

Glossary

This glossary defines terms that are used in several parts of the book and that may be new to the reader. The definitions are stated to the level of precision needed for this book, and some may be incomplete compared to a glossary for a biology text. Very familiar terms are omitted, as are specialized terms that are only used in the same section of the book where they are defined. When a particular term is discussed or defined more completely in the body of the book, the page where that discussion begins is given in parentheses after the definition.

- allele** One of two or more forms that a substring of DNA (often a gene) can take on.
- algorithm** A high level description of a mechanistic way to solve a problem or compute a function. The description must lay out the logic of the method, but should avoid many low-level programming details needed to implement it on a computer. Often, in the biological literature, “algorithm” and “program” are used interchangeably, but this is not correct. A program contains all the implementation detail needed to make the method work in a specific computer programming language on specific computers. That level of detail usually obscures the logic and the ideas behind the algorithm.
- alpha helix** Helical structure in protein. The alpha helix is one of two common secondary structures in protein. The other is the beta sheet. (pages 248, 361)
- amino acid** Molecules that form the building blocks of proteins. There are twenty common amino acids found in proteins. Each amino acid is coded in DNA by a “codon” (see genetic code and codon). (page 14)
- amino acid alphabet** A twenty-character alphabet consisting of the characters A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y, each representing one of the twenty amino acids coded for by DNA.
- amino acid substitution matrix** A matrix specifying the scores to use for character-specific matches and mismatches. The two most widely used classes of amino acid substitution matrices are the PAM matrices and the BLOSUM matrices. (pages 381 and 386)
- analogous protein** See homologous protein.
- antibody** A protein that binds to a foreign molecule. Also known as an **immunoglobulin**. They are used in nature to identify and destroy foreign bodies. They are used in biochemistry as assays to identify specific proteins.
- antigen** A molecule that is recognized as foreign and causes a response by antibodies.
- autosome** Any chromosome other than a sex chromosome. In humans, this is any chromosome other than the X and Y chromosomes.
- base** In the context of molecular biology, a base refers to a single nucleotide (DNA or RNA). The DNA bases are abbreviated A, T, C, and G. The RNA bases are abbreviated A, U, C, and G. See nucleotide.
- base pair** Two bases (nucleotides) that bond to form **complementary** pairs. In DNA these are A and T; C and G. In RNA they are A and U; C and G. These pairs are also called **WATSON-CRICK** pairs.
- beta sheet** One of two common secondary structures in protein. The other is the alpha helix.
- Big-Oh O and related notation** For two non-negative functions $f(n)$ and $g(n)$, if $\lim_{n \rightarrow \infty} \frac{f(n)}{g(n)} < \infty$, then we say $f(n) = O(g(n))$. Intuitively (but not exactly) this means that $f(n)$ grows no faster than $g(n)$, ignoring small values of n and any multiplicative constants.
- If $\lim_{n \rightarrow \infty} \frac{g(n)}{f(n)} < \infty$, then we say $f(n) = \Omega(g(n))$. Intuitively (but not exactly) this means that $f(n)$ grows at least as fast as $g(n)$, ignoring small values of n and any multiplicative constants.
- If $f(n) = O(g(n))$ and $f(n) = \Omega(g(n))$, then we say that $f(n) = \Theta(g(n))$. Intuitively

(but not exactly) this means that $f(n)$ grows at the same rate as $g(n)$, ignoring small values of n and any multiplicative constants. See [32] or [112] for more detail.

bioinformatics A made-up word that refers to a wide range of computational efforts in the human, and other, genome projects. It encompasses data base development, communication software, sequence analysis, development of automated laboratory notebooks, development of standards, and more. The great virtue of the term is that it has no previous meaning so that it can be almost anything. The downside is that people who do informatics are called "bioinformaticists". Although the term is very inclusive, many people identified as bioinformaticists self-label themselves as being in "computational molecular biology".

catalyst A molecule that enables or accelerates a specific chemical reaction. Proteins are the primary catalysts in the cell, but RNA can act as a catalyst as well.

cDNA Complementary or copy DNA. This is DNA made from messenger RNA (through reverse transcription). When the mRNA is from eukaryotes, the cDNA lacks the introns that are contained in the DNA substring from which the mRNA is transcribed. (page 238)

chromosome Large linear or circular structure made of DNA in complex with certain proteins. The hereditary information of an organism is organized into a number of chromosomes.

cladistics A particular systematic approach to building evolutionary trees and interpreting their meaning. Generally, the cladistic approach attempts to reconstruct the branching pattern (order) of the evolutionary tree but not estimate the passage of time on any particular branch.

clone A population of organisms, genes, or molecules consisting of many exact copies of the originating entity. As a verb, "to clone" means to make many copies. To "clone a sequence" also refers to the process of extracting a specific DNA sequence or gene from a larger body of DNA.

clone library See DNA library.

cloning vector An organism, often a virus or a plasmid, or more recently an artificial chromosome (for example, a YAC), that carries a fragment of DNA into a host cell that can clone (replicate) the inserted DNA fragment along with its own DNA. (page 416)

coding region An interval of DNA containing the codons that specify a particular protein.

In Eukaryotes, the coding region is broken into disjoint exons. However, coding regions for different proteins can sometimes overlap; hence an interval should be defined as a coding region for a specific protein. (page 238)

codon A string of three nucleotides in DNA or mRNA that specifies a particular amino acid. See also **genetic code** and **translation**. Two codons are called "synonymous" if they specify the same amino acid. See genetic code. (page 14)

complementary base In DNA, the complementary base pairs are *A, T* and *C, G*. In RNA they are *A, U* and *C, G*. See also **base pairs**.

complementary nucleotide sequence A nucleotide sequence *S* is the complementary sequence of *S'* if each base in *S* is the complementary base of the corresponding base of *S'*.

cosmid An artificially constructed cloning vector capable of holding inserts of lengths around 40,000 bases. (pages 416 and 419)

crossing over The process, during meiosis, where two homologous chromosomes align together, break, and exchange fragments of DNA to form two hybrid chromosomes. Each hybrid chromosome contains alternating intervals of DNA from the two original chromosomes. Normally in a pair of exchanged fragments, each fragment in the pair is the same length as the other. But in **unequal crossing over** one fragment is larger than the other, so the two new chromosomes are of unequal length. See **recombination**.

denaturation A large disruption of the natural structure of protein or DNA due to heating or exposure to chemicals. In double-stranded DNA, denaturation often refers to splitting apart of the two strands into two single strands. In protein, it often refers to destroying the three-dimensional structure of the protein.

Depth-First Search (dfs) A very common recursive algorithm used to explore a graph. In this book it is only used to explore trees. Depth-first search first visits the root of the tree (the first "current node") and then successively executes a *dfs* computation on the tree rooted at each child of the current node. When all those recursive executions are completed, the search backs up from the current node to its parent node. See [10], [32], [112] or [401] for more details.

diploid An organism is diploid if it contains two copies of each chromosome. The copies need not be identical, as each can contain different alleles. In humans, most cells other than

- germ-line cells (sperm and egg) are diploid. See also **haploid**.
- DNA** A chain of nucleotides (bases) in a single molecule. The bases of DNA are **adenine (A)**, **thymine (T)**, **cytosine (C)**, and **guanine (G)**. DNA is the basic carrier of genetic information. DNA usually consists of two strands of complementary nucleotide sequences that are base paired (Watson-Crick) to each other. Hence, in describing the two strands, one typically only specifies one of the strands and its orientation. DNA in humans forms a linear chain, but DNA can also form a circular molecule (see **plasmid**). (page 14)
- DNA library** A physical collection of unordered cloned fragments of DNA, possibly cDNA obtained from mRNA. The fragments can come from the entire genome, but often the fragments in a particular DNA library come from a particular tissue or chromosome region and represent only a subset of the entire DNA of an organism.
- DNA sequencing** The process of determining the complete base pair sequence of a target DNA string. (page 415)
- Drosophila melanogaster** Fruit fly commonly used in genetic studies.
- enhancer sequence** A DNA sequence that binds to regulatory proteins. This binding affects the rate at which certain intervals of DNA are transcribed, and hence it affects the rate at which certain proteins are made (expressed). The concept of an enhancer sequence is often confused with a **promoter** sequence, but unlike the latter, the enhancer sequence can be very far away from the DNA that it regulates. The opposite of an enhancer is a **repressor**.
- electrophoresis** A process in which molecules (DNA, RNA, or protein) with different properties (such as charge, length, or size) are separated. The molecules are put into a gel and an electric field (sometimes alternating) is applied. Molecules with different properties move through the gel at different rates in response to the field.
- Escherichia coli (E. coli)** Bacteria found in the human gut. *E. coli* is one of the widely studied model organisms in genetics and molecular biology.
- enzyme** A protein catalyst, working to accelerate specific chemical reactions in the cell.
- EST (expressed sequence tag)** An STS derived from a cDNA molecule. Therefore, an EST is an STS that is found in a gene, rather than in a noncoding region of DNA. (page 61)
- eukaryote** An organism whose DNA is enclosed in a nucleus. Includes all "higher-order" life. Often in this book, the important feature of a eukaryote is that its genes are broken into alternating exons and introns. Contrast with **prokaryote**.
- exon** In Eukaryotes, genes are typically broken up into alternating regions of exons and introns. The exons contain information that is represented in the messenger RNA (mRNA) and ultimately used to produce a protein (or sometimes RNA). (page 237)
- exon-intron splice site** A point in a eukaryotic gene where an intron adjoins an exon. Nature knows how to identify and splice out the introns from transcribed RNA, but no complete explanation of how splice sites are recognized is known. (pages 238, 248)
- expression** A gene is expressed if the protein it codes for is produced. Sometimes we also say a protein is expressed.
- extant organism** The word "extant" means currently existing. I've found that this term is not well known to computer scientists, and I know of no computer science setting where it is used. It is a standard word in evolutionary biology.
- fushi tarazu** One of the coolest names for a mutant allele in *Drosophila*, an organism where many mutants have cool names. Japanese for "missing a stripe".
- gene** A contiguous interval of DNA that contains the information needed to code for a protein, or less often, for some RNA. Genes form the basic units of heredity.
- gene regulatory protein** A protein that binds to DNA to affect the expression of a gene.
- genetic code** The correspondence between specific nucleotide triplets (**codons**) and specific amino acids. Since there are $4^3 = 64$ codons but only 20 standard amino acids, some amino acids are coded by more than one codon. Therefore, the genetic code is said to be **degenerate**. For example, the amino acid alanine is specified by codons GCA, GCC, GCG, GCU. There are also three codons that do not code for any amino acid, but usually specify the end of the polypeptide. These are called **stop codons**. (page 14)
- genome** The entire genetic information contained in an organism.
- genotype** The description of a specific organism in terms of its genome. This is opposed to **phynotype**, which is a description of an organism in terms of its expressed features.
- germ cells** In humans, sperm or egg cells.

- These cells are haploid, containing only a single copy of each chromosome. See also **somatic cells**.
- globular protein** A protein that forms a round shape.
- haploid** A cell containing only one copy of each chromosome.
- hemoglobin** The primary functional protein of red blood cells. It is involved with binding and transporting oxygen.
- homeobox** A DNA sequence to which regulatory proteins bind, affecting major features of the organism's development. These sequences are highly conserved across many species. (page 230)
- homologous chromosomes** The two copies of the same chromosome in a diploid organism.
- homologous protein** Two proteins that are related by common evolutionary history. For example, human cytochrome c is homologous to duckbill platypose cytochrome c. However, sometimes the term is used even when the shared history is unclear. Two proteins in different organisms that have common biological or chemical features (function, structure, motifs) are often referred to as "homologous", but the word "analogous" seems more appropriate, expressing similarity of the proteins without any implied cause. Unfortunately, in some subareas of biology the phrase "analogous protein" has been defined to mean that the two proteins are similar due to *convergent* evolution, i.e., they are known to have no significant shared history. This deviation from normal English robs biology of an important common word. Sometimes biologists will use the phrase "same protein" (in quotes) to capture the normal English meaning of "analogous protein". In this book, I have tried to use the term "homologous protein" only when a shared history is implied.
- homology** A similarity due to common evolutionary history. "Sequence homology" is sometimes used interchangeably with "sequence similarity", although the later term does not imply a common evolutionary history. A phrase such as "degree of homology" is commonly used, but it is not meaningful under the strict definition given above.
- hybridization** The base-pairing (bonding) of two complementary DNA or RNA molecules. Hybridization of a single stranded **probe** to a longer sequence of DNA is often used to determine where the complement of the probe resides in the longer sequence.
- intron** See **exon**.
- linkage** The degree to which two markers in DNA are inherited together. Linkage provides a crude reflection of physical distance. The closer two markers are on a single chromosome, the more likely they are to be inherited together and not separated by a crossing-over (or recombination) event.
- linkage map** Same as genetic map.
- leucine zipper** A common motif for a transcription factor. (page 62).
- ligate** To join together two molecules.
- ligand** A molecule that binds to a specific location on another molecule.
- marker** A feature of a chromosome that can be detected.
- meiosis** The process during which germ cells (sperm or egg) are created from parent cells. Four haploid cells are created from one diploid cell during two rounds of cell division. Typically, during meiosis each pair of homologous chromosomes come together, cross over/recombine, and replicate, creating four hybrid chromosomes.
- messenger RNA (mRNA)** The template RNA that codes for a single protein. Each codon of the mRNA specifies a single amino acid in the protein sequence. The mRNA is processed by a **ribosome**, which, with the aid of **transfer RNA**, strings together the prescribed amino acids of the protein.
- mitochondria** Bacteria-sized organelle inside eukaryotic cells. It is the energy center of the cell and contains its own DNA. It is an unsolved puzzle how it came to co-exist with eukaryotic cells. Mitochondrial DNA is inherited only from one's mother, and mitochondrial DNA sequences are often used to infer evolutionary history.
- nucleic acid** Essentially a chain of nucleotides; an RNA or DNA molecule.
- nucleotide** Adenine, thymine, cytosine, or guanine in DNA; Adenine, uracil, cytosine, or guanine in RNA. See base.
- oligonucleotide** (or **oligo** for short) A short nucleic acid chain (DNA or RNA).
- oncogene** A gene that acts to promote the development of cancer, when active. In contrast, an **antioncogene** is a gene that normally acts to suppress the development of cancer, allowing cancer when inactive.
- open reading frame (ORF)** A substring in DNA that contains no stop codons (*UAA*, *UAG*, or *UGA*) when read in a single reading frame. Sometimes an ORF is alternately defined as a substring in DNA between the start codon *AUG* and the first stop codon

found in the same reading frame as the start codon. That definition is more appropriate for prokaryotes and is confusing for eukaryotes, because eukaryotic genes are sometimes described as containing several ORFS [192], meaning several exons. A single exon does satisfy the first definition of an ORF, but need not satisfy the alternate definition. However, there seems to be no absolute standard, and two different chapters of [192] use the two different definitions, even though both chapters concern eukaryotic genes. Identifying the open reading frames is often the first step in trying to locate genes in a stretch of anonymous DNA.

PAM Short for “point accepted mutation” or “percent accepted mutations”. PAM is used as a unit of evolutionary distance and also as an identifier of specific amino acid substitution matrices. (page 381)

palindromic sequence in DNA A DNA string that becomes a palindrome in the normal English use of the word after half of its string is replaced by its complementary sequence. (page 138)

polymerase chain reaction (PCR) A process used to make copies (or amplify) the DNA in an interval of DNA defined between two **primer** sequences. The DNA can be *in vitro* (i.e., outside of a living cell), and the two primer sequences are chosen by the experimenter. However, the primers must be located within an “acceptable” distance from each other. PCR proceeds in cycles, and in each cycle the number of copies of the defined interval of DNA roughly doubles. The ability to quickly and cheaply make essentially unlimited quantities of DNA between any two (relatively close) and user-defined points on a DNA molecule, has revolutionized the practice of molecular biology. PCR has been described as “being to genes what Gutenberg’s printing press was to the written word” [90]. The inventor of PCR, Kary Mullis, was awarded the Nobel Prize in chemistry in 1994. A very amusing account by Mullis of that invention appears in [333].

phenotype The total, observable set of features of an organism. See also **genotype** for contrast.

phylogeny The tree-like evolutionary history of a set of taxa. (Chapter 17)

plasmid A circular molecule of DNA found outside the normal genome of the organism. Typically found in bacteria. Often used as a cloning vector. (page 12).

point mutation A single change of one nucleotide (in DNA or RNA) or one amino acid residue.

polypeptide A long chain of linked amino acids. Sometimes used for “protein”.

polymere A long linear molecule made up of identical or similar subunits.

positional cloning A fairly recent approach to identifying disease-related genes. First, using genetic mapping, find an interval in the genome likely to contain the gene; then sequence parts of the interval in disease afflicted individuals and unaffected relatives to find systematic differences that indicate the desired gene has been found. (page 397)

prokaryote An organism whose DNA is not enclosed in a nucleus. Contrast with **eukaryote**.

primary structure The sequence description of a molecular (DNA, RNA, or amino acid) string.

promoter A substring in DNA upstream of a gene, where the RNA polymerase (which helps transcribe the DNA to RNA during transcription) binds to the DNA.

polymerase An enzyme that facilitates the creation of a nucleic acid molecule complementary to an existing single-stranded nucleic acid molecule (called a template).

protease An enzyme that cuts protein.

protein A polypeptide, a chain of linked amino acids. The structural material and workhorse molecule of the cell (DNA proposes, but protein disposes). Enzymes are protein catalysts, working to enable or accelerate chemical reactions in the cell. Until the late 1940s and early 1950s it was generally believed that proteins would also be found to be the molecule encoding hereditary information. It was very surprising when DNA was shown to play that role.

proto-oncogene A normal gene that can be converted to an oncogene by one of several mechanisms.

purine In the context of this book, a purine is either adenine (A) or guanine (G). How do you remember this? Biochemistry students at U. C. Davis (the AGgies) are taught the mnemonic: *AGgies are PURE*.

pyrimidine In the context of this book, a pyrimidine is either cytosine (C) or thymine (T).

rate-limiting step One of several favorite phrases used in biochemistry but not in computer science. In computer science we would call it a “bottleneck”.

reading frame One of three places to start

- reading when translating a string from the DNA alphabet into the amino acid alphabet. If the direction of the string is also not established, then one often refers to six reading frames. (page 14)
- recombination** A general term for several mechanisms resulting in the breaking or cutting and resplicing of intervals of DNA. Occurs both naturally (for example, during meiosis via crossing-over) and in the laboratory.
- repressor** A protein that reduces (or eliminates) the expression of another protein by interfering with the transcription of its DNA. A typical mechanism is for the repressor to bind close to the DNA encoding the repressed protein, hence interfering with its transcription.
- residue** A single unit in a polymer. Used both for a single nucleotide in a DNA molecule or a single amino acid in a protein.
- restriction enzyme** An enzyme that cuts DNA at locations containing specific short (usually palindromic) DNA sequences.
- retrovirus** A virus with an RNA genome that must be transcribed to a double-stranded DNA molecule in order to reproduce.
- reverse transcription** The process of creating a double-stranded DNA molecule encoding a sequence from a single-stranded RNA molecule.
- reverse transcriptase** Enzyme that creates a double-stranded DNA "copy" of a single-stranded RNA molecule.
- ribosome** A complex made of RNA and protein where the protein defined by a messenger RNA is synthesized.
- RNA** ribonucleic acid molecule. See **nucleic acid**.
- satellite DNA** Short DNA substrings that are highly repetitive in a genome. These are further subdivided into mini- and micro-satellite by the length of the repeated substring. (pages 140 and 138)
- sequence tagged site (STS)** Roughly, a short DNA sequence that occurs only once in the genome. More exactly, a pair of PCR primers within a bounded distance, with the property that PCR succeeds using those primers at only one location in the genome. STSs provide markers throughout the genome, but they need not be located in genes. See also **EST**. (pages 61 and 398)
- silent mutation** A mutation in a DNA codon that does not change the specified amino acid. Most often, a silent mutation is in the third nucleotide in the codon. For example, TCN codes for the amino acid serine, where N is any of the four DNA nucleotides. So a mutation in the third nucleotide is a silent mutation.
- start codon** Codon that signals the start of a sequence to be translated to protein. Frequently *AUG*, but it can vary in different organisms.
- stop codon** Codon that signals the end of a sequence to be translated to protein. Frequently *UAA*, *UAG* or *UGA*.
- TATA box** The name given to the commonly occurring substring "TATA" that appears in the promoter region of many genes.
- telomere** The DNA forming each end of a chromosome. It contains highly repetitive short substrings. (page 138)
- transcription** The process by which an RNA molecule is synthesized complementary to the DNA in a gene. In eukaryotes, both the introns and the exons are transcribed into the RNA. The introns are later spliced out, creating an mRNA.
- transcription factor** General term for a protein that aids in initiating or regulating transcription. See also **leucine zipper** or **zinc finger**. (page 62).
- transfer RNA (tRNA)** The RNA molecule that transports a specific amino acid to a growing amino acid chain, as directed by a specific codon in the mRNA.
- translation** The process by which a protein is synthesized, according to the "blueprint" given by a messenger RNA molecule.
- Yeast Artificial Chromosome (YAC)** An artificially created cloning vector used to hold DNA sequences up to one or two million bases long. (page 416)
- You aren't expected to absorb this** Phrase used by biologists in talks when displaying a 35-mm photographic slide containing unreadable DNA or amino acid sequences.
- You aren't expected to absorb this** Phrase used by computer scientists in talks when displaying an overhead transparency containing unreadable C or C++ code.
- You aren't expected to absorb this** Phrase used by mathematicians or statisticians in talks when filling the blackboard with inscrutable equations.
- zinc finger** A common motif for a transcription factor. See **transcription factor** (page 62)