科爱传播
KE'AI COMMUNICATIONS
生命科学

·导读版·

Foundations of Comparative Genomics

比较基因组学基础

Arcady R. Mushegian

ELSEVIER
原版引进

科学出版社
www.sciencep.com

FOUNDATIONS OF COMPARATIVE GENOMICS

比较基因组学基础

Arcady R. Mushegian

科 学 出 版 社
北 京

图字:01-2007-3519 号

This is an annotated version of
Foundations of Comparative Genomics
Arcady R. Mushegian
Copyright © 2007, Elsevier Inc.
ISBN-13: 978-0-12-088794-1
ISBN-10: 0-12-088794-0

图书在版编目(CIP)数据

比较基因组学基础＝Foundations of Comparative Genomics:英文/(俄)马希特贾恩(Mushegian, A. R.)编著. —北京:科学出版社,2007
ISBN 978-7-03-019464-0

Ⅰ. 比… Ⅱ. 马… Ⅲ. 基因组-对比研究-英文 Ⅳ. Q343.1

中国版本图书馆 CIP 数据核字(2007)第 113608 号

责任编辑:孙红梅 贾明月/责任印制:钱玉芬/封面设计:耕者设计工作室

科学出版社 出版
北京东黄城根北街16号
邮政编码:100717
http://www.sciencep.com
中国科学院印刷厂 印刷
科学出版社发行 各地新华书店经销
*
2007 年 8 月第 一 版 开本:787×1092 1/16
2007 年 8 月第一次印刷 印张:17 1/2
印数:1—2 000 字数:415 000

定价:75.00 元
(如有印装质量问题,我社负责调换〈科印〉)

导　读

近些年来，基因组学研究技术得到了快速发展。随着全基因组测序不断大规模地进行，生命科学领域的研究也随之翻开了崭新的一页。庞大的基因组数据信息源源不断地从一系列新技术中产生，而且未来还会进一步呈指数倍增长。但海量的信息无法直接促进我们对生命本质的解读，探索对这些信息进行解析的方式已成为这一时代生命科学研究最大的挑战之一。机遇与挑战同在，丰富的信息量、广泛的涉及内容，结合迅猛发展的计算机、网络技术，“生物信息学”孕育而生。比较基因组学是其中最引人瞩目的领域，它通过对不同物种的基因组数据进行比较分析，揭示彼此的相似性和差异性，以了解不同物种间进化上的差异。综合这些信息能进一步帮助我们了解物种形成的机制、基因或基因组上非编码区的功能。进行基因组比较分析时，研究并不仅限于基因编码区，还扩展到对序列相似性的分析、基因位置的比较、基因编码区长度或外显子数的变异、基因组上非编码区的比例、进化关系较远的物种间高度保守区域的比较分析等等(例如从最简单的细菌到非常复杂的人类基因组之间的比较)。

此前已有多部论述生物信息学资源信息(数据库和工具)、分析方法、算法构建等相关内容的学术专著，但是仍然缺乏一本深入浅出地向广大研究生、科技工作者介绍比较基因组学基本原理、方法及研究进展的专著。本书的出版弥补了这一空缺。本书并非旨在提供比较基因组或其他计算生物学相关领域全面的资料信息。尽管进行比较基因组学分析时要面对众多不同的数学模型、方程式和算法，但本书并没有过多涉及数学、统计学和计算机学的问题。这本书主要是想传达科学研究中思考问题的方法，这在从事计算生物学研究尤其是其中最令人瞩目的比较基因组学研究时非常重要。本书还对比较基因组学的研究进行了全景式的介绍，对其发展历史和基本概念进行了梳理。

之前的生物信息学更多地是注重开发数据管理和分析的实用工具，如基因组信息的组织、显示和序列分析。随着我们获得了越来越多的全基因组数据，结合广泛的基因表达、调控和相互作用的信息，研究方法和思路正发生着重大的转变。基因组是高容量的遗传学单位，但是一个基因组比它简单的各部分之和更为复杂，我们必须采用新方法和新思路来进行研究。我们知道，基因或蛋白质不是孤立存在的，它们很少单独起作用，而是倾向于成组地通过网络相互作用来影响生物系统的功能。对基因功能的研究必须分析其相互作用网络，把基因组或蛋白质组看作一个系统来进行分析，实现从以前“一个基因、一种蛋白、一类表型”到“基因相互作用网络”的观念的转化；从过去对各种生物分子进行分析、获取有关生物学知识，转变为综合多种生物分子及其相互作用的知识以了解生命系统的功能。我们进入了一个以综合为特征、研究意义比基因组信息学更为深刻的时期。当前的研究更侧重于从低维到高维进行综合的尝试，通过对一系列生物知识的综合，使我们加深对生命基本规律的认识和理解。

本书作者 Arcady R. Mushegian 是长期从事计算生物学研究的著名学者，积累了丰富的经验，并对相关问题进行了深刻的思考。本书乃至比较基因组学均是基于序列相似

性比较而构建的。正如第 2 章所陈述的:离开序列比较分析,基因组学毫无意义。序列相似性比较分析是确立同源性的关键(第 3 章),例如在病毒全基因组研究中就发挥了至关重要的作用(第 4 章)。目前的生物信息学研究中,无论技术上还是方法上,都还存在着诸多问题和难点。对此,作者没有采取回避或简单化的态度,而是深入探讨了这些问题存在的根源,以及会对研究结果产生的影响。比较基因组学研究中面临的第一种实际情况,在进化过程中,蛋白质序列变异具有的较强的回复性,(第 5 章),使得基于序列相似度比较搜索全基因组信息的方法敏感而有效;而第二种实际情况(第 6 章)正好与之相反,序列和结构比较分析揭示,存在行使相同功能的不同蛋白质,且它们并非源于同一祖先基因。基于全基因组信息的新陈代谢系统重建(第 7、8 章),作者揭示了如何基于同源性和非同源性方法了解基因的功能。全基因组时代可以通过比较更多的分子特征进行系统发育关系重建(第 11、12 章)。此外作者还探讨了未来如何发展基于某些重要特征相似性比较分析的非传统生物信息学和系统生物学研究的问题(第 14 章)。本书全面回顾了比较基因组学发展至今所取得的进展,从学科发展的角度,呈现了从早先基于单个基因的序列比较分析到如今基于全基因组序列数据的基因组所编码的全部基因的分析比较,向读者展示了进行基因及基因组分析研究的方法和思路,并列举了本研究领域先行者们多方面的探索与尝试。

这是一本特别适合那些想了解或进入比较基因组学领域的研究生或科研人员的好书,同时对于从事基因组学领域相关研究的专业人员也是一本具有基础性、前沿性、权威性的参考书。

何静　张亚平
中国科学院昆明动物研究所
遗传资源与进化国家重点实验室
昆明 650223
E-mail:zhangyp1@263. net. cn
zhangyp@mail. kiz. ac. cn

前　言

很多时候，我内心总有一个疑惑，人们是如何选择一种职业作为自己一生的事业的，这真的很奇妙。经过多年实验病毒学，植物、微生物遗传学的研究和生命科学的教学工作，我最终转而从事计算生物学的研究。在此期间，热心于生命科学研究的同时，我同样感兴趣于不同研究者思考课题的不同方式。怎样选择感兴趣的课题进行研究？为什么有一些科学问题看起来有趣且意义重大，而另一些却不呢？为什么在决定什么问题有趣且值得探究时，不同的人会有不一样的观点？

我认识到，当两个过去认为毫无关联的事物忽然间呈现某种意义的相似时（自然科学或其他领域都一样），通常预示着这个问题有趣并值得深入探究。这样的发现，就像发现原先以为完全相同的事物间存在差异一般令人十分欣喜。于是我意识到，我真正感兴趣的是事物间的相似性和差异性，无论是在模式构建或基本组成上。如果这也正是你孜孜以求的，那么计算生物学将是最好的选择。

作为在研究所里少数几个从事生物信息学的专业人士，我更多的时候是与非生物信息学专业的研究者为伍。同僚们时常谈论，相比实际应用或具体操作规程，他们更看重思考科学问题的方法。他们经常会问："我阅读了数据库搜索的相关资料，我想我了解那些数学模型是如何实现数据搜索的，但请告诉我，当序列相似度很低时，你如何确定其中一些较之其他的而言是更具有重要意义的！"看起来针对已经发表的众多生物信息学文章，需要建立一个读者联盟，集中讨论其中存在的个人偏见、偏好或优选性等相关问题。

这也正是我写此书的目的。本书并非旨在提供比较基因组或其他计算生物学相关问题全面的资料信息。书中更多提及的是我所感兴趣的问题：例如本书很大篇幅是关于蛋白质研究的，而核酸序列分析几乎没有涉及。尽管进行生物信息分析时要面对众多不同的数学模型、方程式和算法，但本书并没有过多涉及数学、统计学和计算机学的问题。其实这本书主要是想传达研究中思考问题的方法，这在进行计算生物学研究，尤其是其中最令人注目的比较基因组学研究时非常重要。同时我试图揭示，全基因组测序的大规模开展预示着一个生命科学全新时代的来临，而比较基因组学是其中最闪亮的一颗星。

没有朋友们的支持与合作，这本书是无法完成的，多少年来我在与 Eugene Koonin 和 Alexey Kondrashov 的讨论交流中获得了很大的帮助。Luna Han，我的编辑，在本书撰写过程中给予了很多帮助，帮助我确定本书应该如何进行并保持对的方向。还要感谢 Stowers 医学研究所生物信息学中心的所有成员为本书倾注的无数心血。

最后要感谢我的家人一直以来的宽容、理解和支持。感谢我的妻子 Irina Sorokina，因为有你，我的生命变得美好，还有我最爱的孩子们 Alexandra，Nikolai 和 Natalia。

（何静　张亚平　译）

To my parents

Preface

Many times I would find myself wondering how people choose what to do in their professional lives. After giving years of work to experimental virology, and some more to botany, microbial genetics, and teaching biological sciences, I settled at a research career in computational biology. And all this time, I was interested as much in biology as in the different ways in which people think about their research. How do we decide which problem to study? Why do some scientific questions sound interesting and important, and others do not? And why do different people have different opinions on what is interesting and important?

I noticed that much of what turns out to be interesting and important—in science and elsewhere—happens when two seemingly unrelated things suddenly reveal some sort of similarity. The pleasure of such discovery, of course, is only comparable to the joy of finding a difference between two things that were previously thought to be the same. Thus, I realized that I am interested in similarities and differences, and in patterns and motifs. And if this is what you are after, then computational biology is a good line of work.

As a "local bioinformatics specialist" at my institute, I spend a lot of time talking to the "noncomputational" biologists. My colleagues often tell me that they are more interested in ways to think about science than in actual applications and protocols. Remarks such as "I have read about database search statistics, and I think I understand how this algorithm works—but tell me how you decide which of these weak sequence similarities are more important than the others!" are common. So, it seems that the myriads of bioinformatics texts that are published these days need a reader's companion, which talks about prejudices, preferences, and priorities.

This is my attempt on such a companion. It is not intended as a comprehensive source on genome comparisons or other issues of computational biology. I wrote mostly about things that are of interest to me: For example, most of this book is concerned with the protein world, and there is almost no discussion of nucleotide sequence analysis. There is also very little mathematics, statistics, or computer science in the book, even though the practice of bioinformatics requires dealing with models, equations, and algorithms. Rather, this book is about scientific ideas that I believe to be the most important in computational biology and in its most accomplished branch, comparative genomics. I am also trying to show that the era of completely sequenced genomes is a truly novel age of biology, and that comparative genomics is the science for this age.

This book would not be possible without collaboration, friendship, and, throughout the years, many conversations with Eugene Koonin and Alexey Kondrashov. Luna Han, my editor, helped me to define where this book should be going and gently persuaded me to stay on track, and the members of the Bioinformatics Center at the Stowers Institute for Medical Research held the fort all the while.

Most important, my family put up with everything. I thank my wife Irina Sorokina—all the good things in my life for so many years are because of you, my love—and my children Alexandra, Nikolai, and Natalia.

目　　录

（何静　张亚平　译）

Contents

1

The Beginning of Computational Genomics

Historians of science may disagree about when computational evolutionary genomics started in earnest. Some may associate the starting point with the work of geneticists Alfred Sturtevant and Theodosius Dobzhansky or statistician Sir Ronald Fisher. Others may say that genomics is incomplete without the molecular-level analysis and mark the beginning of the era with the following citation from Francis Crick (1958):

> *Biologists should realize that before long we shall have a subject which might be called "protein taxonomy"—the study of amino acids sequences of proteins of an organism and the comparison of them between species. It can be argued that these sequences are the most delicate expression possible of the phenotype of an organism and that vast amounts of evolutionary information may be hidden away from them.*

However, I believe that most people would agree that several papers published from 1962 to 1965 by Linus Pauling and Emile Zuckerkandl were extremely important. One article in particular, "Molecules as Documents of Evolutionary History" (Zuckerkandl and Pauling, 1965), set the scene for most of the future work that is described in this book. The circumstances of its publication are also of some interest: Although written in 1963, it first appeared in 1964 as a Russian translation in a monograph dedicated to Alexandr Ivanovich Oparin, a true pioneer of experimental study of abiotic protein synthesis (Oparin, 1953) who, sadly, also endorsed and helped enforce Lysenkoist pseudo-science during his service at the Soviet Academy of Sciences from the 1940s to 1960s (Lewontin and Levins, 1976; Jukes, 1997).

The research first announced in that unlikely place (the original English language version of Zuckerkandl and Pauling's paper followed in 1965) sounds prophetic. If we outline the main ideas of that work, the density of novel ideas in that 10-page article is staggering:

1. The authors use the root "semantics" 72 times when speaking of genes and gene products. They called DNA, RNA, and proteins "semantides," or sense-carrying units. Unlike some of the modern uses of this word, which essentially equates semantics with postmodern relativism (e.g., "let us discuss the substance and not argue about semantics"), Pauling and Zuckerkandl took semantics seriously. So should we: By definition (and as understood by their readers in the early 1960s), *semantics* is the study of the meaning of sense-carrying units in a language or in other code. The meaning of words—and of genes—is exactly what we want to know.

2. There are dissimilarities between even closely related sense-carrying molecules. These dissimilarities are produced by genetic processes, such as nucleotide substitutions,

insertions, deletions, and rearrangements of large DNA fragments. Sense, or meaning, of genes and their products may be extracted by comparing related molecules, detecting the differences between them, and computing something about these differences.

3. Biopolymers contain information about evolution. It is threefold: (1) the time of existence of the ancestral molecule,(2) what the sequence was, and(3) the line of descent from the ancestor to each of the contemporary molecules.

4. Some sense-carrying units carry less sense than others. For example, simple biopolymers, build by repetition of a few blocks (nucleotides or amino acids), may not be a good source of information about complex evolutionary processes.

5. Changes in biopolymers may be of different types. Some of the changes are beneficial and favored by selection, whereas others have no phenotype and are "cryptic polymorphisms." One reason why some genetic changes have no phenotype is the degeneracy of genetic code: The same amino acid can be coded by different combinations of nucleotides. Another reason is degeneracy of protein sequence with regard to the three-dimensional structure and, ultimately, to the protein function: The same structure and function can be achieved by different combinations of amino acids. Analysis of these different solutions to the same problem may result in a better understanding of the relationships between genotype and phenotype.

6. Gene mutations and duplications of whole genes may put some genes into a "dormant" state. It is plausible that dormant genes may be reactivated after they accumulate changes, and this reactivation may be an important source of evolutionary novelty.

7. Sequences outside the protein-coding regions may have a regulatory function and may evolve differently than in the coding regions. Other noncoding regions may have no function, and mutations in these regions will be free of selection.

8. Chemical compounds may be synthesized by more than one biochemical pathway. Thus, functional convergence at the molecular level is expected, both at the level of the pathways and at the level of individual biochemical reactions.

Thus, the authors cast evolutionary molecular biology as information science and thought that particular attention should be given to distinguishing signals from noise in the sense-carrying units. Biologists, chemists, engineers, mathematicians, and computer scientists who work on in genome analysis today are in fact implementing the research program that, unbeknownst to some of them, was started by Zuckerkandl and Pauling.

This book is no exception. Nearly every chapter addresses an issue that can be traced back to an idea set forth in Zuckerkandl and Pauling's seminal paper. Chapters 2 and 3 discuss practical approaches to sequence comparison (point 2 as outlined previously). Evolutionary inferences from these comparisons (point 3) and the relationship between signal and noise in sequence comparison (point 4) are discussed in nearly every chapter. The issues of functional convergence (point 8) are of central importance in Chapters 6, 7, and 9. Cryptic polymorphism (point 5) is discussed in Chapters 9 and 10 in connection with sequence–structure–function degeneracy. Finally, "what the ancestors were" (point 3) is the central theme of Chapters 11–13. Even Chapter 14, which deals with genome-wide numerical data, draws inspiration from approaches to comparative sequence analysis foreseen by Pauling and Zuckerkandl.

The techniques of biological sequence comparison were not discussed at any length in "Molecules as Documents of Evolutionary History," but the central goal of *finding pairs of similar sequence fragments* was stated very clearly.

Sequence similarity lies at the heart of all biology, not just comparative genomics. The following statement has even been called "the first fact of biological sequence analysis" by Dan Gusfield (1997) at the University of California at Davis:

> *In biomolecular sequences high sequence similarity usually implies significant functional or structural similarity.*

This "first fact" may qualify as one of the most fundamental facts of our understanding of life. Most biologists, however, would not hesitate to add the following:

In biomolecular sequences, high sequence similarity also usually implies evolutionary relationship.

The two statements, though similar in form, are actually distinct, and in a quite fundamental way. The structure of a biological molecule, such as a protein, is something that can be physically defined. If we have a pure sample of this protein, a quiet place for growing crystals, and a synchrotron beamline, we can determine a structure of a protein molecule, at least in principle. Technical details aside, the same equipment would generally do the job for all proteins. Indeed, as I write this, the challenges of high-throughput protein structure determination are being met by the structural genomics projects (Chandonia and Brenner, 2006). Function, however, is not a physical characteristic but, rather, a description of some process, so function can be defined only in a biological context. At the bare minimum, function of a protein involves interactions with other molecules, which have to be identified and included in the description of function. Often, in order to define the biological function of a sequence, we need to monitor the interactions of many components in a cellular extract, in the whole cell, in a living organism, or in an ecosystem of which this organism is a part. As the protein function is performed, its structure may change. Thus, when we casually say "structure and function," in fact we are talking about many different things already. And the fact that sequence similarity can be used to make inferences about all those different properties of a sense-carrying unit—from physical properties of the molecule to its relationships with its environment—is not at all trivial. The "second fact" is also nontrivial: Unlike more or less directly observable structural and functional properties, the common ancestor of two molecules cannot be directly observed (with the exception of rare cases in which the ancestral DNA or protein have survived in ancient proteins or in biopsies), and yet we do not hesitate to infer such an ancestor from the sequence similarity.

Thus, on the basis of sequence similarity, we make conclusions about (1) similar structure, (2) similar function, and (3) common ancestry. These inferences are at the heart of computational biology; most biologists make them every day, and almost every theme in this book is based on such inferences. But how do we make them in practice?

At first glance, the statements about structure and function seem to follow from sequence similarity quite naturally. And without doubt, these statements are amenable to direct experimental corroboration. But in fact, structural and functional inference is inseparable from evolutionary inference. Indeed, when comparing sequences of two biopolymers, our path from sequence similarity to the conclusion about structural or functional similarity is never direct. Instead, we always infer common ancestry of these sequences first, and only from there can we proceed to making structural and functional inferences. This logic is not obvious when the similarity is very high, but if the two sequences are more distantly related to each other (as is the case with most sequence comparisons today), this chain of thought becomes explicit. Indeed, we measure similarity between sequences and immediately use statistics to compare the observed similarity with what would be expected by chance (discussed in Chapter 2). If the similarity is too high to occur by chance, this is usually sufficient for making predictions about protein function (discussed in Chapters 5–8) and structure (see Chapter 9). But the only reason why such reasoning works is because the only way for nonrandom sequence similarity to occur is by descent from a common ancestor of the two sequences. This is the homology inference (see Chapter 3). Thus, the inference of evolutionary relationship, which seems to be the least observable of all, turns out to be a prerequisite of proposing other, directly observable, relationships, such as similarity of structure and function.

Consider the alignment of three sequences, A′, A″, and A‴ (here and elsewhere in this book, I use capital letters in regular font to indicate genes and italicized capitals to indicate species in which these genes are found). Suppose that three sequences come from three different species, one from each, and only the function of A′ has been studied. Suppose that A′ and A″ are almost identical, and the third sequence, A‴, is less similar but still quite close to A′ and A″. Do we use the same information to infer common ancestry and common function of all these sequences? It seems that we do not really need every amino acid residue that is conserved between A′ and A″ to determine that they share a common ancestor; for example, we may not care about the sites conserved exclusively between A′ and A″ because we do not need these residues in order to recognize similarity between A′ and more distantly related A‴, as well as between A″ and A‴. On the other hand, when we are making the inference, "closely related A′ and A″ are more likely to have the same function, but a more distant A‴ may have different function," we, in effect, are using the information about the sites conserved exclusively between A′ and A″ but not between each of them and A‴. Thus, evolutionary, structural, and functional information is intertwined in sequence in subtle ways.

The reverse of the "first fact of sequence analysis" is not true: Functionally similar proteins do not have to have similar sequences, and proteins with similar structures also may have dissimilar sequences (this is discussed in much more detail in Chapters 6 and 9). Neither is the reverse of the "second fact" true: There may be an evolutionary connection between two sequences, but, if these sequences have diverged too far, the sequence similarity between them may not be discernible from the random-level similarity (this is discussed in more detail in Chapter 2). Note that in the case of the "reverse-second" fact, we are dealing with a relationship that still exists, even if the sequence similarity has already blended with the noise. The "reverse-first" fact, however, is more dramatic. Functionally similar proteins may have had lost sequence similarity, but, on the other hand, they may have never shared sequence similarity but converged to the same function from completely different, evolutionarily unrelated sequences. This principle applies to structures as well: Similarity of structures in the absence of sequence similarity may represent either extreme divergence of initially similar sequences or convergence of sequences that were not similar in the first place (discussed in Chapters 6, 9, and 10). Distinguishing between divergence and convergence at the molecular level is one of the most important problems of computational biology.

All these considerations are different facets of the most important postulate of Pauling and Zuckerkandl: Biopolymers contain information about their evolution, structure, and function, and these three types of signals may interact in different ways, sometimes enhancing and in other cases interfering with each other. In a sense, whole biology for the past few decades has been dominated by the quest for ways to extract and analyze signals contained in molecular sequences. Genomics is a continuation of these efforts for our times, when complete genetic makeups of many species are known. At the same time, genomics offers even more. Many times in this book, I will return to the argument that with complete genome sequences, we can answer many questions that we could not answer, or even could not think of asking, before. This is the new era in biology—the era of complete genomes.

Sequences of genes, genomes, and proteins are not the only kinds of data that are of interest to genomics. New technologies allow us to collect information about the occurrence and spatial organization of genes and regulatory sequences; the concentration of different molecules in cells, organs, and biological samples (measurement of mRNA levels, collected with the help of gene expression arrays, is the most famous, but by no means unique, example of this class of data); cellular morphology and physiological responses; and so on. This information often takes the form of rows and columns of numbers. It may seem that Zuckerkandl and Pauling did not have much to say about these data, which were not in the form of sense-carrying units anyway. But in Chapter 14, I argue that the analysis of these genomewide measurements also owes a lot to our experience in sequence comparison.

2

Finding Sequence Similarities

As discussed in Chapter 1, Pauling and Zuckerkandl in their seminal work outlined the research program of studying the complicated ways in which structural, functional, and evolutionary information is convoluted within a molecular sequence. It was clear to them that the comparison of sequences is a clue to uncovering all these types of information. Paraphrasing the famous quote from Theodosius Dobzhansky (1973), almost nothing in computational biology makes any sense except in light of sequence comparison.

Before the deciphering of genetic code and the advent of DNA cloning, the most common order of business in protein science was to isolate a protein, study its biological properties, and only then, motivated by its biological importance, attempt to sequence this protein using rather inefficient methods of direct peptide sequencing. The accumulation of novel protein sequences in those times was slow and deliberate. Even when methods of DNA cloning and sequencing came about in the late 1970s, they were applied mostly to one protein at a time, also guided by biological interest in the gene or its product or, in many cases, by the ease with which a gene could be isolated. Thus, proteins and mRNAs that were abundant or homogeneous, such as cytochrome C homologs, immunoglobulins, or virus capsid proteins, were studied at the sequence level much earlier than other families of proteins. And the biological, biochemical, and other properties of proteins usually were quite well studied by the time the sequence was determined.

But what about evolutionary relationships—how can we infer the common ancestry of the "sense-carrying units" without knowing their sequences? In fact, we can do it just fine in many cases. For example, the favorite subjects of comparative evolutionary biochemistry for most of the 20th century were globins, the main protein constituents of vertebrate red blood cells. Years of work in the lab have shown similarity of many physicochemical and biological properties of globins. At the same time, the anatomical, histological, and biochemical similarity of most components of vertebrate blood and circulatory systems was demonstrated. Altogether, this was the overwhelming evidence of common origin of globin genes and their protein products. In this context, sequencing of globins could be perceived more as a confirmation of the phylogenetic hypothesis than a way of proposing their common origin in the first place. Here again, Pauling and Zukerkandl were ahead of their time when they emphasized that sequences of biopolymers are the real foundation for comparing all of their other properties, and that phylogenetic hypotheses may be put forward on the basis of sequence analysis alone, before inferring other shared properties of genes and proteins. This is a dramatic shift in the way we look at genetic information.

Pauling and Zuckerkandl did not discuss at any length how exactly we should compare sequences and how to measure the strength of signals that this comparison may provide. This was an algorithmic problem in the area of pattern matching, and solving it required the help of mathematicians, computational scientists, and statisticians.

Sequence comparison, particularly the crucial role played in it by one class of algorithms, namely dynamic programming, is discussed in almost every book on computational biology and bioinformatics. David Sankoff was one of the most important figures in the field, and reviewed the early work in a short, vivid paper (Sankoff, 2000). Other reviews can be found in Mount (2004), which is also one of the most detailed introductions to the mechanics of database search and sequence alignments, and in Jones and Pevzner (2004). Succinct primers on dynamic programming and other basic elements of sequence analysis (e.g., substitution matrices and hidden Markov models) can be found in notes by Sean Eddy (2004a–2004d), a thorough review of combinatorial and algorithmic aspects of sequence analysis is provided in Gusfield (1997), and the best introduction to the probabilistic aspects of the same is the book by Durbin *et al.* (1998). Finally, the redoubtable family of BLAST programs has been thoroughly covered in a corpus of work by Steven Altschul (Karlin and Altschul, 1990; Altschul, 1991; Altschul and Gish, 1996; Schaffer *et al.*, 2001; Altschul *et al.*, 1990, 2001, 2005). Newer programs suitable for the era of complete genome sequencing, assembly, and multigenome alignment are discussed in Miller (2001), Kent and Haussler (2001), Schwartz *et al.* (2003), Blanchette *et al.* (2004), and Ovcharenko *et al.* (2005).

My goal in this chapter is not to repeat what is written in these excellent books and articles. Rather, I present five challenges of biological sequence analysis that receive relatively little attention but can make a major difference in sequence analysis, and I try to show how some of the well-known sequence comparison approaches address these challenges. In dealing with these concerns, I mostly talk about protein molecules, which, of course, are sequences of amino acid characters drawn, in the first approximation, from the 20-letter alphabet. I only briefly mention comparison of nucleotide sequences, which consist of four nucleotide characters, and other types of comparisons, such as comparison of gene orders in different genomes, when the alphabet may include hundreds or thousands of characters.

Challenge 1. The methods of sequence alignment are often classified into "local" or "global" methods, or, more accurately, into methods that produce local or global alignments. (In a global alignment, each character is forced to be aligned with something, and in a local alignment some characters are not considered. Many special cases of alignment can be given more rigorous definition; Gusfield, 1997.) In one sense, this distinction is important because statistics of local alignments is well-defined, which is not the case for global alignments (Altschul, 2006). In a different sense, this distinction is a red herring because the goal of comparative sequence analysis is really not "to construct an *alignment*." Rather, the objective is to find evolutionary, functional, and structural signals in biological sense-carrying units—the signals that, as discussed in Chapter 1, are revealed by sequence similarity. Thus, algorithms may be set up to produce either local or global *alignment*, whereas in fact the most important question is whether the *similarity* between sequences is global or local.

Challenge 2. Each method of sequence alignment tries to find an extremum of some value, such as the minimal number of operations required to convert one sequence into another or the maximal matching score (which is most commonly sought and which will mostly concern us in this chapter). This solves an optimization problem but may not do much to solve a biological problem (i.e., to find signals in sense-carrying units). Biological knowledge enters into the picture by way of the scoring function, which is the way of measuring similarities/differences between sequences. For example, if we thought that 4 amino acid residues represented by vowels of the Latin alphabet (A, E, I, and Y) are less important in proteins than the other 16 residues, and decided to only consider matches between the latter

16, any alignment algorithm would work with such a scoring system without complain—even though the idea is absurd on its face. All improvements in sensitivity of sequence analysis are in fact the improvements in measuring similarity between sequences—from less sensitive to more sensitive substitution matrices and then to probabilistic models of multiple sequence alignments. The theory of similarity/distance between sense-carrying units, however, is in its infancy, notwithstanding some important insights (see Altschul, 1991; Zharkikh, 1994).

Challenge 3. Sequence alignment algorithms, even when provided with good scoring schemes, will align any strings of allowed symbols and produce the highest scoring match between any two sequences, whether they contain biological signals or not. But these algorithms will not tell whether this highest match is "high enough" to indicate the presence of a signal we are looking for. To pick out matches that represent biologically important signals, one needs a statistical theory that evaluates alignments and compares them to some kind of a standard. Such theory is available in an exact form for ungapped alignments (Karlin and Altschul, 1990; Altschul, 2006) and in an approximate, yet apparently quite accurate, form for alignments with gaps (Mott, 2000). But even with this theory in hand, and with good scoring schemes, there are many alignments that remain in the "twilight zone" of borderline statistical significance and cannot be directly used to infer the presence of a biological signal. The problem of how to validate (or reject) the alignments in the twilight zone is still not fully solved.

Challenge 4. Related to challenges 2 and 3 is the problem of nontransitivity of sequence similarity scores. The simplest way to state nontransitivity is for the case of three sequences: If sequences A and B can be matched (aligned) with a high score, and sequences B and C can also be matched with a high score, this does not tell us anything about the score between A and C. That score can also be high according to our statistical theory or it can be low—so low as to be indistinguishable from the noise. In the context of the database searches, most matches indistinguishable from the noise are not reported to the investigator, so we may not know about similarity between A and C unless we first know about similarity between A and B. Of course, we can increase sensitivity of sequence comparison, for example, by replacing a single-sequence query by a probabilistic model of a protein family to which this sequence belongs or by aligning two family models instead of two representative sequences. This will pull some of the twilight zone similarities into the high-similarity zone (i.e., some "sequences C" will become directly linked to A), but other sequences and sequence families may remain low scoring with regard to some query A yet pass the significance threshold with a query B that itself is high scoring with regard to A. This nontransitivity problem is not fully solved in any method of sequence comparison.

Challenge 5. Any textbook on bioinformatics will discuss differences between pairwise alignments and multiple alignments. It is important to know what these differences are: For example, some of the theory that is worked out in considerable detail for the case of two sequences cannot be easily generalized to multiple alignments, and some alignment methods that have acceptable speed of execution on two sequences are computationally prohibitive when many sequences are involved. But there is another distinction, which is sometimes overlooked; this distinction is between different types of pairwise alignments. Indeed, we may use methods of pairwise alignment as a tool for discovering similarity that was not known before, but we also can apply alignment methods to study similarity between sequences that are already known to be related. The first type of pairwise alignment, in principle, does not have to be biologically optimal: Arguably, it has to score just high enough to stand out from the background. At the same time, this "type I" alignment has to be arrived at with high efficiency, because discovery of sequence similarity is typically done in the context of database searches, in which a query sequence is matched to all, or at