Rui Jiang
Xuegong Zhang
Michael Q. Zhang *Editors*

# Basics of Bioinformatics

## Lecture Notes of the Graduate Summer School on Bioinformatics of China

TSINGHUA UNIVERSITY PRESS

Springer

# Basics of Bioinformatics

Rui Jiang • Xuegong Zhang • Michael Q. Zhang
Editors

# Basics of Bioinformatics

## Lecture Notes of the Graduate Summer School on Bioinformatics of China

TSINGHUA UNIVERSITY PRESS

Springer

*Editors*
Rui Jiang
Xuegong Zhang
Department of Automation
Tsinghua University
Beijing
China, People's Republic

Michael Q. Zhang
Department of Molecular and Cell Biology
The University of Texas at Dallas
Richardson, TX, USA

Tsinghua National Laboratory
for Information Science and Technology
Tsinghua University
Beijing, China, People's Republic

ISBN 978-3-642-38950-4 ISBN 978-3-642-38951-1 (eBook)
DOI 10.1007/978-3-642-38951-1
Springer Heidelberg New York Dordrecht London

Jointly published with Tsinghua University Press, Beijing
ISBN: 978-7-302-32359-4 Tsinghua University Press, Beijing

Library of Congress Control Number: 2013950934



Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

# Foreword

This ambitious volume is the result of the successful 2007 Graduate Summer School on Bioinformatics of China held at Tsinghua University. It is remarkable for its range of topics as well as the depth of coverage. Bioinformatics draws on many subjects for analysis of the data generated by the biological sciences and biotechnology. This foreword will describe briefly each of the 12 chapters and close with additional general comments about the field. Many of the chapters overlap and include useful introductions to concepts such as gene or Bayesian methods. This is a valuable aspect of the volume allowing a student various angles of approach to a new topic.

Chapter 1, "Basics for Bioinformatics," defines bioinformatics as "the storage, manipulation and interpretation of biological data especially data of nucleic acids and amino acids, and studies molecular rules and systems that govern or affect the structure, function and evolution of various forms of life from computational approaches." Thus, the first subject they turn to is molecular biology, a subject that has had an enormous development in the last decades and shows no signs of slowing down. Without a basic knowledge of biology, the bioinformatics student is greatly handicapped. From basic biology the authors turn to biotechnology, in particular, methods for DNA sequencing, microarrays, and proteomics. DNA sequencing is undergoing a revolution. The mass of data collected in a decade of the Human Genome Project from 1990 to 2001 can be generated in 1 day in 2010. This is changing the science of biology at the same time. A 1,000 genome project became a 10,000 genome project 2 years later, and one expects another zero any time now. Chromatin Immunoprecipitation or ChIP allows access to DNA bound by proteins and thus to a large number of important biological processes. Another topic under the umbrella of biological sciences is genetics, the study of heredity and inherited characteristics (phenotypes). Heredity is encoded in DNA and thus is closely related to the goals of bioinformatics. This whole area of genetics beginning with Mendel's laws deserves careful attention, and genetics is a key aspect of the so-called genetic mapping and other techniques where the chromosomal locations of disease genes are sought.

Chapter 2, "Basic Statistics for Bioinformatics," presents important material for the understanding and analysis of data. Probability and statistics are basic to bioinformatics, and this chapter begins with the fundamentals including many classical distributions (including the binomial, Poisson, and normal). Usually the observation of complete populations such as "all people in China over 35 years old" is not practical to obtain. Instead random samples of the population of interest are obtained and then inferences about parameters of the population are made. Statistics guides us in making those inferences and gaining information about the quality of the estimates. The chapter describes techniques such as method of moments, maximum likelihood, and Bayesian methods. Bayesian methods have become indispensable in the era of powerful computing machines. The chapter treats hypothesis testing which is less used than parameter estimation, but hypothesis testing provides understanding of $p$-values which are ubiquitous in bioinformatics and data analysis. Classical testing situations reveal useful statistics such as the $t$-statistic. Analysis of variance and regression analysis are crucial for testing and fitting large data sets. All of these methods and many more are included in the free open-source package called *R*.

Chapter 3, "Topics in Computational Genomics," takes us on a tour of important topics that arise when complete genome information is available. The subject did not begin until nearly 2000 when complete genome sequences became a possibility. The authors present us with a list of questions, some of which are listed next. What are the genes of an organism? How are they turned off and on? How do they interact with each other? How are introns and exons organized and expressed in RNA transcripts? What are the gene products, both structure and function? How has a genome evolved? This last question has to be asked with other genomes and with members of the population comprising the species. Then the authors treat some of the questions in detail. They describe "finding protein coding genes," "identifying promoters," "genomic arrays and a CGH/CNP analysis," "modeling regulatory elements," "predicting transcription factor binding sites," and motif enrichment and analysis. Within this last topic, for example, various word counting methods are employed including the Bayesian methods of expectation maximization and Gibbs sampling.

An alert reader will have noticed the prominence of Bayesian methods in the preceding paragraphs. Chapter 4, "Statistical Methods in Bioinformatics," in this collection focuses on this subject. There is a nice discussion of statistical modeling and then Bayesian inference. Dynamic programming, a recursive method of optimization, is introduced and then employed in the development of Hidden Markov Models (HMMs). Of course the basics of Markov chains must also be covered. The Metropolis-Hastings algorithm, Monte Carlo Markov chains (MCMC), and Gibbs sampling are carefully presented. Then these ideas find application in the analysis of microarray data. Here the challenging aspects of multiple hypothesis testing appear, and false discovery rate analysis is described. Hierarchical clustering and bi-clustering appear naturally in the context of microarray analysis. Then the issues of sequence analysis (especially multiple sequence analysis) are approached using these HHM and Bayesian methods along with pattern discovery in the sequences.

Discovering regulatory sequence patterns is an especially important topic in this section. The topics of this chapter appear in computer science as "machine learning" or under "data mining"; here the subject is called statistical or Bayesian methods. Whatever it is named, this is an essential area for bioinformatics.

The next chapter (Chap. 5), "Algorithms in Computational Biology," takes up the formal computational approach to our biological problems. It should be pointed out that the previous chapters contained algorithmic content, but there it was less acknowledged. It is my belief that the statistical and algorithmic approaches go hand in hand. Even with the Euclid's algorithm example of the present chapter, there are statistical issues nearby. For example, the three descriptions of Euclid's algorithm are analyzed for time complexity. It is easy to ask how efficient the algorithms are on randomly chosen pairs of integers. What is the expected running time of the algorithms? What is the variance? Amazingly these questions have answers which are rather deep. The authors soon turn to dynamic programming (DP), and once again they present clear illustrative examples, in this case Fibonacci numbers. Designing DP algorithms for sequence alignment is covered. Then a more recently developed area of genome rearrangements is described along with some of the impressive (and deep) results from the area. This topic is relevant to whole genome analysis as chromosomes evolve on a larger scale than just alterations of individual letters as covered by sequence alignment.

In Chap. 6, "Multivariate Statistical Methods in Bioinformatics Research," we have a thorough excursion into multivariate statistics. This can be viewed as the third statistical chapter in this volume. Here the multivariate normal distribution is studied in its many rich incarnations. This is justified by the ubiquitous nature of the normal distribution. Just as with the bell-shaped curve which appears in one dimension due to the central limit theorem (add up enough independent random variables and suitably normalized, one gets the normal under quite general conditions), there is also a multivariate central limit theorem. Here detailed properties are described as well as related distributions such as the Wishart distribution (the analog of the chi-square). Estimation is relevant as is a multivariate $t$-test. Principal component analysis, factor analysis, and linear discriminant analysis are all covered with some nice examples to illustrate the power of approaches. Then classification problems and variable selection both give platforms to further illustrate and develop the methods on important bioinformatics application areas.

Chapter 7, "Association Analysis for Human Diseases: Methods and Examples," gives us the opportunity to look more deeply into aspects of genetics. While this chapter emphasizes statistics, be aware that computational issues also drive much of the research and cannot be ignored. Population genetics is introduced and then the important subjects of genetic linkage analysis and association studies. Genomic information such as single-nucleotide polymorphisms (SNPs) provide voluminous data for many of these studies, where multiple hypothesis testing is a critical issue.

Chapter 8, "Data Mining and Knowledge Discovery Methods with Case Examples," deals with the area of knowledge discovery and data mining. To quote the authors, this area "has emerged as an important research direction for extracting useful information from vast repositories of data of various types. The basic

concepts, problems and challenges deals with the area of knowledge discovery and data mining that has emerged as an important research direction for extracting useful information from vast repositories of data of various types. The basic concepts, problems and challenges are first briefly discussed. Some of the major data mining tasks like classification, clustering and association rule mining are then described in some detail. This is followed by a description of some tools that are frequently used for data mining. Two case examples of supervised and unsupervised classification for satellite image analysis are presented. Finally an extensive bibliography is provided."

The valuable chapter on Applied Bioinformatics Tools (Chap. 9) provides a step-by-step description of the application tools used in the course and data sources as well as a list of the problems. It should be strongly emphasized that no one learns this material without actually having hands-on experience with the derivations and the applications. This is not a subject for contemplation only!

Protein structure and function is a vast and critically important topic. In this collection it is covered by Chap. 10, "Foundations for the Study of Structure and Function of Proteins." There the detailed structure of amino acids is presented with their role in the various levels of protein structure (including amino acid sequence, secondary structure, tertiary structure, and spatial arrangements of the subunits). The geometry of the polypeptide chain is key to these studies as are the forces causing the three-dimensional structures (including electrostatic and van der Waals forces). Secondary structural units are classified into α-helix, β-sheets, and β-turns. Structural motifs and folds are described. Protein structure prediction is an active field, and various approaches are described including homology modeling and machine learning.

Systems biology is a recently described approach to combining system-wide data of biology in order to gain a global understanding of a biological system, such as a bacterial cell. The science is far from succeeding in this endeavor in general, let alone having powerful techniques to understand the biology of multicellular organisms. It is a grand challenge goal at this time. The fascinating chapter on Computational Systems Biology Approaches for Deciphering Traditional Chinese Medicine (Chap. 11) seeks to apply the computational systems biology (CSB) approach to traditional Chinese medicine (TCM). The chapter sets up parallel concepts between CSB and CTM. In Sect. 11.3.2 the main focus is "on a CSB-based case study for TCM ZHENG—a systems biology approach with the combination of computational analysis and animal experiment to investigate Cold ZHENG and Hot ZHENG in the context of the neuro-endocrine-immune (NEI) system." With increasing emphasis on the so-called nontraditional medicine, these studies have great potential to unlock new understandings for both CSB and TCM.

Finally I close with a few remarks about this general area. Biology is a major science for our new century; perhaps it will be the major science of the twenty-first century. However, if someone is not excited by biology, then they should find a subject that does excite them. I have almost continuously found the new discoveries such as introns or microRNA absolutely amazing. It is such a young science when such profound wonders keep showing up. Clearly no one analysis subject can

solve all the problems arising in modern computational molecular biology. Statistics alone, computer science alone, experimental molecular biology alone, none of these are sufficient in isolation. Protein structure studies require an entire additional set of tools such as classical mechanics. And as systems biology comes into play, systems of differential equations and scientific computing will surely be important. None of us can learn everything, but everyone working in this area needs a set of well-understood tools. We all learn new techniques as we proceed, learning things required to solve the problems. This requires people who evolve with the subject. This is exciting, but I admit it is hard work too. Bioinformatics will evolve as it confronts new data created by the latest biotechnology and biological sciences.

University of Southern California
Los Angeles, USA
March 2, 2013

Michael S. Waterman

# Contents

# Chapter 1
# Basics for Bioinformatics

**Xuegong Zhang, Xueya Zhou, and Xiaowo Wang**

## 1.1 What Is Bioinformatics

Bioinformatics has become a hot research topic in recent years, a hot topic in several disciplines that were not so closely linked with biology previously. A side evidence of this is the fact that the 2007 Graduate Summer School on Bioinformatics of China had received more than 800 applications from graduate students from all over the nation and from a wide collection of disciplines in biological sciences, mathematics and statistics, automation and electrical engineering, computer science and engineering, medical sciences, environmental sciences, and even social sciences. So what is bioinformatics?

It is always challenging to define a new term, especially a term like bioinformatics that has many meanings. As an emerging discipline, it covers a lot of topics from the storage of DNA data and the mathematical modeling of biological sequences, to the analysis of possible mechanisms behind complex human diseases, to the understanding and modeling of the evolutionary history of life, etc.

Another term that often goes together or close with bioinformatics is computational molecular biology, and also computational systems biology in recent years, or computational biology as a more general term. People sometimes use these terms to mean different things, but sometimes use them in exchangeable manners. In our personal understanding, computational biology is a broad term, which covers all efforts of scientific investigations on or related with biology that involve mathematics and computation. Computational molecular biology, on the other hand, concentrates on the molecular aspects of biology in computational biology, which therefore has more or less the same meaning with bioinformatics.

X. Zhang (✉) • X. Zhou • X. Wang
MOE Key Laboratory of Bioinformatics and Bioinformatics Division, TNLIST/Department of Automation, Tsinghua University, Beijing 100084, China
e-mail: zhangxg@tsinghua.edu

R. Jiang et al. (eds.), *Basics of Bioinformatics: Lecture Notes of the Graduate Summer School on Bioinformatics of China*, DOI 10.1007/978-3-642-38951-1_1,

Bioinformatics studies the storage, manipulation, and interpretation of biological data, especially data of nucleic acids and amino acids, and studies molecular rules and systems that govern or affect the structure, function, and evolution of various forms of life from computational approaches. The word "computational" does not only mean "with computers," but it refers to data analysis with mathematical, statistical, and algorithmatic methods, most of which need to be implemented with computer programs. As computational biology or bioinformatics studies biology with quantitative data, people also call it as quantitative biology.

Most molecules do not work independently in living cells, and most biological functions are accomplished by the harmonic interaction of multiple molecules. In recent years, the new term systems biology came into being. Systems biology studies cells and organisms as systems of multiple molecules and their interactions with the environment. Bioinformatics plays key roles in analyzing such systems. People have invented the term computational systems biology, which, from a general viewpoint, can be seen as a branch of bioinformatics that focuses more on systems rather than individual elements.

For a certain period, people regarded bioinformatics as the development of software tools that help to store, manipulate, and analyze biological data. While this is still an important role of bioinformatics, more and more scientists realize that bioinformatics can and should do more. As the advancement of modern biochemistry, biophysics, and biotechnologies is enabling people to accumulate massive data of multiple aspects of biology in an exponential manner, scientists begin to believe that bioinformatics and computational biology must play a key role for understanding biology.

People are studying bioinformatics in different ways. Some people are devoted to developing new computational tools, both from software and hardware viewpoints, for the better handling and processing of biological data. They develop new models and new algorithms for existing questions and propose and tackle new questions when new experimental techniques bring in new data. Other people take the study of bioinformatics as the study of biology with the viewpoint of informatics and systems. These people also develop tools when needed, but they are more interested in understanding biological procedures and mechanisms. They do not restrict their research to computational study, but try to integrate computational and experimental investigations.

## 1.2 Some Basic Biology

No matter what type of bioinformatics one is interested in, basic understanding of existing knowledge of biology especially molecular biology is a must. This chapter was designed as the first course in the summer school to provide students with non-biology backgrounds very basic and abstractive understanding of molecular biology. It can also give biology students a clue how biology is understood by researchers from other disciplines, which may help them to better communicate with bioinformaticians.

### 1.2.1 Scale and Time

Biology is the science about things that live in nature. There are many forms of life on the earth. Some forms are visible to human naked eyes, like animals and plants. Some can only be observed under light microscope or electron microscope, like many types of cells in the scale of 1.100 μm and some virus in the scale of 100 nm. The basic components of those life forms are molecules of various types, which scale around 1.10 nm. Because of the difficulty of direct observation at those tiny scales, scientists have to invent various types of techniques that can measure some aspects of the molecules and cells. These techniques produce a large amount of data, from which biologists and bioinformaticians infer the complex mechanisms underlying various life procedures.

Life has a long history. The earliest form of life appeared on the earth about 4 billion years ago, not long after the forming of the earth. Since then, life has experienced a long way of evolution to reach today's variety and complexity. If the entire history of the earth is scaled to a 30-day month, the origin of life happened during days 3–4, but there has been abundant life only since day 27. A lot of higher organisms appeared in the last few days: first land plants and first land animals all appeared on day 28, mammals began to exist on day 29, and birds and flowering plants came into being on the last day. Modern humans, which are named homo sapiens in biology, appeared in the last 10 min of the last day. If we consider the recorded human history, it takes up only the last 30 s of the last day. The process that life gradually changes into different and often more complex or higher forms is called evolution. When studying the biology of a particular organism, it is important to realize that it is one leaf or branch on the huge tree of evolution. Comparison between related species is one major approach when investigating the unknown.

### 1.2.2 Cells

The basic component of all organisms is the cell. Many organisms are unicellular, which means one cell itself is an organism. However, for higher species like animals and plants, an organism can contain thousands of billions of cells.

Cells are of two major types: prokaryotic cells and eukaryotic cells. Eukaryotic cells are cells with real nucleus, while prokaryotic cells do not have nucleus. Living organisms are also categorized as two major groups: prokaryotes and eukaryotes according to whether their cells have nucleus. Prokaryotes are the earlier forms of life on the earth, which includes bacteria and archaea. All higher organisms are eukaryote, including unicellular organisms like yeasts and higher organisms like plants and animals. The bacteria *E. coli* is a widely studied prokaryote. Figure 1.1 shows the structure of an *E. coli* cell, as a representative of prokaryotic cells.

Eukaryotic cells have more complex structures, as shown in the example of a human plasma cell in Fig. 1.2. In eukaryotic cells, the key genetic materials, DNA, live in nucleus, in the form of chromatin or chromosomes. When a cell is

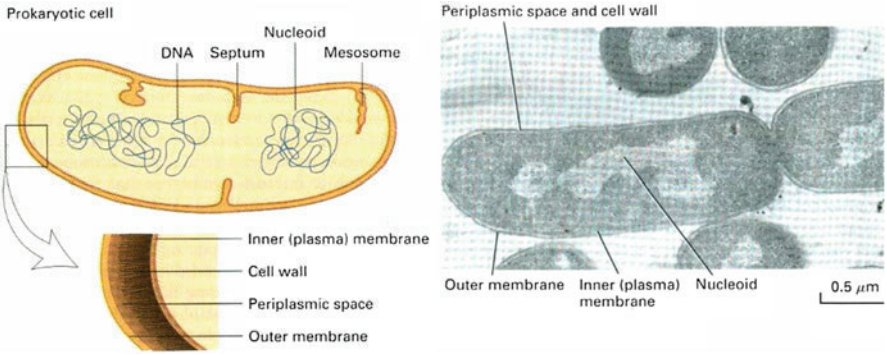


**Fig. 1.1** A prokaryotic cell

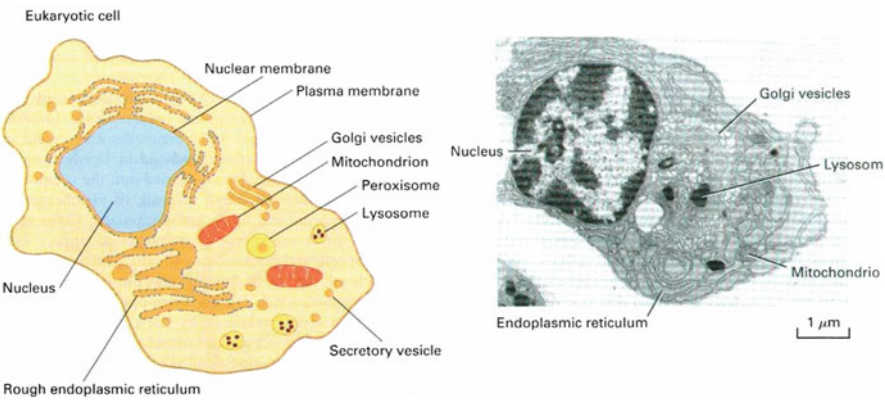


**Fig. 1.2** An eukaryotic cell

not dividing, the nuclear DNA and proteins are aggregated as chromatin, which is dispersed throughout the nucleus. The chromatin in a dividing cell is packed into dense bodies called chromosomes. Chromosomes are of two parts, called the P-arm and Q-arm, or the shorter arm and longer arm, separated by the centromere.

### *1.2.3 DNA and Chromosome*

DNA is the short name for deoxyribonucleic acid, which is the molecule that stores the major genetic information in cells. A nucleotide consists of three parts: a phosphate group, a pentose sugar (ribose sugar), and a base. The bases are of four types: adenine (A), guanine (G), cytosine (C), and thymine (T). A and G are purines with two fused rings. C and T are pyrimidines with one single ring. Besides DNA,

```
5'-A T T A C G G T A C C G T-3'
3'-T A A T G C C A T G G C A-5'
```

**Fig. 1.3** An example segment of a double-strand DNA sequence

there is another type of nucleotide called RNA or ribonucleic acid. For RNA, the bases are also of these four types except that the T is replaced by the uracil (U) in RNA.

DNA usually consists of two strands running in opposite directions. The backbone of each strand is a series of pentose and phosphate groups. Hydrogen bonds between purines and pyrimidines hold the two strands of DNA together, forming the famous double helix. In the hydrogen bonds, a base A always pairs with a base T on the other stand and a G always with a C. This mechanism is called base pairing. RNA is usually a single strand. When an RNA strand pairs with a DNA strand, the base-pairing rule becomes A-U, T-A, G-C, and C-G.

The ribose sugar is called pentose sugar because it contains five carbons, numbered as $1'$–$5'$, respectively. The definition of the direction of a DNA or RNA strand is also based on this numbering, so that the two ends of a DNA or RNA strand are called the $5'$ end and the $3'$ end. The series of bases along the strand is called the DNA or RNA sequence and can be viewed as character strings composed with the alphabet of "A," "C," "G," and "T" ("U" for RNA). We always read a sequence from the $5'$ end to the $3'$ end. On a DNA double helix, the two strands run oppositely. Figure 1.3 is an example of a segment of double-strand DNA sequence. Because of the DNA base-pairing rule, we only need to save one strand of the sequence.

DNA molecules have very complicated structures. A DNA molecule binds with histones to form a vast number of nucleosomes, which look like "beads" on DNA "string." Nucleosomes pack into a coil that twists into another larger coil and so forth, producing condensed supercoiled chromatin fibers. The coils fold to form loops, which coil further to form a chromosome. The length of all the DNA in a single human cell is about 2 m, but with the complicated packing, they fit into the nucleus with diameter around 5 $\mu$m.

### *1.2.4 The Central Dogma*

The central dogma in genetics describes the typical mechanism by which the information saved in DNA sequences fulfills its job: information coded in DNA sequence is passed on to a type of RNA called messenger RNA (mRNA). Information in mRNA is then passed on to proteins. The former step is called transcription, and latter step is called translation.

Transcription is governed by the rule of complementary base pairing between the DNA base and the transcribed RNA base. That is, an A in the DNA is transcribed to a U in the RNA, a T to an A, a G to a C, and vice versa.

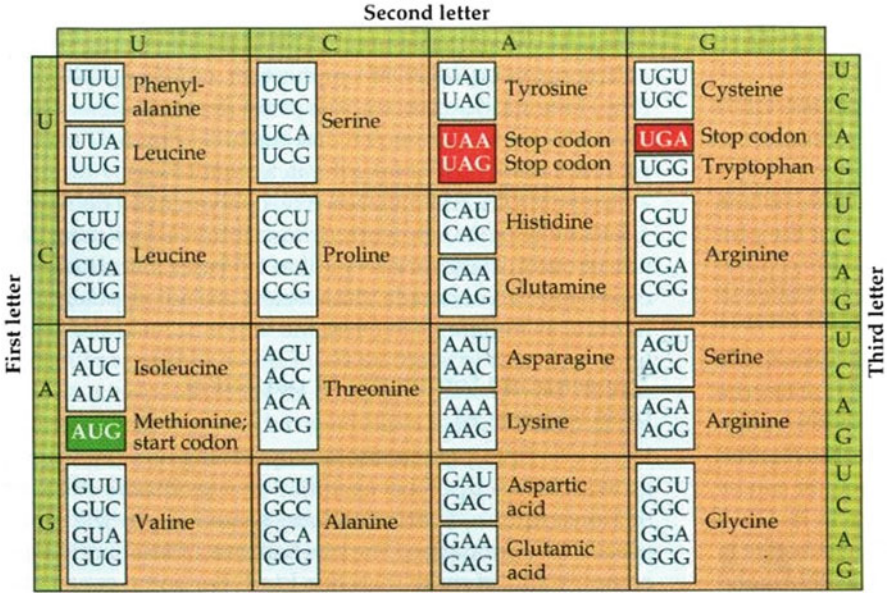


**Fig. 1.4** The genetic codes

Proteins are chains of amino acids. There are 20 types of standard amino acids used in lives. The procedure of translation converts the information from the language of nucleotides to the language of amino acids. The translation is done by a special dictionary: the genetic codes or codon. Figure 1.4 shows the codon table. Every three nucleotides code for one particular amino acid. The three nucleotides are called a triplet. Because three nucleotides can encode 64 unique items, there are redundancies in this coding scheme, as shown in Fig. 1.4. Many amino acids are coded by more than one codon. For the redundant codons, usually their first and second nucleotides are consistent, but some variation in the third nucleotide is tolerated. AUG is the start codon that starts a protein sequence, and there are three stop codons CAA, CAG, and UGA that stop the sequence.

Figure 1.5a illustrates the central dogma in prokaryotes. First, DNA double helix is opened and one strand of the double helix is used as a template to transcribe the mRNA. The mRNA is then translated to protein in ribosome with the help of tRNAs.

Figure 1.5b illustrates the central dogma in eukaryotes. There are several differences with the prokaryote case. In eukaryotic cells, DNAs live in the nucleus, where they are transcribed to mRNA similar to the prokaryote case. However, this mRNA is only the preliminary form of message RNA or pre-mRNA. Pre-mRNA is processed in several steps: parts are removed (called spicing), and ends of 150–200 As (called poly-A tail) are added. The processed mRNA is exported outside the nucleus to the cytoplasm, where it is translated to protein.

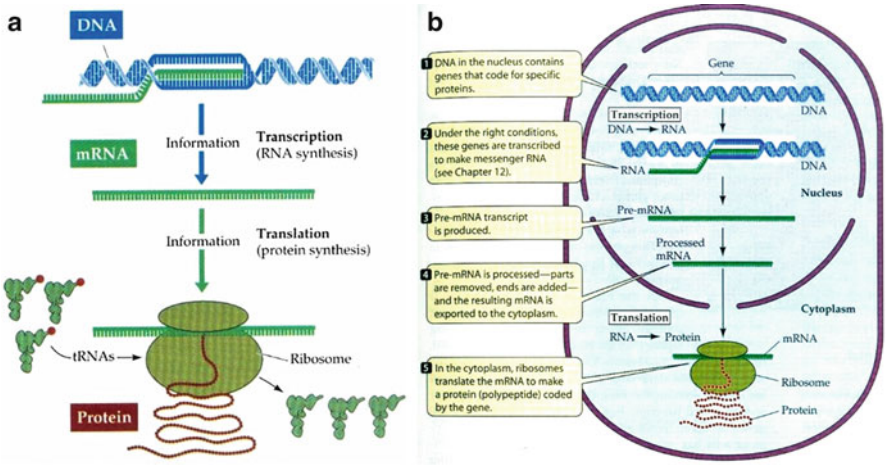


**Fig. 1.5** The central dogma

The procedure that genes are transcribed to mRNAs which are then translated to proteins is called the expression of genes. And the abundance of the mRNA molecules of a gene is usually called the expression value (level) of that gene, or simply the expression of the gene.

### *1.2.5 Genes and the Genome*

We believe that the Chinese translation "基因"of the term "gene" is one of the best scientific term ever translated. Besides that the pronunciation is very close to the English version, the literal meaning of the two characters is also very close to the definition of the term: basic elements. Genes are the basic genetic elements that, together with interaction with environment, are decisive for the phenotypes.

Armed with knowledge of central dogma and genetic code, people had long taken the concept of a gene as the fragments of the DNA sequence that finally produce some protein products. This is still true in many contexts today. More strictly, these DNA segments should be called protein-coding genes, as scientists have found that there are some or many other parts on the genome that do not involve in protein products but also play important genetic roles. Some people call them as nonprotein-coding genes or noncoding genes for short. One important type of noncoding genes is the so-called microRNAs or miRNAs. There are several other types of known noncoding genes and may be more unknown. In most current literature, people still use gene to refer to protein-coding genes and add attributes like "noncoding" and "miRNA" when referring to other types of genes. We also follow this convention in this chapter.

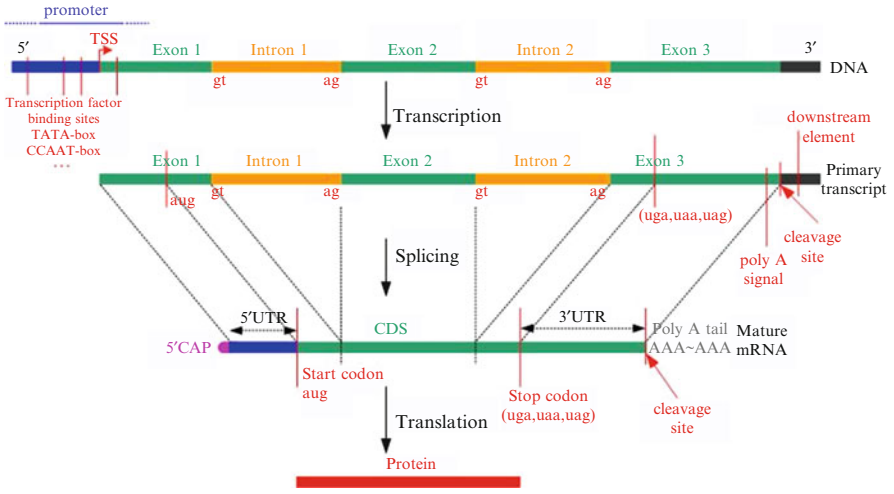


**Fig. 1.6** The structure of a gene

The length of a DNA segment is often counted by the number of nucleotides (nt) in the segment. Because DNAs usually stay as double helix, we can also use the number of base pairs (bp) as the measurement of the length. For convenience, people usually use "k" to represent "1,000." For example, 10 kb means that the sequence is of 10,000 bp. A protein-coding gene stretches from several hundreds of bp to several k bp in the DNA sequence. Figure 1.6 shows an example structure of a gene in high eukaryotes.

The site on the DNA sequence where a gene is started to be transcribed is called the transcription start site or TSS. The sequences around (especially the upstream) the TSS contain several elements that play important roles in the regulation of the transcription. These elements are called cis-elements. Transcription factors bind to such factors to start, enhance, or repress the transcription procedure. Therefore, sequences upstream the TSS are called promoters. Promoter is a loosely defined concept, and it can be divided into three parts: (1) a core promoter which is about 100 bp long around the TSS containing binding sites for RNA polymerase II (Pol II) and general transcription factors, (2) a proximal promoter of several hundred base pairs long containing primary specific regulatory elements located at the immediately upstream of the core promoter, and (3) a distal promoter up to thousands of base pairs long providing additional regulatory information. In eukaryotes, the preliminary transcript of a gene undergoes a processing step called splicing, during which some parts are cut off and remaining parts are joined. The remaining part is called exon, and the cut part is called intron. There can be multiple exons and introns in a gene. After introns are removed, the exons are connected to form the processed mRNA. Only the processed mRNAs are exported to the cytoplasm, and only parts of the mRNAs are translated to proteins. There may be

untranslated regions (UTRs) at both ends of the mRNA: one at the TSS end is called 5′-UTR, and the other at the tail end is called 3′-UTR. The parts of exons that are translated are called CDS or coding DNA sequences. Usually exons constitute only a small part in the sequence of a gene.

In higher eukaryotes, a single gene can have more than one exon-intron settings. Such genes will have multiple forms of protein products (called isoforms). One isoform may contain only parts of the exons, and the stretch of some exons may also differ among isoforms. This phenomenon is called alternative splicing. It is an important mechanism to increase the diversity of protein products without increasing the number of genes.

The term "genome" literally means the set of all genes of an organism. For prokaryotes and some low eukaryotes, majority of their genome is composed of protein-coding genes. However, as more and more knowledge about genes and DNA sequences in human and other high eukaryotes became available, people learned that protein-coding genes only take a small proportion of all the DNA sequences in the eukaryotic genome. Now people tend to use "genome" to refer all the DNA sequences of an organism or a cell. (The genomes of most cell types in an organism are the same.)

The human genome is arranged in 24 chromosomes, with the total length of about 3 billion base pairs ($3 \times 10^9$ bp). There are 22 autosomes (Chr.1.22) and 2 sex chromosomes (X and Y). The 22 autosomes are ordered by their lengths (with the exception that Chr.21 is slightly shorter than Chr.22): Chr.1 is the longest chromosome and Chr.21 is the shortest autosome. A normal human somatic cell contains 23 pairs of chromosomes: two copies of chromosomes 1.22 and two copies of X chromosome in females or one copy of X and one copy of Y in males. The largest human chromosome (Chr.1) has about 250 million bp, and the smallest human chromosome (Chr.Y) has about 50 million bp.

There are about 20,000–25,000 protein-coding genes in the human genome, spanning about 1/3 of the genome. The average human gene consists of some 3,000 base pairs, but sizes vary greatly, from several hundred to several million base pairs. The protein-coding part only takes about 1.5–2 % of the whole genome. Besides these regions, there are regulatory sequences like promoters, intronic sequences, and intergenic (between-gene) regions. Recent high-throughput transcriptomic (the study of all RNA transcripts) study revealed that more than half of the human genomes are transcribed, although only a very small part of them are processed to mature mRNAs. Among the transcripts are the well-known microRNAs and some other types of noncoding RNAs. The functional roles played by majority of the transcripts are still largely unknown. There are many repetitive sequences in the genome, and they have not been observed to have direct functions.

Human is regarded as the most advanced form of life on the earth, but the human genome is not the largest. Bacteria like *E. coli* has genomes of several million bp, yeast has about 15 million bp, Drosophila (fruit fly) has about 3 million bp, and some plants can have genomes as large as 100 billion bp. The number of genes in a genome is also not directly correlated with the complexity of the organism's

complexity. The unicellular organism yeast has about 6,000 genes, fruit fly has about 15,000 genes, and the rice that we eat everyday has about 40,000 genes. In lower species, protein-coding genes are more densely distributed on the genome. The human genome also has a much greater portion (50 %) of repeat sequences than the lower organisms like the worm (7 %) and the fly (3 %).

### *1.2.6 Measurements Along the Central Dogma*

For many years, molecular biology can only study one or a small number of objects (genes, mRNAs, or proteins) at a time. This picture was changed since the development of a series of high-throughput technologies. They are called high throughput because they can obtain measurement of thousands of objects in one experiment in a short time. The emergence of massive genomic and proteomic data generated with these high-throughput technologies was actually a major motivation that promotes the birth and development of bioinformatics as a scientific discipline. In some sense, what bioinformatics does is manipulating and analyzing massive biological data and aiding scientific reasoning based on such data. It is therefore crucial to have the basic understanding of how the data are generated and what the data are for.

### *1.2.7 DNA Sequencing*

The sequencing reaction is a key technique that enables the completion of sequencing the human genome. Figure 1.7 illustrates the principle of the widely used Sanger sequencing technique.

The technique is based on the complementary base-pairing property of DNA. When a single-strand DNA fragment is isolated and places with primers, DNA polymerase, and the four types of deoxyribonucleoside triphosphate (dNTP), a new DNA strand complementary to the existing one will be synthesized. In the DNA sequencing reaction, dideoxyribonucleoside triphosphate (ddNTP) is added besides the above components, and the four types of ddNTPs are bound to four different fluorescent dyes. The synthesis of a new strand will stop when a ddNTP instead of a dNTP is added. Therefore, with abundant template single-strand DNA fragments, we'll be able to get a set of complementary DNA segments of all different lengths, each one stopped by a colored ddNTP. Under electrophoresis, these segments of different lengths will run at different speeds, with the shortest segments running the fastest and the longest segments running the slowest. By scanning the color of all segments ordered by their length, we'll be able to read the nucleotide at each position of the complementary sequence and therefore read the original template sequence. This technique is implemented in the first generation of sequencing machines.