王槐春　编著

蛋白质与核酸序列分析基础

人民军医出版社

蛋白质与核酸序列分析基础

Essentials of Sequence Analysis

王槐春　编著

人民军医出版社

1994 · 北京 ·

内容提要

蛋白质与核酸序列的计算机分析是分子生物学研究的一项新技术，并逐渐形成一门由分子生物学和计算机信息处理技术相结合的交叉学科——生物信息学。本书介绍了蛋白质和核酸数据库，各种基本的序列分析方法和应用程序，并用大量的图表和实例加以说明，力图使不熟悉计算机的分子生物学家容易接受。本书可供生物化学和分子生物学科研人员及生物技术软件设计人员参考，也可作为大学生物物理、生物化学专业的研究生和高年级本科生的参考书。

责任编辑　苗　芳

*

图书在版编目(CIP)数据

蛋白质与核酸序列分析基础／王槐春编著.—北京:人民军医出版社,1994.12

ISBN 7-80020-424-3

Ⅰ.蛋… Ⅱ.王… Ⅲ.①蛋白质-序列-分析方法②核酸-序列-分析方法 Ⅳ.①Q51-34②Q52-34

中国版本图书馆 CIP 数据核字(94)第 11968 号

人民军医出版社

(北京复兴路 22 号甲 3 号　邮政编码: 100842)

中国人民解放军第 1201 工厂印刷

新华书店总店科技发行所发行

开本: 850×1168mm 1／32 印张:8 字数 204 千字

1994 年 10 月第 1 版　1994 年 10 月第 1 次印刷

印数:1—2 000 册　定价:10.00 元

ISBN 7-80020-424-3／R·365

PREFACE

No aspect of modern biology has provided more understanding about life on Earth than the determination of macromolecular sequences. Indeed, the past forty years have witnessed unparalleled advances in our understanding of not only how organisms are related, but, more to the point, how they have come to possess the protein equipment with which they face the world. Thus, the development of life on Earth has largely been the result of a vast succession of gene duplications followed by waves of smaller scale changes in the form of base replacements. The reconstruction of these past events is being made possible by the computer analysis of macromolecular sequences. Such analyses demand not only modern computers, programs and databases, but also informed biologists. It is in the last-named realm that this book makes its important contribution. It is not that the technical aspects of this field are so difficult. Rather, it is simply that they are not familiar to most molecular biologists. It is an area that requires some elementary statistics, on the one hand, and some general background in macromolecular structure, on the other. The author has provided both components in a simple and straightforward way.

It is unusual to be asked to write a preface for a book written in a language that one cannot read. In the case at hand, I have only been able to read a translation of an outline of the contents, but it has been sufficient to convince me that this is a timely and up-to-date volume. Huaichun Wang has gathered together an

impressive assortment of topics. Chinese molecular biologists are fortunate indeed to have these very recent methods gathered into a single volume in their own language. The rapid pace of the biological sciences demands that we all have a basic understanding of these phenomena.

Russell F. Doolittle

序

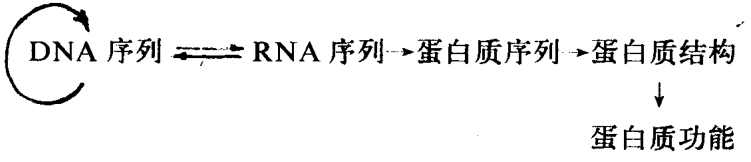
现代生物学的任何一个方面都不能象确定大分子序列那样能提供给我们更多的有关地球上生命的知识。确实，在过去的 40 年里，我们已经目睹了人类不仅在认识生物体是如何相关的，而且，更确切地说，在了解它们是何以获得使它们面对这个世界的蛋白质方面，已经取得了前所未有的进展。因此，地球上生命的发生大体上是一系列的基因复制伴随着以碱基替换为形式的较小规模的改变。可以用计算机对大分子序列的分析来重建这些过去的事件。这样的分析要求不仅有现代的计算机设备、程序和数据库，而且要有在行的生物学家。正是在后一个意义上，本书作出了它的重要贡献。问题不在于本领域的技术方面是如何的困难，而只是多数分子生物学家对它们不熟悉。它是这样的一个领域，一方面需要一些基本的统计学方法，另一方面需要某些大分子结构的背景知识。本书作者以简单和直接的方式提供了这两方面的知识。

要求一个人为一本用他所不懂的语言写作的书来写序言是不寻常的。目前在手边的，我仅能阅读本书目录概要的翻译，但是它足以使我相信这是一本及时的和最新的卷本。王槐春把各类有趣的论题综合在一起。中国的分子生物学家确实是幸运的，能读到一本用他们自己的语言写作的收集了最新方法的书。生命科学的快速发展需要我们都要具备对这些现象的基本认识。

R.F.杜利特尔
加州大学圣迭戈分校
生物学和化学教授

前　言

运用计算机进行蛋白质和核酸的序列分析是分子生物学研究的一个较新的发展，该项技术已越来越多地用于研究新测定的生物大分子序列或大量积累的序列数据。今天可以说，没有一篇发表关于序列的分子生物学研究论文不具有诸如此类的语句：经检索核酸或蛋白质序列数据库，“该序列与数据库中的某某序列同源”，或者“该序列与序列库中的任何序列未见明显类似”，等等。然而，序列分析不单是序列的类似性检索或比较，它还包含很多内容：二级结构预测、亲疏水性分析、序列模式识别、结构域或编码区识别以及蛋白质家族分类和进化树的构建。因此，大分子序列数据是分子生物学研究的一个重要资源，序列分析已成为当代分子生物学家必须掌握的技术。事实上，有关序列分析方法及其应用的研究已形成一个新的交叉学科——生物信息学(bioinformatics)或计算分子生物学（computational molecular biology)。它的基本出发点是通过计算方法由 DNA 和蛋白质的序列推导出它们的结构和功能：



序列分析不同于常规的分子生物学方法（如蛋白质分离纯化、分子克隆和序列测定），它要求研究者具备两方面的知识，其一是基本的数学和统计学方法以及计算机程序应用的知识，其二是分子生物学特别是生物大分子结构的背景知识。由于这第一个因素，许多分子生物学家望而生畏；由于第二个因素，专业计算机软件人员和数学、统计学家难以入门，从而阻碍了序列分析程序的开发和方法的应用。在我国，虽然已经有许多有关分子生

物学实验技术的书籍，但是迄今为止，尚未见到一本关于核酸和蛋白质序列的计算机分析的专著。为了填补这一空白，本人在多年从事生物大分子数据库和软件开发的基础上，编写了这本《蛋白质与核酸序列分析基础》，期望能促进分子生物学家熟悉和掌握序列分析技术，也为计算机软件人员开发分子生物学软件提供指南。

全书共分七章。第 1 章介绍序列数据库，第 2 和第 3 章介绍序列对数据库的类似性检索、序列比较和同源性构建，第 4 章描述蛋白质家族、分子钟和进化树的概念和研究方法，第 5 和第 6 章分别介绍蛋白质和核酸的信号检索、结构预测和功能分析，第 7 章介绍利用电子网络(Internet)进行序列分析。本书的最后有 9 个附录，它们收集了许多有用的信息，包括核苷酸碱基和氨基酸的标准符号、密码子字典、序列数据库和序列分析软件、计算机硬件的选择等。

本书力图从一个生物学家的角度来描述各种基本的序列分析方法和程序，并提供尽量多的图表和实例，而不是过多地去叙述算法和公式。其目的是使读者知道如何分析一个新的 DNA 或蛋白质序列，如何利用现有的数据库，有哪些程序和方法可以使用，它们能得到哪些结论，这些方法还有哪些限制或不足之处。毫不讳言，本书在较大程度上类似于 Doolittle1987 年出版的《OF URFS AND ORFS – A Primer on How to Analyze Derived Amino Acid Sequences》和 Gribskov 与 Devereux1990 年发表的《Sequence Analysis Primer》，因为它们与本书一样都是序列分析入门的书籍。但是本书也包含了一些新近发展起来的方法，如氨基酸组成分析及其用于纯化蛋白质的鉴定、蛋白质稳定性预测、寡核苷酸引物和探针的检索；另外，本书尽可能地收进我国学者在序列分析方面所开展的工作。

我要特别感谢美国加州大学 R. F. Doolittle 教授为本书所写的序言，以及允许我使用他的上述著作中的图表资料。作为序列

数据库和序列分析研究的先驱者之一，他用序列比较方法发现了许多序列同源和进化关系，特别是 1983 年他发现了病毒癌基因 v-sis 产物与人血小板源性生长因子同源。他于 1990 年为著名的《酶学方法》系列丛书编辑了第 183 卷《分子进化：蛋白质与核酸序列的计算机分析》。读者将从本书中了解到 Doolittle 的更多的贡献。我还要感谢美国 Brookhaven 国家实验室的 T.Koetzle、美国 NIH 全国生物技术信息中心的 D.Benson、瑞士日内瓦大学的 A.Bairoch、日本东京理科大学的 A.Tsugita 和意大利热那亚国家癌症研究所的 P.Romano 等向我提供了 PDB、Entrez: Sequences、EMBL、ATLAS 和 MPDB 等核酸和蛋白质数据库。军事医学科学院基础医学研究所唐佩弦研究员为本书题写书名，该院情报研究所廖应昌、缪其宏研究员和中国科学院生物物理研究所徐军同志对本书的出版给予很多帮助，在此一并表示感谢。

王槐春

1994 年 4 月 30 日

目　录

Content