蛋白质分析与数学

——生物、医学与医药卫生中的定量化研究

（上 册）

沈世镒 胡 刚 王 奎 高建召 张 拓 著

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

(上册)

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定量化的阶段. 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. 其中前者是按共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定量化研究. 上册/沈世镒等著. —北京：科学出版社，2014. 6

ISBN 978-7-03-040840-2

Ⅰ. ①蛋 Ⅱ. ①沈… Ⅲ. ①蛋白质–结构分析–生物数学–研究 Ⅳ. ①Q510.1

中国版本图书馆 CIP 数据核字 (2014) 第 116773 号

责任编辑：李 欣 赵彦超 / 责任校对：刘小梅
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：34
字数：665 000

定价：198.00 元

(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进行分析讨论, 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对拓扑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.

目　录

第一部分　概　论

第二部分　蛋白质的三维结构分析

第三部分　蛋白质结构的动力学分析

第 1 章　生物信息学与信息动力学

在前言中, 我们已经对**定量化**研究的目标与本书的主要内容作了简单的说明, 在本章中再围绕其中的有关内容作进一步的补充与讨论.

1.1　生物信息学与定量化研究

探讨生命科学中的定量化研究问题是本书的主要目的. 主要的研究方法与手段是**生物学**、**生物信息学**以及其他有关的学科, 其中包括**动力学**中的一系列问题, 近年来对这些问题的研究虽有许多重大的发展, 但仍然存在许多理论与难解问题需要解决.

1.1.1　定量化的研究目标、内容与途径

前言中已经说明, 本书所讨论的定量化研究主要是指**生物大分子**(如 DNA 与 RNA 的核苷酸序列, 蛋白质、肽链的氨基酸序列等) 在它们的形成过程、结构、功能与相互作用等关系问题, 对其中的这些问题作更精确的、有定量化指标的描述与研究.

由于生物学的研究进展、生物信息学诞生与发展、多学科的介入等因素, 这种定量化的研究有可能更深入地进行. 国际上把这种研究看做生命科学的最新发展, 并向生物、医学与医药等应用领域拓展.

1. 从蛋白质的结构分析看定量化研究的特征

蛋白质是重要的生物大分子, 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. 因此在本书中, 我们把它当作定量化研究的实例进行分析, 由此也可以看到定量化研究的特征与存在的问题. 从以下例子可以看到定性化与定量化研究的区别.

(1) **蛋白质的定义与判定**. 按生物学界的定义: **蛋白质是由核酸序列转译, 并具有一定的空间形态结构与功能的氨基酸序列**. 这是一种定性化的描述. 定量化的研究就是讨论能否给出一些**定量化指标**, 这就是一个氨基酸序列, 要满足什么样的结构条件才有可能形成具有空间折叠结构, 并产生生物功能的蛋白质.

这种定量化的指标如何确定, 它们能否成为蛋白质的基本度量与判定指标, 如何在**基因识别**, 或其他**蛋白质设计**、**构造**或**改造**等研究中应用.

(2) **生命语言的解读**. 生物信息数据库是对多种不同类型生物大分子的大量观察、测量与记录的结果, 因此是**生命语言的记录与汇合**. 对这些语言能否解读, 如何解读, 每一种数据库都可看做一种特定的语言, 它们是否存在各自的**词法与语法关系**, 这些关系如何表达, 这与人类自然语言有哪些异同.

(3) 尤其是在生物、医学中存在多种重要语言之间的**应答关系**的表达. 如**免疫机制**、**酶的催化过程**、**配体与受体**的相互作用、**基因组与蛋白质**之间的转换关系等. 这是生命科学中的基本关系, 能否用生命现象中的语言关系来表达.

(4) **生物大分子的多级结构研究**. 无论是核酸序列还是蛋白质序列, 它们都存在一、二、三、四级的多级结构, 不同的结构之间的相互关系, 不同的结构如何产生, 形态与功能的关系如何 (如何形成功能效果), 其中的动力学因素又是什么.

在蛋白质结构数据库中, 有许多比较完整的多级结构数据, 也有许多功能的说明. 这正是我们选择蛋白质作定量化分析的理由.

这些问题都是解读生命语言的重要内容, 因此使生命科学的研究变得十分复杂与困难. 对这些问题虽有许多研究与成果, 但从定量化的角度来看, 仍有许多未解决的问题. 在深入到定量化的研究中必然会存在许多新的困难, 这个研究过程必然是一个逐步深入的研究过程, 现在还只能算是刚刚开始的起步阶段.

2. 蛋白质结构与功能分析中的定量化研究

蛋白质是一个由数百到数十万个原子所组成的生物大分子, 对它们的研究存在一系列定量化的问题, 因此可以成为定量化研究的切入点.

(1) **蛋白质一级结构分析**. 蛋白质一级结构是由氨基酸序列组成的数据库, 如果把它的数据库看成蛋白质语言的文库, 那么就存在对该文库的语法分析与语义解读问题.

(2) **空间质点系的描述**. 蛋白质是由几十到几十万 (最多可达几千万) 个原子组成的生物大分子, 因此是一个十分复杂与不规则的空间质点系. 这里, 首先要讨论它们的结构特征问题, 在生物学中称为**空间结构**.

在空间结构中又可分**三维结构**与**空间形态**结构. 其中三维结构是指由共价键连接的空间结构, 而对空间形态又可分**总体结构**特征、**内部**与**表面**等结构特征.

在三维结构中又分**二级结构**、**超二级结构**等类型. 在蛋白质空间结构的研究中又分**形成过程**与**最后的形态特征**问题. 在生物学中有多种模型讨论, 如**自由能的最小化收敛性理论**、**熔球态模型**、**一级结构**与**空间结构**的关系问题, 如 Alignment 的一级结构确定空间结构理论的讨论.

(3) **结构与功能的关系**. 各种不同类型的蛋白质执行不同的生物功能, 这些功能与蛋白质的空间结构密切相关, 如何分析它们之间的相互关系是蛋白质研究中的重要问题. 在生物学的研究中已经确定, 不同蛋白质都有各自的**活性特征**与**活性中**