生命科学前沿及应用生物技术

蛋白质结构预测
——支持向量机的应用

孙向东　刘拥军　黄保续　谢仲伦　编著

科学出版社

生命科学前沿及应用生物技术大系·典藏版

蛋白质结构预测
——支持向量机的应用

孙向东　刘拥军　黄保续　谢仲伦　编著

科学出版社
北京

内 容 简 介

应用生物技术大系和现代生命科学前沿系列图书分别被列为“十一五”和“十二五”国家重点图书出版规划项目。本丛书针对生命科学领域前沿重点发展方向以及应用生物技术领域的新成果、新思路、新方法和新技术，全面展示了其最新的发展动态，涵盖了基础理论和主要技术方法，呈现了新的概念与理论、技术，在更深层次上阐明了生命的本质规律，给人们提供了新的认识生命本质的手段，也为生物技术服务于人类开辟了新的途径。涉及领域包括生物医药、干细胞技术、工业微生物学、蛋白质及蛋白质组、系统生物学、合成生物学、生物材料、农业生物技术、环境生物技术、海洋生物技术、生物资源与安全等。

图书在版编目（CIP）数据

生命科学前沿及应用生物技术大系：典藏版/舒红兵等编著. —北京：科学出版社, 2016

ISBN 978-7-03-047487-2

Ⅰ.①现… Ⅱ.①舒… Ⅲ. ①生命科学–研究②生物工程—研究 Ⅳ.①Q1-0②Q819

中国版本图书馆 CIP 数据核字(2016)第 043876 号

责任编辑：王 静 李 悦
责任印制：张 伟 / 封面设计：刘新新

科学出版社出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

北京虎彩印刷科技有限公司 印刷
科学出版社发行 各地新华书店经销
*
2016 年 7 月第 一 版 开本：787×1092 1/16
2016 年 7 月第一次印刷 印张：2108
字数：49 985 000
定价：8900.00 元

(如有印装质量问题，我社负责调换)

前　言

蛋白质由氨基酸残基线性序列构成, 折叠成特定的空间构象后, 蛋白质就具有相应生物学活性和功能. 了解氨基酸残基序列与其空间结构的关系, 是全面认识蛋白质结构和其生物学功能的关系的重要前提. 近些年来, 蛋白质序列数据库的数据积累速度非常快, 与之相比, 蛋白质结构数据库的数据积累速度远不及序列数据库的数据积累速度. 尽管蛋白质结构测定技术有了较为显著的进展, 但是通过实验方法确定蛋白质结构的过程仍然非常复杂, 实验周期很长.

另外, 随着 DNA 测序技术的发展, 人类基因组及很多模式生物基因组已经或将要完全测序, DNA 序列数量将会急增. 由于 DNA 序列分析技术和基因识别方法的进步, 人们可以从 DNA 序列直接推导出大量的蛋白质序列, 这将导致蛋白质序列数据数量急剧增加. 了解了这些序列的结构, 可以使它们直接为人类服务.

氨基酸残基序列的结构分析是对生物学家的极大挑战. 20 世纪 60 年代后期, Anfinsen 首先发现去折叠蛋白或者说变性蛋白质在允许重新折叠的实验条件下可以重新折叠到原来的结构, 这种天然结构对于蛋白质行使生物功能具有重要作用, 蛋白质只有在折叠成其天然结构的时候才能具有完全的生物活性. 因此 Anfinsen 提出了蛋白质折叠信息隐含在蛋白质的一级结构中的观点. 以这种观点为基础, 通过对蛋白质一级结构的研究, 发现其折叠密码后, 仅通过一级结构信息就能预测蛋白质空间结构.

蛋白质结构预测主要有两大类方法. 一类是蛋白质分子特性理论分析方法或从头算方法, 通过理论计算 (如分子力学、分子动力学计算) 进行结构预测. 该类方法假设折叠后的蛋白质取能量最低的构象. 从原则上来说, 人们可以根据物理、化学原理, 通过计算来进行结构预测. 另一类蛋白质结构预测的方法是统计学方法. 该类方法对已知结构的蛋白质进行统计分析、建立序列到结构的映射模型、进而根据映射模型对未知结构的蛋白质直接从氨基酸序列预测结构. 这是进行蛋白质结构预测较为成功的一类方法. 这类方法包括经验性方法、结构规律提取方法、同源模型化方法等. 统计学方法本身就是不确定性方法, 目前虽然还不能完全替代第一类方法而成为预测蛋白质结构的主要方法, 但是发展前景很广阔. 其中以统计学习理论为基础的支持向量机预测蛋白质结构的方法发展非常迅速.

统计学习理论是在 20 世纪 90 年代逐渐成熟的机器学习理论, 以这种理论为基础的支持向量机与以往的学习机器相比具有支持小样本、不会陷入局部势井、具有很好的鲁棒性以及运算成本低等优势. 实现这种理论的支持向量机算法已经成

为机器学习和知识挖掘的重要工具. 从 2001 年支持向量机首次被运用进行蛋白质二级结构的预测以来, 这种算法已经被用于对于蛋白质的结构类型、亚细胞结构和膜蛋白的结构等领域的预测中.

本书一共包含 8 章, 阐述三部分内容, 包括生物信息学基本知识、蛋白质结构预测基本知识、蛋白质二级结构和结构域预测技术、支持向量机算法以及相应软件的使用方法和实验步骤, 由浅及深, 步步深入, 系统阐述了运用支持向量机预测蛋白质二级结构和结构域的基本原理和过程. 有兴趣的读者可以按照本文描述的实验步骤和相应参数完全重复整个实验过程. 第一部分包括第 1 章和第 2 章, 主要对蛋白质二级结构和结构域预测以及知识背景进行简要介绍. 第二部分包含第 3 章到第 6 章, 系统阐述了统计学习理论、以这种理论为基础的学习算法 —— 支持向量机、支持向量机构造方法以及实现支持向量机算法的程序 Libsvm. 最后一部分包括第 7 章和第 8 章, 这一部分详细论述了运用支持向量机方法进行蛋白质二级结构预测和结构类型预测的实验过程和最终结果.

由于作者水平有限、成文仓促, 文中难免出现这样那样的疏漏和错误. 书中欠妥之处敬请读者批评指正!

孙向东　刘拥军　黄保续　谢仲伦

2008 年于北京

目　录

第 1 章　蛋白质结构预测概述

1.1　蛋白质预测基本方法简介

生物信息学是近年来最有活力的生物学研究领域之一, 人们从生物信息的研究中获得了对生命本质更丰富的知识和更深刻的理解. 核酸序列中蕴含着生命的基本信息, 这些信息是自然界留给人类的、解读生命的 "天书". 理解这本天书是最终了解自然、了解生命、了解人类自身的重要途径, 是人类从必然王国到自由王国飞跃的基本前提之一.

由基因决定的蛋白质执行着生物体内各种重要的功能, 如生物化学反应的催化、营养物质的输运、生长和分化控制、生物信号的识别和传递等. 基因确定了组成蛋白质的氨基酸序列. 虽然蛋白质由氨基酸的线性序列组成, 但是它们只有折叠成特定的空间构象才能具有相应的活性和相应的生物学功能. 了解蛋白质的空间结构不仅有利于认识氨基酸残基序列与空间结构的关系, 也有利于认识蛋白质的结构与其生物学功能的关系.

根据近些年来的经验, 蛋白质序列数据库数据积累速度非常快, 而且还有加快的趋势. 尽管蛋白质结构测定技术有了较为显著的进展, 但是通过实验方法确定蛋白质结构的过程仍然非常复杂, 实验周期很长. 另外, 随着 DNA 测序技术的发展, 人类基因组及很多的模式生物基因组已经或将要被完全测序, DNA 序列数量将会剧增, 由于 DNA 序列分析技术和基因识别方法的进步, 人们可以从 DNA 序列直接推导出大量的蛋白质序列. 这意味着已知序列的蛋白质数量和已测定结构的蛋白质数量 (如蛋白质结构数据库 PDB 中的数据) 的差距将会越来越大. 面对这种蛋白质结构信息与 DNA 序列信息发展速度的不平衡, 人们希望找到一些预测方法, 通过这些方法加快蛋白质结构产生速度, 缩小二者之间的差距.

为了缩小这种差距, 要么改进现有的蛋白质测序技术和结构预测方法, 要么发展新的理论分析方法, 这是对生物学家的极大挑战. 20 世纪 60 年代后期, Anfinsen 首先发现去折叠蛋白质或者说变性蛋白质在允许重新折叠的实验条件下可以重新折叠到原来的结构, 这种天然结构对于蛋白质行使生物功能具有重要作用, 蛋白质只有在折叠成其天然结构的时候才能具有完全的生物活性. 因此 Anfinsen 提出了蛋白质折叠的信息隐含在蛋白质的一级结构中的观点. 基于这种观点, 人们相信通过对蛋白质一级结构的研究, 发现其折叠密码后能够仅通过一级结构信息就能预测蛋白质空间结构.

到目前为止, 科学家对于蛋白质结构预测进行了大量的研究, 已经尝试了一些预测蛋白质结构的方法. 蛋白质结构预测主要有两大类方法. 一类是蛋白质分子特性的理论分析方法或从头算方法, 通过理论计算 (如分子力学、分子动力学计算) 进行结构预测. 该类方法假设折叠后的蛋白质取能量最低构象. 从原则上来说, 人们可以根据物理、化学原理, 通过计算来进行结构预测. 但是这种方法可操作性很差, 主要有几个原因:

(1) 自然的蛋白质结构和未折叠的蛋白质结构之间的能量差非常小 (1kcal/mol 数量级);

(2) 蛋白质可能的构象空间庞大, 针对蛋白质折叠的计算量非常大;

(3) 计算模型中蛋白质及溶剂系统的力场参数的不准确性、无法从数学上解决局部势井问题, 因此无法证明某蛋白质分子的构象是全局自由能最小的构象.

另一类蛋白质结构预测的方法是统计学方法. 该类方法对已知结构的蛋白质进行统计分析、建立序列到结构的映射模型、进而根据映射模型对未知结构蛋白质直接从氨基酸序列预测结构. 这是进行蛋白质结构预测较为成功的一类方法. 这一类方法包括经验性方法、结构规律提取方法、同源模型化方法等. 但是这类方法不可能是完全独立的, 它们不能脱离对蛋白质分子的物理、化学和生物性质的研究. 统计学方法本身就是不确定性方法, 目前还不可能替代第一类方法而成为预测蛋白质结构的最终方法, 而只能是一种辅助方法.

1.2　蛋白质二级结构和结构域预测方法简介

蛋白质结构预测已经有了几十年的历史. 通过对已知空间结构蛋白质分子的研究和分析, 人们发现, 尽管一条多肽链采取构象的数目是相当大的, 但在蛋白质分子中由三级结构组装而形成的一定空间结构的方式却是有限的. 蛋白质二级结构是这种组装的基本单位, 蛋白质二级结构预测和由二级结构构成的结构域预测就成了解决由蛋白质的一级结构序列预测其空间结构这一问题的关键步骤.

蛋白质二级结构的预测开始于 20 世纪 60 年代中期, 到目前为止人们已经提出几十种预测蛋白质二级结构的方法. 这些方法大体分为三代, 第一代是基于单个氨基酸残基统计分析, 从有限的数据集中提取各种残基形成特定二级结构的倾向, 以此作为二级结构预测的依据, 这种方法的代表是 Chou-Fasman 方法. 第二代预测方法是基于氨基酸片段的统计分析, 使用大量的数据作为统计基础, 统计的对象不再是单个氨基酸残基, 而是氨基酸片段, 片段的长度通常为 11~21 个氨基酸. 片段体现了中心残基所处的环境. 在预测中心残基的二级结构时, 以残基在特定环境中形成特定二级结构的倾向作为预测依据. 这种方法的代表是 GOR 方法. 二级结构预测的第三代方法运用蛋白质序列的长程信息和蛋白质序列的进化信息, 使二级

结构预测的准确程度有了比较大的提高, 特别是对 β 折叠的预测准确率有较大的提高, 预测结果与实验观察趋于一致. 这种方法的代表是人工神经网络方法.

Chou-Fasman 方法是一种基于单个氨基酸残基统计的经验参数方法, 由 Chou 和 Fasman 在 20 世纪 70 年代提出. 通过统计分析, 获得每个残基出现于特定二级结构构象的倾向性因子, 进而利用这些倾向性因子预测蛋白质的二级结构. Chou-Fasman 方法构象参数的物理意义明确, 方法中二级结构的成核、延伸和中止规则可能真实地反映了真实蛋白质中二级结构形成的过程, 并且可以较简单地用手工完成一个蛋白质分子的二级结构预测, 预测准确率约为 50%.

GOR 方法是一种基于信息论和贝叶斯统计学的方法, 方法的名称以三个发明人姓名的第一个字母组合而成 (Garnier, Osguthorpe, Robson). GOR 方法也是建立在对已知的氨基酸构象分析统计基础上的, 计算被预测结构的位置特异的概率. GOR 方法给出了 20 种氨基酸残基出现在不同位置时的直接信息表. 假定相邻阶段所含的信息可以近似表示为若干个直接信息的简单加和, 根据这一公式和相应的直接信息表, 就可以对一条肽链中任一位置残基的构象进行预测. 这种方法的预测准确率约为 63%.

人工神经网络是一种复杂的信息处理的机器学习模型. 这种模型最早在 20 世纪 80 年代末用于蛋白质二级结构的预测、蛋白质结构的分类、折叠方式的预测以及基因序列的分析等. 将神经网络用于二级结构预测最早是由 Qian 和 Sejnowskit 提出的, 他们受到神经网络在文字语言处理方面应用的启发, 将蛋白质序列看作是由各种氨基酸字符组成的字符序列, 将氨基酸残基片段作为输入的一串语言字符, 二级结构即为对应的输出结果. 神经网络可以有效地学习蛋白质二级结构形成的复杂规律或模式, 提取更多的信息, 并利用所掌握的信息进行预测. 利用神经网络方法可以提高二级结构预测准确率. 神经网络方法利用多序列比对的信息, 能够得到超过 70%的二级结构预测准确率. 最近 Petersen 等以位置特异性得分矩阵作为输入, 使二级结构预测的准确率达到更高的水平.

支持向量机方法是最近刚刚发展起来的蛋白质结构预测技术. 2001 年, 支持向量机首次应用于蛋白质二级结构预测, 马上就显示出这种方法的优势. 通过支持向量机方法得到的蛋白质二级结构预测准确率达到了令人惊奇的 73.5%. 之后几年, 科学家又向前走了一步, 预测准确率达到了 75.2%.

蛋白质结构域要比二级结构复杂, 预测结构域也比预测二级结构的不确定性大些. 目前蛋白质结构域的预测主要在于其折叠类型的预测. 对于蛋白质折叠类型没有一个统一的标准, 因此定义也较为混乱. 总的来说蛋白质的结构类型可以分为 α 螺旋、β 折叠、$\alpha+\beta$ 结构和 α/β 结构.

在自然状态下, 蛋白质的折叠类型不超过 1000 种, 蛋白质相互作用的数量也是有限的. 由于不同蛋白质之间的相互作用和蛋白质与相应配体之间的相互作用

都由它们的三维结构决定, 所以收集、探索和挖掘蛋白质结构数据库中的这类信息对于生命本质研究至关重要. 然而, 对于生物体基因序列的研究、这些基因可以表达的生物分子的结构的研究以及这些结构可以表现出来的功能的研究之间存在不平衡. 一方面, 沉淀在序列数据库中的数据越来越多, 通常这些序列是功能不很清楚的原始数据; 另一方面, 在蛋白质数据库 (protein data bank) 中的结构信息积累相对缓慢, 计算方法就成为预测蛋白质结构的实验方法以外的重要补充.

在蛋白质结构域的折叠类型预测方法中, 氨基酸组分方法和双组件效果的氨基酸组分方法的研究最充分. 仅依赖序列中氨基酸成分, 即仅依赖氨基酸残基在序列中的百分比而不考虑其他因素的影响, 预测准确率就可以达到 80%. 在这种方法上发展起来了双组件效果的氨基酸组分方法和双组件算法. 近十年来, 使用双组件算法用于预测蛋白质结构类可以达到很高的准确率.

然而这个准确率仍然不能满足人们的需要, 相对于 X 射线衍射方法和核磁共振方法得到的准确率仍然有一定的差距. 蛋白质二级结构预测识别率不高的原因复杂. 全面提高蛋白质二级结构预测的准确率是一个系统涉及多领域、多学科的系统工程.

第 2 章　相关知识背景

2.1　生物信息学

2.1.1　生物信息学的定义、目的、内容和发展趋势

生物信息学是一门边缘学科, 它的知识体系中包含了生物学 (生物化学、遗传学、结构生物学等)、计算机科学 (计算理论、人工智能、机器学习以及动态规划等)、物理化学 (热力学、分子建模等) 及数学 (算法、建模技术、概率论与数理统计等) 等方面的知识. 自从生物信息学这个研究领域被开辟以来, 它就以极快的速度发展并快速延伸其学科范围, 并逐渐建立了与多个学科之间的联系, 因此很难明确地界定生物信息学中各个学科之间的界限. 生物信息学主要的研究领域涉及基因组学、蛋白质组学、生物化学、数据挖掘、分子进化、分子建模以及算法等 [1].

简单、直观地从字面意思上来看, 生物信息学由 "生物" 和 "信息" 两部分组成. "生物" 部分一般指的是分子生物学, 包括进化论和遗传学; "信息" 部分指的是计算机科学. 这样把这两部分链接在一起指的就是用计算机科学的方法解决分子生物学的问题 [2]. Luscombe 在 2001 年给出一个明确的定义: "生物信息学是根据分子 (从物理化学的角度) 和信息技术 (源自应用数学、计算机科学和统计学的原则) 的应用来理解和组织与这些分子相关的大规模的信息, 即生物信息学是分子生物学的信息管理系统和诸多实践上的应用"[3]. 生物信息学可以简明扼要地定义为利用计算机方法理解、组织和解析分子生物学研究中的信息. 其实 "生物" 和 "信息" 两种学科之间结合的本质原因是有机体的生理学和行为大体上由它的基因导向, 有机体本身的生长和发育受各种信息的指挥和调节, 而生物学本身就可以认为是一种信息技术 [3].

生物信息学的研究是由大量数据驱动的, 各种各样的数据处理方法和工具被应用于分子生物学的研究中. 数据处理方法和工具的革新会推动生物信息学的发展, 计算技术和实验技术的革新使得数据快速沉淀到公共数据库中, 高容量的存储器和高技术处理器的高速处理能力提高了传统实验室数据处理效能. 在生物信息学领域另一个起到决定作用的是互联网的产生和快速发展, 通过互联网人们可以很容易地访问和交流海量的生物信息数据. 这些是生物数据急剧增长的主要原因. 由于生物信息学数据库中的数据急剧增加, 很多生物学问题实际成了计算问题. 计算机是一种理想的工具, 它不但可以大量处理数据, 而且利用恰当的软件包还能寻找这些数

据的复杂的动力学规律 [3].

生物信息学技术的发展使生物分析向两个方向发展：深度和宽度. 深度方面, 主要目标在于药物理性设计. 它的目标是取得单独的蛋白质并对其进行彻底地分析以透彻地理解这个蛋白质在生物体中的功能. 为了高效地实现这一目标, 人们设计了一整套方法. 首先对基因组进行测序, 并从中找到可读框. 然后运用恰当的方法、依据可读框翻译的蛋白质一级序列预测该蛋白质的高级结构. 利用几何计算可以确定蛋白质表面的形状, 模拟计算可以确定周围受力区域. 最后使用分子对接算法鉴别和设计可能与蛋白质结合的配体, 为药物设计铺平了道路. 然而这一整套方法中所应用的技术有些目前还不成熟, 其中有些技术还处于探索阶段, 利用这些技术一般难以得到精确的预测结果. 因此虽然使用计算工具理解生物分子的结构和功能比实验更加方便, 但是确定分子结构和功能的最可靠途径还是通过直接的实验. 从广度方面来分析, 首先是把基因同其他的基因进行比较, 以确定基因在生物进化中的位置和在有机体中可能发挥的功能. 其次, 对于蛋白质结构进行预测、研究蛋白质结构与功能的关系也是生物信息学发展的重要方向.

生物信息学的目的主要在于三个方面 [4]：

(1) 组织信息. 生物信息学组织数据的目标之一是使得查询者可以得到存在的信息并提交他们获得的新数据. 数据储存仅是生物信息学的一项基本任务, 这些存储的数据在分析之前还不能发挥作用.

(2) 数据分析. 寻找新的工具和信息源来分析数据. 例如, 把一个蛋白质序列与已知的特征序列比较, 这就不仅仅需要直接的数据查询. 生物信息学的分析工具还必须能分析有机体的基因组和蛋白质组之间有意义的共同之处, 做到这一点就需要广泛地汇聚计算理论方面的知识以及分析者对生物的生理生化规律的透彻理解.

(3) 信息释义. 使用合适的工具分析并且解释所得数据的生物学含义, 从而发现新的知识. 传统上, 生物学详细考察单个系统, 并且比较与之相关的少数几个系统. 然而对有机体的生物信息学分析则必须从当前可以得到的数据中对该有机体以及与其相关的生物系统进行全面的比较, 以便揭示涉及多个系统的一般规律和这些系统的重要特征.

从目前生物信息学的研究情况来看, 国际上公认的生物信息学的研究内容, 大致包括以下几个方面 [5]：

(1) 生物信息的收集、存储、管理与提供. 包括建立国际基本生物信息库和生物信息传输的国际互联网系统、建立生物信息数据质量的评估与检测系统、生物信息的在线服务以及生物信息可视化和专家系统.

(2) 基因组序列信息的提取和分析. 包括基因的发现与鉴定, 基因组中非编码区的信息结构分析, 提出理论模型, 阐明该区域的重要生物学功能. 进行模式生物完整基因组的信息结构分析和比较研究. 利用生物信息研究遗传密码起源、基因组

结构的演化、基因组空间结构与 DNA 折叠的关系以及基因组信息与生物进化关系等生物学的重大问题.

(3) 功能基因组相关信息分析. 包括与大规模基因表达谱分析相关的算法、软件研究, 基因表达调控网络的研究. 与基因组信息相关的核酸、蛋白质空间结构的预测和模拟以及蛋白质功能预测的研究.

(4) 生物大分子结构模拟和药物设计. 包括 RNA(核糖核酸) 的结构模拟和反义 RNA 的分子设计, 蛋白质空间结构模拟和分子设计, 具有不同功能域的复合蛋白质以及连接肽的设计, 生物活性分子的电子结构计算和设计, 纳米生物材料的模拟与设计, 基于酶和功能蛋白质结构、细胞表面受体结构的药物设计, 基于 DNA 结构的药物设计等.

(5) 生物信息分析的技术与方法研究. 包括发展能支持大尺度作图与测序需要的软件、数据库以及若干数据库工具, 如电子网络等远程通信工具. 改进现有的理论分析方法, 如统计方法、模式识别方法、隐马尔可夫过程方法、分维方法、神经网络方法、复杂性分析方法、密码学方法、多序列比较方法、统计学习理论方法等. 创建一切适用于基因组信息分析的新方法、新技术, 包括引入复杂系统分析技术、信息系统分析技术等. 建立严格的多序列比较方法. 发展与应用密码学方法以及其他算法和分析技术, 用于解释基因组的信息, 探索 DNA 序列及其空间结构信息的新表征, 发展研究基因组完整信息结构和信息网络的研究方法等, 发展生物大分子空间结构模拟、电子结构模拟和药物设计的新方法与新技术.

(6) 应用与发展研究. 汇集与疾病相关的人类基因信息, 发展患者样品序列信息检测技术和基于序列信息选择表达载体、引物的技术, 建立与动植物良种繁育相关的数据库以及与大分子设计和药物设计相关的数据库.

生物信息学发展的未来趋势主要在以下几个方面 [6]: ① 计算基因组学, 包括高通量基因组测序、模型化和注释; ② 计算结构生物学, 包括模型比较和蛋白质折叠解析; ③ 计算大分子化学, 包括解析低分辨率的折叠拓扑和高分辨率的结构; ④ 分子识别的计算分析, 包括分子对接和分子结构仿真; ⑤ 计算细胞生物学 [7].

2.1.2 基因组学

从生物信息学的数据处理的性质来看, 生物信息学包括基因组学和蛋白质组学两个方面 [8]. 20 世纪 90 年代初, 人类基因组组织很多国家的科学家和分子生物学研究机构着手展开人类基因组计划 [9], 这个计划开启了基因组时代的曙光. 人类基因组计划的目的是要测出人类每一条染色体的完整 DNA 序列, 它的主要研究工作集中于大规模的基因组测序. 第一个微生物*H. influenza* 的完整基因组测序工作完成于 1995 年. 第二年, 测序工作进程有所加快, 三个基因组*S. cerevisiae* [9], *M. jannaschii* [10] 和*M.genitalium* [11] 的测序工作相继完成. 测序技术的完善和互联网

技术的发展是基因组测序的进程加快的主要原因. 人类基因组草图于 2000 年中期完成, 于 2001 年公开发表 [12]. 在这个草图中包含了绝大部分的功能基因组和未表达的蛋白质组信息. 虽然它仅仅是草图, 仍然可以从中发现很多有用的信息.

人类基因组中大约包含 30 亿个碱基对, 人们预测包含 3 万 ~4 万个蛋白质编码基因, 其中包含和关于人类的发展、生理、医药和进化方面的重要、有用的信息. 因此人们需要有效的工具从这些数据中发现信息并快速处理积累的信息 [13]. 目前已经测序的 DNA 序列数据都在互联网上公布, 任何人都可以免费下载这些实验数据.

2.1.3　蛋白质组学

蛋白质组指的是对有机体的整个生命过程起作用的一切蛋白质的总称. 随着人类基因组草图的绘制完成, 生物信息学的研究进入后基因组时代, 并打开了蛋白质组学研究的序幕. 人类基因组计划完成后, 基因的功能和作用并未阐明, 而绘制决定生命体多样性、复杂性及其功能的蛋白质组图谱, 将使人类基因组中绝大部分基因的功能得到揭示和阐述. 人类蛋白质组研究对揭示生命活动规律和本质、探索人类重大疾病发生、发展机制具有深远的意义, 由此必将广泛推动生命科学、生物技术以及信息、分析、材料等科技领域的发展.

蛋白质组是生命活动的执行体, 是基础研究与应用研究、生命科学与医药产业及生物经济的纽带和桥梁, 是极为重要而又有限的生物战略资源. 蛋白质组研究不仅可以实现与基因组的对接与确认、直接揭示生命活动的规律和本质特点以及人类重大疾患发生与发展的病理机制, 而且可广泛推动和促进生命科学基础学科以及分析科学、信息科学、材料科学等应用学科的发展. 随着人类基因组计划的完成, 蛋白质组的研究已经成为 21 世纪生命科学发展的先导, 成为生命科学乃至自然科学最活跃的学科领域.

2003 年底, 国际人类蛋白质组计划正式启动, “人类肝脏蛋白质组计划” 和 “人类血浆蛋白质组计划”、“人类脑蛋白质组计划”、“大规模抗体计划” 和 “蛋白质组标准计划” 五大项目首先开始执行. 其中“人类肝脏蛋白质组计划”由中国科学家领导执行 [14]. 2004 年 10 月 25 日, 以 “蛋白质组学 —— 基因组的诠释” 为主题的第三届国际人类蛋白质组大会在北京隆重开幕, 2000 余位科学家齐聚一堂共同探讨人类蛋白质组研究. 会上, 安捷伦科技蛋白质组市场开发经理 Rudy Grimm 博士说: “人类基因工程成功的关键在于发展自动化程度高、易于操作, 且能够快速、大批量进行基因排序的科学技术. 然而, 目前蛋白质组学研究的开展远没有达到大批量和大产出的程度, 同时还面临着自动化程度较低, 缺乏高水平专业技术人员的局面 ……”[15].

蛋白质组学的研究比基因组学的研究更加困难. 基因的功能由碱基序列完全确

定, 而蛋白质的功能则是通过一级序列确定的、不同空间结构来实现的. 结构完全不同的蛋白质可能具有类似的氨基酸序列, 同时结构相同的蛋白质其序列差别可能很大 [16]. 生物体通常通过复制具有某种基因的多拷贝, 并且不同种类的生物当它们在进化过程中分化时通过遗传使它们具有等价的或相似的蛋白质. 在结构水平上, Chothia 预测蛋白质三维结构的数量是有限的, 这个数目在 1000~10 000[17]. 虽然 PDB 数据库中的蛋白质结构呈指数增长, 但是发现新折叠类型的速率却在下降 [18]. 因为蛋白质的折叠种类大大小于基因的种类, 蛋白质折叠的分类对于基因组的内容提供了一个坚实的简化 [19,20]. 这个基本的发现就是人们通过计算机从蛋白质一级序列预测蛋白质高级结构的依据. 管理这一层面的信息在于发展评估不同生物分子相似性的方法以及鉴别它们的相似性 [18].

2.1.4 数据库

Kanehisa 认为 "发现受数据驱动" 是后基因组时代的特征 [21]. 因此发现、递交、整理和分析数据是生物信息学的重要任务. 人们已经建立了数目庞大、种类众多的各种生物信息数据库. 这些数据库主要包括了基因序列数据库和蛋白质序列数据库, 另外还有一些数据库既收集基因序列也收集蛋白质序列. 近些年来, 人们投入了很大的人力、物力对生物信息数据进行收集和整理, 因为大量数据的存入, 使得当前的生物信息数据以指数速率膨胀. 图 2-1 和图 2-2 直观描述了 PDB 数据库和 GenBank 数据库的增长情况. 从两个图中可以看出两个数据库中的数据量都显示了呈指数增长的趋势. 造成这种现象的原因在于更新、效率更高的分析基因组

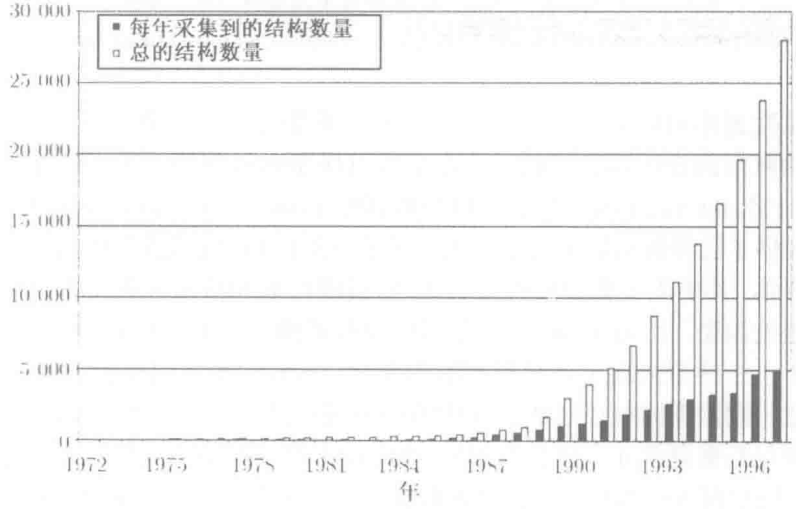

图 2-1 protein data bank 中的数据每年呈指数增长示意图
(图中数据来自 http://www.rcsb.org/pdb/holdings.html)