Modeling and Optimization in Science and Technologies

Gautam B. Singh

Fundamentals of Bioinformatics and Computational Biology

Methods and Exercises in MATLAB

Springer

Modeling and Optimization in Science and Technologies

Volume 6

About this Series

The book series *Modeling and Optimization in Science and Technologies (MOST)* publishes basic principles as well as novel theories and methods in the fast-evolving field of modeling and optimization. Topics of interest include, but are not limited to: methods for analysis, design and control of complex systems, networks and machines; methods for analysis, visualization and management of large data sets; use of supercomputers for modeling complex systems; digital signal processing; molecular modeling; and tools and software solutions for different scientific and technological purposes. Special emphasis is given to publications discussing novel theories and practical solutions that, by overcoming the limitations of traditional methods, may successfully address modern scientific challenges, thus promoting scientific and technological progress. The series publishes monographs, contributed volumes and conference proceedings, as well as advanced textbooks. The main targets of the series are graduate students, researchers and professionals working at the forefront of their fields.

More information about this series at http://www.springer.com/series/10577

Gautam B. Singh

Fundamentals of Bioinformatics and Computational Biology

Methods and Exercises in MATLAB

Springer

Gautam B. Singh
Department of Computer Science
and Engineering
Oakland University
Rochester, Michigan
USA

ISSN 2196-7326 ISSN 2196-7334 (electronic)
ISBN 978-3-319-11402-6 ISBN 978-3-319-11403-3 (eBook)
DOI 10.1007/978-3-319-11403-3

Library of Congress Control Number: 2014949497

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

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

To my family

Preface

The integration of computers in life sciences has been growing for the last two decades. While the first release of GenBank contained a mere half a million DNA sequence bases in 1982, the current release of GenBank has exceeded 100 giga bases of data. With data comes computational challenges for analysis, interpretation, visualization and integration of information. That in a nutshell is the reason to familiarize undergraduate students in computer science and engineering with the nature and use of biological data and thus become prepared to meet the demands of high tech careers in the twenty-first century.

The intended audience of this textbook are students in computer science, engineering and information technology at the undergraduate or lower graduate level. The material is primarily presented in a simplified manner and extensive details are left out. However, pointers to appropriate references should guide those who are interested in exploring specific topics in greater detail.

Topics in this textbook are organized into three parts. **Part I** of this book provides some background to the field of bioinformatics and an introduction to molecular biology and genetics. A survey of biological databases is also included. The material in this part is considered to be fairly fundamental and should be covered in all courses, graduate and undergraduate.

Part II of the book covers methodologies for retrieving information from biological databases and covers simple boolean searches, sequence alignment algorithms, protein alignment, scoring matrices, alignment tools and biolinguistic methods. Undergraduate students should cover basic retrieval techniques and advanced topics such as PAM and BLOSUM may be included based on the amount of time available and level of preparation of the students.

Part III of the book covers the topics related to sequence analysis and covers algorithms for finding patterns and detecting genes.

Part IV focuses on topics in phylogenetics and systems biology and covers the algorithms for distance, character and probabilistic methods for inferring

phylogeny. Also described are some key algorithms for analyzing micro-chip data.

The book is an offshoot of our project aimed at creating bioinformatics educational resources for undergraduates in computer science and engineering. This project is sponsored by the National Science Foundation, USA. Additional details for the project and bioinformatics educational resources are available from `http://bioflow.secs.oakland.edu`.

The author would like to acknowledge the efforts by students who participated in creating resources for the NSF sponsored bioinformatics project: Kenneth DeMonn, Nirmala Venkatraman, David Poe, Guy Lima, Kellie McGowan and Ashwin Kottam. Their work influenced the content and the presentation in this text.

For the *BioFlow* project we are also analyzing student learning styles in collaboration with Professor Christine Hansen, Department of Psychology, Oakland University. The results from obtained from the student assessment studies were very valuable in providing insight into methods that made this text more comprehensible for computer science and engineering undergraduates.

Rochester, MI
USA

Gautam B. Singh
June 2014

Contents

Part I
Background

Chapter 1
Introduction to Bioinformatics

Sequencing of biomolecules began in 1951 when Sanger and Tuppy deduced the thirty residue protein from the insulin B-chain. It was only after 25 years that *real* DNA sequencing methodologies were developed by Maxim & Gilbert and by Sanger *et al.* Today, we are sequencing tens of millions of bases of DNA sequences a year and undertaking the sequencing of genomes from whole organisms. During these times, the sequence databases have continued their exponential growth rate. The computational research in bioinformatics aims at enhancing the retrieval, analysis and interpretation of information that is embedded within the biological databases containing the DNA and protein sequences.

In a manner similar to the transformation of physics and chemistry, the study of biology has been undergoing a transformation since the 1990s. This transformation is aimed at the integration of computational sciences and information technology into the study of life sciences. This transformation has been driven by the computational requirements of genomic research. The experiment-rich field of biology has been generating data at an exponential rate, while there is a dearth of tools for information analysis and visualization. Furthermore, the challenges faced in bioinformatics stem largely from the fact that the languages and techniques utilized in the field of molecular biology are descriptive and experimental in nature, while the methodology in computer science and mathematics is generally based on analytical and precise formulations.

1.1 What Is Bioinformatics?

Bioinformatics is an emerging discipline that draws upon the strengths of computer sciences, mathematics, and information technology to determine and analyze genetic information. Bioinformatics leverage synergies between computational and biological sciences. Although the field of bioinformatics originally aimed at extracting information embedded within the 3 billion

G.B. Singh, *Fundamentals of Bioinformatics and Computational Biology,*
Modeling and Optimization in Science and Technologies 6, DOI: 10.1007/978-3-319-11403-3_1

bases of human DNA, the field has evolved to realize its capabilities for studying *information content* and *information flow* in biological systems and processes in general.

Earlier bioinformatics research aimed at mapping individual genomes and calculating differences to estimate population diversity. The study of bioinformatics now encompasses genomes from other species besides humans. For example, other genomes of interest are those of microbes, plants and fungi. An analysis of plant genome databases leads to advances in the agricultural arena by helping produce plants that are resistant to diseases and have higher yields. Understanding microbial genomics facilitates the development of new therapies for combating infectious diseases. Furthermore, computational analysis of fungal and microbial genomes serve as valuable smaller scale model organisms that lead to understanding functional genomics and improving technological development for applications at a larger "human" scale.

Originally, in the mid-1980s, the field of bioinformatics was defined as the *subject of genetic data collection, analysis and dissemination.* Bioinformatics has come a long way since then and now aims at complex mathematical modeling and simulation as well. For example, bioinformaticians now aim at developing computational models for differential regulation of gene expression that occurs *in-vivo* in a tissue-specific manner. Thus, the field of biology now turns towards mathematics and computation sciences to help make further advancements for understanding its core theoretical principles.

Such a dependence by biological science on the algorithms and formulations provided by analytical sciences was borne out of the necessity to analyze and make interpretations from the large volume of data generated by the Human Genome Project. Biologists look towards computational sciences to help bridge the gap between experimental observations and gaining an understanding of how living systems and processes perform their functions. Such a creation of novel biological *knowledge* in the case of diseases, for example, could lead to our predictive inference on the prognosis and preventive therapeutic treatment based on our understanding of diagnostics data through integrative bioinformatics models.

From an information technology perspective, therefore, bioinformatics may be defined as a scientific discipline encompassing acquisition, storage, processing, analysis, interpretation and visualization of biological information. It encompasses frameworks, theories, algorithms, techniques and tools from mathematics, computer science and biology with the aim of understanding the significance of a variety of biological data.

1.2 The Human Genome Project

Leveraging the technological advancements in molecular biology and genetics in the mid-1980s, the Human Genome Project was initiated with the goal of enabling progress and benefits in biomedicine. The two main goals of the

Human Genome Project were to identify all the approximately 25–30,000 human genes, and to determine the sequences of the 3 billion chemical base pairs that make up human DNA. Completed in 2003, the Human Genome Project (HGP) was a 13-year project coordinated by the U.S. Department of Energy and the National Institutes of Health. Coincidentally, the completion of the human DNA sequence in the spring of 2003 also marked the $50^{th.}$ anniversary of Watson and Crick's discovery of fundamental structure of DNA.

The HGP was truly an international effort, although significant advancements of its goals were accomplished in the United States. Other contributors included the Wellcome Trust (U.K.) who was major partner during the early years of its inception in the 1990s. Contributions to its advancements are also attributed to the commitments by countries such as Japan, France, Germany, China, and others. Being an international effort, the HGP was supported by the technological advancements in database and information retrieval with networking technologies that supported international collaborations.

The completion of the Human Genome Project was celebrated in April 2003 and also marks the completion of the sequencing of the human genome. However, this event also marked the beginning of the *post-genome informatics* era where scientists are collaboratively engaged in understanding the biological function in an integrative manner. As an illustration, consider the initial analysis of the draft human genome sequence that was published in 2001 by the International Human Genome Sequencing Consortium which estimated that the human genome contains only about 20,000 to 30,000 protein-coding genes, an estimate that was significantly lower than previous estimates of around 100,000.

This lower estimate came as a shock to many scientists because counting genes was viewed as a way of quantifying genetic complexity. With around 25,000, the human gene count would be only about 25% greater than that of the simple roundworm *C. elegans* at about 20,000 genes. Thus, science today is embarking upon unraveling the complexity in coordination and regulation of genetic networks in contrast to the number of genes a being the determining factor in an organism's complexity.

Thus, an understanding of biological function is not possible using the sequence information alone. Complete understanding of life's complex processes often necessitates a convergence of computational modeling and experimental sciences. This, incidentally is also the recipe for effective bioinformatics research.

1.3 Genome Data Statistics

The National Center for Biotechnology Information, or the NCBI, is part of the National Library of Medicine. The National Library of Medicine is a component of the National Institutes of Health and the U.S. Department of Health and Human Services. The NCBI was established in 1988 as a national

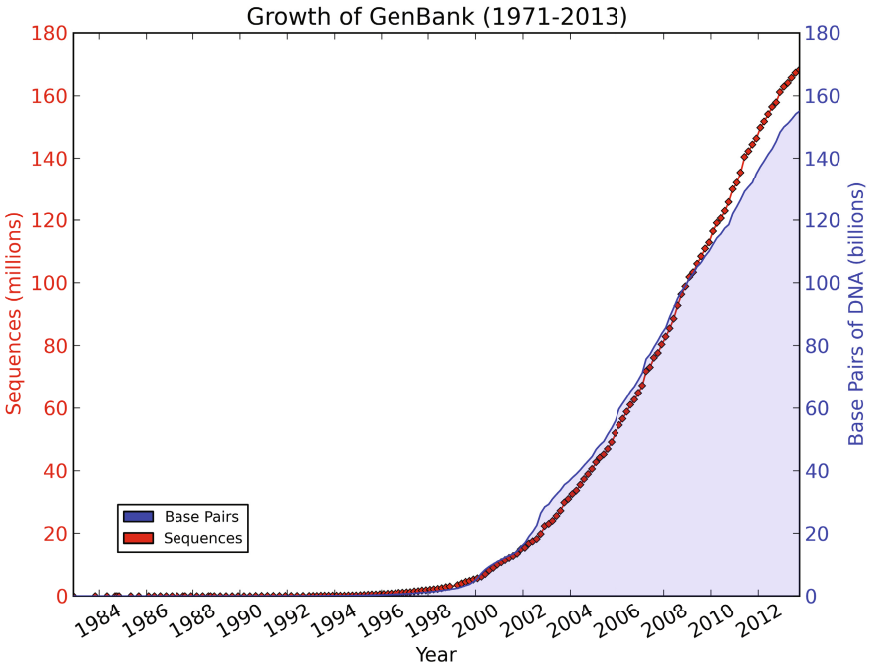

Fig. 1.1 The growth of the nucleotide database GENBANK since its inception till March 2006. The size of the database as well as the number of sequences deposited has been undergoing an exponential growth ever since GENBANK's inception. *Source of data:* `http://www.ncbi.nlm.nih.gov/GenBank/genbankstats.html`

resource for molecular biology information. NCBI creates public databases and develops software tools for analyzing genome data.

GenBank is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences. NCBI is host to the GenBank nucleotide sequence database. Submitters to GenBank contribute over 3-4 million new DNA sequences per month to the database. As of 2013, there are over 150 billion bases in approximately 170 million sequence records in the various GenBank dataset divisions. The data represents both individual genes and partial and complete genomes of over 165,000 organisms. GenBank provides sequences from single genes from organisms as diverse as humans, elephants, earthworms, fruit-flies, apple trees, and bacteria as well as the sequences for organisms' complete genomes.

As illustrated in Fig. 1.1 the total database size and the number of sequences in the nucleotide database GenBank has been growing exponentially. This growth rate has continued despite the completion of the genome project and is fueled by scientists' interest in continuing to sequence other organisms and plants and mankind's interest in biological diversity.

1.4 Applications of Bioinformatics

There are many applications of the burgeoning field of bioinformatics. Bioinformatics is not limited in its applications and in general can be applied to *any* computational inquiry for solving biological problems. Some of the common applications of bioinformatics are listed below. However, it must be emphasized that the list below is by no means exhaustive. At its core of bioinformatics is an application of computational and mathematical problem solving algorithms to further the understanding of a biological system. Thus, the applications of bioinformatics is really an open ended list.

- **Pattern Discovery:** The need to discover patterns stems from the observation that whenever nature finds a mechanism (the evolution of the eye for example) which bestows a differential fitness to an organism, the mechanism is often reused. Such a reuse of successful recipes also occurs at the molecular level implying that biological sequences belonging to distant species share common patterns in their genetic makeup. Through computational analysis and modeling, bioinformatics algorithms help in discovery and functional interpretation of these biological patterns.
- **Protein Folding:** After a gene has been transcribed and translated, the linear polypeptide chain folds into a three dimensional protein within a matter of seconds or minutes. The protein can only function after it has acquired a three dimensional structure as its interactions with other molecules is largely governed by the protein's shape. Even after a concerted effort over the past few decades, computational algorithms for predicting the protein structure from a protein sequence continues to be an unsolved problem. A solution to this problem will enable scientists to accurately model biochemical pathways and lead to effective drug design *in-silico* without the need for extensive experimentation.
- **Alignment and Homology:** Sequence alignment is a methodology for arranging biological sequences to identify regions of similarity between them. The extent of similarity, or homology, between the DNA from a variety of organism may be used to determine evolutionary relationships and degrees of divergence between them. The type of ancestry would let us establish if a certain structures observed evolved from some structure in a common ancestor (such as arms in humans and wings on bats) and also infer function of an anonymous sequence based on the known function of a homologous sequence. There are a number of computational challenges in developing newer and integrative homology algorithms that can infer homology using data from sequence and structure.
- **Orthologs and Paralogs:** Homologous features share an evolutionary history. Orthologs and paralogs are both homologs, but have more specific meanings: Orthologs have diverged because of a speciation event such as a gene like *Antp* in fruit fly *D. melanogaster* diverged from the gene *mab-5* in roundworm *C. elegans* – that is these two genes evolved from the same gene in their common ancestor. In contrast, paralogs have diverged

because of a gene duplication event in an ancient ancestor as is probably the case with *mab-5* and *lin-39* in *C. elegans*. Also consider the example of the gamma-globin genes in the Anthropoidea which duplicated before the new world monkeys diverged from the lineage to humans and apes. That would make gamma-1-globin and gamma-2-globin paralogous. However, the gamma-1 in humans is orthologous to gamma-1 in chimpanzees. Thus, it is sometimes difficult to distinguish orthologs from paralogs. Sophisticated computational models are needed to effectively develop taxonomies and gene trees and discover novel instances of convergent evolution.

- **Information Retrieval and Data Mining from Biological Databases:** The explosive growth of sequence and biological information has created new challenges for data representation, access, and analysis. With the size and the need for access to this data continually increasing, maintaining database performance and availability is a challenging task. In-silico biology, is a term that life science companies use to refer to the computational tools that translate raw experimental data into workable models or simulations to identify targets for drug development. Bioinformatics tools for processing the overwhelming amount of data gathered through genome research, such as the Human Genome Project, are essential to fuel the *in-silico* life science research. Genome database searching entails developing computational tools for identification of protein-encoding regions of a genome and assign functions to these genes on the basis of sequence similarity with other genes of known function.
- **Data Integration from Multiple Modalities:** Researchers in molecular biology and medicine rely heavily on progress in data management and quality assurance as the necessary underpinnings for effectuating progress. Advancements in life sciences, and particularly in areas like drug development, systems biology, or personalized medicine, are dependent on integration of data from myriad experiments, longitudinal studies, and levels of detail. Numerous unsolved problems, for example the integration of data from proteomics mass spectroscopy and gene sequences, requires fulfilling a vast information gap-filling through data rationalization. These issues are moving towards the forefront as advancements in instrumentation is rapidly generating larger quantities of data in a high throughput manner. The next big challenge is expected to be the integration of genomics, proteomics and individualized medical data routinely collected by hospitals.
- **Analysis of Biological Sequences and Pattern Discovery:** A good understanding of the probabilistic nature of biological sequences is the key to statistical modeling of biological sequences. Genome sequence data analysis leads to the design of novel tools for effective prediction of biological function. Sequence analysis research focuses on finding patterns in biological sequences and associating these patterns with functions necessary for pathways that support life forms. Some of the pattern detection encompasses looking for short range patterns such as binding and initiation sites, while other computational techniques seek to model long range

patterns such as genes, locus control regions, matrix attachment regions, and helitrons where a consensus sequence or even a consensus structure of the model is not yet known. Inductive learning approaches are needed to analyze known data and induce mathematical models that learn as new discoveries are made.

- **Micro-arrays and Differential Gene Expression:** The thousands of genes and their products expressed in a given living organism function in a coordinated and complex manner. The discovery of genes has been followed by various techniques for detecting genes based on their expression levels as gene expression is correlated with the tissue type. For example, the genes that are expressed in the liver are not the same set of genes expressed in kidneys, with the exception of certain housekeeping genes required for processes common to all cells that are expressed in all cells. Methods of identifying differential expression in genes have been developed to establish the levels of gene expression in various tissue types. This can be particularly useful in cancer studies, where mutations that can amplify or turn off gene expression occur in malignant samples. These days the gene chip or microarray technology is greatly accelerated the speed of performing a differential gene expression analysis. With gene chips one can monitor the whole genome using a single chip allowing researchers to better understand the interactions among thousands of genes simultaneously. Among the many applications of micro-arrays or gene chips are gene discovery, disease diagnosis, drug discovery, and toxicology research. Gene expression studies using gene chips draw upon the advances in computational analysis and statistics, image processing and data management.
- **Gene Regulatory Networks:** Gene regulatory networks (GRNs) are the on-off switches that act like rheostats and control the level of expression for each gene by controlling whether and to what level the gene will be transcribed into RNA. These networks are a collection of DNA segments that interact with each other and other substances in the cell and govern the overall rate at which the network is transcribed into mRNA. In a graph theoretic model, the nodes of a GRN are proteins and edges represent individual molecular reactions or protein interactions. Edges may have arrowheads and thus be inductive where the increase in the concentration of one leads to the interaction of the other, or inhibitory (with a circle) where the increase in one leads to the decrease in the other. GRNs thus capture the chemical dynamics of the cell. Construction and simulation of GRNs offer many challenges in mathematical modeling, simulation and visualization.
- **Metabolic Pathway Models:** Metabolic network reconstruction and simulation enables us to gain insight into molecular mechanisms that relate the genome to molecular physiology. The metabolic pathway breaks down a metabolic cycle such as glycolysis, Krebs cycle, or pentose phosphate pathway into their respective reactions and enzymes and analyzes their interactions. Enzymes and genes are correlated by searching genomic

databases. Continual validation of metabolic pathways is needed to keep the pathway database consistent.

Further Readings

1. Searls, D.B.: An online bioinformatics curriculum. PLoS Comput. Biol. 8(9), e1002632 (2012)
2. Maojo, V., Kulikowski, C.A.: Victor Maojo and Casimir A Kulikowski. Bioinformatics and medical informatics: collaborations on the road to genomic medicine? J. Am. Med. Inform. Assoc. 10(6), 515–522 (2003)
3. Li, J., Doyle, M.A., Saeed, I., Wong, S.Q., Mar, V., Goode, D.L., Caramia, F., Doig, K., Ryland, G.L., Thompson, E.R., Hunter, S.M., Halgamuge, S.K., Ellul, J., Dobrovic, A., Campbell, I.G., Papenfuss, A.T., McArthur, G.A., Tothill, R.W.: Bioinformatics pipelines for targeted resequencing and whole-exome sequencing of human and mouse genomes: a virtual appliance approach for instant deployment. PLoS One 9(4), e95217 (2014)
4. Hartwell, L.H., Hopfield, J.J., Leibler, S., Murray, A.W.: From molecular to modular cell biology. Nature 402(6761 suppl.), C47–C52 (1999)
5. Csete, M.E., Doyle, J.C.: Reverse engineering of biological complexity. Science 295(5560), 1664–1669 (2002)
6. Ouzounis, C.A.: Rise and demise of bioinformatics? promise and progress. PLoS Comput. Biol. 8(4), e1002487 (2012)
7. Dymond, J.S., Scheifele, L.Z., Richardson, S., Lee, P., Chandrasegaran, S., Bader, J.S., Boeke, J.D.: Teaching synthetic biology, bioinformatics and engineering to undergraduates: the interdisciplinary build-a-genome course. Genetics 181(1), 13–21 (2009)

Chapter 2
Introduction to Molecular Biology

Molecular biology overlaps the fields of biology and chemistry and mainly aims at developing an understanding of the interactions between the various systems of a cell, including the interrelationship of DNA, RNA and protein synthesis as well as with uncovering the manner in which these interactions are regulated.

Researchers in molecular biology use specific techniques native to molecular biology. However, there is considerable diffusion of ideas from other disciplines such as genetics and biochemistry; there does not seem to be a clear boundary that delineates these fields anymore as the researchers borrow techniques and methodologies from all these related fields.

Biochemistry is defined to be the study of the chemical substances and vital processes occurring in living organisms. Genetics concerns itself with the effect of genetic differences on organisms which often is associated with the absence of a genes as in the study of "mutants." Mutants are organisms which lack one or more functional genes with respect to the so-called "wild type."

Molecular biology synthesizes the above viewpoints by studying the molecular underpinnings of the processes related to genetics. Specifically, those related to the replication, transcription and translation of the genetic material – the so called central dogma of molecular biology discussed later in this chapter.

2.1 Cell Structure

Every cell typically contains hundreds of different kinds of macromolecules that function together to generate the behavior of the cell. Computer graphic of a typical animal cell and its contents or the organelles is shown in Fig. 2.1.

The big and round structure is the nucleus, which carries the cell's genes in the form of DNA and controls cellular activities via genes. The nucleolus is located within the nucleus and is the site for ribosome synthesis. The oval

G.B. Singh, *Fundamentals of Bioinformatics and Computational Biology,*
Modeling and Optimization in Science and Technologies 6, DOI: 10.1007/978-3-319-11403-3_2

bodies are mitochondria which are responsible for the production of ATP through the oxidation of carbohydrates and provide the cell with energy.

The folded, dotted structures are rough endoplasmic reticulum where the folds of membrane carry ribosomes that synthesize proteins. The smooth endoplasmic reticulum is involved in the lipid synthesis. The round bodies containing small particles are vesicles are the lysosomes or peroxisome. Lysosome contains hydrolytic enzymes for intracellular ingestion, while peroxisome is responsible for hydrogen peroxide synthesis and degradation and expelling some of the matter through the cell's outer plasma membrane. The vesicles are made by the golgi apparatus which is the folded structure is the packaging center and the site for manufacture of carbohydrates.

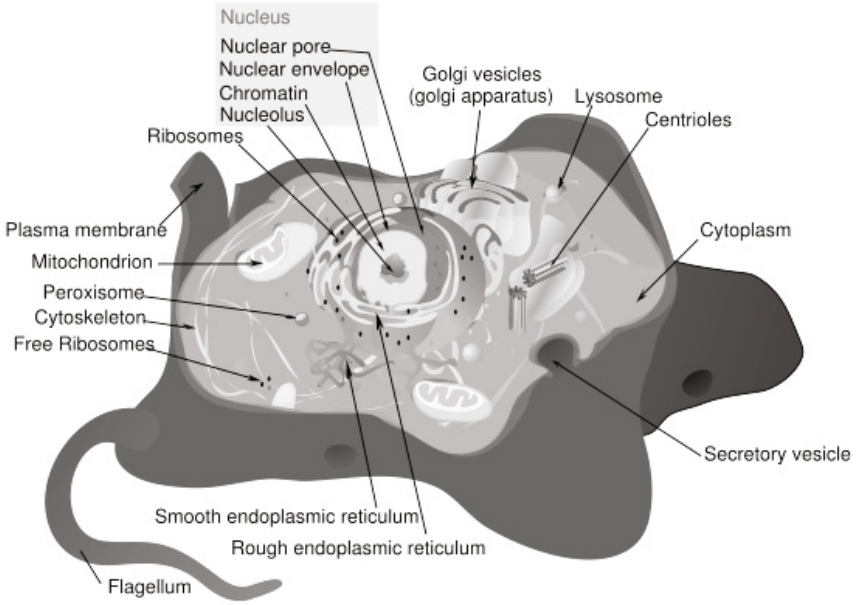

Fig. 2.1 The structure of an animal cell

Most proteins are synthesized by ribosomes in the cytoplasm. This process is also known as protein biosynthesis or simply protein translation. Some proteins, such as those to be incorporated in membranes (membrane proteins), are transported into the ER during synthesis and are also processed further in the golgi apparatus.

2.1.1 Genome

The term genome was coined by Hans Winkler, a Professor of Botany at the University of Hamburg, in 1920 to collectively refer to the complete genetic material of an organism. The genome contains an organism hereditary information encoded in the DNA and packaged into the chromosome. The genome encompasses both the genes and the non-coding sequences of the DNA. More precisely, the genome of an organism is a complete DNA sequence of one set of chromosomes; for example, one of the two sets that a diploid individual carries in every somatic cell.

The term is sometimes qualified in a manner to refer to the complete genetic material is in *nuclear genome* might be used to refer to the complete set of nuclear DNA, can be applied to organelles such as *mitochondria genome* or *chloroplast genome* to refer to the complete contents of the DNA found in mitochondria or chloroplast respectively.

Every somatic cell contains two copies of each chromosome where one is inherited from each parent. Each pair of chromosome is often referred to as a homologous pair where both the chromosomes contain a gene for the same trait at exactly the same loci. A given location in the homologous chromosome pair thus contains a gene pair representing the two *alleles* for that gene with each allele originating from the two parents. For example, one allele in a fruit fly may code for the flies to have long wings, while the other allele may code for flies to be wingless. Alternatively, both the alleles may code for long wings, or may code for no wings at all. Where the two alleles are different the genotype or the genetic makeup is referred to as *heterozygous* while the genotype is referred to as *homozygous* when the two alleles are the same.

The phenotype is the actual expression of the genotype. Purple flowers, brown eyes, or wingless fruit flies are all examples of phenotypes. In a heterozygous individual, the phenotype of the organism is determined by which of the allele is dominant. In the example, whether a heterozygous fruit fly has long wings or no wings depends upon which of the two alleles, the one for long wings or the one for no wings, is dominant. The trait of the dominant allele is expressed, while that of the recessive allele is suppressed. For a homozygote, the expression or the phenotype is the same as that coded by the two identical alleles.

The biological information contained in a genome is encoded in its deoxyribonucleic acid (DNA). The DNA is a macromolecule that contains discrete units of protein coding segments called the genes. Current estimates place the number of genes in the humans to be around 25,000 encoded in on 23 chromosomes. The total length of the human nuclear genome is 3×10^9 base pairs. Note that the length of DNA only considers the of one of the homologous chromosome pairs.

The chromatin, the constituent DNA in a chromosome, becomes observable under a microscope when it is packed in the shape shown in Fig. 2.2 during the metaphase of a cell division or mitosis. While the diameter of a strand

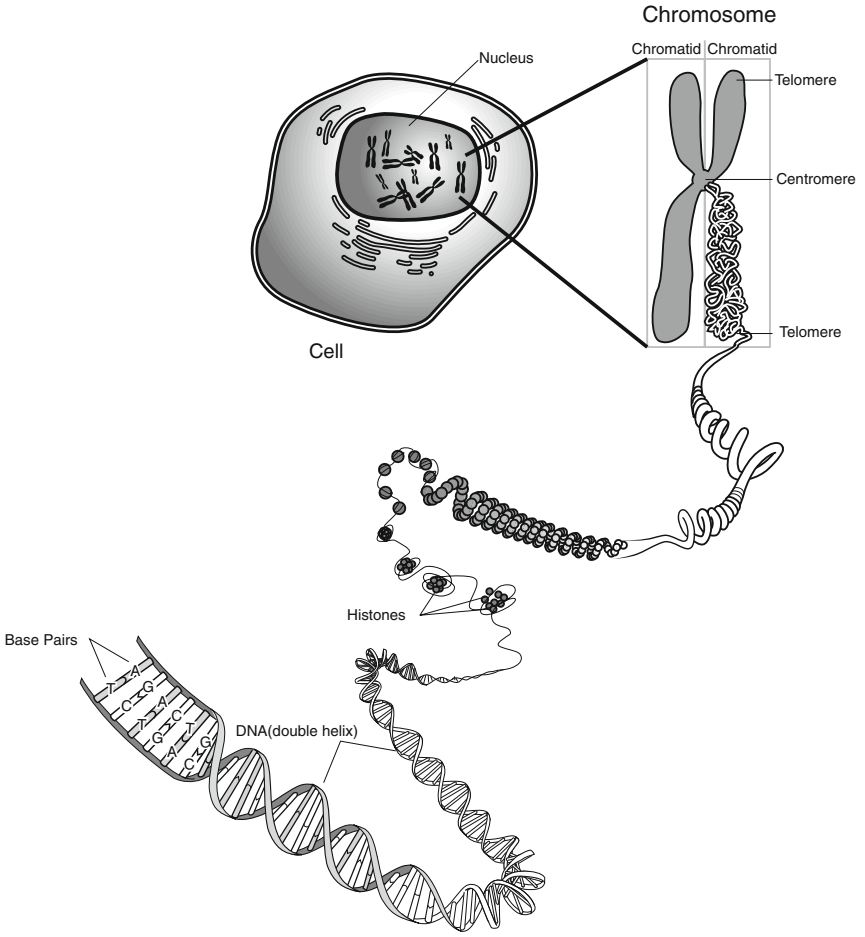

Fig. 2.2 Each chromosome contains a single molecule of DNA organized into several orders of packaging. The packaging of a chromosomes in metaphase is shown where the compact packaging causing length of the chromosome to be about 0.0001 times the length of the DNA molecule. *Reprinted from National Human Genome Research Institute's Educational Resources.*

of DNA is only 2 nm, several levels of packaging facilitated through the packaging proteins called the histones, the chromosomes themselves have a diameter of 1400 nm or 1.4 microns. In order for a gene to be expressed, the loci of the genome containing the gene must be in an open conformation and not tightly packaged as in a metaphase chromosome.

2.1.2 DNA: Deoxyribonucleic Acid

The DNA molecules consist of two complementary chains that are twisted together to form a double helix. The DNA molecule is comprised of four nucleotide bases belonging to two classes. These four bases are adenine (A), guanine (G), cytosine (C) and Thymine (T). The two classes to which these bases belong are purine and pyrimidine. The bases A and G belong to the class purine while C and T belong to the class pyrimidine. The DNA double helical structure is formed by the hydrogen bonding between purines and pyrimidines. The DNA is often thought of as a spiral staircase where the sides of this ladder consist of deoxyribose residues linked together with phosphate bonds and the "rungs" of the ladder are made up of a hydrogen bonded purine and pyrimidine pair.

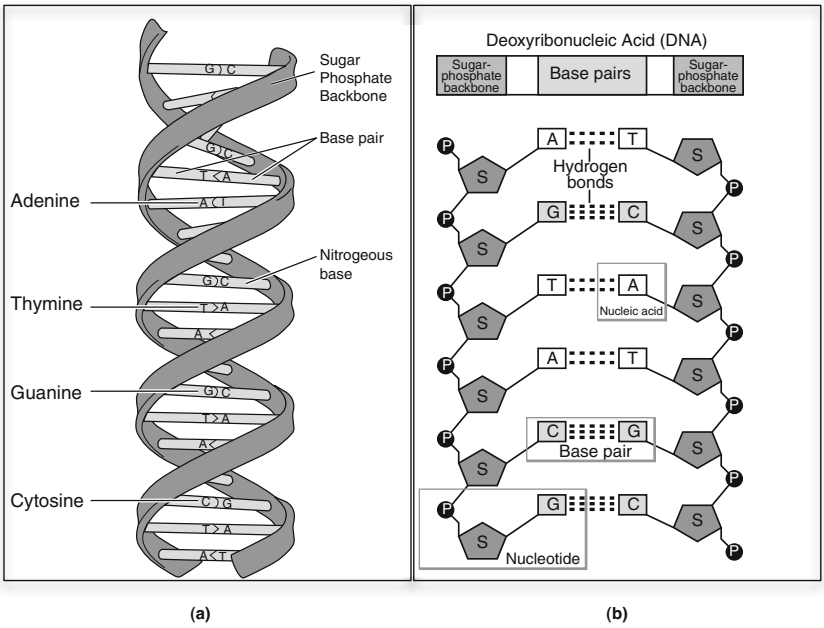

Fig. 2.3 DNA molecule and nucleotide bases. (a) Cytosine (G) always pairs with Guanine (G) while Adenine (A) always pairs with Thymine (T). (b) Triple hydrogen bonds are formed between every C:G pairings while two hydrogen bond exists between each A:T pairs. *Reprinted from National Human Genome Research Institute's Educational Resources.*

Fig. 2.3 shows the structure of a DNA molecule. The DNA is a double stranded molecule with base C pairing with the base G and base A pairing with T. The pairing is not chemical in nature but rather mediated with

hydrogen bonding. This enables the DNA to adopt a single stranded conformation with relative ease as is needed during the cellular processes leading to replication and transcription. The bonding between bases C and G is stronger as characterized by a triple hydrogen bond between this pair. The bases A and T on the other hand are paired using a double hydrogen bond. Generally, due to the C:G bond being stronger, there is ample evidence to suggest that genes are located in the CG rich areas of the genome.

The DNA is read from the direction of 5' (*five-prime*) to 3' (*three-prime*). These labels are indicative of the free carbon atom on the sugar phosphate backbone of the DNA molecule. The 5' carbon on the reverse, or complementary strad is located directly opposite from the nucleotide base attached to the 3' end of the forward strand and vice versa. Further, as the DNA occurs in a double stranded conformation, the length along the molecule is measured in units of *base pairs* or **bp**.

Example 2.1

Consider a 10-bp DNA sequence `TAAGCCTGTA`. WIthout more information we will assume that the sequence provided is for a 5' to 3' read for forward strand. This corresponds to a left to right read. The forward and reverse strands would be shown as follows:

```
5'     forward strand      3'
T  A  A  G  C  C  T  G  T  A
|  |  |  |  |  |  |  |  |  |
A  T  T  C  G  G  A  C  A  T
3'     reverse strand      5'
```

The reverse strand, also read from the 5' to 3' direction, will thus be read from right to left. This would correspond to the sequence `TACAGGCTTA`.

End of Example

2.1.3 Genes

Genes are discrete functional units located on the genome. A gene codes for a protein; that is, a gene is comprised of the blueprint of how a particular protein is to be synthesized. This blueprint is written in an alphabet comprised of the four characters {A,C,T,G}. Only about 1–2% of the genome codes for the approximately 20,000–25,000 genes in the human genome. The percentage of coding regions in other eukaryotes[1] is comparable.

The function of the remainder of the DNA is relatively unknown but is generally believed to be associated with differential regulation of gene expression

[1] Organisms where the DNA is packaged inside the nucleus are called eukaryotes, while DNA floats freely in the cytoplasm in a prokaryote. Higher level life forms, such as mammals, plants, reptiles and fungi are eukaryotes. Lower level life forms such as bacteria are prokaryotes.