高等院校生物类专业系列教材

生物信息学

BIOINFORMATICS

主　编　叶子弘
副主编　张文英　柴　惠　贺平安

ZHEJIANG UNIVERSITY PRESS
浙江大学出版社

高等院校生物类专业系列教材

生物信息学

BIOINFORMATICS

主　编　叶子弘
副主编　张文英　柴　惠　贺平安

ZHEJIANG UNIVERSITY PRESS
浙江大学出版社

内 容 简 介

生物信息学是一门新兴的交叉学科，它融合了生物学、计算机科学与数学等知识，被誉为21世纪生物科学发展的主导学科。本教材首先简要介绍了国内外生物信息学的发展现状和生物信息分析研究方法的最新动向，介绍了生物信息学的基础知识和成熟的技术方法（序列比对、基因特征分析、引物设计、蛋白质结构预测等），阐述了生物信息数据库及分子生物信息相关的分析技术，包括生物信息数据库的结构分析和模拟构建技术，介绍了计算机辅助药物设计技术和化学计量学方法，辅以实例说明，并在每章后面罗列了相关分析方法、软件、工具及知识的重要免费网站作为知识拓展，以供参考。

本教材可作为非生物信息学专业的本科学生的生物信息学课程教材，也可作为生物学、农学、药物设计等领域工作者的参考用书。

图书在版编目(CIP)数据

生物信息学/叶子弘主编. —杭州：浙江大学出版社，2011.9

ISBN 978-7-308-09011-7

Ⅰ.①生… Ⅱ.①叶… Ⅲ.①生物信息论－教材 Ⅳ.①Q811.4

中国版本图书馆CIP数据核字（2011）第169229号

生物信息学

叶子弘 主编

丛书策划 樊晓燕 季 峥

责任编辑 季 峥(really@zju.edu.cn)

封面设计 林智广告

出版发行 浙江大学出版社

（杭州市天目山路148号 邮政编码310007）

（网址：http://www.zjupress.com）

排 版 杭州大漠照排印刷有限公司

印 刷 德清县第二印刷厂

开 本 787mm×1092mm 1/16

印 张 20.5

字 数 525千

版 印 次 2011年9月第1版 2011年9月第1次印刷

书 号 ISBN 978-7-308-09011-7

定 价 40.00元

前　言

生物信息学是一门新兴的交叉学科。它融合了生物学、计算机科学与数学等知识，被誉为21世纪生物科学发展的主导学科。随着测序技术的不断发展，目前已经测序的生物基因组数量超过1000个，生物学数据正在急速和海量地积累。这些海量的生物学数据中隐藏着大量人类目前尚未知的生物学信息和知识。如何充分挖掘这些海量数据的内涵，并从中获取有用的信息，揭示生命的奥秘，是生物信息学的重要使命和亟待解决的问题。

近年来，国内生物信息学专业发展较快，很多综合性高等院校均设置了该专业。有的院校虽然没有设置此专业，但也在生物类与信息科学类等专业开设相关课程，并作为模块进行毕业设计。同时，很多生物专业学生希望报考生物信息专业的研究生。此外，生物信息学相关知识和技术已经成为目前从事生物学、农学、药物设计、生物制品等相关领域研究和开发的重要知识背景和必要手段。因此，需要为更广阔的具有不同学科背景的非生物信息学专业的学生提供生物信息学相关内容的学习课程，从而有利于其更好地参与需要生物信息学知识背景或技术手段的相关研究与工作。从长远来讲，这对加强生物信息学的人才培养与研究工作队伍建设具有非常重要的意义。编写本书的目的是使非生物信息学专业，如生物学、农学、计算机科学、数学与统计科学等专业的学生能够了解生物信息学的基本内涵、发展趋势以及常用的分析工具和分析方法。

本书是集编写老师们的集体智慧撰写完成的，具体分工为：第1章由叶子弘（中国计量学院）编写；第2章由姚玉华（浙江理工大学）编写；第3章由张林（浙江中医药大学）编写；第4、9章由柴惠（浙江中医药大学）编写；第5、8章由贺平安（浙江理工大学）、张文英（长江大学）编写；第六章由金园庭（中国计量学院）编写；第7章由赵彦宏（鲁东大学）编写；第10章由聂作明（浙江理工大学）编写；第11章由阮松林（杭州市农业科学院）编写；第12章由徐路（大理学院）编写。全书由叶子弘和张文英统稿和定稿。

生物信息学是一门新兴的交叉学科，相关的新技术、新进展不断涌现，资料浩瀚。由于编者水平有限，在本书编写过程中，难免出现疏漏和错误之处，恳请同行专家和读者批评指正，不胜感激。

叶子弘

2011年5月于杭州

目　　录

第1章

生物信息学概述

“生物信息学”的名词最早出现在1956年美国田纳西州的Gatlinburg召开的首次“生物学中的信息理论讨论会”上。生物信息学(bioinformatics)是建立在数学、计算机科学和生命科学基础之上的一门交叉科学。它包括生物信息的获取、加工、存储、分发、分析和解释等各方面,综合运用数学、计算机科学和生物学的各种工具,来阐明和理解大量数据所包含的生物学意义。随着相关生物技术的革命性发展和生物学相关信息量呈现的“革命性爆炸”,生物信息学已成为当今最具发展前途的学科之一。

生物信息学的出现极大地推动了分子生物学、基因组学、蛋白质组学和代谢组学等的发展,已经成为医学、农学、生物学等学科发展的强大推动力,也是药物设计、环境监测等的重要技术支撑。生物信息学在基因的功能发现、疾病基因诊断、蛋白质结构预测、基于结构的药物设计、药物合成和制药工业中起着极其重要的作用,生物信息学的应用大大加快了药物的研究开发进程。

本章介绍了生物信息学的发展史、主要研究内容,并对生物信息学的作用及发展方向等进行了展望,以期让学生对生物信息有个总体的了解和认识。

1.1 生物信息学的发展史

生物信息学是建立在分子生物学的基础上的。早在20世纪50年代,生物信息学就已经开始孕育,其间,科学家已经通过实验测定一些蛋白质的序列。例如,1947年测出短杆菌的五肽结构;1951年重构胰岛素的30个氨基酸。几乎在同一时期,科学家认识到DNA是遗传物质。1949年,发现了DNA链中A═T、G═C的规律;1951年,Pauling和Corey提出蛋白质的α-螺旋和β-折叠结构;1953年,Watson和Crick根据Franklin和Wilkins得到的X射线衍射数据提出DNA的双螺旋结构模型,揭开了分子生物学研究的序幕。

1956年在美国田纳西州的Gatlinburg召开了首次“生物学中的信息理论研讨会”。在20世纪60年代,虽然当时没有具体地提出生物信息学的概念,但是一些计算生物学家开始进行相关研究,做了许多生物信息搜集和分析方面的工作。在这个时期,生物大分子携带信息成为分子生物学的重要理论,生物分子信息在概念上将计算生物学和计算机科学联系起来。大量的生物分子序列成为丰富的信息源,科学家们开始应用计算方法分析这些信息。相关或者同源蛋白质序列之间的相似性首先引起人们的注意。1962年,Zucherkandl和Pauling研究了序

列变化与进化之间的关系，开创了一个新的领域——分子进化。随后，通过序列比对确定序列的功能及序列分类关系成为序列分析的主要工作。氨基酸序列的收集是这个时期的一项重要工作，1967 年，Dayhoff 研制出蛋白质序列图集，该图集后来演变为著名的蛋白质信息源 PIR。20 世纪 60 年代是生物信息学形成雏形的阶段。

生物信息学是一门相当年轻的学科，一般认为，生物信息学的真正开端是 20 世纪 70 年代。从 20 世纪 70 年代到 80 年代初期，随着生物化学技术的发展，产生了许多生物分子序列数据，而在这个阶段数学统计方法和计算机技术都得到较快的发展，这促使一部分计算机科学家应用计算机技术解决生物学问题，特别是与生物分子序列相关的问题。他们开始研究生物分子序列，研究如何根据序列推测结构和功能。这时，生物信息学开始崭露头角。

从 20 世纪 70 年代初期到 80 年代初期，出现了一系列著名的序列比对方法。其中，Needleman 和 Wunsch 于 1970 年提出的序列比对算法是对生物信息学发展最重要的贡献。同年，Gibbs 和 McIntyre 发表的矩阵打点作图法也是进行序列比对的一个著名方法，该方法可用于寻找序列中的重复片断，从而推测其功能。Dayhoff 提出的基于点突变模型的 PAM 矩阵是第一个广泛使用的比较氨基酸相似性的得分矩阵，它大大地提高了序列比对算法的性能。1980 年，《*Science*》杂志发表了关于计算分子生物学的综述。1981 年，Smith 和 Waterman 提出了著名的公共子序列识别算法。同年，Doolittle 提出关于序列模式的概念。1983 年，Wilbur 和 Lipman 发表了数据库相似序列搜索算法。1985 年，出现了快速的蛋白质序列搜索算法 FASTP/FASTN。1988 年，Pearson 和 Lipman 发表了著名的序列比对算法 FASTA。1990 年，快速相似序列搜索算法 BLAST 问世。1997 年，BLAST 的改进版本 PSI－BLAST 投入实际应用。

在 20 世纪 70 年代，还不断涌现出许多生物信息分析方法。1972 年，Gatlin 将信息论引入序列分析，证实自然的生物分子序列是高度非随机的。1977 年，出现了将 DNA 序列翻译成蛋白质序列的算法。1975 年，继第一批 RNA(tRNA)序列发表后，Pipas 和 McMahon 首先提出运用计算机技术预测 RNA 二级结构。1978 年，Gingeras 等人研制出核酸序列中限制性酶切位点的识别软件。

20 世纪 80 年代以后，出现了一批生物信息服务机构和生物信息数据库。1982 年，核酸数据库 GenBank 第 3 版公开发行。1986 年，日本核酸序列数据库 DDBJ 诞生。1986 年，蛋白质数据库 SWISS－PROT 问世。1988 年，美国国家卫生研究所和美国国家图书馆成立国家生物技术信息中心 NCBI。同年，欧洲分子生物学网络 EMBnet 成立，专门发布各种生物数据库。

20 世纪 90 年代后，科学家们开始进行大规模的基因组研究。1986 年，出现了基因组学(genomics)概念，即研究基因组的作图、测序和分析。1990 年，国际人类基因组计划启动，该计划被誉为生命科学的“阿波罗登月计划”。1993 年，Sanger 中心成立，专门从事基因组研究。1995 年，第一个细菌基因组被完全测序。1996 年，酵母基因组被完全测序。1996 年，Affymetrix 生产出第一块 DNA 芯片。1998 年，第一个多细胞生物——线虫的基因组被完全测序。1999 年，果蝇的基因组被完全测序。1999 年年底，国际人类基因组计划联合研究小组宣布人类第一次获得一对完整的人类染色体——第 22 对染色体的遗传序列。2000 年 6 月 24 日，人类基因组计划联合研究小组的六个国家研究机构在全球同一时间宣布已完成人类基因组的工作框架图。生物信息学在人类基因组计划的推动之下迅速发展。

图 1－1 描绘了 1973—2000 年生物医学文献数据库 PubMed 中搜集的与生物信息学相关

论文的历年统计结果。该图用有关生物信息学论文数量的变化来说明何时是生物信息学的形成初期，何时是生物信息学的迅速发展期。

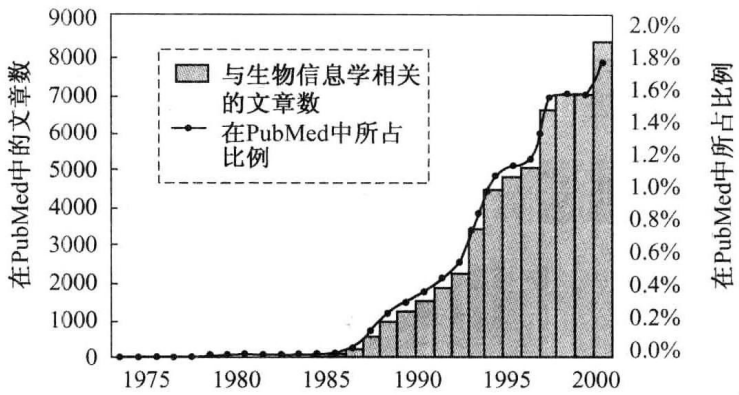

图 1-1　1973—2000 年 PubMed 中与生物信息学相关论文统计

生物信息学的核心内容是研究如何通过对 DNA 序列的统计计算分析，更加深入地理解 DNA 序列、结构、演化及其与生物功能之间的关系。其研究课题涉及分子生物学、分子演化及结构生物学、统计学及计算机科学等许多领域。生物信息学是内涵非常丰富的学科，其核心是基因组信息学，包括基因组信息的获取、处理、存储、分配和解释。基因组信息学的关键是“读懂”基因组的核苷酸顺序，即全部基因在染色体上的确切位置以及各 DNA 片段的功能；同时在发现了新基因信息之后进行蛋白质空间结构模拟和预测，然后依据特定蛋白质的功能进行药物设计。了解基因表达的调控机理也是生物信息学的重要内容。基因表达调控的研究目标是揭示“基因组信息结构的复杂性及遗传语言的根本规律”，解释生命的遗传语言。

无论从理论上来讲还是从实际情况来看，生物信息学的实质就是利用计算机科学和技术来解决生物学问题。生物信息学的诞生是由生物学对大量数据处理和分析的需求而引发的，是历史的必然。作为一门交叉学科，生物信息学的发展依赖于计算机科学技术和生物技术的发展，而生物信息学的研究成果又促进了生物学特别是分子生物学的发展。生物信息学已成为整个生命科学发展的重要组成部分，成为生命科学研究的前沿。

1.2　生物信息学主要研究内容

纵观当今生物信息学界的现状可以发现，虽然生物信息学诞生较晚，但在短短几十年间，已经形成了多个研究方向，研究内容涵盖基因组、蛋白质组、蛋白质结构以及与之相结合的药物设计等方面，其研究内容大致包括以下几个方面：

1.2.1　序列比对

序列比对(sequence alignment)是生物信息学的基础，是生物信息学的核心研究内容之一。在生物学研究过程中，为了确定新测序列的生物属性，经常需要进行序列同源性分析，就是将新序列加入到一组与之同源，但来自不同物种的序列中进行多序列同时比较，以确定该序

列与其他序列间的同源性大小。这是理论分析方法中最关键的一步。完成这一工作通常使用序列比对的方法(图 1-2)。不仅如此,对于蛋白质结构预测等,序列比对也是最为重要的一种方法。

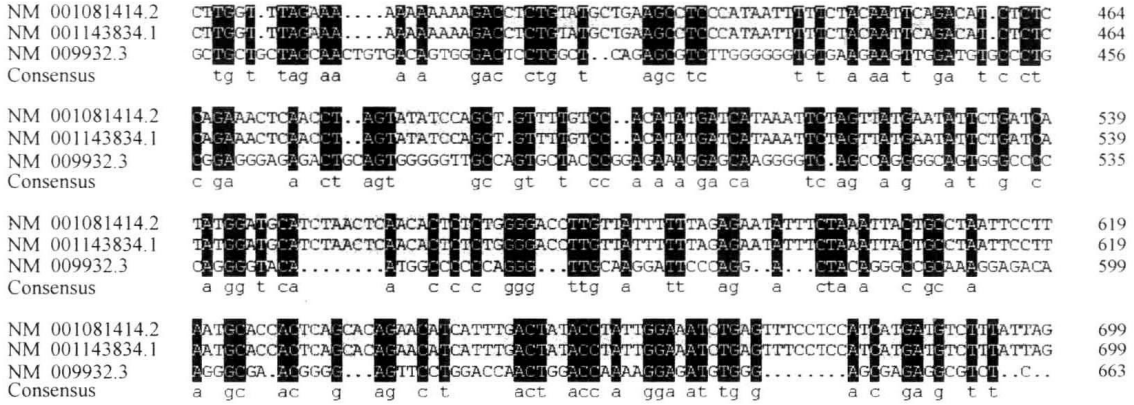

图 1-2 序列比对示意图

序列比对的理论基础是进化学说。如果两个序列之间具有足够的相似性,就推测两者可能有共同的进化祖先,经过序列内残基的替换、残基或序列片段的缺失以及序列重组等遗传变异过程分别演化而来。序列相似和序列同源是不同的概念,序列之间的相似程度是可以量化的参数,而序列是否同源需要有进化事实的验证。

序列比对的基本问题是比较两个或两个以上符号序列的相似性或不相似性。从生物学的初衷来看,这一问题包含了以下几个意义:从相互重叠的序列片断中重构 DNA 的完整序列;在各种试验条件下从探测数据(probe data)中决定物理和基因图存贮;遍历和比较数据库中的 DNA 序列;比较两个或多个序列的相似性;在数据库中搜索相关序列和子序列;寻找核苷酸(nucleotides)的连续产生模式;找出蛋白质和 DNA 序列中的信息成分序列;比对考虑 DNA 序列的生物学特性;如序列局部发生的插入、删除(InDel)和替代;依据序列的目标函数,获得序列之间突变集最小距离加权和或最大相似性和对齐的方法包括全局对齐、局部对齐、代沟惩罚等。两个序列比对常采用动态规划算法,这种算法在序列长度较小时适用,然而对于海量基因序列(如人的 DNA 序列高达 109Mb),这一方法就不太适用。因此,启发式方法的引入势在必行,著名的 BALST 和 FASTA 算法及相应的改进方法均是从此前提出发的。

1.2.2 蛋白质结构比对和结构预测

蛋白质的结构与功能是密切相关的,一般认为,具有相似功能的蛋白质结构一般相似。蛋白质是由氨基酸组成的长链,一般有 50～3000 个氨基酸残基,蛋白质具有多种功能,如作为酶和抗体、物质的存贮和运输、信号传递等等。氨基酸的序列决定了蛋白质的三维结构。一般认为,蛋白质有四级不同的结构。蛋白质结构比对的基本问题是比较两个或两个以上蛋白质分子空间结构的相似性或不相似性。

基于氨基酸序列预测蛋白质自然结构仍然是分子生物学中最重要且尚未解决的问题。其主要理论依据是假定蛋白质的自然构象处于自由能的极小位置。大量的蛋白质从变性状态重新折叠的实验都给予这个假设以事实的支持:如果改变蛋白质的外界环境条件,比如温度、压

力或者溶剂条件，那么蛋白质就会失去折叠，并且失去活性；但是一旦环境条件恢复到正常生理状态，蛋白质又会自发地折叠成其天然结构，并且恢复活性。因此，蛋白质的折叠过程很明显是一个热动力学过程，而且形成蛋白质天然三维结构所需要的全部信息都包含在相应的蛋白质序列当中。蛋白质二级结构的预测通常被认为是蛋白质结构预测的第一步，是根据预测的局部结构，对蛋白质序列中的氨基酸的二级结构类型进行分类。但是蛋白质的二级结构在一定程度上受远程残基的影响，尤其是 β-折叠。从理论上来说，局部信息仅包含二级结构信息的 65%左右，因此，可以想象只用局部信息的二级结构预测方法，其准确率不会有太大的提高。而蛋白质的三维结构比其一级结构在进化过程中得到更稳定的保留，同时也包含了较氨基酸序列更多的信息。蛋白质三维结构研究的前提假设是内在的氨基酸序列与三维结构一一对应(不一定全真)，物理上可用最小能量来解释。

蛋白质结构预测的方法主要有演绎法和归纳法两种。前者主要是从一些基本原理或假设出发来预测和研究蛋白质的结构和折叠过程。分子力学和分子动力学属这一范畴。后者主要是从观察和总结已知结构的蛋白质结构规律出发来预测未知蛋白质的结构。同源建模和指认(threading)方法属于这一范畴。同源建模用于寻找具有高度相似性的蛋白质结构(超过 30% 氨基酸相同)；后者则用于比较进化族中不同的蛋白质结构。蛋白质结构和预测的研究具有重要的研究意义和应用价值，医药上可以用来理解生物的功能，寻找 dockingdrugs 的目标，农业上可以藉此通过基因工程获得更好的农作物，工业上相关研究将有助于酶的合成。虽然已经过 30 余年的努力，目前的蛋白结构预测研究仍远远不能满足实际需要，相关研究亟待进一步发展。

1.2.3　基因识别和非编码区分析

1. 基因识别

基因识别是生物信息学的一个重要分支，通过使用生物学实验或计算机等手段识别 DNA 序列上的具有生物学特征的片段。基因识别的对象主要是蛋白质编码基因，也包括其他具有一定生物学功能的因子，如 RNA 基因和调控因子。基因识别是基因组研究的基础。识别具有生物学功能的片段与判定该片段(或其对应的产品)的功能是两个不同的概念，后者通常需要通过基因敲除等实验手段来决定。不过，生物信息学的前沿研究正使得由基因序列预测基因功能变得可能。

基因识别方法主要有两种：一种是基于序列比对方法；另一种是所谓的从头计算方法。基于序列比对方法考虑到同源蛋白质基因结构具有相似性，通过比较位置序列和已知基因之间的相似性，判断未知序列是否为基因，采用的工具有 FASTA 和 BLAST。从头计算方法考虑到基因结构具有保守性的特点，分析已知基因结构特征，提取信息参量，建立理论模型，设计算法达到识别基因的目的。

早期基因识别的主要手段是基于活细胞或生物实验。通过对若干种不同基因的同源重组速率的统计分析，我们能够获知它们在染色体上的顺序。若进行大量类似的分析，我们可以确定各个基因的大致位置。现在，由于人类已经获得了巨大数量的基因组信息，依靠较慢的实验分析已不能满足基因识别的需要，而基于计算机算法的基因识别得到了长足的发展，成为了基因识别的主要手段。目前在计算机辅助基因识别方面已有了数十种算法，有十种左右重要的

算法和相应软件提供网上免费服务。原核生物计算机辅助基因识别相对容易些，结果可靠些。从具有较多内含子的真核生物基因组序列中正确识别出起始密码子、剪切位点和终止密码子是个相当困难的问题，目前的研究现状不令人满意，仍有大量的工作要做。

2. 非编码区分析

众所周知，生命是由基因组决定的。每个生物都具有基因组，携带着构成和维持该生物体生命形式所必需的所有生物信息。不论是原核生物基因组还是真核生物基因组，都由两部分组成：编码区和非编码区。DNA 序列作为一种遗传语言，既包含在编码区，又隐含在非编码序列中。非编码区包含如下类型的 DNA 成分或由其表达的 RNA 成分：内含子、卫星 DNA、小卫星 DNA、微卫星 DNA、非均一核 RNA(hmRNA)、短散置元(SINE)、长散置元(LINE)、伪基因等。除此之外，顺式调控元件，如启动子、增强子等也属于非编码序列。非编码区曾被称为“垃圾”DNA，其实它们并非“垃圾”，只是我们暂时还不知道其重要的功能。

非编码序列在原核生物基因组与真核生物基因组中所占的比例有很大的区别。原核生物基因组较小，非编码蛋白质的区域只占整个基因组序列的 10%～20%，编码区在基因组中所占的比例很高(80%～90%)，基因常以操纵子形式组织。相比之下，真核生物，例如人类基因组含有约 32 亿对碱基对，其中仅有 3%～5%编码约 2 万～2.5 万个基因，而其余 95%～97%均为非编码区，这表明这些非编码序列必定具有重要的生物功能。普遍的认识是，它们与基因的表达调控有关。关于编码区的相关研究已经缔造了数十名诺贝尔奖获得者，非编码区蕴含的成果将是十分可观的，因此寻找这些区域的编码特征、信息调节与表达规律是未来相当长时间内的热点课题。

分析非编码区 DNA 序列目前没有一般性的指导方法。对非蛋白编码区进行生物学意义分析的策略主要有两种：一种是基于已被实验证实的、功能已知的 DNA 元件的序列特征，预测非蛋白编码区中可能含有的功能已知的 DNA 元件，从而预测其可能的生物学功能，并通过实验进行验证；另一种则是通过数理理论直接探索非蛋白编码区的新的未知序列特征，并从理论上预测其可能的信息含义，最后同样通过实验验证。侦测密码区的方法包括测量密码区密码子(codon)的频率、一阶和二阶马尔可夫链、ORF(open reading frames)、启动子(promoter)识别、HMM(hidden markov model)和 GENSCAN、Splice Alignment 等等。

1.2.4 分子进化和比较基因组学

1. 分子进化

分子进化和比较基因组学是生物信息学最重要的课题之一。分子进化是利用不同物种中同一基因序列的异同来研究生物的进化，构建进化树。既可以用 DNA 序列也可以用其编码的氨基酸序列来研究分子进化，甚至还可通过相关蛋白质的结构比对来进行，其前提是假定相似种族在基因上具有相似性。通过比较可以在基因组层面上发现哪些是不同种族中共同的，哪些是不同的。早期研究方法常采用外在的因素，如大小、肤色、肢体的数量等等作为进化的依据。随着近年来较多模式生物基因组测序的完成，人们可从整个基因组的角度来研究分子进化。在匹配不同种族的基因时，一般需处理三种情况：orthologous——不同种族、相同功能的基因；paralogous——相同种族、不同功能的基因；xenologous——有机体间采用其他方式传递的基因，如被病毒注入的基因。这一领域常采用的方法是构造进化树，通过基于特征(即

DNA 序列或蛋白质中的氨基酸的碱基的特定位置)和基于距离(对齐的分数)的方法和一些传统的聚类方法(如 UPGMA)来实现。

2. 比较基因组学

比较基因组学是基于基因组图谱和测序基础,对已知的基因和基因组结构进行比较,来了解基因的功能、表达机理和物种进化的学科,利用模式生物基因组与人类基因组之间编码顺序上和结构上的同源性,克隆人类疾病基因,揭示基因功能和疾病分子机制,阐明物种进化关系及基因组的内在结构。目前从模式生物基因组研究中已得出一些规律:模式生物基因组一般比较小,但编码基因的比例较高,重复顺序和非编码顺序较少;其 GC 含量比较高;内含子和外显子的结构组织比较保守,剪切位点在多种生物中一致;DNA 冗余,即重复;绝大多数的核心生物功能由相当数量的 orthologous 蛋白承担;synteny 连锁的同源基因在不同的基因组中有相同的连锁关系等。模式生物基因组研究协助揭示了其他生物基因的功能,利用基因顺序上的同源性克隆人类或其他生物重要基因,如人类疾病基因。利用模式生物实验系统上的优越性,应用比较作图分析复杂性状,加深对基因组结构的认识。

1.2.5　序列重叠群装配

一般来说,根据现行的测序技术,每次反应只能测出 500 或稍多一些碱基对的序列,如人类基因的测量就采用了鸟枪法(shotgun)。这就有一个把大量的较短的序列拼接成一个较大的、完整序列的任务。显然,为了正确拼接,短的序列之间应有一部分重叠区(contigs)。所有相互部分重叠的序列全体构成了重叠群。逐步把它们拼接起来形成序列更长的重叠群,直至得到完整序列的过程称为重叠群装配。拼接 EST 数据以发现全长新基因也有类似的问题。从算法层次来看,序列的重叠群是一个 NP -完备性算法问题。

1.2.6　遗传密码的起源

遗传密码为什么是现在这样的?这一直是一个谜。对遗传密码的研究通常认为,密码子与氨基酸之间的关系是生物进化历史上一次偶然的事件而造成的,并被固定在现代生物的共同祖先里,一直延续至今。不同于这种"冻结"理论,有人曾分别提出过选择优化、化学和历史等三种学说来解释遗传密码。各种生物基因组测序工作的完成,为研究遗传密码的起源和检验上述理论的真伪提供了新的素材。

早在遗传密码破译以前,盖莫夫就曾对遗传密码的起源进行了假设。关于遗传密码的起源,主要有两种相互对立的假说:① 克里克提出的偶然冻结理论认为,三联体密码子与相应的氨基酸的密码关系完全是偶然的,而这种关系一旦建立就立即冻结,保持不变。由于这种假说难以用实验进行验证,至今尚无有力的证据。② 伍斯的立体化学理论认为,三联体密码子与相应的氨基酸之间的密码关系起源于它们之间特殊的立体化学相互的作用。近 30 年来对遗传密码起源的研究主要是从这个角度进行的。大量的研究结果表明,氨基酸与反密码子的直接作用以及疏水-亲水相互作用在遗传密码的起源中可能具有重要意义。

近 10 年来,分子生物学特别是核酸化学的一系列进展,证明了从核酸特别是从 RNA 途径研究遗传密码起源的准确性,为遗传密码起源问题的最终解决提供了可能性。塞克和艾尔

麦恩发现核酸酶这一事实表明，在生命起源的早期，在蛋白质酶的生命系统之前，存在一种基于 RNA 的自我复制系统，它使著名的“鸡-蛋”悖论倾向于先有 RNA。从原则上讲，集催化与信息功能于一身的 RNA 可能催化自身的复制。且核酸酶不会仅限于自身或自身的互补链作模板。某些 RNA 具有 RNA 加工酶的活性，如 RNaseP；某些 RNA 分子可能结合一个氨基酸作为原始的 tRNA；某些 tRNA 分子可能促进邻位的两个 tRNA 的结合，而在 RNA 模板上催化肽键的形成。随着多肽与蛋白质的形成，其中一些可能与核酸酶相互作用促进或调节其活性。因此，基于核酸酶的作用与功能，可以预见一个较现实的复制翻译过程。以上想法近年来已取得了某些实验证据。

原始 tRNA 比现代 tRNA 分子小得多，而且可能是由反密码环与氨基酸接受臂构成。虽然原始 tRNA 可能是随机形成的，但只要其特定的氨基酸处于正确的位置上，即能识别相应的氨基酸。这些特定的核苷酸就是反密码子。由于酶只能改变反应的速率，不能改变反应的平衡和性质，故在原始 tRNA 与相应的氨基酸的相互识别中，没有特定酶的催化，以上反应也能进行。原始地球条件下相对长的反应时间以及已经发现的一些原始的催化作用也有可能有利于以上反应的进行，而不是像有的科学家认为的，原始 tRNA 与相应氨基酸的特异性完全是由特定的氨酰 tRNA 合成酶决定的。

因此，遗传密码的起源既非偶然的冻结，也非简单地源于三联体密码子或反密码子与相应核苷酸的直接相互作用，而可能来源于氨基酸与相应原始 tRNA 之间的立体化学相互作用。在这种相互作用中，反密码子决定了作用的特异性，并进而决定了与它相应氨基酸的特定密码关系。人们正通过实验检验以上观点，即按照已知 tRNA 的核苷酸排列顺序及识别位点，人工合成各种识别位点的 tRNA 片段以检验它们与各种氨基酸的亲和性。尽管如此，关于遗传密码起源和进化的问题至今仍未得到令人满意的诠释，特别是在如何解释遗传密码的分配原则以及与生命体多样性的关系上，相关研究仍在进行中。

1.2.7 基于结构的药物设计

人类基因工程的目的之一是要了解人体内约 10 万种蛋白质的结构、功能、相互作用以及与人类各种疾病发生之间的关系，寻求各种治疗和预防方法，包括药物治疗。基于生物大分子结构及小分子结构的药物设计是生物信息学中的极为重要的研究领域。为了抑制某些酶或蛋白质的活性，在已知其蛋白质三级结构的基础上，可以利用分子对齐算法，在计算机上设计抑制剂分子，作为候选药物。基于结构的药物设计是计算机辅助药物设计的重要分支。这一领域研究的目的是发现新的基因药物，有着巨大的经济效益。

以基于靶标分子结构的药物设计为例，了解基于结构的药物设计的基本过程是：① 确定药物作用的靶标分子（如蛋白质、核酸等）；② 用生物技术手段对靶标分子加以分离纯化；③ 确定靶标分子的三维结构以及依据相关理论或经验提出的一系列假定的配体与靶标分子复合物的三维结构；④ 依据这些结构信息，利用相关的计算机程序和法则（如 DOCK）进行配体分子设计，模拟出最佳的配体结构模型；⑤ 合成出这些模拟出来的结构，进行活性测试，如果对测试结果感到满意，则进行后面的临床实验等研究，反之，重复以上过程，直至满意为止。

基于结构的药物设计有助于合成化学先导物及其优化。但是与其他领域相比，基于结构的药物发现在抗菌药物领域成功的例子相对还比较少，这主要与细胞透过困难有关。基于结

构的药物设计可提高化合物的活性和选择性，但在细胞透过性方面需要结合其他技术，如突变技术等。对细胞转运机制的深入了解有助于提高化合物的细胞穿透能力。

1.2.8　基因和蛋白质芯片

1. 基因芯片

基因芯片(gene chip)又称 DNA 芯片，是专门用于核酸检测的生物芯片，也是目前运用最广泛的微阵列芯片。它是指在固相载体上按照待定的排列方式固定大量序列已知的 DNA 片段，形成 DNA 微矩阵。所谓基因芯片就是按特定的排列方式固定有大量基因探针/基因片段的硅片、玻片、塑料片(图 1-3)。探针 DNA 是指被有序地点样固定在玻片或硅片上的 DNA 片段，可直接合成，或通过 PCR 技术扩增获得。这些大小和序列不同的片段分别经过纯化后，被高密度有序地点样固定在玻片或硅片上，从而制备成 DNA 微阵列，用于检测待测样品中是否有与之互补的序列。待测样品中的 mRNA 被提取后，通过反转录反应过程获得标记荧光的 cDNA，与包含上千个基因的 DNA 微阵列进行杂交反应，将玻片上未互补结合反应的片段洗去，再对玻片进行激光共聚焦扫描，测定微阵列上各点的荧光强度，推算出待测样品中各种基因的表达水平。若要比较不同的两个细胞系或不同组织来源的细胞中基因表达的差异，则从不同的两个细胞系或不同的组织来源中提取 mRNA。反转录反应过程中标记上不同颜色的荧光等量混合后，与包含上千个基因的 DNA 微阵列进行杂交反应，对玻片进行激光共聚焦扫描，比较两种荧光在各点阵上的强度，推算出各基因在不同细胞系中的相对表达水平。

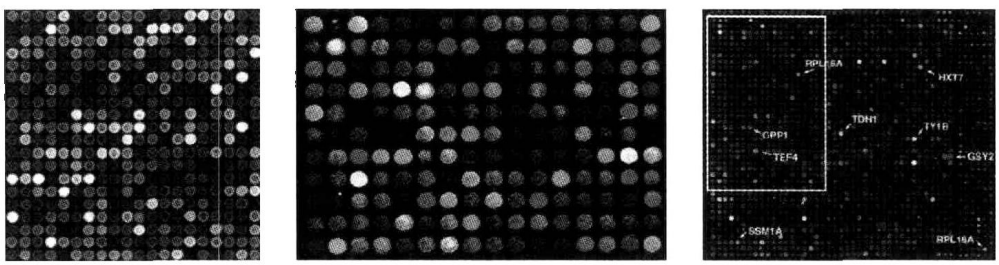

图 1-3　基因芯片示意图

基因芯片技术的核心原理与 Southern blot 相同。但 Southern blot 是将待测样品固定于尼龙膜上，与一个特定的经标记的 DNA 探针杂交，每次只能对一个靶序列进行检测；而基因芯片技术则是将大量的 DNA 探针固定于固相基质上，与待测的经标记的 DNA 样品杂交，只需一次实验，便能够将成千上万的基因的表现形式记录下来。基因芯片技术具有多靶并行处理能力、分析速度快、所需样品量少、污染少等优点，近年来对临床诊断、药物筛选、寻找新基因等研究领域带来巨大影响。

2. 蛋白质芯片

蛋白质芯片是一种新型的生物芯片，是由固定于不同种类支持介质上的蛋白微阵列组成的。阵列中固定分子的位置及组成是已知的，用未经标记或已标记(荧光物质、酶或化学发光物质等标记)的生物分子与芯片上的探针进行反应，让待测样品通过芯片表面与其接触，经过洗脱把非特异性结合的蛋白洗掉，然后通过特定的扫描装置对特异性地结合在上面的蛋白质