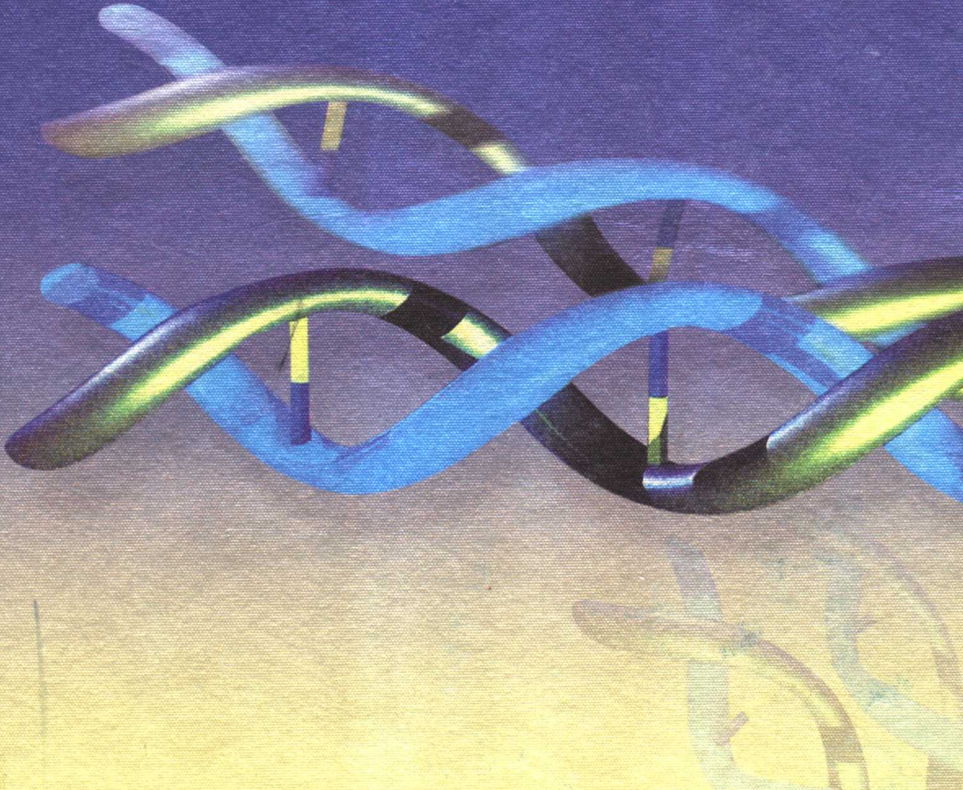
Chinese Science Series

Theoretic-Physical Approach to Molecular Biology

Liaofu Luo
Inner Mongolia University



Shanghai Scientific & Technical Publishers

Theoretic-Physical Approach to Molecular Biology

Liaofu Luo
Inner Mongolia University

江苏工业学院图书馆
藏书章

Shanghai Scientific & Technical Publishers
Shanghai

Copyright © 2004 by Shanghai Scientific & Technical Publishers

All rights reserved. Reproduction or translation of any part of this work without the permission of the copyright owner is unlawful. Requests for permission of the copyright information should be addressed to Shanghai Scientific & Technical Publishers, 450 Ruijin Er Road, Shanghai. Printed in Shanghai, the People's Republic of China

图书在版编目(CIP)数据

分子生物学的理论物理途径 = Theoretic-Physical Approach to Molecular Biology/罗辽复著. —上海:
上海科学技术出版社,2004.9
(中国科学丛书)
ISBN 7-5323-7479-3

Ⅰ.分… Ⅱ.罗… Ⅲ.理论物理学—应用—分子生物学—研究—英文 Ⅳ.Q7

中国版本图书馆 CIP 数据核字(2004)第015174号

责任编辑:王桂华 赵玲丽

世纪出版集团
上海科学技术出版社 出版、发行
(上海瑞金二路450号 邮政编码200020)
新华书店上海发行所经销
上海汇杨印刷厂印刷
开本 787×1092 1/16 印张 38.5 插页 4
字数 696 000
2004年9月第1版
2004年9月第1次印刷
印数 1—1 000
定价: 108.00元

本书如有缺页、错装或坏损等严重质量问题,
请向本社出版科联系调换

TO THE MEMORY OF
MY PARENTS

Foreword

The research of traditional biology is from morphology to cytology and then to the atomic and molecular level, from physiology to microscopic regulation, and from phenotype to genotype. However, the recent development of life science shows that the process might be reversed. W. Gilbert, Nobel Prize winner, wrote in *Nature* (1991), "The new paradigm, now emerging, is that all the genes will be known (being resident in databases available electronically) and the starting point of a biological investigation will be theoretical. An individual scientist will begin with a theoretical conjecture, only then turning to experiment to follow or test that hypothesis." Sequence — Structure — Function, this is a possible line of reverse biology. It begins with the research on genes and moves to molecular sequence, then to molecular conformation, from structure to function. In the meantime, it sets about with a unifying principle and extensively uses mathematical tools to quantitatively clarify the ever-changing phenomena of life. Obviously, the establishment of such reverse biology will completely change the appearance of life sciences and allow biology to gain more benefits from deductive inferences of mathematics. Some scientists summed up three causes for the low level of natural science research in medieval times: firstly, the scholars were in awe of authorities; secondly, the level of mathematical research was low; thirdly, the experimental nature of sciences was unknown at that time. This point of view is very informative. If introducing mathematics into biology, strengthening interaction between mathematics, physics and biology, and reforming biology into a systematical science, making it in harmony with the rational spirits of mathematics, we can greatly improve the research level and the prediction potential of life science. When discussing the great book

此为试读，需要完整PDF请访问：www.ertongbook.com

Nature, the founder of modern science, Galileo said it is written in mathematical language, and without mathematics, one can not but hesitate in the maze of darkness, which also holds true for today's life science. Reviewing the history, we find that in the early twentieth century, with the advent of two revolutionary theories: relativity theory and quantum theory, theoretical physics, an experimental science in origin, as a new branch came into being and separated itself from traditional physics. The reason for this separation was that physics had developed to such a stage that theoretical analysis must be carried out independently and systematically, and mathematical methods were to be creatively used as well. Now, a similar process of separation has begun to appear in the research of life science. For example, how is life formed and evolved from lifeless nature? How does life display its function as the aggregate of molecules and atoms according to the physical laws? How does the genetic information express itself under the precise control? How does the brain recognize, learn, memorize and think? etc. Answers to these mysteries of life phenomena, on one hand, depend upon the accumulation of experimental data, on the other hand, rely on a deep theoretical analysis and summary of experimental materials.

The modern scientific tradition starts from Galileo-Newton age. What are the main features of this new paradigm? First is experimentalism. Galileo inherited the enlightening ideas of Renaissance such as any reliable knowledge is the product of experience; nature was accurately representable only by virtue of careful observation; the learning not originated from and unable to be tested by experience, the learning obtained without the involvement of any sensory organ at any stage is often groundless and erroneous. Without adequate emphasis on experiments, the fetters of Medieval Scholastic Philosophy can never be completely shattered. It is through the synthesis of observation and theory that Galileo established himself the founder of modern Science. The second key feature of modern science is rationalism. The best and most rigorous and effective deduction tool is mathematics. In *Philosophiae Naturalis Principia Mathematica* Isaac Newton developed an overall scheme and formulated the physical laws of Mechanics through almost a copy of the logic methods of Euclid's *The Elements*, i. e. following the pattern: from definition to axiom and then to theorem. The broad application of mathematics makes possible the accurate portrayal of the real physical world. Without rigorous mathematical deduction, how can we prove that the gravitational force between the earth and the moon and the force determining the parabolic motion on the Earth belong to the same force — Universal Gravitational Force? It is the integration of experimentalism and rationalism that constitutes the recent four hundred years scientific tradition and paradigm. This integration is represented most fully and fruitfully in physics, especially in the 20th century's modern physics. With invincible power, it is conquering every aspect of natural science. The permeation of this paradigm into life science is thus inevitable. As the result of this diffusion, life science must be not only experimental but also theoretical and comprehensive.

This book is the summary of our group's work in theoretical biology in the past 15 years (1987—2002). By choosing *Theoretic-Physical Approach* as its title instead of using a more common one *Theoretical approach*, the author intends to emphasize the significance of exploring the essence of the phenomena and finding the 'physics' behind them rather than only resorting to mathematical methods. Richard Feynman once said, "Physicists always have the habit of taking the simplest example of any phenomenon and calling it 'physics'." Although the life phenomena are extremely complex, we expect that this method of constructing the simplified model, neglecting all non-essential factors and extracting the 'coarse-grained' laws for the model system, can work effectively on the main topics of molecular biology. The characteristic feature of life which differentiates from an inanimate piece of matter is the large amount of information contained in it. Different from "matter" and "energy", "information" constitutes the third category in natural sciences. The central task of a mature theoretical approach to molecular biology thus lies in exploring the law on the formation, storage, expression and transmission of life information in each step, from DNA to mRNA, to amino acid sequence and finally to protein structure and function. Of course, to attain the goal, there is still a long way to go. The choice of the subject matter in this book only reflects on the author's personal point of view and his present research experience. Therefore it cannot be expected to cover every aspect of this field. In fact, in such a broad and rapidly developing field, even a relative comprehensiveness is hard to achieve. So it is probably more proper to consider this book a research report on the progress of theoretical biology.

Unexceptionally, every mature discipline has its own relatively stable scientific community and a rather consistent value system. Being a new discipline, theoretical biology is forming its community. I consider myself fortunate to be assigned to the quiet border of the desert to carry out my research. Here I enjoy the freedom of creation and the ability to proceed with multi-level independent thinking within different areas with little interruption. Most research works collected in this book were finished in this unique environment, among which, some have not been formally published yet, or only with preliminary results published in Inner Mongolia University Journal. In order to systematically, comprehensively present our work and contribute it in time to the visual field of the scientific community of theoretical biology, to publish this monograph might be necessary.

Life exists as an apparatus of genes. For human beings, in addition to biological genes, there is another still evolving gene-like structure that can be copied, transmitted: cultural genes. Genes and cultural genes are the only two things we can leave to our offspring. Plato demonstrated such a philosophy in his *Symposium*: " Every mortal creature is seeking as far as possible to be everlasting and immortal: and this is only to be attained by generation, because generation always leaves behind a new existence in the place of the old. " He highly praised "the love of generation and of birth in beauty, whether of body or soul. " "Because to the mortal creature, generation is a sort of eternity and immortality. " The desire for immortality through creation is very moving, "this procreation is a divine thing; for conception and generation are an immortal principle in the mortal creature, and in the inharmonious they can never be. " Only creation can bring happiness, only creative beings are valuable beings. During the process of writing the book, I found this pleasure of creation.

I should like to thank many of my students and colleagues who had worked with me in exploring the topics relevant to this book. Their names can be found in the references under each section heading. I owe my completion of section §3.7 to the inspiration from Prof. Lee Hoong-Chien's manuscript. My thank also goes to the support of National Science Foundation of China, especially, the support of grant No. 90103030 for my work related to Chapter 4. Meanwhile, I feel grateful to the Shanghai Scientific & Technical Publishers for their efforts in publishing this book. The main parts of this book had been published in Chinese under the title of *The Physical Aspects of Life Evolution*. This time, however, in translating it into English, some major revisions have been made and supplements has been added. The Chinese manuscript had been reviewed by Professor Fang Tianqi in the Biology Department of Inner Mongolia University, and Professor Hao Bailin and

Professor Liu Jixing in Theoretic Physics Institute of the Chinese Academy of Sciences. I would like to express my sincere gratitude for their encouragement and valuable suggestions. I thank Dr. Xie Jiang and Laiou for their great help in language correction of some sections of the manuscript. Finally, I greatly appreciate Dr. Guo Weisheng's contribution to a large part of the graphic works.

Only those literatures closely related to our work are listed in the references after each chapter. However, due to the duration of the studies, the references cited in our earlier works may be incomplete from the present view. The author deeply apologizes for any omission and unaccredited quotation.

Liao-fu Luo

Inner Mongolia University, Hohhot, P. R. China
December 2002

Contents

此为试读，需要完整PDF请访问：www.ertongbook.com

Chapter 1
The Logic of Genetic Code

The genetic code is one of the greatest creations of the evolution of Nature because it contains not only the important working principles of life, but also abundant information on formation and evolution of life. In 1943, after Schrödinger summarized Delbruck's experiment on the mutation of bacteriophage induced by X-ray — the change of genetic substance occurs in a volume of ten atomic-distances-cubes and with an energy of 1 or 2 electron volts — he proposed the concept of "miniature code" as the molecular picture of gene which contains a large amount of genetic information [Schrödinger, 1944]. However, which kind of biological macromolecules is the genetic information stored in? In the early stage more attentions were concentrated on proteins but the nucleic acids had been ignored. Avery *et al.* discovered in 1943 that DNA in toxic strain *S pneumococci* can transform strain R into strain S and proved for the first time that DNA has the capacity of changing the bacterial heredity. Then in 1950, by using labeled atoms ^{35}S and ^{32}P Hershy and Chase established that while no (^{35}S-labeled) protein penetrates a phage-infected bacterial cell, (^{32}P-labeled) DNA was able to penetrate the cells. This observation indicated that it is the DNA injected into the cell that causes the infection and proved that DNA is the genetic material of the phage. During the same period, the publications of a series of important experimental and theoretical results promoted the birth of double-helix model of DNA. Among which, the first was the determination of G + C content of DNA equal to its A + T content by Charguff; the second was the point of view suggested by Pauling that the α helix in protein is formed by hydrogen bond formation; and the last and the most important one was the DNA crystal structure data obtained from X-ray diffraction by Franklin and Wilkins. Based on these works, in 1953 Watson and Crick proposed the double helix model of DNA which declared the emergence of a vital new science — molecular biology [Watson and Crick, 1953].

As a nuclear physicist and an astrophysicist with broad interests, Gamow's attention was attracted immediately to the model of DNA structure and its relation to genetic code. In 1954 Gamow firstly formulated the problem of genetic code [Gamow, 1954]. He hypothesized that the protein chain is synthesized directly on the DNA double helix, while every amino acid is located in a rhombus groove between four nucleotides, two of which belong to one chain of the helix and two to the other chain. One nucleotide in the first chain forms a Watson-Crick pair with one nucleotide in the second chain. Gamow demonstrated that there are exactly 20 different rhombuses, each containing four nucleotides and corresponding to one amino acid. Although his idea on DNA double helix as the template of protein synthesis is invalid (as reviewed by Crick) , and as proved later, the genetic code is a non-overlapping co-linear triplet-code (co-linear means that the codons in a nucleic acid and the corresponding amino acid residues in a protein are arranged in the same linear sequence) instead of an overlapping quadruplet in his diamond code model, Gamow's work is of great importance because it is the first truly abstract theory of code that contains no superfluous and unnecessary chemical details.

Nirenberg, Matthaei and Ochoa initiated the deciphering of the genetic code in 1961, and it was completed between 1965—1966 by Nirenberg and Khorana through the direct experimentation on protein biosynthesis [Nirenberg, 1973; Khorana, 1973]. They established the corresponding relationships between 64 codons and 20 amino acids. Now that the genetic code has been deciphered completely, what is the deeper implication of the code? Why the 20 amino acids and three terminators are arranged in such a way in the prevalent code dictionary? Is there any logic that should be obeyed in order to deduce the code table? What is the logic behind the universality of the genetic code? We will discuss these questions in the following sections.

§ 1. 1 The Degeneracy Rule of the Genetic Code*

The constancy of the genetic code among different organisms is one of the most striking, interesting and challenging phenomena in life. The mathematical relation behind the constancy intrigued many biologists and physicists. Historically, there are two different kinds of theories regarding the origin and evolution of the genetic code [Yockey, 1992; Giulio, 1997]. The first

* Luo, 1988; Luo, 1989; Luo, 2000.

approach originated from Gamow [1954]. His "Diamond code" model opened up a way to explain the origin of the universal amino acid code through the stereochemical interactions between codons or anticodons and amino acids [Woese *et al.*, 1966; Woese, 1967; Yarus and Christian, 1989; Shimizu, 1995]. The second approach is called "frozen accident" theory. The term "frozen accident", used firstly by Crick, means that all living organisms evolved from an ancient single ancestor, and after the evolutionary expansion of the descendants started, changes in the amino acid assignments of codons were not possible [Crick, 1968]. Recently, Trifonov *et al.* suggested that GCU is the first codon from the observation that mRNA sequences carry a hidden periodical pattern $(GCU)_n$ that may be considered a remnant of sequence organization of early evolution [Trifonov and Bettecken, 1997]. This is a variant of the frozen accident theory. In fact, the two theories can explain part of observations and experiments from their own standpoints, but they are by no means comprehensive. For example, the point that the universal code is superior due to some specific fit or affinity between each amino acid and its codon (through tRNA molecule as an adaptor) [Crick *et al.*, 1961] has never been proved rigorously. The deviant codon assignments discovered since 1979 demonstrate that other different codes are also possible [Jukes and Osawa, 1991]. On the other hand, the formation of the genetic code could not be explained as a fully accidental event. The hydrophobic order of amino acids consistent with that of their anticodonic dinucleotide is an important fact [Lacey *et al.*, 1983; 1992], which shows that the codon assignments may be required thermodynamically and some stereochemical relations may exist between the amino acids and the codons. So, the historical accident and the stereochemical constraint both exist and play their roles together in the formation of prevalent code. In fact, it has been proved that coevolution exists between amino acids and codes. The coevolution theory suggests that early on in the genetic code, only precursor amino acids were codified and later, as these precursors gave rise to new products, their codons underwent subdivision and some of the codons of each precursor were transferred to its product [Wong, 1975; 1988]. In recent years, new efforts have been made to explain the construction of genetic code in the following directions. The first is to find the group-theoretic symmetry, for example sp(6) symmetry [Hornos and Hornos, 1993], behind genetic code. However, should such a high symmetry sp(6) among four nucleotides exist in code, it must be seriously broken. But the decomposition of sp(6) symmetry did not reflect the real process of temporal refinement in the codon recognition [Nieselt-Struwe and Wills, 1997]. Perhaps, the most relevant group-theoretic description is the Klein 4-group [Findley *et al.* , 1982; Jimenez-Montano, 1999; Luo, 2000]. The second, initiated by Swanson [1984], investigates the gray code representation (rather than binary code) of genetic code. It emphasizes the order of importance of the bits in a code-word [Jimenez-Montano *et al.* , 1996]. In other words, the definite order of four nucleotides —UCGA— is an important factor in understanding the broken symmetry. We proposed a similar concept. Four nucleotides are represented by dual Yin-Yang states (a term borrowed from ancient Chinese philosophy) [Luo, 1992]. The third stresses the role of codon-anticodon interaction in the arrangement of amino acids in the code table. The thermodynamical measurement on codon-anticodon Gibbs energy reveals a possible explanation that the smallness of some energies makes the codons synonymous [Klump *et al.* , 1991; Klump, 1993; Jimenez-Montano, 1999]. The fourth optimizes the physicochemical distances between amino acids [Di Giulio *et al.* , 1994]. It could explain some properties of the code but the deduced distribution of amino acids in the table is much different from the prevalent genetic code. An alternative point is the introduction of "mean square" measure to quantify the relative efficiency of any given code [Freeland and Hurst, 1998]. They indicated that only one in every million random alternative codes generated is more efficient than the natural code and thus the natural genetic code is extremely efficient at minimizing the effects of errors. The fifth emphasizes the balanced accomplishment of robustness and changeability as an essential factor on the origin of genetic codes [Maeshiro and Kimura, 1998].

Since 1988 we have proposed an alternative approach to the problem. The central idea of the approach is the evolutionary stability of the code. It is generally accepted that the current standard code evolved from some archetypal amino acid code, simpler and more incomplete than its present form [Crick *et al.* , 1976; Jukes, 1983]. However, after it reached its current form, the standard code remained unchanged for billions of years. This is called the evolutionary stability of the genetic code, which means the prevalent code, as compared with other ideal codes, is more stable. It resists best the effects of mutations, which are the equivalent of noise or errors inherent to all information systems [Labouygues and Figureau, 1984]. We call these lethal effects of mutations mutational deterioration (*MD*, dangerousness, insecurity) [Luo, 1988; 1989]. Evolutionary stability means that from the minimization of MD we are able to deduce the prevalent genetic code. More plainly, deducing the arrangement of 64 codons in a code table comprises three problems: 1) 64 codons correspond to 20 amino acids (and terminators). The case of several codons corresponding to one amino acid should occur. We call it the degeneracy of codons — a term borrowed from physics. What is the degeneracy rule of the code? To find the degeneracy rule of codons in a multiplet that code for an amino acid or terminators, this is the first problem. The problem is also called the redundancy distribution in code. 2) The second problem is to find the distribution of amino acids (and terminators) in the code table, given the degree of degeneracy and a rule for each multiplet. 3) The third problem is to investigate the possible variation of degree of degeneracy and find the distribution of amino acids (terminators) under the new assignment of degeneracy degrees. We shall discuss the first problem in this section. Through introduction of the concept of mutational deterioration we shall demonstrate that the arrangement of each codon multiplet in code table (the degeneracy rule) is MD minimal and therefore obeys the stability principle.

From the genetic code (see Figure 1.1-1) we find that each degenerate codon doublet is located on the upper side or lower side of one of the 4 × 4 blocks in the table, that is, their first two nucleotides are the same and the third ones are related by a transitional mutation, a mutation not changing the purine or pyrimidine-type of the nucleotide. We also find each degenerate codon quartet located in one of the 4 × 4 blocks, namely, their first two nucleotides are common in the quartet. The triple and hexamerous codon multiplets (including terminators) all have their particular degeneracy rules. Ile and terminators both are triplet but their codons are arranged differently in the code table. The codons in three hexamerous multiplets, Leu, Ser, and Arg, have different arrangements too. How can these rules be explained? Our theory is based on the following assumptions [Luo, 1988; 1989]:

1. The prevalent code is a product of long-term evolution. The universality of the code over a wide range of organisms indicates its evolutionary stability. In particular, the codon arrangement in a multiplet obeys a stability principle. That is, as compared with other ideal arrangements, the real code is the most stable. Mathematically, for each ideal codon multiplet, one can define a mutational deterioration function that represents the mutational frequency of the multiplet and the deterioration

此为试读,需要完整PDF请访问: www.ertongbook.com