

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

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

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.

Chapter 1 The Logic of Genetic Code 1
 §1.1 The Uniqueness Rule of the Genetic Code 2
 §1.2 The Two Factors of Stop Codons in the Genetic Code 10
 §1.3 The Two Factors of Stop Codons in the Genetic Code 16
 §1.4 The Two Factors of Stop Codons in the Genetic Code 16
 §1.5 The Multiplicity of Amino Acid and the Abnormal Genetic Code 27
 §1.6 The Uniqueness and Duality of the Genetic Code 42
 §1.7 A Statistical Theory of Amino Acid Mutation 52
References 60
Chapter 2 Gene Informatics and Sequence Evolution (1) 61
 §2.1 The Evolution of Protein Sequence and Rate of Molecular Evolution 65
 §2.2 Sequence Correlation of Nucleotides and Its Short-Range Correlation 68
 §2.3 The Evolutionary Behavior of Gene Information 83
 §2.4 Evolutional Studies in DNA Sequences 89
 §2.5 Molecular Evolutionary Time 104
 §2.6 The Statistical Analysis of DNA Sequence and Long-range Correlation of Bases 124
 §2.7 DNA Walk and the Diagrammatic Representation for DNA Sequences 140
 §2.8 The Markovian Chain and Probability Matrix of Nucleotide Sequences 150

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.

Chapter 1 The Logic of Genetic Code 1
 §1.1 The Uniqueness Rule of the Genetic Code 2
 §1.2 The Two Factors of Stop Codons in the Genetic Code 10
 §1.3 The Origin of the Genetic Code 15
 §1.4 The Origin of the Genetic Code Table 18
 §1.5 The Origin of the Genetic Code Table 20
 §1.6 The Origin of the Genetic Code Table 22
 §1.7 The Origin of the Genetic Code Table 24
 December 2002

Chapter 2 The Origin of the Genetic Code 27
 §2.1 The Origin of the Genetic Code 27
 §2.2 The Origin of the Genetic Code 29
 §2.3 The Origin of the Genetic Code 31
 §2.4 The Origin of the Genetic Code 33
 §2.5 The Origin of the Genetic Code 35
 §2.6 The Origin of the Genetic Code 37
 §2.7 The Origin of the Genetic Code 39
 References 60

Chapter 2 Gene Informatics and Sequence Evolution (1) 61
 §2.1 The Evolution of Protein Sequence and Rate of Molecular 65
 §2.2 Sequence Correlation of Nucleotides and Its Short- 68
 §2.3 The Evolutionary Behavior of Gene Information 83
 §2.4 Evolutional Studies in DNA Sequences 89
 §2.5 Molecular Evolutionary Time 104
 §2.6 The Statistical Analysis of DNA Sequence and Long-range 124
 §2.7 DNA Walk and the Diagrammatic Representation for DNA 140
 §2.8 The Markovian Chain and Probability Matrix of Nucleotide 150