<<生物数据整合与挖掘>>

图书基本信息

书名：<<生物数据整合与挖掘>>

13位ISBN编号：9787309066142

10位ISBN编号：7309066146

出版时间：2009-5

出版时间：复旦大学出版社

作者：朱扬勇，熊S　著

页数：240

字数：282000

前言

自然科学研究宇宙和生命，所用的方法是证明和实验，证明依靠的是数学，实验依靠的是观测。
由于观测具有不确定性，因此依靠数学更能促进科学的发展。
历史上，一旦某个研究领域采用了数学就会快速发展（例如，物理学采用数学后产生了数学物理），可以把这种现象称为“数学化”，从这个角度可以将数学看成是自然科学的工具。
时至现今，几乎所有的学科都或多或少地依靠数学。
后来，计算机出现了，这是建立在数学基础上的机器，计算机突破了人在运用数学时的局限性（如问题规模较大时，只能抽样）。
于是，当一个领域采用了计算机也会快速发展（如数学本身采用计算机后产生的计算数学），这种现象称为“信息化”。

生命科学一直是以实验为主的，很难“数学化”，然而却能够“信息化”，信息化后形成了生物信息学。
生物信息学应用计算机对各种生物数据进行存储、管理、处理和分析，以期发现生物数据所反映的生物规律，促进生命科学的发展。

生物数据主要来自于生命科学领域的实验，实验产生了巨量的生物数据，其中尤其是基因组计划产生的数据最具代表性。
这些巨量的生物数据保存在世界各地的相关研究机构中，或隐含在浩瀚的科学文献里。
这种方式存放的生物数据也常常被称为生物数据库，但是它们和计算机领域所用的数据库可能是完全不同的。
这些数据有用文本文件方式存储的，也有用各种数据库管理系统存储的。
它们反映了生命科学研究的整体进展和成果，有重叠更相互补充，这需要将这些生物数据整合在一起。

<<生物数据整合与挖掘>>

内容概要

生物信息学应用计算机技术对各种生物数据进行管理和分析，以期发现生物数据所反映的生物规律，促进生命科学的发展。
一方面，生命科学实验产生的巨量的生物数据保存在世界各地的相关研究机构中，或隐含在浩瀚的科学文献里。
这些数据反映了生命科学研究的整体进展和成果，有重叠更相互补充，这就需要将这些生物数据整合在一起。
另一方面，生物信息学也希望采用数据挖掘技术对生物数据进行分析，以期发现生物规律，因此根据生命科学的需要和领域知识，设计出有效的生物数据挖掘算法和软件工具是一个重要的研究内容。
本书较为系统地介绍了生物数据整合与挖掘的技术框架，主要介绍了作者在这方面的研究成果，包括：生物数据抽取技术、生物数据整合技术、生物序列数据挖掘、基因表达谱芯片数据挖掘、转录因子及顺式调控元件挖掘、生物数据模型和数据库管理系统等内容，还介绍了一个生物数据整合系统、一个基因表达谱芯片数据库和数据挖掘系统、一个转录因子及顺式调控元件的挖掘分析平台等等的设计与实现。
本书的读者对象为从事生物信息学研究的科学工作者。
本书也可以作为生物信息学专业研究生的教学参考书和生物软件工程技术人员的参考书。

作者简介

朱扬勇，1963年生，浙江武义人。
1994年于复旦大学获计算机软件专业理学博士学位。
现为复旦大学计算机科学技术学院教授；上海市政府信息化专家；上海生物信息技术研究中心学术委员会委员；上海市计算机学会理事；上海市生物信息学会理事等。
长期从事数据库、数据挖掘、生物

<<生物数据整合与挖掘>>

书籍目录

<<生物数据整合与挖掘>>

章节摘录

第1章　背景知识　　诺贝尔奖获得者Dulbecc0于1986年在《Science》杂志上发表的一篇短文中率先提出了人类基因组计划。
该计划在探讨生命奥秘的过程中，使得自动化的DNA测序技术、生物数据挖掘分析技术、基因组数据库和分析软件、基因芯片技术的一些工具性技术获得了快速发展，并使生物信息学作为一个学科领域获得了公认。
本章介绍生物信息学、数据整合与数据挖掘方面的背景知识和基本概念。
1.1　生物信息学　　生命科学实验产生了大量生物数据，如何在数学、计算机科学等的支持下充分利用这些生物数据更有效地开展生命的探讨是一个很有意义的问题。
于是，生物数据处理技术获得了发展，并最终产生了生物信息学。
1.1.1　基本概念　　生物信息学（Bioinformatics）是指生命科学与数学科学、计算机科学和信息科学等交汇融合所形成的一门交叉学科。
它应用先进的数据管理技术、数学分析模型和计算机软件对各种生物数据进行提取、存储、处理和分析，旨在掌握复杂生命现象的形成模式与演化规律。
该定义是Rashidi等人于2000年给出的。
由于生命科学研究者各自从事的具体领域不同，对其存在不同的理解，因此至今仍没有一个关于生物信息学的统一定义。
但其基本的研究内容和研究方法还是比较统一的，就是通过研究生物数据来促进生命科学的研究。
随着生命科学研究的深入，生物信息学也受到广泛关注。
事实上，生物信息学起源要早很多。
1953年4月25日，Waston和Crick提出DNA（Deoxyribo Nucleic Acid）双螺旋结构和自我复制机制，揭开了分子生物学研究的新篇章。
1956年，在美国田纳西州盖特林堡召开首次“生物学中的信息理论研讨会”，萌生了生物信息学概念。
20世纪60年代，研究者开始搜集生物信息，并应用计算方法对其进行分析，发现其中反映生命现象的重要规律。
随后，生物学的研究手段发生了革命性的变化，由单纯的观察和实验研究转向与生物数据分析相结合。
70年代到80年代初，数学统计方法和计算机技术得到了较快发展，研究者开始应用计算机技术解决生物学问题，生物信息学初步形成。
1986年，美国科学家首次提出“人类基因组计划”（Human Genome Project，HGP），促进了生物信息学的迅速发展。
1987年，Hwa A.Lim博士　　首次将这一学科命名为“Bioinformatics”（生物信息学）。
正如Dulbecc0 1986年所说：“人类的DNA序列是人类的真谛，这个世界上发生的一切事情，都与这一序列息息相关。
”但这些由数以亿计ACGT符号组成的DNA序列中包含着什么信息？
基因组中的这些信息怎样控制有机体的发育？
基因组本身又是怎样进化的？
要完全破译这一序列以及相关的内容，人类还有相当长的路要走。
生物信息学成为可能揭开谜底的重要方法之一。
1.1.2　研究内容　　生物信息学的目标是指导生命科学研究，以揭示生物数据中蕴含的生物学知识和规律，读懂基因组的遗传信息。
其研究内容主要包括以下两大方面。
1.生物数据的存储、管理和整合　　生物数据主要有生物序列数据（如DNA序列、蛋白质序列等）、生物分子结构数据、芯片及基因表达数据、生物网络数据（如蛋白质相互作用网络、调控网络、代谢网络等）、生物文献数据等。

目前在国际上总共约有1 000多个生物数据库，存放数百TB（tera byte）的生物数据。
由于大多数生物数据的含义目前还不为人们所知，因此大量的生物学研究将基于生物数据进行。
生物学研究手段由单纯的观察和实验转向现代信息学方法，即将生物的实验变成了数据的计算。
　　生物数据是一种非结构化数据，数据量巨大、种类繁多、数据操作类型复杂等是其主要的特征。
其表达和存储方式是生物数据访问和处理的关键。
目前，生物数据的存储方式有两种：一种是采用文本文件方式存储；另一种是采用关系数据库、XML（eXtensible Markup Language）数据库或者面向对象数据库等存储方式，但是由于没有合适的数据模型或数据类型，生物数据在这种存储方式中也只是用数据库管理系统（Database Management System，DBMS）中提供的文本字段来存储。
就是说，两者本质上是一样的，都是文本方式。
文本方式对复杂的生物数据操作（如：生物序列相似性查询、MOTIF查询等）而言，处理效率是难以令人满意的，也即目前的数据库技术（包括XML数据库技术）都不适合生物数据的存储、管理和处理，这直接影响了生物信息学软件的有效性和实用性，进而影响了生命科学和生物技术的发展。
另外，文本方式的存储在生物数据的处理能力和处理性能上也都不能满足要求。
因此，如何有效地管理和处理生物数据是一个亟待解决的问题。
针对生物数据的特点，建立生物数据库管理系统是一个关系生命科学与技术发展的重要课题。
　　由于生物数据产生于世界各地的研究机构，存储在各种生物数据库中，因此为完成一项研究工作，需要整合这些分散在各研究机构中的生物数据。
但因为生物数据库数量众多且规模庞大，所以生物数据整合是一项艰巨的计算机工程任务。
　　2.生物数据挖掘和分析　　生物信息学领域的核心内容是研究如何通过对生物数据的分析，以期发现生物数据中的规律（如DNA序列、结构及其与生物功能之间的关系等），并对分析结果进行解释和可视化，其研究范围涉及基因组学、蛋白质组学、系统生物学、比较基因组学等，挖掘和分析的内容包括生物序列数据的分析和挖掘、蛋白质结构数据的分析和挖掘、生物网络系统的分析和挖掘、芯片和基因表达数据分析等内容。
　　（1）生物序列数据的分析和挖掘　　①序列比对：序列相似性研究是生物序列数据分析和挖掘研究的核心内容，其中一个主要的应用问题是给定一条生物序列，在序列数据库中查询与其相似程度大于一定阈值的序列（比较两个或两个以上的序列的相似性），即生物序列相似性查询。
序列比对是最基本、最重要的方法之一，它根据给定的相似矩阵（PAM250，BLOSUM62等），同时考虑可能的插入、删除和突变，找出序列间的最优联配。
序列比对主要有全局比对和局部比对两种策略：全局比对是对序列的全长进行比对，适用于全局水平上相似性程度较高的序列；典型的算法有Needle—man-Wunsch算法等；局部比对是寻找序列间相似性最大的子序列，典型的算法有基于动态规划思想的Smith-Waterman算法以及启发式的两序列比对数据库相似性搜索算法FASTA和BLAST（Basic Local Alignment Search T001）等。
多序列比对是将一组序列同时进行比对，发现序列间的相似程度，大多采用启发式算法，具有代表性的主要是渐进比对方法和迭代比对方法。
　　②功能元件分析：基因识别是识别DNA序列上的具有生物学特征的片段，识别对象包括蛋白质编码（即基因的范围和在序列中的位置），也包括其他具有一定生物学功能的功能元件，如转录因子、顺式调控元件等。
功能元件能够表征序列的功能特征。
序列上的功能元件主要包括编码序列元功能片段和非编码序列元功能片段等。
其中，编码序列可被转录并执行一定的生物学功能；调控序列控制编码序列的动态行为，如转录调控序列控制编码序列的表达速率等。
目前，“DNA元件百科全书”（Encyclopedia of DNA Elements，ENCODE）计划已开展人类基因组中功能元件的分析工作，但该计划正处于初期，积累的数据仍然较少。
　　（2）蛋白质结构数据的分析和挖掘　　人类基因工程的目的之一是要了解人体内蛋白质的结构、功能、相互作用以及与各种人类疾病之间的关系。
虽然蛋白质由氨基酸的线性序列组成，但是只有折叠成特定的空间构象才能具有相应的生物学功能。

由于蛋白质的三维结构比其一级结构在进化中更稳定，同时也包含了较氨基酸序列（一级结构）更多的信息，因此，蛋白质结构分析和预测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。
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，因此可通过对已知结构的蛋白质结构的分析来预测未知蛋白质的结构。
在蛋白质结构数据的分析和挖掘中，同源建模（homology modeling）方法是具有代表性的方法。
另一个目的是从蛋白质的氨基酸序列预测蛋白质结构，即从头预测（abinitio）方法，根据物理、化学原理通过理论计算（如分子力学、分子动力学）进行蛋白质的结构预测。
该方法假设折叠后的蛋白质取能量最低的构象。
但是从头预测方法在实际中常常不合适。
（3）生物网络系统的分析和挖掘　　分子生物学研究揭示，复杂生命现象是大量基因活动且相互作用的结果。
“DNA元件百科全书”计划的研究表明，人类基因组蓝图是一个复杂的网络系统。
认识和解读人类基因组蓝图是后基因组时代生命科学领域面临的最具挑战性的热点问题之一。
从全局和系统水平研究和分析生物学系统，阐述人类基因组中所有基因间的关系是系统认识人类基因组蓝图的重要步骤。
系统生物学是研究生物系统中所有组成成分（基因、mRNA、蛋白质等）的构成，以及在特定条件下这些组分间的相互关系的学科，将在基因组序列的基础上完成由生物体内各种组成成分鉴别及其相互作用的研究到途径、网络和模块的构建，这是现代生物学的研究前沿，已成为21世纪生物学的核心驱动力。
构建生物系统组成成分间的生物网络是系统生物学的重要研究内容，这对揭示基因功能、解析序列间的相互作用、认识生命活动的规律有重要意义。
通过实验识别生物网络是一种高耗费的方法。
因此，有必要研究新的生物信息学方法，以对生物网络进行有效识别，从而为实验生物学研究提供重要的指导信息。

版权说明