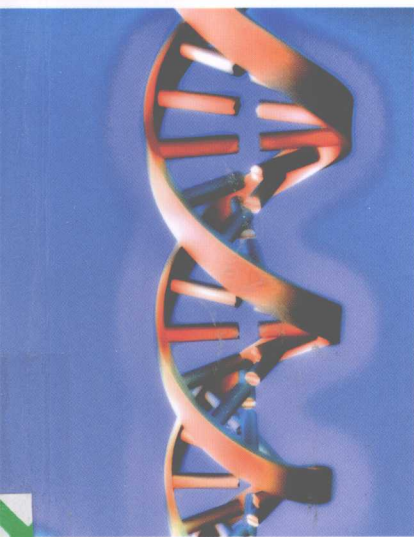


国家自然科学基金和“863计划”资助项目

生物数据整合与挖掘

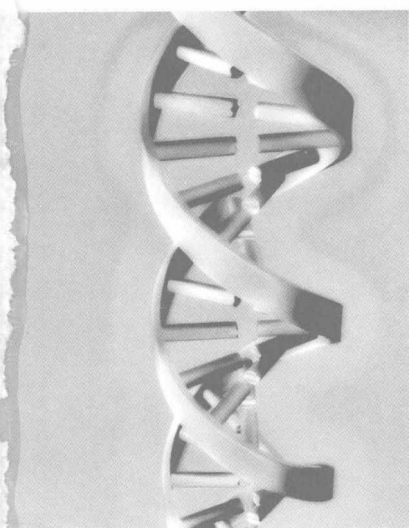
朱扬勇 熊赞 著



国家自然科学基金和“863计划”资助项目

生物数据整合与挖掘

朱扬勇 熊赞 著



复旦大学出版社

图书在版编目(CIP)数据

生物数据整合与挖掘/朱扬勇,熊贇著. —上海:复旦大学出版社,2009.5
ISBN 978-7-309-06614-2

I. 生… II. ①朱…②熊… III. 生物信息论-数据处理 IV. Q811.4-39

中国版本图书馆 CIP 数据核字(2009)第 062334 号

生物数据整合与挖掘

朱扬勇 熊 贇 著

出版发行 **复旦大学出版社** 上海市国权路 579 号 邮编 200433
86-21-65642857(门市零售)
86-21-65100562(团体订购) 86-21-65109143(外埠邮购)
fupnet@fudanpress.com <http://www.fudanpress.com>

责任编辑 范仁梅

出品人 贺圣遂

印 刷 上海崇明南海印刷厂
开 本 787×960 1/16
印 张 15.75
字 数 282 千
版 次 2009 年 5 月第一版第一次印刷

书 号 ISBN 978-7-309-06614-2/Q·72
定 价 39.00 元

如有印装质量问题,请向复旦大学出版社发行部调换。

版权所有 侵权必究

内 容 提 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，因此根据生物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。

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。时至今日,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有代表性。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。由于数据源

在内容、形式和管理方式上的多样性,因此生物数据整合是一项困难的工作,且工作量巨大。

生物信息学希望通过对巨量生物数据的分析来发现生物规律。数据挖掘是从大量数据中寻找其规律的技术,是目前最强有力的计算机数据分析技术之一,因此成为生物信息学采用的主要数据分析技术。数据挖掘工作输入的是大量数据,产出的是用户认为有用的数据规律。对于生物数据挖掘而言,这包含两个方面:一方面设计数据挖掘算法和相应的软件工具,这是计算机领域的工作;另一方面确定什么样的数据规律是有用的,这是生命科学领域的工作。因此,如何根据生命科学领域的需要和知识,设计出有效的生物数据挖掘算法和软件工具,是生物信息学研究的重要内容。

本书较为系统地介绍了生物数据整合与挖掘的技术框架,主要介绍了我们在这方面的研究成果,希望能够给生物信息学领域的研究者一些启示。对于有生命科学背景的研究者来说,生物数据整合与挖掘是一项先进的计算机工程,类似于建造国家体育场(鸟巢),需要大量的人力、物力投入,也需要技术攻关;对于有计算机科学背景的研究者来说,生物数据整合与挖掘是生物信息学研究的必需,生物数据表象简单但其内涵却是未知的,生物学家同样希望知道这些数据的内容,并希望通过数据挖掘来解决问题。因此,生物数据整合与挖掘需要生命科学和计算机科学两方面的科学家和工程师的良好合作。

我们的工作还存在不成熟的地方。书中的不妥之处全属作者水平有限所致,如蒙读者指正,不胜感谢。

朱扬勇 熊 赟

2008年12月27日于复旦大学

目 录

第 1 章 背景知识	1
1.1 生物信息学	1
1.1.1 基本概念	1
1.1.2 研究内容	2
1.1.3 研究方法	5
1.1.4 研究机构	6
1.2 数据整合	6
1.2.1 数据资源	7
1.2.2 数据整合的动因	8
1.2.3 数据整合的概念	9
1.2.4 数据整合的内容	12
1.3 数据挖掘	13
1.3.1 数据挖掘的定义	13
1.3.2 数据挖掘的任务	14
1.3.3 数据挖掘的类型	21
1.3.4 相关技术的差异	23
第 2 章 数据整合与数据挖掘方法	26
2.1 数据整合的方法	26
2.1.1 数据整合的方式	26
2.1.2 数据整合的步骤	28
2.2 数据挖掘的方法	33
2.2.1 数据挖掘过程示例	33
2.2.2 数据挖掘过程模型	34
2.2.3 数据挖掘应用方式	36

2.3 数据清洁技术	38
2.3.1 数据的质量问题	38
2.3.2 数据清洁的主要工作	39
2.4 数据仓库技术	40
2.4.1 数据仓库的概念	41
2.4.2 数据组织	41
2.4.3 主题设计	43
2.4.4 数据加载	44
2.4.5 数据规约	46
2.5 小结	46
第3章 生物数据源	47
3.1 生物数据	47
3.1.1 生物序列数据	47
3.1.2 生物分子结构数据	48
3.1.3 芯片及基因表达数据	49
3.1.4 生物网络数据	49
3.2 生物数据组织	50
3.2.1 生物数据的数据库组织形式	51
3.2.2 生物数据的互联网组织形式	53
3.3 生物数据库	54
3.3.1 生物序列数据库	55
3.3.2 基因组数据库	57
3.3.3 结构数据库	58
3.3.4 芯片和基因表达数据库	58
3.3.5 生物文献数据库	59
3.4 生物数据源的特征	60
3.5 小结	62
第4章 复杂生物数据源的数据抽取	63
4.1 生物数据抽取	63
4.1.1 生物数据抽取面临的问题	63
4.1.2 包装器的要素	65
4.1.3 抽取算法	66

4.1.4 元数据生成与包装器生成工具	66
4.2 包装器的设计	67
4.2.1 基于实例切分的抽取算法	67
4.2.2 基于定位器多结点共享的数据抽取模型	68
4.2.3 数据抽取模型描述	68
4.2.4 元数据的生成和维护	69
4.2.5 数据抽取模型表达能力	69
4.3 包装器解决方案	69
4.3.1 面向无噪声复杂数据源的解决方案	69
4.3.2 面向含噪声复杂数据源的解决方案	70
4.3.3 ReDE 和 L-树包装器生成工具的架构	71
4.3.4 ReDE 和 L-树包装器生成工具的实现技术	73
4.4 L-树匹配:面向复杂数据源的数据抽取算法	74
4.4.1 L-树上的数据映射机制	74
4.4.2 L-树匹配算法的相关概念	76
4.4.3 L-树匹配算法	78
4.4.4 L-树匹配算法举例	79
4.5 基于 L-树的包装器生成工具	82
4.5.1 将 ERE 扩充成数据抽取脚本语言	82
4.5.2 可视化编辑调试环境	83
4.5.3 ERE 的可视化构建	83
4.5.4 ERE 的逻辑检查	86
4.5.5 抽取结果的可视化评价	86
4.5.6 以 XML 格式输出抽取结果	87
4.6 小结	88
第 5 章 生物数据整合案例	89
5.1 生物数据整合系统的设计	89
5.1.1 生物数据整合的关键问题分析	89
5.1.2 生物数据整合目标的确立	90
5.1.3 生物数据整合方式和技术的的设计	91
5.2 基于 GO 的数据整合	92
5.2.1 GO 简介	93
5.2.2 DB2GO 表	95

5.2.3 语义相似数据库表	95
5.2.4 以 GO 统一数据的逻辑和语义	96
5.3 数据抽取和增量更新	98
5.3.1 数据抽取	98
5.3.2 数据的增量更新	100
5.4 基于 GO 的查询技术	101
5.4.1 异构生物数据库的语义查询	101
5.4.2 BioDW 中语义查询的体系结构	101
5.4.3 GO 语义相似性度量方法	102
5.4.4 语义相似性查询	104
5.5 BioDW 系统	106
5.5.1 BioDW 的系统结构	106
5.5.2 BioDW 的系统的规模	107
5.5.3 BioDW 的数据查询	107
5.6 小结	115
第 6 章 生物序列数据挖掘进展	116
6.1 生物序列数据挖掘的基本概念和内容	116
6.1.1 生物序列相似性	117
6.1.2 生物序列模式挖掘	118
6.1.3 生物序列聚类分析	119
6.1.4 生物序列分类分析	120
6.1.5 生物序列关联分析	120
6.1.6 生物序列异常分析	120
6.2 生物序列数据挖掘的研究阶段	121
6.2.1 基于统计技术的数据挖掘方法的应用阶段	121
6.2.2 一般化数据挖掘方法的应用阶段	122
6.2.3 专门数据挖掘技术的设计阶段	122
6.3 生物序列数据挖掘研究与应用现状	122
6.3.1 生物序列模式挖掘方面	123
6.3.2 生物序列聚类分析方面	127
6.3.3 生物序列分类分析方面	130
6.3.4 生物序列关联分析方面	132
6.3.5 生物序列异常分析方面	132

6.4 生物序列数据挖掘研究趋势	133
6.5 小结	134
第7章 生物序列数据挖掘技术	135
7.1 序列数据源	135
7.2 生物序列模式挖掘	137
7.2.1 生物序列模式挖掘问题	137
7.2.2 基于多支持度的生物序列模式挖掘框架	139
7.2.3 基于多支持度的生物序列模式挖掘算法	141
7.3 生物序列聚类分析	147
7.3.1 生物序列聚类问题分析	147
7.3.2 蛋白质序列聚类	149
7.3.3 基因序列聚类	154
7.4 生物序列分类分析	167
7.4.1 生物序列分类问题分析	167
7.4.2 转录因子分类	167
7.4.3 基于支持向量机的转录因子分类算法	169
7.5 小结	171
第8章 基因芯片数据挖掘	172
8.1 基因表达谱芯片数据挖掘	172
8.1.1 基因表达谱数据分析	172
8.1.2 基因表达相似性分析	173
8.1.3 基因表达共发生分析	173
8.1.4 基因表达路径分析	173
8.1.5 特殊表达基因分析	174
8.2 基因表达谱数据库建设	175
8.2.1 基因表达谱芯片数据的标准	175
8.2.2 基因表达谱数据库建设的难点	176
8.2.3 数据库结构设计	177
8.2.4 数据加载与数据管理	179
8.2.5 自动导入数据	180
8.3 基因表达谱数据挖掘系统	182
8.3.1 数据挖掘框架	182

8.3.2	BDMAPA 架构扩展	183
8.3.3	基因表达谱芯片数据挖掘系统	184
8.4	小结	186
第 9 章	转录因子、顺式调控元件挖掘系统	187
9.1	转录因子、顺式调控元件挖掘原理	187
9.1.1	转录因子、顺式调控元件挖掘原理	188
9.1.2	顺式调控元件文本挖掘原理	188
9.2	转录因子、顺式调控元件挖掘系统设计	194
9.2.1	数据挖掘软件	194
9.2.2	数据分析服务	195
9.2.3	综合的转录因子、顺式调控元件数据库	200
9.3	小结	201
第 10 章	生物序列数据库管理系统	202
10.1	生物数据处理面临的问题	202
10.1.1	生物数据存储方式	202
10.1.2	生物序列数据库的查询需求	204
10.2	生物序列数据模型 BioSeg	205
10.2.1	数据结构	206
10.2.2	代数操作	210
10.2.3	Open BUILT-IN 函数	224
10.2.4	等价规则	225
10.2.5	BioSeg 模型的特点	227
10.3	生物序列数据库管理系统的设计	229
10.3.1	代数查询实例	229
10.3.2	查询语言	232
10.3.3	体系结构	234
10.4	小结	235
参考文献		236
致谢		239

第1章

背景知识

诺贝尔奖获得者 Dulbecco 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 Dulbecco 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库模型或数据类型,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 Needleman-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 Tool)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(ab initio)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。认识和解码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。

(4) 芯片和基因表达数据分析

芯片技术已经为生物信息学研究积累了相当数量并且蕴含丰富生物学知识的表达数据。对芯片和基因表达数据的主要处理技术之一是聚类分析,其目的是得到在一定条件或环境影响下具有共调控(共表达)功能的基因簇,从而揭示基因序列间的调控机制。表达数据通常是在多个条件(如组织、细胞周期的不同阶段、药物作用时间等)下度量的,它具有高维性。低维空间中经常采用的 L_k -范数在很多情况下不适合用作高维空间的距离函数,基于全空间距离函数的聚类方法也会失效。由于一个感兴趣的生物过程通常只包含基因中的一部分,且过程也可能只发生在少数条件下,即高维空间中的对象在某个子空间中相似,其中,那些在部分条件下具有保守变化趋势但表达水平值强度存在差异(即在一个维集上表现出一致的模式,但在各维上的值并不一定接近的情况)的基因,对于揭示基因序列间的共调控关系有重要的生物意义,因此,限制在只对原始空间的子空间搜索而不是在一个新的维度(例如由原始的维度线性组合成新的维度)上搜索的方法具有重要意义。定义合适的相似性度量是影响具有高维特性的芯片和基因表达数据聚类质量的关键。

1.1.3 研究方法

生物信息学的基本工具是数学科学与计算机科学技术。与生物信息学关系密切的数学领域有统计学、运筹学、计算数学、组合数学等;与生物信息学密切相关的计算机科学技术主要有数据库技术、数据挖掘技术、人工智能技术、网络技术等。其中,由于复杂的生物数据特点、多样的生物学分析需求以及激增的生物数据规模,数据挖掘在生物信息学中扮演着重要的角色。

1. 数学方法

统计学是收集、分析和解释数据并从数据中得出结论的科学,已成为生物信息学的数学基础之一。例如,支持向量机(Support Vector Machine, SVM)在生物信息学中得到了广泛应用,如预测蛋白质相互作用位点、蛋白质结构和功能等,其研究背景是 Vapnik 的统计学习理论,通过最大化两个数据集的最大间隔实现分类。对于非线性问题,则采用核函数将数据集映射至高维空间而又无需显式描述数据集在高维空间的性质。

概率论与随机过程理论,如隐 Markov 链模型(Hidden Markov Models, HMM),在生物序列分析方面中有重要应用;其他如用于序列比对的运筹学,在蛋白质空间结构预测和分子拼接研究中采用的最优化理论,研究生物网络的拓扑学,研究遗传密码和 DNA 序列的对称性方面的群论,等等。然而,并非所有的数