FR3 Bioinformatics Primer: Glossary of terms*

(*from various sources including Discovering Genomics, Proteomics, & Bioinformatics 2nd edition, Campbell and Heyer; The Internet and the New Biology by Peruski and Peruski; and Bioinformatics for Dummies by Claverie and Notredame)

Accession number identification number given to every DNA and protein sequence submitted to NCBI or equivalent database.

Algorithm step-by-step procedure for solving a problem (e.g. aligning two sequences) or computing a quantity (e.g. %GC). Typically written in Perl or another computer language.

Alignment representation of two or more protein or nucleotide sequences where identical amino acids or nucleotides are in the same columns while missing amino acids or nucleotides are replaced with gaps.

Allele frequency prevalence of a gene variant in a population.

Annotate a genome is annotated once it has been analyzed for gene content. A gene is considered annotated if it has been assigned information pertaining to a cellular role.

Antibody array An **antibody** microarray is a specific form of protein microarrays, a collection of capture antibodies are spotted and fixed on a solid surface such as glass, plastic or silicon chip, for the purpose of detecting antigens.

Antisense technology molecular method that uses a nucleic acid sequence complementary to an mRNA so that the two bind and the mRNA is effectively neutralized.

Array an orderly pattern of objects. In genomic studies, there are microarrays and macroarrays. Microarrays are small spots of DNA or protein and the identity of the spotted material is known. Macroarrays are bacterial, yeast or similar colonies on plates used to determine functional consequences of genomic manipulations.

Anonymous FTP A file transfer protocol (FTP) that allows the retrieval of files from public sites.

Archive a collection of data, text, programs or other electronic information stored for other parties to access or retrieve, typically without charge.

ASCII stands for American Standard Code for Information Interchange, sets a unique numeric definition for each of the most commonly used characters in Western language, including numerals, letters and some symbols. An ASCII set is these characters and their numbers. In the context of a file, an ASCII file is one that contains only text characters:

numbers, letters and standard punctuation.

Bacterial artificial chromosome (BAC) cloning vector replicated in bacteria that can hold an insert of about 150,000 base pairs.

BAM file a binary version of a SAM file, the format is .bam.

BankIt a computer program developed by the NCBI for submitting your own sequences to GenBank.

Binomial a probability model for counting the number of ‘successes’ in an experiment with two possible outcomes.

Bioinformatics a field of study that extracts biological information from large data sets such as sequences, protein interactions, microarrays, etc. This field also includes the area of data visualization.

Biological process coined by Gene Ontology to describe broad cellular outcomes, such as mitosis or energy production, that are accomplished by ordered assemblies of molecules.

BLAST the protein and nucleic acid sequence search engine developed at NCBI that allows you to search sequence databases. BLASTn searches for nucleotide sequences, BLASTp searches for amino acid sequences; BLAST2 compares 2 sequences.

BLOSUM, block scoring matrix, a popular substitution matrix for aligning protein sequences.

Bootstrap analysis a statistical technique used to assess the reliability of the branching structure in phylogenetic trees.

Bowtie software that maps short DNA sequences to a reference genome.

Browser client software that reads HTML-formatted documents and displays them on your computer screen.

cDNA, see complementary DNA

CDS acronym for ‘coding sequences’

Cellular component a term coined by Gene Ontology to describe subcellular structures, locations and macromolecular complexes such as nucleus, telomere, and mitotic spindles.

ChIP-Seq, chromatin immunoprecipitation sequencing combination of chromatin immunoprecipitation and massively parallel sequencing used to survey interactions between protein, DNA and RNA.

Chromatogram a four-colored graph produced from nonradioactive dideoxy sequencing methods including cycle sequencing.

Clade group of related species and their common ancestors.

Cladogram phylogenetic tree showing the relationship between the species.

Client In its simplest form a client is a software program that an individual uses to send requests to a server. In a broader definition it can mean a computer or computer program that requests a service of a host computer or program.

Clone noun or verb. A clone is any molecule/cell/organism present in more than one identical copy. To clone something means to produce more than one copy of the original molecule/cell/organism.

ClustalW the most popular program for making multiple sequence alignments.

Clustering action of grouping sequences according to their similarity

Clusters of orthologous groups COG NCBI compilations of evolutionarily related gene sequences from several microbial genomes. This site allows you to search by gene or cellular role and produce dendrograms to show sequence similarities.

Comparative genomics analysis based on comparisons of entire genomes.

Complementary DNA abbreviation used at NCBI to indicate which bases constitute the open reading frame.

Consensus a pseudo sequence where each residue summarizes (using the majority rule) the content of a column in a multiple sequence alignment.

Conserved domain (CD) a domain that has been retained during evolution presumably due to its essential role within the protein's structure. Conserved domain searches are a part of the BLAST search.

Constitutive always active; constitutive promoters or genes are active at all times.

Contiguous (contigs) overlapping DNA segments that as a collection form a longer and gapless segment of DNA.

Correlation coefficient a measure of the correspondence between two sets of data.

Covariance analysis the study of a sequence that follows the evolution of two different positions and looks for some correlation.

Coverage based on the number of bases sequenced in a genome, the coverage represents how many times an average base was sequenced; finished genomes frequently have 8X coverage.

Cufflinks software that accepts aligned RNAseq reads and estimates the relative abundances of transcripts. Also tests for differential expression between samples.

Cy3 and Cy5 fluorescent green and red dyes, respectively, that are commonly used for microarray experiments.

Curate oversee/annotate/manipulate gene data (often manually). Includes fine annotation of genes, filtering of genes according to certain parameters and construction and manipulation of gene models.

Dendrogram another name for phylogenetic tree. The ClustalW guide tree is often referred to as a dendrogram, it is a file with a .dnd extension.

Distance matrix matrix that contains all the pairwise distances between sequences in a data set. Distance matrices afe often u sed to compute phylogenetic trees. Not a substitution matrix.

DNA microarray or DNA chip synonyms for gene sequences spotted on glass slides used to measure simultaneously the level of transcription of many genes.

Domain (computational) a network or portion thereof that operates under a single administrative umbrella such as an institution, a group of institutions, a geographical region, or a country.

Domain (molecular) a protein unit of at least 50 amino acids that can fold on its own, can also be used for protein sequences that occur in various contexts.

Dot matrix dot plot, a method for representing the similarity between two sequences without using an alignment.

Downstream a relative direction for a DNA sequence; towards the 3' end.

Draft sequence a description of the degree of confidence in a DNA sequence. When the DNA has been sequenced only 4 times it is described as a 'draft sequence' as opposed to a 'finished sequence' which has been sequenced 8 times. With a draft sequence about 95% of the genes should be identifiable.

dsDNA double stranded DNA, formed when two complementary DNA sequences bind each other.

dsRNA double stranded RNA, formed when two complementary RNA sequences bind to each other.

E value (Expect value) when performing a BLAST search, you will obtain an E-value for each sequence that is retrieved. An E-value can be thought of as the probability that two sequences are similar to each other by chance. Therefore E-values are best when they are small (e.g. 1 x 10^{-12}).

EC number, Enzyme Commission number a numerical classification scheme for enzymes based on the chemical reactions they catalyze.

EBI, European Bioinformatics Institute the European counterpart to the NCBI in the U.S.

EMBL, European Molecular Biology Laboratory, this acronym refers to the nucleotide database that the laboratory maintains.

Entrez the NCBI database querying system that is similar to SRS at the EBL.

Epitope a part of an antigen that is recognized by the immune system.

EST, expressed sequence tag a short subset of a cDNA sequence.

Expansionist a person who takes a systems approach to understanding complex, interconnected parts working as a whole; opposite of a reductionist.

ExPASy a server maintained by the Swiss Institute of Bioinformatics. It is the home of SWISS-PROT, the annotated protein database.

FASTA simple text format for DNA or protein.

Field in the context of databases, a well-delimited part of an entry containing information of a precise nature (author, date, etc.). You can normally search databases by explicitly targeting specific fields.

Fileserver a computer that provides files to other computers via a network.

Finished sequence DNA has been sequenced at least eight times and there are no gaps, and contains no more than 1 error in 10,000 base pairs.

FPKM Fragments Per Kilobase Of Exon Per Million Fragments Mapped; approximates the relative abundance of transcripts in terms of fragments observed from an RNAseq experiment.

FTP, file transfer protocol, an internet protocol that defines a commend method of transferring files across networks and between remote computers.

GA II Illumina sequencing platform that generates 40 Mb sequence per run with paired read lengths of 70 bp.

Gap in sequencing a segment of DNA that has not been sequenced but is flanked by sequenced DNA. Also used as a synonym for insertion or deletion in sequence alignment.

GenBank developed and housed at NCBI, GenBank is the U.S. repository for all DNA and protein sequences.

Gene ID a unique alphanumeric identifier assigned to a gene.

Gene ontology a collaborative effort of investigators to unify and standardize terms associated with the role a gene or protein plays in an organism. Represented model organisms include *Drosophila melanogaster, Saccharomyces cervisiae, Schizosaccharomyces pombe, Mus musculus, Arabidopsis thaliana, Caenorhabditis elegans, Rattus norvegicus and Dictyostelium discoideum.*

Genome the complete set of genes or genetic material present in a cell or organism.

Genome browser is a graphical interface for display of information from a biological database for genomic data.

Genomic colocation comparison of genome structure between two or more genomes.

Genomics a vague term that encompasses the study of reference genome sequences, variations within a species' genome, DNA microarrays, circuits and systems biology. Some people include wider areas such as proteomics, metabolomics, etc., under the genomics umbrella.

GFF, generic feature format a data format for identifying the features of a sequence that involves data featured in a tab-delimited file with one feature per line to enable text manipulation and data analysis tools that work with tabular data.

Global alignment an alignment of two sequences where no amino acid or nucleotide is discarded. They are all either aligned with other amino acids/nucleotides or aligned with gaps.

Haplotype a collection of alleles in one individual that are located on one chromosome. Alleles within a haplotype are often inherited as a single unit from one generation to the next. In SNP studies, haplotypes refer to a group of genomic variations found repeatedly in many individuals within a population.

Heat map graphical representation of data where the individual values contained in a matrix are represented as colors.

Hierarchical clustering a method for organizing large numbers of genes, tumors or other objects into dendrograms.

High-throughput methods that produce large volumes of data and can process many samples quickly. Robots and computerized data collection are common themes in high-throughput methods.

HiSeq Illumina sequencing platform that generates 120 Gb sequence per run with paired read lengths of 125 bp.

HiSeq2500 Illumina sequencing platform that generates 1,000 Gb sequence per run with paired read lengths of 125 bp.

Hits shorthand for sequences returned when searching a database such as NCBI.

HMM, Hidden Markov Model, a mathematical formulation of a succession of hidden, mutually exclusive properties (such as intron or exon) associated with one sequence or a multiple sequence alignment. In the context of proteins HMM is used interchangeably with the words 'motifs' or 'profiles'. In the context of DNA, HMMs are used to predict the location and structure of genes.

Homolog/homology – two sequences (DNA or amino acid) that are similar due to evolutionary relatedness.

Horizontal transfer the movement of DNA from one species to another without sexual transmission; mechanism uncertain.

Host A computer that allows users to communicate with other computers or hosts on a network.

HTML hypertext markup language, the standard language used for creating hypermedia documents within the World Wide Web.

HTTP hypertext transfer protocol, the standard language that World Wide Web clients and servers use to communicate.

Hydropathy plot (Kyte-Doolittle) a acomputer-generated graph that uses the amino acid sequence to predict whether or not the protein will span a membrane.

Hypermedia Hypertext that includes links to other forms of media.

Hypertext The linking of one document to another document or to another location within the same document.

Identity percentage of identical amino acids or nucleotides in the alignment of two sequences. Usually the identity is measured on the aligned residues with gaps ignored.

Illumina a company located in San Diego, CA that provides sequencing platforms such as MiSeq, HiSeq, HiSeq2500, NextSeq and GA II.

In silico experimental process performed on a computer and not by bench research.

Indels collective noun that refers to insertions or deletions of bases in DNA sequences.

Induced a gene with increased transcription

Intergenic sequence DNA sequence between two genes, sometime referred to as ‘junk DNA’.

Internet a global collective of computer networks running TCP/IP.

InterPro a federative database of profiles, patterns and HMMs for detecting protein domains.

Intron portion of the gene and initial RNA transcript that will be excised and not included in the mRNA; usually begins with the sequence GT and ends with AG.

Ion Proton II sequencing platform available through Gibco Life Technologies, generates 60 Gb sequence per run, with paired read lengths of 150 bp.

Ion Torrent sequencing platform available through Gibco Life Technologies, generates 30-120 Mb sequence per run, with paired read lengths of 400 bp.

Isoforms/isozymes two versions of highly similar proteins, isoforms refer to any proteins, isozymes is used for enzymes.

Isoelectric point (pI) when the net charge of a protein is zero the pH of the local environment will be equivalent to the isoelectric point of the protein.

Iteration when a process is repeated in an attempt to reach the ideal outcome. Each iteration is slightly different from the previous one since we learn from the first and improve the second iteration.

Jalview a popular tool for editing and analyzing multiple sequence alignments.

Java applet a small piece of software that your browser installs automatically and that runs on your computer.

Kb kilobase, 1,000 base pairs of DNA sequence.

KEGG, Kyoto Encyclopedia of Genes and Genomes, a world-famous Japanese database on genomes and biochemical pathways.

Kyte-Doolittle plot, see Hydropathy plot.

LAN local area network, two or more computers connected together via cabling.

Lateral transfer, see Horizontal transfer, a gene acquired from another organism as opposed to inheriting it from an ancestor.

Linkage when two genes are located near each other on the same chromosome.

Local alignment an alignment of the most similar segments between sequences. Dissimilar sequences are not considered and are removed in the final output. BLAST does local alignments.

Low-complexity filtering removal of low-complexity sequences when doing a database search.

Low complexity sequence a sequence that contains a few elements repeated many times.

Macroarrays/Microarrays, see Arrays

Mapping graphical representation of where genes are found on chromosomes in relation to each other.

Mass spectrometry, MS a technique that allows investigators to separate proteins based on their mass to charge ration (m/z). The m/z for each protein allows them to be identified and quantified from complex mixtures. This proteomics tool is often used in pairs and called tandem mass spectrometry (MS/MS). Protein samples are first ionized then inserted into MS/MS by either laser-based methods (MALDI, SELDI) or an electrospray method (ESI).

Megabase (Mb) 1,000 kilobases or 1 million bases of DNA.

Metabolic pathway the sequential reactions that taken together lead from one substance to another.

Metabolome term coined to encompass the entire metabolic content of a cell or organism.

Metadata Data concerning or describing some core data, as distinct from the primary data that is being described. This includes metadata on the origin, source, history, ownership or location of some thing.

Microarray technology used for parallel gene expression and DNA homology analysis using an orderly arrangement of thousands of identified sequenced genes printed on a solid chip.

Microsatellite a short segment of DNA (2-50 bases) repeated multiple times. Microsatellites vary in length and base composition which makes them useful tools for distinguishing members of a population.

MiSeq Illumina sequencing platform that generates 15 Gb sequence per run with paired read lengths of 300 bp.

Molecular function coined by Gene Ontology, describes tasks performed by individual gene products such as transcription factor or calcium transportation.

Motif short recurring patterns in DNA that are thought to have a biological function.

Multidimensional scaling the purpose of multidimensional scaling (MDS) is to provide a visual representation of the pattern of proximities (i.e., similarities or distances) among a set of objects. The relationship between input proximities and distances among points on the map is positive: the smaller the input proximity, the closer (smaller) the distance between points, and vice versa.

Multiplex when a series of reagents are mixed in a single tube so more than one outcome will be produced simultaneously; multiplex PCR produces several different sized bands of DNA that can be detected.

NCBI National Center for Biotechnology Information a federally funded part of the U.S. National Library of Medicine. NCBI is the home of GenBank, BLAST, COG, and many other genomic databases and computational tools.

Neighbor joining, NJ the most popular method for reconstructing phylogenetic trees.

NextSeq Illumina sequencing platform that generates 120 Gb sequence per run with paired read lengths of 150 bp.

Normal distribution data have an average value around which all the individuals are clustered. A graph of normal distribution results in the classic ‘bell curve’.

Normalized data that has been corrected or standardized, for example, by subtracting the sample mean from each observation, and dividing the result by the sample standard deviation.

NR the non-redundant protein database that contains all the putative protein sequences contained in the nucleotide databases. Its European equivalent is TrEMBL.

Offline describes actions or tasks performed when not connected to another computer via a network.

Oligonucleotide (oligos) ssDNA polymers of unspecified length. The oligo sequence is determined by the investigator and synthesized in vitro. Oligos are used to probe blots, prime sequencing reactions, and PCR, as well as for spotting on DNA microarrays.

Online describes actions or tasks performed when connected to another computer via a network.

ORF, Open reading frame a portion of a cDNA or gene that begins with a start codon and ends with the stop codon. Synonym for coding sequence (CDS) on GenBank results.

Overexpress when genes are bioengineered to produce excessive amounts of protein, they overexpress their encoded proteins.

Orthologue two genes in different species that are evolutionarily related.

PacBio RSII sequencing platform available through Pacific Biosciences in Menlo Park, CA that generates 250 Mb sequence per run with an average read length of 4,500 bp.

Pairwise alignment alignment of two sequences.

Paralogue two genes within the same species are called paralogs if they evolved from the other (evolved from one ancestral gene).

Parsimony a technique for reconstructing phylogenetic trees.

Patterns conserved residues that one can use as a functional signature.

Perl script a program written in Perl, which is optimized for manipulating and finding patterns in sequences.

Pfam a collection of profiles for detecting domains and protein families.

PIR protein information resource, an annotated protein database similar to SWISS-PROT. PIR is also the name of a sequence format that is similar to FASTA.

Polymorophism alternative forms of genes and other sequences.

Postgebnomic era the time in biology after entire genomes have been sequenced routinely. The postgenomic era began around the year 2000.

Postranscriptional control after a gene has been transcribed the RNA can be modified and regulated, processes that constitute posttranscriptional control.

Protein data bank (PDB) database of every protein for which the 3D structure is known, it also contains a few non-protein structures.

Protein mass fingerprinting (PMF) identification of proteins by matching the masses of peptides contained in the protein with those of known proteins in a database.

Protein microarrays proteomic methods similar to DNA microarrays in size and scale' proteins are spotted onto glass and are used to determine protein interaction or to identify and quantify molecules found in various solutions.

Protein profile a tool for visualizing a particular property (hydrophathy, charge, etc.) along a particular sequence by using a sliding window technique.

Protein export domains sequence of nucleic acids in a transcript that facilitate its transport through nuclear pores.

Protein microarray a high-throughput method used to track the interactions and activities of proteins, and to determine their function, and determining function on a large scale. Its main advantage lies in the fact that large numbers of proteins can be tracked in parallel. The chip consists of a support surface such as a glass slide, nitrocellulose membrane, bead, or microtitre plate, to which an array of capture proteins is bound. Probe molecules, typically labeled with a fluorescent dye, are added to the array. Any reaction between the probe and the immobilised protein emits a fluorescent signal that is read by a laser scanner.

Proteome the complete collection of proteins in a cell/tissue/organism at a particular time. Unlike genomes that are stable over the lifetime of the organism, proteomes change rapidly as each cell responds to its changing environment and produces new proteins and at different amounts.

Proteomics the study of proteomes that includes determining the 3D shapes of proteins, their roles inside cells, the molecules with which they interact, and defining which proteins are present and how much of each is present at a given time.

Pseudogenes segments of DNA that resemble genes by their sequence of bases but are nonfunctional. Pseudogenes often have transposons inserted in them, or they may have other mutations that led to their inability to encode a functional protein.

Public domain software that can be freely use, copied, distributed and modified (freeware, shareware).

Pubmed an extensive database of biomedical literature hosted by NCBI that is searchable. You can subscribe to PubCrawler and automatically search PubMed and receive e mail results on a schedule of your dhoosing.

p-value probability associated with a statistical test of the difference between populations. Populations are considered significantly different if the associated p-value is small (typically 0.1 or smaller).

Query the sequence used to scan a database for sequence similarity.

qPCR quantitative PCR, synonym for real-time PCR.

real time PCR (RT-PCR) a molecular method that is sometimes confused with reverse transcriptase PCR. Real time PCR uses the specificity of PCR to measure the number of template molecules in your starting material.

Reductionist a person who dissects a complex system into increasingly smaller parts in order to understand it.

Reference a reference genome was sequenced first for a species and thus represents a standard but not necessarily 'normal' example. The term 'reference' implies that variations exist within the population, but the reference is used as a common point for comparison.

Reiterative a process can be described as reiterative if you keep trying to improve the quality with each successive attempt. For example, models that describe how genes are regulated are reiterative because each version of the model is built upon more information so the model gradually approaches the truth.

Repressed when the level of gene transcription is reduced.

reverse genetics beginning with a gene sequence and deducing its function afterwards; the opposite of traditional genetics.

RNAi RNA interference, short dsRNA capable of inactivating genes by blocking the production of the encoded proteins via microRNA (miRNA) and short inhibitory RNA (siRNA).

RNAseq use of deep sequencing technologies for transcriptome profiling.

Router a device that directs or routes data between networks or different parts of a network.

RPKM , Reads per kilobase per million reads, a measure of relative molar RNA concentration from an RNAseq sample.

SAM file a text format for storing sequence data in a series of tab delimited ASCII columns.

Sanger sequencing dideoxy DNA sequencing method.

Scaffold a collection of contigs lumped together into one larger contig.

Secretome all the secreted proteins of a cell/tissue/organism.

Sequence-tagged site (STS) unique locus in a genome defined by PCR primers to amplify a single locus, used as markers to define chromosomal positions and to map genomes.

Server a combination of both computer hardware and software that supplies a service to requests submitted by client computers. The server processes the requests and then returns the results to the client. In short, it is a computer that makes services available on a network. For example a fileserver makes files available.

Shotgun sequencing a strategy for sequencing whole genomes pioneered by the company Celera. Genomes are cut into very small pieces, cloned into plasmids, sequenced, and then assembled into whole chromosomes or genomes. This method is faster than hierarchical shotgun sequencing but is more prone to assembly errors.

Signal peptide or signal sequence hydrophobic in nature, the first 20 amino acids of proteins synthesized that pause translation until the ribosome docks with the rough endoplasmic reticulum.

Signal transduction conveyance of information from the outside to the inside of the cell. When a ligand binds to its receptor, the information is conveyed to the rest of the cell through a complex pathway of signal tranasduction that involves second messengers.

Similarity percent of similar amino acids in the alignment of two sequences. Two amino acids are similar if they have similar physicochemical properties.

Single nucleotide polymorphisms/SNPs very similar to point mutations except SNPs are considered to represent the genetic variation present in wild type genotypes. By definition, SNPs differ from the reference sequence of a species.

Singleton any object that is not included in a collection of objects of its type, usually because it is not similar to any of the objects under consideration.

Small nuclear RNAs (snRNAs) example of a noncoding RNA.

Small nucleolar RNAs (snoRNA) example of a noncoding RNA.

SMTP simple mail transfer protocol, defines the manner by which e-mail is transferred between computers on the Internet.

SNP, **single nucleotide polymorphism** DNA sequence variations that occurs when a single nucleotide in the genome sequence is altered.

Splice site junction intron/exon boundaries determined by transcriptome analysis of genes.

Stochastic defines a program that uses chance to compute its results, common in genetic algorithms.

Submission the data you send to a server.

Substitution matrix matrix that contains a numerical score associates with the cost or reward of each possible mutation or conservation. Unlikely mutations are penalized with a negative score.

SWISS-PROT one of the most extensive annotated protein databases available.

Synteny multiple generic loci from different species located on a chromosomal region of common evolutionary ancestry.

Synthetic biology the use of engineering principles to construct small biological devises by assembling new DNA circuits inside cells.

Systems biology/systems approach coined to denote the new perspective for research in the postgenomic era. Systems biology studies whole cells/tissues/organisms not by a traditional reductionist approach but by holistic means in a reiterative attempt to model the complete cell/tissue/organism.

tBLASTx compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

TCP/IP transmission control protocol/internet protocol,
The software communication protocol of the Internet. One computer communicates with another computer through the Internet using TCP/IP format.

The Institute for Genomic Research (TIGR) Maryland-based genomics and proteomics nonprofit organization that produced public domain genome sequences and analysis software.

Transcript a sequence of RNA produced by transcription.

Transcriptome coined to describe the complete RNA content of a cell/tissue/organism and often measured by DNA microarrays or RNAseq.

Transmembrane domain the portion of a protein that spans a phospholipid bilayer, typically about 20 amino acids long and predominately hydrophobic.

Transposons sometimes referred to as ‘jumping genes’, segments of DNA that can move from one place in a genome to another.

TrEMBL translated EMBL, European equivalent to NR.

t-test statistical method for determining whether a mean or the difference between two means is significantly different than the hypothesized value.

URL, uniform resource locator, a standardized way of representing different documents, media, and network services on the World Wide Web. It gives every document of the Internet its own unique address. URLs are most commonly associated with Web sites but also apply to e-mail servers, Gopher servers and any other computer with an Internet address.

untranslated region (UTR) portion of mRNA that is not translated’ found on 5’ and 3’ ends of mRNA.

Upstream a relative direction for nucleic acids often used to describe the location of a promoter relative to the start transcription site. For example, the start codon is upstream of the stop codon.

Venn diagram a diagram representing mathematical or logical sets pictorially as circles or closed curves within an enclosing rectangle (the universal set), common elements of the sets being represented by the areas of overlap among the circles.

WAIS, wide area information server, a service that allows users to intelligently search for information among databases distributed throughout the Internet.

WAN, wide area network, a group of geographically separated computers connected via dedicated lines or satellite links.

Webmaster the administrator responsible for the management and often design of a World Wide Web site.

Whole genome shotgun (WGS) a genome sequencing strategy that skips the mapping stage and uses computers to reassemble huge numbers of independent sequencing runs that were generated from many random fragments from a genome.

Wild-type (*wt*), an allele, genotype or phenotype that is considered to be the standard for a given strain or species. Wild type alleles encode functional proteins and produce typical phenotypes.

WWW, World Wide Web, the initiative to create a universal hypermedia-based method to access information and resources on the Internet.

Yeast artificial chromosome (YAC) a cloning vector that replicates in yeast and can contain inserts about 1 Mb in size.