# CHAPTER 1

# INTRODUCTION



## 1.1 BASIC CONCEPTS IN MOLECULAR BIOLOGY

We introduce some basic and central concepts in modern molecular biology in this section to help readers understand the related problems discussed in the later chapters. Note that this is a very general and brief introduction, and arranged mainly for computer scientists and mathematicians who are trying to acquire a reading knowledge about molecular biology. Biology-oriented researchers can skip the details in this section. For more detailed and systematic biological knowledge, readers can refer to professional books (e.g., [Sta02], [Kar02], [Bro02], [Sad07]).

All living things, whether simple or complex organisms, are composed of cells, which are the basic units of structure and function in an organism [Sta02]. Each cell is a complex system consisting of many different building blocks. According to their sizes and types of internal structures, cells are classified as prokaryotic cells and eukaryotic cells, which, in turn, distinguish organisms into prokaryotic organisms (or prokaryotes) and eukaryotic organisms (or eukaryotes). Prokaryotic organisms, represented by bacteria and blue algae, are made up of prokaryotic cells that are smaller and have simpler internal structures, whereas eukaryotic organisms such as fungi, plants, and animals are composed of structurally complex eukaryotic cells [Kar02]. The distinction between eukaryotes and prokaryotes leads to the vast differences between many cellular building blocks and life processes in these two organism types.

Both eukaryotic and prokaryotic cells contain a nuclear region with the genetic materials of living organisms. However, the genetic materials of a prokaryotic cell

*Biomolecular Networks*. By Luonan Chen, Rui-Sheng Wang, and Xiang-Sun Zhang

are contained in a nucleoid without a boundary membrane, whereas a eukaryotic cell has a nucleus that is separated from the rest of the cell by a complex membranous structure or nuclear envelope. Note that besides nuclear membrane, both prokaryotes and eukaryotes have cell membranes or plasma membranes, which regulate the flow of nutrients, energy, and information in and out of the cell and play important roles in signal transduction. Despite this difference, eukaryotic cells have a molecular chemistry composition similar to that of prokaryotic cells. For example, both eukaryotic and prokaryotic organisms possess a genome in their cell that contains the biological genetic information needed to maintain life in that organism. Another essential feature of most living cells is their ability to reproduce and grow in an appropriate environment through cell division. New cells are generated from the reproduction of existing cells to maintain the life in living beings.

Cells consist of four basic types of molecules: (1) small molecules, (2) DNA, (3) RNA, and (4) protein. Small molecules in cells include water, sugars, fatty acids, amino acids, and nucleotides. They are either the basic building blocks of the macromolecules (DNA, RNA, proteins) or independent units with important roles, such as signal transduction and energy sources. Most eukaryotic and prokaryotic genomes consist of deoxyribonucleic acid (DNA), but a few viruses have ribonucleic acid (RNA) genomes [Bro02]. DNA and RNA are polymeric large molecules made up of chains of monomeric subunits.

DNA is the hereditary material in almost all organisms. Most DNA is located in the cell nucleus, but a small amount of DNA can also be found in the mitochondria. DNA is a linear polymer of four chemically distinct nucleotides consisting of three components: 2′-deoxyribose (a type of sugar composed of five carbon atoms labeled from 1′ to 5′), a phosphate group attached to the 5′-carbon of the sugar, and a nitrogenous base. Four kinds of nucleotides differ in their nitrogenous bases: adenine (A), cytosine (C), guanine (G), and thymine (T), which are usually referred to as *bases*, denoted by their initial letters, A, C, G, and T (Fig. 1.1). Hence, a DNA sequence can always be denoted by a string of A, C, G, T. Individual nucleotides are linked by phosphodiester bonds between their 5′-carbon and 3′-carbon in any order to form a DNA chain called a polynucleotide. A DNA molecule is actually double-stranded, and its nucleotide bases on two strands form complementary pairs: A pairing with T, and C pairing with G. The orientations of DNA strands are determined by the carbons at their ends which conventionally start from the 5′ ends to the 3′ ends (Fig. 1.1). The two strands are tied together and form a stable structure known as the *DNA double helix*, which was identified in 1953 in Cambridge by Watson and Crick. (Fig. 1.2).

RNA is also a polynucleotide, and its structure is similar to that of DNA except for two main differences [Bro02]: (1) the sugar in a RNA nucleotide is ribose rather than deoxyribose, and (2) RNA contains uracil (U) instead of thymine (T). In addition, the structure of RNA generally does not form a double helix as does the structure of DNA. The functions of DNA and RNA for living cells are also different. Generally, DNA is responsible for encoding genetic information and performs one essential function, while several types of RNA perform different functions, such as ribosomal RNAs and transfer RNAs. RNA also contains 3′–5′ phosphodiester bonds, but these

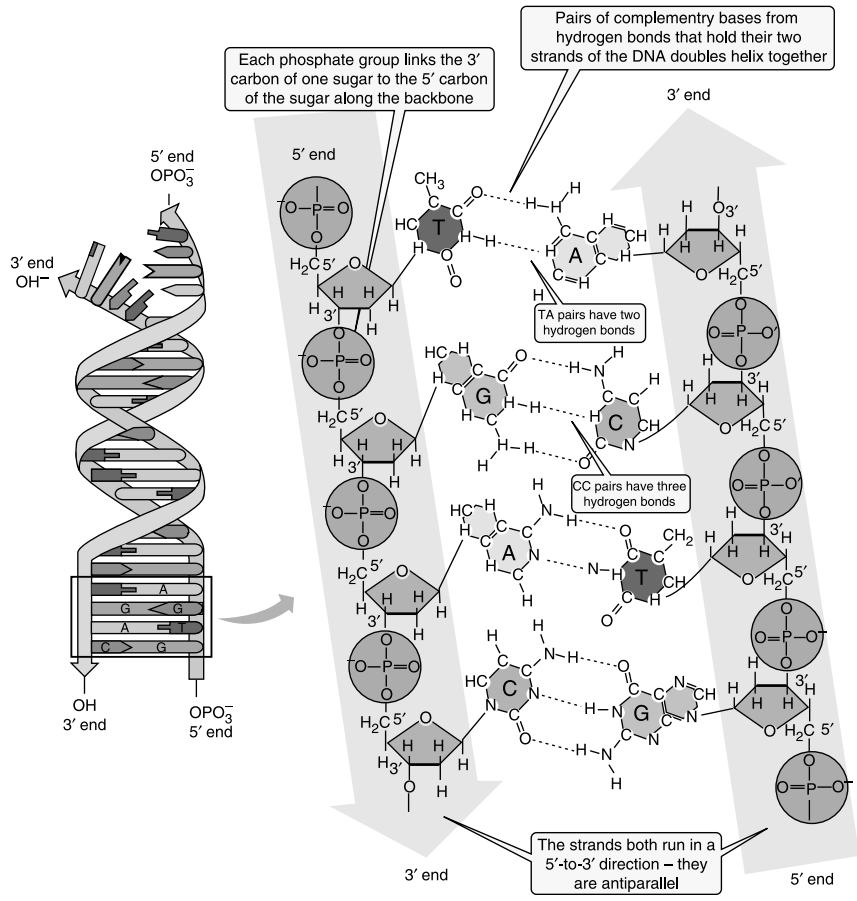


**Figure 1.1.** The double-helix DNA backbone with complementary base pairs. [Reprinted from [Sad06] © (2006) with permission of Sinauer Associates, Inc.]

bonds are not as stable as those in a DNA polynucleotide [Bro02]. In RNA polynucleotide, A complements or "pairs" with U, and C pairs with G. Such complementary base-pairing leads to folded structures of RNA that help RNA molecules carry out their functions in the expression of genes.

DNA encodes RNA and protein molecules through a law dominating the whole biology, which is called as the central "dogma" of molecular biology (Fig. 1.3). It provides a framework for understanding the flow of information from DNA via RNA and then to protein. Three important biological processes in the central "dogma" of molecular biology are replication, transcription, and translation. First, certain contiguous DNA segments containing biological information must be duplicated through a replication process to transmit the genetic information from parents to progeny. Then, the information contained in a section of DNA is transferred to a newly assembled piece of messenger RNA (mRNA) through a transcription process,

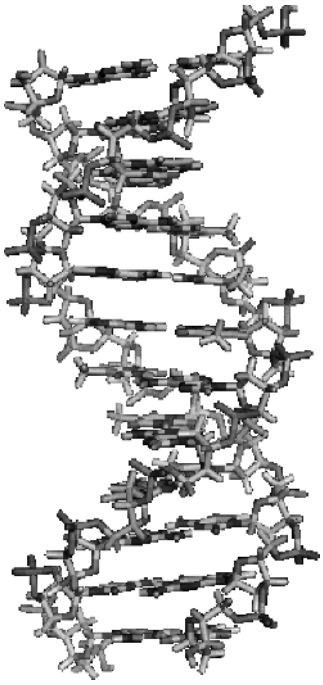

**Figure 1.2.** The double-helix structure of a DNA.

in which RNA polymerase and transcription factors play an important role. This transcription process is completed in the cell nucleus with the synthesis of RNA molecules. Finally, mRNAs are transported into a protein-synthesizing "factory" (i.e., ribosome) and read by the ribosome as triplet codons through a translation

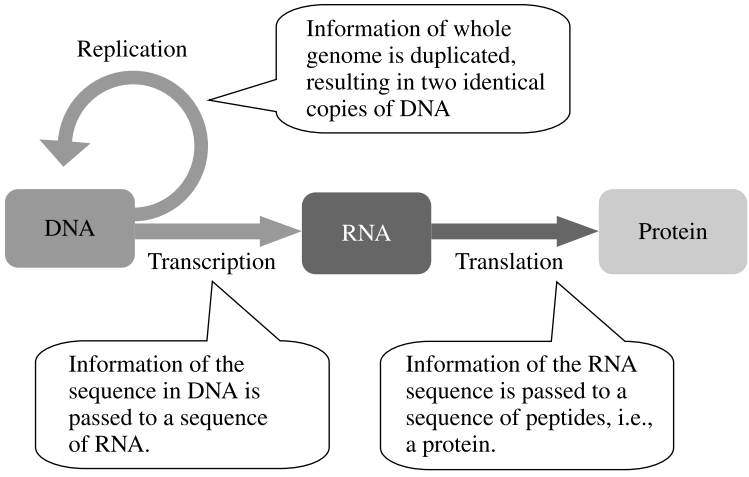


**Figure 1.3.** The central "dogma" of molecular biology.

process, which further synthesizes proteins. In Sections 1.1.1–1.1.3 we will describe these biological processes in detail.

### 1.1.1 Genomes, Genes, and DNA Replication Process

According to the number of cells that they contain, organisms may be unicellular or multicellular. Bacteria and baker's yeast are representative examples of unicellular organisms that consist of only one cell. Most organisms consist of two or more cells. Each cell contains one or more DNA molecules. A chromosome is formed from a single DNA molecule. In prokaryotes, DNA is organized in the form of a circular chromosome. In eukaryotes, chromosomes have a complex structure where DNA is wound around structural proteins called histones. Most of the DNA in eukaryotes is located in the cell nucleus and is called chromosomal DNA. But a small amount of DNA can also be found in the mitochondria, which is called mitochondrial DNA. Both chromosomal and mitochondrial DNA in a cell constitute a genome. Owing to DNA replication in the process of cell division, all cells in an organism contain identical genomes with few rather special exceptions. The total number of chromosomes and genome size differ quite considerably in different organisms. For example, each cell in Homo sapiens has 23 pairs of chromosomes, whereas a fruit fly has 4 pairs and a yeast has 12 pairs of chromosomes. The human genome has about 3 billion base pairs. Determining the four-letter order for a given DNA molecule is known as *DNA sequencing*. Since the first full genome for a bacterium was sequenced in 1995, genomes of many organisms have been sequenced. The well-known Human Genome Project was completed in 2001, and a draft human genome was obtained.

As mentioned earlier, information encoded in static DNA is passed to functional protein molecules through transcription and translation processes. However, not all portions of DNA are used for encoding proteins. A continuous stretch of DNA molecule that contains the information necessary to encode a protein is called a gene. Other portions are termed "junk DNA," which is actually not real "junk"; such noncoding portions have been found to perform important functions [Soo06, Lev07]. In cells, genes consist of a long strand of DNA that contains an important region for controlling gene transcription called a promoter. In addition to promoter regions, genes in eukaryotic organisms contain regions called introns and exons (Fig. 1.4). The introns will be removed from mRNAs in a process called *splicing*. The regions encoding gene products are called exons, which are interspersed with noncoding introns. The number and size of introns and exons differ considerably between different genes and different species. In eukaryotes, a single gene can encode multiple proteins through different alternative splice variants, that is, the same pre-mRNA produces different mRNAs by different arrangements of exons known as alternative splicing. In prokaryotes, genes seldom have introns and thereby there is no splicing.

DNA replication is the process of copying a double-stranded DNA molecule or a whole genome, a process essential in all known life forms. The general mechanisms of DNA replication are also different in prokaryotic and eukaryotic organisms. As each DNA strand holds the same genetic information, both strands can serve as templates for the reproduction of the opposite strand. The template strand is preserved in its

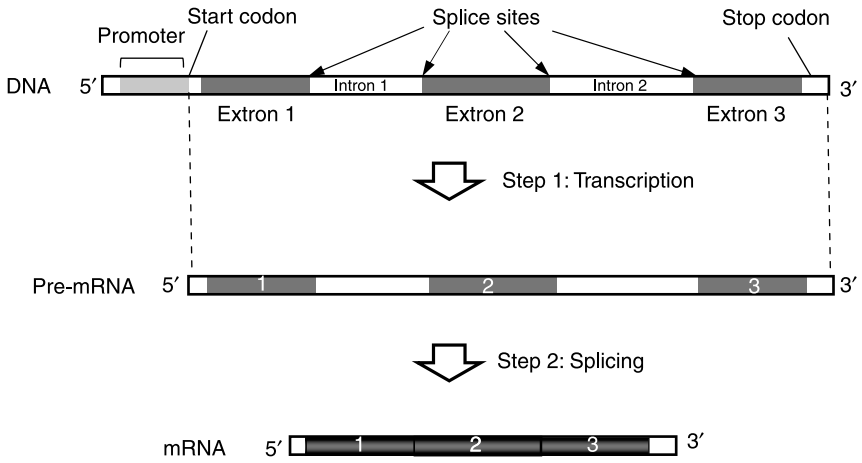


**Figure 1.4.** The structure of eukaryotic genes and splicing process.

entirety and the new strand is assembled from nucleotides. This process is called semiconservative replication. The resulting double-stranded DNA molecules are identical; proofreading and error-checking mechanisms exist to ensure extremely high fidelity. In a cell, DNA replication must occur before cell division. Prokaryotes replicate their DNA throughout the interval between cell divisions. On the other hand, the replication of eukaryotic cells progresses through a regular cycle of growth and division termed as cell cycle, consisting of four phases: S phase, during which DNA is synthesized; M phase, during which the actual cell division or mitosis occurs; and two gap phases, G1 and G2, which fall between M and S phases and between S and M phases, respectively. In other words, the replication timings of DNA in eukaryotes are highly regulated, and this occurs during the S phase of the cell cycle, preceding mitosis.

### 1.1.2 Transcription Process for RNA Synthesis

In all organisms, there are two major steps necessary for DNA producing proteins: (1) information of the DNA on which the gene resides is transcribed to messenger RNA (mRNA), and (2) information on the mRNA is translated to the protein. Transcription is the process of producing mRNA using genes as templates. In the transcription process, one strand of DNA molecule is copied into a complementary pre-mRNA by an enzyme called RNA polymerase II. To initiate transcription, the two-stranded double-helix structure of DNA molecule is "unzipped." The DNA strand whose sequence matches that of the RNA is known as the coding strand and the strand to which the RNA is complementary is the template strand. Then, RNA polymerase II first recognizes and binds a promoter region of the gene. It begins to read the template strand in the $3'-5'$ direction, splice the introns, and synthesize the primary transcript mRNA from $5'$ to $3'$. It is worth noting that the splicing of introns present within the

transcribed region is unique to eukaryotes. In prokaryotes, transcription occurs in the cytoplasm. In contrast, transcription in eukaryotes necessarily occurs in the nucleus. After such a transcription process, mRNA is synthesized and will be transported to ribosomes to form proteins. However, the mature mRNA may be further modified by other biochemicals, such as noncoding RNA, before the translation.

The process of producing functional molecules such as RNA or protein is called *gene expression*. In addition to transcription and translation, the steps in the gene expression process may be further modulated, including the posttranscriptional regulation of an mRNA and the posttranslational modification of a protein. Messenger RNA can be quantitatively measured by many techniques such as DNA microarray technology, which is now widely adopted to study many problems in biology.

### 1.1.3 Translation Process for Protein Synthesis

Translation is a process of forming proteins by using a mature mRNA molecule as a template. It is the second stage of protein biosynthesis and an important part of gene expression. Translation takes place in the cytoplasm where ribosomes are located. In the translation process, mRNA is decoded to produce a specific polypeptide according to the rules known as triplet or genetic code, which specifies the mapping from mRNA nucleotide bases (codons) to 20 specific amino acids (Fig. 1.5). There are start and stop codons to indicate the beginning and ending of a gene. Since there are 64 codons and only 20 amino acids, the code is redundant; that is, an amino acid may be represented

Second letter

| First letter | U | | C | | A | | G | | Third letter |
|---|---|---|---|---|---|---|---|---|---|
| U | UUU | Phenyl | UCU | Serine | UAU | Tyrosine | UGU | Cysteine | U |
| | UUC | alanine | UCC | | UAC | | UGC | | C |
| | UUA | Leucine | UCA | | UAA | Stop codon | UGA | Stop codon | A |
| | UUG | | UCG | | UAG | | UGG | Tryptophan | G |
| C | CUU | Leucine | CCU | Proline | CAU | Histidine | CGU | Arginine | U |
| | CUC | | CCC | | CAC | | CGC | | C |
| | CUA | | CCA | | CAA | Glutamine | CGA | | A |
| | CUG | | CCG | | CAG | | CGG | | G |
| A | CUU | Isoleucine | ACU | Threonine | AAU | Asparagine | AGU | Serine | U |
| | CUC | | ACC | | AAC | | AGC | | C |
| | CUA | | ACA | | AAA | Lysine | AGA | Arginine | A |
| | CUG | Methionine; Start codon | ACG | | AAG | | AGG | | G |
| G | GUU | Valine | GCU | Alanine | GAU | Aspartic | GGU | Glycine | U |
| | GUC | | GCC | | GAC | acid | GGC | | C |
| | GUA | | GCA | | GAA | Glutamic | GGA | | A |
| | GUG | | GCG | | GAG | acid | GGG | | G |

**Figure 1.5.** The mapping rules (genetic codes) from codons to amino acids.

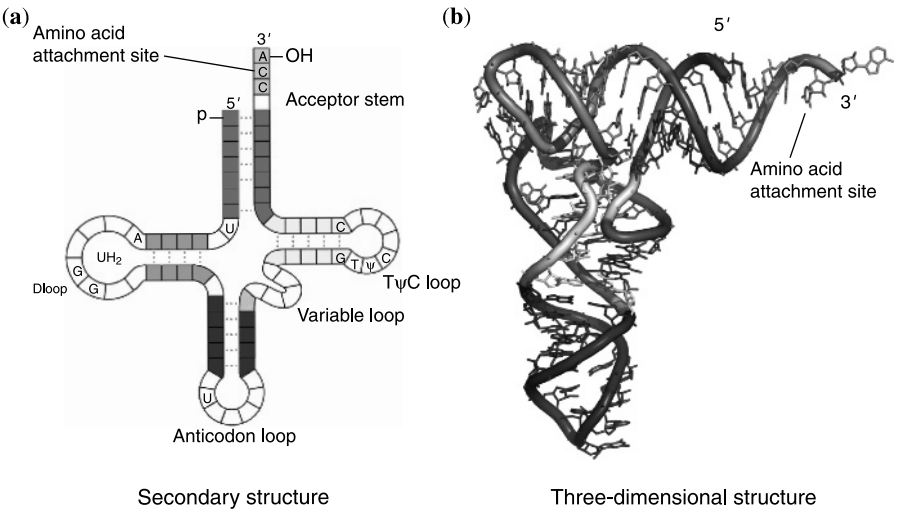


**Figure 1.6.** The structure of tRNA.

by more than one codon. For example, histidine is encoded by CAT and CAC, but a single codon can represent only one amino acid.

After the transcription process, mRNA carries genetic information encoded as a ribonucleotide sequence from chromosomes to ribosomes. In cytoplasm, mRNA forms a complex with ribosomes. Transfer RNA (tRNA) is a small noncoding RNA chain that transports amino acids to the ribosome and makes the connection between a codon and the corresponding amino acid (Fig. 1.6). Ribosome and tRNA molecules read the ribonucleotides by translational machinery and guide the synthesis of a chain of amino acids to form a protein. After the translation process, gene expression is completed. The final product of gene expression is a protein. The protein is still subject to multiple posttranslational biochemical modifications before becoming a mature, active, and functional molecule, such as degradation, dimerization, and phosphorylation. It is worth noting that, as a result of alternative splicing and posttranslational modifications, one gene can produce multiple proteins. After its synthesis, the new protein folds to its active three-dimensional structure before carrying out cellular functions.

## 1.2 BIOMOLECULAR NETWORKS IN CELLS

Through the transcription and translation processes, gene products such as mRNA and protein are produced. Gene, mRNA, and protein are known as biological molecules or basic components. The complicated relations and interactions between these components are responsible for diverse cellular functions. At the genome or DNA level, transcription factors (TFs) function as DNA-binding proteins and can activate

or inhibit the transcription of genes to synthesize mRNAs by regulating the activities of genes. Since these TFs themselves are products of genes, the ultimate effect is that genes regulate each other's expression as part of a transcription (or transcriptional) regulatory network (TRN) or gene regulatory network (GRN). Similarly, at the proteome or protein level, proteins can participate in diverse posttranslational modifications of other proteins or form protein complexes and pathways together with other proteins that assume new roles. Such local associations between protein molecules are called protein–protein interactions (PPIs), which form a protein interaction network. The biochemical reactions in cellular metabolism can likewise be integrated into a metabolic network whose fluxes are regulated by enzymes that catalyze the reactions. In many cases, these interactions at different levels are integrated into a signaling network. For example, external signals from the exterior of a cell are first mediated to the inside of that cell by a cascade of protein–protein interactions of the signaling molecules. Then, both biochemical reactions and transcription regulations including protein–DNA interactions trigger the expression of some genes to respond the signals [Alb05]. In short, although cells consist of various biological molecules, their cellular processes and functions are actually achieved by biomolecular networks with the collaborative effects of those individual components. Figure 1.7(b) illustrates several typical molecular networks at different levels in cellular systems, which are the backbone of network systems biology. From the viewpoint of network architecture, main ingredients in this book are molecules, interactions, pathways, and networks. Their hierarchical relations are conceptually shown in Figure 1.7(a), where a cellular system can also be viewed to be formed conceptually from individual molecules, to pairwise interactions, to local structures (including network motifs, modules, pathways, and subnetworks), and eventually to global networks. In other words, basic components in a cellular system are individual molecules, which affect each other by their pairwise interactions. A cascade of those pairwise interactions forms a local structure (i.e., linear pathway or a subnetwork) which transforms local perturbations into a functional response. And all of linear pathways or subnetworks are assembled into a global biomolecular network which eventually generates global behaviors and holds responsibility for complicated life in a living organism. In terms of interactions, each type of molecular network is assembled by the following different pairwise interactions: transcription regulatory network: TF–DNA interactions; gene regulatory network: gene–gene interactions (or genetic interactions); protein interaction network: protein–protein interactions; metabolic network: enzyme–substrate interactions; signaling network: molecule–molecule interactions.

The completion of the Haemophilus influenzae genome sequence in 1995 marked the beginning of the genomic era [Fle95]. The advent of whole-genome sequencing technologies leads to hundreds of complete genome sequences. Especially after the release of the draft version of the human genome sequence [Ven01], we are now entering into a postgenomic era and begin to analyze the transcriptome and the proteome of many model organisms. In this era, various high-throughput experimental techniques in molecular biology can provide genome-scale measurements from biological molecules that exist within the cell such as genes (DNA), proteins, RNA, metabolites,

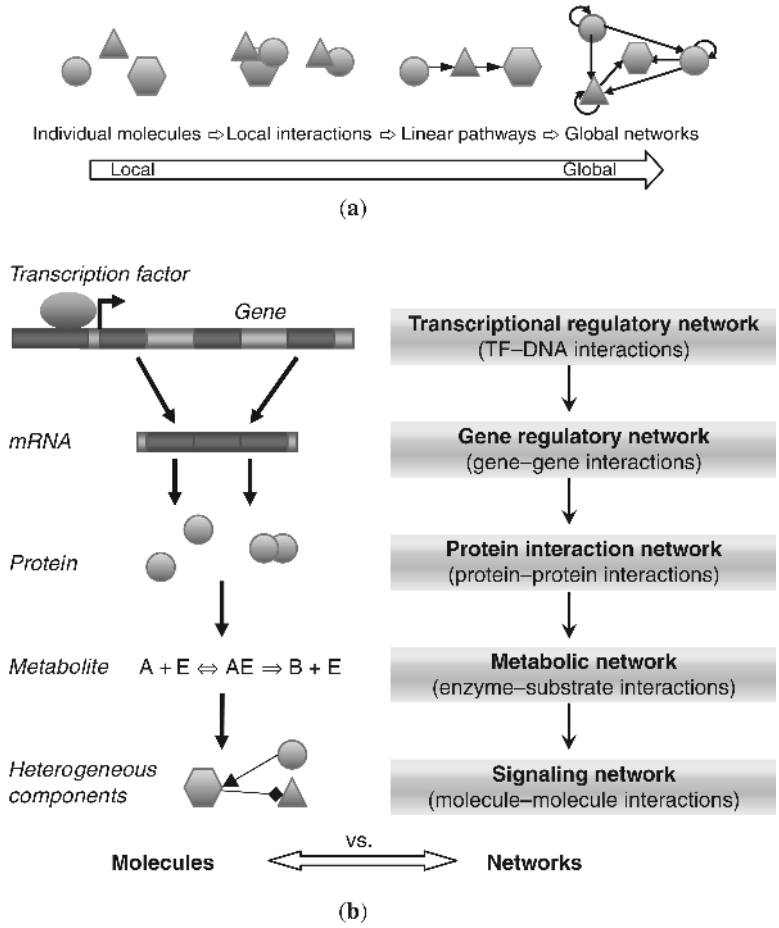


**Figure 1.7.** Ingredients in cellular systems in terms of network architecture. (a) Hierarchical relations of molecules, interactions, pathway, and networks. (b) Hierarchical relations of various biomolecular networks. In (a), "Local interactions" are mainly pairwise interactions, and "Linear pathways" are local network structures, including pathways, modules, communities, network motifs and subnetworks.

and other molecules, and have resulted in an enormous amount of component data. In addition, the functional genomic and proteomic approaches have generated a variety of protein–protein, protein–DNA, and other component–component interaction mappings, which make it possible to study biomolecular networks mentioned above. The resulting datasets by these experimental techniques run through the information flow of the central dogma of molecular biology, and include genome, transcriptome, proteome, metabolome, localizome, and interactome components, which are collectively referred to as "omic" data and provide comprehensive descriptions of all

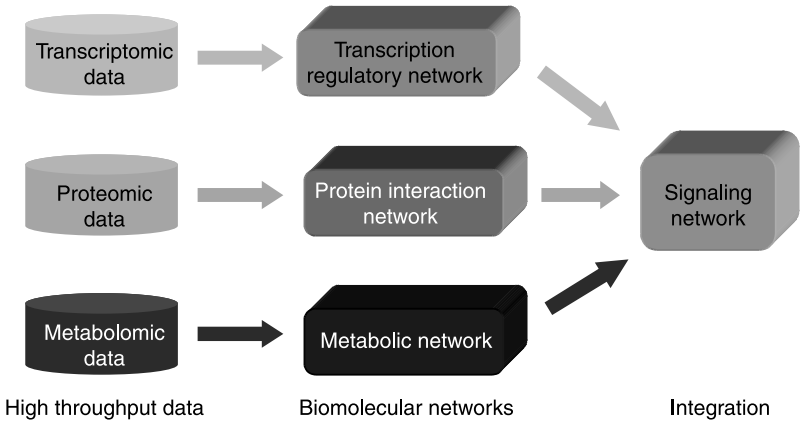


**Figure 1.8.** Omic data and biomolecular networks.

components and interactions within the cell [Joy06]. Figure 1.8 illustrates the relations between omic data and biomolecular networks.

- *Transcriptomic Data–Transctiption Regulatory Network*. Transcriptome profiling is one of the first omic approaches developed. DNA chips, microarrays and serial analysis of gene expression (SAGE) are the most widely used approaches for examining the expression of thousands of genes simultaneously under various experimental conditions and have generated large amounts of mRNA transcripts [Har05]. Such data have been applied to many fields, such as identifying differentially expressed genes in stem cells, classifying the molecular subtypes of human cancers, and monitoring the host cell transcriptional response to pathogens. Gene expression is the result of transcription factors regulating target genes; hence it is possible to retrieve the interaction relationships between different genes from a large amount of gene expression data. Such pairwise interaction relationships are combined into gene regulatory networks. In addition, the ChIP-chip technique helps determine protein–DNA interactions [LeT02], which constitute transcription regulatory networks describing special functional modules of interest. In addition, transcription factors regulate genes by binding to upstream and downstream regulatory regions of transcription start sites. With the availability of whole-genome sequences, identification of regulatory regions and transcription factor binding sites has become feasible from a computational viewpoint.
- *Proteomic Data–Protein Interaction Network*. Although the analysis of proteomics has lagged behind that of transcriptomics, the functions of all proteins and how they form complexes during various conditions are now beginning to be systematically explored. Two-dimensional gel electrophoresis (2DE) and mass spectroscopy (MS) have been used to identify and quantify the activity, binding, and other cellular levels of proteins [Par03]. For protein spot detection,

conventional staining techniques such as colloidal Coomassie Brilliant Blue (CBB) and silver staining are being popular. Yeast two-hybrid (Y2H) is one of the first methods for high-throughput protein–protein interaction mapping and has been used to determine the interactomes of many organisms. Besides Y2H, tandem affinity purification (TAP) and phage library display are also used. Such protein–protein interactions can be represented as a protein interaction network, from which much useful knowledge can be extracted. For example, protein interactions provide rich information for protein function and signaling pathway information.

- *Metabolomic Data–Metabolic Network.* As one of the new types of omic data, the methods used to generate the complete set of metabolites of many organisms are still being refined. MS, nuclear magnetic resonance (NMR) spectroscopy, and vibrational spectroscopy have been used to analyze the metabolite contents that are extracted from isolated cells or tissues [Joy06]. The resulting data make it possible to study the dynamic metabolic response of living systems to environmental stimuli or genetic perturbations through analyzing metabolic networks, in which the nodes denote metabolites and the edges represent reactions or enzymes. A metabolic network provides not only a list of metabolite components but also a functional readout of the cellular state. Given the highly diverse set of biomolecules and the large dynamic range of metabolite concentrations, sophisticated computational techniques are needed to reconstruct and analyze various biochemical reaction pathways and networks.
- *Integrated Data–Signaling Network.* Integrating the above mentioned interaction data at different levels leads to a signaling network or a hierarchical molecular network. A signaling network involves the transduction of a variety of signals such as energy and stimuli from the outside to the inside of the cell. It is one of the main parts of cellular communication and relies on an underlying series of biochemical reactions, transcription regulations, and protein interactions. Except in a very few cases, experimentally determining a complete signaling network is a time-consuming and also costly task. However, with the increasing deposition of various types of data, reconstructing a signaling network from multiple information sources is becoming a promising topic and feasible task that attracts much attention from the researchers in systems biology and computational biology. Depending on the types of data, the integrated system may be not only a hierarchical but also a heterogeneous molecular network with diverse substructures.

In contrast to component data such as genomic and proteomic data providing a specific molecular content of a cellular system, pairwise interaction data include protein–DNA interactions, protein–protein interactions, and protein–ligand (enzyme–substrate) interactions, which determine the local connectivity that exists among the molecular species, and provide a network scaffold within the cell system [Joy06]. The subsequent function data are closely related to the interaction data since many biological processes in cells are not performed by individual components but through gene regulations, signal transduction, and interactions between

biomolecules. It is the local interactions of those components that are assembled into a global network and are ultimately responsible for an organism's form and functions [Bar04, Har99]. Generally, a living organism can be viewed as a huge nonlinear biochemical reaction system characterized by the interactions of biomolecules, including genes, RNAs, proteins, and metabolites. Such local and pairwise interactions are often represented by a global biomolecular network in which each node is a biological molecule or complex, and each edge represents an interaction or association of two molecules. Generally, biomolecular networks include gene regulatory networks (gene–gene interactions), transcription regulatory networks (TF–gene interactions), signaling networks (integrated interactions among molecules), protein interaction networks (protein–protein interactions), metabolic networks (enzyme–substrate interactions), and hybrid networks. These biomolecular networks indispensably exist in cell systems and play fundamental and essential roles in giving rise of life.

In short, with various interaction data available, the focus of biological research is being transformed from analyzing individual components to studying global networks from a systematic perspective. Without depreciating the importance of individual molecules, the more recent research results indicate that a cellular function is actually the contribution of various kinds of interactions between a myriad of cellular constituents. In particular, a cellular system can be viewed as a networked biological system. Therefore, an important challenge for biology is to understand the cell's function organization by investigating the structure, function, and dynamics of complex biomolecular networks in living cellular systems.

## 1.3 NETWORK SYSTEMS BIOLOGY

To elucidate the essential principles of cellular systems, study of biomolecular networks is increasingly attracting much attention from various science and engineering communities. High-throughput experimental methods in molecular biology have resulted in an enormous amount of data, including interactions, networks and pathways [Bar04]. Hence, it is crucial that mathematicians and computer scientists provide computational tools to reveal the essential biological mechanisms from a system perspective. To meet such a challenge, rather than analyzing individual components or partial aspects of the organism, network systems biology, is to study an organism viewed as a dynamical interaction network of genes, proteins, and biochemical reactions by developing sophisticated theoretical methodologies and computational tools.

The goal of network systems biology is to mine knowledge on the basis of the networked data generated from high-throughput techniques by exploiting special features of the biological system, and gain biological insight by further interpreting them in a systematic manner. For example, understanding the process of specific gene regulations and signal transduction provides deep insight into the mechanisms of cellular systems. From a computational viewpoint, modeling the gene regulation process and signal transduction by appropriate mathematical models will enhance such knowledge. Given a large amount of gene expression data from microarray techniques, identifying gene–gene interactions and signaling pathways is by no means a

trivial thing. A large number of genes with few timepoints are a main characteristic of microarray data that hinders us from achieving this task. With the ChIP-chip data, estimating the activity profiles of transcription factors (TFs) is also a very important task, since measuring the activity of a TF is still experimentally difficult owing to chemical modifications after the translational process. At the same time, inferring the relationships between TFs and their targets from experimental data is of utmost importance for understanding direct interactions as well as the complex regulatory mechanisms in cellular systems.

A living organism or a cell is a highly organized system of interacting macromolecules and metabolites, which can be viewed as a huge molecular network formed by those local interactions of molecules, as shown in Figure 1.7. Therefore, as a discipline related closely to systems biology, network biology emphasizes local interactions and global networks of molecules **that characterize various biological systems**, and attempts to understand biology from the viewpoint of the global and local systems properties of molecular networks by **offering a quantifiable description of the interactions and networks**. The research subject of network systems biology or network biology is rich and diverse. For example, the huge deposit of gene expression profiles and protein–DNA interactions makes it possible to quantitatively study the regulatory relationships between genes. Reverse engineering of regulatory networks is one of main computational problems in this field. Protein interaction data from high-throughput techniques are highly "noisy" and incomplete with unknown portion of false positives. We are required to systematically integrate these data and further estimate their confidence by statistical techniques. This problem is known as protein interaction prediction. Interacting proteins are believed to have similar functions. The annotation of proteins is currently far from complete. We can enhance it by employing protein interaction data and other biological sources. This is the challenge of protein function prediction. With the current technologies, experimental determination of protein complexes and functional modules is not cost-effective. Actually, it is not only expensive but also time-consuming, and the result is not reliable. Can this be done by mining the data from protein interaction networks to provide a rough estimation for biologists? This is the problem defined as functional module detection. Given protein interaction networks or metabolic networks from multiple species, how can we compare these networks and extract important knowledge related to evolution? Given a concerned pathway, how can we find a similar one in a protein interaction network? These are problems of network alignment and query. In addition, in order to understand the structure and function of a living cell, we also should investigate the structure and dynamics of these biological networks. This is the challenge of dynamical modeling or qualitative analysis of biomolecular networks. In contrast to qualitative studies, quantitative simulation can directly predict the dynamic behaviors of living cells, and is an important topic related to the development of highly efficient computation algorithms on both stochastic and deterministic dynamic models. Do biomolecular networks have topological properties similar to those of other complex networks? Are the topological patterns (e.g., network motifs, modules, or hubs) of biomolecular networks related to specific biological functions? How can we reconstruct metabolic pathways and identify active subnetworks from a large set of biochemical reactions?

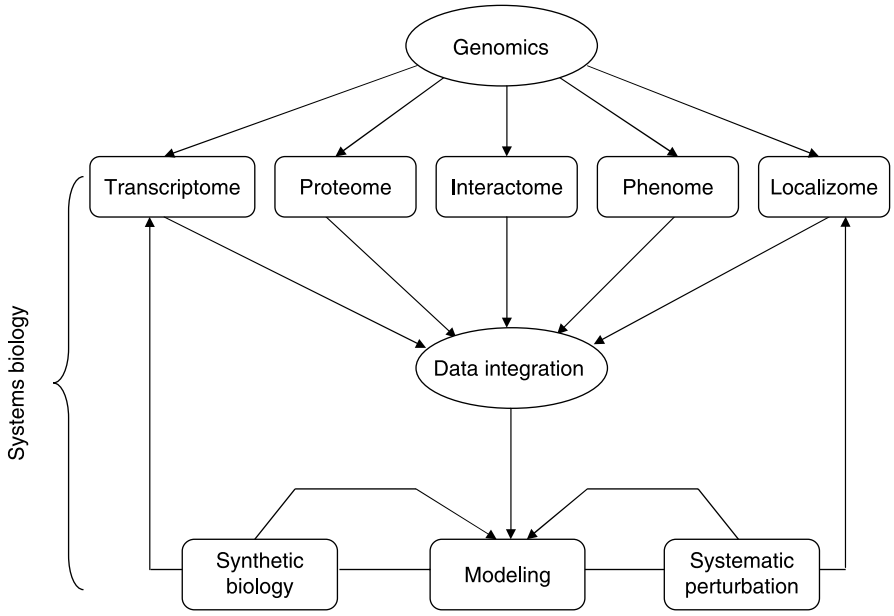


**Figure 1.9.** Systems biology focusing on integrating omic data. [Reprinted from [GeH03] © 2003 with permission from Elsevier.]

How can we detect signal transduction pathways or drug targets from data of perturbed biological experiments? Can we design or construct a synthetic biological network representing a whole or partial cellular system (i.e., forward engineering of biomolecular networks)? All of these problems are expected to be solved by computational means, which are the main research focuses of network systems biology.

With increasingly accumulated data from high-throughput technologies, biomolecular networks and their functional roles have been studied extensively from various aspects of living organisms. This research not only helps scientists understand complicated biochemical phenomena but also reveals the fundamental mechanisms of living organisms from a system perspective. Hence, systems biology with emphasis on networks is anticipated to enhance our understanding of cellular systems by integrating comprehensive data of molecular components in different layers and studying how the multitudes of interactions facilitate the complicated biological functions within a cell. Figure 1.9 illustrates the main research focus of systems biology. In this book, we particularly emphasize four aspects for network systems biology, represented by four keywords—network, dynamics, system, and integration:

- *Network.* A cellular system can be viewed as a huge biochemical reaction network that orchestrates the sophisticated and complex functions of the cells and thus gives rise to life. Living organisms differ from each other not only because of the differences of their constituting proteins but also because of the architectures of their molecular networks. The availability of genome sequences for

hundreds of organisms including humans leads to a transition from molecular biology to modular biology [Har99]. Since most gene products function in unison, cellular processes are considered to be the results of complex networks of individual components. Therefore, to elucidate fundamental cellular behaviors, it is essential to focus on the interactions between individual components and the functional states of these networks resulting from the assembly of all such local and pairwise interactions. It has been recognized that a complicated living organism cannot be fully understood by merely analyzing individual components, and that the global network of those components is ultimately responsible for an organism's form and governs the organism's behavior.

- *Dynamics.* Life is dynamic, and dynamics exists in living organisms at each level. Cellular systems are commonly modeled by nonlinear dynamical systems such as ordinary differential equations [ChL02, ChL04, WaR08], or stochastic processes such as the chemical master equation [ChL05, WaR08], based on mass action law and enzyme reaction kinetics. The ultimate goal of network systems biology is to understand a complex biological process in sufficient detail to enable us to build a computational network model for the process and gain deep insight into the principles of living organisms. Clearly, dynamic simulations of a cellular system can provide a more thorough quantitative understanding of its principles, mechanism and function [Zho05, LiC07a]. From both theoretical and experimental perspectives, it is a very challenging problem in biological science to model, analyze, and further predict the dynamic behaviors of biosystems [Zho08]. One of the most widely best studied dynamic or rhythmic phenomena so far is circadian oscillation [WaR08], which is assumed to be produced by limit cycle oscillators at the molecular level from the gene regulatory feedback loops or protein interaction loops. With the rapid advances in mathematics and experiments concerning the underlying regulatory mechanisms, developing more sophisticated theoretical models and general quantitative simulation techniques is increasingly necessary for elucidating dynamical behaviors in a cell at the system level.
- *System.* In cells, an individual component always receives signals and outputs information. There is a regular communication between different cells [ChL05, Zho05], which mediates their collective behaviors. All the components are wound together to form into a complex cellular system that collectively performs biological function and system behavior [WaR08, LiC06a]. The system behaviors are essentially coordinated responses resulting from the local interactions of individual components in both prokaryotes and eukaryotes. Such a mechanism is an absolute requisite to ensure appropriate and robust coordination of cell activities at all levels of organisms under an uncertain environment. Each cellular process can be studied in a systematic way, which means not only that the interactions of homogeneous components are important but also that the functional relationships between heterogeneous components are essential. In addition, perturbation is an approach often adopted in systems biology. When genes or proteins in a cellular system are systematically perturbed, responses

from other parts of the system can be recorded and the information obtained can be incorporated into the basic model [GeH03].

- *Integration.* Integration has multiple implications, including integrating different data sources, integrating different levels of systems, integrating different technologies, integrating different disciplinary areas, and even integrating different human resources. To make full use of high-throughout data, clearly we need to integrate not only heterogeneous data sources but also different methodologies and different levels of systems. Most cellular processes involve some components of gene regulation as well as protein interactions. For example, a membrane protein receiving an external signal may trigger a cascade of protein interactions that results in one or more genes being expressed in the genome. Understanding this interplay between proteins and DNA clearly requires data integration. When proteins interact to accomplish a specific process, some of the genes encoding them may be expressed in a coordinated manner. Therefore, integrating microarray gene expression data and protein interaction network can provide new insights. In another view, the knowledge on most biological processes studied for many years in individual labs is fragmentary and stored in countless scientific papers. Hence, integrating various sources of knowledge on interactions and regulations from scientific literature is also necessary and imperative. Finally, different computational methods and tools have their own advantages and limitations, and integrating methodologies could also enhance our ability to analyze the large amount of data in an accurate and robust manner.

Except for the main characteristics of systems biology, sophisticated computational and analytical methods are definitely indispensable as tools of network systems biology. For instance, reverse-engineering gene regulatory networks requires optimization models to fit time-series experimental data. Mining useful knowledge from the topological properties of biomolecular networks cannot be done without machine learning and data-mining methods. Dealing with noise and uncertainty underlying experimental data demands appropriate probabilistic or statistical methods. In particular, integrating all kinds of omic data necessitates efficient data integration techniques such as kernel methods and Bayesian methods. Finally, a network itself is a kind of graph, and thus graph-theoretic methods are an effective tool for systems biology. Combining these computational methods with the systematic framework of studying cell mechanism, systems biology stressing on networks, or network systems biology is an increasingly promising discipline for studying complex life phenomena (Fig. 1.10).

One of the great challenges in this area is to build a complete and high-resolution description of molecular topography and connect biomolecular interactions with physiological responses. By investigating the relationships and interactions between various parts of a biological system such as gene regulatory systems, protein interaction networks, metabolic pathways, organelles, cells, physiological systems, and organisms, we expect to eventually develop an understandable model of the whole

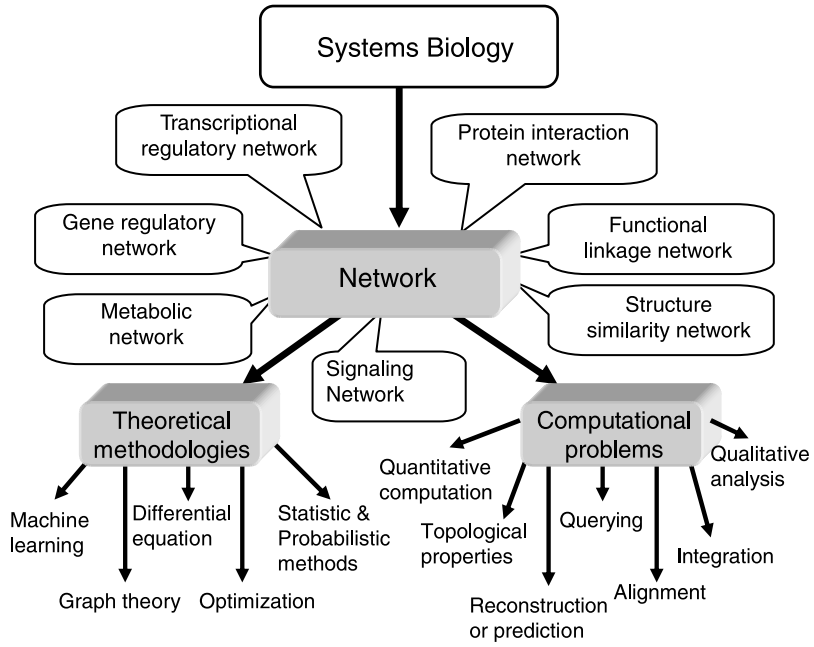


**Figure 1.10.** The research focus of network systems biology.

cellular system, which is critical for a thorough understanding of the essential mechanisms of living organisms.

## 1.4 ABOUT THIS BOOK

This book focuses primarily on various kinds of biomolecular networks, with particular emphasis on computational problems, methods, and applications in bioinformatics and systems biology. It provides a general theoretical and methodological framework for analyzing biomolecular networks. In the book, many mathematical concepts and methods, such as graph theory, optimization theory, probability theory, statistics, thermodynamical theory and differential equations, are adopted to solve the computational problems in bioinformatics and systems biology (Fig. 1.11), and these methods play important roles in many interesting and sophisticated applications. In contrast to conventional bioinformatics, which studies mainly the individual components or local interactions of biological systems, this book presents a new research area, where machine learning and computation techniques, such as text mining, classification, clustering, and visual techniques, find their alternative applications on global networks. Computer scientists or mathematicians will find that it is in high demand to develop new techniques suited for this increasingly important area, to enable them to deal with large-scale networks, heterogeneous data, and intense computation. This book also gives a comprehensive survey for biomolecular networks and the biological

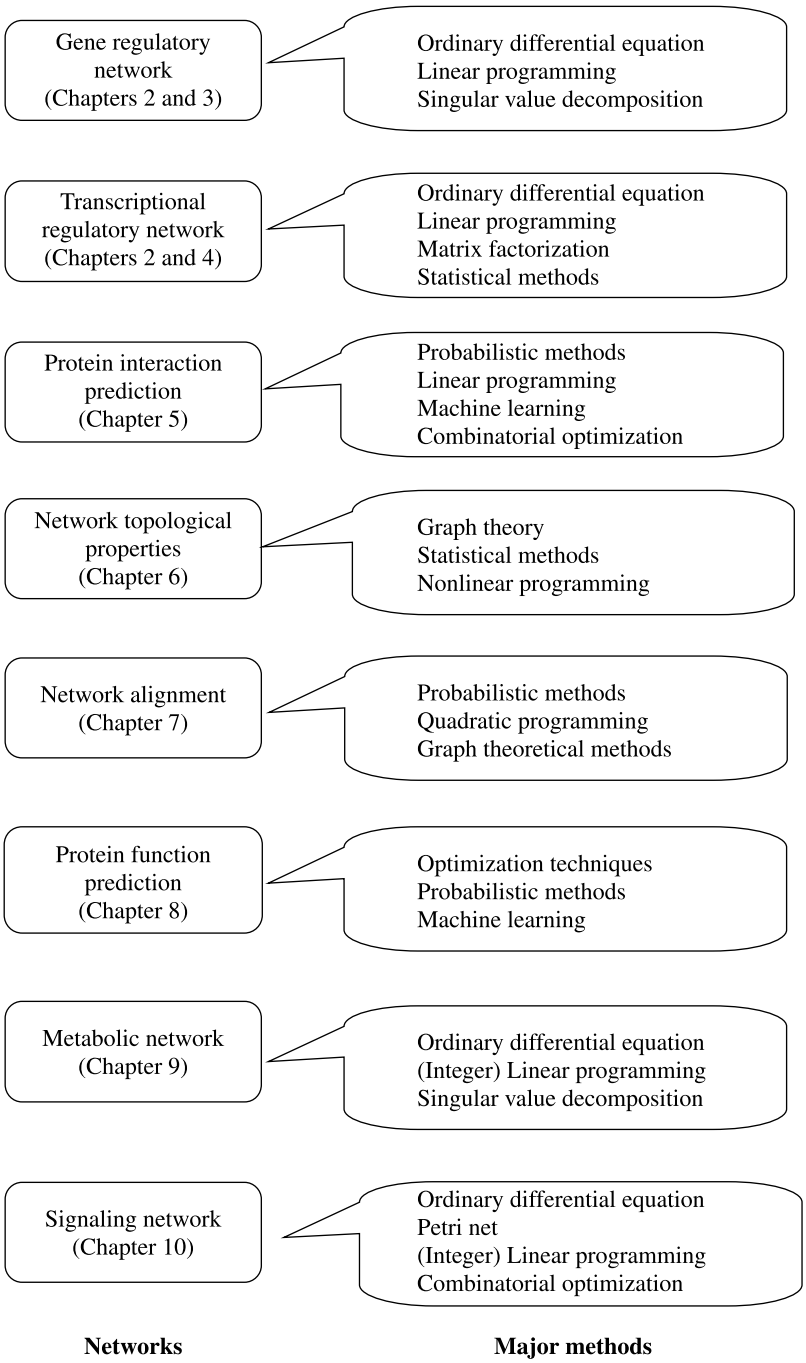


**Figure 1.11.** Biomolecular networks with major computational tools applied in this book.

relations between them, which can significantly enhance the understanding and knowledge of the essential mechanisms of living organisms from a system viewpoint. On the other hand, biologists will find in this book many useful tools, and algorithms that can be utilized to analyze biological networks and further design biological experiments, such as network inference tools, network alignment tools, functional module detection tools, and drug target detection tools. The book assumes basic knowledge of molecular biology with each chapter covering the necessary materials. The objective of this book is to help biology-oriented researchers and other researchers understand the state-of-art techniques in bioinformatics and systems biology. It covers extensive topics related to biomolecular networks and the latest trends such as

- Reverse engineering of gene regulatory networks
- Inferring transcriptional interaction and regulator activity*
- Protein–protein interaction prediction
- Topological analysis of biomolecular networks
- Alignment of biomolecular networks
- Network-based function prediction and annotation
- Uncovering signal transduction networks
- Metabolic network modeling and reconstruction
- Drug target detection in metabolic networks
- Integration of heterogenous data and heterogeneous networks

The remainder of this book is organized as follows:

- Chapter 2 introduces basic concepts related to gene regulation and gene expression, as well as microarray technologies. In this chapter, gene regulatory networks and transcription regulatory networks are described with basic kinetic models.
- Chapter 3 first presents several mathematical models for modeling gene regulatory networks, and then reviews several representative methods for reconstructing gene regulatory networks, particularly emphasizing the approaches integrating multiple datasets.
- Chapter 4 describes the basic principle of ChIP-chip technology and reports the most recent advances of quantitative studies on transcription regulation, including how to infer transcriptional interactions, reveal combinatorial regulation mechanisms, and reconstruct TF activity profiles.
- Chapter 5 is one of the main parts of this book. In this chapter, we first introduce some experimental techniques for determining protein–protein interactions and then focus on computational prediction methods based on domain information. Among various approaches, we introduce probabilistic approaches, optimization

*Identifying the regulatory roles of microRNAs in the process of post transcription regulation.

methods, and other topics. Finally, several methods, particularly designed for domain interaction prediction, are discussed.

- Chapter 6 discusses the statistical properties of protein interaction networks from the topological viewpoint. In particular, hubs, motifs, and modularity are described and utilized with their explorative biological roles. In addition, a new criterion, modularity density $D$ for characterizing the modularity structure of complex networks is also described.
- Chapter 7 introduces a variety of network alignment methods including pairwise network alignment and multiple network alignment. In particular, a quadratic programming approach for pairwise network alignment is discussed in detail. Subnetwork and pathway query methods are also covered.
- Chapter 8 focuses on protein function prediction and is also an important part of this book. First, methods for detecting functional modules and for creating protein function linkages are extensively explored, and then several protein function prediction methods based on high-throughput data are discussed, including optimization methods and machine learning methods. Finally, we report the most recent advances of annotation approaches for domain function.
- Chapter 9 describes the biological principle of metabolism and introduces some analysis methods for metabolic pathways and networks in living organisms. Computational approaches for reconstructing and simulating metabolic networks are included. In addition, on the basis of available metabolic networks, we introduce an effective drug target detection method.
- Chapter 10 first introduces the biological principles of signal transduction and discusses mathematical modeling of signaling pathways. Then we describe several computational methods for uncovering signaling networks from high-throughput data sources or experimental evidences.
- Chapter 11 discusses some new and promising topics on both artificial and real networks to conclude this book, such as protein structure networks, integrated heterogeneous networks, and the posttranscriptional regulation networks for noncoding RNAs (e.g., microRNAs).