现代生物技术前沿

INTRODUCTION TO COMPUTATIONAL MOLECULAR BIOLOGY

〔巴西〕J.塞图宝
J.梅丹尼斯 著

计算分子生物学导论

朱浩等译

科学出版社
www.sciencep.com

现代生物技术前沿

INTRODUCTION TO COMPUTATIONAL MOLECULAR BIOLOGY

〔巴西〕J.塞图宝
J.梅丹尼斯 著

计算分子生物学导论

朱浩等译

科学出版社
北京

图字:01-2003-0482

内 容 简 介

分子生物学的迅速发展,产生了大量的数据,而其间的关系也日趋复杂。计算分子生物学就是处理这些数据的新学科。

本书主要是介绍分子生物学中具有代表性的计算问题以及某些求解这些问题的有效算法。具体包括:分子生物学的基本概念;两个重要的数学对象,即串和图,和算法的基本概念等;序列比较和经典的动态程序设计算法;DNA片段组装技术;DNA的物理作图问题和一种物理作图的近似算法及启发式;与种系发生树构造有关的一些数学问题和某些用于构造特定类型种系发生树的算法;用以研究DNA中序列差异的数学模型和用于RNA结构预测的动态程序设计法以及蛋白质比配方法;最后还介绍了DNA计算。

本书可供研究基因组学与分子生物学的生物学、数学、计算机科学等专业的科研人员、教师、研究生等参考。

João Setubal, João Meidanis

Introduction to Computational Molecular Biology. 1st ed.

ISBN 0-534-95262-3

图书在版编目(CIP)数据

计算分子生物学导论 /(巴西)塞图宝,(巴西)梅丹尼斯著;朱浩等译. —北京:科学出版社,2003.8

(现代生物技术前沿)

ISBN 7-03-011493-0

Ⅰ.计… Ⅱ.①塞…②梅…③朱… Ⅲ.分子生物学-计算方法 Ⅳ.Q7

中国版本图书馆CIP数据核字(2003)第042257号

策划编辑:周辉 王静/文案编辑:彭克里 吴慧涵/责任校对:朱光光

责任印制:刘士平/封面设计:王浩

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

源海印刷有限责任公司 印刷

科学出版社发行 各地新华书店经销

*

2003年8月第 一 版 开本:787×1092 1/16

2003年8月第一次印刷 印张:15

印数:1—2 000 字数:339 000

定价:36.00元

(如有印装质量问题,我社负责调换〈新欣〉)

译者序

计算生物学(包括生物信息学)是生物科学与技术在世纪之交迅猛发展的助推器。可以预见它将会发挥越来越广泛和重要的作用。如何在国内推动这个交叉学科的健康发展是一个紧要的问题。我们以为,好书所起的作用是不可低估的。在我们所接触过的书中,这一本书篇幅适中,内容全面,行文严谨,取材经典,是一本值得介绍的好书。特别要说明的是,原书的两位作者均是计算机科学家,读者也可从本书的行文风格体察得到。当他们多年前涉足计算生物学这条当时的"偏径"时,恐怕并不曾想到该学科今天能有这般红火。计算机和数学专业人员的参与推动了计算生物学近年在国外的迅速发展,我们希望国内也有越来越多的"外行"参与进来。

翻译是难以完美的,我们的原则首先是忠实原著,然后是力求通顺。术语的命名和索引的使用非常重要,我们推敲了所有术语的命名并重建了中文索引,希望该索引有助于读者的进一步阅读。需要说明的是,原文的索引并不全面,未收录术语的所有出处,我们并未对其作任何修改;其次,一些术语的译名与时下在其他书中采用的译法或用法不完全一致;最后,原书有一个勘误表,我们已根据勘误表校正了有关译文。对译文中可能存在的不当,我们事先请读者包涵。我们建议最好是阅读原文,而更好的办法则是:投身到计算生物学中来。

本书既可作为计算机专业、数学专业、生物医学工程专业高年级学生和研究生的教材,也可作为医学、生物学学生和工作者的参考书。本书基本上是纲领性质的,所呈示的内容是计算分子生物学的骨架。在弄懂了本书内容后,读者应能顺利阅读其他计算分子生物学和生物信息学文献。我们认为,尽管该领域发展很快,但大部分文献是关于应用的,因此从原理上弄通主要的脉络将大有裨益。

第一军医大学赵永忠博士翻译了部分初稿,国防科技大学计算机学院周斌博士、新加坡生物信息学研究所李国彬博士分别校对了第 2 章和第 3 章的译稿,在此特别致谢。同时感谢科学出版社和本书的责任编辑为本书的出版所付出的辛勤努力。

译校者　朱　浩

2002 年 12 月于新加坡生物信息学研究所

审校者　赵克森

2003 年 6 月于广州第一军医大学

前　言

生物学中有着至少500年也解决不完的有趣问题。

——Donnald E. Knuth

自从1953年DNA结构被揭示以来，分子生物学取得了巨大的进展。随着我们对生物大分子序列操纵能力的增强，科研工作已经产生并仍在产生大量的数据。处理来自全世界不同实验室所产生的大量数据，并使其可用于进一步的科研，产生了全新的在本质上具有多学科交叉性质的问题。生物科学家是这些数据的创造者和最终的用户。然而，由于数据的规模和复杂性，从创造到使用这些数据都需要多学科特别是数学和计算机科学的参与。这种要求产生了一个新的领域，通常叫做计算分子生物学。

广义地讲，计算分子生物学包括开发和使用数学与计算机技术，以帮助解决分子生物学中的问题。使用若干例子有助于说明。产生的所有信息都是以数据库的形式来存储的。国际上已存在好几个序列数据库，但生物科学家已认识到针对分子生物学的特殊需求，人们需要有新的数据库模型。例如，随着研究工作的进行，这些数据库应能记下我们对分子序列理解的变化，而目前的数据库尚不能满足这个要求。接下来，理解分子序列需要新的复杂的模式识别技术，而这种技术正是人工智能目前研究的课题。在数据库搜索过程中还遇到复杂的统计问题，这也需要新的专门工具。

而有一类问题最需要的是有效算法。简单地说，一个算法就是一个一步一步的过程，它能在限定时间内完成求解一个严格定义的问题。出于效率，算法在求解问题时不能花费“过长”的时间，即便问题相当的大。分子生物学中一个可以用算法求解的经典问题是序列比较：给定两个生物分子序列，我们要知道它们的相似性是多少。这样的问题每天要处理上千次，因此必须使用非常有效的算法。

本书的目的是介绍分子生物学中具有代表性的计算问题以及某些求解这些问题的有效算法。一些问题已相当清楚，某些算法也应用多年。另一些问题则困难些，尚未建立起满意的算法。对后一类问题，我们专注于阐释某些数学模型，这些模型可作为将来算法发展的基础。

读者应当知道，解决分子生物学问题的算法颇为奇特，它需要满足两类用户：分子生物学家和计算机科学家。前者要求算法切用，即要能求解带各种误差和不确定性的实际问题；后者则注重算法有效性的证明，证明它能有效地求解一个严格定义的问题，并愿意

为可证明性而牺牲切用性。我们试图在相互矛盾的二者之间寻求平衡，但时常倾向于后者，因为我们毕竟是计算机科学家。然而，我们希望本书对分子生物学家和计算科学家都有益。

这本书是个导论，这意味着我们的指导原则之一是尽可能展示我们认为简单的算法。对于书中的某些问题，存在着更有效亦更复杂的算法，有关这些算法的线索通常在每一章最后的文献提要中给出。尽管秉持简单性原则，本书中的一些算法和模型仍可能被认为是复杂的，这通常反映了对应问题所固有的复杂性，我们通常在标题前用“*”号将这些内容标出。这本书的导论性质还意味着对于某些课题，我们的介绍仅是个起点，这些课题完全可能需要一整本书的内容来讨论。

这本书的主要读者是数学和计算科学的学生。我们假定他们仅具有中学水平的分子生物学知识，因而提供专门一章简要介绍本书中用到的基本概念。对分子生物学不熟悉的读者应尽可能涉猎超出本书的内容，拓展他们的生物学知识。有关书籍在第一章的末尾给出。

我们希望本书在一定程度上也对生物科学学生有益，并假定这类读者在大学受过离散数学与算法方面的训练。为了帮助不熟悉这些内容的读者，我们安排了一章来简要介绍本书中用到的基本概念。

计算分子生物学进展迅速，更好的算法层出不穷，新的领域不断涌现。我们尽可能地覆盖主要领域，并认为大多数的材料具有永久价值。为满足想进一步研究的读者的需求，我们还提供了对若干信息源的索引（特别是在最后一章的文献提要，包括有用的网站）。当然，这些提示不是全面的。此外，我们不能保证所提供的网络 URL 仍然有效。尽管已测试过所有这些地址，但由于网络的动态特性，它们可能会有变动。

全书概述

第 1 章介绍分子生物学的基本概念，主要包括蛋白质和核酸的基本结构和功能，分子遗传学的机制，研究生物基因组的主要实验室技术，以及现有序列数据库的概述。

第 2 章主要讲述串和图，它们是本书用到的两个最重要的数学对象。同时简短叙述了算法的基本概念，如何分析算法，NP-完全性的定义。

后续的章节主要针对分子生物学的特殊问题。第 3 章是序列比较，研究了基本的两序列比较并给出经典的动态程序设计算法。我们接着扩展了该算法，使之可用于该问题更一般的情形，其中一个是多序列比较，另一个是数据库搜索。我们亦讨论了有关的杂项问题。

第 4 章是片段组装。当 DNA 序列被打碎成小段时便引发了该问题，这些小段需要重新组装成原来的片段。这个技术被广泛应用于大规模测序项目，如人类基因组计划。我们揭示了各种因素可使该问题非常复杂。接着，介绍了几种简化的模型，并在最后部分给出了基于这些模型的算法和启发式。

第 5 章是 DNA 的物理标图问题，它可看做是更大尺度上的片段组装问题。由于片段更大，组装技术也完全不同，其目标是获得 DNA 分子上某些标记的位点。我们对主要

技术作了回顾并给出了模型,接着描述了连续 1 问题的算法,该抽象问题在物理标图中起重要作用。本章结尾部分讨论了一种物理标图的近似算法和启发式。

蛋白质和核酸随年代而进化,理解种系如何发生的一个重要工具是种系发生树,种系发生树亦能帮助理解蛋白质功能。第 6 章描述与种系发生树构造有关的一些数学问题,并给出某些用于构造特定类型种系发生树的算法。

计算生物学中一个新近出现的重要研究领域是基因重排。已经发现一些生物具有遗传差异,它们并非在序列水平有大的差异,而是大的序列块在各自 DNA 中的排序不同。已有一些有趣的数学模型用于研究这种差异。第 7 章专门论述这个问题。

理解分子的生物功能是计算生物学的核心问题。因为分子在三维空间里折叠,且因为其功能取决于这种折叠的方式,在过去几十年里科学家关心的一个主要议题是发现这些大分子的三维结构,特别是 RNA 和蛋白质的三维结构,这孕育出基于分子原始序列预测其结构的方法。第 8 章描述用于 RNA 结构预测的动态程序设计方法,概述了蛋白质结构预测的困难性,并介绍了该领域一个重要的近期进展,即蛋白质比配,它试图把蛋白质的序列和已知的结构作比较。

结束本书的第 9 章介绍一个令人兴奋的新领域:DNA 计算。我们介绍了能使用 DNA 分子求解困难问题的生物学实验,并在理论上探讨了对另一个难题的求解。

我们已给出的一个约定是,标题带"*"的内容是作者认为较有难度的内容。在概念定义中,我们用到另一个约定,即第一次出现的全书均用到的术语以黑体表示,其他术语在定义中用斜体表示(译者注:译著中以楷体表示)。大多数算法首先以英语叙述,然后以伪代码描述(伪代码约定在第二章第二节列出)。在部分例子中,伪代码已相当具体,有助于感兴趣的读者将其实现。

较长的章节给出了小结。

习 题

每一章的结尾有习题。标记星号"*"的习题难度较大,可能需要超出本书的计算机知识,但一天之内应能求解。标记"**"的习题曾是研究课题,但现已解决,在文献提要中给出了解决该习题的原始文献。最后,标记"◇"的习题据作者所知到目前为止尚未解决。

书的最后我们给出部分习题的答案或提示。

错 误

尽管作者已尽其所能,错误仍在所难免。如果你发现错误,或有任何改进之建议,请不吝赐教。请将错误报告或其他评论寄至 bio@dcc.unicamp.br 或

J.Meidanis/J.C.Setubal
Instituto de Computacao, C.P.6176
UNICAMP
Campinas,SP 13083-970

Brazil
作者的私人电子邮箱是：meidanis@dcc.unicamp.br 或 setubal@dcc.unicamp.br。我们衷心感谢感兴趣的读者提出改正意见。查出的错误将在下面的网址公布：http://www.dcc.unicamp.br/~bio/ICMB.html。

致　谢

本书是作者原先以葡萄牙文写于 1994 年并在巴西出版的第一部拙作的续篇。第一本书的出版要归功于两年一度的巴西计算机科学会议“Escola de Computacao”，如果没有这个会议，我们就根本没有机会在此写序。

本书的诞生要感谢 Mike Sugarman、Bonnie Berger 和 Tom Leighton，他们给予我们很多鼓励和有帮助的建议。特别是 Bonnie，她甚至把她早期阶段的讲义拷贝给我们。

我们有幸得到来自 FAPESP 和 CNPq（巴西研究局）的资助，他们给我们多方面的帮助。FAPESP 基金用于“算法和组合学实验室”项目，提供了计算机设备；CNPq 基金是通过提供个人奖金以及通过 PROCOMB 和 TCPAC 资助研究访问的形式提供资助。

感谢我们学生对手稿的校对，特别是 Nalvo Franco de Almeida Jr. 和 Maria Emilia Machado Telles Walter。Nalvo 还制作了许多图表并给出很多建议。

我们和同事 Jorge Stolfi 进行了许多有益的讨论，他还给予至关重要的排版辅助。Fernando Reinach 和 Gilson Paulo Manfio 帮助我们编写了第 1 章。我们同 Jim Orlin，Martin Farach，Sampath Kannan 以及其他一些佚名的评阅人讨论了本书的目标和一般性问题，部分建议已写入书中。我们在 UNICAMP 计算研究所的同事给我们以鼓励并提供了令人上进的工作环境。

下列同行友好地寄给我们研究论文：Farid Alizadeh，Alberto Caprara，Martin Farach，David Greenberg，Dan Gusfield，Sridar Hannenhalli，Wen-Lian Hsu，Xiaoqiu Huang，Tao Jiang，John Kececioglu，Lukas Knecht，Rick Lathrop，Gene Myers，Alejandro Schäffer，Ron Shamir，Martin Vingron（他还寄来了讲义），Todd Wareham 和 Tandy Warnow。本书的一些部分很大程度上基于上述的部分论文。

非常感谢 Erik Brisson，Eileen Sullivan，Bruce Dale，Carlos Eduardo Ferreira 和 Thomas Roos，他（她）们以不同的方式提供了帮助。

J.C.S 感谢他的妻子 Silvia（a.k.a Teca）和孩子 Claudia，Tomás 和 Caio，没有他们的支持本书可能无法完成。

作者以 Leslie Lamport 的 LATEXA2 系统排版本书。

前言开头引用的语句出自 Don Knuth 在 1993 年 12 月 7 日同 Computer Literacy Bookshops，Inc. 的谈话。

João Carlos Setubal
João Meidanis

目　录

译者序 …… (i)
前言 …… (iii)
全书概述 …… (iv)
习题 …… (v)
错误 …… (v)
致谢 …… (vi)
第 1 章　分子生物学的基本概念 …… (1)
1.1　生命 …… (1)
1.2　蛋白质 …… (2)
1.3　核酸 …… (4)
1.3.1　DNA …… (4)
1.3.2　RNA …… (6)
1.4　分子遗传学机制 …… (6)
1.4.1　基因和遗传密码 …… (7)
1.4.2　转录、翻译和蛋白质合成 …… (8)
1.4.3　无用 DNA 和读框 …… (9)
1.4.4　染色体 …… (10)
1.4.5　基因组类似计算机程序? …… (11)
1.5　基因组是如何被研究的 …… (11)
1.5.1　图谱和序列 …… (11)
1.5.2　特殊技术 …… (13)
1.6　人类基因组计划 …… (16)
1.7　序列数据库 …… (17)
习题 …… (23)
文献提要 …… (23)
第 2 章　串、图和算法 …… (24)
2.1　串 …… (24)
2.2　图 …… (25)
2.3　算法 …… (28)
习题 …… (32)
文献提要 …… (33)
第 3 章　序列比较与数据库搜索 …… (34)
3.1　生物学背景 …… (34)

3.2 比较两个序列 …… (35)
3.2.1 全局比较—基本算法 …… (35)
3.2.2 局部比较 …… (40)
3.2.3 半全局比较 …… (41)
3.3 基本算法的扩展 …… (42)
3.3.1 空间节省 …… (42)
3.3.2 一般空隙罚分函数 …… (45)
3.3.3 仿射空隙罚分函数 …… (47)
3.3.4 比较相似序列 …… (49)
3.4 比较多个序列 …… (52)
3.4.1 SP 度量 …… (52)
3.4.2 星比对 …… (58)
3.4.3 树比对 …… (59)
3.5 数据库搜索 …… (60)
3.5.1 PAM 矩阵 …… (60)
3.5.2 BLAST …… (63)
3.5.3 FAST …… (65)
3.6 其他问题 …… (67)
* 3.6.1 相似性与距离 …… (67)
3.6.2 序列比较中的参数选择 …… (72)
3.6.3 串匹配和确切序列比较 …… (74)
小结 …… (75)
习题 …… (76)
文献提要 …… (77)
第 4 章 DNA 片段组装 …… (80)
4.1 生物学背景 …… (80)
4.1.1 理想情形 …… (81)
4.1.2 复杂情形 …… (81)
4.1.3 DNA 测序的其他方法 …… (86)
4.2 模型 …… (87)
4.2.1 最短公共超串 …… (87)
4.2.2 重构 …… (88)
4.2.3 多连叠 …… (89)
* 4.3 算法 …… (91)
4.3.1 交叠的表示 …… (91)
4.3.2 通路产生超串 …… (91)
4.3.3 作为通路的最短超串 …… (93)
4.3.4 贪婪算法 …… (95)

4.3.5 无环子图 …… (96)
4.4 启发式 …… (100)
4.4.1 发现交叠 …… (102)
4.4.2 排序片段 …… (103)
4.4.3 比对与表决 …… (104)
小结 …… (106)
习题 …… (106)
文献提要 …… (107)
第 5 章 DNA 物理作图 …… (109)
5.1 生物学背景 …… (109)
5.1.1 限制位点作图 …… (110)
5.1.2 杂交作图 …… (111)
5.2 模型 …… (112)
5.2.1 限制位点模型 …… (112)
5.2.2 区间图模型 …… (113)
5.2.3 连续 1 特性 …… (115)
5.2.4 算法启示 …… (115)
5.3 一个 C1P 问题的算法 …… (116)
5.4 带错杂交作图的一种近似 …… (122)
5.4.1 一个图模型 …… (123)
5.4.2 一个保证 …… (124)
5.4.3 实际计算 …… (126)
5.5 杂交作图的启发式 …… (128)
5.5.1 筛选嵌合克隆 …… (128)
5.5.2 获得一个好的探针次序 …… (129)
小结 …… (130)
习题 …… (130)
文献提要 …… (132)
第 6 章 种系发生树 …… (133)
6.1 性状状态和完全种系发生问题 …… (135)
6.2 二值性状状态 …… (138)
6.3 两个性状 …… (141)
6.4 种系树的简约性和相容性 …… (144)
6.5 距离矩阵算法 …… (146)
6.5.1 重构可加树 …… (146)
* 6.5.2 重构超度量树 …… (149)
6.6 种系树之间的一致 …… (155)
小结 …… (158)

习题 …… (159)
文献提要 …… (160)
第 7 章 基因组重排 …… (163)
7.1 生物学背景 …… (163)
7.2 有向块 …… (165)
7.2.1 若干定义 …… (166)
7.2.2 断点 …… (167)
7.2.3 现实与期望图式 …… (168)
7.2.4 交叉图 …… (172)
7.2.5 坏组分 …… (175)
7.2.6 算法 …… (176)
7.3 无向块 …… (178)
7.3.1 条 …… (179)
7.3.2 算法 …… (181)
小结 …… (182)
习题 …… (182)
文献提要 …… (183)
第 8 章 分子结构预测 …… (185)
8.1 RNA 二级结构预测 …… (185)
8.2 蛋白质折叠问题 …… (190)
8.3 蛋白质比配 …… (191)
小结 …… (195)
习题 …… (195)
文献提要 …… (195)
第 9 章 结语:DNA 计算 …… (197)
9.1 哈密顿通路问题 …… (197)
9.2 可满足性 …… (199)
9.3 问题与展望 …… (202)
习题 …… (202)
文献提要 …… (203)
习题选解 …… (204)
参考文献 …… (209)
索引 …… (220)

第 1 章　分子生物学的基本概念

本章介绍分子生物学的基本概念，旨在为读者提供足够的信息，以便他们能从容应对本书中涉及的生物学背景以及一般的计算分子生物学文献。受过严格科学训练的读者应清楚在分子生物学中没有百分之百绝对的东西，每一条规则都有例外。对于一般的规则，我们已尽力指出某些主要的例外，而在其他情形下则常常省略不提。所以读者不宜将本章视为分子生物学的教科书。

1.1　生命

自然界包括生命和非生命物质。相对于非生命物质而言，生命物质具有移动、繁殖、 1[①]
生长和摄食等能够主动参与环境的特性。20 世纪的研究已表明这两类物质都是由同样的原子组成并遵循同样的物理和化学规律。那么，区别何在？在长久的人类历史中，人们认为生命体被赋予了某些特殊的物质，使得它们具备主动的特性，即它们被这种物质“激活”了。可是，这类物质却从未被发现。相反，我们现在知道生命体的行为缘于它们体内一系列复杂的化学反应。这些反应从不停止，且常常是一个反应的产物即刻被另一个反应利用，使系统往复不止。活着的生物持续不断地与周围环境进行物质和能量交换；反之，任何与环境达到平衡的东西都被认为是死亡的（某些生物形态则是例外，如种子、病 2
毒，它们可长时间没有活性，但并不死亡）。

现代科学研究表明生命起源于 35 亿年前，地质时间上稍晚于（按地质术语）地球本身的形成。最初的生命形式非常简单，但在几十亿年的称之为进化的主动过程的作用下，生命发生了演绎并产生了多样性，因此今天我们可以同时看到非常复杂和非常简单的生命体。

复杂和简单的生物有着相似的分子化学或生物化学。生物化学中的主要角色是被称为**蛋白质**（protein）和**核酸**（nucleic acid）的分子。粗略地说，蛋白质决定一个生物是什么和做什么（著名的科学家 Russell Doolittle 曾写过“我们是我们的蛋白质”），核酸则负责编码产生蛋白质所必要的信息，并把这种信息传给后代。

分子生物学的研究基本上致力于理解蛋白质和核酸的结构与功能，这些分子因而成

① 为了方便读者查阅，我们在正文切口空白处加上原书的原始页码。

为本书的研究对象，下面我们将简要介绍当前有关蛋白质和核酸的基本知识。

1.2 蛋白质

我们身体的大部分物质是各种各样的蛋白质。蛋白质包括很多种，结构蛋白是组织的构成单元，酶是化学反应的催化剂。催化剂是指能够加速化学反应的物质。如果没有催化剂，大多数生物化学反应进行得很慢甚至不能开始。酶可呈数量级地加速这种过程，使生命成为可能。酶非常特别，通常一种酶仅催化一类生物化学反应。考虑到为维持生命需要有大量的生化反应，我们就需要大量的酶。蛋白质的其他功能还包括氧气运输和抗体防御等。究竟什么是蛋白质？它们如何组成？它们如何发挥功能？这一部分将简要回答这些问题。

蛋白质是由一类称为**氨基酸**(amino acid)的简单分子组成的。图 1.1 是氨基酸的例子，每一个氨基酸有一个中心碳原子，谓之 α 碳原子或 C_α。C_α 连接 1 个氢原子(H)、1 个氨基($-NH_2$)、1 个羧基(—COOH)和 1 个侧链。正是侧链决定了氨基酸间的差异。侧链可以是简单的氢原子(如甘氨酸)，也可以是复杂的两个碳环(如色氨酸)。我们在自然界只发现了 20 种不同的氨基酸，表 1.1 列出了这些氨基酸，这是蛋白质中最常见的 20 种，另外还有几种非标准的氨基酸。

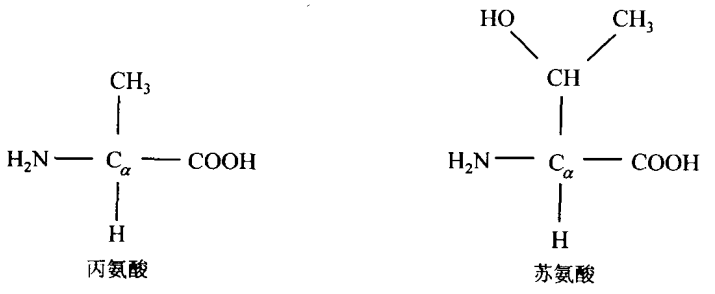

图 1.1 氨基酸的例子

在蛋白质中，氨基酸通过肽键相连。因此，蛋白质是多肽链。在肽键中，属于氨基酸 A_i 的羧基的 C 原子与氨基酸 A_{i+1} 的 N 原子连接。在该键形成中，因羧基的氢原子和氧原子与氨基的氢原子结合而脱去 1 个水分子。因此我们在肽键内部所发现的仅仅是一个**残基**(residue)。由此我们说蛋白质有 100 个残基而不是说有 100 个氨基酸。典型的蛋白质含大约 300 个残基，但也有少至 100 个残基和多达 5 000 个残基的氨基酸。

表 1.1 蛋白质中发现的 20 种常见氨基酸

	单字母缩写	三字母缩写	名 称
1	A	Ala	丙氨酸(alanine)
2	C	Cys	半胱氨酸(cysteine)
3	D	Asp	天冬氨酸(aspartic acid)
4	E	Glu	谷氨酸(glutamic acid)

续表

	单字母缩写	三字母缩写	名称
5	F	Phe	苯丙氨酸(phenylalanine)
6	G	Gly	甘氨酸(glycine)
7	H	His	组氨酸(histidine)
8	I	Ile	异亮氨酸(isoleucine)
9	K	Lys	赖氨酸(lysine)
10	L	Leu	亮氨酸(leucine)
11	M	Met	甲硫氨酸(methionine)
12	N	Asn	天冬酰胺(asparagine)
13	P	Pro	脯氨酸(proline)
14	Q	Gln	谷氨酰胺(glutamine)
15	R	Arg	精氨酸(arginine)
16	S	Ser	丝氨酸(serine)
17	T	Thr	苏氨酸(threonine)
18	V	Val	缬氨酸(valine)
19	W	Trp	色氨酸(tryptophan)
20	Y	Tyr	酪氨酸(tyrosine)

肽键使得每个蛋白质都有一个骨架(backbone),它是重复的基本单元—N—C_α—(CO)—,每个 C_α 有一个侧链,图 1.2 是一个多肽链的图示。因为在骨架的一端是一个氨基,另一端是一个羧基,我们因此可以区别多肽链的两端并给它定一个方向,习惯上多肽始于氨基(N 端),止于羧基(C 端)。

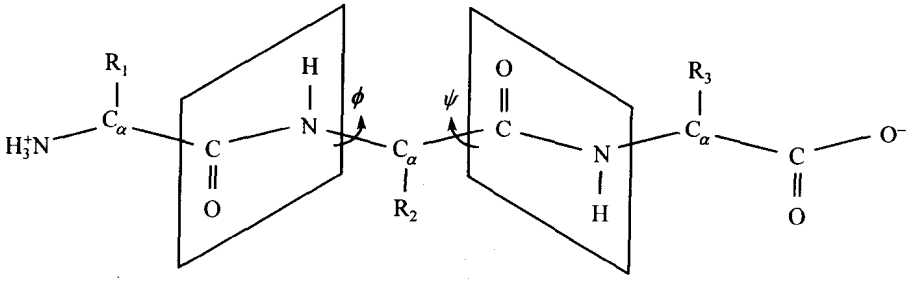

图 1.2　一个多肽链,R_i 侧链表示氨基酸,每个四边形中的原子在同一平面上,能够根据角度 ϕ 和 ψ 旋转

蛋白质并不仅仅是氨基酸残基的线性序列,这种序列称为**一级结构**(primary structure)。蛋白质实际上在三维空间中折叠,形成**二级**(secondary)、**三级**(tertiary)和**四级**(quaternary)结构。蛋白质的二级结构是通过骨架原子间的相互作用形成的,并导致"局部"结构,如螺旋(helix)等。三级结构是二级结构在更大的范围内堆积的结果。更高层次的堆积,即一组不

同亚基(译者注:原书此处是"protein")的堆积,形成四级结构。图 1.3 显示了这些结构。

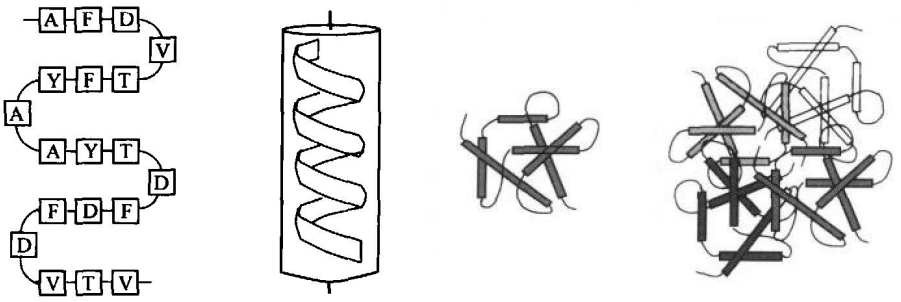

图 1.3 蛋白质的一级、二级、三级、四级结构(取自参考文献[28]中的图)
译者注:第二个图显示了一个左手螺旋,而实际上蛋白质中几乎都是右手螺旋

蛋白质之所以能够在三维空间中折叠是因为 C_α 原子和 N 原子之间的肽键平面可以旋转,C_α 原子和其他 C 原子之间的平面也可以旋转,如图 1.2 所示,这两类旋转的角度分别表示为 ϕ 和 ψ。侧链也可以运动,但相对于骨架旋转而言属于次级运动。因此只要我们规定了蛋白质中所有 ϕ-ψ 对的值,便可确切知道蛋白质的折叠。确定蛋白质的折叠或三维结构是分子生物学的一个重要领域,原因有三点。首先,蛋白质的三维结构与其功能相关;其次,蛋白质由 20 种氨基酸组成,这使得三维结构非常复杂,缺乏对称性;第三,目前没有简单精确的确定三维结构的方法。这些因素促成了第 8 章内容的形成,我们在该部分讨论分子结构的预测方法,这些方法试图通过分子的原始序列预测其结构。

蛋白质的三维形状能以下述方式决定其功能。折叠的蛋白质具有不规则的形状,这意味着它具有各种隐窝和膨凸,这些形状使得蛋白质能够与其他特定分子形成紧密的接触或**结合**(bind)。哪些分子可以结合取决于蛋白质的形状。例如,蛋白质的形状可以容许自身几个拷贝的结合,形成一个像发丝的结构;或其结构可使得分子 A 和 B 与其结合并交换原子,A 和 B 之间发生了化学反应,而蛋白质起到了催化剂的作用。

我们是如何获得蛋白质的?蛋白质是在一个称为核糖体的细胞结构中合成的。在核糖体中,根据**信使 RNA**(messenger ribonucleic acid,mRNA)分子所包含的信息,蛋白质的氨基酸一个接一个地装配起来。为解释这是如何发生的,我们需要说明什么是核酸。

1.3 核酸

生物体包含两类核酸:**核糖核酸**(ribonucleic acid),简写为 RNA;**脱氧核糖核酸**(deoxyribonucleic acid),简写为 DNA。我们首先描述 DNA。

1.3.1 DNA

类似于蛋白质,DNA 分子也是由小分子组成的链。实际上 DNA 是双链,但我们首先认识一下 DNA 的**单链**(strand),它是由重复的基本单元组成的骨架。这种基本单元由一个称做 2′-脱氧核糖的糖分子和一个磷酸残基组成。糖分子含有 5 个碳原子,标记为 1′→

5′(图 1.4)。形成骨架的键是在一个单元的 3′碳原子和下一个单元的 5′碳原子之间。因此,DNA 分子具有**方向性**(orientation),一般从 5′开始到 3′结束。当我们在技术文献、书籍和序列数据库看到一个单链 DNA 分子时,若无特别说明,均是按 5′→3′方向书写的。

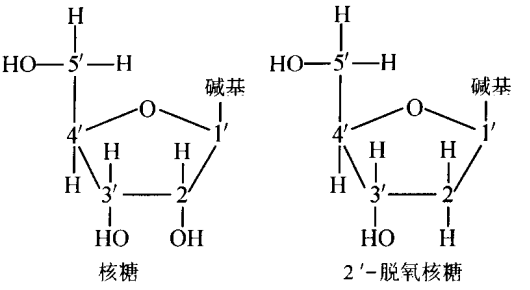

图 1.4　核酸中的糖,符号 1′→5′表示碳原子。两种糖的惟一差别是 2′碳原子中的氧,RNA 中是核糖,DNA 中是脱氧核糖

与骨架中与 1′碳原子相连的分子为**碱基**(base)。在 DNA 分子中有 4 种碱基,分别是:腺嘌呤(adenine,A)、鸟嘌呤(guanine,G)、胞嘧啶(cytosine,C)和胸腺嘧啶(thymine,T)。图 1.5 显示了每种碱基的分子结构,图 1.6 为单链 DNA 分子的示意图。碱基 A 和 G 是嘌呤,而碱基 C 和 T 属于嘧啶。DNA 分子的

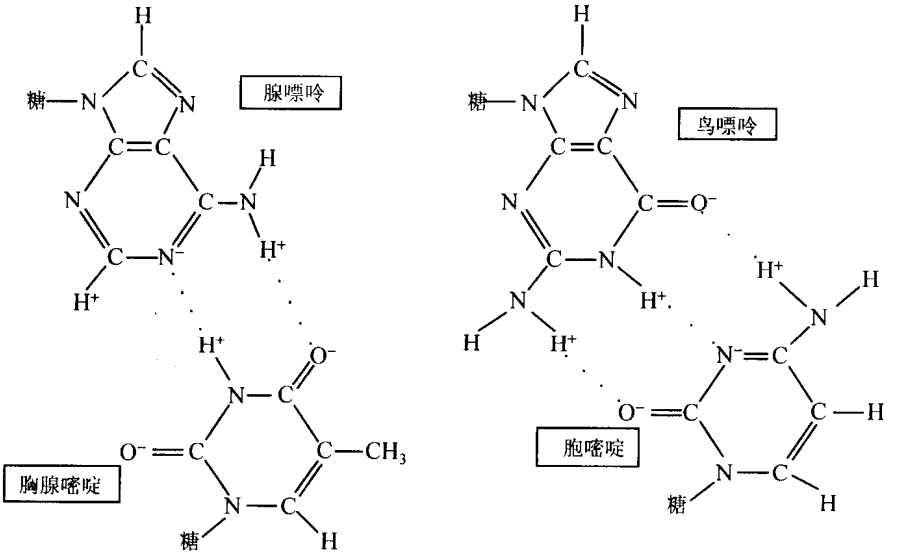

图 1.5　DNA 中氮化的碱基。注意腺嘌呤和胸腺嘧啶、鸟嘌呤和胞嘧啶之间所形成的键,图中用点画线表示

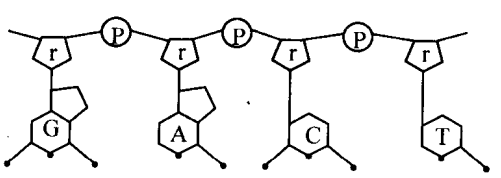

图 1.6　一个 DNA 链分子结构的图示

基本单元由糖、磷酸和碱基组成,该基本单元谓之**核苷酸**(nucleotide)。虽然碱基和核苷酸不同,但我们可以称 DNA 分子有 200 个碱基或由 200 个核苷酸组成。仅含有几个或几十个核苷酸的 DNA 分子称为寡核苷酸。自然界中 DNA 分子非常长,长度远超过蛋白质。在人的细胞中,DNA 分子由上亿个核苷酸组成。

前已述及,DNA 分子是双链结构。两条链缠绕在一起形成双螺旋,此著名的双螺旋(double helix)结构是由 James Watson 和 Francis Crick 在 1953 年发现的。两条链结合的