



蛋白质分析与数学

——生物、医学与医药卫生中的 量化研究

(上册)

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

(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 1 章 生物信息学与信息动力学	3
1.1 生物信息学与定量化研究	3
1.1.1 定量化的研究目标、内容与途径	3
1.1.2 不同学科的作用与贡献	5
1.1.3 生物信息学概论	7
1.1.4 生物信息学的近期发展与未来动向	10
1.2 第二、三代测序技术	13
1.2.1 第一代核酸序列测量技术的发展原理	13
1.2.2 第二代测序技术的原理与特征	15
1.2.3 第二代测序平台	16
1.2.4 第二代测序的应用	18
1.2.5 测序技术的临床应用及展望	20
1.3 信息动力学概述	22
1.3.1 ID 的目的与意义	22
1.3.2 ID 与物理学、生物学的关系	24
1.3.3 对数据库的说明	28
1.4 ID 的基本原理与方法	31
1.4.1 ID 的基本方法之一: 信息统计法	31
1.4.2 几种重要类型的 IDF	33
1.4.3 由 IDF 产生的词法分析	35
1.4.4 ID 的基本方法之二: 组合分析法	39
1.4.5 点线图的基本知识	43
1.5 语义分析概要	46
1.5.1 词与词法分析要点	46
1.5.2 词与句的关系数据库	48
1.5.3 由关系数据库做有关语法问题的讨论	50
1.5.4 词与句的网络结构	52
1.5.5 PIDF 的因子分解理论	53

1.5.6	PIDF 的运动分析与其他类型的分析	56
第 2 章	蛋白质一级结构数据库的 ID 的计算与分析	59
2.1	蛋白质一级结构数据库的 ID 计算	59
2.1.1	蛋白质结构分析概论	59
2.1.2	SP'06 数据库的一般性质	61
2.1.3	ID 的计算结果与初步分析	65
2.2	词法与句法分析	72
2.2.1	词法分析	72
2.2.2	词库 $\mathcal{D}$ 的极小化与极大化网络结构	78
2.2.3	词与句的关系数据库	81
2.3	相似蛋白质组的网络结构分析	84
2.3.1	对相似蛋白质组的搜索计算	84
2.3.2	相似蛋白质组的布尔网络结构	89
2.3.3	利用相似蛋白质组做人造蛋白质的构造设计	91
2.3.4	一些重要的小肽、寡肽与小蛋白	95
2.4	蛋白质的 IDF 与 PIDF 分析	102
2.4.1	计算结果与初步分析	102
2.4.2	计算结果的初步统计分析	104
2.4.3	协方差矩阵与相关矩阵的计算结果	106
第 3 章	分子生物的参数控制系统与 IDF 的控制问题	107
3.1	分子生物的参数控制系统	107
3.1.1	一些基本概念	107
3.1.2	对因子分解理论的补充说明	108
3.1.3	生物参数控制系统的数学模型	110
3.1.4	分子运动的动力学非线性控制系统	112
3.2	数据阵列 PIDF 的运动分析	113
3.2.1	PIDF 运动分析的内容与意义	113
3.2.2	数据阵列 $\mathcal{K}_s$ 的运动方程	115
3.2.3	所有 $\mathcal{K}_s$ 数据阵列的运动区域	119
3.2.4	驱动因子的运动分析	123
3.3	蛋白质的判定问题	123
3.3.1	训练集与检测集	123
3.3.2	判定方法与它的依据	125
3.3.3	蛋白质判定的计算结果与分析	126
3.3.4	若干问题的分析与讨论	127

3.4 蛋白质 M-PIDF 的频谱分析	129
3.4.1 频谱分析概论	129
3.4.2 M-PIDF 的 Fourier 变换	131
3.4.3 不同长度蛋白质的频谱结构分析	133
3.4.4 蛋白质数据库的频谱分析	136
第 4 章 预备知识	141
4.1 分子的空间结构表示	141
4.1.1 分子结构的描述与表达	141
4.1.2 有关数学工具的说明	142
4.1.3 分子官能团的组合与分解	144
4.2 空间结构的稳定性分析	145
4.2.1 分子官能团的稳定性的定义	145
4.2.2 不稳定质点系的描述与参数类型	148
4.2.3 分子空间结构的综合分析	149
4.3 分子结构的点线图表示	149
4.3.1 分子点线图的定义	149
4.3.2 分子点线图的分解与组合	153
4.3.3 几种典型稳定的分子官能团的图表示	156
4.3.4 点线图的一些子图结构	159
4.3.5 点线图的组合与分解	162
4.4 活动坐标系理论	165
4.4.1 活动坐标系的定义与性质	165
4.4.2 活动坐标系的构造	166
4.4.3 分子点线图的其他坐标参数	168
4.4.4 活动坐标系中的旋转变换理论	169
第 5 章 四原子与多原子空间结构的几何模型	171
5.1 四原子点的空间几何结构	171
5.1.1 空间四原子点的结构表示与它们的参数系	171
5.1.2 基本参数系的相互关系	173
5.1.3 四原子点的位相分析	176
5.1.4 四原子点参数的稳定性问题	177
5.2 多原子点结构分析的几何理论	178
5.2.1 有关记号与类型	178
5.2.2 五原子点的参数系	179
5.2.3 若干特殊四原子或五原子点	180

5.2.4 一般多原子点的参数系	183
5.3 四原子点分子在溶液中的随机运动	184
5.3.1 随机运动的基本特征	184
5.3.2 四原子点在溶液中的随机运动模型与参数分析	185
5.3.3 关于转动角度取值范围的讨论	187
5.3.4 扭角取值范围移动后的效果讨论	188

第二部分 蛋白质的三维结构分析

第 6 章 氨基酸的一般性质与它的分子官能团	195
6.1 氨基酸概论	195
6.1.1 氨基酸的化学成分与性质	195
6.1.2 氨基酸的相互连接	200
6.1.3 遗传密码子	202
6.2 氨基酸的分子成分与结构特征分析	204
6.2.1 氨基酸的分子结构模型	204
6.2.2 氨基酸侧链中的非氢原子骨架图	206
6.2.3 氨基酸侧链中非氢原子的层次函数表示	208
6.3 氨基酸空间结构的分解与分析	211
6.3.1 氨基酸中所有原子的表示	211
6.3.2 氨基酸中存在基团的构造与类型	213
6.3.3 氨基酸中的原子结构全图	215
6.4 氨基酸中 42 种基本基团的结构计算	216
6.4.1 基本基团的结构类型	216
6.4.2 对丙氨酸的结构分析与计算	218
6.4.3 对其他氨基酸花盆结构的计算与分析	219
6.4.4 甘氨酸中的原子结构	220
6.4.5 脯氨酸的原子结构	222
第 7 章 氨基酸中具有稳定空间结构基团的计算与分析	225
7.1 重要基团的计算结果	225
7.1.1 对氨基酸不变部分原子结构的计算与分析	225
7.1.2 具有中心点基团的计算	227
7.1.3 具有环形环的侧链结构	229
7.1.4 氨基酸侧链中有关基团镜像的讨论	232
7.1.5 在活动坐标系下的计算	232
7.2 氨基酸与氨基酸侧链的运动与变化模型	234

7.2.1	氨基酸的全着色图	234
7.2.2	氨基酸侧链的参数表达	237
7.2.3	扭角分布的计算结果	238
7.2.4	扭角取值分布曲线类型的讨论	239
7.3	扭角在不同层次与活动坐标系中的取值分布	241
7.3.1	扭角在不同层次中的运动与变化	241
7.3.2	关于镜像的分布计算与分析	243
7.3.3	氨基酸侧链的珠链模型	246
7.3.4	氨基酸侧链第 2 层非氢原子点的类型与分析	247
7.3.5	氨基酸侧链第 2 层非氢原子点的运动状况在活动坐标下的表示	250
7.3.6	计算结果与分析	252
7.3.7	氨基酸其他层次原子的扭角类型与计算	256
7.4	氨基酸侧链参数表达的因子分解	260
7.4.1	氨基酸侧链参数表达的基本数据与它们的特征数计算	260
7.4.2	参数表达的主因子分析	263
7.4.3	因子分解的计算结果与分析	265
7.5	氨基酸空间结构分析中的其他问题	268
7.5.1	氨基酸中氢原子位置的预测	268
7.5.2	氨基酸空间结构分析小结	270
7.5.3	氨基酸梢点的空间运动计算与分析	272
第 8 章	蛋白质主链的三角形拼接带	276
8.1	概论	276
8.1.1	蛋白质三维结构概述	276
8.1.2	主链三角形拼接带的类型与参数	277
8.1.3	三角形拼接带的有关性质	280
8.1.4	计算结果与分析	281
8.2	对大、小三角形拼接带的结构分析	283
8.2.1	小三角形拼接带的基本特征	283
8.2.2	扭角分布与氨基酸的关系分析	285
8.2.3	大、小三角形拼接带的关系分析	289
8.3	三角形拼接带的其他性质	292
8.3.1	三角形拼接带上下边的性质	292
8.3.2	三角形拼接带的扭角转动	293
8.3.3	三角形拼接带的平面展开	295
8.3.4	计算公式与计算结果	296

8.3.5 蛋白质主链小三角形拼接带的参数表达与因子分解	297
第 9 章 主链的中位点曲线分析	299
9.1 中位点曲线的定义与性质	299
9.1.1 中位点曲线的定义记号与一般性质	299
9.1.2 中位点曲线的有关参数的性质	300
9.1.3 计算模型与结果	303
9.1.4 初步统计分析结果	304
9.2 利用中位点曲线对蛋白质二级结构关系的讨论	306
9.2.1 蛋白质空间结构中的一些特殊结构	306
9.2.2 中位点曲线在二级结构分析中的应用	307
9.2.3 实例分析	310
9.2.4 对血红蛋白的分析	314
9.3 中位点曲线的特征分析	319
9.3.1 中位点曲线的平面展开图的特征分析	319
9.3.2 蛋白质中位点曲线的一些重要指标	322
9.3.3 对蛋白质总体指标的计算结果与说明	324
9.4 对计算结果的分析与说明	327
9.4.1 对 α 螺旋结构的分析与判定	327
9.4.2 对一些不同类型蛋白质实例的分析	329
第 10 章 部分原子的空间结构分析	331
10.1 二氨基酸序列中部分原子的空间结构	331
10.1.1 部分已知结论的回顾	331
10.1.2 关于 A, C, O, H', N', A' 六原子点的特性讨论	333
10.1.3 部分原子点位置的预测	335
10.1.4 二氨基酸序列双底座原子集合空间结构的讨论	337
10.1.5 对三氨基酸序列中部分原子的计算	339
10.1.6 出现脯氨酸的情形	340
10.2 蛋白质中位点曲线的参数系数与蛋白质的判定算法之二	342
10.2.1 由二肽或三氨基酸序列对中位点曲线转角扭角的计算	342
10.2.2 二氨基酸序列的折叠系数分析	344
10.2.3 蛋白质折叠系数的定义与计算	347
10.2.4 蛋白质的判定条件之二与判定结果	348
10.3 多氨基酸序列侧链的运动	350
10.3.1 多氨基酸序列主链与侧链的关系图	350
10.3.2 主要计算结果与初步统计分析	352

10.3.3	两组六原子点的结构分析	353
10.3.4	计算结果与分析	355
10.4	侧链在活动坐标系中的运动分析	359
10.4.1	二肽与三氨基酸序列的活动坐标系	359
10.4.2	B 与 Γ 原子的运动坐标	361
10.4.3	B 原子点与梢点 Γ 在不同二肽与三肽的不同类型	364
第 11 章	结构域与侧链的修饰	371
11.1	概论	371
11.1.1	部分重复序列与结构域的构造问题	371
11.1.2	侧链修饰的研究	375
11.1.3	重复序列在蛋白质数据库中的表达	376
11.2	在 α 螺旋中侧链的修饰	377
11.2.1	α 螺旋的结构模型	377
11.2.2	计算结果	379
11.2.3	计算结果的分析	383
11.3	β 折叠与其他特殊结构的讨论分析	388
11.3.1	β 折叠的结构分析	388
11.3.2	关于 Ω 结构的讨论	392
11.3.3	蛋白质局部三维结构的综合讨论	393
11.4	结构域与 Model 的结构特征问题	394
11.4.1	结构域的类型与特征	395
11.4.2	血红蛋白的结构域	395
11.4.3	蛋白质中的 Model 结构	397
11.4.4	蛋白质三维结构的预测与 CASP 比赛	399

第三部分 蛋白质结构的动力学分析

第 12 章	有关分子动力学的基础知识	407
12.1	有关统计力学的一些基本知识	407
12.1.1	统计力学中的一些基本概念与公式	407
12.1.2	原子与分子在溶液中的随机运动	409
12.1.3	原子与分子之间的相互作用	411
12.1.4	原子与分子之间的一些能量参数	412
12.2	化学反应的基础知识	415
12.2.1	原子与分子的特征与化学反应的基本方程	415
12.2.2	化学反应的基本类型	416

12.2.3	化学反应的基本规律	418
12.2.4	化学反应中的能量分析	419
12.2.5	化学反应中的动力学指标、平衡系数与反应速率	420
12.3	几种重要的生物分子官能团与它们的化学反应	424
12.3.1	一些重要的分子官能团	424
12.3.2	几种重要的化学反应的类型与方程式	426
12.3.3	化学反应的动力学特征分析	428
12.3.4	催化反应简介	430
12.4	水与溶液的分子动力学特征	432
12.4.1	水分子的形态特征	432
12.4.2	水分子与其他分子官能团的相互作用	433
12.4.3	水与溶液分子的动力学	435
12.4.4	与溶液有关的动力学指标	436
第 13 章	蛋白质三维结构中的动力学问题	439
13.1	蛋白质空间结构形成过程中动力学的几个基本观点	439
13.1.1	关于蛋白质空间结构形成过程的讨论	439
13.1.2	自由能与结合能	441
13.1.3	动力学模型中的基本特征	443
13.1.4	运动方程的可计算性与收敛性问题	446
13.2	关于自由能的讨论	447
13.2.1	有关自由能定义的讨论	447
13.2.2	用负 KL-互熵作分子内部自由能定义的合理性问题	449
13.2.3	KL-互熵的可计算性问题	449
13.2.4	KL-互熵的近似计算	452
13.3	KL-互熵的估计与计算	453
13.3.1	简化计算的考虑依据	453
13.3.2	计算结果与初步讨论分析	454
13.3.3	蛋白质判定条件之三	456
13.3.4	蛋白质的 KL-互熵与 PIDF 的关系讨论	458
第 14 章	蛋白质的分子动力学特征分析	461
14.1	蛋白质分子动力学的特性要点	461
14.1.1	蛋白质空间结构中的结合能	461
14.1.2	蛋白质的活性特征分析	462
14.1.3	蛋白质三维折叠速率与瞬时速率的因素分析	465
14.1.4	蛋白质空间结构内部的结合能	466

14.2 化学键的动力学特性	467
14.2.1 化学键的一些基本特征	467
14.2.2 蛋白质中化学键的类型分析	469
14.2.3 其他非固定共价键的讨论	472
14.2.4 非固定化学键的动力学	475
14.3 氢键与范德华力的动力学特性	476
14.3.1 氢键的形成条件与特征	477
14.3.2 蛋白质中可能产生氢键的类型分析	478
14.3.3 范德华力的动力学分析	478
14.4 氨基酸与蛋白质中极性问题的讨论	480
14.4.1 极性的一般理论	480
14.4.2 氨基酸与蛋白质中部分原子的极性讨论	482
第 15 章 对氢键、离子键与共价键的搜索、计算与讨论	484
15.1 搜索计算方法与结果	484
15.1.1 对不同类型键的搜索与判别	484
15.1.2 搜索与计算结果	485
15.1.3 计算结果的初步分析	486
15.2 对化学键与氢键的进一步分析	488
15.2.1 对二硫键与离子键的分析	489
15.2.2 关于非固定共价键的分析	492
15.2.3 对氢键的补充分析	495
15.3 氨基酸的动力学倾向性因子与分子聚合团的分析	496
15.3.1 氨基酸的动力学倾向性因子分析	496
15.3.2 双氨基酸的动力学倾向性因子	498
15.3.3 分子聚合团的定义与计算	501
参考文献	507
索引	518

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、多学科的介绍等因素, 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起步阶段.

2. 蛋白质结构与功能分析中的定量研究

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**