计算机科学与技术学科前沿丛书　计算机科学与技术学科研究生系列教材(中文版)

生物信息学导论

——面向高性能计算的算法与应用

王勇献　王正华　编著

清华大学出版社

北　京

内 容 简 介

本书主要针对生物信息学中的典型应用，从计算方法角度介绍相关算法的原理及应用；内容分成生物学及数理基础、生物序列分析、蛋白质组学分析以及大规模生物学网络分析等四个专题，涉及生物分子序列分析、基因发现、分子进化分析、蛋白质结构预测、蛋白质肽测序、生物学网络模块划分等具体问题的求解原理及算法实现。

本书的读者对象是具有现代分子生物学及计算机科学基本知识的研究生及相关科研人员，在附加习题后也可作为生物信息学方面的入门及进阶教材，供生物医学工程、计算机应用等专业学生使用。

图书在版编目(CIP)数据

生物信息学导论：面向高性能计算的算法与应用/王勇献, 王正华编著. — 北京: 清华大学出版社, 2011.6

(计算机科学与技术学科前沿丛书　计算机科学与技术学科研究生系列教材（中文版))

ISBN 978-7-302-25022-7

Ⅰ. ①生…　Ⅱ. ①王…　②王…　Ⅲ. ①生物信息论-研究生-教材　Ⅳ. ①Q811.4

中国版本图书馆 CIP 数据核字 (2011) 第 045333 号

责任编辑：焦　虹　徐跃进
责任校对：李建庄
责任印制：

出版发行：清华大学出版社　　**地　　址**：北京清华大学学研大厦 A 座
http://www.tup.com.cn　　**邮　　编**：100084
社 总 机：010-62770175　　**邮　　购**：010-62786544
投稿与读者服务：010-62795954, jsjjc@tup.tsinghua.edu.cn
质量反馈：010-62772015, zhiliang@tup.tsinghua.edu.cn

印 刷 者：
装 订 者：
经　　销：全国新华书店
开　　本：185×260　　**印　张**：32.5　　**字　数**：810 千字
版　　次：2011 年 6 月第 1 版　　**印　次**：2011 年 6 月第 1 次印刷
印　　数：
定　　价：0.00元

产品编号：040630-01

“十一五”国家重点图书　　计算机科学与技术学科前沿丛书

计 算 机 科 学 与 技 术 学 科 研 究 生 系 列 教 材

序

未来的社会是信息化的社会，计算机科学与技术在其中占据了最重要的地位，这对高素质创新型计算机人才的培养提出了迫切的要求。计算机科学与技术已经成为一门基础技术学科，理论性和技术性都很强。与传统的数学、物理和化学等基础学科相比，该学科的教育工作者既要培养学科理论研究和基本系统的开发人才，还要培养应用系统开发人才，甚至是应用人才。从层次上来讲，则需要培养系统的设计、实现、使用与维护等各个层次的人才。这就要求我国的计算机教育按照定位的需要，从知识、能力、素质三个方面进行人才培养。

硕士研究生的教育须突出“研究”，要加强理论基础的教育和科研能力的训练，使学生能够站在一定的高度去分析研究问题、解决问题。硕士研究生要通过课程的学习，进一步提高理论水平，为今后的研究和发展打下坚实的基础；通过相应的研究及学位论文撰写工作来接受全面的科研训练，了解科学研究的艰辛和科研工作者的奉献精神，培养良好的科研作风，锻炼攻关能力，养成协作精神。

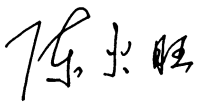
高素质创新型计算机人才应具有较强的实践能力，教学与科研相结合是培养实践能力的有效途径。高水平人才的培养是通过被培养者的高水平学术成果来反映的，而高水平的学术成果主要来源于大量高水平的科研。高水平的科研还为教学活动提供了最先进的高新技术平台和创造性的工作环境，使学生得以接触最先进的计算机理论、技术和环境。高水平的科研也为高水平人才的素质教育提供了良好的物质基础。

为提高高等院校的教学质量，教育部最近实施了精品课程建设工程。由于教材是提高教学质量的关键，必须加快教材建设的步伐。为适应学科的快速发展和培养方案的需要，要采取多种措施鼓励从事前沿研究的学者参与教材的编写和更新，在教材中反映学科前沿的研究成果与发展趋势，以高水平的科研促进教材建设。同时应适当引进国外先进的原版教材，确保所有教学环节充分反映计算机学科与产业的前沿研究水平，并与未来的发展趋势相协调。

中国计算机学会教育专业委员会在清华大学出版社的大力支持下，进行了计算机科学与技术学科硕士研究生培养的系统研究。在此基础上组织来自多所全国重点大学的计算机专家和教授们编写和出版了本系列教材。作者们以自己多年来丰富的教学和科研经验为基础，认真研究和结合我国计算机科学与技术学科硕士研究生教育的特点，力图使本系列教材对我国计算机科学与技术学科硕士研究生的教学方法和教学内容的改革起引导作用。本系列教材的系统性和理论性强，学术水平高，反映科技新发展，具有合适的深度和广度。同时本系列教材两种语种（中文、英文）并存，三种版权（本版、外版、

合作出版）形式并存，这在系列教材的出版上走出了一条新路。

相信本系列教材的出版，能够对提高我国计算机硕士研究生教材的整体水平，进而对我国大学的计算机科学与技术硕士研究生教育以及培养高素质创新型计算机人才产生积极的促进作用。



序　言

生物信息学 (又称计算生物学) 是计算机、生物、数学等多学科交叉而新兴的学科分支，以借助计算机科学、信息科学等领域的算法与工具解决生命科学中的问题为主要特征。随着生命科学研究的深入和后基因组时代的来临，生物信息学所研究的问题已经发生了巨大的变化，新的研究越来越需要借助于高性能计算环境的支持，然而综合生命科学与高性能计算两个领域的知识、有效解决现实中的应用问题并不是一件容易的事，目前也缺乏系统深入的相关图书资料。编写这本《生物信息学导论》，作者希望在向读者介绍生物信息学中与高性能计算结合最密切的一些基础性问题，讨论并总结相关的求解算法与应用技术。

尽管本书定位在生物信息学“导论”的层次，但是作者并没有打算从生物信息学领域的概念内涵、发展现状入手，也没有追求内容上的面面俱到、或者对生物信息学领域内容进行泛泛介绍，而是有所选择地介绍了生物分子序列分析、基因发现、分子进化分析、蛋白质结构预测、蛋白质肽测序、生物学网络模块划分等具体问题的求解算法及原理实现。从这个意义上讲，也可以将本书看成是生物信息学部分专题内容的汇集。在每部分专题内容中，既有对经典方法的详细讨论，也融入了作者及其合作研究者最近几年研究的创新成果，既注重理论方面的方法 (例如：利用谱分析挖掘 PPI 网络中的典型模式)，也强调具体应用方面的实现 (例如：利用 MPI 实现并行计算)。值得说明的是，结合作者知识背景与研究兴趣，本书在介绍各类生物信息学问题的求解方法时，特别关注了如何跟高性能计算技术相结合 (例如：关于序列比对的并行计算)。

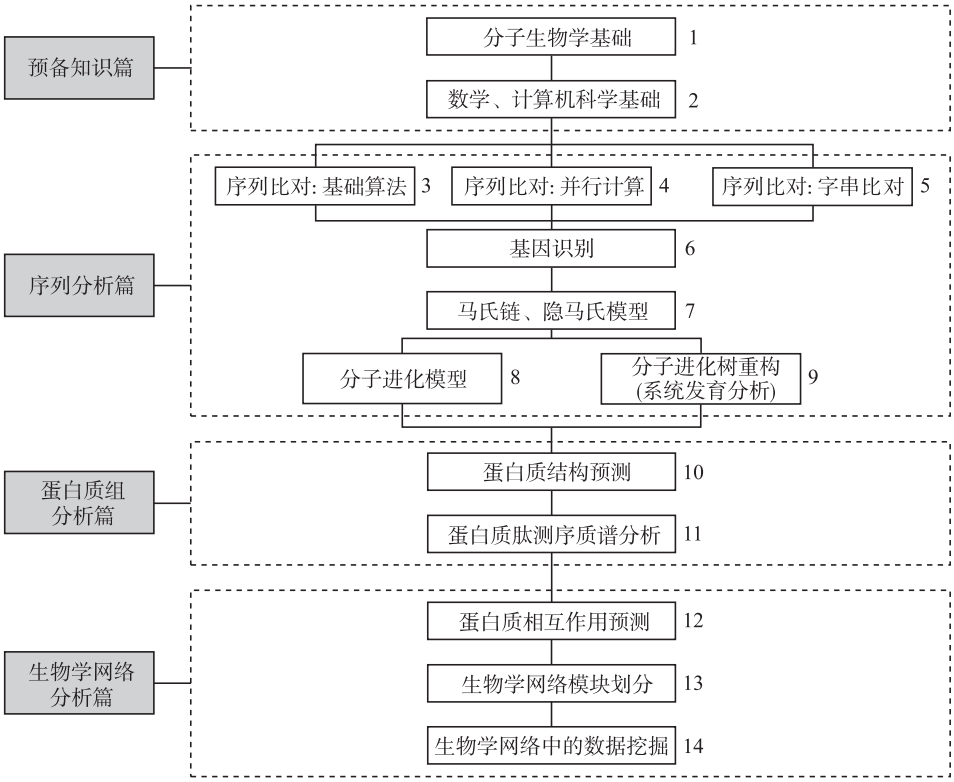
全书正文各章节结构如下图所示，共分为“预备知识篇”、“序列分析篇”、“蛋白质组分析篇”和“生物学网络分析篇”等四部分。

“预备知识篇”(包括第 1 章和第 2 章) 提供了生物信息学分析中涉及的常用分子生物学、线性代数、概率统计以及计算机算法等方面的基础知识，不同知识背景的读者可以选择自己所需的内容。读者阅读时也可以直接跳过这两章，在后续章节需要相关知识时，再回头翻阅这些内容。

“序列分析篇”是全书的重点内容之一，共包括七章内容 (第 3 章到第 9 章)，主要涵盖生物大分子的序列比对、DNA 序列上的基因识别 (或称基因发现) 和分子系统发育分析等。其中，序列比对共有三章，分别从经典基础算法、高性能并行实现以及基于字符串的模式匹配三个层面进行介绍；分子系统发育分析共有两章，分别介绍分子进化模型和进化树重构方法等内容；作为序列分析中最常用的一种模型，我们在本篇内容中专门介绍了马氏链和隐马氏模型 (第 7 章)，该模型在后续各篇中也有应用。

“蛋白质组分析篇”包括两章内容 (第 10 章和第 11 章)，分别介绍了基于序列预测蛋白质结构的计算分析，以及基于质谱数据分析的蛋白质序列测定方法等内容。

“生物学网络分析篇”是全书的最后一部分，共包含三章内容 (第 12 章到第 14 章)，重点以蛋白质相互作用为例，介绍了蛋白质相互作用预测、蛋白质相互作用网络的模块划分与功能预测等应用中的计算方法，最后一章系统总结了一般生物学网络中的数据挖掘方法与结果。



本书的读者对象是具有现代分子生物学及计算机科学基本知识的研究生及相关科研人员，在附加习题后也可作为生物信息学方面的入门及进阶教材，供分子生物学、计算机应用等专业的读者使用。为了照顾不同学科背景知识读者的需求，本书在开头还简要介绍了阅读后续章节所需要的分子生物学、数学及计算机科学基础知识；全书最后附有详细的主题索引、人名索引及以字母排序的参考文献，每条文献还特意标注了在正文中被引用处的页码，以方便读者检索。

本书是在为国防科技大学硕士研究生开设的“生物信息学导论”课程讲义以及作者从事国家自然科学基金项目研究成果的基础上整理而成。作者感谢 2005—2010 各学年选修这门课程学习的所有同学，他们参与课程讨论的许多内容构成了本书的基本素材；本书在选题与出版方面得到了国家自然科学基金 (60603054)、湖南省自然科学基金 (08JJ4021) 及国家重点基础研究发展计划课题 (2009CB723803) 的资助。

在书稿准备出版过程中，清华大学出版社的广大员工给予了大力支持与帮助，在此

一并表示感谢。

全书由王勇献主编，王正华对内容进行了统稿并提出了改进意见。由于作者水平所限，书中还有很多错误和不足之处，希望读者批评指正 (作者联系方式：湖南长沙国防科技大学计算机学院，邮编：410073，电子邮箱：yxwang@nudt.edu.cn)。

作者于长沙

2010 年 12 月

目　录

第 一 篇

预备知识篇

第 1 章　分子生物学基础

在过去的十多年间，计算机技术在生命科学和医学的各个领域中发挥了前所未有的重要作用。特别是人类基因组计划实施以来，新的高效实验技术 (尤其是 DNA 测序技术) 的广泛应用，使得人们已经从细胞、器官、生物个体甚至生物群体等不同的生物层次获取到了海量的数据信息。巨大的数据量对于实验设计、数据处理和结果解释等方面的计算机技术支持提出了严峻挑战，融合生命科学与计算机、信息科学于一体的计算生物学 (或称生物信息学) 就是在这种背景下产生并迅速发展起来的。

生物信息学产生至今，已经在数据存储与管理、信息获取与分析等方面 (如基因序列拼接与组装、DNA 微阵列数据、蛋白质结构预测) 取得了极大的成功。随着生命科学的迅猛发展，人们已将研究重点从基因组转向了蛋白质组、从序列转向了功能，生命研究进入了所谓的后基因组时代 (蛋白质组时代)。新的实验对象和研究领域的发展要求有强大的计算机信息技术的支持, 从而对生物信息学提出了更高要求和更大挑战。

在今后相当长的一段时期内，生物医学领域里的数据量仍将呈几何级数增长，同时，科学家们更加关注这些多样化的数据之间的复杂集成关系，以便尽快挖掘与提取出使人类直接受益的知识与信息。以基因组测序工程为例，随着全球各个测序项目的不断进展, 研究重点正逐步地从积累原始数据转移到如何解释这些数据上来，未来生物学的新发现将不仅依赖于对传统领域 (如分子遗传、转基因、个体克隆等) 的继续关注，而且更依赖于在多尺度、全方位上对生物数据的组合和关联分析能力。序列数据将与结构和功能数据、基因表达数据、生化反应途径数据、表现型和临床数据等一系列数据相互集成，生物信息学的基础研究也将更加致力于解决生命科学中与系统和集成相关的问题。所有这些问题的解决都需要研究者探索各种新的思路和方法。

从另一角度看，传统的计算机科学也从生命科学的新挑战中获得了新的活力，计算机信息科学的研究领域日益扩大，一批交叉学科的新概念大量产生，并成为当前研究的热点，例如，遗传算法、人工神经网络、人造免疫系统、DNA 计算以及 VLSI-DNA 混合芯片等各类“仿生”方法研究。国际知名出版商 Springer-Verlag 公司从 2003 年起，就在其出版的计算机科学讲义 (lecture notes in computer science, LNCS) 丛书中，新增了生物信息学讲义 (lecture notes in bioinformatics, LNBI) 子系列，将其与原来的人工智能讲义 (lecture notes in artificial intelligence, LNAI) 子系列并列至同等重要的位置。此外，美国电气和电子工程师学会 (IEEE) 每年主办或资助的学术会议中，主题包含生物信息学内容的占了相当比例。所有这些表明，新的交叉领域研究不仅丰富了生物信息学科的内容，也扩展了各个独立学科的相关领域。可以预计，生物信息学在未来的几十

年中会得到进一步发展，正如 Baldi 所说的那样：基于“碳”的生物体信息处理和基于“硅”的电子化信息处理之间的界限，从概念到实现都已经开始在逐渐淡化了 [18]。

要学习计算机生物学，显然离不开经典的生物学、特别是现代分子生物学的知识。我们假定本书的读者具有这方面的初步知识，即使不是这样也没关系，本章将列出一些分子生物学方面的基础知识，主要是提供给不熟悉现代分子生物学的读者，也仅列出了为了理解本书内容可能用到的知识，现代分子生物学领域更深入的内容请读者参阅相关的教材或参考书。

1.1 生命的演化与分类

地球上有生命的生物体的演化历史大体可划分为几个阶段，进程的简要示意如图1.1所示。如果将地球产生以来的时间比作一个月 30 天，那么生命产生于第 3 天，多细胞生物出现于第 24 天，人类产生于最后 10 分钟，有文字记载的历史只相当于最后 5 秒钟时间。

- 生命诞生、细胞的形成
  - 海洋孕育原始生物 38∼35 亿年前
- 单细胞生物在地球上繁衍、原始生态系统的建立
  - 原核生物占主体的阶段 35∼20 亿年前
  - 真核生物占主体的阶段 20∼6 亿年前
- 多细胞生物出现和多样化表达、生物圈覆盖整个地球
  - 多细胞植物、多细胞动物的诞生 6∼5.5 亿年前
  - “寒武纪大爆发”(物种“爆发”) 约 5.5∼5.3 亿年前
  - “志留纪大爆发”(生命从海洋扩展到陆地) 约 4.3 亿年前
- 人类诞生、文明的发展
  - 人类诞生 约 400 万年前
  - 文化史 100 万年前

图 1.1 地球生命演化进程

自从有生命诞生以来，地球上的生物种类日益庞大，为了对生物进行系统的研究，瑞典博物学家林奈(Karl von Linnee)建立了一套对所有生物进行多级分类管理的标准命名体系，这套所谓的“双命名法(binomial nomenclature)”成为生物学研究中的规范而被沿用至今。在该体系中，将所有生物的分类划分为树状的 7 个层次，从高到低依次为：界、门、纲、目、科、属、种，同时根据需要，在每个层次中，可使用两个前缀词：超

(super-) 或亚 (sub-)，如表1.1所示。

表 1.1 林奈分类体系中的层次 (以人为例)

界	kingdom			动物界	Animalia
门	phylum	脊索动物门	Chordata	脊椎动物亚门	Vertebrata
纲	class	哺乳动物纲	Mammalia	真兽亚纲	Eutheria
目	order	灵长目	Primates	类人猿亚目	Anthropoidea
科	family			人科	Hominidae
属	genus			人属	Homo
种	species			人种	Sapiens

每种生物归类到树状层次的最底层 (即“种”这一层)，并采用拉丁文为其命名，称为该生物的学名。学名都采用相同的构造法，即

$$\text{学名} = \text{属名称} + \text{种名称}$$

例如，现代人的学名 (中文称为智人) 记作 *Homo sapiens*，或缩略为 *H. sapiens*。学名要求，属名称的首字母采用大写形式，种名称全部采用小写形式；在特殊情况下，学名的右端还可标注发现的地名或发现者的名字。在书写与印刷时，学名总使用斜体表示。学名之所以采用拉丁文，主要是因为它是一种目前没有大面积人群使用的、不再演化发展的语言，因而语法固定、不易产生歧义。

由于物种数量庞大，生物学家在研究时，根据研究目标的不同，通常选取一些具有代表性的样板生物进行研究，通常称之为模式生物(model organism)。常见的一些模式生物包括 (括号内为其学名或俗名)：噬菌体 (*Bacteriophage*)、病毒 (Virus)、大肠杆菌 (*Escherichia coli*)、酵母 (*Saccharomyces cerevisiae, yeast*)、秀丽线虫 (*Caenorhabditis elegans, worm*)、果蝇 (*Drosophila melanogaster*)、拟南芥 (*Arabidopsis thaliana*)、水稻 (*Oryza sativa*)、非洲爪蟾 (*Xenopus lavias*)、斑马鱼 (*Brachydanio rerio*) 和小鼠 (*Mus musculus*) 等。

1.2 核酸：DNA 与 RNA

核酸与蛋白质是生物体内两类重要的生物大分子，其中核酸是生物遗传信息的载体，遗传信息从上一代传递到下一代，核酸分子在其中起到核心作用。

核酸的基本组成单元是核苷酸，每个核苷酸的化学组成都包括一个磷酸基团、一个糖基团和一个含有氮原子的碱基基团 (见图1.2(a))。生物体内共存在两种类型的糖基团，根据这两种糖基团的不同，可将核苷酸分为脱氧核糖核苷酸和核糖核苷酸，相应的核酸也分为脱氧核糖核酸和核糖核酸，分别简称为DNA(脱氧核糖核酸)和RNA(核糖核酸)。

我们先以 DNA 的核苷酸为例，所有 DNA 核苷酸的磷酸基团和糖基团都是相同的，但共有 4 种碱基类型：腺嘌呤、鸟嘌呤、胞嘧啶和胸腺嘧啶，分别用字母 A、G、C 和

T 表示。根据碱基的不同，也可以将整个 DNA 核苷酸分成腺嘌呤脱氧核糖核苷酸、鸟嘌呤脱氧核糖核苷酸、胞嘧啶脱氧核糖核苷酸和胸腺嘧啶脱氧核糖核苷酸4 种类型，为了简化书写，通常仍分别用表示碱基类型的 A、G、C、T 四个字母来表示相应的脱氧核糖核苷酸。核苷酸糖基团上的碳原子依次编号为 $1', 2', \cdots, 5'$(见图1.2(b))，多个核苷酸可以相互连接形成一个长的 DNA 单链，其中每个核苷酸中环状糖基的 $5'$ 位置与上一个核苷酸相连， $3'$ 位置与下一个核苷酸相连 (见图1.2(c))。由于 DNA 单链在结构上并不对称，因此具有方向性，习惯上按照从 $5'$ 到 $3'$ 的方向书写和使用。

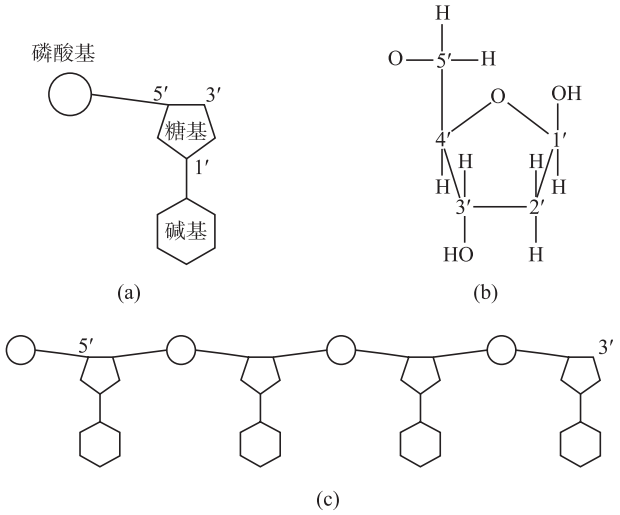


图 1.2 DNA 的组成与一级结构

不同 DNA 单链上的一对核苷酸，或者同一 DNA 单链上不相邻的一对核苷酸可以通过各自碱基之间形成的氢键相结合，从而形成碱基对(或核苷酸对)，这种配对具有特异性，即 A 只能与 T 形成配对，C 只能与 G 形成配对，称为碱基的互补配对。通过碱基互补配对，两条互相平行的 DNA 单链盘绕形成双链，其中两条互补的单链在方向上是相反的，每条单链称为另一条单链的反向互补链。通常的 DNA 都是由互补配对的双链组成的，在空间结构上，DNA 分子中的脱氧核糖与磷酸交替连接，排在双链空间结构的外链，而碱基排列在双链空间结构的内侧，并通过碱基互补配对形成的氢键将两条单链固定在一起，这种复杂的空间结构称为 DNA 的双螺旋结构，如图1.3所示。

通常在不强调 DNA 空间结构的时候，可将 DNA 序列中每条单链写成字母表 $\Sigma = \{\mathrm{A}, \mathrm{G}, \mathrm{C}, \mathrm{T}\}$ 上的字符串的形式，对应于从 $5' \to 3'$ 方向顺序，这称为DNA 序列，或DNA 的一级结构。为了表达 DNA 双链的特征，在书写 DNA 序列时，经常将两条序列同时上下并列写出，并标注方向。习惯上，将 $5' \to 3'$ 方向的序列放在上方，互补的序列放在下方。

例 1.1 假定 $s = \mathrm{ACGCTGC}$ 为 DNA 序列中的一条单链，则其反向互补链为 $\overline{s} = \mathrm{GCAGCGT}$，写成双链形式如下

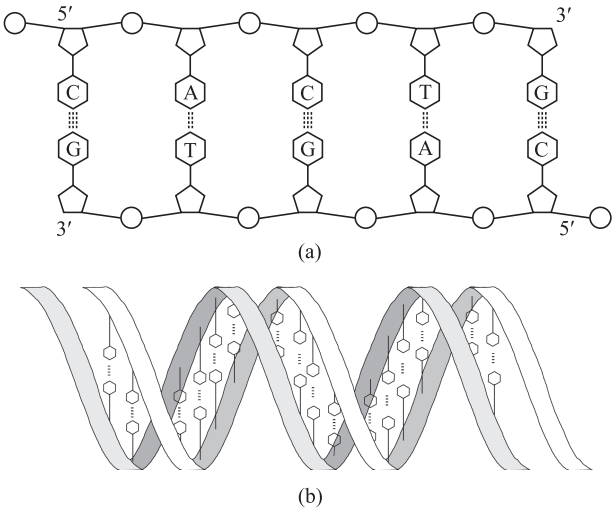


图 1.3 DNA 双螺旋空间结构

$$
\begin{array}{lccccc}
s: & 5' & \cdots & \texttt{ACGCTGC} & \cdots & 3' \\
\overline{s}: & 3' & \cdots & \texttt{TGCGACG} & \cdots & 5'
\end{array}
$$

由于两条单链只要已知其中之一，另一条便可唯一确定，因此实践中经常只列出一条 DNA 序列。

度量 DNA 序列长度的常用方法是用 DNA 分子中所包含的核苷酸碱基数目 (即表示 DNA 双链序列的字符总数)，由于 DNA 通常具有的双链结构，计算长度的单位也通常用“碱基对”数目，简写为bp(base pair)。

尽管与 DNA 一样，RNA 也是一条比较长的核酸链，但它与 DNA 具有不同的性质。

(1) RNA 通常是单链形式 (single stranded, ssRNA)；

(2) RNA 中的糖是核糖，而不是脱氧核糖；

(3) RNA 中只有尿嘧啶(Uracil, U)而没有胸腺嘧啶(thymine, T)；

(4) DNA 主要存在于细胞核内，但 RNA 则在细胞核外的细胞质中也存在。

1.3 蛋 白 质

蛋白质是由氨基酸用肽键相连接起来的线性聚合物。蛋白质的平均长度为 200 个左右的氨基酸，不过大些的蛋白质可以达到上千个氨基酸。在更大粒度上，细胞通常含有多种蛋白质，其重量将占细胞干重的一半以上。蛋白质决定细胞的形状与结构，同时蛋白质也是分子识别及催化作用的主要主体。

每个氨基酸是由一个位于中心的碳原子 (记为 C_α) 以及用共价键跟它相连的四个基团 (或原子) 组成的，这四个基团 (或原子) 中，一个为氨基团 (NH_2)，一个为羧基团

(COOH)，还有一个为氢原子，最后一个为称侧链基团(用 R 表示)，如图1.4所示。正是由于 R 基团的不同，可以将氨基酸划分为很多种，它们之间的主要差别在于 R 位上连接的基团不同。当多个氨基酸相互连接形成蛋白质时，每个氨基酸上相同的部分称为骨架或主链骨架，而可变的 R 基团称为侧链。可见氨基酸之间的差异主要体现在侧链部分，这也同时决定了不同氨基酸的物理化学性质。

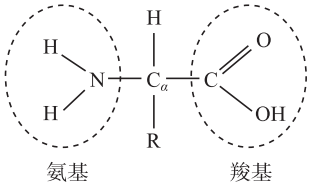


图 1.4 氨基酸的结构与组成

自然界存在有很多种氨基酸，但组成蛋白质标准元件的氨基酸共有 20 种，如表1.2所示，其中氨基酸的极性是刻画其与水溶剂亲和性的一种指标，氨基酸的极性对于蛋白质的结构具有重要的影响。表 1.2 中，残基表示失去一分子水的氨基酸；极性一栏中，(P) 表示亲水性，(H) 表示疏水性。

表 1.2 20 种氨基酸

英文名称	中文名称	三字母缩写	单字母缩写	分子式	残基平均质量 (Da)	极性
Alanine	丙氨酸	Ala	A	C_3H_5NO	71.0788	(H)
Arginine	精氨酸	Arg	R	$C_6H_{12}N_4O$	156.1875	(P)
Asparagine	天冬酰氨	Asn	N	$C_4H_6N_2O_2$	114.1038	(P)
Aspartic acid	天冬氨酸	Asp	D	$C_4H_5NO_3$	115.0886	(P)
Cysteine	半胱氨酸	Cys	C	C_3H_5NOS	103.1388	(P)
Glutamic acid	谷氨酰氨	Glu	E	$C_5H_7NO_3$	129.1155	(P)
Glutamine	谷氨酸	Gln	Q	$C_5H_8N_2O_2$	128.1307	(P)
Glycine	甘氨酸	Gly	G	C_2H_3NO	57.0519	(P)
Histidine	组氨酸	His	H	$C_6H_7N_3O$	137.1411	(P)
Isoleucine	异亮氨酸	Ile	I	$C_6H_{11}NO$	113.1594	(H)
Leucine	亮氨酸	Leu	L	$C_6H_{11}NO$	113.1594	(H)
Lysine	赖氨酸	Lys	K	$C_6H_{12}N_2O$	128.1741	(P)
Methionine	甲硫氨酸	Met	M	C_5H_9NOS	131.1926	(H)
Phenylalanine	苯丙氨酸	Phe	F	C_9H_9NO	147.1766	(H)
Proline	脯氨酸	Pro	P	C_5H_7NO	97.1167	(H)
Serine	丝氨酸	Ser	S	$C_3H_5NO_2$	87.0782	(P)
Threonine	苏氨酸	Thr	T	$C_4H_7NO_2$	101.1051	(P)
Tryptophan	色氨酸	Trp	W	$C_{11}H_{10}N_2O$	186.2132	(H)
Tyrosine	酪氨酸	Tyr	Y	$C_9H_9NO_2$	163.1760	(P)
Valine	缬氨酸	Val	V	C_5H_9NO	99.1326	(H)

多个氨基酸通过肽键相连接形成链状，称为肽链。一个氨基酸的氨基与相邻氨基酸

的羧基在形成肽键时，会缩去一分子水，如图1.5所示。肽链上所有氨基酸的次序称为肽链氨基酸序列，也称为蛋白质的一级结构。与 DNA 分子一样，肽链序列也有方向性。肽链的一端具有自由氨基基团(N 端)，另一端以羧基基团结束 (C 端)。整条肽链的主链骨架上依次为“氨基、 α 碳原子、羧基、氨基、……”排列，而侧链则分别连接到 α 碳原子上去。

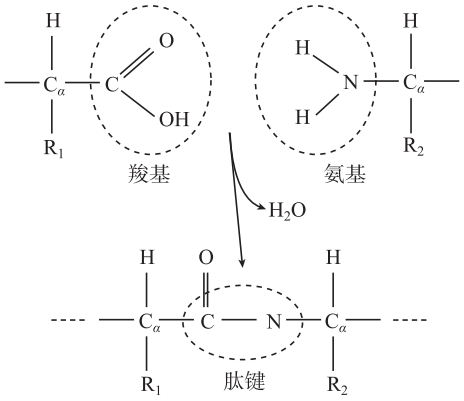


图 1.5 氨基酸的结构与组成

通常，一个完整的蛋白质是由多个肽链组成的。蛋白质具有复杂的空间结构，按其复杂程度，可将蛋白质结构划分为四个层次。组成蛋白质链的氨基酸序列称为其一级结构(primary structure)。序列的不同区域可以形成一些局部的二级结构(secondary structure)，例如，α 螺旋是氨基酸的单链螺旋，而β 片层则由序列片段“纺织”形成平面片状结构。将这些二级结构压缩打包成一个或多个三维的域 (domain)，就构成了蛋白质的三级结构(tertiary structure)。最后，蛋白质可能由多个蛋白域组成，这个层次的结构称为是蛋白质的四级结构(quaternary structure)。

蛋白质的结构 (一级到四级) 是由氨基酸序列及其之间的物理化学作用决定的。可以说，蛋白质的折叠结构是由其遗传物质唯一确定的，事实上，蛋白质的三维结构具有最小的自由能量。另一方面，蛋白质的结构同时又会影响到其功能。

1.4 DNA 的复制

DNA 的复制是指以原始的 DNA 分子为模板合成出相同分子的过程。DNA 复制过程的最大特点是“半保留复制”，即完成复制后所形成的两个子代 DNA 分子中，每一个分子中都有一条单链是来自亲代的旧链，另一条则是新合成的新链。DNA 复制过程中，首先在一些酶的作用下将作为模板的 DNA 分子的双链进行解螺旋，以便形成两条单链，这样在已完成解螺旋与未完成解螺旋的接口处形成一个三叉形，称为复制叉，由于解螺旋过程是随着时间动态进行的，因此复制叉的位置也沿着 DNA 序列不断推进。两条解螺旋后的单链都可以作为模板指导合成出与之反向互补的新链，并重新组合成双螺旋结

构。两条模板单链中，一条称为前导链，另一条称为滞后链。对于前导模板链，复制时沿着 $3' \to 5'$ 方向进行，按照反向互补配对原则合成新链的方向则是 $5' \to 3'$，新链合成的方向与复制叉进行的方向保持一致；对于滞后模板链，复制时沿 $5' \to 3'$ 方向进行，新链延伸的方向则为 $3' \to 5'$，合成方向与复制叉推进方向相反，因此只能采用分段接合的不连续方式进行合成。

1.5 基因与染色体

构成蛋白质的氨基酸序列是由 DNA 序列所编码决定的，特别地，将 DNA 上负责编码多肽的区域称为基因。1977 年，分子生物学家发现，大多数的真核生物的基因中存在着称为外显子(exon)的编码区域和非编码区域 (内含子(intron))，二者在基因中相间分布。在人类基因组中，编码区大约只占到 2%∼3%，其余 97%∼98% 的区域为非编码区。过去由于对非编码区的功能理解不足，一度认为它们对遗传效应没有作用，故将它们称为垃圾 DNA(junk DNA)。近些年，人类逐渐认识到非编码区对于基因调控等具有重要作用，但对非编码区的完整功能目前仍不是很清楚。

通常一个 DNA 分子中包含有多个基因区域，这样的 DNA 分子又称为染色体。在较为高等的真核生物细胞核中，染色体经常成对出现，因此它们互称为同源染色体。在同源的一对染色体中，一个来自亲代母本，另一个来自亲代父本。生物体所有染色体中包含的所有遗传信息合称为该种生物的基因组(genome)。除了少数例外，多细胞真核生物的每个细胞都包含了相同的全部基因组信息；由于基因表达会在不同组织细胞上有差别，因此不同组织中细胞的功能各异。人类基因组由 3×10^9 个碱基对(base pair)组成，分布在 23 对同源染色体上。其中含有 22 对常染色体及 1 对性染色体：XX 或 XY，这 24 种染色体的大小分布在 $50 \times 10^6 \sim 250 \times 10^6$bp 之间。不同生物所含 DNA 的总量是不同的。例如，*Amoeba dubia*(变形虫，一种单细胞生物) 的基因组大小约是人类的 200 多倍。

C 值是用每单倍体染色体组中的 DNA 含量来表示基因组的大小。一个物种的 C 值是恒定的，不同物种之间则差别较大。以前认为，随着生物的进化，生物体的结构和功能越来越复杂，其 C 值也就越大；然而现在认为，生物体复杂性和 DNA 的含量之间的关系变得模糊了。在结构功能很相似的同一类生物中，甚至亲缘关系十分接近的物种之间，C 值也可相差数十倍到上百倍。这种无法解释的现象称为C 值矛盾，或C 值悖论(C value paradox)。

1.6 基 因 表 达

在真核生物中，尽管编码蛋白质的 DNA 分子主要分布在细胞核之内，但蛋白质的最终合成场地核糖体却是位于细胞核之外的。遗传信息是如何从细胞核内的 DNA 传递到细胞核外的蛋白质呢？整个过程分成两个步骤，并主要以 mRNA 作为中介：先以