1

Biology in a Nutshell

The goal of computational genomics is the understanding and interpretation of information encoded and expressed from the entire genetic complement of biological organisms. The complete inventory of all DNA that determines the identity of an organism is called its **genome**. Biological systems are complicated, interacting multivariate networks. They can be considered at a number of different levels, ranging from populations of organisms to molecules. At the present time, computational biology emphasizes biological phenomena at levels of complexity ranging from molecular to cellular, though other levels in the hierarchy are also explored, especially in evolutionary contexts. The nature, anatomy, structure, physiology, biochemistry, and evolutionary histories of organisms define the types of problems to be solved. There are significant medical and evolutionary reasons to emphasize understanding human biology. Understanding the biology of other organisms, a worthy goal in its own right, also serves as a guide for interpreting the human genome and gene expression.

In this brief introduction we can only outline some key biological principles. For more details consult the monographs and Web sites listed at the end of the chapter.

1.1 Biological Overview

Zoos do not give a correct impression of what life on Earth is like because they over-represent mammals and other vertebrates. Organisms range from bacteria to multicellular plants and animals, and these organisms may employ a variety of strategies for extracting energy from their environment, ranging from reducing dissolved sulfate to produce H_2S and ultimately pyrite (fool's gold), to photosynthesis, to aerobic respiration. Some organisms can exist at temperatures near the boiling point (at atmospheric pressure) or below the freezing point of water. Others may be found in rocks 3 km below Earth's surface (lithotrophic bacteria) or flying over the Himalayas (snow geese). Nevertheless, analysis of ribosomal RNA sequences suggests that there are

three major related domains of organisms: **eubacteria** (organisms such as *Escherichia coli* or *Bacillus subtilis*), **Archaea** (bacteria notable for the extreme environments in which they can live), and **eukaryotes** (organisms with true nuclei, hierarchically structured chromosomes complexed with histones, and membrane-bound organelles—organisms such as humans or fungi). Relationships between these groups and between their representative members are indicated in Fig. 1.1. Two of the three major domains of life are **prokaryotic** (eubacteria and **archaebacteria**).

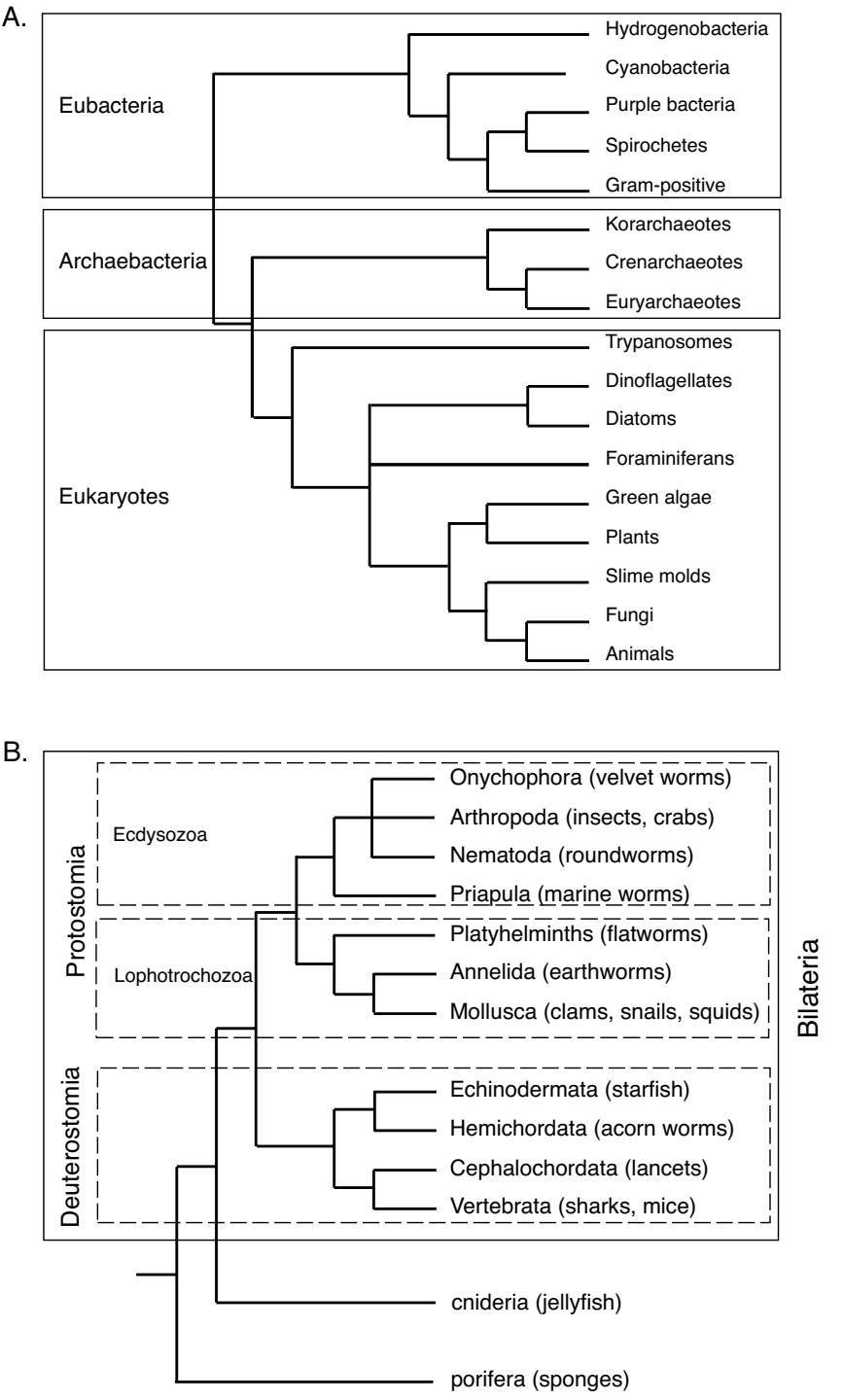
Prokaryotes do not contain a true nucleus or membrane-bound organelles, and their DNA is not as highly structured as eukaryotic chromosomes. Given the wide range of environments bacteria can inhabit and their abundance from the ancient past up to the present, bacteria as a group are considered to be the most successful life form on this planet.

Among the eukaryotes, there is an abundance of unicellular forms, called protists. Most of these are marine organisms. Ultrastructural and molecular data indicate that different types of protists may differ from each other more than plants differ from animals. Nevertheless, the unicellular eukaryotes are conventionally lumped into a kingdom called "Protista." Major multicellular groups are fungi, plants, and animals. There are about 300,000 described species of plants and about 1,000,000 described species of animals. This is a biased sample of the planet's biodiversity. Among animals, mammals represent a rather small number of species. There are about 5000 species of mammals, but there are three times as many known species of flatworms. Three-quarters of all described animal species are insects. In terms of numbers of species, insects might be considered the most successful form of land animal.

There are similarities shared by all organisms on this planet:

- The basic unit of life is the cell.
- Chemical energy is stored in ATP.
- Genetic information is encoded by DNA.
- Information is transcribed into RNA.

Fig. 1.1 ***(opposite page).*** Phylogenetic relationships among organisms (panel A) and among animals (panel B). Ancestor-descendant relationships are shown as a **tree** (see Chapter 12) with shared common ancestors corresponding to nodes to the left of descendant groups. The tree has been greatly simplified. Any given "twig" on the right can be further split to indicate descendant groups in more detail. There is usually a bifurcation at each node, but in those cases for which the branching order is unknown, there may be three (or more) descendant groups emanating from a particular node. Panel B indicates groupings of animals based upon body plans (bilateria), processes of embryological development (protostomes or deuterostomes), and physiological or anatomical features. Ecdysozoa shed their outer covering, lophotrochozoa share a type of feeding appendage or larval type, and chordata possess a notochord at some stage of development. Data from Pennisi (2003) and Knoll and Carroll (1999).


A.
Eubacteria
Hydrogenobacteria
Cyanobacteria
Purple bacteria
Spirochetes
Gram-positive
Archaebacteria
Korarchaeotes
Crenarchaeotes
Euryarchaeotes
Eukaryotes
Trypanosomes
Dinoflagellates
Diatoms
Foraminiferans
Green algae
Plants
Slime molds
Fungi
Animals
B.
Protostomia
Ecdysozoa
Onychophora (velvet worms)
Arthropoda (insects, crabs)
Nematoda (roundworms)
Priapula (marine worms)
Lophotrochozoa
Platyhelminths (flatworms)
Annelida (earthworms)
Mollusca (clams, snails, squids)
Deuterostomia
Echinodermata (starfish)
Hemichordata (acorn worms)
Cephalochordata (lancets)
Vertebrata (sharks, mice)
Bilateria
cnideria (jellyfish)
porifera (sponges)

- There is a common triplet genetic code (with a few exceptions).
- Translation into proteins involves ribosomes.
- There are shared metabolic pathways (e.g., glycolysis), with steps catalyzed by proteins.
- Similar proteins are widely distributed among diverse groups of organisms.

These shared properties reflect the evolutionary relationships among organisms, which can be useful for understanding the significance of shared biological processes. For example, there are relationships between the pathways for bacterial photosynthesis with photosynthesis seen in cyanobacteria and plants. Some characters, such as the basic biochemical pathways, are so central to life that they are found in nearly all organisms. Processes such as replication, DNA repair, and glycolysis (extraction of energy by fermentation of glucose) are present and mechanistically similar in most organisms, and broader insights into these functions can be obtained by studying simpler organisms such as yeast and bacteria. It is unnecessary to start all over again from scratch in trying to understand functions encoded in genomes of new experimental organisms.

For efficient study, biologists have typically focused on model organisms that conveniently embody and illustrate the phenomena under investigation. Model organisms are chosen for convenience, economic importance, or medical relevance. Studies on such organisms are often applicable to other organisms that might be difficult to study experimentally. For example, the generation of antibody diversity in humans is equivalent to the process that can be genetically studied in mice. It was initially surprising to discover that developmental genes (hox genes) controlling segment specification in *Drosophila* (fruit flies) were mirrored by similar genes in mammals, including humans. Model organisms include bacteria such as *E. coli* and *B. subtilis* (now joined by many other bacteria whose genomes have been sequenced), fungi such as the yeasts *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*, simple animals such as the nematode *Caenorhabditis elegans*, insects such as *Drosophila melanogaster* (fruit fly), rapidly reproducing vertebrates such as *Danio rerio* (zebrafish) and mice (*Mus musculus*), and plants such as *Arabidopsis thaliana* (mustard weed). In addition to these are agriculturally important plants and animals (e.g., corn, or *Zea mays*) and of course humans (for medical reasons).

After this brief description of the complexity and scope of biological systems and organisms, in the rest of this chapter we will turn to those levels of complexity most pertinent to computational biology. First, we discuss cells, and we follow that with an introduction to informational macromolecules. We close by indicating some of the experimental methods that define the structure and scope of computational approaches.

1.2 Cells

Except for viruses, all life on this planet is based upon cells. Cells typically range in size from 2×10^{-6} m to 20×10^{-6} m in diameter (some cells, such as neurons, can be much larger). Cells sequester biochemical reactions from the environment, maintain biochemical components at high concentrations (which facilitates appropriately rapid reaction rates), and sequester genetic information. As mentioned above, structurally there are two different types of cells: prokaryotic and eukaryotic. **Prokaryotes** have cell membranes and cytoplasm, but the DNA is not separated from the cytoplasm by a nuclear membrane. Instead, the DNA is condensed into the *nucleoid*, which is less highly structured than eukaryotic chromosomes and appears as a disorganized "blob" in thin sections viewed by electron microscopy. Prokaryotes also lack membrane-bound organelles such as mitochondria and chloroplasts. Prokaryotic cells are typically small, and they may have generation, or doubling, times as short as 20–30 minutes. **Eukaryotes** (fungi, flies, mice, and men) have a true nucleus and membrane bound organelles. Most eukaryotes have observable **mitochondria**, where major steps in aerobic respiration occur. Plant cells may contain **chloroplasts**, where plant photosynthesis occurs. They also may have prominent vacuoles and cell walls composed of cellulose. The typical doubling time of eukaryotic cells from complex organisms is significantly longer than it is for prokaryotes: for a mammalian cell in tissue culture, this is about 24 hours (although some cells, such as neurons, may not divide at all).

Cells are organized into a number of components and compartments (Fig. 1.2). The plasma membrane—the "face" that the cell shows to the outside world—is decorated with transporter proteins capable of moving particular classes of molecules into and out of the cell. Because of their more complicated structure, eukaryotic cells have a more complex spatial partitioning of different biochemical reactions than do prokaryotes. For example, translation of particular mRNA molecules (ribonucleic acid copies of DNA coding for proteins) occurs on the endoplasmic reticulum, and processing of polypeptides may occur in the Golgi apparatus. The cellular **cytoskeleton** (composed of microtubules, microfilaments, and other macromolecular assemblages) aids in the trafficking of proteins and other cellular components from point to point in the cell. **Respiration** (the production of the energetic molecule ATP by oxidation of carbon compounds in the presence of oxygen) is localized on the membranes of mitochondria. All of these features imply that particular proteins may be localized for function in some compartments of the cell, but not others.

The simplest "food" requirements are seen with bacteria. For example, *E. coli* can grow in water containing ammonium chloride, ammonium nitrate, sodium sulfate, potassium phosphate, and magnesium sulfate (NH_4Cl, NH_4NO_3, Na_2SO_4, KH_2PO_4, and $MgSO_4$, respectively) at pH 7.2 with glucose as the sole source of carbon and energy. The water usually contains other

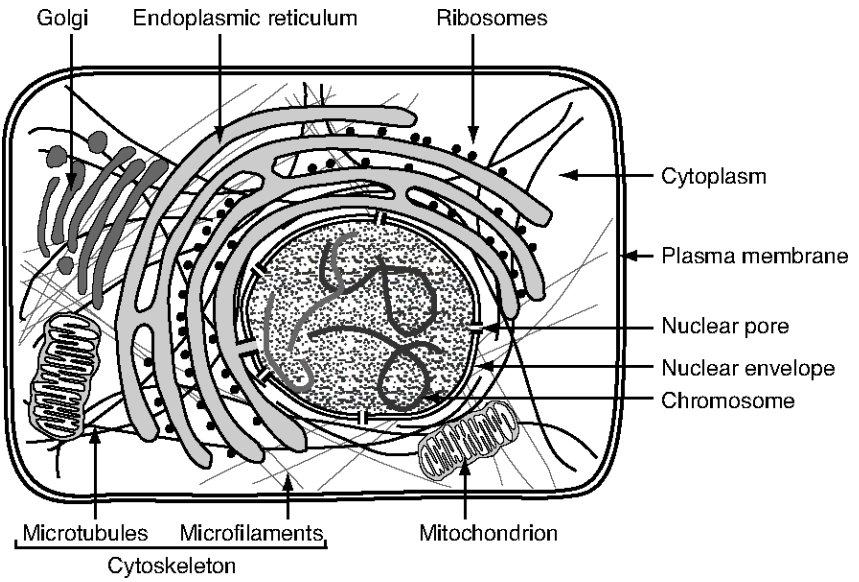


Fig. 1.2. Some major components of an animal cell (not necessarily drawn to scale). Some features (e.g., intermediate filaments, centrioles, peroxisomes) have not been depicted. In multicellular organisms, cells are frequently in contact and communication with other cells in tissues, but intercellular junctions and contacts with the extracellular matrix are not shown.

necessary metal ions in trace amounts. These substances flow into the cell through the inner and outer membranes. From a single type of sugar and inorganic precursors, a single bacterial cell can produce approximately 10^9 bacteria in approximately 20 hours; *E. coli* is also capable of importing other organic compounds into the cell from the outside, including other types of sugars and amino acids, when available.

To grow animal cells (e.g., human cells) in tissue culture, it is necessary to supply not only glucose and inorganic salts but also about a dozen amino acids and eight or more vitamins (plus other ingredients). Eukaryotic cells must import a large variety of components from the outside environment (matter flow). Because eukaryotic cells typically are 10 times larger in linear dimension than prokaryotic cells, their volumes are approximately 10^3 larger than volumes of prokaryotic cells, and diffusion may not suffice to move molecules into, out of, or through cells. Therefore, eukaryotic cells employ mechanisms of protein and vesicle transport to facilitate matter flow.

Another defining characteristic of eukaryotes is the machinery required for managing the genome during mitosis and meiosis (described below). Unlike prokaryotes, eukaryotes package their DNA in highly ordered chromosomes, which are condensed linear DNA molecules wrapped around octamers of proteins called histones. Since there are often many chromosomes, mechanisms

to ensure that each daughter cell receives a complete set are needed. This involves formation of the mitotic spindle, which includes microtubules that also contribute to the cell cytoskeleton. In addition, regulatory mechanisms are required to coordinate mitosis with DNA synthesis and the physiological state and size of the cell. These are fundamental processes that are shared by all eukaryotic cells.

This section has briefly presented a variety of information about the structure and biochemistry of cells. The DNA, RNA, and protein sequences with which computational biologists deal are important primarily because of the functions that they have within the cell. As we shall see, relating functions to macromolecules and sequences is one of the problems addressed in computational biology.

1.3 Inheritance

1.3.1 Mitosis and Meiosis

Each eukaryotic chromosome contains a single duplex DNA molecule bound with histone proteins to form a macromolecular complex. The sequences of bases contained on chromosomal DNA molecules are the result of a set of evolutionary processes that have occurred over time. These processes are intimately connected with how the chromosomes recombine and how they are copied during the DNA synthesis that precedes cell division.

Prokaryotes are typically **haploid** when they are not actively dividing, and they often (but not in every instance) have a single circular chromosomal DNA containing 10^6–10^7 bp (base pairs) of DNA. The DNA is typically inherited *vertically*, meaning that transmission is from parent to daughter cells. Under conditions of rapid growth or prior to cell division, there may be multiple copies of all or part of the prokaryotic chromosome, and except for low-frequency replication errors, their DNA sequences are usually identical. In such circumstances, recombination does not produce new assemblages of genes. Inheritance is *clonal* in the sense that descendants are more or less faithful copies of an ancestral DNA. This seemingly static mode of inheritance can be modified by transposable elements, by conjugation systems, and by acquisition of external DNA (transformation), but these interesting phenomena are beyond the scope of this introduction.

Sexual organisms such as mammals are usually **diploid**, which means that they contain N *pairs* of chromosomes (visible by light microscopy as stained chromatin). If the haploid chromosome number of an organism is N, the body (somatic) cells of that organism contain $2N$ chromosomes. There are two functional types of chromosomes: **autosomes**, which are not associated with sex determination, and sex chromosomes. Humans, for example, have 22 pairs of autosomes and two sex chromosomes: two X chromosomes for females, and one X + one Y for males. During the reproductive cycle of sexual organisms,

the **germline** tissues produce haploid sex cells, or **gametes**: ova from females and spermatozoa from males. Fusion of the gametes after mating produces a **zygote**, which will undergo development to form a new organism.

The sexual cycle involves an alternation between cells having $2N$ chromosomes or N chromosomes:

$$\begin{array}{lll} \text{Parent 1:} & 2N \rightarrow \text{Gamete 1:} & N \\ & & + \qquad \rightarrow \text{Zygote: } 2N \\ \text{Parent 2:} & 2N \rightarrow \text{Gamete 2:} & N \end{array}$$

The process of replication and reduction of chromosome numbers from $2N$ to N is called **meiosis**, which is confined to germline cells. Meiosis reduces the number of chromosomes by half because one chromosome doubling is followed by *two* cell divisions. Growth and development of the zygote is largely through a repeated process of chromosome doubling followed by one cell division—a process called **mitosis**. Cells destined to become germline cells are ordinarily subject to different sets of controls than typical body, or **somatic cells**. Mitosis of somatic cells is not genetically significant except for contributions that those cells may make to reproductive success (e.g., mitosis leading to colorful plumage in some male birds). Genetic mechanisms operate primarily during the formation and fusion of gametes.

Figure 1.3 follows two chromosomes during meiosis. Particularly important processes occur during prophase I, the beginning of the first meiotic division. As a result of the DNA synthesis that occurred during interphase, each chromosome has already been duplicated to generate a pair of *sister chromatids*. (Chromatids are precursors of chromosomes that have not yet been separated by meiosis.) Corresponding chromosomes from each parent (maternal and paternal copies of chromosome 7, for example) align with each other, and recombination occurs between corresponding maternal and paternal chromatids (i.e., between *nonsister* chromatids). **Recombination** is a process of breaking and joining chromosomes or DNA, and when this occurs at corresponding regions of a pair of similar molecules, the result is an exchange of these regions between the two molecules. This type of recombination is so-called *homologous recombination*—recombination between nearly identical sequences.

Overview of meiosis (See Fig. 1.3)

Step A: Chromatids from corresponding partner chromosomes from each parent recombine. Step B: Recombining chromosome partners (called bivalents) line up in the middle of the cell in preparation for the first meiotic cell division. Step C: Bivalents break apart, and one chromosome of each type moves to the opposite poles of the dividing cell. One or both chromatids may be recombinant. Step D: Completion of the first meiotic division produces two cells, each containing a haploid number of chromosomes, but each chromosome has two chromatids. Step E: Chromosomes line up at the center of the cell in preparation for the second meiotic division. Step F: During the second

meiotic division, bivalents in each duplicated chromosome are split, and one of each type is directed to one of the two daughter cells. The resulting cells are haploid with respect to chromosome number, and they contain only one genome equivalent.

Chromosomes are replicated only once, prior to prophase I. Thus there are four copies of each chromosome per cell at the beginning of a process that, through two cell divisions, will increase the number of cells by 2^2. Metaphase I/anaphase I leads to separation of homologous *chromosomes*, while metaphase II /anaphase II leads to separation of *sister chromatids*. Recombination (prophase I) may involve multiple crossovers with both chromatids. Note that at anaphase I and anaphase II, chromosomes originating from one parent need not migrate to the same pole: assortment is independent and random. Only one of the meiosis II products becomes the egg in vertebrate females.

1.3.2 Recombination and Variation

Recombination between nonsister chromatids has extremely important genetic consequences. The frequencies and constraints of this process determine the **genetic map**, the **haplotypes**, and blocks of **conserved synteny**. (We will define these terms in the next paragraphs.) These are properties important in genetics, population genetics, and genome analyses. Each DNA or chromosome may contain alternative forms of given genes (an alternative form of a particular gene is called an **allele** of that gene). As a result of recombination during meiosis, the allele combinations in the gamete chromosomes are usually different from the combinations found in parental chromosomes. Thus, each gamete produced by parents drawn from a population represents a novel combination of alleles that were present in the population, and the resulting variation produced in successive generations is evolutionarily "tested" against the environment. Another source of variation is the production of new alleles by mutation (change in base sequence; see below). Moreover, it is possible for normal recombination processes to "derail," leading to insertions, deletions, or duplications of longer stretches of chromosomal DNA sequence. These changes also are raw material for evolutionary change.

Chromosomes analyzed during genome projects display features that reflect recombination processes. One of the first tasks is to establish a correspondence between the DNA sequence and the genetic map. The **genetic map** records the order of genes and the approximate distances between them on their respective chromosomes. Genes are identified in classical genetics by particular mutations, sometimes called *genetic markers*. The order of genes is determined by *genetic crosses* (intentional mating of organisms having mutant genes), and the distances are measured in terms of recombination frequencies (often measured in **centimorgans**). A centimorgan corresponds to a recom-

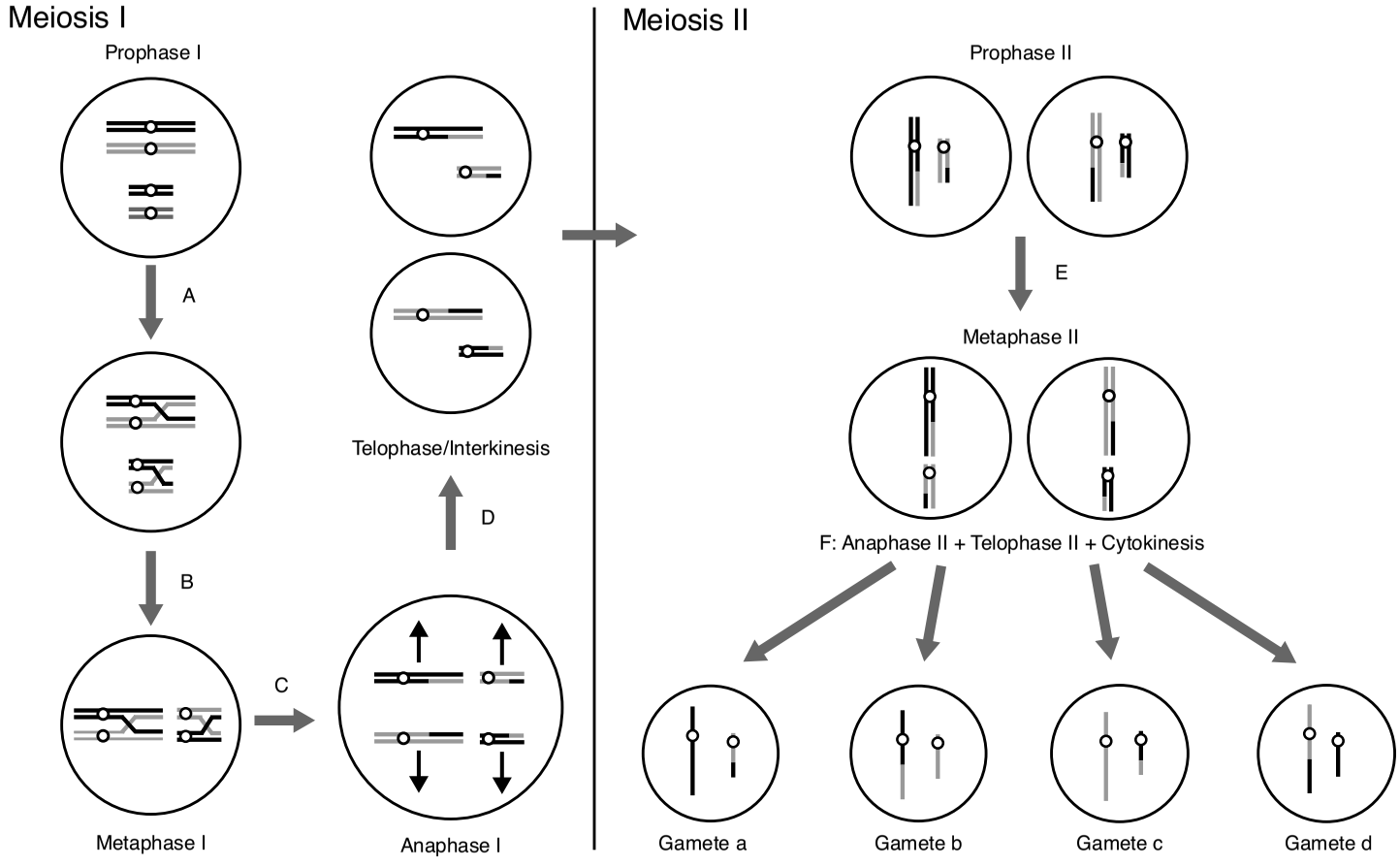


Fig. 1.3. Schematic summary of steps in meiosis (DNA replication and intermediate details not shown). In this diagram, the haploid chromosome number is 2 (one large and one small chromosome). Black chromosomes came from one parent, and grey chromosomes came from the other. For a description of processes A–F, see the accompanying box.

bination frequency of 1%, which means that two markers or genes that appear together on the same chromosome are separated from each other by recombination at a frequency of 0.01 during meiosis. Recombination is more likely to separate two distant markers than two close ones, and the recombination frequency between two markers is related to the physical distance separating them. Genes that tend to be inherited together are said to be genetically **linked**. If genetically linked alleles of several genes on a chromosome are so close together that they are rarely separated by recombination, this constellation of alleles may persist for a long period of time. Particular combinations of alleles carried on single chromosomes are called **haplotypes**, and frequencies of various haplotypes within a population characterize the structure of populations and can allow reconstruction of the evolutionary history of a population.

Over a longer timescale, recombination may shuffle the genetic maps of related species. For example, if species B and C are both descendants of ancestor A, the order of genes on the chromosomes of B and C might not be identical. Nevertheless, there may be groups of linked genes on a single chromosome in B and that also are linked on a particular chromosome of C. This circumstance is called **conserved synteny** (Fig. 1.4A; see Glossary for alternative definition). If the order of a set of genes is the same in both B and C, this set of genes is described as a **conserved segment,** and if high-density "landmarks" appear in the same order on a single chromosome in each of the two species, this set of landmarks defines a **syntenic segment**. (In some contexts, *conserved segments* and *syntenic segments* are also referred to as *conserved linkages* or *collinear gene clusters*). A set of adjacent syntenic segments is called a **syntenic block**, which may contain inversions and permutations of the syntenic segments of C compared with B (Fig. 1.4B). The numbers and sizes of such syntenic blocks are revealed when genome sequences of two organisms are compared, and these blocks are signatures of the evolutionary events separating B and C from A. It is possible to compare genomes of several related organisms and make inferences about their evolutionary relationships (i.e., comparative degrees of relatedness). One of the significant computational problems is the construction of phylogenetic trees based upon sequences or gene orders.

Even if there were no recombination, the DNA of the gametes would differ from the DNA of the parent cells because of errors that occur at low frequency during DNA replication. These errors occur at a frequency of 10^{-6}–10^{-10} per base pair for each cell division (depending upon the cell, the genome, and the DNA polymerase involved). If the errors occur within a gene, the result may be a recognizable **mutation** (alteration of the base sequence in a normal gene or its control elements). Base changes at the DNA sequence level do not always lead to recognizable phenotypes, particularly if they affect the third position of a **codon** (three successive bases of DNA that encode a particular amino acid during translation). As a result of mutations occurring over time, a position in the DNA (whether in genes or in other regions of the genome)

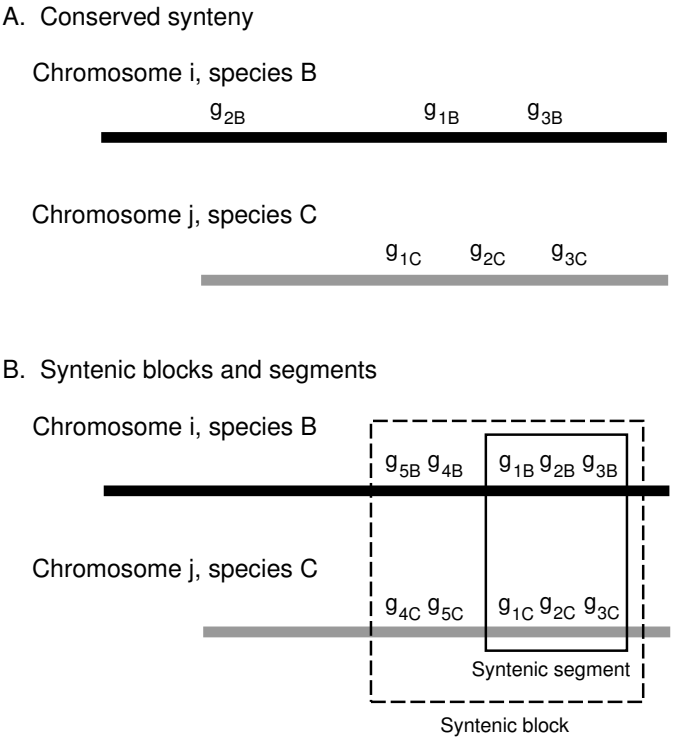


Fig. 1.4. Co-occurrence of genes or landmark sequences within single chromosomes or chromosome regions when chromosomes from each of two different organisms are compared. Panel A: Conserved synteny. In this case, $g_{\mathrm{B}1}, \ldots, g_{\mathrm{B}3}$ represent genes in species B that have homologs $g_{\mathrm{C}1}, \ldots, g_{\mathrm{C}3}$ in species C. Panel B: Syntenic segments and syntenic blocks. In this case, $g_{\mathrm{B}1}, \ldots, g_{B5}$ and the similar sequences in species C refer to landmark sequences on the genome, which can be more numerous than genes to produce a higher marker density. Syntenic segments are conceptually similar to *conserved segments*, except that in the latter case there may be microrearrangements undetected because of the low marker density.

may contain different base pairs in different representatives of a population, and this variation can be measured at particular nucleotide positions in the genomes from many members of that population. This variation, when it occurs as an isolated base-pair substitution, is called a **single-nucleotide polymorphism**, or **SNP** (pronounced "snip").

1.3.3 Biological String Manipulation

As indicated above, DNA is not immutable. During the copying or replication process, errors can occur (hopefully at low frequency, but at significantly high frequency in the case of reverse transcription of the HIV genome, for

example). In the human genome, the substitution rate at each nucleotide position averages $\sim 2.2\times10^{-9}$ per year (MGSC, 2002). The genome sequences of contemporary organisms contain a record of changes that have occurred over time. The types of changes that may have occurred include:

Deletion: Removal of one or more contiguous bases.
Insertion: Insertion of one or more contiguous bases between adjacent nucleotides in a DNA sequence. This is often associated with insertion of *transposable elements*.
Segmental duplication: Appearance of two or more copies of the same extended portion of the genome in different locations in the DNA sequence.
Inversion: Reversal of the order of genes or other DNA markers in a subsequence relative to flanking markers in a longer sequence. Within a longer sequence, inversion replaces one strand of the subsequence with its complement, *maintaining 5′ to 3′ polarity*.
Translocation: Placement of a chromosomal segment into a new sequence context elsewhere in the genome.
Recombination: In vivo joining of one DNA sequence to another. When similar sequences are involved, this is called *homologous recombination*.
Point mutation: Substitution of the base usually found at a position in the DNA by another as a result of an error in base insertion by DNA polymerase or misrepair after chemical modification of a base.

Results from some of these processes are diagrammed in Fig. 1.5. Point mutation is closely related to processes of DNA replication and repair for the generation or fixation of mutations. The other processes may also involve DNA copying, but they also involve other processes of DNA breaking and joining. Figure 1.6 makes an analogy between these processes and the menu for a computerized text editor, indicating the enzymes that may be involved in the various steps.

1.3.4 Genes

In the nineteenth century, Gregor Mendel observed that units of inheritance are discrete. A **gene** is now usually defined as a DNA segment whose information is expressed either as an RNA molecule or as a polypeptide chain (after translation of mRNA). Genes usually are found at defined positions on a chromosome, and such a position may be referred to as a **locus**. (A locus may correspond to a gene, but there are some loci that are not genes.)

Genes are identified biologically (i.e., in organisms studied in laboratories) when a mutation occurs that alters a phenotype or character. **Characters** are properties of an organism that can be observed or measured, and the phenotype corresponds to a particular state of a character. For example, a mutation in *Escherichia coli* may render the organism incapable of using lactose as a carbon source, or a mutation in *Drosophila melanogaster* may cause the eye

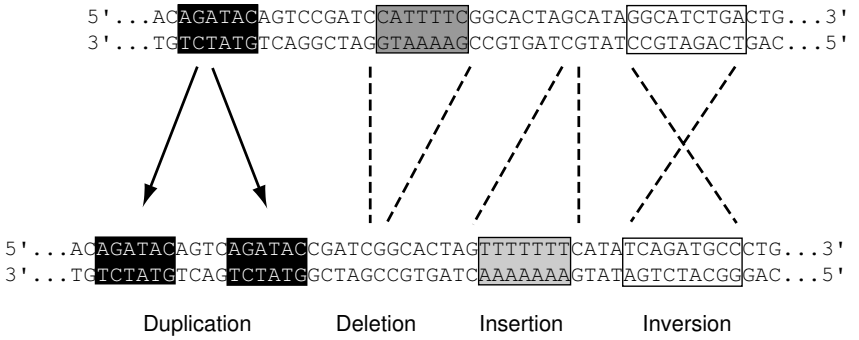


Fig. 1.5. Processes modifying multiple positions on a duplex DNA molecule. Although modifications of small numbers of basepairs are depicted, such modifications can involve much larger stretches of DNA (thousands to millions of bp or more). The processes named below the bottom molecule apply if the top molecule represents the starting condition. If the initial molecule were the one at the bottom, the DNA segment marked as "insertion" would be a "deletion" at the corresponding position in the top molecule. If it is unknown which molecule represents the initial state, such an insertion or deletion is called an "indel."

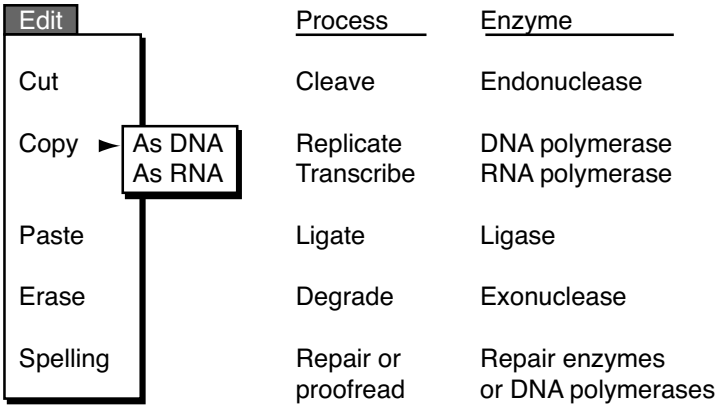


Fig. 1.6. The DNA text-editing "menu" (left) and associated enzymes.

color to change from red to white. In the latter case, the character is *eye color*, and the phenotype is *red* eye color. However, genes and phenotypes are not always in one-to-one correspondence. For example, in *E. coli* there are seven genes involved in the biosynthesis of the amino acid tryptophan from a precursor molecule. Mutations in any of these seven genes might lead to a Trp^- phenotype (which requires the addition of tryptophan for growth in the minimal medium). Similarly, a phenotype such as height or stature in *Homo sapiens* is controlled by a number of different genes, this time in a quantitative rather than all-or-none manner.

A given gene (corresponding to a particular locus) may have alternative forms called **alleles**, as described in Section 1.3.2. These alleles differ in DNA sequence, and these differences lead to differences in amino acid sequence. For example, individuals affected by sickle cell anemia have a beta-globin gene in which the glutamine normally present as the sixth amino acid residue is replaced by valine. This altered beta-globin gene is one allele of beta-globin, and the normal gene (wild-type gene) is another allele. There are other beta-globin alleles in the human population.

Genes are transcribed from DNA to form RNA molecules (including mRNA, a very important class of RNA). The DNA strand that is complementary to the mRNA is called the **template strand**, while the DNA strand whose sequence corresponds to the mRNA sequence is called the **coding strand**. DNA features (elements) found $5'$ relative to the coding sequence are considered to be "upstream," and elements that occur $3'$ relative to the coding sequence are referred to as "downstream." Diagrams of prokaryotic and eukaryotic genes are presented in Fig. 1.7.

Prokaryotic genes have a number of component elements. Going in the $5'$ to $3'$ direction relative to the direction of transcription, we encounter sites for binding proteins that control expression of the gene, a **promoter** where transcription initiates, the uninterrupted coding sequence (that eventually is translated into an amino acid sequence), and translational terminators. Sometimes the coding sequences for two or more polypeptide chains will be transcribed in succession from the same promoter, and such genes are said to be **polycistronic**. (*Cistrons* are identified by a particular type of genetic test, but they roughly correspond to coding sequences for polypeptide chains.) Polycistronic genes are relatively common in prokaryotes.

Eukaryotic genes are much more complicated than prokaryotic genes. They contain **exons**, which are segments of gene sequences that are represented in the processed mRNA. All segments of the coding sequence that will eventually be translated into protein appear in exons. **Introns** are noncoding DNA segments that separate the exons, and the RNA corresponding to introns is removed from the initial transcript by an excision process called **splicing**. Eukaryotic genes have more extensive control regions that include binding sites for transcription factors and enhancer sequences (that together regulate the level of transcription) and the "core" promoter where the transcription complex assembles. Eukaryotic transcription terminates nonspecifically down-

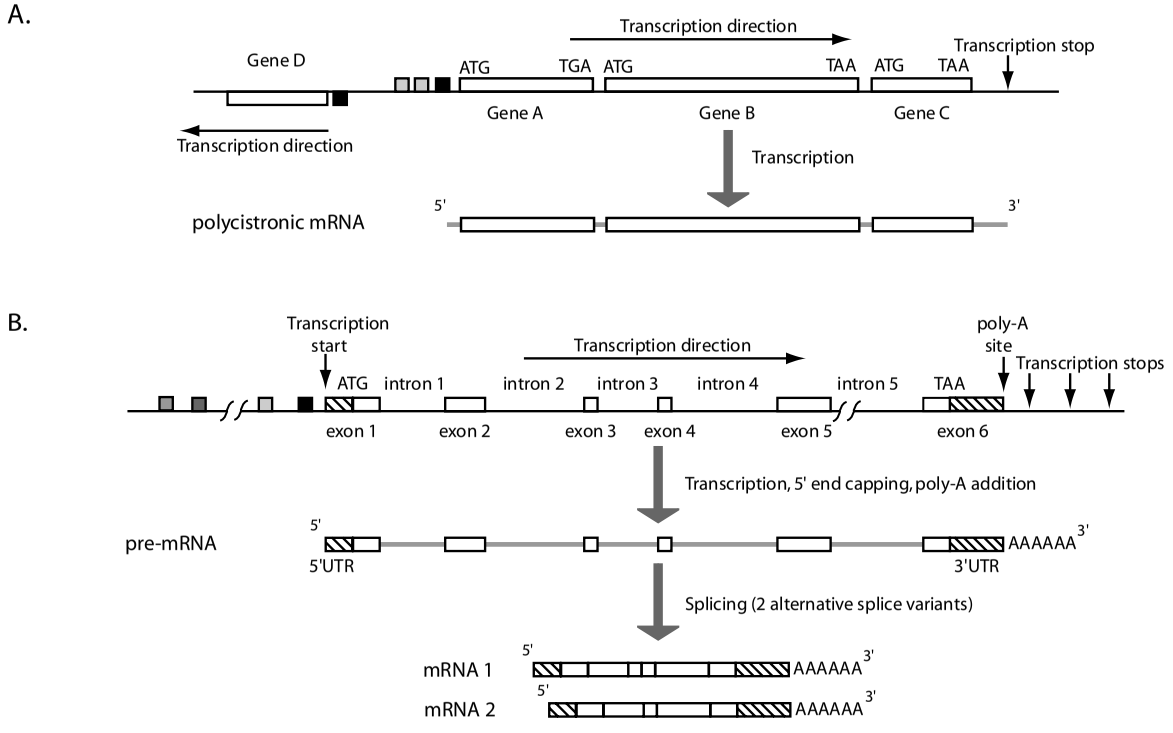


Fig. 1.7. Structures of prokaryotic and eukaryotic genes in duplex DNA (horizontal black line). White boxes above the solid black line correspond to coding sequences present on the "top" strand, while boxes below the line correspond to similar features on the "bottom" strand. Regulatory sequences (grey boxes) or core promoter sequences (black boxes) are indicated, but sizes and spacings are not to scale. DNA sequences for start and stop codons (`ATG`, `UAG`, `UAA`; `UGA` not shown) are indicated relative to the coding sequences. Panel A: Prokaryotic gene. Panel B: Eukaryotic gene. For a more complete description, see the accompanying box.

stream of the DNA "signal" that specifies the post-transcriptional addition of a string of `A` residues at the 3′ end of the message. (The addition of these `A` residues is called *polyadenylation*.)

Organization of prokaryotic and eukaryotic genes (See Fig. 1.7)

Prokaryotic genes
Genes in bacteria are normally close together and may be organized as operons (Genes A, B, and C) or may be individually transcribed (Gene D). Gene D is transcribed from the strand opposite to genes A, B, and C. Transcription (5′ to 3′ polarity with respect to the coding sequence) is initiated at promoter sequences, with initiation controlled by one or more operator sequences (grey boxes) to which repressor proteins bind. The mRNA product of the ABC operon (bottom of panel A) is ready for immediate translation.

Eukaryotic genes
The eukaryotic gene schematically depicted in panel B has transcription factor binding sites (grey boxes) upstream of the promoter. Promoter regions may be extensive, and introns may also be long, as indicated by the interruptions in the black line. The gene illustrated contains six exons and five introns. Exons 1 and 6 contain regions that will not be translated (5′ UTR in exon 1 and 3′ UTR in exon 6, hatched boxes). Transcription does not terminate at a unique position (vertical arrows). The immediate product of transcription is a pre-mRNA (grey line with open boxes) that is modified at the 5′ and 3′ ends. The poly-A tail is added to the end of the message after transcription has occurred. Splicing removes the introns to produce mature mRNA species that are ready for translation. Alternative splicing may or may not occur, but when it does, a single gene region may be responsible for two or more different (related) polypeptide chains (mRNA 1 and mRNA 2).

With the arrival of the "genome era," genes can now be identified by analyzing DNA sequence. For prokaryotes, this may be relatively easy because the average prokaryotic gene is around 1000 bp long, and approximately 90% of a typical prokaryotic genome codes for gene products. For eukaryotes, this can be a significant computational problem because eukaryotic genes usually are much larger, are interrupted by several introns (in higher eukaryotes), and occur as a much smaller fraction of their genomes (around 1.2% in the case of *H. sapiens*). For example, the average human gene is about 27,000 bp in extent and contains 9–10 exons of average length 145 bp. The rest of the gene region corresponds to extensive control regions, untranslated regions, and the intronic regions, which may be thousands of base pairs in extent.

1.3.5 Consequences of Variation: Evolution

In the last two sections, we described types of change that can occur in DNA sequences, and we alluded to the biochemical mechanisms leading to these

changes. The result of such processes occurring in a large number of interbreeding individuals is genetic variation within the **population** (i.e., a localized collection of individuals in a species that reproductively transmits genes). This means that different versions (alleles) of the same gene may be found within the population. Some members of that species may be more reproductively successful than others, depending on their genetic constitution (**genotype**). Every organism's genotype is tested against environmental conditions through the phenotypes specified by that genotype. The conditions that a population experiences allow some genotypes to be more efficiently transmitted to the succeeding generations, leading to enrichment of some alleles at the expense of others. This process is called natural selection. The change in population gene or allele frequencies over time is called evolution.

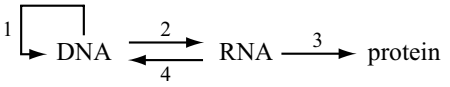
There are two related statistical and computational problems in dealing with populations: (1) characterization of genetic variation within and between populations in terms of allele frequencies or nucleotide sequence variation, and (2) analysis of the trajectory of population parameters over time. These two approaches are practically and conceptually interwoven. The first activity, which is the concern of population genetics, employs gene or locus frequency measurements taken from samples of populations alive today. Population genetics is usually (but not always) concerned with variation within species over relatively shorter periods of time. The second activity, known as molecular evolution, invokes evolutionary models to describe molecular (often sequence) data in a parsimonious manner. Molecular evolution usually (but not always) focuses on variation among species or higher-level taxa over comparatively longer periods of time. The key idea in evolutionary thought is that all species are related by descent from shared, common ancestral species that lived at some time in the past. These relationships are often represented in terms of phylogenetic trees (see Chapter 12). Thus, today's human and chimpanzee populations arose from the same ancestral primate population that existed about 6 million years ago, while humans and mice arose from an ancestral population that existed 80–85 million years ago (dates are estimated by a combination of molecular and fossil evidence). However, there were populations of organisms in the past that left no contemporary descendants (e.g., hadrosaurs, or "duck-billed" dinosaurs): they have become extinct. The biota of today are as much a result of extinction as of speciation. More than 99% of all species that have ever lived are now extinct. On average, species exist for about 2–4 million years before they succumb to extinction (some species last for much shorter, others for much longer times). Causes of extinction are varied and unpredictable. The organisms on Earth today resulted from a particular sequence of environmental conditions that occurred in a unique temporal sequence on this one planet. This pattern of evolution was not predictable and is not repeatable.

1.4 Information Storage and Transmission

An important segment of computational biology deals with the analysis of storage and readout of information necessary for the function and reproduction of cells. This information is used to code for structural proteins (e.g., cytoskeletal proteins such as actin and tubulin) and catalytic proteins (e.g., enzymes used for energy metabolism), for RNA molecules used in the translational apparatus (ribosomes, transfer RNA), and to control DNA metabolism and gene expression.

An organism's genome (defined at the beginning of the chapter) is, approximately, the entire corpus of genetic information needed to produce and operate its cells. As indicated above, eukaryotic cells may contain one, two, or three types of genomes: the nuclear genome, the mitochondrial genome, and the chloroplast genome (in plants). The vast majority of eukaryotes contain both nuclear and mitochondrial genomes, the latter being much smaller than the nuclear genome and confined to mitochondria. When one speaks of "the X genome" (e.g., "the human genome"), the nuclear genome is usually the one meant. Mitochondrial and chloroplast genomes in some ways resemble genomes of the prokaryotic symbionts from which they were derived.

The information flow in cells (already alluded to above) is summarized below.



The processes are identified as follows:

1. DNA replication, where a DNA sequence is copied to yield a molecule nearly identical to the starting molecule;
2. Transcription, where a portion of DNA sequence is converted to the corresponding RNA sequence;
3. Translation, where the polypeptide sequence corresponding to the mRNA sequence is synthesized;
4. Reverse transcription, where the RNA sequence is used as a template for the synthesis of DNA, as in retrovirus replication, pseudogene formation, and certain types of transposition.

Most biological information is encoded as a sequence of residues in linear, biological macromolecules. This is usually represented as a sequence of Roman letters drawn from a particular alphabet. Except for some types of viruses, DNA is used to store genomic information. RNA may be used as a temporary copy (mRNA) of information corresponding to genes or may play a role in the translational apparatus (tRNA, spliceosomal RNA, and rRNA). Proteins are polypeptides that may have catalytic, structural, or regulatory roles.

1.4.1 DNA

The genomes of free-living (nonviral) organisms are composed of DNA. The subunits (nucleotides) of these macromolecules are deoxyribonucleotides of four types: deoxyadenosine 5′-phosphate (`A`), deoxycytidine 5′-phosphate (`C`), deoxyguanosine 5′-phosphate (`G`), and thymidine 5′-phosphate (`T`). The 5′ position on the sugar of each nucleotide is connected via a phosphate group to the 3′ position on the sugar of the immediately preceding nucleotide. Each DNA strand has a 5′ end, corresponding to the phosphate group attached to the 5′ position on the sugar molecule of the first nucleotide, and a 3′ end, corresponding to the -OH group at the 3′ position on the sugar of the last nucleotide. For double-stranded DNA (Fig. 1.8), the two strands are antiparallel, which means that the two polynucleotide chains have opposite orientations or polarities. Base-pairing rules are usually observed: `A` base pairs with `T` and `G` base pairs with `C` (Fig. 1.9). Two strands whose sequences allow them to base pair are said to be complementary. A duplex DNA molecule can thus be represented by a string of letters drawn from {`A`, `C`, `G`, `T`}, with the left-to-right orientation of the string corresponding to the **5′ to 3′** polarity. The other strand is implied by the base-pairing rules. If the string corresponds to a single strand, then this should be explicitly stated. If a strand is written in the 3′ to 5′ direction, then this should be explicitly indicated. DNA molecules are encountered with numbers of bases or base pairs ranging from ∼20 (oligonucleotide primers) to hundreds of millions (panel A of Table 1.1). For example, the DNA molecule in human chromosome 1 has a size of 285,000,000 base pairs. The number of bases or base pairs may be colloquially referred to as "length," and units may be given in kilobases (kb = 1000 bases or base pairs) or megabases (Mb = 1,000,000 bases or base pairs).

The organization of DNA in cells can be considered at four different structural levels: constituent nucleotides, DNA, chromatin, and chromosomes. As an indicator of scale, the "length" of a nucleotide is approximately 1×10^{-9} m. The diameter of the DNA helix is 2×10^{-9} m, and the pitch of the helix is 3.4×10^{-9} m. In eukaryotes, the DNA is wrapped around histones to form nucleosomes (diameter 11×10^{-9} m). The chain of nucleosomes is further wrapped into higher-order structures that constitute the chromosomes, which are located in the nucleus. A typical nucleus might have a diameter of 0.5×10^{-5} m and would represent approximately 10% of the cell volume. Notice that a DNA molecule may be orders of magnitude longer than the diameter of the cell that contains it. For example, the length along the contour of the DNA in human chromosome 1 is approximately 9.5 cm (!), while the cell diameter is approximately 10^{-3} cm. The small diameter of the DNA helix and the hierarchical packing of the nucleosome structures allow packing of these long molecules into nuclei.

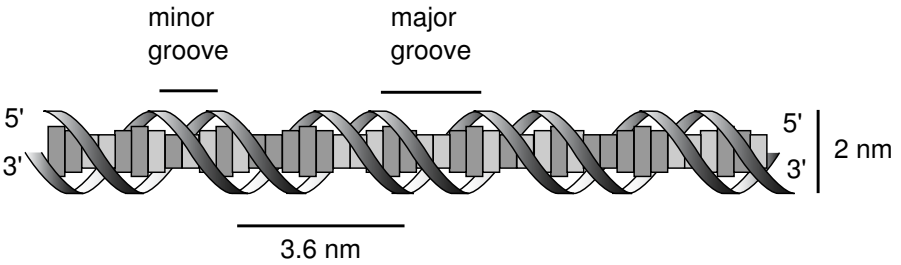


Fig. 1.8. Structure of duplex DNA. Ribbons represent the phosphodiester backbones of the two antiparallel strands, and the rectangular elements in the middle of the duplex represent the stacked base pairs. Connections of these base pairs to the phosphodiester backbone are not indicated. The gradients in size of the rectangles indicate that sometimes the base pairs are being viewed "edge-on" and other times "end-on" as they lie at different degrees of rotation about the helix axis. The major and minor grooves and relevant dimensions are indicated. Major and minor grooves are distinguished from each other by the spacing between the two phosphodiester backbones and the depth from the outside of the molecule to the edges of the base pairs.

1.4.2 RNA

RNA differs from DNA in two primary ways: the residues contain hydroxyl groups at the $2'$ position of the sugar (and thus are not "deoxy"), and uracil (`U`) replaces the thymine base `T`. Thus RNA molecules are composed of the monomers adenosine $5'$-phosphate, cytidine $5'$-phosphate, guanosine $5'$-phosphate, and uridine $5'$-phosphate. RNA is written as a string of Roman letters drawn from the alphabet {`A`, `C`, `G`, `U`}, with the left-to-right orientation corresponding to the $5'$ to $3'$ polarity. In most cases, RNA is encountered as a single strand, but often it will form intrastrand base pairs to form *secondary structures* that may be functionally important. RNA secondary structures are the result of intrastrand base pairing to form sets of "hairpins" and loops. The prediction of secondary structures for RNA molecules is a significant computational problem that takes into account free energies of base-pair formation and constraints on loop sizes. Duplex RNA molecules can be functional, either as genomes of some viruses or as interfering RNA (RNAi) that helps regulate gene expression in plants and animals. Sizes of RNA molecules typically range from approximately 100 to a few thousand nucleotides (not bp)—see panel B of Table 1.1.

1.4.3 Proteins

As indicated above, proteins are directly involved in the functioning of the cell. They are polypeptides, composed of strings of amino acid residues polymerized with the loss of one H_2O molecule per joined pair of amino acid residues. They

Table 1.1. Examples of DNA, RNA, and protein molecules. DNA molecules differ primarily by base composition and length and are structurally similar (but not identical) to each other. RNA molecules differ by length, base composition, and secondary and tertiary structure. Proteins are much more structurally diverse: the data represent only a sample of the diversity of structure types for proteins. Identical protein types from different organisms may differ in sequence; see the accession numbers for the source organism.

A: DNA

Name (GenBank acc. num.)	Number of bp	Base composition (%G + C)
Mouse mtDNA (NC_001569)	16,295	36.7
Bacteriophage λ (J02459)	48,502	49.9
E. coli K-12 chromosome (U00096)	4,639,221	50.8
Human chromosome 1	285,000,000	41.0

B: RNA

Name (GenBank acc. num.)	Number of nt	Base composition (%G + C)
$tRNA_{Ala}$(M26928)	73	60.3
18S rRNA(X04025)	1826	53.8
HIV-1 (AF443114)	9094	41.9

C: Protein

Name (PDB acc. num.)	Polypeptides (number/molecule)	Number of residues	Molecular weight
Ribonuclease A (1FS3)	A (1)	124	13,674
Total:	1	124	13,674
Hemoglobin (2HCO)	A (2)	141	15,110
	B (2)	146	15,851
Total:	4	574	61,922
Ubiquinol oxidase (1FFT)	A (1)	663	74,359
	B (1)	315	34,897
	C (1)	204	22,607
	D (1)	109	?
Total:	4	1291	148,000 (est)
Glutamine synthase (1FPY)	A (12)	468	51,669
Total:	12	5616	620,028

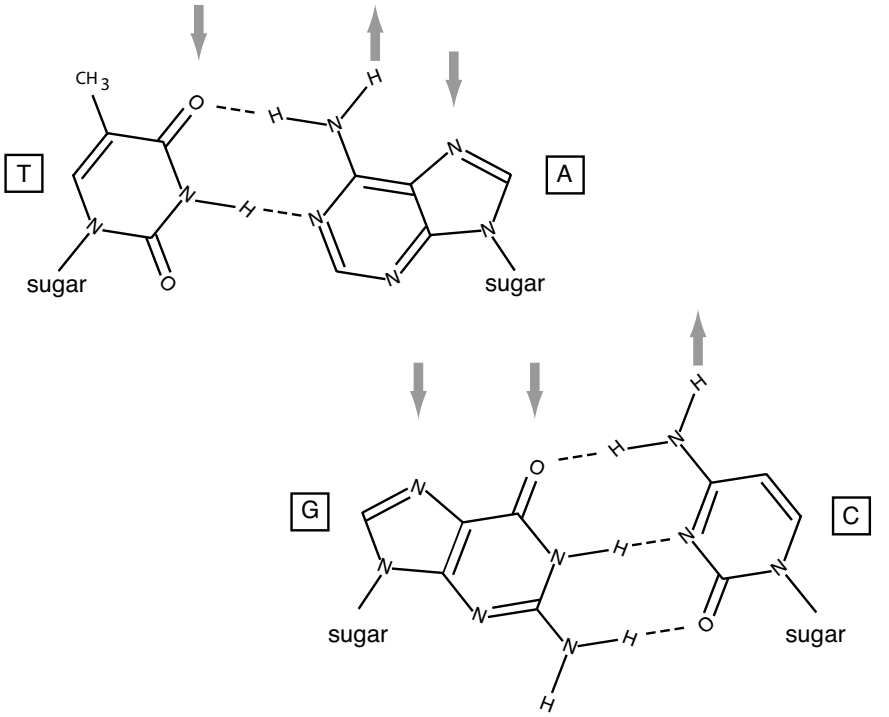


Fig. 1.9. Structure of Watson-Crick base pairs. Only relevant atoms are indicated. Junctions between straight segments in the rings correspond to locations of carbon atoms. The -CH_3 of `T` is called a methyl group. Broken lines connecting `A` and `T` or `G` and `C` correspond to hydrogen bonds that help stabilize the base pairs. Grey arrows pointing toward atoms on the rings indicate that these atoms are hydrogen bond acceptors, and grey arrows pointing away correspond to hydrogen bond donors. Only donors and acceptors on the major groove edges of the base pairs are shown. The bonds extending to the sugar residues in the phosphodiester backbone are on the minor groove side of the base pairs.

are usually represented as a string of letters drawn from an alphabet of twenty, written in the direction from the amino-terminal to the carboxy-terminal ends:

NH_2-AEGLV$\cdots$WKKLAG-COOH

This also may be written as

NH_2-Ala-Glu-Gly-Leu-Val-$\cdots$-Trp-Lys-Lys-Leu-Ala-Gly-COOH

using the three-letter abbreviation for the amino acid residues. Each polypeptide chain usually corresponds to a gene. Polypeptides in proteins usually have between 50 and 1000 amino acid residues, with 300 to 400 residues being the typical average length of polypeptides in many organisms. For small proteins (often in the range 100 to 200 amino acid residues), the active molecule may

be composed of a single polypeptide chain, but many proteins are composed of a precisely defined set of polypeptide chains. The simplicity of representation of polypeptides as a string of letters belies the profound structural complexity of protein molecules: the prediction of protein structure from the amino acid sequence is a difficult computational and theoretical problem.

Panel C of Table 1.1 lists some examples of proteins, illustrating the ranges in size and numbers of polypeptides. The sequence of amino acid residues constitutes the *primary structure*. There are a limited number of types of *secondary structures* (involving interactions between nearby amino acid residues on the same polypeptide chain), including alpha helix and beta pleated sheet. Secondary structure elements may combine to form a particular fold of a polypeptide segment. There are thought to be 1000–2000 different types of folds. Each individual polypeptide chain is folded into a particular three-dimensional structure (called *tertiary structure*). It is a general observation that complex proteins are often composed of multiple polypeptide subunits. Sometimes these are all identical, as is the case with glutamine synthase (12 identical polypeptide chains), but in other cases these subunits are all different (e.g., ubiquinol oxidase; see Table 1.1C). The aggregate structures formed from multiple polypeptide chains are called *quaternary structures*.

1.4.4 Coding

The DNA alphabet contains four letters but must specify polypeptide chains with an alphabet of 20 letters. This means that combinations of nucleotides are needed to code for each amino acid. Dinucleotides are combinations of two: `AA`, `AC`, `AG`,..., `TC`, `TG`, `TT`. There are 4^2, or 16, possible dinucleotides—not enough to code for all 20 naturally occurring amino acids. Trinucleotides (triplets) are combinations of three nucleotides: `AAA`, `AAC`, `AAG`, ..., `TTC`, `TTG`, `TTT`. There are 4^3, or 64, possible trinucleotides. The genetic code is a triplet code, and the code triplets in mRNA are called **codons**. These may be written in their DNA form with `T` instead of `U` (when looking for genes in DNA) or in their RNA form with `U` instead of `T` (when we are concerned about the actual translation from mRNA). Triplets that specify "stop translation" are `UAG`, `UGA`, and `UAA`. Translational starts usually occur at an `AUG` codon, which also specifies the amino acid methionine. One representation of the genetic code is given in Appendix C.2.

Since there are three stop codons out of 64 triplets, there are 61 triplets coding for the 20 amino acids. This means that amino acids may be specified by more than one triplet. For example, leucine (`Leu`) is encoded by `CUG`, `CUA`, `CUC`, and `CUU`. As we will see later, these codons are not used with equal frequencies in various genes and organisms, and the statistics of codon usage is a characteristic that can sometimes be used to distinguish between organisms.

Successive amino acid residues in a polypeptide chain are specified by the sequence of triplets. For example, if the first amino acid is specified by a triplet beginning at nucleotide i in the mRNA, the second one will be specified by

the triplet beginning at nucleotide $i + 3$, and the third one will be specified by the triplet beginning at nucleotide $i + 6$, and so on. But what determines the location of the initial triplet at i? Prokaryotes contain a hexanucleotide at the $5'$ end of the mRNA that sets the stage for translation beginning with the next `AUG`. Eukaryotes have a $5'$ cap structure, and translation begins at the next `AUG` of the mRNA.

When examining DNA for protein-coding regions, we initially look for **open reading frames** (ORFs). An **ORF** (pronounced "orf") is a stretch of sequence that does not contain stop codons. In a random DNA sequence that is 50% `G+C`, on average one would expect a stop codon to occur after a stretch of 21 codons. The median exon size for humans is about twice as large as this (122 bp), and for prokaryotes the average gene size is approximately 1000 bp, so longer-than-average ORFs are indications of possible protein-coding regions. As a first approximation to gene recognition in prokaryotes, one looks for (1) an `AUG` start codon followed by (2) an open reading frame long enough to code for a reasonably sized polypeptide (> 50 amino acid residues) and having (3) the characteristic codon usage frequencies. The ORF ends at one of the three stop codons. As we will see later, gene recognition in eukaryotes requires much more analysis to replace point (2) above. Duplex DNA contains six possible reading frames, as illustrated in Fig. 1.10: three on the "top" strand and three on the "bottom" strand. When searching for genes in DNA, we must examine all six reading frames.

1.5 Experimental Methods

Data used in computational biology often can be traced back to a few microliters of an aqueous solution containing one or more types of biological macromolecules. The methods for studying biological materials determine the types of data available and the computational approaches required. The structures of DNA, RNA, and proteins were presented in Section 1.4. Here we discuss these types of molecules from an experimental perspective. As computational biologists, we should clearly understand what quantities are measured and how the measurements are made. It is appropriate to examine the raw data (e.g., output from a sequencing machine, autoradiogram, pattern of restriction fragments on a gel) to help understand the type and quality of the data.

1.5.1 Working with DNA and RNA

Most DNA and RNA molecules, regardless of their source, have similar properties and can be purified by using only minor variations from a standard set of protocols: alkaline minipreps (appropriate for DNA, not RNA), ultracentrifugation in CsCl gradients, or ethanol precipitations, for example. The experimental approaches depend on the amounts and molecular sizes of each type of macromolecule in a typical cell.

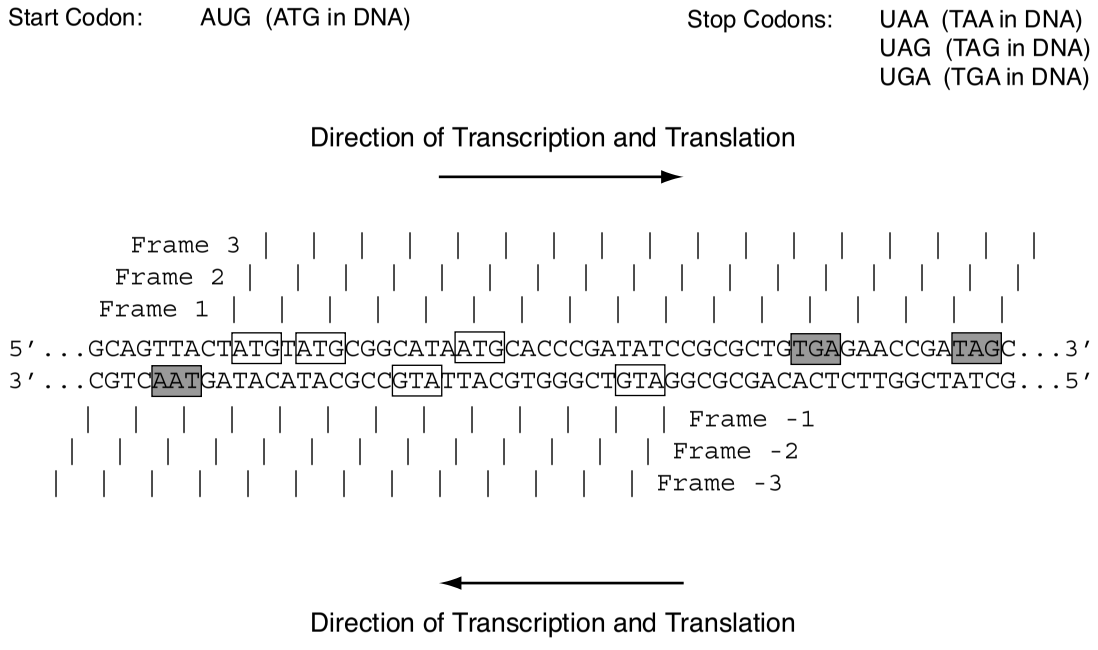


Fig. 1.10. Relationships among open reading frames, coding or "sense" strands, and template strands in duplex DNA. For frames 1, 2, and 3 above the DNA sequence, the "top" strand is the coding strand and the bottom strand is the template. For the frames written below the sequence, the situation is reversed. For actual open reading frames in prokaryotes, the number of codons is, on average, approximately 300. Processed eukaryotic mRNA molecules are defined largely by their exons, which average approximately 50 codons in humans.

We can calculate the abundance of a 1000 bp segment of single-copy DNA in diploid eukaryotic cells. Taking each base pair of the sodium salt of DNA to have a molecular mass of 662, and recognizing that there are two copies of each molecule per cell, we calculate that 10^3 bp of single-copy DNA in a genome corresponds to about 2×10^{-18} grams of DNA/cell. A 1 liter culture of mammalian cells grown in suspension (at about 10^6 cells/mL in a tissue culture flask, higher in a bioreactor) would contain 2×10^{-9} g of this 1000 bp region. In contrast, a DNA segment of similar length in mitochondrial DNA (mtDNA, present at 10^3-10^4 copies/mammalian somatic cell) will be at least a thousand times more abundant. The molecular sizes also matter. A 1000 bp DNA segment of eukaryotic DNA from a particular chromosome is part of a long, linear DNA molecule that cannot be easily purified without fragmentation, which destroys long-distance relationships to other regions of the molecule. In contrast, a similar length segment in mitochondrial DNA can be easily purified on an intact molecule because the mtDNA molecules are small, circular molecules that can be purified without fragmentation.

For routine molecular biology procedures (e.g., restriction mapping, in vitro mutagenesis), a laboratory technician requires about 10^{-7} to 10^{-8} g of DNA. Because only small quantities of any particular nuclear DNA sequence are isolated directly (see above), DNA is often amplified. The most common amplification methods are the polymerase chain reaction (**PCR**) and cloning. PCR employs in vitro enzymatic DNA synthesis in repeated cycles, such that if amount A of a DNA is originally present, after n cycles the amount present will be approximately $A2^n$. With extreme care to exclude contaminating DNA, it is technically possible to amplify as little as one molecule (!) by PCR. There is no DNA repair system to correct polymerase copying errors during PCR; consequently, after repeated cycles of amplification, some of the copies will have slightly altered sequences compared with the original, unamplified molecules.

Cloning (see Fig. 1.11) involves enzymatic joining of the desired DNA sequence to a cloning vector and its propagation in a convenient host organism (bacteria, yeast, or eukaryotic cells in culture). The **cloning vector** is a DNA molecule chosen to have appropriate replication properties (copy number, control of copy number), insert size, and host specificity (vectors must be matched to the host cell) for producing the desired amount and length of DNA product. If a genome is being analyzed, the genome may be represented as collections of clones, called **libraries**. In genome sequencing projects, clone collections based upon three different types of cloning vector are not uncommon. The number of clones to be collected and stored may range from 10^4 to 10^7, depending upon the genome size of the organism being analyzed, the chosen vector, and the cloned fragment size. This of course raises practical issues of physically archiving the clones (e.g., sample storage in freezers) and recording pertinent descriptive information about each (e.g., cloning vector, date prepared, shelf location in a particular freezer).

Microarray studies of gene expression focus on types and amounts of different RNA species in cells or tissues (see Chapter 11). Mammalian genomes may have ∼25,000 genes, which corresponds to 25,000–75,000 or more possible different mRNA species when alternative splice variants are taken into account. Different mRNA species can have very different levels of abundance, and they are often unstable (some may be degraded in vivo within seconds to minutes). mRNA molecules are extremely unstable in vitro as well. Unlike DNA, they are sensitive to alkaline hydrolysis. Stable ribonucleases present in cells (and on laboratory apparatus and fingerprints of technicians who don't wear protective gloves) represent an even bigger problem. Sequences represented on mRNA molecules can be enzymatically converted to DNA form by in vitro reverse transcription using retroviral reverse transcriptases. Since the DNA strands that are reverse-transcribed are complementary to the mRNA molecules, these products are referred to as **cDNA** (Fig. 1.11, right). Such molecules may not contain copies of the complete mRNA sequence, however. The molecules are usually converted to duplex cDNA for eventual manipulation (e.g., cloning). Collections of cloned cDNAs representing transcripts from a particular cell type are called cDNA libraries. Of course, cDNA molecules can be further amplified by PCR as required.

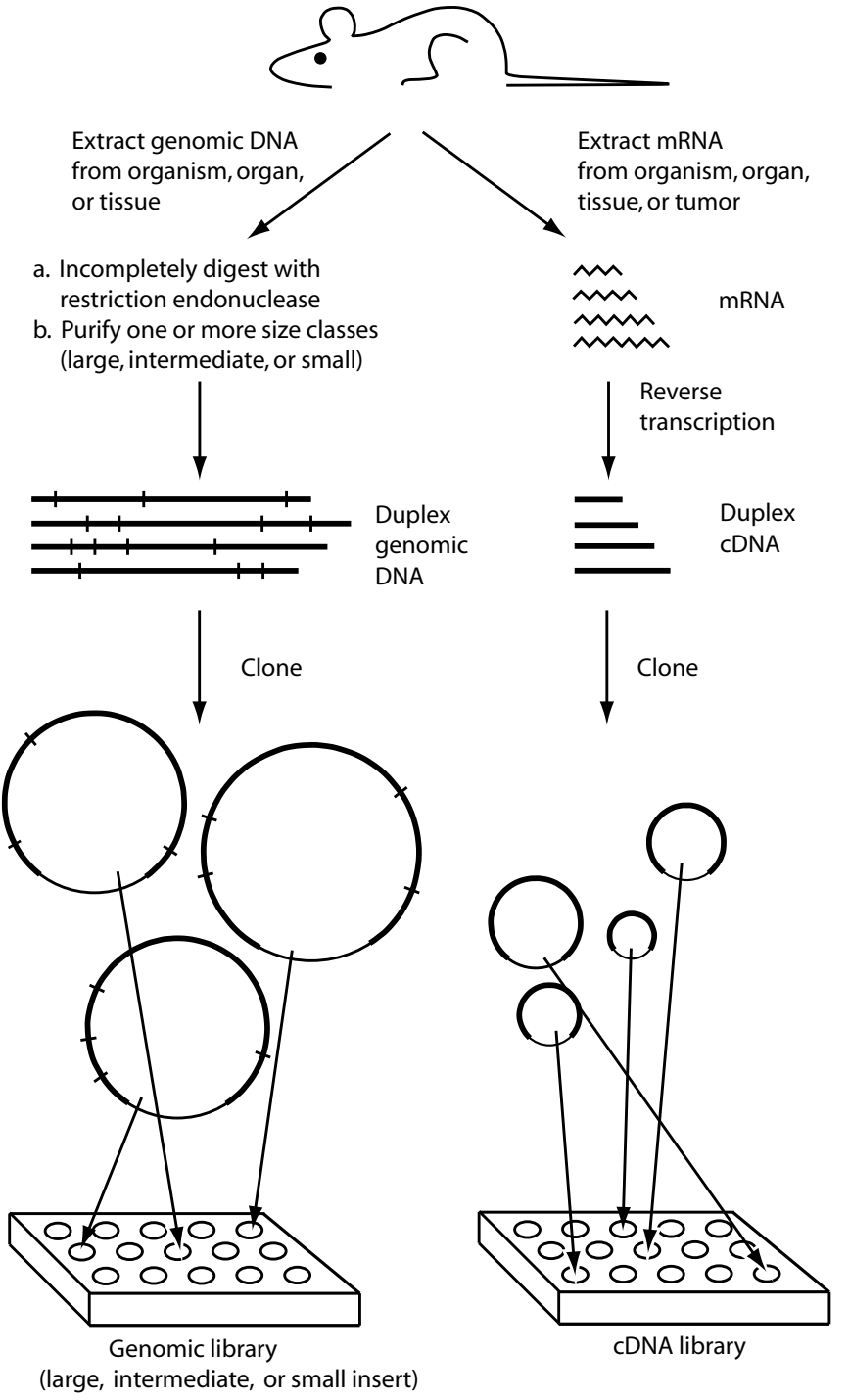
Sometimes short DNA sequences (∼200 nucleotides) are obtained from large collections of cDNA clones to provide sequence labels for genes expressed under a particular set of conditions for a particular cell or tissue type. These short sequences are called expressed sequence tags (**ESTs**). The sequences of ESTs can be used to design PCR primers that can assist in mapping these sequences to genomic locations.

1.5.2 Working with Proteins

Different proteins can have substantially different abundances. For example, proteins such as histones or cytoskeletal proteins are abundant in cells and are readily purified. Other proteins have very low abundances (< 100 molecules per cell). Unlike DNA and RNA (represented as cDNA), currently there is no method for amplifying rare proteins in a cell extract. To obtain rare proteins

---

Fig. 1.11 ***(opposite page).*** Capturing genetic information in genomic or cDNA libraries. After extraction or conversion to DNA, DNA fragments are cloned and introduced into an appropriate host organism for propagation. Either DNA molecules or clones may be archived for future use. cDNA clones are usually small because mRNA from which cDNA is derived is usually hundreds of nucleotides to a few thousand nucleotides in length. Genomic clones may have small inserts (∼2 kb, useful for DNA sequencing), intermediate-sized inserts (10 kb, useful for DNA sequence assembly), or large inserts (100–300 kb, useful for long-range sequence assembly). Molecules in the appropriate size range are produced by a size selection step prior to cloning. Different cloning vectors are required for different insert sizes.


Extract genomic DNA from organism, organ, or tissue
Extract mRNA from organism, organ, tissue, or tumor
a. Incompletely digest with restriction endonuclease
b. Purify one or more size classes (large, intermediate, or small)
mRNA
Reverse transcription
Duplex genomic DNA
Duplex cDNA
Clone
Clone
Genomic library (large, intermediate, or small insert)
cDNA library

directly from cells, it may be necessary to employ many liters of cell culture and multiple purification steps to obtain even 10^{-3} g of purified protein. Large amounts of a particular protein can be obtained by cloning its coding sequence (possibly derived from a full-length cDNA) onto an **expression vector**—a vector with a promoter (constitutive or inducible) and sequences specifying the addition of poly`A` to transcripts. Large amounts of the desired protein are produced from such clones because of the increased gene copy number and increased transcription rate. However, a polypeptide chain produced in this manner may or may not be properly processed, folded, glycosylated, or assembled if it is expressed in a nonnative cell type.

If a particular protein has never before been purified, protein biochemists know a number of procedures to *try* (e.g., precipitation by $(NH_4)_2SO_4$, ion-exchange chromatography, high-pressure liquid chromatography, molecular sieve chromatography, etc.), but in general multiple steps are required, each of which must be optimized before milligram amounts of purified and active protein can be produced. During purification, proteins that function as parts of macromolecular complexes may be separated from other proteins in the complex and thus may lose biological activity. In addition, proteins **denature** (lose their natural structure) more readily than do nucleic acids. This can occur because of binding to surfaces (a problem particularly in dilute solutions), oxidation of sulfhydryl groups, unfolding at elevated temperatures, or exposure to detergents. (Nucleic acids are indifferent to detergents.) Similarly, some membrane proteins may be difficult to purify in soluble form because of their hydrophobic character.

A lack of amplification methods (short of expression cloning) and the wide range of protein properties determined by the various possible amino acid sequences influence methods for studying proteins. Two-dimensional polyacrylamide gel electrophoresis (**2DE**) is an established method for displaying (in principle) any cellular proteins in cell lysates or cell fractions (Fig. 1.12). Sample preparation disrupts the protein structures, producing a solution of denatured polypeptide chains. Therefore, functional proteins composed of multiple polypeptide chains (e.g., RNA polymerases) are "deconstructed" to their constituent polypeptides. Separation of the polypeptides on the polyacrylamide gel matrix depends upon two different properties associated with each polypeptide chain: the molecular mass and the isoelectric point (pI). Molecular mass is the summation of the molecular masses of the constituent amino acid residues, and since polypeptides generally differ in length and composition, each polypeptide species has a characteristic molecular mass. The isoelectric point of a protein or polypeptide is the pH at which its average charge is zero. The isoelectric point is related to the amino acid composition by the number and kind of acidic and basic residues that the polypeptide chain or protein contains. An excess of basic residues leads to a pI greater than 7.0, and an excess of acidic residues leads to a pI less than 7.0.

With 2DE (Fig. 1.12), the protein mixture is first resolved into a series of bands by **isoelectric focusing**, which is electrophoresis in a stationary

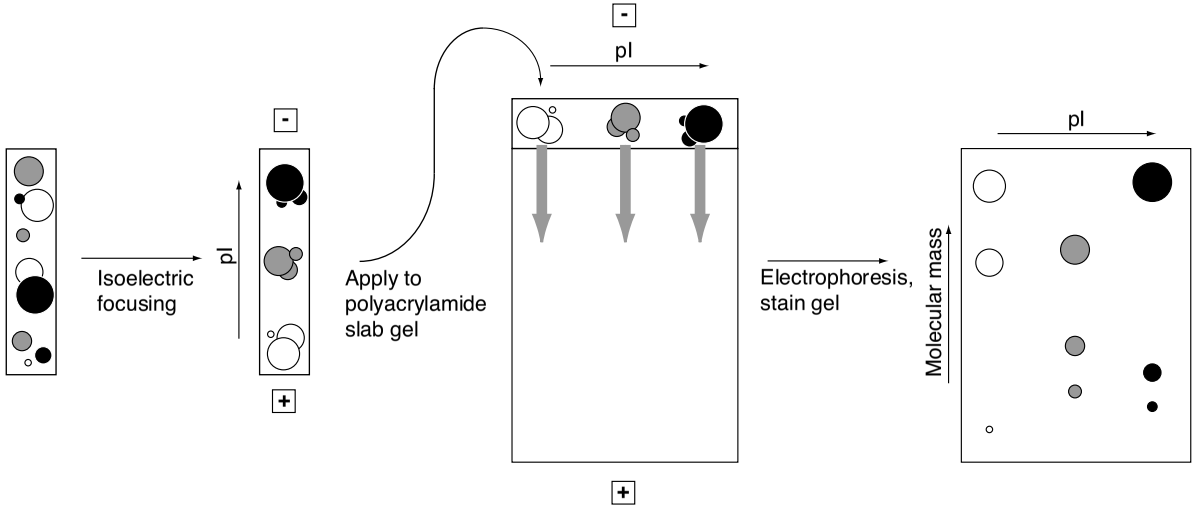


Fig. 1.12. Schematic illustration of two-dimensional gel electrophoresis. Circles represent proteins in the mixture, with different sizes representing different molecular masses. The shadings correspond to the isoelectric points of the proteins. Isoelectric focusing (left) separates the molecules into three different classes, each of whose members share the same pI. SDS-polyacrylamide gel electrophoresis (right) is conducted in an orthogonal direction, leading to the separation of proteins in each pI class into different protein spots.

pH gradient (i.e., a pH gradient for which the pH at any position is time-invariant). Proteins having an isoelectric point of 8, for example, migrate to that point in the pH gradient where the pH is 8 and then stop migrating because their average charge is zero at that point. Proteins whose isoelectric point is 4.5 migrate to the position in the pH gradient where the pH is 4.5. Isoelectric focusing is performed in a "tube gel" (typical tubes are 0.3 cm in diameter and 15 cm long) or on plastic-backed gel strips containing immobilized pH gradients. After the bands have been formed by isoelectric focusing, the gel or strip is equilibrated with a solution containing the strong detergent SDS. SDS is negatively charged. All polypeptides bind SDS at approximately the same mass of SDS/mass of protein, and they become extended and negatively charged with approximately the same charge-to-mass ratio in solution. The gel or strip is then placed on a slab gel, perpendicular to the direction of the electric field to be applied. Electrophoresis through the slab polyacrylamide gel resolves polypeptides based upon their extension in space, which is related to the molecular mass. After the electrophoresis step, spots corresponding to individual polypeptides can be visualized by staining or by autoradiography or phosphorimaging (if proteins were labeled with radioisotopes). Figure 1.13 shows an example of a stained gel. With typical protein loads, up to 1000 to 1500 polypeptides can be resolved in two dimensions.

Often, we wish to detect a specific macromolecule in the presence of others. For DNA and RNA, this is relatively easy because Watson-Crick base pairing is a specific and efficient mechanism for a probe DNA molecule to "recognize" a molecule containing the complementary sequence. However, there currently are no easy methods for detecting specific protein sequences, except for methods using antibodies and antibody-like molecules. These and similar methods are powerful and sensitive but are experimentally demanding, as described in the box below.

Antibodies and specific protein recognition

Antibodies (Ab) or **immunoglobulins** are proteins that are produced by vertebrate immune systems to bind "foreign" molecules that may be present in the body (e.g., bacteria). A complex combinatorial process produces antibodies that are capable of binding virtually any specific **antigen** (a molecule that elicits the immune response) that the organism might encounter. Usually (but not always), an antibody that recognizes and binds to antigen `x` will not recognize and bind to antigen `y`, and vice versa.

There are two labor-intensive steps in the production of antibodies: production of the antigen and production of the antibody. We have already discussed earlier the issues related to purification of proteins to be used as antigens. The second issue, antibody production, can be attacked in different ways. Traditionally, antibodies are made by repeated injection of an antigen into rabbits or goats and bleeding the animals several weeks later. This produces a sample of **polyclonal antibodies** (a *mixture* of different immunoglobu-

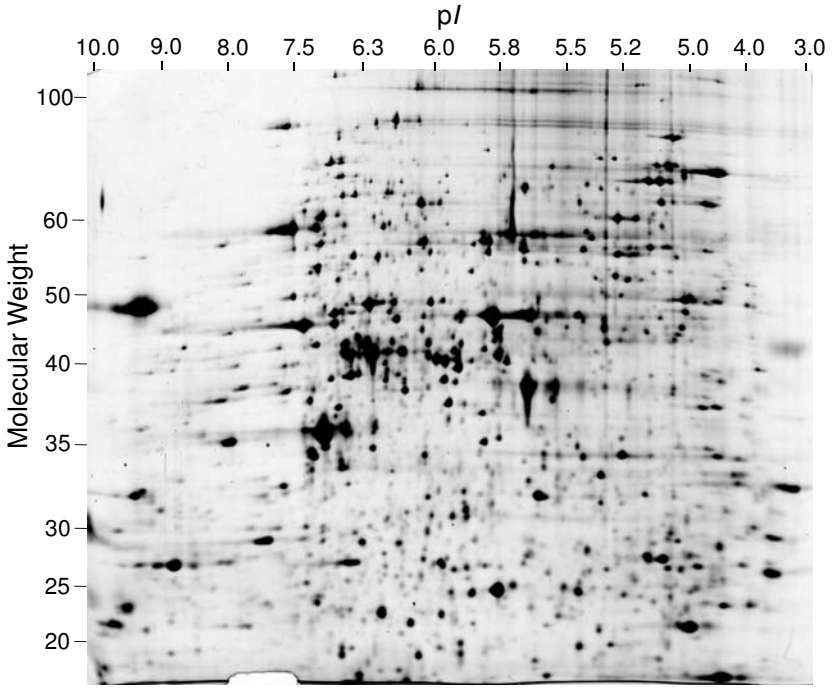


Fig. 1.13. Proteins extracted from yeast resolved by two-dimensional gel electrophoresis and visualized by silver staining. Molecular weights are in kilodaltons. Over 1000 spots were detected, corresponding to about 20% of all yeast genes. The abundance of proteins in the gel is approximately proportional to the spot intensity. Reprinted, with permission, from Gygi SP et al. (1999) *Molecular and Cell Biology* 19:1720–1730. Copyright 1999, the American Society for Microbiology. All rights reserved.

lin species directed against the antigen). The different immunoglobulins may "recognize" or bind to different specific regions of the antigen (different protein domains, for example). Different regions of the antigen recognized by distinct antibody species are called **epitopes**. A preferable but more expensive and time-consuming procedure is to produce a **monoclonal antibody**, which is a single immunoglobulin species that recognizes a single epitope. In this procedure, a mouse is immunized with the antigen, and cells from the spleen (where antibody-producing cells mature) are removed and fused with an "immortal" multiple myeloma (tumor) cell line. This produces hybrid cell lines (hybridomas), and individual hybridoma cell lines may produce a single antibody directed toward a single epitope from the original antigen. Such cells can be used to produce the antibody in tissue culture, or they can be injected into mice to produce the desired antibody in vivo. This procedure obviously

entails costs for vivarium facilities and for a hybridoma tissue culture facility. The procedure usually requires a few months.

An alternative to monoclonal antibodies are probe molecules identified by **phage display** approaches. With phage display, DNA encoding the antibody binding regions is cloned into one of the coat protein genes of a filamentous *E. coli* bacteriophage such as fd or M13. When the mature phage is produced, it "displays" the antigen binding region on its surface. From a mixture of bacteriophages that recognize a large number of antigens, a phage species capable of recognizing a particular antigen can be purified after repeated amplification and enrichment steps. This approach only requires standard cloning and biochemical expertise, but of course it still requires purified antigen. Production of antibodies can also be avoided by using small polypeptides specifically designed by protein engineering for specific protein binding, and such technology is commercially available (e.g., Affibody AB, Sweden).

1.5.3 Types of Experiments

In general, questions addressed by computational biology are subsets of the overall question, "How do cells and organisms function?" We may think of the larger question as being composed of three components:

- Characterizing the genome;
- Identifying patterns of gene expression and protein abundance under different conditions;
- Discovering mechanisms for controlling gene expression and the biochemical reactions in the cell.

We expand on these topics in the introductions to subsequent chapters, but here we provide an overview. Computational biologists should be concerned about experimental details because computational approaches must be tailored to the structure of the experimental data.

Genomes can be characterized by genetic and physical mapping, analyzing evolution of single-copy and repeated DNA sequences, identifying genes and their organization, and building inventories of genes by type. This is the realm of restriction mapping, cloning, DNA sequencing, pattern recognition, and phylogenetic tree building. Each of these topics is addressed in a subsequent chapter. Typically, DNA is isolated from an appropriate representative strain or lineage of a particular type of organism and cloned as shown in Fig. 1.11 (left). Clones stored in libraries may be digested with **restriction endonucleases** (individually or in combination) to produce maps of the various clones. Alternatively, the ends of the cloned inserts may be sequenced directly. Depending upon the purpose of the experiment, the clones may be screened by hybridization to find those having sequences similar to particular probe sequences (e.g., to DNA from a gene of interest). Ultimately the result

is the sequence of a DNA molecule annotated for any genes or regulatory signals (e.g., transcription factor binding sites) that it may contain. Comparison with similar gene regions in model organisms may provide insight into gene function. Investigators interested in those genes associated with a particular genetic disease may focus on a few genes, but with genome sequencing approaches, the entire panoply of genes is examined, usually in an evolutionary context.

Gene expression studies seek to measure the amounts of mRNA or protein species in a particular cell or tissue type under a particular set of physiological conditions. The **transcriptome** is the entire collection of transcripts, and the **proteome** is the entire collection of proteins for a particular cell and set of conditions. The transcriptome is studied by a variety of methods for measuring (directly or indirectly) mRNA levels, including **spotted microarray** experiments, "gene chip" experiments, serial analysis of gene expression (**SAGE**), and total gene expression analysis (**TOGA**). For eukaryotes, this may involve purification of RNA and preparing cDNA (Fig. 1.11, right). Each cDNA clone corresponds to a particular expressed sequence (mRNA). For spotted microarrays, gene or intergenic DNA samples spotted onto solid substrates are hybridized to labeled cDNA mixtures prepared from mRNA extracts (see Chapter 11). Proteomes can be analyzed by resolving protein extracts from cells by using 2DE and subjecting particular polypeptides to tandem mass spectrometry. Array technologies also are being devised to identify and quantify protein species in cell extracts.

Gene regulation may depend upon sites on DNA that bind regulatory proteins and also can depend upon protein-protein interactions. Sites on DNA that bind regulatory proteins can be identified on a DNA fragment by **gel-shift** or **footprinting** methods. Gel-shift experiments are lower-resolution electrophoretic assays that depend upon the retardation ("gel-shift") of DNA-protein complexes relative to DNA having no bound protein. Alternatively, "footprinting" methods may be used to locate the position and extent of the binding region. These methods rely on reagents that cleave DNA within one or the other of the two strands. Proteins bound to the DNA protect it from cleavage. Fragmented DNA strands are resolved by gel electrophoresis, and the "footprint" is the region of the gel lacking cleaved fragments. Protein-DNA complexes formed in vitro can also be investigated by immunoprecipitation of complexes using antibodies specific for the bound protein.

Gene regulation can also depend upon protein-protein interactions, which can be studied in vivo by using yeast "two-hybrid" genetic systems. Protein-protein interactions can also be studied in vitro by chemical cross-linking. To detect proteins that interact with protein P, the extract containing it and possible binding partners is subjected to chemical cross-linking. Then reagents that specifically bind to P (a specific antibody, for example) are used to purify complexes containing P, and reversal of cross-linking releases proteins interacting with it. Such data are helpful for identifying the components and connectivity of protein interaction networks.

The particular combination of experiments employed will depend upon the reason for the study. The study may be part of an investigation of a particular genetic disease conducted in a medical school, or it may have been initiated within a biotechnology company to produce a profitable therapeutic agent. The study might be a comparison of a particular set of genes among a variety of organisms or might encompass the entire genome of a single organism. A wide range of methods derived from genetics, chemistry, biochemistry, and physics may be applied to each individual problem or project. Computational biologists should be aware of concepts associated with a wide range of disciplines.

References

Alberts B, Lewis J, Raff M, Johnson A, Roberts K (2002) *Molecular Biology of the Cell* (4th edition). London: Taylor & Francis, Inc.

Branden C, Tooze J (1998) *Introduction to Protein Structure* (2nd edition). London: Taylor & Francis, Inc.

Calladine CR, Drew HR (1997) *Understanding DNA: The Molecule and How It Works* (2nd edition). San Diego, CA: Academic Press.

Griffiths AJ, Lewontin RC, Gelbart WM, Miller JH, Gelbart W (2002) *Modern Genetic Analysis* (2nd edition). New York, NY: W. H. Freeman Company.

Knoll AH, Carroll SB (1999) Early animal evolution: Emerging views from comparative biology and geology. *Science* 284:2129–2137.

Mouse Genome Sequencing Consortium [MGSC] (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* 420:520–562.

Pennisi E (2003) Modernizing the tree of life. *Science* 300:1692–1697.

`http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Books`
Online books on genetics, molecular biology, and cell biology presented by the National Center for Biotechnology Information (NCBI). Includes some of the best standard textbooks.

`http://www.ucmp.berkeley.edu/alllife/threedomains.html`
An excellent resource covering evolution and the diversity of life from both biological and paleontological perspectives. Maintained by the University of California, Berkeley Museum of Paleontology.

`http://web.mit.edu/esgbio/www/7001main.html`
Biology hypertextbook. Information is highly compressed but rapidly accessible.

`http://www.genome.gov/10002096`
Glossary of terms presented by the National Human Genome Research Institute. Includes illustrations for glossary entries.