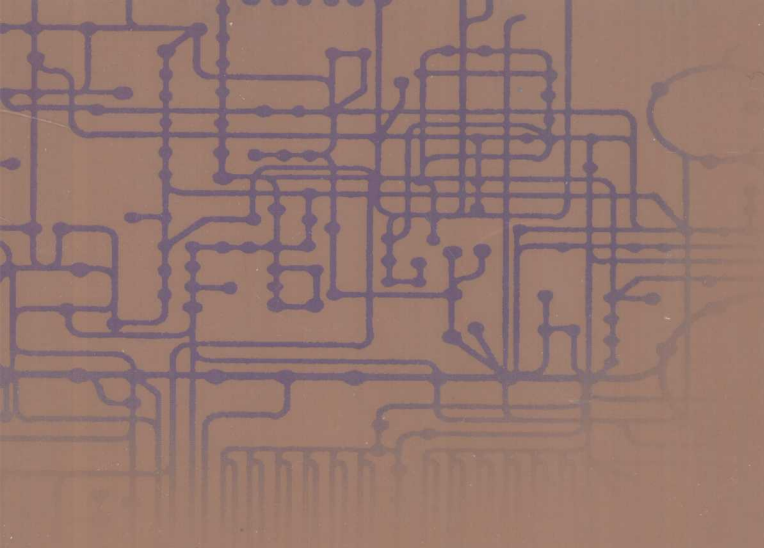

Post-genome Informatics

Minoru Kanehisa

Post-Genome Informatics

Minoru Kanehisa

Institute for Chemical Research, Kyoto University, Japan

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford OX2 6DP
Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in

Oxford New York

Athens Auckland Bankok Bogotá Buenos Aires Calcutta
Cape Town Chennai Dares Salaam Delhi Florence Hong Kong Istanbul
Karachi Kuala Lumpur Madrid Melbourne Mexico City Mumbai
Nairobi Paris São Paulo Taipei Tokyo Toronto Warsaw

with associated companies in Berlin Ibadan

Oxford is a registered trade mark of Oxford University Press
in the UK and in certain other countries

Published in the United States
by Oxford University Press Inc., New York

First published 2000

A catalogue record for this book is available from the British Library

Library of Congress Cataloging in Publication Data
(Data available)

ISBN 0 19 850327 X (Hbk)
ISBN 0 19 850326 1 (Pbk)

Typeset by J&L Composition Ltd, Filey, North Yorkshire
Printed in Great Britain
on acid-free paper by

Biddles Ltd, Guildford & King's Lynn

Preface

The Human Genome Project was initiated in the late 1980s as the result of technological developments in molecular biology and with the expectation of biomedical benefits. The project's goal was to determine the entire sequence of three billion nucleotides in the human genome and to identify and understand a whole repertoire of human genes. At the same time, genome projects were conceived and undertaken for a number of organisms from bacteria to higher eukaryotes. Both the public and private sectors are spending unprecedented amounts of resources in order to quickly decipher the genomes and to claim discovery of the information. However, the determination of the complete genome sequence is not the end of the story. It is actually the beginning of 'post-genome informatics', especially in view of the fact that the biological function cannot be inferred from the sequence information alone for roughly one half of the genes in every genome that has been sequenced.

Conceptually, whole genome sequencing represents an ultimate form of reductionism in molecular biology. It is hoped that complex processes of life can be explained by simple principles of genes. In experimental reality, DNA sequencing requires drastic reductions from higher to lower dimension—to destroy the cell and to extract the DNA molecules. We do not question how much information is lost in these procedures, but simply accept the common wisdom that the genome, or the entire set of DNA molecules, contains all the necessary information to make up the cell. Post-genome informatics is then considered as an attempt at synthesis from lower to higher dimension, whereby the functioning biological system of a cell is reconstructed from the entire complement of genes.

The genome projects have transformed biology in many ways, but the most impressive outcome is the emergence of computational biology, also known as bioinformatics. It is no longer possible to make advances in biology without integration of informatics technologies and experimental technologies. Here we like to distinguish between genome informatics and post-genome informatics. Genome informatics was born in order to cope with the vast amount of data generated by the genome projects. Its primary role is therefore to support experimental projects. In contrast, post-genome informatics, as we define here, represents a synthesis of biological knowledge from genomic information toward understanding basic principles of life, as well as for practical purposes in biomedical applications. Post-genome informatics has to be coupled with systematic experiments in functional genomics using DNA chip and other technologies. However, the coupling is the other way around—informatics plays more dominant roles of making predictions and designing experiments.

This book is an introduction to bioinformatics, an interdisciplinary science encompassing biology, computer science, and physics. In fact the major motivation

for writing this book is to provide conceptual links between different disciplines, which often share common ideas and principles. The content is in part a translation of my book in Japanese *Invitation to genome informatics* (Kyoritsu Shuppan, Tokyo, 1996), which originated from my lecture notes on theoretical molecular biology for undergraduate students in the Faculty of Science, Kyoto University. The first chapter is a concise introduction to molecular biology and the Human Genome Project. The second and third chapters provide an overall picture of both database and computational issues in bioinformatics. They are written for basic understanding of underlying concepts rather than for acquiring the superficial skills of using specific databases or computational tools. Because most algorithmic details are deliberately left out in order to cover a wide range of computational methods, it is recommended that the reader consult the references in the Appendix where necessary.

The last chapter, which has been specially written for this English edition, is the essence of post-genome informatics. It introduces the emerging field of network analysis for uncovering systemic functional information of biological organisms from genomic information. KEGG (Kyoto Encyclopedia of Genes and Genomes) at *www.genome.ad.jp/kegg/* is a practical implementation of databases and computational tools for network analysis. It is our attempt to actually perform synthesis of biological systems for all the genomes that have been sequenced. Since the field of network analysis is likely to evolve rapidly in the near future, KEGG should be considered as an updated version of the last chapter.

The very concept of post-genome informatics grew out of my involvement in the Japanese Human Genome Program. I have been supported by the Ministry of Education, Science, Sports and Culture since 1991 as princpal investigator of the Genome Informatics Project. This book is the result of active collaborations and stimulating discussions with the many friends and colleagues in this project. I am grateful to Chigusa Ogawa, Hiroko Ishida, Saeko Adachi, and Toshi Nakatani for their work on the drawings and to Stephanie Marton for checking the text of the English edition. The support of the Daido Life Foundation is also appreciated.

With a diverse range of Internet resources publicly available, it is not difficult for anyone interested to start studying post-genome informatics. I hope this book will help students and researchers in different disciplines to understand the philosophy of synthesis in post-genome informatics, which is actually the antithesis of the extreme specialization found in current scientific disciplines. The study of post-genome informatics may eventually lead to a grand synthesis—a grand unification of the laws in physics and biology.

Minoru Kanehisa

Kyoto, Japan
May 1999

Contents

1

Blueprint of life

Gene and genome

Life is a complex system for information storage and processing. Information is transmitted 'vertically' from cell to cell and from generation to generation, while at the same time information is expressed 'horizontally' within a cell in the ontogenesis of an individual organism. The information transmission from parent to offspring must have been recognized vaguely as heredity since the early days of human history. However, it was Gregor Mendel's experiments on the garden pea, performed in the mid-1800s but not recognized until 1900, that provided a first glimpse of its mechanism. A hereditary unit, later called a gene, was found to determine a particular characteristic, or a trait, of an organism. The Mendelian law of inheritance has established, first of all, that genes are inherited more or less independently in the vertical flow of information transmission and, secondly, that there is a set of competing genes (alleles) so that what is inherited (genotype) is not necessarily what is observed (phenotype) as the result of the horizontal flow of information expression.

Biology is the science of life that aims at understanding both functional and structural aspects of living organisms. In the latter half of the nineteenth century, great advances were made not only in genetics, a branch of biology analysing functional appearances of heredity, but also in cell biology based on microscopic observation of cellular structures. Most importantly, it was discovered that the chromosome in the nucleus of a cell contains the hereditary material. The entire set of genes in the chromosome, or more precisely in the haploid chromosome, was later named the genome. In view of the dual flow of information in life, the genome can be defined as the structural and functional unit of the information transmission and the gene as the structural and functional unit of the information expression (Table 1.1).

The disciplines of genetics and cell biology that emerged in the nineteenth century are the roots of modern biology, especially the golden era of molecular biology in the latter half of the twentieth century. The elaboration of

Table 1.1. Genome and gene

Entity	Definition	Molecular mechanism
Genome	Unit of information transmission	DNA replication
Gene	Unit of information expression	Transcription to RNA Translation to protein

experimental technologies in molecular biology has established the molecular basis of genes and genomes, uncovered the structure–function relationships of biological macromolecules in a diverse range of cell processes, and ultimately led to the conception of the Human Genome Project, a project working towards deciphering the blueprint of life. Let us quickly overview these developments.

DNA and protein

The chromosome is a molecular complex made from deoxyribonucleic acids (DNAs) and proteins. Originally, the protein was suspected to be the genetic material, but by the mid-1900s it became apparent that the DNA contained the information transmitted and the protein was synthesized within the cell. In 1953 James Watson and Francis Crick proposed the double helix model for the DNA structure. The model was constructed from the X-ray diffraction data for the DNA fibres, which had been obtained by Rosalind Franklin and Maurice Wilkins, together with the experimental observation that in any DNA the composition of adenine (A) plus thymine (T) was equal to the composition of guanine (G) plus cytosine (C). In essence, a DNA molecule contains two chains, where each chain is a linear polymer consisting of the repeating units of the four nucleotides, A, T, G, and C (Tables 1.2 and 1.3). The structural complementarity and the resulting hydrogen bonding between A and T and between G and C stabilize the assembly of the two chains into the double helix structure (Fig. 1.1). The genetic informa-

Table 1.2. Nucleic acid and protein

Macromolecule		Backbone	Repeating unit	Length	Role
Nucleic acid	DNA	Phosphodiester bonds	Deoxyribonucleotides (A, C, G, T)	10^3–10^8	Genome
	RNA	Phosphodiester bonds	Ribonucleotides (A, C, G, U)	10^3–10^5 10^3–10^4 10^2–10^3	Genome Messenger Gene product
Protein		Peptide bonds	Amino acids (A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y)	10^2 – 10^3	Gene product

Table 1.3. Nucleotide codes

A	Adenine	W	Weak (A or T)
G	Guanine	S	Strong (G or C)
C	Cytosine	M	Amino (A or C)
T	Thymine	K	Keto (G or T)
U	Uracil	B	Not A (G or C or T)
R	Purine (A or G)	H	Not G (A or C or T)
Y	Pyrimidine (C or T)	D	Not C (A or G or T)
N	Any nucleotide	V	Not T (A or G or C)

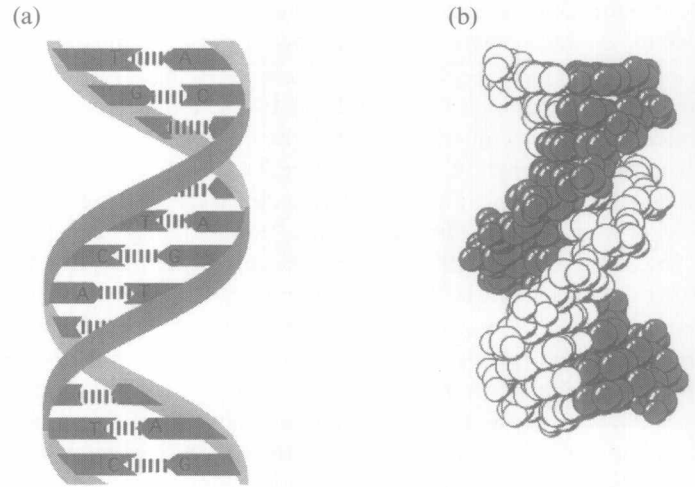

Fig. 1.1. DNA double helix. (a) A schematic diagram of the Watson–Crick model. (b) The three-dimensional structure of a synthetic DNA, 'CGCGAATTCGCG' (PDB:7BNA).

tion is encoded in the sequence of nucleotides that forms the polymer, and once the nucleotide sequence of one chain is given, the other sequence is automatically determined by the complementarity. Thus, the DNA double helix model has a profound implication for the molecular mechanism of heredity. The transmission

Table 1.4. Amino acid codes

Ala	A	Alanine
Arg	R	Arginine
Asn	N	Asparagine
Asp	D	Aspartic acid
Cys	C	Cysteine
Gln	Q	Glutamine
Glu	E	Glutamic acid
Gly	G	Glycine
His	H	Histidine
Ile	I	Isoleucine
Leu	L	Leucine
Lys	K	Lysine
Met	M	Methionine
Phe	F	Phenylalanine
Pro	P	Proline
Ser	S	Serine
Thr	T	Threonine
Trp	W	Tryptophan
Tyr	Y	Tyrosine
Val	V	Valine
Asx	B	Asn or Asp
Glx	Z	Gln or Glu
Sec	U	Selenocysteine
Unk	X	Unknown

of genetic information is realized by the replication of DNA molecules (Table 1.1) in which the complementary chains guarantee correct copying of the information.

The protein is a linear polymer of the 20 different kinds of amino acids, which are linked by peptide bonds (Tables 1.2 and 1.4). The three-dimensional structure of a protein that results from the folding of the polypeptide chain is far more complex than the double helical DNA structure. This complexity reflects the variety and specificity of protein functions. The amino acid sequence, or the primary structure of a protein, was first determined for insulin by Frederick Sanger in 1953. The three-dimensional (3D) structure, or the tertiary structure, was first elucidated for myoglobin by John Kendrew in 1960 using the X-ray diffraction of protein crystals. The X-ray diffraction was refined by Max Perutz and became known as the isomorphous replacement method. The tertiary structure of myoglobin is an assembly of six α-helices as shown in Fig. 1.2. The model structure of an α-helix was originally proposed for homopolymers of single amino acids by Linus Pauling in 1951, who also predicted the existence of α-helices in proteins. As more 3D structures were resolved general principles emerged for the functioning of proteins. It was again the structural complementarity, as in the case of the

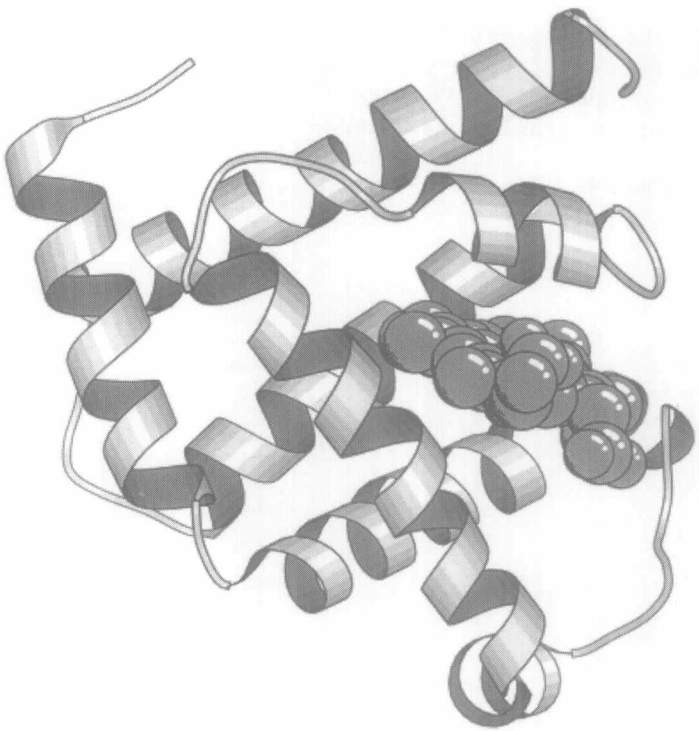

Fig. 1.2. The three-dimensional structure of sperm whale myoglobin (PDB:1MBN).

DNA double helix, for example, that enabled enzymes to recognize and react on substrates. The enzyme–substrate complementarity is like the matching of a key-hole with a specific key, but the analogy should not be taken in a strict sense since molecular structures are flexible and inducible. It is the chemical complementarity in the structural framework that ultimately determines the interaction of two molecules.

Central dogma

DNAs and proteins are the biological macromolecules that play key roles in the living cell. Both are linear polymers of repeating units (Table 1.2). The genetic information is encoded and stored in the sequence of four types of nucleotides, and DNA replication is the molecular mechanism for genetic information transmission. In contrast, the actual functioning of the cell is mostly performed by proteins. Thus, the conversion of a DNA sequence with four possible 'letters' into a protein sequence with twenty possible 'letters' is required. This is a crucial step in the genetic information expression. The molecular mechanism of this conversion is the translation where a triplet of nucleotides, or a codon, is translated into an amino acid. The translation table of 64 codons into 20 amino acids is specified by the genetic code shown in Table 1.5. Originally the genetic code was considered to be universal among all species, but now a number of variations are known, as shown in Table 1.6.

Table 1.5. Standard genetic code

1st Position	2nd position								3rd Position
	U		C		A		G		
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys	C
	UUA	Leu	UCA	Ser	UAA	*Stop*	UGA	*Stop*	A
	UUG	Leu	UCG	Ser	UAG	*Stop*	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC	Leu	CCC	Pro	CAC	His	CGC	Arg	C
	CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg	A
	CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg	G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser	C
	AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg	A
	AUG	Met, *Start*	ACG	Thr	AAG	Lys	AGG	Arg	G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly	C
	GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly	A
	GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly	G

Table 1.6. Variation of genetic codes

	T1	T2	T3	T4	T5	T6	T9	T10	T12	T13	T14	T15
CUU	Leu	-	Thr	-	-	-	-	-	-	-	-	-
CUC	Leu	-	Thr	-	-	-	-	-	-	-	-	-
CUA	Leu	-	Thr	-	-	-	-	-	-	-	-	-
CUG	Leu	-	Thr	-	-	-	-	-	Ser	-	-	-
AUU	Ile	-	-	-	-	-	-	-	-	-	-	-
AUC	Ile	-	-	-	-	-	-	-	-	-	-	-
AUA	Ile	Met	Met	-	Met	-	-	-	-	Met	-	-
AUG	Met	-	-	-	-	-	-	-	-	-	-	-
UAU	Tyr	-	-	-	-	-	-	-	-	-	-	-
UAC	Tyr	-	-	-	-	-	-	-	-	-	-	-
UAA	*Stop*	-	-	-	-	Gln	-	-	-	-	Tyr	-
UAG	*Stop*	-	-	-	-	Gln	-	-	-	-	-	Gln
AAU	Asn	-	-	-	-	-	-	-	-	-	-	-
AAC	Asn	-	-	-	-	-	-	-	-	-	-	-
AAA	Lys	-	-	-	-	-	Asn	-	-	-	Asn	-
AAG	Lys	-	-	-	-	-	-	-	-	-	-	-
UGU	Cys	-	-	-	-	-	-	-	-	-	-	-
UGC	Cys	-	-	-	-	-	-	-	-	-	-	-
UGA	*Stop*	Trp	Trp	Trp	Trp	-	Trp	Cys	-	Trp	Trp	-
UGG	Trp	-	-	-	-	-	-	-	-	-	-	-
AGU	Ser	-	-	-	-	-	-	-	-	-	-	-
AGC	Ser	-	-	-	-	-	-	-	-	-	-	-
AGA	Arg	*Stop*	-	-	Ser	-	Ser	-	-	Gly	Ser	-
AGG	Arg	*Stop*	-	-	Ser	-	Ser	-	-	Gly	Ser	-

T1, Standard code; T2, vertebrate mitochondrial code; T3, yeast mitochondrial code; T4, mould, protozoan, and coelenterate mitochondrial code and mycoplasma and spiroplasma code; T5, invertebrate mitochondrial code; T6, ciliate, dasycladacean and hexamita nuclear code; T9, echinoderm mitochondrial code; T10, euplotid nuclear code; T12, alternative yeast nuclear code; T13, ascidian mitochondrial code; T14, flatworm mitochondrial code; T15, blepharisma nuclear code.

The conversion of DNA information into protein information is not direct; DNA is first transcribed to RNA (ribonucleic acid) which then is translated to protein. This particular type of RNA is called messenger RNA (mRNA), since other types of RNAs also exist including transfer RNA (tRNA) and ribosomal RNA (rRNA). The RNA is another linear macromolecule that is closely related to DNA; the only differences are that the sugar backbone is made of ribose rather than deoxyribose and that thymine (T) is substituted by uracil (U) (Tables 1.2 and 1.3). In a symbolic representation, the transcription is to change only the letter T in the DNA sequence to the letter U to obtain the RNA sequence. Thus, there is a unidirectional flow of information expression from DNA to mRNA to protein. This flow, together with the flow of information transmission from DNA to DNA, forms the central dogma of molecular biology (Fig. 1.3) as first enunciated by Francis Crick in 1958.

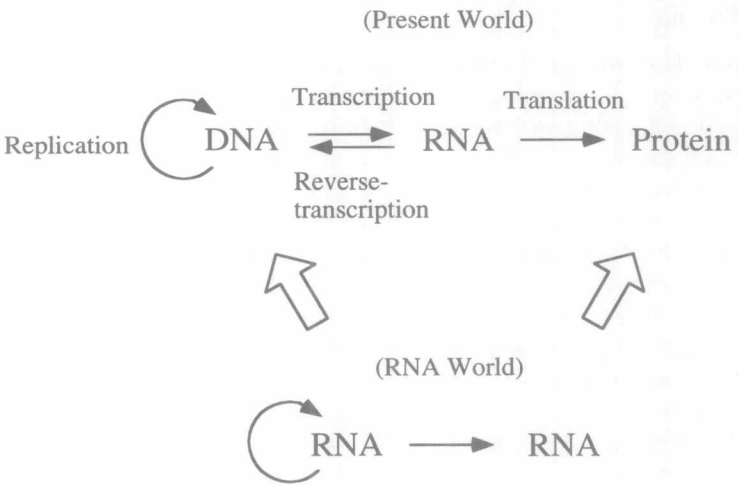

Fig. 1.3. The central dogma and its evolution.

However, it soon became known that in a special virus called a retrovirus there appeared to be an opposite flow of information—from RNA to DNA. This reverse-transcription was predicted by Howard Temin and the enzyme responsible for the reaction, reverse-transcriptase, was discovered in 1970. Retroviruses are an intriguing class of viruses that include many cancer viruses and the AIDS virus, and they store the genetic information in genomic RNA rather than DNA. The genetic information transmission thus requires the reverse-transcription from RNA to DNA before the DNA replication.

In 1977 another surprising discovery was made. The mRNA transcript was not entirely used for the following step of translation. The protein coding gene was found to be divided into pieces, which is called the exon/intron structure. There is an additional step of mRNA processing where introns are removed and exons are spliced to form a mature mRNA molecule. This mRNA splicing is prevalent in higher species, and occurs in the nucleus after transcription from genomic DNA. It is then followed by the nuclear export of mature mRNA and the translation of protein on the ribosome. RNA splicing is also known to exist in tRNA genes and rRNA genes in certain species. Furthermore, in contrast to RNA splicing that is an operation of nucleotide deletions, another type of processing called RNA editing that involves insertions and replacements of specific nucleotides has been discovered in a limited group of species. The post-transcriptional RNA processing does not contradict the central dogma in terms of the direction of information flow, but it certainly has caused modifications on the complexity of RNA information processing.

RNA world

In almost all of the species that presently exist on Earth, the information of the genome is stored and transmitted in DNA and the information of the gene is expressed to the functional molecule of protein. Is RNA's main role, as the central dogma implies, simply to be a messenger for the more important molecules of DNA and protein? Quite the contrary. It is widely believed that life started with RNA and there was an RNA world when neither DNA nor protein yet existed. RNA must have been playing a dual role both as the genome in the information transmission and as the functional molecule in the information expression. In fact, RNA can be a genome because there still exist viruses with RNA genomes. RNA can be a functional molecule, first because tRNAs and rRNAs are functional molecules expressed from genes, but more importantly because a new class of RNAs with catalytic activities was discovered by Thomas Cech in 1981. The discovery of this ribozyme transformed the idea that the catalysis of chemical reactions was exclusively performed by protein enzymes. The repertoire of catalysis by ribozymes has since been increasing, partly from additional discovery of natural ribozymes but mostly from the design and selection of artificial ribozymes by *in-vitro* molecular evolution experiments.

If the RNA world existed, then the central dogma would have evolved as shown in Fig. 1.3. In terms of the stability of information storage and the fidelity of copying, DNA is a more favourable molecule than RNA. In terms of the variety and specificity of catalytic functions, protein is a more favourable molecule than RNA. These could have been the reasons that DNA and protein, with their separate roles, respectively, for information storage (template) and information processing (catalyst), replaced the dual roles of RNA. The bizarre phenomena of RNA splicing and RNA editing could be remnants of RNA processing in the RNA world to produce catalytic molecules from templates in genomic RNA. The fact that ribose rather than deoxyribose is often seen in biologically active chemicals, such as ATP, cyclic AMP, NAD, and coenzyme A, could also be considered a remnant of the RNA world. The variety of genetic codes (Table 1.6) could suggest that the central dogma evolved somewhat independently in different species.

Then, how could the transition from the RNA world to the DNA/protein world happen? The transition from RNA to DNA is conceivable because they are highly similar molecules and they can be used in an identical manner for information storage. In contrast, RNA and protein are quite different molecules. However, what needs to be preserved here is not the capacity of information storage but the capacity of information processing. Therefore, as long as the catalytic activity was preserved, the switch from RNA to protein could happen in the evolution of life, like switching from cathode ray tube to transistor in the early days of computer development. Since the catalytic activity is highly related to the structural and chemical complementarity of interacting molecules, the keyhole–key relation of interface structures must have been preserved during the transition of the two worlds.

Cell

As we have seen, a most fundamental aspect of life is information storage and processing. The information is stored in molecules and the information is processed by molecules. The information processing to copy the storage is called information transmission, and the information processing to make catalysts from templates in the storage is called information expression. While the central dogma neatly summarizes the molecules that are involved in the two types of information processing, it does not state when and where the processing should take place. In fact, life would not exist without the cell, which provides the field for space-dependent and time-dependent processing.

The concept that all biological organisms are made up of structural and functional units, or cells, was established by the early nineteenth century. The cell contains smaller, or subcellular, units such as the nucleus, which stores genetic information. Biological organisms have traditionally been classified into two domains by the cellular architecture—eukaryotes that are made up of cells with a nucleus and prokaryotes that are made up of cells without a nucleus. Note, however, that biological organisms can also be divided into three domains of life by the

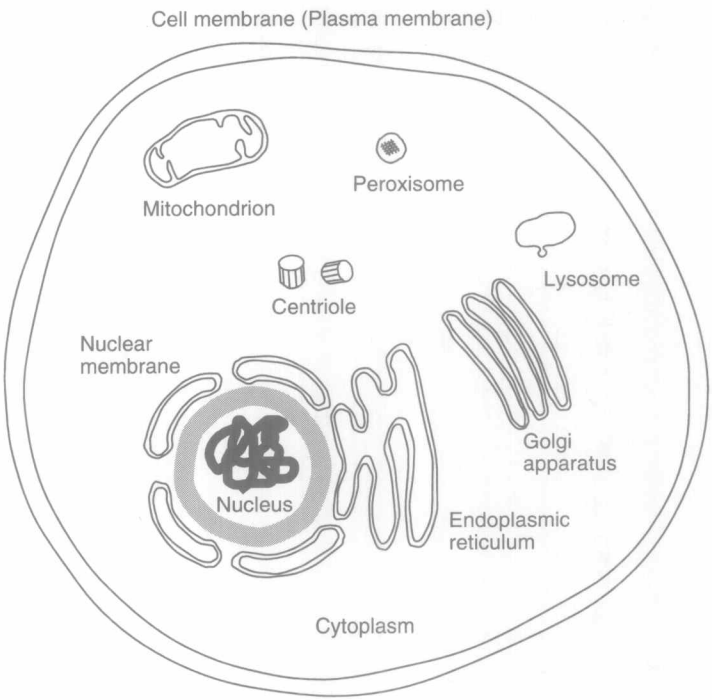

Fig. 1.4. A schematic illustration of a eukaryotic cell.