA template by reverse transcriptase.

**cDNA library:** A collection of double-stranded DNA sequences that are generated by copying mRNA molecules. Because these sequences are derived from mRNAs, they contain primarily protein-coding DNA.

**CDS:** Coding sequence. A segment of genomic DNA or mRNA that codes for a protein. This abbreviation is used extensively in sequence database records.

**Characters and character states:** In phylogenetics, characters are homologous features in different organisms. The exact condition of that feature in a particular individual is the character state. As an example, the character “hair color” can have the character states “gold,” “red,” and “yellow.” In molecular biology, the character states can be one of the four nucleotides (A, C, T, G) or one of the twenty amino acids. Note that some authors define character to mean the character state as defined here.

**Chimeric read:** In sequencing, a read-containing sequence from two noncontiguous regions of the target or vector. Chimeric reads can be the result of multiple inserts ligating into the same vector during library construction, or sequence from a mixture of two clones that have regions in which each of the clones is more obvious.

**Chromatogram:** A file containing raw data and ancillary information about a single DNA sample that has been run through an automated DNA sequencing instrument. Also see base caller.
**Client**: A computer, or the software running on a computer, that interacts with another computer at a remote site (server). Note the difference between client and user (see below).

**cM**: Centimorgan. The genetic distance between two markers that recombine at a frequency of 1%. In humans, 1 cM is approximately equivalent to $1 \times 10^6$ bp.

**Coding statistic**: A mathematical function that computes a real number related to the likelihood that a given DNA sequence codes for a protein.

**Codon**: The triplet of bases in either a DNA or RNA sequence that ultimately codes for a specific, single amino acid. See also *stop codon*.

**Comparative proteomics**: The general approach of comparing proteomes from two or more cellular states, then using phylogenetic techniques or mass spectrometry to identify the proteins or peptides that differ between them.

**Complementary**: Two DNA or RNA sequences that can form uninterrupted base pairs.

**Consensus**: In sequencing, the predicted sequence of the original DNA used to create a shotgun library. In alignments, the base or amino acid most likely to occur at any given position; consensus sequences can be used to characterize protein families. Also see *shotgun sequencing*.

**Conservative substitution**: The replacement of one residue by another having similar properties (e.g., size, charge, hydrophobicity).

**Contig**: Short for *contiguous*. Refers to a contiguous set of overlapping DNA sequences in the context of a sequencing project.

**Cytogenetic map**: The representation of a chromosome on staining and examination by microscopy. Visually distinct light and dark bands give each chromosome a unique morphological appearance and allow for the visual tracking of cytogenetic abnormalities, such as deletions or inversions (see below).

**Da**: Dalton. A unit of molecular mass equal to the mass of a hydrogen atom. On average, amino acids are $\sim 110$ Da and DNA/RNA bases are $\sim 330$ Da.

**Deletion**: A mutation in which one or more bases is lost from a given region of a chromosome.

**Deletion-insertion polymorphism**: Alleles (see above) that are represented by one or more bases that are present in one sequence and absent in the other.

**Descriptor**: Information about a sequence or set of sequences whose scope depends on its placement in a record. Placed on a set of sequences to reduce the need to save multiple redundant copies of information.

**Difference gel electrophoresis**: A type of two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) in which two samples are run on a single gel. Also see *2D gel electrophoresis*.

**DIGE**: See *difference gel electrophoresis*.

**DIP**: See *deletion-insertion polymorphism*. 

---

**Glossary of Bioinformatics Terms**

A.2.3
**Domain name:** Refers to one of the levels of organization of the Internet and used to both classify and identify host machines. Top-level domain names usually indicate the type of site or the country in which the host is located.

**Dotplot:** A visual technique for comparing two sequences with one another, allowing for the identification of regions of local alignment, direct or inverted repeats, insertions, deletions, or low-complexity regions.

**Download:** The act of transferring a file from a remote host to a local machine via FTP.

**Dynamic programming:** A computational technique to solve complex problems by decomposing the problem into successively smaller subproblems, then solving them recursively. The solution of a subproblem of given complexity is dependent on the solutions already computed for subproblems of lesser complexity.

**Electrospray ionization:** A method for creating ions in mass spectrometry that consists of forming a spray of charged droplets containing analyte molecules, then desolvating the droplets, leaving charged ions for analysis.

**Energy minimization:** A computational method for reducing the calculated covalent and noncovalent energy of a molecule with a given geometry. Energy minimization uses highly parameterized Newtonian descriptions of molecular bonds and atomic interactions. Energy minimization frequently is used to refine or fix protein structures determined from X-ray, nuclear magnetic resonance (NMR), or homology modeling.

**e-PCR:** Electronic polymerase chain reaction. A computational method that predicts the location of sequence tagged sites (STSs) in DNA by searching for subsequences that closely match the PCR primers used to make the STS; these subsequences must also have the correct order, orientation, and spacing such that they could prime the amplification of a PCR product of the correct molecular weight.

**EST:** Expressed sequence tags. These are usually short (300 to 500 bp), single reads from mRNA (cDNA) that usually are provided in large numbers. They represent a snapshot of what is expressed in a given tissue or at a given developmental stage. They represent tags (some coding, others not) of expression for a given cDNA library.

**Exon:** The part of a gene that remains in a mature mRNA transcript after any introns have been spliced out.

**Family trio:** DNA samples from mother, father, and child.

**FAQ:** Frequently asked questions. A compiled list of questions and answers intended for new users of any computer-based resource, such as mailing lists or newsgroups.

**Feature:** Annotation on a specific location on a given sequence.

**Filtering:** See masking.

**Firewall:** A computer separating a company or organization’s internal network from the public part, if any, of the same network. Intended to prevent unauthorized access to private computer systems.

**Flanking sequence:** Sequences 5’ or 3’ of a core DNA or RNA sequence of interest.

**FTICR or FT/MS:** Fourier transform mass spectrometry (MS) or Fourier transform ion cyclotron resonance MS. A highly accurate ion measurement used especially for large molecules of up to $1 \times 10^6$ Da or more. Electromagnetic forces are used to cycle ions in
a chamber, which are then measured by the frequency at which they resonate. To convert from frequency to a mass-charge ($m/z$) spectrum, a Fourier transform is applied.

**FTP:** File transfer protocol. The method by which files are transferred between hosts.

**Gap:** Used to improve alignments between sequences. Gaps represent insertions and deletions between sequences being studied.

**Genetic map:** Gives the relative positions of known genes, markers, or both. Markers must have two or more alleles that can be distinguished.

**Genetic marker:** A DNA feature whose physical location is known and that can be used to (indirectly) deduce the method of inheritance of a gene.

**Genome:** All of the DNA found within each of the cells of an organism. Eukaryotic genomes can be subdivided into their nuclear genome (chromosomes found within the nucleus) and their mitochondrial genome.

**Genotype:** (1) The alleles present in a given individual’s DNA for a particular genetic marker or set of markers. (2) The unique genetic makeup of an organism. See also *phenotype*.

**Graph:** A set of vertices (also called nodes) and a set of edges connecting those vertices. Can be visualized as a set of points connected by lines.

**GSS:** Genome survey sequences. This DDBJ/EMBL/GenBank division is similar in nature to the expressed sequence tag (EST) division, except that its sequences are genomic, rather than originating from cDNA (mRNA). The GSS division contains (but is not limited to) the following types of data: random “single-pass read” genome survey sequences, single-pass reads from cosmid/bacterial artificial chromosome (BAC)/yeast artificial chromosome (YAC) ends (these can be chromosome specific, but need not be), exon-trapped genomic sequences, and Alu PCR sequences.

**GUI:** Graphical user interface. Refers to software that relies on pictures and icons to direct the interaction of users with the application.

**Haplotype:** A set of closely-linked genes or genetic markers on a single chromosome.

**Heuristic algorithm:** An economical strategy for deriving a solution to a problem that is not guaranteed to find the optimal solution. Required when the algorithm for the optimal solution is computationally impractical.

**Hidden Markov Model (HMM):** A probabilistic method for the linear analysis of sequences. HMMs are used in bioinformatics for almost any task that can be described as a process which analyzes sequences from left to right. Applications of HMMs to biological data include gene prediction, protein secondary structure prediction, and the detection of sequence signals such as translation initiation sites.

**Homologs/homologous:** In phylogenetics, particular features in different individuals that are descended genetically from the same feature in a common ancestor are termed homologous. Also see orthologs and paralogs.

**Homology modeling:** A method for predicting the tertiary structure of a protein by using an existing homologous protein structure as a template.

**Homoplasy:** Similarity that has evolved independently and is not indicative of common phylogenetic origin.
Host: Any computer on the Internet that can be addressed directly through a unique IP address.

**HTG/HTGS:** High-throughput genome sequences. Various genome sequencing centers worldwide are performing large-scale, systematic sequencing of human and other genomes of interest. The databases have deemed it beneficial to put the unfinished sequences that are the result of such sequencing efforts in a separate division. HTG sequence entries undergo a maturation process. In Phase 0, the entry contains a single-to-few pass read of a single clone. In Phase 1, the entry contains unfinished sequence, which may be unordered or contain unoriented contigs or a large number of gaps. In Phase 2, the entry still contains unfinished sequence, but is ordered, with oriented contigs that may or may not contain gaps. In Phase 3, the entry contains finished sequence, with no gaps; at this point, the entry is moved into the appropriate primary DDBJ/EMBL/GenBank division. In all cases, irrespective of phase, these records (sets of sequences) are assigned a single accession number that is maintained throughout the maturation of the DNA sequence in question.

**HTML:** Hypertext markup language. The standard, text-based language used to specify the format of Internet documents. HTML files are translated and rendered through the use of Web browsers. Also see Java.

**Hyperlink:** A graphic or text within an Internet document that can be selected using a mouse. Clicking on a hyperlink transports the user to another part of the same Web page or to another Web page, regardless of location.

**Hypertext:** Within a Web page, text which functions as a hyperlink and is differentiated either by color or by underlining. Also see URL.

**ICAT:** Isotope coded affinity tagging. A method used to compare the quantities of proteins directly from two different cellular states by mass spectrometry analysis of a shotgun digest of a proteome and appropriate isotopic labeling. Also see shotgun proteomics.

**Identity:** A quantitative measure of how related two sequences are to one another, assessed as the fractional number of exact matches in a pairwise sequence alignment.

**Indel:** Acronym for insertion or deletion. Applied to length-variable regions of a multiple alignment when it is not specified whether sequence length differences have been created by insertions or deletions.

**Insertion:** A mutation in which one or more bases are inserted into a region of DNA.

**Interaction:** Any relationship, physical or otherwise, between biological entities (e.g., proteins, cells, amino acids) that can be defined experimentally.

**Internet:** A system of linked computer networks used for the transmission of files and messages between hosts. Also see intranet and LAN.

**Intranet:** A computer network internal to a company or organization. Intranets are often not connected to the Internet or are protected by a firewall. Also see Internet and LAN.

**Intron:** The part of a primary RNA transcript that is removed by splicing and, therefore, does not appear in the messenger RNA.

**Ion Trap Mass Spectrometer:** A mass spectrometer that traps ions in an oscillating electric field to measure mass-to-charge ratios. Multiple stages of mass spectrometry (MS^n) can be performed in ion traps to deduce the structure of ions.
IP address: The unique numeric address of a computer on the Internet.

Isoelectric point: The pH value at which the net charge of a protein or peptide is neutral, as determined by isoelectric focusing.

Java: A programming language developed by Sun Microsystems that allows small programs (applets) to be run on any computer. Java applets are typically invoked when a user clicks on a hyperlink on a Web page. Also see HTML.

LAN: Local Area Network. A network that connects computers in a small, defined area, such as the offices in a single wing or a group of buildings. Also see Internet and intranet.

Library: In sequencing, a collection of insert-containing clones. Sequencing libraries are created from a sequencing vector (see plasmid) and a set of inserts obtained by fragmentation of a larger piece of DNA.

Linkage: Genes or genetic markers that are physically close to one another on a chromosome and that tend to be inherited together.

Linkage disequilibrium: A state resulting when alleles at two defined loci are linked more frequently than would otherwise be expected.

Liquid chromatography: A method for separating sample proteins or peptides in preparation for mass spectrometry; separations may be based on properties such as hydrophobicity (reversed phase), surface charge (ion exchange), or diffusion rate (size exclusion/gel filtration).

LOD score: Log odds. A statistical estimate of the linkage between two loci on the same chromosome.

Low-complexity region: Regions of biased composition, usually homopolymeric runs, short-period repeats, or the subtle overrepresentation of several residues in a sequence.

MALDI: Matrix-assisted laser-desorption ionization. A common method for generating ions from analyte molecules for analysis by mass spectrometry.

Masking: The technique by which low-complexity regions (see above) are removed from protein sequences, or LINE, SINE, Alu, and similar repetitive sequences are removed from nucleotide sequences before database searches.

Mass spectrometry: A collection of exquisitely sensitive and accurate analytical techniques that precisely measure molecular masses through a process of ionization and subsequent mass-to-charge measurement.

microRNA: A short (∼20- to 25-nucleotide) single-stranded noncoding RNA molecule that has the ability to regulate gene expression at the translational level.

Microsatellite: Regions of DNA containing short tandem repeats of a simple nucleotide sequence.

mmu: One millimass unit or one thousandth of a Dalton (see above).

Molecular clock: The hypothesis that nucleotide or amino acid substitutions occur at a more or less fixed rate over evolutionary time like the slow ticking of a clock. It has been proposed that given a calibration date and a constant molecular clock, the amount of sequence divergence can be used to calculate the time that has elapsed since two molecules diverged.
**Molecular complex:** A stable complex of molecules functioning as a biological unit.

**Monoisotopic mass:** The base mass of a protein or peptide containing none of the more rare, higher mass isotopes in any of its constituent atoms.

**Motif:** Relatively conserved sequences within proteins or DNA that usually correspond to structural or functional regions.

**MS:** See *mass spectrometry*.

**MS/MS:** See *tandem mass spectrometry*.

**Mutation:** A modification to a chromosome. Mutations can involve single bases or entire regions of a chromosome. Mutations can be neutral (i.e., have no effect), harmful, or beneficial. As such, mutations drive evolutionary change.

**Neutral mass:** The actual mass, in Daltons, of a measured protein or peptide after deconvolution and subtraction of any associated ion mass. It refers to the neutral (noncharged) state of the analyzed molecule.

**Noncoding DNA:** A region of DNA that does not code for a protein.

**Nucleotide:** The basic component of both DNA and RNA. Nucleotides consist of a base, a sugar molecule, and a phosphoric acid molecule.

**Oligo:** For oligonucleotide, a short, single-stranded DNA or RNA.

**OnO (or O&O):** Ordered and oriented. The particular order and direction (complemented or uncomplemented) of each contig from an assembly is known and specified.

**ORF:** Open reading frame. A DNA sequence that has the potential to encode a protein sequence because it contains no in-frame stop codons.

**Orthologs/orthologous:** Homologous sequences are said to be orthologous when they are direct descendants of a sequence in the common ancestor, that is, without having undergone a gene duplication event. See also *homologs* and *paralogs*.

**PAM matrix:** PAM (percent accepted mutation) and BLOSUM (blocks substitution matrix) are matrices that define scores for each of the 210 possible amino acid substitutions. The scores are based on empirical substitution frequencies observed in alignments of database sequences and, in general, reflect similar physico-chemical properties—e.g., a substitution of leucine for isoleucine (two amino acids of similar hydrophobicity and size) will score higher than a substitution of leucine for glutamate.

**Paralogs/paralogous:** Homologous sequences that have arisen due to a gene duplication event. Also see *homologs* and *orthologs*.

**Pathway:** A set of interactions (see above) among biological entities; these interactions may or may not be linear or ordered in any way. Usually defined by a perturbation to a biological system whose output can be measured experimentally.

**Pedigree:** A tree representation of a family (cohort) showing the relationships between members and the pattern of inheritance of a given trait.

**Peptide mass fingerprint:** A protein identification method that works by enzymatically digesting a protein to produce a distinctive fingerprint of masses. The fingerprint is matched against putative fingerprints from protein or nucleotide databases to identify the unknown.
Phenotype: The outwardly observable characteristics of an organism. See also genotype.

Phylogenetic profile: A profile capturing the existence of orthologs of a gene across genomes. Genes with similar phylogenetic profiles are hypothesized to interact physically or at least to be linked functionally.

Physical map: A genome map showing the exact location of genes and markers. The highest-resolution physical map is the DNA sequence itself.

pI: See isoelectric point.

Plasmid: A circular, self-replicating piece of DNA. Numerous artificially designed plasmids contain priming sites, making them suitable for cloning and sequencing segments of DNA that range from 2000 to 10,000 bases in length.

Platform: Properly, the operating system running software on a computer, for example, Unix or Windows. More often used to refer to the type of computer, such as a Macintosh or PC.

PMF: See peptide mass fingerprint.

Polymorphism: Common differences in DNA sequence among individuals that can be used as markers for linkage analysis. Also see linkage.

Positional cloning: Relies on the identification of a gene through pedigree analysis, genetic and physical mapping, and mutation analysis. Does not require extensive knowledge of the biochemistry of the disease to determine the gene responsible for the disease. The opposite of functional cloning.

Primary structure: The linear sequence of a protein or RNA.

Primary transcript: The RNA molecule resulting from the transcription of the DNA sequence encoding a gene. In eukaryotes, it contains both the introns and exons prior to processing.

Primer: An oligonucleotide used to initiate polymerase-mediated replication of a strand of DNA.

Promoter: The region of DNA upstream of a gene where transcription is initiated.

Pseudogene: DNA that is similar to a normal, coding gene but that is not functional (may or may not be expressed). Pseudogenes are incapable of producing functional gene products.

Quaternary structure: The arrangement or positioning of multiple polypeptide chains (with defined tertiary structures) in larger protein complexes.

R factor: Residual disagreement. Used in X-ray crystallography as a measure of agreement between the experimentally measured diffraction amplitudes and those calculated using the protein coordinates. Perfect agreement corresponds to an R factor of 0.0. Total disagreement corresponds to an R factor of 0.59. Most good quality protein structures have R factors between 0.15 and 0.20.

Ramachandran plot: A scatterplot showing the disposition of backbone phi (\(\phi\)) and psi (\(\psi\)) torsion angles for each residue in a protein or set of proteins. Certain combinations of \(\phi\) and \(\psi\) angles are preferred strongly or are repeated over a series of residues, and these patterns can be easily detected in a Ramachandran plot.
**Reference SNP:** Reference single nucleotide polymorphism. Curated dbSNP records that define a nonredundant set of markers used for annotation of reference genome sequence and integration with other National Center for Biotechnology Information resources. Each refSNP record provides a summary list of submitter records in dbSNP and a list of external resource and database links. Also see *polymorphism*.

**Repetitive DNA:** DNA sequences of variable length that occur in multiple copies in the human and other eukaryotic genomes.

**Restriction fingerprint:** The sizes of the DNA fragments resulting from an endonuclease digestion of the piece of DNA of interest.

**Ribozyme:** A catalytic RNA sequence.

**RMSD:** Root mean square deviation. An archaic term for standard deviation. RMSD is still used in the quantification of the atomic position differences between protein structures. Very similar structures have RMSD values between 0 and 1.5 Å; moderately similar structures have RMSD values between 1.5 and 3.0 Å.

**RNAi:** RNA interference, or RNA-mediated interference. A cellular process in which double-stranded RNAs interfere with the expression of homologous genes through the degradation of complementary mRNA molecules. A technique commonly used to selectively suppress the expression of individual genes.

**Scaffold:** See supercontig.

**Secondary structure:** In proteins, the local, regular backbone structures found in folded proteins (α-helices, β-strands). In RNA, secondary structure is the set of canonical base pairs.

**Sequence polymorphisms:** Differences in DNA sequences that occur naturally among individuals. See also *single nucleotide polymorphism*.

**Server:** A computer that processes requests issued from remote locations by client machines.

**Shotgun proteomics:** An approach to proteome analysis where all components of a sample are first enzymatically digested into peptides, then separated by liquid chromatography (see above), and finally analyzed by MS/MS to identify them.

**Shotgun sequencing:** A sequencing method in which the DNA to be sequenced is broken randomly into many small fragments. The fragments in turn are sequenced individually; based on overlaps between the individual sequences, the pieces can be reassembled and the original sequence can be deduced.

**Similarity:** A quantitative measure of how related two sequences are to one another, usually assessed as the total number of identities and conservative substitutions (see above) in a pairwise sequence alignment (see *alignment* above). Similarity does not imply homology. Also see *alignment* and *conservative substitution*.

**Single nucleotide polymorphism:** Alleles that are represented by single-base changes in a DNA sequence.

**SNP:** See *single nucleotide polymorphism*.

**Stop codon:** Distinct codons (UAA, UAG, and UGA) that do not code for a specific amino acid. Stop codons act as signals that translation of an mRNA sequence into a protein sequence should be terminated. Also called termination codons. Also see *codon*. 

---

**Glossary of Bioinformatics Terms**

**A.2.10**

Supplement 18  
Current Protocols in Bioinformatics
**STS:** Sequenced tagged sites. An operationally unique sequence that identifies the combination of primer pairs used in a PCR assay, generating a reagent that maps to a single position within the genome (see above). STS sequences usually are on the order of 200 to 500 bases in length. STS is a division of GenBank devoted to STS sequences; it is intended to facilitate cross-comparison of STSs with sequences in other divisions for the purpose of correlating map positions of anonymous sequences with known genes.

**Supercontig:** A stretch of DNA sequence composed of one or more contigs (see above) with known order and orientation.

**Synteny:** In comparative mapping, the observation that the order of loci in a chromosomal region of one organism is conserved in a chromosomal region of a second organism.

**Tandem mass spectrometry:** A process whereby a first stage of mass spectrometry is used to select certain components of a sample, which are then broken down for further analysis by a second stage of mass spectrometry. In some instruments, this can be applied repeatedly to yield MS^n separations. Also see *mass spectrometry*.

**Tertiary structure:** The arrangement or positioning of secondary structure elements into compact, nonoverlapping globules or domains. Tertiary structures are the three-dimensional structures of proteins and RNAs.

**Threading:** A method for predicting the most likely fold or topology of a protein by assessing the likelihood that its sequence “fits” into a known three-dimensional fold or a known arrangement of secondary structure.

**Tiling path format:** Format indicating the identities and order of sequences to be included in a targeted region assembly.

**Time-of-Flight Mass Spectrometer (TOF-MS):** A mass spectrometer that measures mass-to-charge ratios by the time required to traverse a set distance.

**Top-down proteomics:** A newer method of proteomics that begins with mass spectrometry of intact proteins, followed by subsequent analysis of their constituents via MS degradation methods or peptide mass fingerprinting.

**Topology:** The map or plan of a physical system or set of connected objects. The topology of proteins generally is described by their backbone tertiary (three-dimensional) structure.

**TPF:** See *tiling path format*.

**URL:** Uniform resource locator. Used within Web browsers, URLs specify both the type of site being accessed (FTP, Gopher, or Web) and the address of the Web site.

**User:** The person using client-server or other types of software.

**Walking:** In sequencing, the extension of nucleotide sequence by generating reads “off the end” of the currently known sequence by using custom primers designed to be complementary (see above) to known sequence.

**Wiki:** A Web site that can be used for collaborative or informational purposes, in which any individual can freely add, edit, or delete information. The term originates from the Hawaiian word for “quick” (“wikiwiki”).

**Word matching:** Computer-based search for small segments (“words”) of identical DNA and protein sequences.
**World Wide Web:** A document delivery system capable of handling various types of nontext-based media.

**YAC:** Yeast artificial chromosome. A vector used to clone segments of DNA up to $1 \times 10^6$ bases in length.

**Z-score, Z-value:** This measures the distance of a value from the mean of a normal or Gaussian distribution in standard deviation units. A Z-score of one means the value is one standard deviation away from the mean. A Z-score of four indicates the value is four standard deviations away from the mean, indicating the value has <99.9% chance of occurring randomly.