

# واژه نامه بيوانفورماتيك

## A

### **Accession number**

An identifier supplied by the curators of the major biological databases upon submission of a novel entry that uniquely identifies that sequence (or other) entry.

### **Active site**

The amino acid residues at the catalytic site of an enzyme. These residues provide the binding and activation energy needed to place the substrate into its transition state and bridge the energy barrier of the reaction undergoing catalysis

### **Adenine**

A purine base found in DNA and RNA

### **Agents**

Independent, autonomous, software modules that can search the Internet for data or content pertinent to a particular application, such as a gene, protein, or biological system.

### **Agricultural biotechnology (AgBio)**

The application of rDNA technology to agriculturally important plants and organisms.

### **Algorithm**

A series of steps defining a procedure or formula for solving a problem, that can be coded into a programming language and executed. Bioinformatics algorithms typically are used to process, store, analyze, visualize and make predictions from biological data.

### **Alignment**

The result of a comparison of two or more gene or protein sequences in order to determine their degree of base or amino acid similarity. Sequence alignments are used to determine the similarity, homology, function or other degree of relatedness between two or more genes or gene products.

### **Allele**

A given form of a gene that occupies a specific position or locus on a chromosome. Variant forms of genes occurring at the same locus are said to be alleles of one another.

### **Alternative splicing**

One of the alternate combinations of a folded protein that are possible due to by recombination of multiple gene segments during mRNA splicing that occurs in higher organisms.

### **Alternative splice-form**

One of the possible alternate combinations of exons into a folded protein that are possible by recombining multiple gene segments during mRNA splicing in higher organisms.

### **Alu family**

A common set of dispersed DNA sequences found throughout the human genome; each is about 300 bases long and they are repeated at least 500,000 times. Alu sequences are speculated to have originated from viral RNA sequences that integrated into human DNA thousands of years ago.

### **Amino acid**

One of the 20 chemical building blocks that are joined by amide (peptide) linkages to form a polypeptide chain of a protein

### **Analogy**

Reasoning by which the function of a novel gene or protein sequence may be deduced from comparisons with other gene or protein sequences of known function. Identifying analogous or homologous genes via similarity searching and alignment is one of the chief uses of Bioinformatics. (See also alignment, similarity search.)

### **Annotation**

A combination of comments, notations, references, and citations, either in free format or utilizing a controlled vocabulary, that together describe all the experimental and inferred information about a gene or protein. Annotations can also be applied to the description of other biological systems. Batch, automated annotation of bulk biological sequence is one of the key uses of Bioinformatics tools.

### **Anticodon**

The triplet of contiguous bases on tRNA that binds to the codon sequence of nucleotides on mRNA. Example: GGG codes for Glycine.

### **Antigen**

Any foreign molecule that stimulates an immune response in a vertebrate organism. Many antigens are proteins such as the surface proteins of foreign organisms.

### **Antisense**

DNA or RNA composed of the complementary sequence to the target DNA/RNA. Also used to describe a therapeutic strategy that uses antisense DNA or RNA sequences to target specific gene DNA sequences or mRNA implicated in disease, in order to bind and physically inhibit their expression by physically blocking them.

### **Assay**

A method for measuring a biological activity. This may be enzyme activity, binding affinity, or protein turnover. Most assays utilize a measurable parameter such as color, fluorescence or radioactivity to correlate with the biological activity.

### **Assembly**

Compilation of overlapping sequences from one or more related genes that have been clustered together based on their degree of sequence identity or similarity. Sequence assembly may be used to piece together "shotgun" sequencing fragments (see shotgun sequencing) based upon overlapping restriction enzyme digests, or may be used to identify and index novel genes from "single-pass" cDNA sequencing efforts.

### **Autoradiography**

A method used to locate radioisotope-labeled materials which have been separated in gels or are present in blots. The location of the radiolabeled material is determined by overlaying the test material with a photographic film that is sensitive to the radioisotope.

## **B**

### **Bacterial artificial chromosome (BAC)**

Cloning vector that can incorporate large fragments of DNA. (see YACS)

### **Bacteriophage**

A virus that infects bacteria. The bacteriophage DNA has served as a basis for cloning vectors, and is also utilized to create phage libraries containing human or other genes.

### **Baculovirus**

An insect virus which forms the basis of a protein expression system

### **Base pair**

A pair of nitrogenous bases (a purine and a pyrimidine), held together by hydrogen bonds, that form the core of DNA and RNA i.e the A:T, G:C and A:U interactions.

## **Beta sheet**

A three dimensional arrangement taken up by polypeptide chains that consists of alternating strands linked by hydrogen bonds. The alternating strands together form a sheet that is frequently twisted. One of the secondary structural elements characteristic of proteins.

## **Bioinformatics**

The field of endeavor that relates to the collection, organization and analysis of large amounts of biological data using networks of computers and databases (usually with reference to the genome project and DNA sequence information)

## **Bivalent**

Having two binding sites; having 2 free electrons available for binding.

## **Blunt-end (ligation)**

The joining of DNA fragments that contain no overhang at either end and consequently no DNA bases available for hybridization (cf. sticky-end ligation).

## **C**

### **Carboxyl group**

The -COOH functional group, acidic in nature, found in all amino acids

### **cDNA (complementary DNA)**

A DNA strand copied from mRNA using reverse transcriptase. A cDNA library represents all of the expressed DNA in a cell.

### **cDNA library**

A set of DNA fragments prepared from the total mRNA obtained from a selected cell, tissue or organism.

### **Cell**

The basic unit of any living organism.

### **Cell Cycle**

The life cycle of a cell which is marked by cell division which is separated into four phases: G1, S, G2, and M. DNA replication is confined to the S(synthesis) phase, and chromosomal separation in the M (mitotic) phase .

### **Chimeric clone**

A cloning artifact created by a foreign gene being inserted into a vector in an incorrect orientation resulting in the expression of a protein consisting of a fusion of two different gene products.

## **Chromat**

Data file output from most popular DNA sequencers. Chromat files consist of the fluorescent traces generated by the sequencer for each of the four chemical bases, A, C, G, and T, together with the sequence and measures of the error in the traces at each sequence position.

## **Chromatin**

The chromosome as it appears in its condensed state, composed of DNA and associated proteins (mainly histones).

## **Chromosome**

The structure in the cell nucleus that contains all of the cellular DNA together with a number of proteins that compact and package the DNA.

## **Clinical trials**

Research studies that involve patients. Biotechnology companies typically use clinical trials to assess the efficacy and safety of new therapies and to answer scientific questions. Typically, there are 3 phases during a clinical trial. Phase I is designed to evaluate the safety of the product in humans; phase II analyses the effects of dose escalation, and phase III definitively evaluates the clinical efficacy of the product.

## **Clone**

A population of genetically identical cells or DNA molecules.

## **Cloning**

The formation of clones or exact genetic replicas.

## **Cluster**

The grouping of similar objects in a multidimensional space. Clustering is used for constructing new features which are abstractions of the existing features of those objects. The quality of the clustering depends crucially on the distance metric in the space. In bioinformatics, clustering is performed on sequences, high-throughput expression and other experimental data. Clusters of partial or complete gene sequences can be used to identify the complete (contiguous) sequence and to better identify its function. Clustering expression data enables the researcher to discern patterns of co-regulation in groups of genes.

## **Coding regions (CDS)**

The portion of a genomic sequence bounded by start and stop codons that identifies the sequence of the protein being coded for by a particular gene.

### **Codon**

A sequence of three adjacent nucleotides that designates a specific amino acid or start/stop site for transcription.

### **Combinatorial chemistry**

The use of chemical methods to generate all possible combinations of chemicals starting with a subset of compounds. The building blocks may be peptides, nucleic acids or small molecules. The libraries of compounds formed by this methodology are used to probe for new pharmaceutical reagents (see high-throughput screening).

### **Complementary determining region (CDR)**

The hypervariable regions of an antibody molecule, consisting of three loops from the heavy chain and three from the light chain, that together form the antigen-binding site.

### **Complexity (of gene sequence)**

The term "low complexity sequence" may be thought of as synonymous with regions of locally biased amino acid composition. In these regions, the sequence composition deviates from the random model that underlies the calculation of the statistical significance (P-value) of an alignment. Such alignments among low complexity sequences are statistically but not biologically significant, i.e., one cannot infer homology (common ancestry) or functional similarity.

### **Configuration**

(in software) The complete ordering and description of all parts of a software or database system. Configuration management is the use of software to identify, inventory and maintain the component modules that together comprise one or more systems or products.

### **Conformation**

The precise three-dimensional arrangement of atoms and bonds in a molecule describing its geometry and hence its molecular function.

### **Consensus sequence**

A single sequence delineated from an alignment of multiple constituent sequences that represents a "best fit" for all those sequences. A "voting" or other selection procedure is used to determine which residue (nucleotide or amino acid) is placed at a given position in the event that not all of the constituent sequences have the identical residue at that position.

### **Constitutive synthesis (expression)**

Synthesis of mRNA and protein at an unchanging or constant rate regardless of a cell's requirements (see housekeeping genes).

## **Contig**

A length of contiguous sequence assembled from partial, overlapping sequences, generated from a "shotgun" sequencing project. Contigs are typically created computationally, by comparing the overlapping ends of several sequencing reads generated by restriction enzyme digestion of a segment of genomic DNA. The creation of contigs in the presence of sequencing errors, ambiguities and the presence of repeats is one of the most computationally challenging aspects of the role of Bioinformatics in genome analysis.

## **Convergence**

The end-point of any algorithm that uses iteration or recursion to guide a series of data processing steps. An algorithm is usually said to have reached convergence when the difference between the computed and observed steps falls below a pre-defined threshold.

## **Cosmids**

DNA vectors that allow the insertion of long fragments of DNA (up to 50 kbases).

## **Crystal structure**

Term used to describe the high resolution molecular structure derived by x- ray crytallographic analysis of protein or other biomolecular crystals.

## **Cytoplasm**

The medium of the cell between the nucleus and the cell membrane.

## **Cytosine**

A pyrimidine base found in DNA and RNA.

## **D**

## **Data Cleaning**

A process whereby automated or semi-automated algorithms are used to process experimental data, including noise, experimental errors and other artifacts, in order to generate and store high-quality data for use in subsequent analysis. Data cleaning is typically required in high-throughput sequencing where compression or other experimental artifacts limit the amount of sequence data generated from each sequencing run or "read."

## **Data Mining**

The ability to query very large databases in order to satisfy a hypothesis ("top-down" data mining); or to interrogate a database in order to generate new hypotheses based on rigorous statistical correlations ("bottom-up" data mining).

## **Data Processing**

Data processing is defined as the systematic performance of operations upon data such as handling, merging, sorting, and computing. The semantic content of the original data should not be changed, but the semantic content of the processed data may be changed.

## **Data Warehouses**

Vast arrays of heterogeneous (biological) data, stored within a single logical data repository, that are accessible to different querying and manipulation methods.

## **Database**

Any file system by which data gets stored following a logical process. (see also relational database)

## **Deconvolution**

Mathematical procedure to separate out the overlapping effects of molecules such as mixtures of compounds in a high-throughput screen, or mixtures of cDNAs in a high density array.

## **Deletion**

A chromosomal alteration in which a portion of the chromosome or the underlying DNA is lost.

## **Deletion mapping**

Process in which different deletions in a region of DNA are created and used to map the functionally critical areas of that DNA. e.g the minimal region of DNA required for a test promoter can be ascertained by systematic deletions in the region of interest.

## **Dendrogram**

A graphical procedure for representing the output of a hierarchical clustering method. A dendrogram is strictly defined as a binary tree with a distinguished root, that has all the data items at its leaves. Conventionally, all the leaves are shown at the same level of the drawing. The ordering of the leaves is arbitrary, as is their horizontal position. The heights of the internal nodes may be arbitrary, or may be related to the metric information used to form the clustering.

## **Dimer**

A composite molecule formed by the binding of two molecules (see homo and heterodimers).

## **Disulphide bond**

Covalent link formed between the sulphur atoms of two different cysteine residues in a protein. Important in maintaining the folded structure of a protein, and also for linking different proteins in a complex.

## **DNA (deoxyribonucleic acid)**

The chemical that forms the basis of the genetic material in virtually all organisms. DNA is composed of the four nitrogenous bases Adenine, Cytosine, Guanine, and Thymine, which are covalently bonded to a backbone of deoxyribose-phosphate to form a DNA strand. Two complementary strands (where all Gs pair with Cs and As with Ts) form a double helical structure which is held together by hydrogen bonding between the cognate bases.

## **DNA fingerprinting**

A technique for identifying human individuals based on a restriction enzyme digest of tandemly repeated DNA sequences that are scattered throughout the human genome, but are unique to each individual.

## **DNA microarrays**

The deposition of oligonucleotides or cDNAs onto an inert substrate such as glass or silicon. Thousands of molecules may be organized spatially into a high-density matrix. These DNA chips may be probed to allow expression monitoring of many thousands of genes simultaneously. Uses include study of polymorphisms in genes, de novo sequencing or molecular diagnosis of disease.

## **DNA polymerase**

An enzyme that catalyzes the synthesis of DNA from a DNA template given the deoxyribonucleotide precursors.

## **DNA probes**

Short single stranded DNA molecules of specific base sequence, labeled either radioactively or immunologically, that are used to detect and identify the complementary base sequence in a gene or genome by hybridizing specifically to that gene or sequence.

## **DNA sequencing**

The technique in which the specific sequence of bases forming a particular DNA region is deciphered.

## **DNase (Deoxyribonuclease)**

One of a series of enzymes that can digest DNA.

## **Domain (protein)**

A region of special biological interest within a single protein sequence. However, a domain may also be defined as a region within the three-dimensional structure of a protein that may encompass regions of several distinct protein sequences that accomplishes a specific function. A domain class is a group of domains that share a common set of well-defined properties or characteristics.

## **Drug**

An agent that affects a biological process. Specifically, a molecule whose molecular structure can be correlated with its pharmacological activity.

## **Drug discovery cycle**

The cycle of events required to develop a new drug. Typically this involves research, preclinical testing and clinical development, and can take from 5 to 12 years.

## **E**

## **Electronic Northern**

The use of an electronic database of cDNA sequences (or probes derived from them) in order to measure the relative levels of mRNAs expressed in different cells or tissues. An example of the use of an electronic Northern might be to identify the differences in the genes expressed in prostate cancer and those in benign prostate hyperplasia, by subtracting the database of one from the other and seeing which cDNAs remain.

## **Electrophoresis**

The use of an external electric field to separate large biomolecules on the basis of their charge by running them through acrylamide or agarose gels.

## **Enhancers**

DNA sequences that can greatly increase the transcription rates of genes even though they may be far upstream or downstream from the promoter they stimulate.

## **Enzyme**

A class of proteins that are capable of catalyzing chemical reactions (the making or breaking of chemical bonds). They do so by orienting their substrates into a suitable geometry in a particular location (the active site) where electrophilic or nucleophilic amino acid residues can participate in the reaction. Enzymes are protein catalyst that speeds up chemical reactions that would otherwise be prohibitively slow under physiological conditions.

## **Epigenomics**

The study of complex expression networks or linkages both spatially (within the body) and temporally (at different times in development).

### **Equilibrium constant**

Value that describes the equilibrium state of the reversible reaction between two molecular species.

### **Eukaryote**

A cell or organism with a distinct membrane-bound nucleus as well as specialized membrane-based organelles (see also prokaryote).

### **Exon**

The region of DNA within a gene that codes for a polypeptide chain or domain. Typically a mature protein is composed of several domains coded by different exons within a single gene.

### **Expressed Sequence Tags (ESTs)**

A small sequence from an expressed gene that can be amplified by PCR. ESTs act as physical markers for cloning and full length sequencing of the cDNAs of expressed genes. Typically identified by purifying mRNAs, converting to cDNAs, and then sequencing a portion of the cDNAs.

### **Expression (gene or protein)**

A measure of the presence, amount, and time-course of one or more gene products in a particular cell or tissue. Expression studies are typically performed at the RNA (mRNA) or protein level in order to determine the number, type, and level of genes that may be up-regulated or down-regulated during a cellular process, in response to an external stimulus, or in sickness or disease. Gene chips and proteomics now allow the study of expression profiles of sets of genes or even entire genomes.

### **Expression profile**

The level and duration of expression of one or more genes, selected from a particular cell or tissue type, generally obtained by a variety of high-throughput methods, such as sample sequencing, serial analysis, or microarray-based detection.

### **Expression vector**

A cloning vector that is engineered to allow the expression of protein from a cDNA. The expression vector provides an appropriate promoter and restriction sites that allow insertion of cDNA.

## **F**

### **Fingerprint**

A fingerprint is a set of motifs used to predict the occurrence of similar motifs, in either an individual sequence or in a database. Fingerprints are refined by iterative scanning of a composite protein sequence database. A composite or multiple-motif fingerprint contains a number of aligned motifs taken from different parts of a multiple alignment. True family members are then easy to identify by virtue of possessing all elements of the fingerprint, while subfamily members may be identified by possessing only part of it.

### **Frameshift**

A deletion, substitution, or duplication of one or more bases that causes the reading-frame of a structural gene to shift from the normal series of triplets.

### **Functional genomics**

The use of genomic information to delineate protein structure, function, pathways and networks. Function may be determined by "knocking out" or "knocking in" expressed genes in model organisms such as worm, fruitfly, yeast or mouse.

### **Fusion protein**

The protein resulting from the genetic joining and expression of 2 different genes (see chimeric)

## **G**

### **Gaps (affine gaps)**

A gap is defined as any maximal, consecutive run of spaces in a single string of a given alignment. Gaps help create alignments that better conform to underlying biological models and more closely fit patterns that one expects to find in meaningful alignment. The idea is to take in account the number of continuous gaps and not only the number of spaces when calculating an alignment. Affine gaps contain a component for gap insertion and a component for gap extension, where the extension penalty is usually much lower than the insertion penalty. This mimics biological reality as multiple gaps would imply multiple mutations, but a single mutation can lead to a long gap quite easily.

### **Gap penalties**

The penalty applied to a similarity score for the introduction of an insertion or deletion gap, the extension of a gap, or both. Gap penalties are usually subtracted from a cumulative score being determined for the comparison of two or more sequences via an optimization algorithm that attempts to maximize that score.

### **Gel electrophoresis**

A technique by which molecules are separated by size or charge by passing them through a gel under the influence of an external electric field.

## **Gene Index**

A listing of the number, type, label and sequence of all the genes identified within the genome of a given organism. Gene indices are usually created by assembling overlapping EST sequences into clusters, and then determining if each cluster corresponds to a unique gene. Methods by which a cluster can be identified as representing a unique gene include identification of long open reading frames (ORFs), comparison to genomic sequence, and detection of SNPs or other features in the cluster that are known to exist in the gene.

## **GenBank**

Data bank of genetic sequences operated by a division of the National Institutes of Health.

## **Gene**

Classically, a unit of inheritance. In practice, a gene is a segment of DNA on a chromosome that encodes a protein and all the regulatory sequences (promoter) required to control expression of that protein.

## **Gene chips (also Gene arrays)**

The covalent attachment of oligonucleotides or cDNA directly onto a small glass or silicon chip in organized arrays. Over 50,000 different DNA fragments can be presented on a single chip providing a high throughput parallel method of probing gene expression, genotype or gene function.

## **Gene expression**

The conversion of information from gene to protein via transcription and translation.

## **Gene families**

Subsets of genes containing homologous sequences which usually correlate with a common function.

## **Gene library**

A collection of cloned DNA fragments created by restriction endonuclease digestion that represent part or all of an organism's genome.

## **Gene product**

The product, either RNA or protein, that results from expression of a gene. The amount of gene product reflects the activity of the gene.

## **Gene therapy**

The use of genetic material for therapeutic purposes. The therapeutic gene is typically delivered using recombinant virus or liposome based delivery systems.

### **Genetic code**

The mapping of all possible codons into the 20 amino acids including the start and stop codons.

### **Genetic engineering (Recombinant DNA technology)**

The procedures used to isolate, splice and manipulate DNA outside the cell. Genetic Engineering allows a recombinantly engineered DNA segment to be introduced into a foreign cell or organism, and be able to replicate and function normally.

### **Genetic marker**

Any gene that can be readily recognized by its phenotypic effect, and which can be used as a marker for a cell, chromosome, or individual carrying that gene. Also, any detectable polymorphism used to identify a specific gene.

### **Genome**

The complete genetic content of an organism.

### **Genomic DNA (sequence)**

DNA sequence typically obtained from mammalian or other higher-order species, which includes both intron and exon sequence (coding sequence), as well as non-coding regulatory sequences such as promoter, and enhancer sequences.

### **Genomics**

The analysis of the entire genome of a chosen organism.

### **Genotype**

Strictly, all of the genes possessed by an individual. In practice, the particular alleles present in a specific genetic locus.

### **Glycosylation**

The addition of carbohydrate groups (sugars) e.g. to polypeptide chains

### **Guanine (G)**

One of the nitrogenous purine bases found in DNA and RNA

## **H**

### **Hairpin**

A double-helical region in a single DNA or RNA strand formed by the hydrogen-bonding between adjacent inverse complementary sequences to form a hairpin shaped structure.

### **Haploid**

A cell or organism containing only one set of chromosomes without the homologous pairs. (cf. diploid)

### **Heterodimer**

Protein composed of 2 different chains or subunits .

### **Heteroduplex**

Hybrid structure formed by the annealing of two DNA strands (or an RNA and DNA) that have sufficient complementarity in their sequence to allow hydrogen bonding.

### **Hidden Markov model (HMM)**

A joint statistical model for an ordered sequence of variables. The result of stochastically perturbing the variables in a Markov chain (the original variables are thus "hidden"), where the Markov chain has discrete variables which select the "state" of the HMM at each step. The perturbed values can be continuous and are the "outputs" of the HMM. A Hidden Markov Model is equivalently a coupled mixture model where the joint distribution over states is a Markov chain. Hidden Markov models are valuable in bioinformatics because they allow a search or alignment algorithm to be trained using unaligned or unweighted input sequences; and because they allow position-dependent scoring parameters such as gap penalties, thus more accurately modeling the consequences of evolutionary events on sequence families.

### **High-throughput screening**

The method by which very large numbers of compounds are screened against a putative drug target in either cell-free or whole-cell assays. Typically, these screenings are carried out in 96 well plates using automated, robotic station based technologies or in higher- density array ("chip") formats.

### **HLA complex**

Another name for the MHC in humans; refers to the "Human Leukocyte Antigen" complex located on chromosome 6.

### **Homeobox**

A highly conserved region in a homeotic gene composed of 180 bases (60 amino acids) that specifies a protein domain (the homeodomain) that serves as a master genetic regulatory element in cell differentiation during development in species as diverse as worms, fruitflies, and humans.

## **Homeodomain**

A 60 amino-acid protein domain coded for by the homeobox region of a homeotic gene.

## **Homeotic gene**

A gene that controls the activity of other genes involved in the development of a body plan. Homeotic genes have been found in organisms ranging from plants to humans.

## **Homology**

(strict) Two or more biological species, systems or molecules that share a common evolutionary ancestor. (general) Two or more gene or protein sequences that share a significant degree of similarity, typically measured by the amount of identity (in the case of DNA), or conservative replacements (in the case of protein), that they register along their lengths. Sequence "homology" searches are typically performed with a query DNA or protein sequence to identify known genes or gene products that share significant similarity and hence might inform on the ancestry, heritage and possible function of the query gene.

## **Housekeeping genes**

Genes that are always expressed (ie. they are said to be constitutively expressed) due to their constant requirement by the cell.

## **Human Anti-Murine Antibody Response (HAMA)**

An immune response generated in humans to antibodies raised in murine (e.g. mouse or rat) cells.

## **Hybridization**

The interaction of complementary nucleic acid strands. This can occur between two DNA strands or between DNA and RNA strands, and is the basis of many techniques such as Southern and northern blots.

## **Hydrogen bond**

A weak chemical interaction between an electronegative atom (e.g. nitrogen or oxygen) and a hydrogen atom that is covalently attached to another atom. This bond maintains the two-helices of DNA together and is also the primary interaction between water molecules.

## **Hydrophilicity**

(lit. water-loving) The degree to which a molecule is soluble in water. Hydrophilicity depends to a large degree on the charge and polarizability of the molecule and its ability to form transient hydrogen-bonds with (polar) water molecules.

## **Hydrophobicity**

(lit. water-hating) The degree to which a molecule is insoluble in water, and hence is soluble in lipids. If a molecule lacking polar groups is placed in water, it will be entropically driven to finding a hydrophobic environment (such as the interior of a protein or a membrane).

I

## **Idiotype**

Antibody variants localized to the variable portion of an immunoglobulin that are recognised by their antigenic determinants. The determinants are composed from the antigen-combining site or CDRs. Every unique antigenic determinant has a specific antibody with its own unique idiotype.

## **Immunoglobulin**

A member of the globulin protein family consisting of two light and two heavy chains linked by disulfide bonds. All antibodies are immunoglobulins.

## **in silico (biology)**

(Lit. computer mediated). The use of computers to simulate, process, or analyse a biological experiment.

## **in situ hybridization**

A variation of the DNA/RNA hybridization procedure in which the denatured DNA is in place in the cell and is then challenged with RNA or DNA extracted from another source. (See also fluorescence in situ hybridization).

## **Integration**

The physical insertion of DNA into the host cell genome. The process is used by retroviruses where a specific enzyme catalyses the process or can occur at random sites with other DNA (eg. transposons).

## **Intracellular signalling**

The communication of a molecular message from the surface of the cell to the nucleus via the participation of a series of molecules, including receptors, enzymes, proteins, and small-molecules. The end result of the signalling process is the up- or down-regulation of a particular series of genes that may be involved in cell growth, division or differentiation.

## **Introns**

Nucleotide sequences found in the structural genes of eukaryotes that are non-coding and interrupt the sequences containing information that codes for polypeptide chains.

Intron sequences are spliced out of their RNA transcripts before maturation and protein synthesis. (cf. Exons)

### **Isoschizomers**

Two different restriction enzymes which recognize and cut DNA at the same recognition site. e.g Sma I and Xma I both recognize and cut the sequence CCCGGG.

### **Isozymes**

Two or more enzymes capable of catalyzing the same reaction but varying in their specificity due to differences in their structures and hence their efficiencies under different environmental conditions.

### **Iteration**

A series of steps in an algorithm whereby the processing of data is performed repetitively until the result exceeds a particular threshold. Iteration is often used in multiple sequence alignments whereby each set of pairwise alignments are compared with every other, starting with the most similar pairs and progressing to the least similar, until there are no longer any sequence-pairs remaining to be aligned.

## **J**

### **Junk DNA**

Term used to describe the excess DNA that is present in the genome beyond that required to encode proteins. A misleading term since these regions are likely to be involved in gene regulation, and other as yet unidentified functions.

## **K**

### **Karyotype**

The constitution (typically number and size) of chromosomes in a cell or individual.

### **Knockout mice (gene targeting)**

Mice which have been engineered to lack a chosen gene. The gene is inactivated in so called embryonic stem cells using the technique of homologous recombination. These cells are then introduced into a early stage embryo (blastocyst) and this is then transplanted into a recipient mouse. The subsequent progeny lack the targeted gene in some cells. This technique is used to determine the function of the chosen gene.

## **L**

### **"Lab on a chip"**

Term describing microdevices that allow rapid, microanalytical analysis of DNA or protein in a single, fully integrated system. Typically, these devices are miniature

surfaces, made of silicon, glass or plastic, which carry the necessary microdevices (pumps, valves, microfluidic controllers, and detectors) that allow sample separation and analysis. These devices are used in drug discovery, genetic testing and separation science.

### **Lead compound**

A candidate compound identified as the best "hit" (tight binder) after screening of a combinatorial (or other) compound library, that is then taken into further rounds of screening to determine its suitability as a drug.

### **Lead optimization**

The process of converting a putative lead compound ("hit") into a therapeutic drug with maximal activity and minimal side effects, typically using a combination of computer-based drug design, medicinal chemistry and pharmacology.

### **Leucine zipper**

Protein motif which binds DNA in which 4-5 Leucines are found at 7 amino acid intervals. This motif is present typically in transcription factors and other proteins that bind DNA.

### **Lexicon**

In Bioinformatics, a lexicon refers to a pre-defined list of terms that together completely define the contents of a particular database. (strict.) The component in the grammar which is in bare form a list of words or lexical entries.

### **Library**

A large collection of compounds, peptides, cDNAs or genes which may be screened in order to isolate cognate molecules.

### **Ligand**

Any small molecule that binds to a protein or receptor; the cognate partner of many cellular proteins, enzymes, and receptors.

### **Linkage**

The association of genes (or genetic loci) on the same chromosome. Genes that are linked together tend to be transmitted together.

### **Linkage map**

A genetic map of a chromosome or genome delineated by mapping the positions of genes to their chromosomes by their linkage to readily identifiable genetic loci.

## **Locus**

The specific position occupied by a gene on a chromosome. At a given locus, any one of the variant forms of a gene may be present. The variants are said to be alleles of that gene.

## **M**

### **Map unit**

A measure of genetic distance between two linked genes that corresponds to a recombination frequency of 1%.

### **Markov chain**

Any multivariate probability density whose independence diagram is a chain. The variables are ordered, and each variable "depends" only on its neighbors in the sense of being conditionally independent of the others. Markov chains are an integral component of hidden Markov models.

### **Meiosis**

A process within the cell nucleus that results in the reduction of the chromosome number from diploid (two copies of each chromosome) to haploid (a single copy) through two reductive divisions in germ cells.

### **Melting (of DNA)**

The denaturation of double-stranded DNA into two single strands by the application of heat. (Denaturation breaks the hydrogen bonds holding the double-stranded DNA together).

### **Messenger RNA (mRNA)**

The complementary RNA copy of DNA formed from a single-stranded DNA template during transcription that migrates from the nucleus to the cytoplasm where it is processed into a sequence carrying the information to code for a polypeptide domain.

### **Methylation**

The addition of -CH<sub>3</sub> (methyl) groups to a target site. Typically such addition occurs on to the cytosine bases of DNA. (see maternal imprinting).

### **Microarray**

A 2D array, typically on a glass, filter, or silicon wafer, upon which genes or gene fragments are deposited or synthesized in a predetermined spatial order allowing them to be made available as probes in a high-throughput, parallel manner.

### **Microfluidics**

The miniaturization of chemical reactions or pharmacological assays into microscopic tubes or vessels in order to greatly increase their throughput, by placing many of them side-by-side in an array.

### **Mimetics**

Compounds that mimic the function of other molecules via their high degree of structural (conformational) similarity, and hence physio-chemical properties.

### **Missense mutation**

A point mutation in which one codon (triplet of bases) is changed into another designating a different amino acid.

### **Mitosis**

The nuclear division that results in the replication of the genetic material and its redistribution into each of the daughter cells during cell division.

### **Modeling**

In bioinformatics, modeling usually refers to molecular modeling, a process whereby the three-dimensional architecture of biological molecules is interpreted (or predicted), visually represented, and manipulated in order to determine their molecular properties. (general) A series of mathematical equations or procedures which simulate a real-life process, given a set of assumptions, boundary parameters, and initial conditions.

### **Monomer**

A single unit of any biological molecule or macromolecule, such as an amino acid, nucleic acid, polypeptide domain, or protein.

### **Monovalent**

Having one binding site; strictly, an atom with only one free electron available for binding in its highest energy shell.

### **Motif**

A conserved element of a protein sequence alignment that usually correlates with a particular function. Motifs are generated from a local multiple protein sequence alignment corresponding to a region whose function or structure is known. It is sufficient that it is conserved, and is hence likely to be predictive of any subsequent occurrence of such a structural/functional region in any other novel protein sequence.

### **Multigene family**

A set of genes derived by duplication of an ancestral gene, followed by independent mutational events resulting in a series of independent genes either clustered together on a chromosome or dispersed throughout the genome.

### **Multiple (sequence) alignment**

A Multiple Alignment of  $k$  sequences is a rectangular array, consisting of characters taken from the alphabet  $A$ , that satisfies the following conditions: There are exactly  $k$  rows; ignoring the gap character, row number  $i$  is exactly the sequence  $S_i$ ; and each column contains at least one character different from "-". In practice multiple sequence alignments include a cost/weight function, that defines the penalty for the insertion of gaps (the "-" character) and weights identities and conservative substitutions accordingly. Multiple alignment algorithms attempt to create the optimal alignment defined as the one with the lowest cost/weight score.

### **Multiplex sequencing**

Approach to high-throughput sequencing that uses several pooled DNA samples run through gels simultaneously and then separated and analyzed.

### **Mutagen**

Any agent that can cause an increase in the rate of mutations in an organism.

### **Mutation**

An inheritable alteration to the genome that includes genetic (point or single base) changes, or larger scale alterations such as chromosomal deletions or rearrangements.

## **N**

### **Naked DNA**

Pure, isolated DNA devoid of any proteins that may bind to it.

### **NCEs (New Chemical Entity)**

Compounds identified as potential drugs that are sent from research and development into clinical trials to determine their suitability .

### **Nested PCR**

The second round amplification of an already PCR-amplified sequence using a new pair of primers which are internal to the original primers. Typically done when a single PCR reaction generates insufficient amounts of product.

### **Neural net**

A neural net is an interconnected assembly of simple processing elements, units or nodes, whose functionality is loosely based on the animal brain. The processing

ability of the network is stored in the inter-unit connection strengths, or weights, obtained by a process of adaptation to, or learning from, a set of training patterns. Neural nets are used in bioinformatics to map data and make predictions, such as taking a multiple alignment of a protein family as a training set in order to identify novel members of the family from their sequence data alone.

### **Nonsense mutation**

A point mutation in which a codon specific for an amino-acid is converted into a nonsense codon.

### **Northern blotting**

A technique to identify RNA molecules by hybridization that is analogous to Southern blotting (see Southern blotting).

### **Nuclease**

Any enzyme that can cleave the phosphodiester bonds of nucleic acid backbones.

### **Nucleoside**

A five-carbon sugar covalently attached to a nitrogen base.

### **Nucleotide**

A nucleic acid unit composed of a five carbon sugar joined to a phosphate group and a nitrogen base.

## **O**

### **Object-Relational Database**

Object databases combine the elements of object orientation and object-oriented programming languages with database capabilities. They provide more than persistent storage of programming language objects. Object databases extend the functionality of object programming languages (e.g., C++, Smalltalk, or Java) to provide full-featured database programming capability. The result is a high level of congruence between the data model for the application and the data model of the database. Object-relational databases are used in Bioinformatics to map molecular biological objects (such as sequences, structures, maps and pathways) to their underlying representations (typically within the rows and columns of relational database tables.) This enables the user to deal with the biological objects in a more intuitive manner, as they would in the laboratory, without having to worry about the underlying data model of their representation.

### **Oligonucleotide**

A short molecule consisting of several linked nucleotides (typically between 10 and 60) covalently attached by phosphodiester bonds.

## **Open reading frame (ORF)**

Any stretch of DNA that potentially encodes a protein. Open reading frames start with a start codon, and end with a termination codon. No termination codons may be present internally. The identification of an ORF is the first indication that a segment of DNA may be part of a functional gene.

## **Operator**

A segment of DNA that interacts with the products of regulatory genes and facilitates the transcription of one or more structural genes.

## **Operon**

A unit of transcription consisting of one or more structural genes, an operator, and a promoter.

## **Ortholog**

Orthologs are genes in different species that evolved from a common ancestral gene by speciation. Normally, orthologs retain the same function in the course of evolution. Identification of orthologs is critical for reliable prediction of gene function in newly sequenced genomes. (See also Paralogs.)

## **Overlapping clones**

Collection of cloned sequences made by generating randomly overlapping DNA fragments with infrequently cutting restriction enzymes.

## **P**

## **Palindrome**

A region of DNA with a symmetrical arrangement of bases occurring about a single point such that the base sequences on either side of that point are identical (if the strands are both read in the same direction) e.g 5' GAATTC 3' whose complementary sequence is 3' CTTAAG 5'.

## **Pattern**

Molecular biological patterns usually occur at the level of the characters making up the gene or protein sequence. A pattern language must be defined in order to apply different criteria to different positions of a sequence. In order to have position-specific comparison done by a computer, a pattern-matching algorithm must allow alternative residues at a given position, repetitions of a residue, exclusion of alternative residues, weighting, and ideally, combinatorial representation.

## **Pathways**

Bioinformatics strives to define representations of key biological datatypes, algorithms and inference procedures, including sequences, structures, biological pathways and reactions. Representing and computing with biological pathways requires ontologies for representing pathway knowledge; User interfaces to these databases; Physico-chemical properties of enzymes and their substrates in pathways; And pathway analysis of whole genomes including identifying common patterns across species and species differences.

### **Paralog**

Paralogs are genes related by duplication within a genome. Orthologs retain the same function in the course of evolution, whereas paralogs evolve new functions, even if these are related to the original one.

### **Parameters**

Parameters are user-selectable values, typically experimentally determined, that govern the boundaries of an algorithm or program. For instance, selection of the appropriate input parameters governs the success of a search algorithm. Some of the most common search parameters in bioinformatics tools include the stringency of an alignment search tool, and the weights (penalties) provided for mismatches and gaps.

### **Peptide**

A short stretch of amino acids each covalently coupled by a peptide (amide) bond.

### **Peptide bond (amide bond)**

A covalent bond formed between two amino acids when the amino group of one is linked to the carboxy group of another (resulting in the elimination of one water molecule).

### **Phage (Bacteriophage)**

A virus that infects bacterial cells and serves as a useful vector for introducing genes into bacteria for a number of purposes.

### **Phage display**

A technique in which phage are engineered to fuse a foreign peptide or protein with their capsid (surface) proteins and hence display it on their cell surfaces. The immobilized phage may then be used as a screen to see what ligands bind to the expressed fusion protein exhibited (displayed) on the phage surface.

### **Pharmacogenomics**

The use of (DNA-based) genotyping in order to target pharmaceutical agents to specific patient populations. Genetic differences are known to affect responses to many types of drug therapy, and pharmacogenomics analysis serves to customize the use of pharmaceuticals for specific subgroups of patients. The rationale for this

approach is that observed gene expression differences may correlate with, and explain, the differences in side effects and efficacy to drugs in humans.

### **Pharmacophore**

The three dimensional spatial arrangement of atoms, substituents, functional groups, or chemical features that together are sufficient to describe the pharmacologically active components of a drug molecule or molecule series.

### **Phenotype**

Any observable feature of an organism that is the result of one or more genes.

### **Phylum**

The segmentation of the animal kingdom into about 30 major groups collectively known as phyla. The members of each phylum share the same basic structure and organization. For instance, fish, birds, and human beings belong to one phylum - the Chordata - because all have spinal cords.

### **Physical map**

A physical map consists of a linearly ordered set of DNA fragments encompassing the genome or region of interest. Physical maps are of two types, macro-restriction maps and ordered clone maps. The former consists of an ordered set of large DNA fragments generated by using restriction enzymes whose recognition sequences are infrequently represented in the genome. An ordered clone map consists of an overlapping collection of cloned DNA fragments. The DNA may be cloned into any one of the available vector systems--YACs, cosmids, phage, or even plasmids. Major advantages of ordered clone maps are that they are of high resolution and directly provide the clones for further study.

### **Plasmid**

Any replicating DNA element that can exist in the cell independently of the chromosomes. Synthetic plasmids are used for DNA cloning. Most commonly found in bacterial cells.

### **Pleiotropy**

The multiple effects on an organism's phenotype due to a single gene or allele e.g the cytokines which can bind to multiple cellular receptors and effect growth and multiple immune pathways.

### **Point mutation**

A mutation in which a single nucleotide in a DNA sequence is substituted by another nucleotide.

## **Poly(A) tail**

The stretch of Adenine (A) residues at the 3' end of eukaryotic mRNA that is added to the pre-mRNA as it is processed, before its transport from the nucleus to the cytoplasm and subsequent translation at the ribosome.

## **Polyadenylation site**

A site on the 3'-end of messenger RNA (mRNA) that signals the addition of a series of Adenines during the RNA processing step and before the mRNA migrates to the cytoplasm. These so-called poly(A) "tails" increase mRNA stability and allow one to isolate mRNA from cells by PCR-amplification using poly(T) primers.

## **Polygenic inheritance**

Inheritance involving alleles at many genetic loci.

## **Polymerase chain reaction (PCR )**

Technique used to amplify or generate large amounts of replica DNA of a segment of any DNA whose "flanking" sequences are known. Oligonucleotide primers which bind these flanking sequences are used by an enzyme (Taq polymerase) to copy the sequence in between the primers. Cycles of heat to break apart the DNA strands, cooling to allow the primers to bind, and heating again to allow the enzyme to copy the intervening sequence lead to a doubling of DNA at each cycle. The reactions are typically carried out on a regulated heating block and consist of 30-35 cycles of repeated amplification of all the DNA present. Single molecules of "target" DNA can be amplified to microgram amounts of DNA. The target DNA can be of any origin.

## **Polymorphism**

(lit. many forms) The existence of a gene in a population in at least two different forms at a frequency far higher than that attributable to recurrent mutation alone. Variations in a population may be measured by determining the rate of mutation in polymorphic genes (see SNPs).

## **Polypeptide**

A single chain of covalently attached amino acids joined by peptide bonds. Polypeptide chains usually fold into a compact, stable form (a domain) that is part (or all) of the final protein.

## **Positional cloning**

Method used to define the location of a gene on a chromosome and use this information to identify and clone the gene. The location of the gene is determined by linkage analysis of DNA from a large family containing afflicted and normal members to identify linkages between the transmission of the disease gene and observable genetic markers. This information is then used to screen (by chromosomal jumping and walking) the location for putative genes. The disease gene must be

compared between the afflicted and normal family members and be shown to be different in the two groups. The full sequencing of the gene will then provide information regarding the characteristics and function of the gene product, and a potential explanation for the cause of the disease.

### **Post-transcriptional modification**

Alterations made to pre-mRNA before it leaves the nucleus and becomes mature mRNA.

### **Post-translational modification**

Alterations made to a protein after its synthesis at the ribosome. These modifications, such as the addition of carbohydrate or fatty acid chains, may be critical to the function of the protein.

### **Primary sequence (protein)**

The linear sequence of a polypeptide or protein.

### **Primary structure (protein)**

see primary sequence.

### **Primer**

A short oligonucleotide that provides a free 3' hydroxyl for DNA or RNA synthesis by the appropriate polymerase (DNA polymerase or RNA polymerase).

### **Probe**

Any biochemical that is labelled or tagged in some way so that it can be used to identify or isolate a gene, RNA, or protein.

### **Profile**

Sequence profiles are usually derived from multiple alignments of sequences with a known relationship, and consist of tables of position-specific scores and gap-penalties. Each position in the profile contains scores for all of the possible amino acids, as well as one penalty score for opening and one for continuing a gap at the specified position. Attempts have been made to further improve the sensitivity of the profile by refining the procedures to construct a profile starting from a given multiple alignment. Other representations for sequence domains or motifs do not necessarily require the presence of a correct and complete multiple alignment, such as hidden Markov models.

### **Prokaryote**

An organism or cell that lacks a membrane-bounded nucleus. Bacteria and blue-green algae are the only surviving prokaryotes (cf. Eukaryote).

## **Promoter (site)**

A promoter site is defined by its recognition by eukaryotic RNA polymerase II; its activity in a higher eukaryote; by experimental evidence, or homology and sufficient similarity to an experimentally defined promoter; and by observed biological function.

## **Protein families**

Sets of proteins that share a common evolutionary origin reflected by their relatedness in function which is usually reflected by similarities in sequence, or in primary, secondary or tertiary structure. Subsets of proteins with related structure and function.

## **Proteome**

The entire protein complement of a given organism.

## **Proteomics**

The study of the proteome. Typically, the cataloging of all the expressed proteins in a particular cell or tissue type, obtained by identifying the proteins from cell extracts using a combination of 2D gel electrophoresis and mass spectrometry. The large scale analysis of the protein composition and function. (cf genomics)

## **Purine**

A nitrogen-containing compound with a double-ring structure. The parent compound of Adenine and Guanine.

## **Pyrimidine**

A nitrogen-containing compound with a single six-membered ring structure. The parent compound of Thymidine and Cytosine.

## **Q**

## **Query (sequence)**

A DNA, RNA or protein sequence used to search a sequence database in order to identify close or remote family members (homologs) of known function, or sequences with similar active sites or regions (analogs), from whom the function of the query may be deduced.

## **R**

## **Rational drug design (Structure based drug design)**

The development of drugs based on the 3-dimensional molecular structure of a particular target.

## **Reading frame**

A sequence of codons beginning with an initiation codon and ending with a termination codon, typically of at least 150 bases (50 amino acids) coding for a polypeptide or protein chain (see ORF and URF).

## **Reagents**

Sources of biological or chemical material that can be used as the starting blocks in laboratory experiments. Reagents can range from chemicals needed to perform a particular chemical reaction, constituents of a laboratory protocol, or clones to be used in a large-scale gene expression study.

## **Recessive**

Any trait that is expressed phenotypically only when present on both alleles of a gene (cf dominant).

## **Recombinant DNA (rDNA)**

DNA molecules resulting from the fusion of DNA from different sources. The technology employed for splicing DNA from different sources and for amplifying the resultant heterogenous DNA.

## **Recombination**

A new combination of alleles resulting from the rearrangement occurring by crossing-over or by independent assortment (see crossing over).

## **Recursion**

An algorithmic procedure whereby an algorithm calls on itself to perform a calculation until the result exceeds a threshold, in which case the algorithm exits. Recursion is a powerful procedure with which to process data and is computationally quite efficient.

## **Regulatory gene**

A DNA sequence that functions to control the expression of other genes by producing a protein that modulates the synthesis of their products (typically by binding to the gene promoter). (cf. Structural gene).

## **Relational Database**

A database that follows E. F. Codd's 11 rules, a series of mathematical and logical steps for the organization and systemization of data into a software system that allows easy retrieval, updating, and expansion. An RDBMS stores data in a database consisting of one or more tables of rows and columns. The rows correspond to a record (tuple); the columns correspond to attributes (fields) in the record. In an RDBMS, a view, defined as a subset of the database that is the result of the evaluation

of a query, is a table. RDBMSs use Structured Query Language (SQL) for data definition, data management, and data access and retrieval. Relational and object-relational databases are used extensively in bioinformatics to store sequence and other biological data.

### **Relational Database Management Systems (RDBMS)**

A software system that includes a database architecture, query language, and data loading and updating tools and other ancillary software that together allow the creation of a relational database application.

### **Repeats (repeat sequences)**

Repeat sequences and approximate repeats occur throughout the DNA of higher organisms (mammals). For example, the *Alu* sequences of length about 300 characters, appear hundreds of thousands of times in Human DNA with about 87% homology to a consensus *Alu* string. Some short substrings such as TATA-boxes, poly-A and (TG)\* also appear more often than by chance. Repeat sequences may also occur within genes, as mutations or alterations to those genes. Repetitive sequences, especially mobile elements, have many applications in genetic research. DNA transposons and retrotransposons are routinely used for insertional mutagenesis, gene mapping, gene tagging, and gene transfer in several model systems.

### **Repetitive elements**

Repetitive elements provide important clues about chromosome dynamics, evolutionary forces, and mechanisms for exchange of genetic information between organisms. The most ubiquitous class of repetitive elements in the DNA sequence in primate genomes is the *Alu* family of interspersed repeats which have arisen in the last 65 million years of evolution. *Alu* repeats belong to a class of sequences defined as short interspersed elements (SINEs). Approximately 500,000 *Alu* SINEs exist within the human genome, representing about 5% of the genome by mass.

### **Replication**

The synthesis of an informationally identical macromolecule (e.g. DNA) from a template molecule.

### **Repressor**

The protein product of a regulatory gene that combines with a specific operator (regulatory DNA sequence) and hence blocks the transcription of genes in an operon.

### **Restriction enzyme (restriction endonuclease)**

A type of enzyme that recognizes specific DNA sequences (usually palindromic sequences 4, 6, 8 or 16 base pairs in length) and produces cuts on both strands of DNA containing those sequences only. The "molecular scissors" of rDNA technology.

### **Restriction fragment length polymorphisms (RFLPs)**

Variation within the DNA sequences of organisms of a given species that can be identified by fragmenting the sequences using restriction enzymes, since the variation lies within the restriction site. RFLPs can be used to measure the diversity of a gene in a population.

### **Restriction map**

A physical map or depiction of a gene (or genome) derived by ordering overlapping restriction fragments produced by digestion of the DNA with a number of restriction enzymes.

### **Reverse Genetics**

The use of protein information to elucidate the genetic sequence encoding that protein. Used to describe the process of gene isolation starting with a panel of afflicted patients (see positional cloning) .

### **Reverse transcriptase**

A DNA polymerase that can synthesise a complementary DNA (cDNA) strand using RNA as a template - a so-called RNA-dependent DNA polymerase.

### **Reverse transcriptase-PCR (RT-PCR)**

Procedure in which PCR amplification is carried out on DNA that is first generated by the conversion of mRNA to cDNA using reverse transcriptase.

### **Ribonucleic acid (RNA)**

A category of nucleic acids in which the component sugar is ribose and consisting of the four nucleotides Thymidine, Uracil, Guanine, and Adenine. The three types of RNA are messenger RNA (mRNA), transfer RNA (tRNA) and ribosomal RNA (rRNA).

## **S**

### **Secondary structure (protein)**

The organization of the peptide backbone of a protein that occurs as a result of hydrogen bonds e.g alpha helix, Beta pleated sheet.

### **Selectivity**

Selectivity of bioinformatics similarity search algorithms is defined as the significance threshold for reporting database sequence matches. As an example, for BLAST searches, the parameter E is interpreted as the upper bound on the expected frequency of chance occurrence of a match within the context of the entire database search. E may be thought of as the number of matches one expects to observe by chance alone during the database search.

## **Sense strand**

The strand of double-stranded DNA that acts as the template strand for RNA synthesis. Typically only one gene product is produced per gene, reading from the sense strand only. (Some viruses have open reading frames in both the sense and the antisense strands).

## **Sensitivity**

Sensitivity of bioinformatics similarity search algorithms centers around two areas: First, how well can the method detect biologically meaningful relationships between two related sequences in the presence of mutations and sequencing errors; Secondly how does the heuristic nature of the algorithm affect the probability that a matching sequence will not be detected. At the user's discretion, the speed of most similarity search programs can be sacrificed in exchange for greater sensitivity - with an emphasis on detecting lower scoring matches.

## **Sequence Tagged Site (STS)**

A unique sequence from a known chromosomal location that can be amplified by PCR. STSs act as physical markers for genomic mapping and cloning.

## **Sexual PCR (Molecular Diversity)**

Sexual PCR is a form of PCR in which similar, but not identical, DNA sequences are reassembled to obtain novel juxtapositions, simulating the result of genetic recombination. The result is the creation of an array of related genes which may possess improved characteristics. By repeated rounds of recombination, selection and PCR-based amplification vastly improved gene-products, such as enzymes with greater activity, may be generated and selected.

[WWW.IBP.IR](http://WWW.IBP.IR)

iranian bioinformatics portal

[babakbabaabasi@gmail.com](mailto:babakbabaabasi@gmail.com)