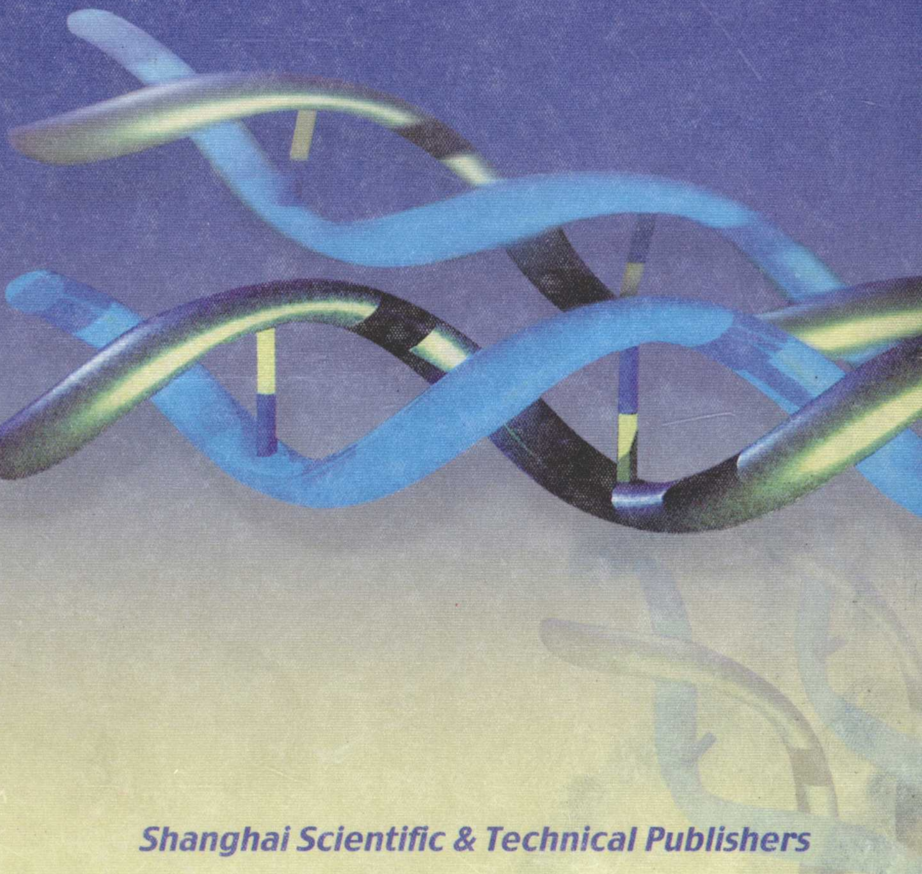
Chinese Science Series

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

Shanghai Scientific & Technical Publishers
Shanghai

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. O7

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
上海江杨印刷厂印刷
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

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

(continued)

Chrom.	Position	Length	Frame	IHI	R(name)	R(position)	R(frame)
O	493 422-494 399	326	c3	47	YOR091W	493 264-494 469	w1
P	141 036-141 722	229	w3	105	YPL217C	139 619-143 170	c2
P	700 779-701 753	325	c3	98	YPR080W	700 590-701 966	w3
P	945 870-946 550	227	c3	365	YPR204W	944 598-947 696	w3
P	945 928-946 608	227	w1	343	YPR204W	944 598-947 696	w3

The novel ORF generally overlaps with a part of another known longer gene, denoted as R in the table, but has different reading frame (at least a part of overlapping sequence being in different frame in the case of longer ORF containing intron). The locations, amino acid numbers, frames and IHI values are listed in the second, third, fourth and fifth columns of the table respectively. The name of corresponding longer genes and their positions and frames are listed in the sixth, seventh and eighth columns.

* This ORF has been reported as YFL013W-A, see Kumar *et al.* (2002).

Coding Potential Measure—Inhomogeneity Index

To identify an ORF as being the coding or non-coding we use a quantity based on Pearson's statistics [Feller, 1971], called inhomogeneity index (IHI). IHI of an ORF or a segment of sequence is defined as follows [Lee and Luo, 1997b]: Let N_j be the number of base j (j = A, C, G, T) in a segment (length N) of DNA sequence. We divide the segment into N/m multiplets with m bases in each multiplet. Let N_{ja} be the number of base j in a-th ($a = 1, 2, \cdots, m$) position of multiplet and $N_a = \sum_j N_{ja} = N/m$.

$$\mathrm{IHI} = \sum_{a}^{m} \sum_{j} \frac{\left(N_{ja} - \dfrac{N_j N_a}{N}\right)^2}{\dfrac{N_j N_a}{N}} \qquad (3.3.5)$$

In present case, the multiplet is the codon and $m = 3$. According to Pearson's theorem, if each base is homogeneously distributed over the three codon positions IHI obeys χ^2 distribution with 6 degrees of freedom as N is large enough. For a homogeneous sequence segment the expectation value of IHI is 6 with standard deviation $\sqrt{12} = 3.46$. The larger the inhomogeneity is, the greater the IHI will be.

We shall use IHI to differentiate ORFs in brewers yeast from intergenic sequences. Fickett *et al.* (1992) proposed four classes of content measures. The first class is based on codon usage (and counts of in-phase words); the second method is related to the encoded amino acid sequence; the third is related to the base compositional bias between codon positions; and the fourth is based on imperfect periodicity in base occurrences, etc. In fact, these measures are related to one another. Some methods are slight variants or special cases of others. For example, the periodicity of coding DNA sequence means a strong resonance at 1/3 position of correlation spectrum. The detailed calculation indicated that the resonance occurs due to four bases not uniformly distributed at three codon positions (see §2.6). So, the base compositional bias at the three codon positions is a key feature in the detection of the coding potential of a DNA segment. The information used for gene recognition is usually classified into three types: signals, content measures and similarity measures [Stormo, 2000]. The "first ATG" and the splice site in intron-exon boundary are examples of signals. The parameter IHI—heterogeneity index—defined above is a quantity related to content measures. The method based on content measures combined with methods based on signals provides an efficient approach to gene recognition with high accuracy [Burge and Karlin, 1998]. Since it, based on Pearson theorem, can differentiate between random and nonrandom occurrence of four bases on three codon positions IHI is important to describe the base compositional bias using an elaborate argument. It can serve as a marker to differentiate between coding and non-coding ORFs. In combination with other methods, this method may provide an effective tool for assessing the coding potential for an ORF, on a gene by gene basis. We shall utilize the IHI to yeast genome. By use of first class ORFs (i.e. known proteins) as positive samples and intergenic sequences as negative samples we are able to differentiate between them by calculation of IHI values. After removing intron-containing and mitochondrial, we choose 3 081 ORFs in the first class as positive samples. We obtained 3 552 intergenic sequences with lengths longer than 300 bp. They comprise the negative samples. For each sample the IHI value has been calculated. The results are summarized in Table 3.3-7. The distributions of IHI ranges are very different for known proteins (the first class ORFs) compared to intergenic sequences. The boundary marker can be defined as IHI = 14. More than 95% positive samples have IHI > 14. Only 152 positive samples (4.9% of 3 081) have IHI ≤14. More than 95% negative samples have IHI ≤14. Only 171 negative samples (4.8% of 3 552) have IHI > 14. So the gene identification by IHI rule (using IHI = 14 as a boundary marker) has sensitivity $S_n = 0.951$ and specificity $S_p = 0.945$. The accuracy (= the average of S_n and S_p) is 94.8%. The IHI for a homogeneous sequence is distributed in the range with expectation value 6 and standard deviation 3.46 according to χ^2 distribution. The negative sample taking IHI $\leqslant 14$ means that the intergenic sequence behaves as homogeneous within two standard deviations. It demonstrates that the symmetry of base composition among three codon positions is a feature of the majority of intergenic sequence and the coding potential of a DNA segment is closely related to the inhomogeneity of its base distribution.

The detection of spurious ORFs in the yeast genome has attracted the attention of many authors [Cliften *et al.*, 2001; Wood *et al.*, 2001]. The IHI parameter is a rapid and simple tool to distinguish coding and non-coding ORFs with high accuracy. We have calculated the IHI for each intron-less ORF in MIPS/GenBank matching database. (The IHI values for intron-containing ORFs are scattered. The more the intron number and the larger the intron length, the lower the IHI. So we have ignored them in statistics.) The results are shown in Table 3.3-8. The total number of ORFs in each class and the number of IHI > 14 and IHI $\leqslant 14$ ORFs are listed in the second, third and fourth lines of the table respectively. Of the 5 950 ORFs there are 5 187 with IHI > 14 and 763 with IHI $\leqslant 14$. As expected, the percentage of IHI $\leqslant 14$ ORFs in the six classes (line 5 of Table 3.3-8) increases from the first class to the sixth class. Considering the 95% confidence level of IHI as a tool to distinguish coding and non-coding ORFs, we can estimate the number of "spurious" ORFs (unlikely to be regarded as coding) from the percentage of IHI $\leqslant 14$ in each class. The results are shown in the last line of Table 3.3-8. For example, for the sixth class one has $385 \times (53.5 - 4.93)\% = 187$ spurious ORFs. The spurious ORF number in the first class is assumed to be zero. So, in the above calculation the IHI $\leqslant 14$ percentage has 4.93% subtracted, which is attributed to the false identification due to 95% confidence level of IHI. As seen from Table 3.3-8, the spurious ORFs can be neglected in the second class. They are distributed mainly in last three classes, namely, in the class of similarity to unknown proteins, no similarity or questionable ORFs. The total number of spurious ORFs is estimated to be 470, which is close to the recent estimation of Woods *et al.* (2001). The distribution of IHI $\leqslant 14$ ORFs in 16 chromosomes is given in Table 3.3-8a. The percentage of ORFs with IHI $\leqslant 14$ is near 13%. Considering the spurious ORFs existed mainly in the fourth, fifth and sixth classes we calculate the ratio (r) of the number of ORFs with IHI $\leqslant 14$ to the number of ORFs in the classes 4, 5 and 6. It is near 0.4 for most chromosomes. But for chromosomes 2 and 3 the ratio is higher than 0.5 and for chromosome 14 the ratio is lower than 0.3. The spurious ORF density seems different for different chromosomes.

Since the IHI defined by us has a deeper meaning in mathematics it should be applicable to other genomes. Using the same rule for *E. coli* we have deduced more than 92.3% positive samples taking IHI > 14 and more than 93.6% negative samples taking IHI ≤ 14. For *B. subtilis*, it is shown that more than 90.6% positive samples have IHI > 15 and more than 91.5% negative samples have IHI ≤ 15. So, the IHI rule appears to be valid at least for some prokaryotic organisms in addition to brewers yeast. Historically, other measures based on compositional bias have been proposed. For example, after defining the frequency of base j in position $a(a = 1, 2, 3)$ as $f(j, a)$, one may introduce three measures $M1$, $M2$ and $M3$ [Fickett and Tung, 1992] as follows:

$$M1 = \sum_j \frac{\max(f(j, 1), f(j, 2), f(j, 3))}{1 + \min(f(j, 1), f(j, 2), f(j, 3))}$$

$$M2 = \sum_{j,a} \left| f(j,a) - \frac{f(j, 1) + f(j, 2) + f(j, 3)}{3} \right|$$

$$M3 = \sum_{j,a} \left[f(j,a) - \frac{f(j, 1) + f(j, 2) + f(j, 3)}{3} \right]^2 \qquad (3.3.6)$$

Consider chromosome 1 of *S. cerevisiae* as an example. It contains 245 positive samples and 236 negative samples. Through calculation of $M1$ $M2$ and $M3$ for each sample we obtain 90% positive samples having $M1 > 0.945$ and 90% negative samples having $M1 \leq 0.945$, 88.6% positive samples having $M2 > 0.355$ and 88.6% negative samples having $M2 \leq 0.355$, and 88.2% positive samples having $M3 > 0.015$ and 87.3% negative samples having $M3 \leq 0.015$. The results show that we can use $M1$, $M2$ and $M3$ as protein coding measure but the accuracy is 90%, 88.6% and 87.8% respectively, lower than IHI about 5 to 7 points. So, IHI is a better parameter to measure the coding potential of sequence. As compared with 'YZ score' method proposed by Zhang and Wang (2000), both two methods use the information on the number of four bases on three codon positions. But the IHI, defined by a nonlinear function of N_{ja}'s, reflects the essence of the difference of coding and non-coding sequences in a simple formula and no learning is necessary in its application. It seems more simple and easily manipulated than YZ score. (In YZ score the linear discriminant equation for ten functions of N_{ja}'s was used and the Fisher's coefficients were determined empirically.) Despite the success of IHI in the classification of coding and noncoding sequences for yeast and lower organisms, its application to higher organisms to determine coding or noncoding sequence are more difficult due to many large introns, small exons, and large intergenic regions. However, in combination with other methods, the generalization of IHI may provide an additional tool for gene recognition. We will discuss the problem further in next section.

Table 3.3-7 The distribution of IHI in first class ORFs and intergenic sequences in yeast genome

First class ORFs

IHI range	ORF number	IHI range	ORF number	IHI range	ORF number
(0 7]	30	(77 84]	125	(147 154]	33
(7 14]	122	(84 91]	124	(154 161]	32
(14 21]	199	(91 98]	103	(161 168]	19
(21 28]	243	(98 105]	85	(168 175]	22
(28 35]	241	(105 112]	89	(175 182]	18
(35 42]	225	(112 119]	74	(182 189]	22
(42 49]	253	(119 126]	62	(189 196]	16
(49 56]	216	(126 133]	47	(196 203]	19
(56 63]	175	(133 140]	46	(203 210]	12
(63 70]	166	(140 147]	39	>210	111
(70 77]	113				

Intergenic sequences

IHI range	Seq. number	IHI range	Seq. number	IHI range	Seq. number
(0 1]	57	(5 6]	432	(10 11]	120
(1 2]	241	(6 7]	350	(11 12]	95
(2 3]	366	(7 8]	278	(12 13]	78
(3 4]	463	(8 9]	202	(13 14]	43
(4 5]	482	(9 10]	169	>14	171

$(a\ b]$ means $a < \text{IHI} \leqslant b$.