



蛋白质分析与数学

——生物、医学与医药卫生中的 定量化研究

(下 册)

沈世镒 胡 刚 王 奎 著
 高建召 张 拓

 科学出版社

国家科学技术学术著作出版基金资助出版

蛋白质分析与数学

——生物、医学与医药卫生中的定量化研究

(下册)

沈世镒 胡 刚 王 奎 高建召 张 拓 著

科学出版社

北 京

内 容 简 介

自生物信息学诞生以来,生物、医学与医药的研究已开始进入量化的阶段,大量生物信息数据的测定为此提供了基础与条件。本书运用多种数学理论与方法,对其中的问题进行研究,寻找其中的规律、特征与应用。

本书分上、下两册,由六部分共 25 章和 3 个附录组成。其中第一部分内容是预备知识与蛋白质一级结构分析,对一些不同学科的知识进行综合性的介绍,同时把蛋白质一级结构数据库看做蛋白质语言的文库,由此对它作相应的语法与语义分析。

第二、四部分内容是对蛋白质作空间结构分析,把蛋白质空间结构分为三维结构与空间形态结构两部分内容。其中前者是按其价键连接关系所产生的空间结构,而后者是蛋白质的空间形态特征。因此讨论它们的目标相同,但采用的数学理论、方法与模型不同。

第三部分内容是蛋白质结构中的动力学分析,其中包括分子动力学与信息动力学,动力学问题是研究蛋白质结构与功能的关键,利用这些讨论可对蛋白质分析中的许多重要问题有更深入的了解。

第五部分内容是应用部分。在此对一些重要的蛋白质作具体的结构与功能分析,并对一些应用热点与难点问题进行讨论。

第六部分内容是附录。对全书所涉及的记号、公式作统一的表达,并对一些重要概念与结论作概要说明。本书涉及大量生物信息数据,其中许多计算结果与彩色图像在光盘(见下册)中给出,也可登陆 www.sciencep.com 的下载区下载。

本书可供从事相关领域的专业人员与研究生学习与参考,尤其适用于有志从事数学与生命科学相结合的相关人员使用。

图书在版编目(CIP)数据

蛋白质分析与数学:生物、医学与医药卫生中的量化研究.
下册/沈世镒等著. —北京:科学出版社,2014.6

ISBN 978-7-03-040878-5

I. ①蛋… II. ①沈… III. ①生物数学-应用-蛋白质-分析-研究 IV. ①Q51 ②Q-332

中国版本图书馆 CIP 数据核字 (2014) 第 119339 号

责任编辑:李欣 / 责任校对:钟洋

责任印制:钱玉芬 / 封面设计:陈敬

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>

北京通州皇家印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2014 年 6 月第 一 版 开本:720×1000 1/16

2014 年 6 月第一次印刷 印张:27

字数:524 000

定价:168.00 元(含光盘)

(如有印装质量问题,我社负责调换)

前 言

定量化研究是一个十分普遍的概念,且针对某种事物特性作较精确的量化表达与研究.随着科学技术的进步,尤其是大数据及云计算技术的发展,定量化研究的过程正在向各个不同的领域拓展.本书对生命科学中所作的定量化研究是指:在分子与原子水平上,对不同的生物分子在形成过程、结构与功能关系,不同分子(或生物体)之间的相互作用及它们的动力学特性等问题作更精确的量化研究.

蛋白质是重要的生物大分子,它具有较系统的数据测定与记录,也有丰富的结构与功能分析,因此我们把它作定量化研究的实例进行分析.

由于近几十年来生物、医学与医药学科的迅速发展,以及生物信息学的诞生,这种定量化研究才得以深入.自20世纪70年代以来,多种不同类型的生物大分子结构在分子与原子水平上被测定、记录与分析.在近三四十年中,生物信息的数据量一直保持以每两三年翻一番的速度在爆炸式增长,由此奠定了生物信息学的发展基础.尤其是在近十年发展起来的第二、三代测序技术,对核酸序列数据测量的数量与价格比每年以千百倍的比例在增长,这种测量技术的惊人发展不仅是当代科学技术的巨大进步,也必然会导致对生命科学研究面貌的改变,其中最直接的影响是对生物、医学与医药领域,许多问题可作更深入的研究与精确分析.

由于生命现象的复杂性,其定量化研究仍然十分困难,其中存在许多奥秘与疑难问题,甚至还会有许多不为人们所了解的未知因素,这方面的典型问题很多,生命科学中的定量化研究现在仅仅是开始,这必将为科学的研究与发展带来巨大的机遇与挑战.

在生物大分子的定量化研究中,最大难点是其中的动力学问题.量子力学与量子化学是微观动力学的基础,它们涉及原子与分子组合中的许多基本关系问题,并由此产生更复杂的分子动力学(包括溶液的分子动力学与统计力学)问题,这些都是我们考虑问题的基础与出发点.由于生物大分子的规模都很大,大量原子与分子相互作用的动力学因素很多,如何对这些规模巨大的生物大分子的结构、功能与动力学问题进行描述与分析,寻找其中的规律是定量化研究的关键.

生物信息数据库为我们提供了大量观察与记录的数据,这样就为推动这种定量化的研究提供了可能性.利用生物信息数据库来寻找生物大分子中所存在的规则,并提出其中的问题,我们称这种动力学的分析法为信息动力学.由信息动力学得到的这些规律大部分是统计分析的规律,对其中的许多问题还不能得到完全确定的解答,这使生命科学的研究出现更多新的难题,这些问题最终还要归结到一些更深层

次的理论与应用问题。

由此可见, 对生命科学的定量化研究需要多学科的综合研究, 不同学科都已为此做了许多贡献, 但都存在许多新的问题与挑战。本书的目的是希望对蛋白质结构(一级结构与空间结构)的分析提供一些数学理论、方法与工具, 并力图与生物、医学及医药中的问题结合, 使其中的一些问题可以得到更深入的讨论。

本书由六部分共 25 章和 3 个附录组成, 其中前四部分分别对蛋白质一级结构、三维结构、空间形态特征与动力学特性等问题进行一般讨论, 采用多种数学理论与方法寻找其特征与规律。第五部分是在前四部分内容讨论与分析的基础上, 结合一些重要的蛋白质或一些重要应用与热点问题分析讨论, 其中涉及免疫系统、神经科学、酶学与流行病传播过程中的一些定量化研究问题。也对一些热点问题, 如生态的多样性、生物能源、基因组学中的一些问题提出讨论。这些问题都是人们关心的问题, 我们对其中可能存在的问题或未来的发展提出一些设想或猜想。

第六部分是附录, 是对前五部分内容的补充与说明。由于本书涉及多学科领域, 故所涉及的记号、概念与公式比较复杂, 在附录中作统一说明。附录 B 是对本书各章节的主要内容、重要概念与结果作概要性的介绍与说明。附录 C 是对附加光盘中的数据结构等内容的说明。有关数学与其他学科的一些补充知识分别在有关章节中穿插补充介绍。读者通过阅读附录可对本书的概貌有一个总体了解。

本书附光盘(见下册)一张, 无论是一些原始数据还是在计算过程中出现的数据, 它们的规模都比较大, 我们只能在光盘中给出。另外, 许多彩色图像可以加深读者对有关内容与结果的理解, 我们也在光盘中给出。即使附了光盘, 仍然不能容纳这些数据与图像文件, 因此只能用压缩文件的方式给出。读者在阅读光盘文件时要解压与存储, 才能阅读。同时光盘中的内容也可登陆 www.sciencep.com 的下载区下载。

本书是由沈世镒执笔, 胡刚、王奎、高建召与张拓协助完成, 其中第二、三代测序技术与蛋白质三维结构预测部分的内容由张拓撰写。部分内容与结果在笔者的一些有关参考文献中给出过, 一些新的问题、模型与结果在这里首次给出, 其中一些问题与观点涉及其他学科是探索性的问题。由于这些问题的复杂性与笔者水平所限, 不足之处在所难免。本书所提出的观点、理论、方法与结果供有关研究人员参考与讨论, 也欢迎读者提出意见或指出不足及错误之处。

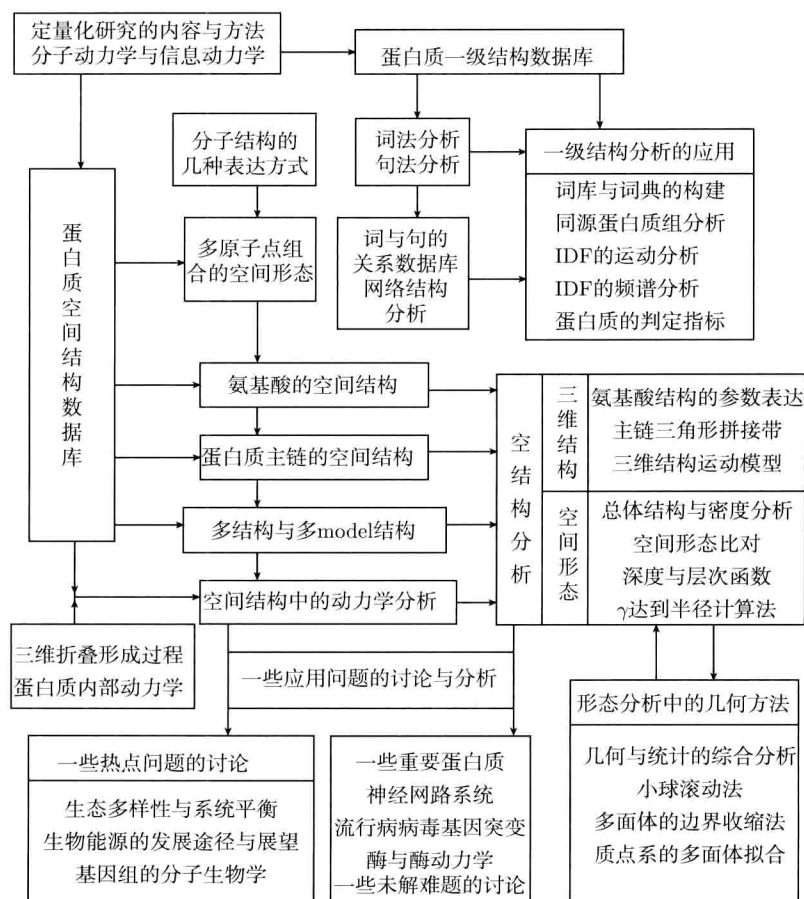
本书的 23.4 节中有关 HIV 数据是由国家疾病预防控制中心邵一鸣、何翔与冯毅等先生提供, 依据他们提供的数据所构造的比拓拓网络结构图是我们与他们合作的结果, 在此第一次正式公开, 特此说明。

本书中涉及的有关蛋白质、病毒、病菌与流行病分析中的一些结论是我们理论推导与分析的结果, 可为实际工作提供思路与参考, 但不能直接使用。如果其中的有些结论能够得到实验的证实, 那将可能为生命科学的发展带来重要影响。

关于生物能源的讨论是我们对人类未来能源发展方向与出路的探讨与猜想,我们还讨论了它们实现的可行性问题与其中的一些发展模式.但要真正实现这个猜想与这些发展模式还有很长的路程要走.但我们相信,这种可能性是存在的,也可能是必然的.

本书得到国家自然科学基金项目(20836005)与国家科学技术学术著作出版基金2013年度出版项目的支持.

感谢郭锡娟女士对沈世镒先生的生活与健康给予精心照顾,使本书能够完成.本书各章的结构关系与研究层次、内容如下图所示.



有关内容的结构关系图

作者

2013年9月

目 录

第四部分 蛋白质的空间形态分析

第 16 章 空间形态分析概论与数学预备知识	531
16.1 蛋白质空间形态结构分析概述	531
16.1.1 蛋白质空间形态结构分析的主要途径	531
16.1.2 蛋白质空间结构的指标类型	532
16.1.3 蛋白质空间结构形态的实例分析	535
16.2 形态分析的指标体系	536
16.2.1 蛋白质形态描述的总体指标	536
16.2.2 蛋白质内部与表面形态特征的指标	540
16.2.3 蛋白质内部其他特征指标	541
16.3 一些数学知识的补充与介绍	543
16.3.1 集合与拓扑空间	543
16.3.2 平面多边形与空间多面体	545
16.3.3 空间质点系与多面体	547
16.3.4 超图与晶格	548
第 17 章 蛋白质空间形态的相似性分析	552
17.1 形态的相似性问题	552
17.1.1 形态的相似性的几种不同的度量指标	552
17.1.2 序列比对算法	554
17.2 蛋白质空间形态的比对	556
17.2.1 三角形拼接带参数序列的比对	557
17.2.2 离散化的蛋白质序列比对	558
17.2.3 对长度在 300 AA 左右的蛋白质的总体相似性分析	560
17.2.4 内部结构的相似性的比对结果与分析	562
17.2.5 长短蛋白质的比对与定位	567
17.3 其他实例计算	569
17.3.1 离散情形下的实例计算	569
17.3.2 同源蛋白质族的实例比对计算	570
17.3.3 其他同源蛋白质的实例比对计算	575

17.3.4 连续情形下的实例计算	575
第 18 章 空间质点系的深度理论	579
18.1 深度的一般理论	579
18.1.1 深度的几种不同定义	579
18.1.2 T-深度与 L-深度的定义与概念	580
18.1.3 T-深度计算的基本定理	582
18.1.4 T-深度的其他性质	583
18.2 L-深度与层次函数	584
18.2.1 L1-深度的性质与推广	584
18.2.2 一些简单图形的 L1-深度, L2-深度计算	586
18.2.3 质点系中的层次函数的定义与性质	591
18.2.4 密度分布的定义	593
18.3 深度函数的计算结果与分析	595
18.3.1 深度的计算结果	595
18.3.2 T-深度与 L1-深度的关系分析	596
18.3.3 氨基酸深度的倾向性因子计算	597
18.3.4 深度倾向性因子的其他分析	600
18.3.5 氨基酸序列在蛋白质中的深度预测	603
18.3.6 层次函数的计算分析	604
18.4 利用 L-深度与密度分布对一些特殊形态的蛋白质作结构分析	606
18.4.1 一些具有特殊形态蛋白质的数据结构	606
18.4.2 图形文件与特征分析	609
18.4.3 有关内部结构的计算分析	612
18.4.4 利用分布密度来说明蛋白质的形态结构特征	614
第 19 章 空间质点系形态分析的几种基本算法	620
19.1 凸闭包与凸壳	620
19.1.1 凸闭包与凸壳的计算	620
19.1.2 凸壳的拼接计算	622
19.1.3 对 PDB2D42 蛋白质的初步分析	623
19.1.4 边界三角形的拼接计算	626
19.1.5 1GVM 蛋白质的计算	629
19.2 小球滚动法	630
19.2.1 空球的定义与分类	630
19.2.2 空球网络结构图	633
19.2.3 空球网络结构的实际计算问题	635

19.3 多面体收缩法	638
19.3.1 多面体结构性质的补充性质	638
19.3.2 多面体收缩法的原理与性质	640
19.3.3 多面体的收缩算法	641
19.3.4 小球滚动法与多面体收缩法的图表示	642
19.4 实例计算与分析	643
19.4.1 对 2D42 蛋白质的第一步小球滚动计算	643
19.4.2 多面体收缩计算	646
19.4.3 对 1GVM 蛋白质的计算与分析	649
19.4.4 空球网络结构图的分析	653
19.4.5 空球网络结构图的结构分析	655
19.4.6 其他实例计算	658
第 20 章 空间质点系的多面体拟合理论	660
20.1 多面体的拟合问题	660
20.1.1 多面体拟合的目的与要求	660
20.1.2 多面体拟合的类型与条件	662
20.1.3 多面体空间结构特征的分析	664
20.1.4 多面体的拟合计算方法	666
20.2 拟合多面体的凹凸结构性质的分析与讨论	668
20.2.1 凹凸深度的指标定义与分析	668
20.2.2 不同结构域的聚类计算	670
20.3 拟合多面体的分解与组合分析法	672
20.3.1 空间质点系的分解与组合	672
20.3.2 不同结构域的交叉重叠区	673
20.3.3 对 2D42 蛋白质的初步计算	674
20.3.4 2D42 蛋白质的活动坐标系计算与分析	675
20.4 对 2D42 蛋白质结构域的多面体收缩计算	677
20.4.1 多面体的收缩计算	677
20.4.2 对 2D42 蛋白质结构域的 T-深度切割计算	680
20.4.3 对 1GVM 蛋白质结构域的分解与组合计算	683
20.5 拟合多面体的动力学分析	689
20.5.1 蛋白质空间形态的动态分析	689
20.5.2 蛋白质与空间分子的啮合问题	690
20.5.3 靶点的动力学分析	691

第五部分 若干应用问题

第 21 章 几种球蛋白与膜蛋白分析	695
21.1 血红蛋白的结构与功能分析	695
21.1.1 血红蛋白结构性质的回顾与补充	695
21.1.2 血色素的构造	698
21.1.3 血红蛋白的变异分析	699
21.1.4 血红蛋白空间结构的数据与图形表示	701
21.2 病毒概论	702
21.2.1 病毒的分布、分类与研究	702
21.2.2 病毒的结构形态与传播过程	703
21.3 免疫系统与免疫球蛋白	705
21.3.1 免疫机制的一般概念	705
21.3.2 免疫球蛋白的空间结构	708
21.3.3 免疫球蛋白的计算分析	709
21.3.4 免疫球蛋白的结构与功能分析	716
21.4 生物膜与膜蛋白	718
21.4.1 生物膜概论	718
21.4.2 几种典型的膜蛋白	719
21.4.3 膜转运体	722
21.4.4 对膜蛋白的一些实例分析	724
第 22 章 流行病传播过程中基因突变的定量化分析	726
22.1 基因突变定量化分析的一般理论	726
22.1.1 流行病传播过程中的几个关键环节	726
22.1.2 序列比对的一般理论与方法	728
22.1.3 多重序列比对	731
22.1.4 序列比对中的一些理论问题	734
22.2 定量化分析的主要方法	737
22.2.1 基因突变的类型分析	737
22.2.2 系统树的构造与分析	738
22.2.3 拓扑距离结构图的构造与分析	740
22.2.4 拓扑距离结构图的正交分解理论	743
22.3 SARS 病毒传播过程的分析	746
22.3.1 SARS 病毒传播过程的概况	746
22.3.2 SARS 病毒传播过程的分析	747

22.3.3 正交分解理论的应用	748
22.4 HIV 病毒传播的基因突变分析	751
22.4.1 HIV 病毒传播的一些基本知识	752
22.4.2 HIV 病毒的基因与蛋白质结构	753
22.4.3 一个特殊的 HIV 病毒传播过程的量化分析	754
22.4.4 拓扑网络结构图的说明与讨论	755
第 23 章 神经网络系统与神经科学	757
23.1 神经网络系统概述	757
23.1.1 不同神经网络系统的研究内容与意义	757
23.1.2 ANNS 的研究概述	758
23.1.3 NNS 的构造与功能	760
23.1.4 HBNNS 概述	762
23.2 感知器	764
23.2.1 感知器模型与它的学习算法	764
23.2.2 感知器模型的推广	767
23.2.3 支持向量机	771
23.3 其他几种 ANNS	772
23.3.1 HNNS	772
23.3.2 HNNS 中的一些理论问题	774
23.3.3 正向与反向的 HNNS	776
23.3.4 有关智能算法的讨论与说明	778
23.3.5 对 ANNS 的评议	778
23.4 对 BNNS 的讨论	780
23.4.1 神经元与神经胶质的细胞特性	780
23.4.2 神经元与神经胶质的网络结构特征	783
23.4.3 不同感觉所产生的信号类型与特征	785
23.4.4 不同感觉所产生的信号的异同点	787
23.5 神经科学的综合研究与应用	789
23.5.1 对不同 NNS 研究的意义	790
23.5.2 感知器模型的推广	791
23.5.3 关于 ANNS 理论推广问题的探讨	794
23.5.4 对神经科学发展与应用的探讨与展望	795
第 24 章 酶与酶动力学	798
24.1 概论	798
24.1.1 酶的发现与特性	798

24.1.2 酶的命名分类	800
24.1.3 酶学杂谈	803
24.2 酶动力学	803
24.2.1 酶的动力学分析原理	803
24.2.2 酶的活力分析	805
24.2.3 酶活力的微观分析	809
24.2.4 酶的实例分析	810
24.2.5 对 ATP 酶的分析	812
24.3 酶的应用与酶工程学概述	814
24.3.1 酶的工业生产	814
24.3.2 酶的应用	816
24.3.3 酶在一些前沿科学研究中的应用	817
第 25 章 若干热点问题的探讨	821
25.1 生态的多样性与系统的平衡问题	821
25.1.1 蛋白质结构类型的多样性	821
25.1.2 从蛋白质的一级结构与空间结构看它的多样性	823
25.1.3 蛋白质多样性与生态系统的平衡性问题	825
25.2 生物能源	825
25.2.1 能源问题概论	825
25.2.2 生物能源概论	829
25.2.3 对第三代生物能源的探讨	830
25.2.4 发展第三代生物能源的其他问题	834
25.2.5 从科学幻想到科学奇迹	835
25.3 关于基因分子生物学的补充说明与介绍	836
25.3.1 基因组与染色体的一般特性	836
25.3.2 染色体与基因组的空间结构	839
25.3.3 细胞的有丝分裂与 RNA 的结构分析	842
25.3.4 基因组序列中的动力学性质	844
25.3.5 基因分子生物学中的疑难问题	846
25.4 其他难点问题	848
25.4.1 一些难点问题	848
25.4.2 造成这些难点问题的生理因素	849
25.4.3 对动力学因素的考虑	851
25.5 结束语	853

第六部分 附 录

附录 A 重要符号的说明	857
A.1 不同类型符号的说明	857
A.1.1 英文大、小写字母的表示	857
A.1.2 希腊字母的表示	859
A.1.3 有关 PDB 数据库中英文字母的表示	860
A.2 有关数学公式的表示	862
A.2.1 随机变量的概率分布与特征数	862
A.2.2 一些数学公式与符号	864
A.2.3 空间多面体与图的记号	864
A.3 尺度指标的数据与比较	865
A.3.1 尺度指标概论	865
A.3.2 一些典型的尺度指标	867
A.3.3 其他综合尺度分析	869
附录 B 一些重要概念与结果的说明	873
B.1 生物信息学与信息动力学	873
B.1.1 生物信息学	873
B.1.2 ID 的理论与方法	874
B.1.3 ID 与蛋白质一级结构的语义分析	875
B.2 一些数学理论、方法与工具	876
B.2.1 图论与拓扑空间	876
B.2.2 凸闭包与空间多面体	878
B.2.3 分子结构的几何分析	879
B.2.4 分子结构的统计分析	880
B.3 氨基酸与蛋白质的空间结构	881
B.3.1 氨基酸的空间结构与蛋白质的三维结构表示	881
B.3.2 蛋白质三维结构研究中的一些主要结论	883
B.3.3 蛋白质的空间结构形态	885
B.3.4 蛋白质的空间结构形态的分析方法	886
B.4 分子动力学理论的分析与讨论	888
B.4.1 蛋白质的形成过程与溶液中的动力学问题	888
B.4.2 蛋白质内部的自由能	890
B.4.3 蛋白质内部的结合能	891
B.5 一些应用问题的说明	892

B.5.1 一些重要蛋白质的分析	892
B.5.2 一些特殊问题的讨论与分析	895
B.5.3 酶与酶动力学	899
B.5.4 有关热点问题的讨论	905
B.5.5 基因分子生物学的补充说明与介绍	907
附录 C 有关数据库与光盘的说明	911
C.1 关于数据库与数据处理中的一些说明	911
C.1.1 关于数据库与数据处理中的一些问题	911
C.2 光盘文件的说明	912
C.2.1 DATA0 文件包	912
C.2.2 DATA1 与 DATA2 文件包	913
C.3 光盘中的所有文件一览表	914
参考文献	921
索引	934

第 16 章 空间形态分析概论与数学预备知识

我们已经说明,蛋白质的空间结构由三维结构与空间形态特征两部分组成,在本书的第三部分中已经对氨基酸、氨基酸序列与蛋白质三维结构进行了分析与讨论.在这部分内容中主要讨论它们的空间形态特征.

16.1 蛋白质空间形态结构分析概述

蛋白质空间的形态可谓五花八门、千奇百怪,因此对它们进行描述与分析十分困难,这就需要相应的数学理论与计算方法,以及相应的图形分析软件,而且需要它们交替使用.

16.1.1 蛋白质空间形态结构分析的主要途径

对蛋白质空间形态结构分析研究大体可分为软件分析与直接从头计算分析两大类,在本书中我们主张交替使用,这样可以知道这些图形结构的来源与其中的度量关系.

1. 蛋白质空间形态结构分析中使用的软件

由于蛋白质空间的形态结构的重要性与复杂性,所以有多种重要的软件包与数据库可供使用.

(1) 如 PDB 数据库,它已给出了几万种(约 90000 种)重要蛋白质的结构信息,如名称、编号与简单的功能分析,还有它们的一、二、三、四级结构数据, PDB 数据库具有统一的格式表达,其中每个蛋白质中的原子都有各自的空间坐标位置与结构特征的说明.因此在此基础上就可形成多种不同类型的软件包.

(2) 具有一些基本功能的软件包.这些基本功能有图形的描述,这些描述可为蛋白质中各种不同的原子与氨基酸提供空间位置的着色标记与在空间中旋转,从不同角度观察它们形态的功能,另外还有二级结构的标记,不同原子或氨基酸名称与图形中距离的标记,不同文件格式的转换与保存等功能.这些基本功能的软件包可在网上下载,免费使用(如 RASWIN 软件,见 <http://www.openrasmol.org> 网站等).

(3) 有些软件包除了具有这些基本功能外,对蛋白质空间结构有更详细的表述,如对由二级结构构成图像的详细描述功能、动漫式的空间形态等,它们具有更多、

更详细、更深入的分析. 这种软件包一般都在专门商业与研究部门使用, 费用十分昂贵.

(4) 利用软件包进行蛋白质空间形态结构分析的主要优点是, 可以直接利用这些工具观察到蛋白质的形态, 从中也可以得到许多结构信息, 尤其是专业型的软件包, 许多图形表达的结果可在药物设计等领域中直接参考使用. 利用软件包的主要缺点是只能在给定范围内得到它们的信息.

2. 直接从头计算分析

蛋白质空间结构数据库已经给出了大量蛋白质空间结构的许多重要的原始信息, 因此可以直接利用这些信息, 对蛋白质的空间形态、形态变化、形成的功能特征与其中的动力学因素等一系列问题进行计算与分析, 我们称这种分析法为直接从头计算分析法.

直接从头计算分析需要较多的数学理论与工具. 如果把蛋白质中各原子看做一个不规则的空间质点系, 那么对这些质点系进行分析在数学中也是一个新的课题. 对不规则的空间质点系的几何计算已是计算机技术发展中的新课题, 有许多论文与著作问世 (如文献 [260] 等). 但针对生物学中大分子结构的研究还不多见.

直接从头计算分析的主要优点是可以对我们所要求的目标进行分析, 所形成的理论、方法与工具也是软件包设计的基础. 所存在的问题主要是数学理论与工具的使用, 这不仅涉及多个数学学科分支, 许多问题还有待深入研究与发展.

16.1.2 蛋白质空间结构的指标类型

蛋白质的空间结构形态十分复杂, 用少量的几种类型或指标来说明是不可能的, 因此需要采用多种不同类型的指标, 作逐步的深入分析与讨论.

1. 总体结构指标与基本图形的类型

在对蛋白质空间形态分析中, 最简单与最基本的指标是它们的总体结构指标, 这些指标是蛋白质的大体形态特征. 例如, 球形、柱形、扁球形、棍形、伞形、蘑菇形、枣核形、枕形与葫芦形等.

在本书的光盘 DATA2/16/16-1 文件夹中, 我们给出了 171 个不同类型的蛋白质, 这些蛋白质的类型在下面还有说明. 从这些图形可以看到蛋白质的大体形态特征.

(1) **球形、柱形与棍形.** 球形的基本特征是存在一个中心点, 由该点沿空间的各方向达到蛋白质边缘的距离大体相同, 典型的球形结构是多种不同类型的球蛋白 (如血红蛋白、免疫球蛋白等), 它们在蛋白质中大量存在. 柱形蛋白存在一条中心轴, 由该轴的不同位置沿空间的各方向达到蛋白质边缘的距离大体相同.

柱形的形态又可分圆柱形与方柱形, 它们的主要区别是在沿中心轴作横截面时, 方柱形的截面形态是矩形, 而圆柱形的截面形态是圆或椭圆. 柱形的拉长就是棍形.

(2) **伞形、蘑菇形、枣核形、枕形与葫芦形**. 这 5 种形态都可看成柱形的变形. 伞形与蘑菇形是在柱形的结构中一头变大; 另一头变小, 称变小的部分为该形态的尾部. 枣核形是在柱形的结构中段变大; 上下两头变小. 枕形与枣核形的特征相反, 它中段变小, 上下两头变大.

葫芦形是在柱形中轴的某些位置到蛋白质边缘的距离变大, 而在中轴另一些位置到蛋白质边缘的距离变小.

(3) **扁球形**. 扁球形是球形的变形, 它的基本特征是仍然存在一个中心点, 由该点沿空间某平面的各方向达到蛋白质边缘的距离大体相同, 而沿该空间平面的垂直方向到蛋白质边缘的距离较短. 带形也是柱形的一种变形, 这时中心轴的长度变得很长, 而中心轴到蛋白质边缘的距离较短.

称这 9 种形态为蛋白质总体结构的基本类型, 这些基本类型还可能发生许多不同的变形, 如球形的各个方向不一定均匀、柱形与棍形会发生弯曲等.

2. 基本类型的变形

这 9 种基本类型存在许多变形. 这些变形也很复杂, 一些简单变形如下.

(1) 总体结构特征的变形. 如在柱形、棍形、枕形中出现弯曲、局部凹凸等变形, 伞形、蘑菇形、枣核形、葫芦形就可看做柱形、棍形、枕形的变形.

(2) 在蛋白质内部出现的**空洞结构**. 这就是在蛋白质内部出现一个较大的区域, 在此区域中不存在该蛋白质中的任何原子点.

(3) 蛋白质内部的**管道**. 这就是在蛋白质内部出现一个半径较大的管道, 如果管道从蛋白质的一侧通向该蛋白质的另一侧, 而且在此管道中不存在该蛋白质中的任何原子点. 这种穿越蛋白质的管道就变成**通道**.

(4) 在蛋白质表面出现各种不同类型**凹凸区域结构**. 这就是在蛋白质表面形成一些较大的凸块, 而在这些凸块之间又会形成一些凹形结构.

(5) 这些简单变形的相互交叉可产生更复杂变形. 如大洞四周套许多中洞或小洞, 大凹槽四周套许多中小凹槽或空洞, 管道有可能形成蛋白质内部的管道网络结构. 这些变形使蛋白质的空间形态更加复杂.

由此可见, 对蛋白质空间结构形态描述分**总体结构基本类型、简单的结构变形与复杂的结构变形等**.

3. 影响蛋白质总体形态结构变化的其他因素

在蛋白质空间结构数据库中, 除了总体结构类型、简单的结构变形与复杂的结