From molecules to a living cell

One of the striking features of life on earth is the universality (as far as we know) of the chemistry of the basic building blocks of cells; this is especially true in the case of the carrier of genetic information, the DNA. This universality suggests that it is in the intrinsic physicochemical properties of these biomolecules where one can find the origins of spatiotemporal organization and functions characteristic of living systems. At the level of molecular interactions, fundamental laws of physics and chemistry apply. However, the emergence of the 'living state' is expected to be associated with ensembles of molecular processes organized spatially in organelles and other cellular compartments, as well as temporally in their dynamics *far from equilibrium*. To help understand these levels of organization, the basic anatomy of cells, the properties of these biomolecules and their interactions are summarized in this chapter. Of central importance is the molecular machinery for expressing genes to proteins; this is a complex but well-orchestrated machinery involving webs of gene-interaction networks, signalling and metabolic pathways. Information on these networks is increasingly and conveniently made available in public internet databases. A brief survey is given at the end of this chapter of the major databases containing genomic, proteomic, metabolomic, and interactomic information. The challenge to scientists for decades to come is to integrate and analyze these data to understand the fundamental processes of life.

Cell compartments and organelles

A diagram of the basic architecture of *eukaryotic* cells is shown in Fig. 2.1. Every eukaryotic cell has a membrane-bound *nucleus* containing its chromosomes. In contrast, a *prokaryotic* cell lacks a nucleus; instead, the chromosome assembly is referred to as a *nucleoid*. A description of the compartments and major organelles in a representative eukaryotic cell is given in this section.

A bilayer phospholipid membrane, called the *plasma membrane*, delineates the cell from its environment. This membrane allows the selective entry of raw materials for the synthesis of larger biomolecules, the transmission of extracellular signals (e.g. from extracellular ligands docking on membrane-receptor proteins), retains or concentrates substances needed by the cell, and the efflux of waste products. Each phospholipid molecule has a hydrophobic (or 'water-hating') end and a hydrophilic (or 'water-loving') end. When these molecules are dispersed in water, they aggregate spontaneously to form a bilayer membrane, both surfaces of the membrane being lined

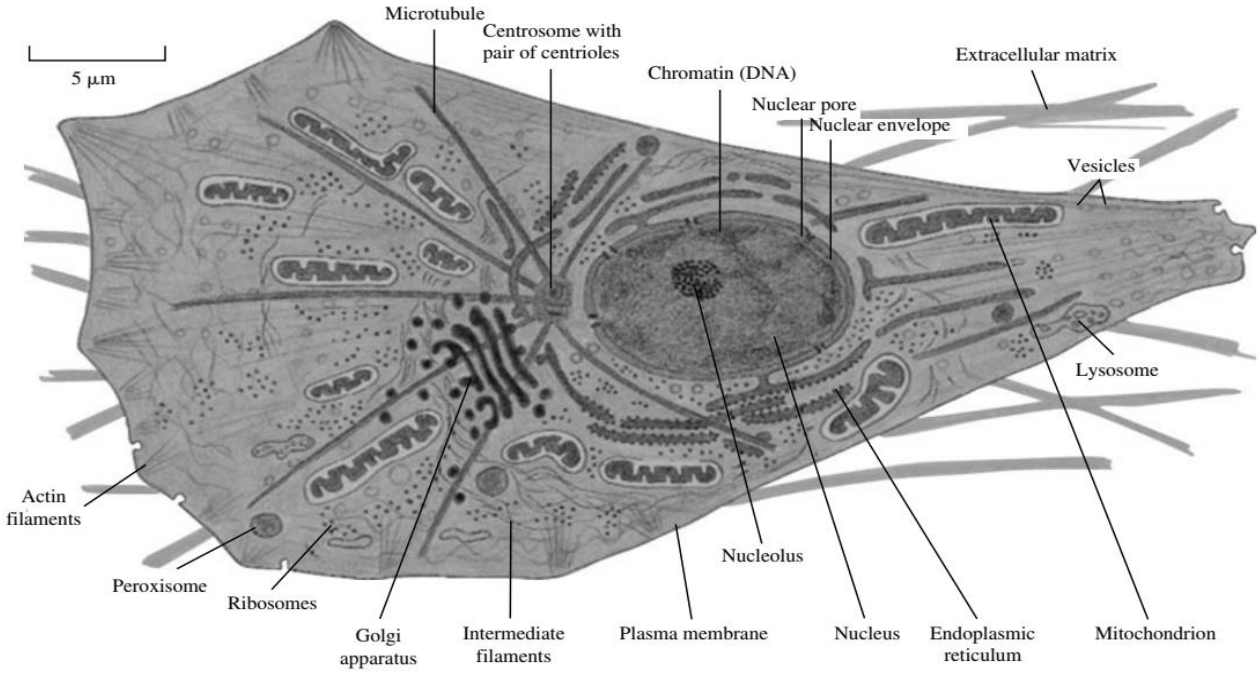

Fig. 2.1 The major compartments and organelles of a typical eukaryotic cell. The plasma membrane, chromosomes (condensed chromatin), ribosomes, nucleolus, mitochondria, centrosome and the cytoskeleton (microtubules and filaments) are described in the text. The Golgi apparatus is referred to as the 'post office' of the cell: it 'packages' and 'labels' the different macromolecules synthesized in the cell, and then sends these out to different places in the cell. Lysosomes are organelles containing digestive enzymes, which is why they are also called 'suicide sacs' because spillage of their contents causes cell death.

by the hydrophilic ends of the lipid molecules, while the hydrophobic ends are tucked in between the surfaces. This is an example of a common observation that many types of biomolecules synthesized by cells possess the ability to self-assemble into structures with specific cellular functions (other examples will be given below).

Proteins that span the plasma membrane, called *transmembrane proteins*, are involved in cell–environment and cell–cell communications. Examples of these proteins are *ion-channel* proteins (e.g. sodium and potassium ion channels involved in regulating the electric potential difference across the plasma membrane) and *membrane-receptor* proteins, whose conformational changes (brought about, for example, by binding with extracellular ligands) usually initiate cascades of biochemical processes that get transduced to the nuclear DNA causing changes in gene expression. Certain membrane proteins are involved in cell–cell recognition that is crucial in the operation of the immune system.

The material between the plasma membrane and the nucleus is called the *cytoplasm*. Encased by the nuclear membrane are the chromosomes that contain the

genome (set of genes) of the organism. Humans (*Homo sapiens*) have 46 chromosomes in their somatic cells. Human sperm and egg cells have 23 chromosomes each.

Although the code for producing proteins is in the chromosomes, proteins are synthesized outside the nucleus in sites that look like granules under the microscope. These sites of protein synthesis are the *ribosomes* (see Fig. 2.1 and Fig. 2.2). As shown in Fig. 2.1, ribosomes are either attached to a network of membranes (called the *endoplasmic reticulum*) or are free in the cytoplasm. A bacterium such as *E. coli* cell has $\sim 10^4$ ribosomes and a human cell has $\sim 10^8$ ribosomes. The assembly of ribosomes originates from a nuclear compartment called the *nucleolus* (see Fig. 2.1).

Besides proteins, many other types of molecules are produced in the cell through enzyme-catalyzed *metabolic* reactions. The organelles called *mitochondria* (Fig. 2.1) are the cell's power plants because most of the energetic molecules – called ATP (*a*denosine *tri*phosphate) – are generated in these organelles. Energy is released when a phosphate bond is broken during the transformation of ATP into ADP (*a*denosine *di*phosphate); this energy is used to drive many metabolic reactions. A typical eukaryotic cell contains ~2000 mitochondria. (Interestingly, mitochondria contain DNA, which suggests – according to the endosymbiotic theory – that these organelles were once free-living prokaryotes.)

As depicted in Fig. 2.1, the shape of the cell is maintained by the *cytoskeleton* that is a network of *microtubules* and *filaments*. These cytoskeletal elements are self-assembled from smaller protein subunits. Rapid disassembly and assembly of these subunits can occur in response to external signals (this happens, for example, when a cell migrates). Of major importance to cell division is the organelle called *centrosome* that is composed of a pair of barrel-shaped microtubules called *centrioles* (Fig. 2.1). Immediately after the chromosomes are duplicated, the centrosome is also duplicated; the two centrosomes are eventually found in opposite poles prior to cell division. The spindle fibers (microtubules) emanating from these two centrosomes carry out the delicate task of segregating the chromosomes equally between daughter cells.

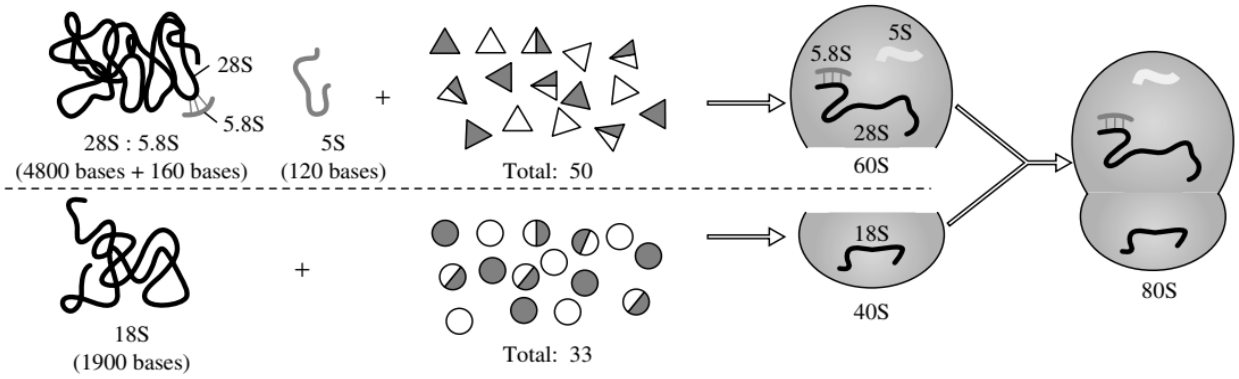

Fig. 2.2 Ribosomes of mammalian cells. Shown are schematic pictures of the components of the large (60S) and small (40S) subunits of the ribosome (80S). The strands represent ribosomal RNAs, and the triangles are the 50 proteins of the large subunit and the 33 proteins of the small subunit.

The components and structures of cell organelles and other large protein complexes have been elucidated. For example, mammalian ribosomes are large complexes of 83 proteins and 4 ribonucleic acids (see Fig. 2.2). Other important examples are the components and the mechanisms of action of various polymerase enzymes in the replication of chromosomes (DNA polymerases) and in decoding genes (RNA polymerases). Many of these macromolecular complexes are being viewed as molecular machines.

To reiterate, a wide variety of the biomolecules synthesized in cells self-assemble spontaneously. The phospholipid molecules of the plasma membrane – products of cell metabolism – form bilayers spontaneously in aqueous solutions. In the construction of the cytoskeleton, tubulin proteins polymerize to form microtubules, actin to microfilaments, and myosin to thick filaments. Recent studies even suggest that the whole eukaryotic nucleus is a self-assembling organelle.

The molecular machinery of gene expression

All known living things on earth use DNA (*d*eoxyribo*n*ucleic *a*cid) as the genetic material (except for some viruses that use *r*ibo*n*ucleic *a*cid or RNA for short). The publication of the structure of DNA by James Watson and Francis Crick in 1953 revolutionized biology. The structure of DNA provides a clear molecular basis for the inheritance of genes from one generation to the next, as described in more detail below.

In each eukaryotic chromosome, DNA exists as two strands paired to form a double helix (Fig. 2.3). Each strand has a sugar–phosphate backbone, and attached to the sugars are four nitrogenous bases, namely, adenine (A), thymine (T), cytosine (C), and guanine (G). The double helix is formed from the Watson–Crick pairing between these bases: A paired to T, and C paired to G. As shown schematically in Fig. 2.3, the specificity of these pairings is due to the number of hydrogen bonds between the bases. Because these hydrogen bonds are weak – unlike the much stronger covalent bonds in molecules – they allow the 'unzipping' of the double helix during DNA replication. Note that the T–A pair has two hydrogen bonds while the G–C pair has three, suggesting that the double helix is easier to unzip where there are more T–A pairs than G–C pairs. It is these Watson–Crick base pairings that elegantly explain the molecular basis of gene inheritance.

For DNA replication to start, the duplex has to 'unzip' to expose single-stranded DNA segments where synthesis of new DNA strands occur according to the Watson–Crick base pairing. This is a highly regulated affair involving dozens of enzymes, including DNA polymerases.

Genes correspond to stretches of sequences of the letters A, T, C, G on the DNA (DNA segments comprising a gene are not necessarily contiguous). *Gene expression* refers to the synthesis of the protein according to the DNA sequence of the gene (also called protein-coding sequence). The gene-expression machinery requires that the DNA sequence is first transcribed to an RNA sequence. RNA molecules also have the A, C, and G bases, but uracil (U) is used instead of T. RNA molecules do not stably form double helices like DNA. However, the pairings of C–G and A–U are observed. The gene-expression machinery is summarized in Fig. 2.4.

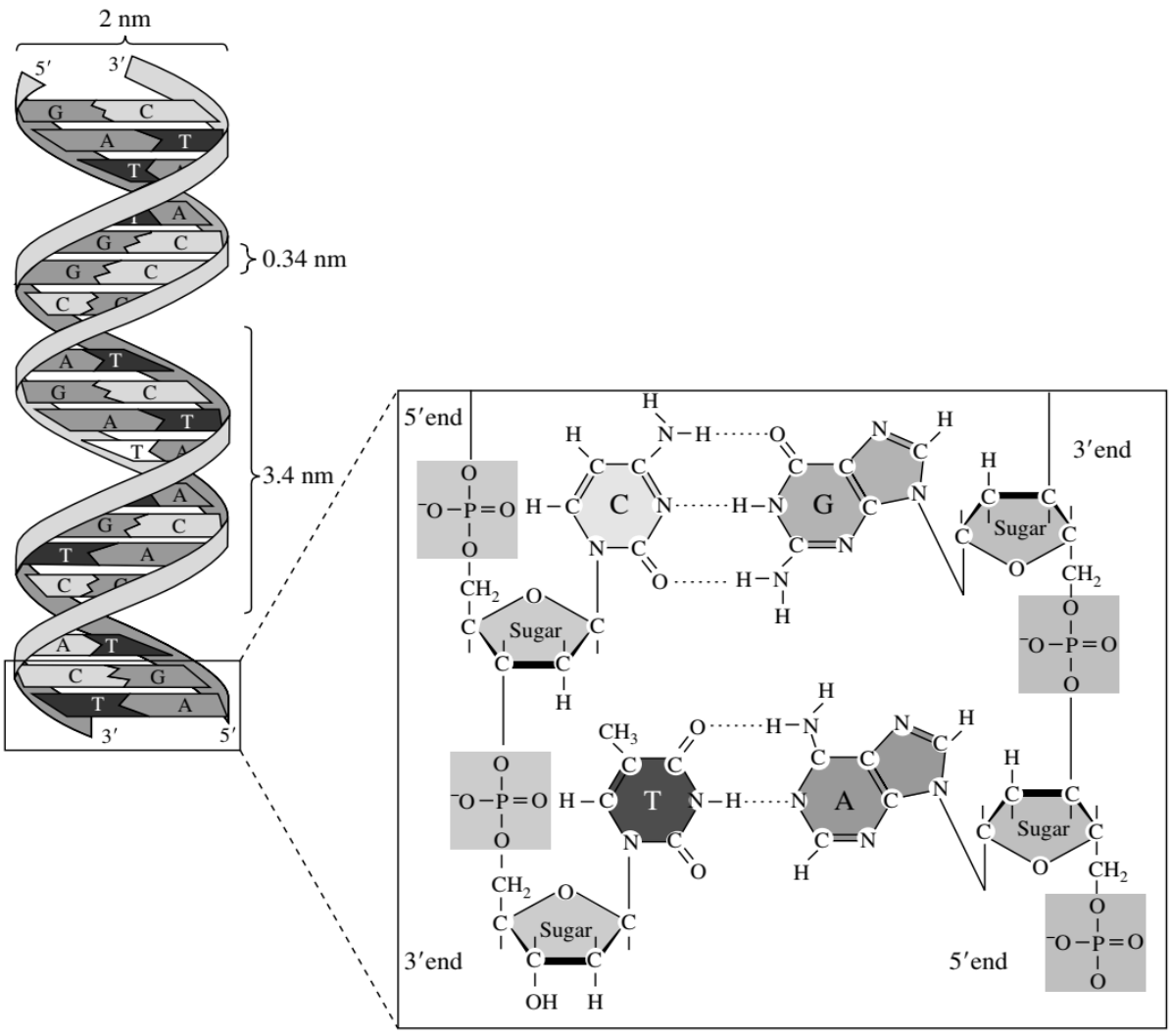

Fig. 2.3 Two DNA molecules form the Watson–Crick double helix where the sugar–phosphate backbones are on the outside and the bases are inside, paired by hydrogen bonds as shown on the right of the figure (A with T, and C with G). The 5′ and 3′ designations of the ends of a DNA strand are based on the numbering of the C atoms on the deoxyribose (sugar).

As depicted in Fig. 2.4, the DNA double helix is unzipped where particular genes are located so that the enzyme called RNA polymerase can transcribe the DNA sequence into RNA. This primary RNA contains sequences called *exons* and *introns*; the latter do not code for proteins and are removed. The remaining exons are then stitched together through a process called *RNA splicing* to form a continuous molecule of mature *messenger RNA* (mRNA). This mRNA relocates from the nucleus to the cytoplasm where it is *translated* in ribosomes. Thus, gene expression is defined as the combination of transcription and translation to the protein product.

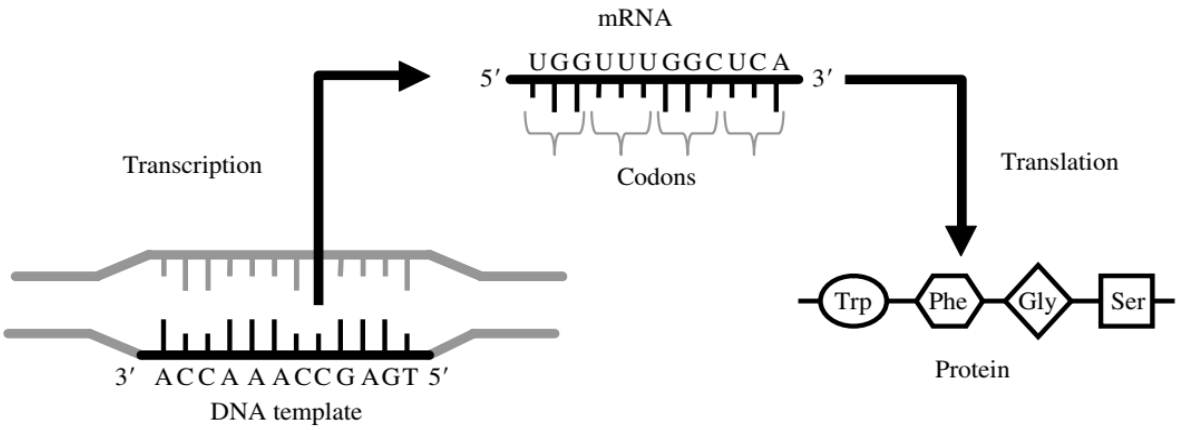

Fig. 2.4 Gene expression is carried out in two steps: *transcription* of DNA to RNA, followed by *translation* of the messenger RNA (mRNA) to protein. The correspondence between a *codon* (a triplet of bases) and the translated amino acid is given by the genetic code (Table 2.1).

A key question is the correspondence between the mRNA sequence and the amino-acid sequence of the protein product. One of the triumphs of molecular genetics is the discovery of the universal *genetic code* shown in Table 2.1. The genetic code gives the correspondence between *codons* (three-nucleotide sequences) on the mRNA and the 20 amino acids found in almost all naturally occurring proteins. There is a total of 4^3 or 64 possible codons, all listed in Table 2.1. The code also specifies codons that signal termination and initiation of translation. The code is degenerate in the sense that more than one codon can specify a single amino acid (but not vice versa). As depicted in Fig. 2.5, small RNAs (composed of 73 to 93 nucleotides) called transfer RNAs (tRNAs) act as adaptor molecules that read the mRNA codons. Each tRNA has a sequence of three nucleotides called an *anticodon* that matches the mRNA codon by Watson–Crick complementarity. The ribosome moves along the mRNA, and the charged tRNAs (i.e. those carrying their specific amino acids) enter in the order specified by the mRNA codons (see Fig. 2.5). The contiguous amino acids are then enzymatically joined to form polypeptides (proteins).

One can conclude that the amino-acid sequences of all cellular proteins are encoded in the DNA. Changes in certain DNA sequences can have drastic consequences on the shape and function of translated proteins. For example, a particular mutation in the hemoglobin gene (namely, a specific GAG sequence in the DNA is changed to GTG) leads to the disease called sickle-cell anemia; here, the corresponding single amino-acid change causes a drastic change in the shape of hemoglobin that compromises the protein's function as carrier of oxygen in red blood cells. The shape of proteins largely determines their biological functions, giving a rationale to many observations that, in the course of evolution, the three-dimensional structures of proteins are better conserved than their one-dimensional amino-acid sequences. Although many advances have been made recently, the problem of predicting three-dimensional structures of proteins from their one-dimensional amino-acid sequence is still not solved.

Table 2.1 The genetic code: from RNA codons to amino acids. A 'stop' codon signifies termination of translation. AUG (Met) is the usual initiator codon, but CUG and GUG are also used as initiator codons in rare instances. The 3-letter symbols in this table are for the following amino acids: L-Alanine (Ala), L-Arginine (Arg), L-Asparagine (Asn), L-Aspartic acid (Asp), L-Cysteine (Cys), L-Glutamic acid (Glu), L-Glutamine (Gln), Glycine (Gly), L-Histidine (His), L-Isoleucine (Ile), L-Leucine (Leu), L-Lysine (Lys), L-Methionine (Met), L-Phenylalanine (Phe), L-Proline (Pro), L-Serine (Ser), L-Threonine (Thr), L-Tryptophan (Trp), L-Tyrosine (Tyr), L-Valine (Val).

First position (5′ end)	Second position: U	Second position: C	Second position: A	Second position: G	Third position (3′ end)
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	stop	stop	A
	Leu	Ser	stop	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu (Met)	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met (start)	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val (Met)	Ala	Glu	Gly	G

Molecular pathways and networks

Although many of the so-called *housekeeping genes* are constitutively expressed for cell maintenance, there are also many other genes whose expressions respond or adapt to conditions of the cell environment. As a specific example, the bacterium *E. coli* can synthesize tryptophan (Trp) if the level of this amino acid in the extracellular medium is low; otherwise the bacterium shuts off its endogenous Trp-synthesizing machinery. The network of molecular interactions regulating Trp synthesis, from the transcription and translation of genes to the metabolic pathway that generates the amino acid, will be analyzed in Chapter 4. The Trp network is a good example of how the expression of genes can be affected by their products – thus forming feedback loops in the network.

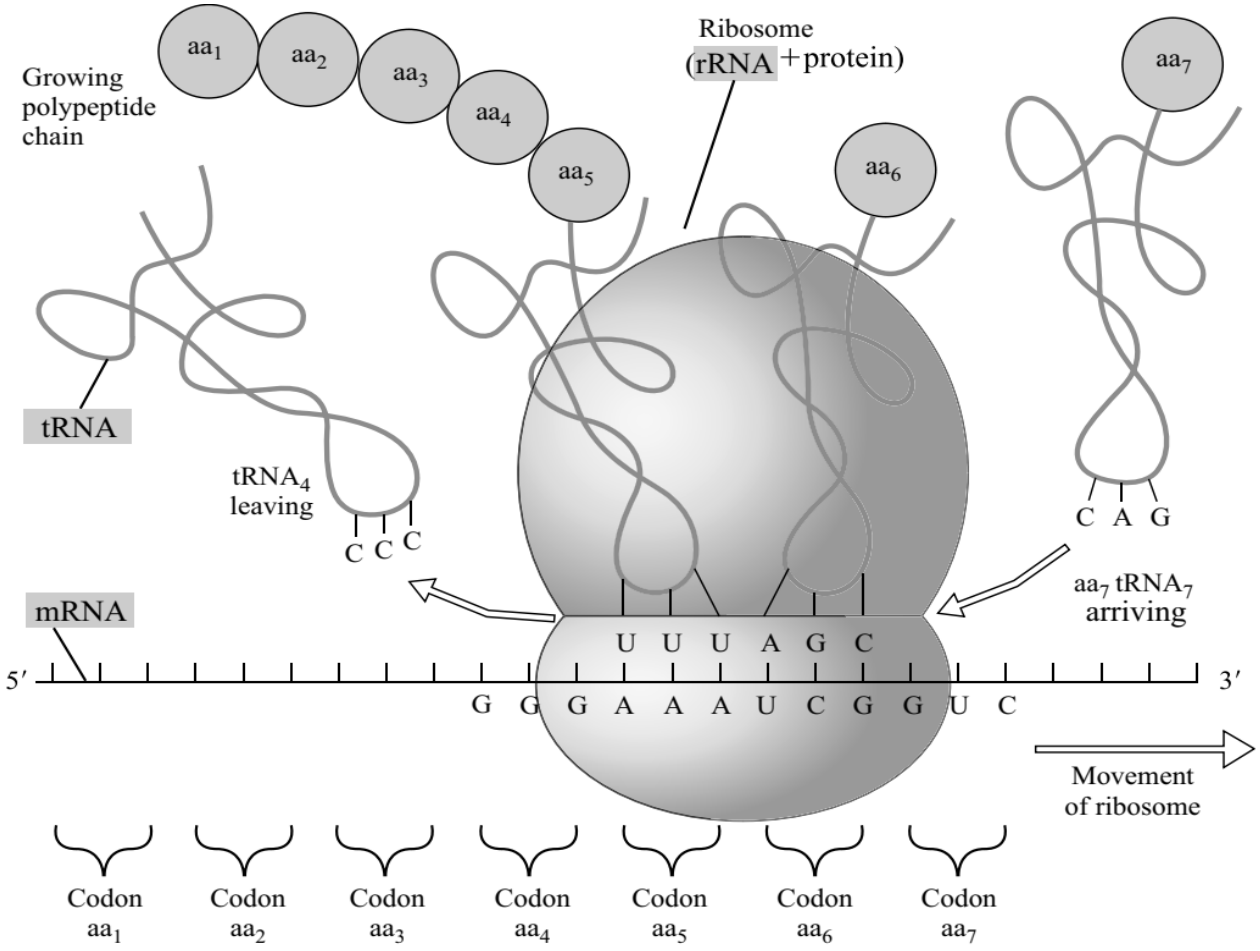

Fig. 2.5 A cartoon of how the ribosome moves along the mRNA to translate the codons to amino acids – in collaboration with tRNAs that are charged with corresponding amino acids (*circles* labelled *aa*s in the diagram).

The metabolic steps in the synthesis of the 20 amino acids in the universal genetic code, as well as other essential biomolecules – nucleotides, lipids, carbohydrates and many others – are coupled in a complex web of metabolic reactions. The steps in the metabolism of these biomolecules require enzymes (proteins) to occur, and therefore one can claim that the set of biochemical reactions in a cell is orchestrated by the information contained in its genome. A glimpse of the complexity of metabolic pathways is shown in Fig. 2.6.

In addition to metabolic networks, many other cellular networks involve the regulation of the activities of enzymes and other proteins. Enzymes are found in both inactive and active states, and the switching between these states involve regulatory networks whose complexity may reflect the importance of the enzyme function. These post-translational protein networks add another layer in the complexity of cellular networks. Figure 2.7 is a broad summary of these networks as they relate to the 'DNA-to-RNA-to-protein' flow of information; the general network shown in the

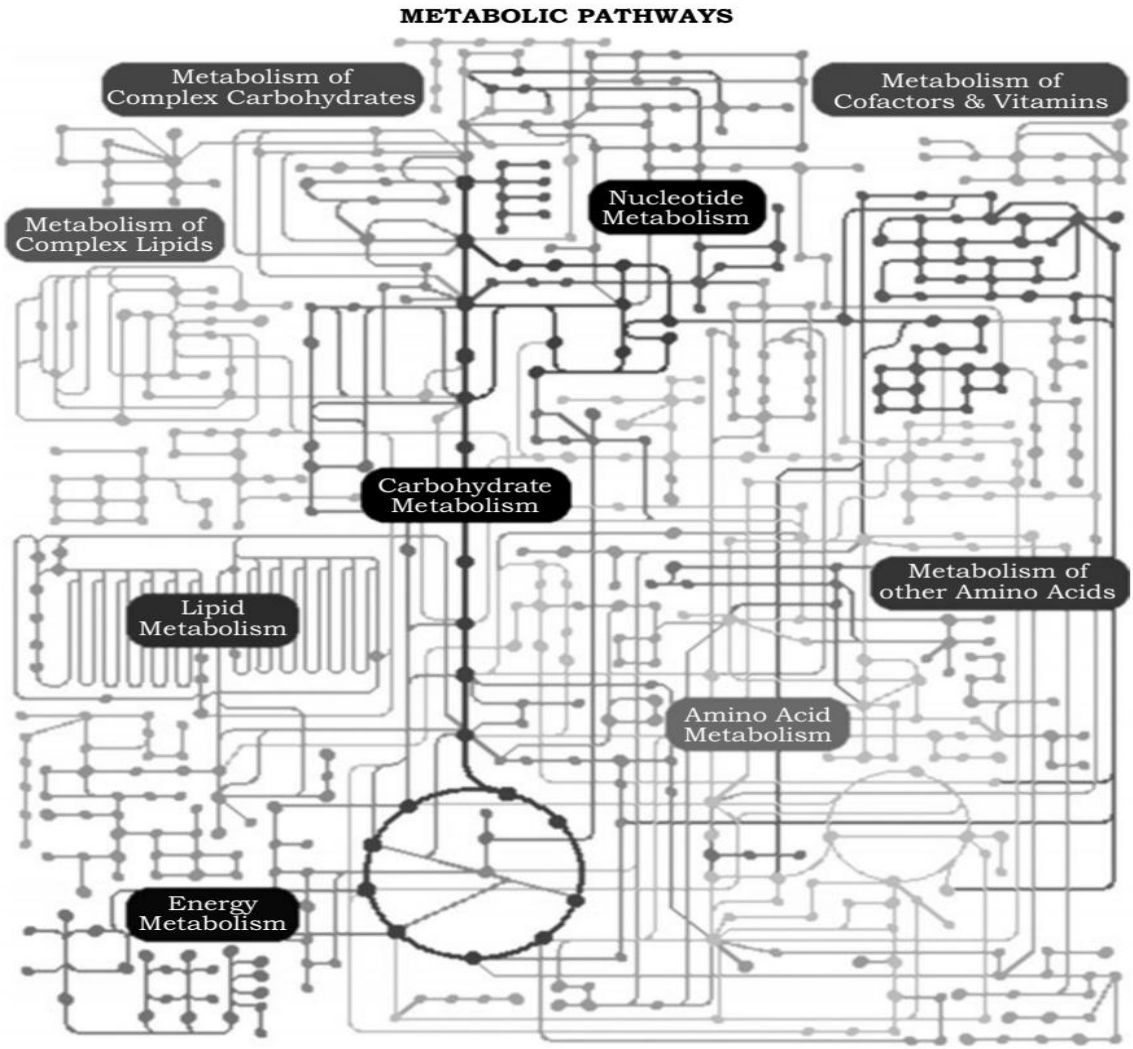

Fig. 2.6 Metabolic pathways from the online database KEGG (Kyoto Encyclopedia of Genes and Genomes, see Table 2.2 for its internet address). Each dot in the above 'wiring' diagram represents a metabolite (usually a small organic molecule). The edge between dots represents a chemical reaction that is catalyzed by an enzyme (which, in turn, is usually synthesized by a cell's gene-expression machinery).

figure is referred to in this book as *gene-regulatory networks* (GRNs). As indicated by the many feedback loops in this diagram, the information flow is not strictly linear; for example, reverse transcription from RNA to DNA is accomplished by retroviruses. Feedback loops may occur at every step during gene expression where

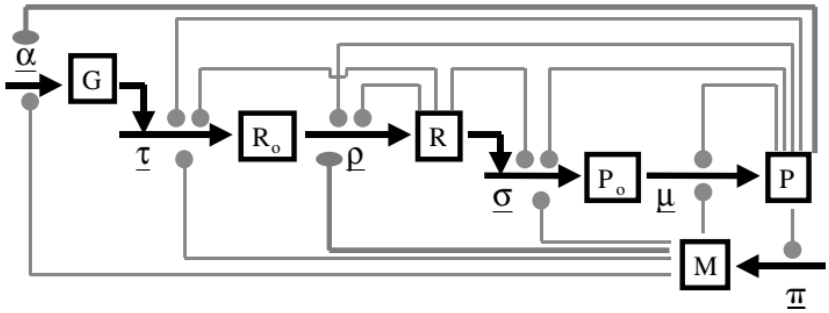

Fig. 2.7 A broad summary of gene-regulatory networks. The *arrow* labelled π represents metabolic networks requiring proteins (P) to catalyze reactions that produce metabolites (M); in general, metabolites are needed in every step of the gene-expression machinery. The *arrow* labelled α represents the replication of the genomic DNA – a process that needs metabolites (nucleotides), proteins (polymerases), and RNA (edge from R is not shown in figure). Transcriptional units (G) in the genome are transcribed in step τ to primary RNA transcripts (R_o) that are processed in step ρ to form mature transcripts (R). Proteins, such as transcription factors, can directly influence the transcription step τ. The translation of mRNAs to proteins (P_o) in step σ requires the co-operation of many proteins, tRNAs, and ribosomal RNAs. Step μ represents post-translational modifications of proteins that render them functional. The edges that end in dots (regulating the steps in the network) represent either activatory or inhibitory influence.

products can influence the rates of information flow, as well as which information is to be transmitted.

The existence of the many feedback loops depicted in Fig. 2.7 presents a formidable challenge in the analysis of GRNs. Many chapters in this book deal with models that implicitly assume a modularization of these large cellular networks – that is, focusing only on subnetworks that are assumed to explain particular cellular phenomena or functions. This reductionist approach is open to question in light of the highly connected property of cellular networks.