

医药分析信息学 及 分析数据处理技术

程翼宇 瞿海斌 等编著



化学工业出版社
生物·医药出版分社

医药分析信息学 及 分析数据处理技术

程翼宇 瞿海斌 等编著



化学工业出版社
生物·医药出版分社

· 北京 ·

图书在版编目 (CIP) 数据

医药分析信息学及分析数据处理技术/程翼宇, 瞿海斌等编著. —北京: 化学工业出版社, 2006. 6
ISBN 7-5025-8896-5

I. 医… II. ①程…②瞿… III. ①药物分析-信息学
②药物分析-信息处理 IV. R917

中国版本图书馆 CIP 数据核字 (2006) 第 061818 号

医药分析信息学及分析数据处理技术

程翼宇 瞿海斌 等编著

责任编辑: 麻雪丽

责任校对: 于志岩 李 林

封面设计: 史利平

*

化学工业出版社 出版发行
生物·医药出版分社

(北京市朝阳区惠新里 3 号 邮政编码 100029)

购书咨询: (010)64982530

(010)64918013

购书传真: (010)64982630

<http://www.cip.com.cn>

*

新华书店北京发行所经销

大厂聚鑫印刷有限责任公司印刷

三河市万龙印装有限公司装订

开本 787mm×1092mm 1/16 印张 18½ 字数 454 千字

2006 年 9 月第 1 版 2006 年 9 月北京第 1 次印刷

ISBN 7-5025-8896-5

定 价: 45.00 元

版权所有 违者必究

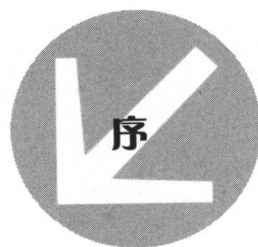
该书如有缺页、倒页、脱页者, 本社发行部负责退换



程翼宇教授简介

程翼宇博士毕业于浙江大学，曾赴美国哈佛大学医学院、FDA / 国家毒理研究中心等欧美许多大学及药物研究单位留学或访问研究。现任浙江大学求是特聘教授、药学院副院长、药物信息学研究所所长；兼任国家技术前瞻研究组专家、973 计划项目专家组成员及《化学学报》等 6 家学刊编委，享受国务院政府特殊津贴。

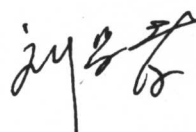
他主要从事药学、中药学及生命信息科学等多学科前沿研究，主持研究完成 973 计划等国家重大课题 16 项，领衔获国家科技进步二等奖 1 项、省级奖 11 项，发表 SCI 收录论文 78 篇及 EI 论文 30 余篇。



生物技术和信息技术是新世纪科学技术领域中最具影响力和快速发展的学科，它们推动了生物医药的高速发展。当今生物医药科技的发展在很大程度上取决于分析测试技术和信息技术的发展，特别是“组学”分析研究过程中形成的海量数据更有赖于数据挖掘、特征提取和统计分析等生物信息技术的科学应用。浙江大学程翼宇教授等编著的《医药分析信息学及分析数据处理技术》一书，无疑将会对推动我国医药分析信息学科的发展起到重要作用。

当代生物医药科技在发生深刻的变化，多学科渗透、交叉和融合促进了医药分析技术的发展，对医药信息采集、变换、传递、辨析及应用已开始形成一个充满生机的研究领域，医药分析信息处理已成为生物医药研究领域的前沿。此书用 10 章的篇幅，从理论到应用，从原理到方法，较系统地论述了医药分析信息学和分析数据处理技术的诸多问题，分别对分析数据处理技术的基础知识、分析信号处理、模式信息的辨别处理、指纹图谱的计算处理、信息智能管理、生物芯片和组学分析信息处理等方面作了阐述，从中看到了学科发展和应用现状，更能看到学科发展的前沿和前景。本人确信该书的出版一定会使读者受益匪浅。

中国工程院院士

 教授

2006. 6. 20



进入 21 世纪以来,生物尖端科技领域风起云涌,创新研究成为竞速发展的核心内容,科技创新正在深刻地改变众多学科发展的传统轨迹。审视正在发生的科技剧变,药理学、分析化学、信息科学、生命科学等学科都面临着严峻的挑战和重大发展契机,处于多学科交汇区的分析信息获取和处理技术已成为生物医药等领域的重要研究前沿和关键技术支撑点,集聚了巨大的创新潜能和机会。

欣逢又一个科技春天降临中华大地,肩承建设创新型国家的历史使命,深受鼓舞和鞭策。站在中华民族波澜壮阔的 15 年自主创新征程端口,程翼宇教授携博士生与瞿海斌教授一起合作,将多年累积的资料和最新研究成果撰写成本书,试图服务于医药与生物科技创新研究领域及分析化学界的同仁们,为读者提供分析信息获取和处理技术方面的知识与应用实例。

本书共 10 章,第 1 章由程翼宇教授撰写,第 2 章由瞿海斌教授负责撰写,第 3 章由博士生马捷撰写,第 4 章由程翼宇教授撰写,第 5 章由陈闽军博士与博士生王毅共同撰写,第 6 章由瞿海斌教授撰写,第 7 章由范骁辉博士撰写,第 8 章由博士生李云飞和王毅共同撰写,第 9 章由博士生王学伟撰写,第 10 章由博士生白景清为主陈闽军博士参与撰写,程翼宇教授负责全书的内容编排及部分章节修改工作。

本书的出版得到了国家 973 计划 2005CB523402 课题资助和化学工业出版社的大力支持,在此一并致以衷心的感谢。

由于作者水平所限及时间仓促,书中不足及疏漏之处在所难免,恳请读者斧正。

作者

2006 年 2 月于浙江大学求是园



第1章 绪论	1
1.1 分析科学发展前沿若干问题	1
1.2 涉及的若干概念、界定及术语定义	2
1.3 分析信息获取问题	3
1.4 生物医药分析技术发展趋势	5
1.5 药物信息学与生物医药分析	5
1.6 复杂物质体系辨析与医药分析信息学	7
第2章 分析数据处理技术基础知识	10
2.1 化学计量学与化学信息学	10
2.1.1 化学计量学	10
2.1.2 化学信息学	10
2.2 多元统计分析技术	11
2.2.1 基本概念	12
2.2.2 常用的多元统计分析方法	13
2.3 机器学习技术	15
2.3.1 机器学习的定义	16
2.3.2 机器学习的类型	16
2.3.3 机器学习的应用领域	17
2.3.4 几种常见的机器学习方法	17
2.4 模式信息处理技术	18
2.4.1 基本概念	19
2.4.2 模式识别基本方法	20
2.5 知识工程技术	24
2.5.1 知识工程定义	25
2.5.2 知识表示方法	25
2.5.3 知识搜索	26
2.5.4 知识推理机制	27
2.5.5 专家系统	28
2.6 人工神经元计算技术	33
2.6.1 神经网络的定义	33
2.6.2 基本概念与原理	33
2.6.3 神经网络的应用	35

2.6.4 几种常见的神经网络模型	35
2.7 数据库与数据挖掘技术	37
2.7.1 数据库技术概述	37
2.7.2 数据库的基本概念	38
2.7.3 数据仓库和数据挖掘技术	38
2.8 小波分析技术	44
2.8.1 基本概念	44
2.8.2 小波分析的基本原理	45
2.8.3 信号小波级数的展开及其变换算法的实现	48
2.9 数据可视化技术	49
2.9.1 数据可视化概念	49
2.9.2 数据可视化过程	50
2.9.3 数据可视化方法	51
2.9.4 数据可视化应用	51
2.9.5 展望	52
2.10 化学信息集成处理技术	53
参考文献	53
第3章 现代仪器分析信号处理	55
3.1 分析信号处理技术概述	55
3.2 分析仪器的信号与噪声特性	56
3.3 数字滤噪方法分类	59
3.4 滤噪效果评价方法	59
3.4.1 整体滤噪效果评价指标	60
3.4.2 局部滤噪效果评价指标	60
3.5 频带滤噪法	61
3.5.1 多点归并平滑法	61
3.5.2 加权平均平滑法	61
3.5.3 SG 数字滤波器	62
3.5.4 可调型低通滤波器	64
3.5.5 几种常用频带滤噪法的性能比较	66
3.6 最佳滤波器	68
3.6.1 匹配滤波器	69
3.6.2 维纳滤波器和卡尔曼滤波器	69
3.7 基于小波变换法的滤噪技术	69
3.7.1 小波基自适应滤噪技术	70
3.7.2 基于小波包基分解与重构的滤噪技术	75
3.8 自适应滤波器	76
3.8.1 最小均方自适应滤波器	77
3.8.2 神经网络滤波器	78

3.9 新型滤噪方法研究实例	79
3.9.1 基于分形理论的自适应中位值滤噪方法研究实例	79
3.9.2 基于光谱信息计算解析的液相色谱滤噪方法研究实例	81
3.10 小波信号压缩	84
参考文献	85
第4章 分析谱图的谱峰辨识方法	86
4.1 谱峰辨识技术概要	86
4.2 谱峰检测方法	87
4.2.1 幅值检峰法	87
4.2.2 一阶导数检峰法	87
4.2.3 二阶导数检峰法	88
4.2.4 分形维检峰法	89
4.3 谱峰识别方法	90
4.3.1 基线漂移法	91
4.3.2 谱峰间距测定法	91
4.3.3 峰高比综合判别法	92
4.4 仿人智能辨识谱峰方法	92
4.5 谱峰自适应辨识方法	93
第5章 多元校正与复杂分析数据解析	95
5.1 多元校正与复杂分析数据解析概述	95
5.2 多元校正	97
5.2.1 直接校正	97
5.2.2 间接校正	100
5.2.3 非线性多元校正	107
5.3 近红外光谱数据的多元校正	114
5.3.1 近红外光谱预处理方法	115
5.3.2 波段选择	116
5.3.3 多元校正建模方法	116
5.4 复杂分析数据解析	125
5.4.1 多元分辨	126
5.4.2 应用实例	129
参考文献	136
第6章 化学与生物模式信息处理	137
6.1 数据预处理方法	137
6.2 常用的模式特征提取方法	138
6.2.1 主成分分析法	139
6.2.2 偏最小二乘法	141
6.2.3 核主成分分析法	142
6.3 复杂化学与生物模式的分步特征提取法	143

6.3.1	复杂化学与生物模式特征的分步提取方法原理	143
6.3.2	特征矢量的分类能力评价指标	145
6.3.3	复杂化学与生物模式特征的分步提取方法应用实例	146
6.4	最优分类特征提取方法	148
6.4.1	最优分类特征提取方法原理	148
6.4.2	最优分类特征提取方法应用实例	150
6.5	聚类分析法	151
6.5.1	相似性测度	152
6.5.2	系统聚类法	153
6.5.3	动态聚类法	155
6.6	判别函数法	155
6.6.1	判别函数	156
6.6.2	参数学习	157
6.6.3	多分类问题	159
6.7	近邻法	160
6.7.1	最近邻法	160
6.7.2	k 近邻法	161
6.8	神经网络模式分类法	162
6.8.1	神经网络基本原理	162
6.8.2	感知器	164
6.8.3	误差反传神经网络	164
6.8.4	RBF 网络	167
6.9	支持向量机	168
6.9.1	支持向量机基本原理	168
6.9.2	多类支持向量机	170
6.9.3	支持向量机应用实例	171
6.10	模糊模式识别方法	174
6.10.1	模糊数学的基本知识	175
6.10.2	模糊模式识别的一般过程	175
6.10.3	模糊聚类方法	177
6.10.4	模糊神经网络	178
6.10.5	应用实例	180
	参考文献	181

第 7 章	化学指纹图谱计算处理	182
7.1	化学指纹图谱计算处理方法简介	182
7.1.1	指纹图谱相似性计算原理	183
7.1.2	模式分类计算原理	184
7.2	化学指纹图谱测量参数的选择	184
7.3	化学指纹图谱配准方法	188

7.4 化学指纹图谱相似性测度的比较	192
7.5 化学指纹图谱模式分类计算方法	195
7.5.1 基于小波基分形参量的化学指纹图谱计算方法	196
7.5.2 基于 Fisher 因子的化学指纹图谱模式分类方法	199
7.5.3 化学指纹图谱类别相似性计算方法	202
参考文献	205
第 8 章 分析信息智能管理	206
8.1 实验室信息管理系统	206
8.1.1 LIMS 的定义和范畴	207
8.1.2 LIMS 的发展过程	207
8.1.3 LIMS 的工作流程	209
8.1.4 LIMS 的功能	209
8.1.5 LIMS 的实施过程	212
8.1.6 LIMS 的标准和认证体系	214
8.2 电子实验记录本	214
8.2.1 电子实验记录本的概念	214
8.2.2 电子实验记录本的功能	215
8.3 中药分析信息智能管理	216
8.3.1 中药化学信息数据库	217
8.3.2 中药化合物数据库	218
8.4 中药指纹图谱数据库管理系统	218
8.4.1 系统功能	218
8.4.2 应用实例	222
8.5 数字中药信息系统	223
8.5.1 数字中药信息系统结构	223
8.5.2 数字中药信息系统功能	224
8.6 中药分析信息的数据挖掘	226
8.6.1 基于因果关系发现的中药有效组分辨识方法	227
8.6.2 中药材关键药效成分辨识方法研究实例	228
8.6.3 中药复方关键药效成分辨识方法研究实例	232
8.6.4 中药复方有效组分配伍优化方法研究实例	234
参考文献	235
第 9 章 生物芯片分析信息处理	237
9.1 生物芯片概述	237
9.2 生物芯片分析过程	238
9.3 芯片数据获取	239
9.3.1 芯片图像处理	239
9.3.2 归一化方法	240
9.3.3 芯片数据形式	242

9.4 芯片数据分析	242
9.4.1 差异分析	242
9.4.2 分类分析	243
9.4.3 聚类分析	244
9.4.4 网络分析	250
9.4.5 可视化方法	252
9.5 芯片分析信息处理发展趋势	255
9.5.1 数据质量评价与归一化	255
9.5.2 差异分析	255
9.5.3 测度选择	255
9.5.4 基因相互关系发现	255
9.5.5 聚类结果验证	255
参考文献	256
第 10 章 组学分析信息处理	257
10.1 组学及组学分析信息处理技术概述	257
10.2 蛋白质组分析信息处理	259
10.2.1 蛋白质组分析信息处理简述	259
10.2.2 基于形态特征的双向电泳图像处理方法	261
10.3 代谢组分析信息处理	267
10.3.1 代谢组信息处理技术简述	267
10.3.2 肾阳虚模型大鼠代谢模式分析	268
10.3.3 乳腺癌代谢物组模式特征发现	270
10.3.4 肺癌患者尿液代谢组分析	274
10.4 展望	276
10.4.1 组学分析信息处理的应用前景	276
10.4.2 组学分析信息学	277
参考文献	278
附录	279
1 线性空间预备知识	279
1.1 基本概念及定义	279
1.2 矩阵代数	279
1.3 矩阵和向量的微积分	280
2 点盒分形维	282
2.1 分形的定义	282
2.2 分形维数定义	283
2.3 点盒维数定义	283

第 1 章 绪 论

生物科技和信息科技是当今世界发展最快和影响最大的学科。迈入 21 世纪的生物医药受其自身发展内在逻辑力量的强力驱动,正从宏观和微观两个方向不断深入拓展。生物、化学和信息三大学科汇聚,交叉融和加速,思路不断突破,学科界限日趋模糊,逐渐形成新的科技体系,进入了前所未有的多学科集群创新时代,这将彻底改变生物科技及医药研究领域的现有竞争格局。当代生物科技及医药研究的领先权,在很大程度上取决于分析测试方法和信息技术手段的先进程度。因此,分析科学、分析信息处理技术理所当然地成为生物医药等领域的重要研究前沿和技术支撑点。

1.1 分析科学发展前沿若干问题

人类对探索大自然奥秘的巨大热情和发现自然界规律的无穷欲望使分析科学与技术始终面临着严峻的研究课题和使命:

运用一切可能的方法和技术手段去分析测试、研究探索、揭示阐明待测物质体系中所需要量测和辨识的一切化学属性及其属性间的联系。

进入 21 世纪以来,各学科所提出的物质体系分析测试和表征问题变得繁杂多样,既有复杂体系成分分析问题,又有不确定性化学模式识别问题;既要检测微乎其微物质的组成和含量,又要对其复杂的结构或形态进行分析;而且还需对复杂物质体系的生物效应及其动态变化过程等进行有效的监测、示踪及时空分辨等。特别是生物体等高度复杂体系不仅物质组成和形态复杂,而且各组分含量极微、性质各异;既存在化学和物理变化,又伴有生物变化;既有线性变化,也有非线性变化;同时,复杂体系中各组分的相互作用交错盘杂,呈现网络状交互关系,这就对分析科学和技术提出了创新发展的巨大需求。

随着物理学、微电子学、计算机科学、生物学以及仪器工程学的发展,各种新型分析仪器相继问世,昔日以化学分析为主的经典分析化学已发展成为一门以众多仪器分析(包括:色谱、光谱、质谱、核磁共振及各类仪器联用等)为主的现代分析化学,其主要任务是研究如何获取待测物质体系的化学成分、数量、结构、分布、各组分间相互作用以及时变信息等。计算机应用技术与分析仪器的结合导致了化学量测数据获取方法的重大进步,这一发展不仅实现了分析仪器数据的自动采集、传递和储存,而且使分析仪器的自动化操作成为现实,从计算机数据采集系统可得到大规模分析测试数据。然而,在多数情形下,峰高、峰位或峰面积等数据并不等同于有价值的信息。由此带来的尖锐问题是如何大批量地快速处理和解析这些仪器分析数据,这就促使人们开始努力研究各类仪器分析数据的处理方法。在此情况下,旨在将数学、统计学与计算机技术应用于化学的分支学科——化学计量学应运而生。作为 20 世纪 80 年代蓬勃兴起的新技术,它运用数学、统计学及计算机等工具,设计或选择化学量测的最优方法,并通过解析化学量测数据,试图最大限度地获取物质体系的化学相关信息。然而,在生命科技、生物医药、环境检测、新材料创制、工业过程质量控制等领域不

断提出的越来越高的各种分析要求面前,与现有的分析化学计量学方法已不相适应,出现了学科聚汇、整合创新的重大契机。有学者认为,分析化学学科正经历着巨大的变革,处在一个高速发展时期,融入了尖端信息技术的新型分析工具和分析方法将蜂拥而至。也有学者认为,分析化学的发展已使之成为一门创造和应用新概念、新原理和仪器策略来测量化学物质体系及其组分的学科,简言之,分析化学已成为一门化学量测科学。还有学者认为,分析化学本身正在急剧分化,而与其他学科产生交叉,展现出多学科融合创新发展的态势;生物学、信息科学和计算机科学的引入,使分析化学进入了一个崭新的境界。

现代分析化学在科学研究和技术进步中所起的重要作用促使和逼迫它自身必须产生新的实质性飞跃以适应未来科技创新发展的需要。人们已经认识到,分析化学不仅承担获取化学量测数据的任务,更重要的是需要解决从大量分析数据中提取有价值信息的问题,学科研究理念起了根本性转变。处于历史性跨越发展时代的分析化学家们不但希望拥有更多的化学量测手段,而且期待创新发展的结果将能为解决生命科学等许多学科所提出的复杂物质体系辨识难题提供更强有力的技术工具。

科学现实告诉我们,许多物质体系辨识问题的复杂性不仅大大超出当代分析化学所能破解的限域,而且已跨出了分析化学、化学计量学及化学信息学目前研究的范畴,这就给研究者们提出了严峻的挑战,同时创造了产生新的学科生长点的历史机遇。以拓展的学科视野看,分析化学、统计学及生命科学正与飞速发展中的信息科学、计算数学、电子科学和计算机科学等学科汇聚交融,伴随着这些学科群的不断渗透、交叉、分化、融合和裂变,在以惊人速度发展的高新技术冲击下,探索者们将突破化学计量学或化学信息学现有的学科框架,孕育铸就新的学科,研究发现化学及生物分子信息采集、变换、传递、获取、储存、辨析及解读等信息过程的一般规律,这是一个庞大而充满创新活力的研究领域。

1.2 涉及的若干概念、界定及术语定义

严格地讲,信号(signal)、数据(data)与信息(information)是三个各具特定含义的术语。信号是某一种被考察的过程所产生的表征或表述某种含意(meaning)的物理量,即传递信息的函数;数据是信号的某种记录形态,亦是信息的载体;信息则是含有某种语义解释(semantic interpretation)的数据,是数据的抽象。同样内容的信息可以用多种不同形式的信号或数据来表达。显然,三者间具有层次关系。以仪器分析为例,在最基本的层次上,色谱和光谱等分析谱图均可视作由检测器输出的化学响应信号,若将此分析信号记录下来即形成一组原始量测数据,故常将分析信号与原始量测数据视为等效,此步称为分析信号采集;在高一层次上,若对信号或原始量测数据作某种变换或其他处理,将形成新的数据群体,此步称为分析信号处理;在更高层次上,若再对数据群体作某种处理并赋予一定的语义解释,则可获取化学定性定量信息,此步称为分析数据处理;在最高一层,可对化学定性定量信息(含有某种语义解释的数据集合)再作某种处理并赋予特定的语义解释,则可获取被测物质体系的某些隐含化学信息,此步称为分析信息处理。因此,信号处理、数据处理与信息处理三者间的区别并不是在处理方法上,而是在层次性质方面存在差异。

由于信号、数据及信息处理方法的作用对象均为具有某种层次性质的数据形态,故在不至于混淆的前提下,本书有时不加区别地将三者统称为分析数据处理方法。

分析数据处理方法分数值计算和非数值计算两大类。基于数值计算的数据处理方法可视

为作用于数值型数据的一系列数学变换群,它通过对数据集合作各种变换,既能剔除数据集中不合理部分,也可用于揭示各数据集之间蕴含的映射关系,其作用对象为数值变量型数据(numerical variables)。基于非数值计算的数据处理方法则依靠符号运算(symbolic operation)、推理演绎(inference)及搜索(searching)实现问题求解,其作用对象主要为逻辑变量型数据(logic variables)。

基于数值计算的数据处理方法分确定性数值计算法和不确定性数值计算法两大类。不确定性数值计算法又分为统计计算处理法和模糊计算处理法。统计计算处理法常用来处理含噪数据,以便去除数据集中的噪声成分,减少噪声导致的随机误差并估计随机误差对计算结果的影响;模糊计算处理法则被用于计算处理模糊变量型数据(fuzzy variables)。

统计计算处理方法可分为描述性统计(descriptive statistics)和推理性统计(inferential statistics)两种。描述性统计法主要用于数据的收集、整理、表达及描述;推理性统计则用来对所考察的数据作出解释、推断和预测,并为决策提供依据。

信号、数据及信息处理技术有硬件法和软件法之分。软件方法代表了信号、数据及信息处理技术的主流方向,本书所述的各类处理技术均属软件法范畴。

分析信号有模拟信号(analog signal)与数字信号(digital signal)两种,时间和幅值均离散化的信号称为数字信号。相应地,化学量测数据分为模拟式数据(analog data)与数字式数据(digital data)两大类。在计算机技术广泛应用的今天,数字信号或数字式数据已占主导地位,本书所论及的信号或数据除特别说明外均指数字信号或数字式数据。

在仪器分析中,实际所得的化学量测数据往往是真实分析信号和各种各样的噪声或干扰等分量相叠加的数据序列。依据统计学的观点,从受噪声污染的化学量测数据中辨取真实分析信号就是对信号的未知参数赋予具有一定统计意义的数值,这一过程称为估计(estimation)。在时刻 t ,当估算分析信号在时刻 t_1 的估计值时,若 $t_1 = t$,该估计问题称为滤波问题;若 $t_1 > t$,称之为预报问题;如果 $t_1 < t$,则称为平滑或内插问题。估计量(包括滤波、平滑或预报)是化学量测数据集的一个函数,这个函数称为估计器;对于一组特定的量测数据,该函数的取值称为估计值。

信号、数据及信息处理过程的计算步骤均称之为算法(algorithms)。从广义上讲,算法定义了作用于数据集的过程以及对数据集进行各种变换或揭示各数据集间映射关系的运算策略。通常,信号和数据处理过程涉及的是具有普遍性意义的算法,而信息处理过程涉及的是针对某一题域的特定算法。在分析信息获取及处理技术方面,算法研究占有极为重要的位置,因而也是最富有活力和创造性的研究领域。

1.3 分析信息获取问题

前已论及,处于历史性跨越发展时代的分析化学家们已深刻认识到,获取化学量测数据与从分析数据中提取有价值信息同样重要;肩负分析学科变革使命的化学计量学专家们也已发现,在越来越难的复杂物质体系辨识问题面前,现有的分析化学计量学方法必须创新发展,建立学科的新基点。探索者们已在努力拓展学科发展空间和推动科研理念蜕变,化学信息(包括生物分子信息等)获取与信息分析技术研究正受到来自生命科学等领域巨大需求的强力驱动。

从本质上讲,分析化学是一门获取、辨析和处理化学信息的大学科,是化学、生物学、药学、农学、环境学、考古学等众多学科的科学数据源。它不仅属于化学学科研究范畴,大

量研究内容涉及信息科学领域的许多相关学科。因此,理应在信息科学的范畴内研究分析化学的创新问题,着力构建化学及生物分子信息获取、辨析及解读等方法学体系。化学信息学的创立及“信息化学”概念的新近提出,正是反映了当今前沿科技发展中的现实需求和分析化学学科的未来趋向。

现代分析化学最基本的研究内容是获取物质体系的定性定量分析信息。众所周知,分析仪器是一类在光、电、磁、热、声等物理原理基础上综合运用生物学、电子学及计算机科学等学科成就所发明的对物质作全面与纵深分析的化学量测设备,其共有特性是能够检测提供蕴含被测样本化学信息的谱图。无论仪器原理差异如何,它们均以电信号形式输出谱图。用数学的观点看,谱图是一组化学量测数据的几何表达。因此,从化学谱图中获取分析信息的工作可归结为对仪器分析数据(即化学量测数据)进行计算处理和解析,其主要包含以下几个方面:

- ① 滤除噪声;
- ② 背景校正与基线修正;
- ③ 谱峰检测与识别;
- ④ 重叠峰解析与分解;
- ⑤ 多元校正与多元分辨;
- ⑥ 组分定量计算;
- ⑦ 结构解析与谱图解释;
- ⑧ 谱图库检索及数据管理。

此外,在分析信息获取方法研究中一项不容忽视的任务是如何优化分析仪器条件,通过选择最佳的仪器操作参数,利用现有仪器获取尽可能多的定性定量分析信息。分析仪器的多机联用是获取更多定性定量分析信息的另一有效途径。仪器联用分析水平的提高不仅取决于联用仪器间的匹配优化,而且依赖于多维分析数据处理方法的进步,这是最大限度地获取分析信息的关键所在。在分析信息获取方法研究中,重点研究方向包括以下几个方面:

- ① 新型分析仪器的研制,特别是发展高灵敏度、高选择性、高通量、在线、原位分析检测技术等;
- ② 分析仪器操作参数优选技术,包括用于分析条件最佳选择的计算机仿真及优化算法;
- ③ 用于各类谱峰识别、分解及定量计算的高效及精准计算方法。
- ④ 用于化合物结构解析智能系统的知识表达方法、知识获取机制、自学方法、问题求解策略及推理构造方法等;
- ⑤ 用于存储、检索及识别谱图的智能数据库构造方法;
- ⑥ 用于工业过程在线分析数据的实时处理方法;
- ⑦ 分析仪器联用及多维分析数据处理方法;
- ⑧ 高维非线性分析数据处理方法,用于解决复杂体系定性定量分析问题的智能集成计算方法等。

上述方向的研究工作,传统上主要由分析化学家承担,只有少量其他领域的专家涉足其中。一般而言,分析化学家们往往习惯于使用本领域的观点和方法来看待和研究分析信息获取过程中的分析仪器数据处理问题,这就容易受到某些限制。因此,多学科集群创新是分析信息获取技术未来发展的突破口。

1.4 生物医药分析技术发展趋势

由于生物医药分析的对象是物质组成复杂的药品及高度复杂的生命体等,故随着生命科学与医药研究的迅猛发展,会对分析测试和表征技术不断地提出新的超越需求和局限挑战。目前,生物医药领域面临的分析难题包括:

- ① 生物体内极微量物质分析;
- ② 超多组分复杂混合物分析;
- ③ 天然药物等复杂药品物质体系的组成及结构分析;
- ④ 药品安全、食品安全及生物安全研究中的分析测试问题;
- ⑤ 药物与生命体相互作用及其过程行为的动态分析;
- ⑥ 生命体的生物分子及其相互作用分析;
- ⑦ 生理活动中分子信息的获取与信息分析;
- ⑧ 机体与药物相互作用的系统性行为的多维、多尺度和多参数分析等。

当代生命科学研究的范围已从局部扩大到整体,研究模式从“分解还原”趋向“系统整合”。基因组学、蛋白质组学、代谢组学以及相互作用组学等研究的兴起正是反映了这种大趋势。在组学分析研究领域,探索者们主要着力破解两方面难题:一是建立对各类生物分子进行全面、准确和高通量分析的方法;二是建立海量数据处理算法及知识发现方法。当前,生物芯片已经成功地应用于基因表达谱研究,但是用于蛋白质、糖类、脂质和细胞等分析仍存在较大困难。此外,代谢组学分析方法尚待突破。由于机体的内源性代谢产物类型多样,含量差异悬殊,常规分析方法难以满足要求;同时,如何对代谢物组分析数据进行处理,从中提取出有价值的隐含信息,发现和获知这些数据集合所蕴含的科学规律,用于诊断疾病、揭示药物的生物效应或评价药物毒性等也是迫切需要解决的难题。

纵观当今生物医药分析技术领域发展态势,深刻的技术变革正在发生:从简单体系分析发展到复杂体系分析,从低通量分析跃迁到高通量分析,从静态分析拓展到动态分析,从生物体外分析延拓到生物体内分析,从物质化学组成分析拓宽到生物效应分析,从单一仪器分析进入到多源信息整合分析时代。复杂样品系统分离分析方法,高灵敏、高选择性、在体、原位和瞬时分析方法,集成式多维多元分析方法,“组学”分析方法,以色谱-波谱联用、多维色谱(电泳)、生物芯片为代表的新型分析设备,医药分析信息处理系统等已成为生物医药分析技术研究的前沿。

1.5 药物信息学与生物医药分析

药物信息学(pharmacoinformatics)是药学与计算科学等各相关学科交叉融合而产生的新兴边缘学科,系运用当代信息技术及计算分析方法收集、管理、辨析和处理药学研究数据,获取相关信息,进而解释药物实验现象、发现新的知识及规律、指导药物研究、加速新药创制以及确保合理用药的科学。药物信息学的出现与生物信息学和化学信息学的迅猛发展密不可分,它既全面汲取了生物信息学和化学信息学的技术精髓,又顺应药物学科自身的重大需求而带有鲜明的药学特点,具有更为广阔的研究范围及发展前景,影响着药物科学、生命科技及生物医药产业的技术走向,是国际医药巨头争相攀登的技术制高点,也是未来生物