生物大分子的数学描述及其应用

Mathematical Description of the Biological Macromolecules and Its Applications

李 春　钱伟懿 著

大连理工大学出版社
DALIAN UNIVERSITY OF TECHNOLOGY PRESS

生物大分子的数学描述及其应用

李 春　钱伟懿 著

大连理工大学出版社
DALIAN UNIVERSITY OF TECHNOLOGY PRESS

图书在版编目(CIP)数据

生物大分子的数学描述及其应用　李春，钱伟懿著．
大连：大连理工大学出版社，2009.2
ISBN 978-7-5611-4611-6

Ⅰ．生… Ⅱ．①李…②钱… Ⅲ．生物高分子－研究
Ⅳ．Q71

中国版本图书馆 CIP 数据核字(2009)第 025836 号

大连理工大学出版社出版
地址：大连市软件园路 80 号　邮政编码：116023
发行：0411-84708842　邮购：0411-84703636　传真：0411-84701466
E-mail:dutp@dutp.cn　URL:http://www.dutp.cn
大连金华光彩色印刷有限公司印刷　大连理工大学出版社发行

幅面尺寸：147mm×210mm　印张：5.625　字数：141 千字
2009 年 2 月第 1 版　2009 年 2 月第 1 次印刷

责任编辑：王伟　责任校对：婕琳
封面设计：胡辑

ISBN 978-7-5611-4611-6　定价：38.00 元

序

随着人类和一些模式生物基因组计划的相继完成或全面实施，生物学研究的重点正从积累数据向分析解释这些数据过渡，生物信息学（也称计算分子生物学）便应运而生。这是一门运用数学、信息科学、计算机科学和系统科学的理论与方法研究生命现象、分析和处理呈指数增长的生物学原始数据并进行加工、分析和建立计算模型的一门学科。生物信息学的研究内容十分丰富，例如，序列比较、计算机辅助基因识别、系统发育分析、RNA 和蛋白质结构预测、遗传密码及其起源、序列重叠群装配、基于结构的药物设计等等，都是生物信息学中重要的研究领域。其中大多数领域的研究工作都有一个共同的需求，就是常常需要给出生物学数据的数学上的描述，因此，生物大分子的数学描述便成为生物信息学中一个非常基础又十分重要的课题。

近年来，围绕生物信息学开展的相关研究在我国受到了广泛的

重视，许多高等院校都为本科生和研究生开设了生物信息学课程，相关的书籍也出了不少，但专门从数学的角度进行研究的却不多见。建立生物学数据以及各种数据间复杂关系的数学模型，然后在此基础上分析和解释相应生物学意义，进而探索其固有的生物学规律并研究相关生物信息学问题，这是本书的特色之处。

本书是在我们实际从事的课题基础上形成的，从这个意义上讲，本书可以说是一份工作汇报。但限于我们的水平，书中难免有不妥或错误之处，恳请读者批评指正。

本书得到了辽宁省重点学科建设项目和国家自然科学基金项目的支持。在本书的编写过程中，得到了硕士研究生邢丽丽、潘琳玉、裴晓、马弘等的帮助，在此一并表示感谢。

李　春

钱伟懿

2009 年 1 月

目　录

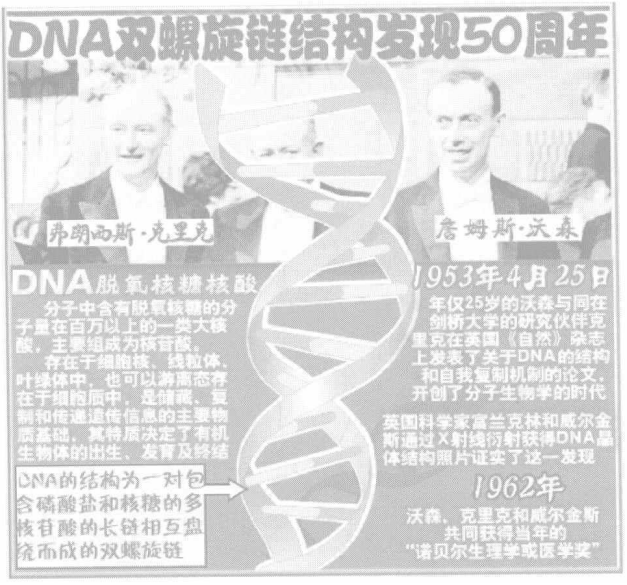
第0章 绪 论

20世纪是科学技术迅速发展的世纪，物理和化学的发展使我们可以清楚地认识物质的组成，从分子、原子、电子等各个层次上深入地了解微观世界；天文技术、空间技术的发展使得我们可以了解地球以外的客观世界；以电子信息技术为龙头的工业技术的飞速发展，使得我们可以不断地改造世界，为人类创造更加舒适的生活条件，而生命科学的发展，则使我们能从器官、组织、细胞、生物大分子等各个层次认识生命的物质基础。

0.1 生物信息学产生的背景

1953年4月25日，詹姆斯·沃森与同在剑桥大学的合作伙伴弗朗西斯·克里克一起，在《自然》杂志上发表了一篇仅两页的论

文，提出了DNA的结构和自我复制机制，揭开了分子生物学的新篇章。



图片来源：新华网

50年后，人们迎来了又一个激动人心的时刻，在2003年4月14日，美、英、日、法、德和中国科学家经过13年努力共同完成了人类基因组计划(HGP，Human Genome Project)，比原计划提前两年，在人类揭示生命奥秘、认识自我的漫漫长路上又迈出重要一步。人类基因组计划是美国在1990年提出实施的一项伟大的科学计划，与阿波罗登月计划、曼哈顿原子弹计划同称为人类自然科学史上的三大计划。其目标是用大约15年时间，完成人类所有染色体中3×10^9个碱基对(bp，base pair)的序列测定。人类基因组计划的成果是一个人类遗传信息数据库，是一本指导人类进化的"说明书"。它不仅可以揭示人类生命活动的奥秘，而且人类几千种单基因遗传性

疾病和严重危害人类健康的多基因易感性疾病的致病机理有望得到彻底阐明，为这些疾病的诊断、治疗和预防奠定基础。同时，人类基因组计划的实施还将带动医药业、农业、工业等相关行业的发展，产生极其巨大的经济效益和无法估量的社会效益。

随着HGP的顺利完成，和诸如大肠杆菌、啤酒酵母、线虫、果蝇、小鼠、鸡、拟南芥、水稻、玉米等模式生物的基因组计划的相继完成或全面实施，DNA/蛋白质序列数据正以惊人的速度增长，在此基础上派生和整理出来的数据库已达500余个。这一切构成了一个生物学数据的海洋。我们知道，生物学是一门实验科学，也是一门发现科学。通过实验发现新的现象、新的生物学规律，经过分析和归纳总结，提炼出新的生物学知识。在这个过程中，需要对实验数据进行处理和理论分析，在此基础上解释实验现象，认识实验现象发生的本质，探索固有的生物学规律，进而了解和掌握生命的物质基础和生命的本质。生物数据积累速度不断加快，对生物数据的科学分析方法和实用分析工具提出了更新、更高的要求。

传统分子生物学实验往往是集中精力研究一个基因、一条代谢路径，手工分析完全能够胜任。然而，一方面，现在我们面对的是海量并且仍在不断迅速增加的生物学数据。一次只分析一个生物分子的传统的生物学已经无法满足要求。换句话说，现在需要的是同时分析成千上万个生物分子，是自动分析。同时，面对这么多生物分子数据，不可能用实验的方法去详细研究每一条序列，必须先进行信息处理和分析，去粗取精，去伪存真。通过预处理，发现有用的线索，在此基础上进行有针对性、有明确目的的分子生物学实验。另一方面，从生物分子数据本身来看，各种数据之间存在着密切的关系，如DNA序列与蛋白质序列、基因突变与疾病等，这些联系反映了生物学的规律。但是，这些关系可能是非常复杂的，是我们未知的，是简单的统计方法难以分析的。对于这些复杂的关系，必须运用现代信息学的方法去分析，去研究。

综上所述，生物学已经不再单单是原来的观察和实验的科学，理论和计算对生物学的进步正在发挥越来越重要的作用，这就催生了生物信息学（又称计算分子生物学，现在人们常常不加区分地使用这两个名称）这门崭新的交叉科学。这是一门运用数学、信息科学、计算机科学和系统科学的理论与方法研究生命现象、分析和处理呈指数增长的生物学原始数据并进行加工、分析和建立计算模型的一门学科。普遍认为，生物信息学是当今生命科学和自然科学中最关键、最重要的部分，是21世纪自然科学的核心领域之一[18]。

0.2 生物信息学的研究对象

生物体是一个复杂的系统，生命过程是一个极端复杂的过程，需要物质和能量的支持。生物体也是一个信息系统，该系统控制着生物的遗传、生长和发育。所有的信息存贮在生物体内，存贮在遗传物质中。在生命科学研究方面，人们已经逐渐认识到，不仅需要用物理、化学和生物学方法研究生命的物质基础、能量转换、代谢过程等，还需要用信息科学方法研究生命信息特别是遗传信息的组织、复制、传递、表达及其作用，否则难以理解生命的工作机制，难以揭示生命的奥秘。从生物学的观点来看，细胞是生命的基本单位，而从信息科学的观点来看，细胞则是存贮、复制和传递遗传信息的系统。

生物系统通过存贮、修改、解读遗传信息和执行遗传指令形成特定的生命活动，生长发育，产生生物进化。从信息学的角度来看，生物分子是生物信息的载体。生物分子至少携带着三种信息，即遗传信息、与功能相关的结构信息、进化信息。俗话说"种瓜得瓜，种豆得豆"，这是对生物遗传现象的生动描述。地球上的所有生物，上至"万物之灵"的人类，下至细菌的"寄生虫"——噬菌体，都表现出遗传现象，能够复制出新的一代，这是生命延续和种族繁衍的保证。生物的复制由基因所决定，复制是生命的基本特征，但不是生命的

全部特征。计算机程序可以自动复制大量的拷贝，但是这些程序不是活动的生命，活动的生命是不断变化的。绝大多数生命体可以从周围的环境中摄取物质，获取能量，并将所摄取的物质转换为其自身的一部分。计算机程序虽然可以拷贝，但是这种拷贝往往是绝对真实的拷贝，毫厘不差。而生物体在繁殖和遗传的过程中并非一成不变，后代与亲代存在着差异。正因为有遗传差异的存在，才有生物的进化。

生物信息学主要研究两种信息载体，即核酸分子(DNA、RNA)和蛋白质分子。

0.2.1 核 酸

核酸是遗传物质。核酸分为脱氧核糖核酸(DNA)和核糖核酸(RNA)。DNA主要存在于细胞核中，但细胞质里的线粒体、叶绿体中也含有少量DNA，RNA则主要分布在细胞质中。遗传的主要物质基础是DNA，但有时也是RNA(如病毒的遗传物质)。

核酸是由称为核苷酸(nucleotide)的小分子生成的聚合物。核苷酸还可以进一步分解成核苷(nucleoside)和磷酸，核苷进一步水解生成碱基(base)和戊糖。所以，核酸的基本结构单位是核苷酸，其组成方式为碱基－戊糖－磷酸(图0-1)。

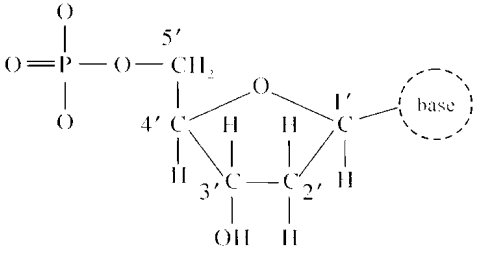


图0-1 核苷酸分子结构示意图

DNA 和 RNA 所含的戊糖不同：前者中的戊糖是脱氧核糖，而后者的则是核糖。DNA 和 RNA 在组成上的另一个区别体现在它们所含的碱基组成上。DNA 中的碱基有 4 种，分别是腺嘌呤（Adenine，简写作 A）、鸟嘌呤（Guanine，简写作 G）、胞嘧啶（Cytosine，简写作 C）和胸腺嘧啶（Thymine，简写作 T）。RNA 中没有胸腺嘧啶 T，取而代之的是尿嘧啶 U（Uracil）。五种碱基的分子结构示意图如图 0-2 所示。

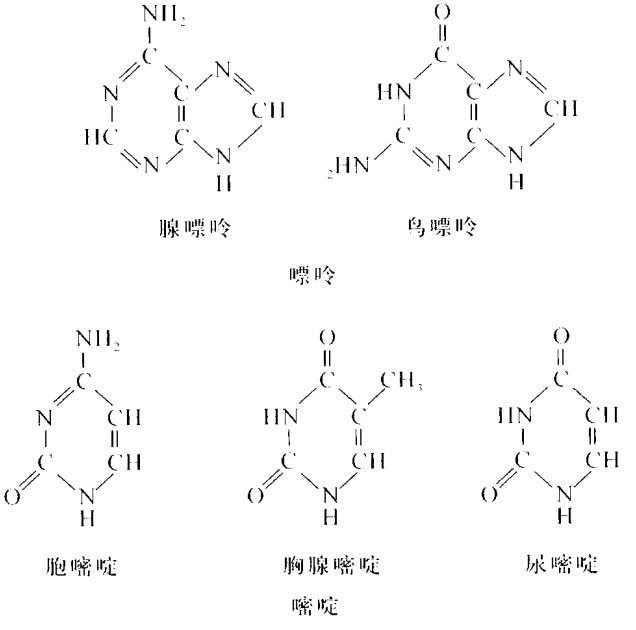


图 0-2　五种碱基 A、G、C、T、U 的分子结构示意图

可见，仅就 DNA 或者 RNA 分子而言，不同核苷酸之间的区别仅在于它们所含的碱基不同。因此，A、G、C、T(U) 也常被用来直接

表示相应的核苷酸。核苷酸相互连接形成长的多核苷酸链。由四种脱氧核苷酸连接而成的长链高分子多聚体为DNA分子的一级结构。DNA分子中第一个核苷酸的3′-羟基与第二个核苷酸的5′-磷酸基脱水形成3′,5′-磷酸二酯键,第二个核苷酸的3′-羟基又与第三个核苷酸的磷酸基脱水形成3′,5′-磷酸二酯键,依此类推,形成线性多聚体。DNA分子中第一个核苷酸的5′-磷酸与最末一个核苷酸的3′-羟基都未参与形成3′,5′磷酸二酯键,故分别称为5′-磷酸端(或5′端)和3′-羟基端(或3′端)。

DNA蕴涵的复制机制的关键特征是互补基对。这就是著名的Watson-Crick配对,即A与T配对,G与C配对。这种配对是由于氢键作用,原理是DNA单链(按从5′到3′的次序)与相反方向写的互补链配对。例如,单链碱基序列5′-ATGGTGCACC 3′和3′-TACCACGTGG 5′配对:

```
5′   A T G G T G C A C C — 3′
     | | | | | | | | | |
3′   T A C C A C G T G G — 5′
```

0.2.2 蛋白质

蛋白质是生物体内占有特殊地位的生物大分子,它是生物体的基本构件,也是生命活动的重要物质基础,几乎一切生命现象都要通过蛋白质的结构与功能而体现出来。因此,在分子生物学中,深刻阐明蛋白质的结构与功能,是探索生命奥秘最基本的任务。

蛋白质是由氨基酸(amino acid)聚合而成的生物大分子。氨基酸是蛋白质的基本组成单位。自然界中的氨基酸种类很多,但参与蛋白质组成的常见氨基酸只有20种。这20种标准氨基酸的英文三字母和单字母表示见表0.1。

表 0-1　20 种标准氨基酸的英文三字母和单字母表示

氨基酸名称	英文缩写	简写	氨基酸名称	英文缩写	简写
甘氨酸	Gly	G	丝氨酸	Ser	S
丙氨酸	Ala	A	苏氨酸	Thr	T
缬氨酸	Val	V	天冬酰胺	Asn	N
异亮氨酸	Ile	I	谷酰胺	Gln	Q
亮氨酸	Leu	L	酪氨酸	Tyr	Y
苯丙氨酸	Phe	F	组氨酸	His	H
脯氨酸	Pro	P	天冬氨酸	Asp	D
甲硫氨酸	Met	M	谷氨酸	Glu	E
色氨酸	Trp	W	赖氨酸	Lys	K
半胱氨酸	Cys	C	精氨酸	Arg	R

氨基酸是带有氨基的有机酸，它的中心碳原子特称为 α 碳(C_α)。C_α 有四个键，分别连着一个氨基(NH_2)、一个羧基(COOH)、一个氢原子和一个 R 基团(图 0-3)。各种 α 氨基酸的区别在于侧链 R 基团不同，R 基团的特异性使不同氨基酸显示出不同的理化性质，进而决定了氨基酸在蛋白质分子的空间结构中可能的位置。

在蛋白质合成时，一个氨基酸的羧基和另一个氨基酸的氨基缩水形成肽键(peptide bond)。所以，蛋白质也是有方向的一维链，带氨基的一头称为 N 端或记为 N'，另一头带羧基称为 C 端，常用 C' 表示。

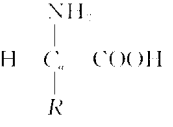


图 0-3　氨基酸分子结构示意图

0.2.3　中心法则和遗传密码

DNA 携带遗传材料，即生物功能所要求的信息(某些病毒除

外，它们的遗传材料是RNA）。信息从基因的核苷酸序列中被提取出来，用来指导蛋白质合成的过程对地球上的所有生物是相同的，分子生物学家称之为中心法则（central dogma）。

生物体的遗传信息以密码形式编码在DNA分子上，表现为特定的核苷酸排列顺序，并通过DNA的复制（replication）使遗传信息从亲代传向子代。在后代的生长发育过程中，DNA分子中的遗传信息转录（transcription）到RNA分子中（即RNA聚合酶以DNA为模板合成RNA），再由RNA翻译（translation）生成体内各种蛋白质，行使特定的生物功能。翻译过程是在核糖体上进行的。这样，通过遗传信息从亲代传向子代，并在子代表达，使得子代获得了亲代的遗传性状。RNA也能通过复制过程合成出与其自身相同的分子。此外，生物界还存在由RNA指导下的DNA合成过程，即逆转录，这一过程发现于逆转录病毒中。通过基因转录和翻译得到的蛋白质分子可以反过来作用于DNA，调控其他基因的表达。分子生物学的中心法则如图0-4所示。它说明遗传信息由DNA到RNA，再到蛋白质的传递过程。

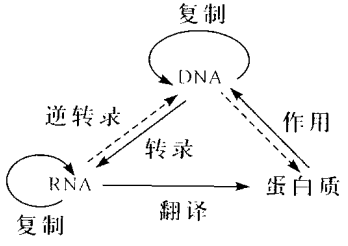


图0-4 分子生物学的中心法则

在翻译过程中，每三个碱基构成一个三联体，对应一个氨基酸或者一个终止密码子。我们称这种对应为遗传编码，可数学地表示