计算机化学

许　禄　　胡昌玉
(中国科学院长春应用化学研究所　长春 130022)

计算机化学是将计算机科学、数学应用于化学的一门新兴的交叉学科,是化学领域的一个重要分支。

计算机化学的英文叫法有多种,如 Computers & Chemistry、Computers in Chemistry 及 Computers on Chemistry。有时文献中亦会出现 Computer Chemistry,但应用较少。计算化学(Computational Chemistry)通常指分子力学及量子化学计算等,与计算机化学有较大区别。

计算机与化学的联姻始于 60 年代。其首先应用领域是分析化学。因为分析化学的最本征特征是借助于诸种手段收集数据及其数据处理。到了 70 年代,计算机化学得以突飞猛进的发展,几乎在化学的每一分支领域都结满了丰硕的成果。

当今的化学几乎无处不用计算机。计算机(包括数学)已是化学的重要工具,同时计算机化学作为一个学科分支也在迅速发展。本文拟就如下几个方面作一简单介绍。

一、 数据库技术[1]

数据库是计算机科学领域中 70 年代出现的新技术。化学中的许多数据库正是在 70 年代历经了由起步、发展,直至成熟的过程。其中,最具代表性的是用于化合物结构解析的谱图数据库。目前,几乎所有的大型分析测试仪器均带有数据库及其检索系统。

各种谱学手段的广泛应用对当代有机化学的发展起到了很大促进作用,因为这些物理方法和手段使人们能较精确地了解化合物的结构。但是,谱图的解释是一较为繁琐,极为费时的工作。然而,随着计算机技术的发展极大地推进了这一领域的革新。

计算机辅助谱图解析方法可粗略地分为两大类:直接谱图库手段,即谱图检索;间接谱图库手段,包括波谱模拟、模式识别和人工智能。目前,应用最广泛的是谱图库检索。此处顺便提及:数据库,英文一般用 database 或 databank 表示,而数据库检索却常用 library searching 一词。所谓谱图库,目前用于结构解析的主要是指质谱、核磁谱和红外光谱。

1.　质谱谱图库

随着计算机的发展,首先引入这一新技术的是质谱谱图的处理。因为,第一,质谱以分立峰的形式出现,其质量峰和强度本身已数字化形式表示,便于计算机处理;第二,如 GC-MS 联用系统,一小时可以产生几十个,甚至上百个质谱图,平均每谱为 75 个峰,由此,每小时所产生的峰数达 6000~ 7000 个。这些背景给质谱的计算机处理提供了条件和动力。

对于质谱谱图库,可以追溯到 40 年代美国石油研究院(API)的工作,当时他们已经进行了谱图的收集和出版。近 30 年来,世界上许多实验室开展了谱图的收集工作,涌现出许多质谱检索系统,如 MIT/KB 系统、MSSS 系统、PBM 系统、Masslib 系统等,不过,当今在国际上影

响最大的是 NIH/EPA/NBS 和 wiley Ms Data 两个系统。在国内,中国科学院大连化学物理研究所也开展了这方面的工作,他们的系统为 ASES/MS。

2. 碳-13NMR 谱图库

C-13NMR 波谱是有机化合物结构解析中三大谱图手段之一。该手段问世于 70 年代初,但发展很快。随着这种技术的发展,国际上先后涌现出许多 C-13 谱图库信息系统。其中比较主要的有:德国 BASF 公司的 C-13 谱图库信息系统,美国 NIH/EPA 系统,法国的 DARC 系统等。

在国内,中国科学院长春应用化学研究所于 80 年代初已经开展了此方面课题的研究,所建立的 CIAC-C-13 谱图库系统于 1985 年投入正式运转,目前拥有谱图 20,000 张。其中的谱图不仅涉及一般的化合物,同时包括中草药及天然化合物类。该系统的主要特色是具有较强的相似检索功能,即未知谱图不在库中,其检索可以给出与之相似的化合物,在实际应用中收到良好的效果。

3. 红外谱图库[1]

从 50 年代初到 70 年代中期,美国 ASTM(The American Society of Testing and Materials)编辑了 150,000 张机读红外光谱,这是世界上最早的,同时也是最大的红外谱图数据库。另外,美国 Sadtler 光谱实验室(现属 Bio-Tad 公司)在 80 年代中期已收录 94,000 张 IR 谱图。目前,国际上主要检索系统为 SPIR、IRGO、IRIS、COSMOSS 等。

在国内,中国科学院上海有机化学研究所于 80 年代初开始组建 IR 谱图库 CISOC-IR 系统,目前该系统拥有谱图库约 100,000 张。

另外,应该提及的是中国科学院化学冶金研究所的研究工作。该所从 80 年代开始先后组建了化工冶金中的许多数据库系统。

二、 有机化合物结构自动解析

该类研究属于人工智能的范畴。人工智能包括的范围很广,如定理证明、语音识别、对奕及专家系统等。对于化学领域,尤以专家系统研究的为最多。所谓专家系统即在规则(常称为“知识库”)的基础上,模拟专家演绎推理的过程,以得到专家水平的应答。在化学中,除结构解析以外,其它专家系统如分离科学、实验方案的最优设计、工业生产的流程控制及计算机辅助合成(见后)等。

世界上第一个专家系统诞生于化学领域,即美国斯坦福大学建造的 DENDRAL 系统[2]。该系统利用低分辨质谱和核磁共振波谱来进行有机化合物的结构解析。这一系统的建造成功对整个人工智能领域产生了重要影响。

早年,专家系统主要建造在中、小型机以上的计算机上。后来出现工作站,但由于价格的昂贵使其应用受到限制。到了 80 年代中期,微机发展极为迅速。目前,世界上至少有 60% 的专家系统建立在微型计算机上。

作为软件,原则上任何一种计算机语言均可作为专家系统设计工具。但是,由于一般的高级语言字符处理能力较差,所以在选用上应首先选用人工智能语言,如 LISP 和 PROLOG。

几十年来,在结构解析领域中涌现出一大批专家系统,除 DENDRAL 外,目前比较有影响的系统为 CHEMICS(日本)[3]、CASE(美国)[4]、PAIRS(美国)[5]等。在国内,从 80 年代初在作者的实验室中就开始了计算机自动结构解析的研究工作。并先后建造了含碳、氢、氧有机化合物结构阐明专家系统及含多种杂原子的结构阐明专家系统。

结构解析专家系统工作的逻辑过程为:

(1) 由实验数据(如质谱、红外光谱和核磁共振谱等) 或者化学信息(如分子式) 出发, 在知识库(如子结构-子光谱相关规则) 作用下获得化合物中可能含有的结构片断集。

(2) 在结构片断集的基础上, 利用知识库(如诸多约束条件), 经结构产生器(进行结构异构体穷举生成的程序部分) 来作整体结构的对接, 所生成的异构体常称为候选化合物。

(3) 在波谱模拟、碳-13 谱峰信息、分子张力能计算、模式识别及人机交换信息作用下, 进行候选化合物的验证。

其中, 核心(也是难点) 部分是结构产生器。对结构产生器的要求为: ①具有穷举性; ②具有无冗余性; ③具有高效性。对于第 ①、②两点的要求是显然的, 如 C_6H_6, 其异构体数为 217 个, 其生成结果不能少一个, 如 216 个, 也不能多出一个, 如 218 个。同时作到穷举和无冗余是不易的, 然而作到高效不仅重要, 且更难。

几十年来, 世界上有众多的实验室开展了有关课题的研究, 但是迄今, 卓有成效的, 可以投入实际运用的系统并不多, 其要害在于没有设计出高效的结构产生器。对结构产生器具有高效性要求的理由有二: ①当分子大时, 其异构体数目庞大。如 $C_{20}H_{42}$的饱和链烷烃异构体数为 366319 个, 当碳原子数增加 1 时, 异构体数大约增加 3 倍。②在整体结构的对接过程中, 无效对接数目庞大。如 2 个 CH_3—和 18 个 —CH_2—, 最大无效组合数达 $2 \times 18!$, 如此浩繁的运算量, 一台普通计算机从安装开始直到报废也无法完成。

在作者的实验室中, 为了验证我们的系统 ESESOC[6~9] 的穷举和非冗余性, 曾就如下化合物系列进行了异构体的穷举生成:

(1) 在图论中, 可根据 Polya 定理计算烷烃 C_nH_{2n+2}、醇和醚 $C_nH_{2n+2}O$ 及醛和酮 $C_nH_{2n}O$ 等各类化合物异构体数目。我们曾运用 ESESOC 由分子式 $C_{21}H_{44}$ 生成了 910726 个异构体, 由分子式 $C_{12}H_{26}O$ 生成了 3057 个醇类异构体, 这与 Polya 定理计算出来的结果完全一致(Polya 定理计算的伯醇数 1238, 仲醇数 1188, 叔醇数 631 之和为 1238+ 1188+ 631= 3057)。

(2) ESESOC 系统就如下系列与 DEDRAL 系统的生成结果进行了比较:

①C_5H_n, $n = 0, 2, 4, 6, 8, 10, 12$, 如 C_5H_6 有 40 个异构体;

②C_6H_n, $n = 0, 2, 4, 6, 8, 10, 12, 14$, 如 C_6H_6 有 217 个异构体;

③含多种杂原子如 O、N、S 的复杂体系, 如 $C_3H_8N_2O$ 有 527 个含三价 N 的异构体。

对于这些化合物, 由 ESESOC 生成异构体与 DEDRAL 的生成结果完全一致。DENDRAL 的算法在数学上已被证明是正确的, 从而佐证了 ESESOC 算法(穷举、非冗余) 的正确性。

另外, 在 ESESOC 系统中, 施加了光谱实测数据的约束条件及最大生成环数和用拓扑等价性分析法以尽早消除冗余的对接, 从而可使 ESESOC 非常高效地运行。

三、 计算机辅助化合物合成[10,11]

有机化合物的合成最早开始于 1895 年, 距今已非常久远。但是, 计算机辅助合成还是近几十年的事。

计算机辅助合成系统在解决问题中, 要用到人工智能技术及专家系统的知识, 即计算机辅助合成系统为一专家系统。

1969 年, 美国哈佛大学的 Corey 和 Wipke 首先报道了他们的系统。之后, 其他系统相继问世。现在, 国际上该类系统已用于工业之中, 特别是药物工业其应用尤为普遍。

计算机辅助合成大体上可分为两种类型:

(1) 经验型 该种类型主要由两部分组成: 知识库, 存贮大量的反应、反应过程、条件和产

物等; 推理机, 即在知识库基础上进行演绎推理的逻辑部分。在该领域中主要系统有: LHASA(Logic and Heuristics Applied to Synthetic Analysis)、SYNCHEM (Synthetic Chemistry) 、SECS(Simulation and E aluation of Chemical Synthesis)、CASP(Computer- Aided Synthesis Program) 和 SOPHIA(System for Organic Reaction Prediction by Heuristic Approach) 等。

(2) 理论型　这种类型不是用已知的、大量的有关反应信息, 而是应用抽象的原子和价键电子模型, 把化学反应在数学上进行公式化。其代表系统有 EROS (Elaboration of Reactions of Organic Synthesis) 和 CAMEO。

目前, 研究得最多、最普遍、最有实用价值的是经验型。后一种类型目前仍处于实验室阶段, 远不够成熟。但其最大优点是可以发现新的反应, 在将来应用中存在较大潜力。下面仅对经验型的工作原理作进一步介绍。

在经验型系统中通常采用逆向推理。即给定一目标分子(即所需产物) 之后, 由计算机自动识别, 或者由化学工作者指定分子子结构及分子断裂的类型。接着, 程序生成目标分子的反应物(称为“前体”) 。但在此步之先要对目标分子的重要特征进行识别和提取, 如环、环之封闭点、官能团、官能团彼此间关系及对称性等, 并根据这些特征选择目标化合物前体的策略。前体化合物可能有多个, 对于每一前体应赋值一优先权。演绎推理是按照优先权的高低进行的, 一但发现某条路径不合理或不可取, 则中断此条路径。对合理的路径继续前推, 由此, 以目标分子为根, 逐渐生长起一棵逆向推理合成树。图 1 为 LHASA 系统显示合成树的一例。图中, 长方框之外部分为程序命令。如选“GET-FAMILY”, 并将光标定位到屏幕结构 17, 则它的前体 18、19 和 20 则被显示出来。若选“GET-LINEAGE”, 并将光标定位到 18, 则从 18 开始, 程序将显示整个合成路径上的所有结点 17、16、15 和 14, 直至目标分子为根结点 2。

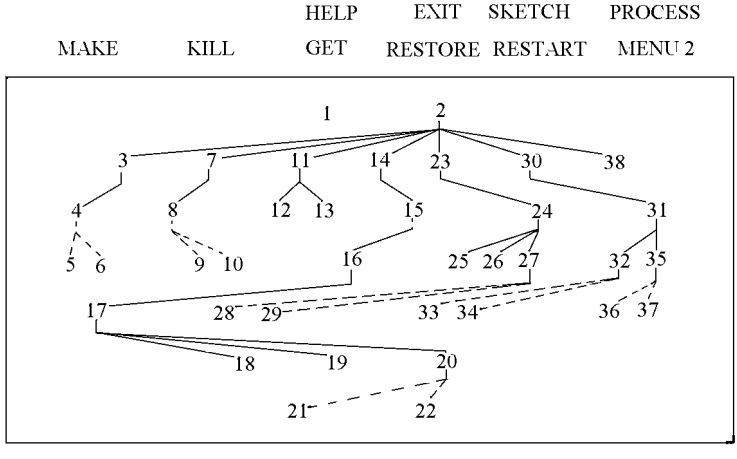
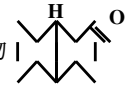


图 **1**　化合物  的 **LHASA** 合成树

四、 分子设计

计算机辅助分子设计是计算机化学的前沿, 目前已用于药物分子设计, 蛋白质、核酸等生物大分子设计及材料科学, 如高分子材料、无机材料和催化剂设计等。下边将侧重以药物分子

设计为例进行介绍。

为分子设计近年来发展了很多种方法,其中,开展得尤为广泛的是定量结构-活性/性质相关性(QSAR/QSPR)研究[12, 13]。这种方法的要点是由分子式结构出发来构造某种数学模型,然后运用这种模型去预测未知化合物的活性/性质,从而为新分子的设计提供理论依据。

60年代, Hansch 根据药物的分配系数、电参数和立体参数的改变,直接影响药物的转运及与受体作用的自由能的一般原理,提出了线性自由能的相关方法。与此同时,还有其他一些方法的提出,如 Free-Wilson 加和模型、模式识别法及量子化学法等。不过几十年来,得到广泛应用的主要是 Hansch 方法。我们知道,药物分子或毒物与受体间相互作用是在三维空间进行的,定量地描述三维结构与生物活性的关系,需要对化合物(毒物或药物)分子以及受体的结构有较精确的表达。Hansch 方法中所用参数虽然涉及到 Taft 立体常数和 Hammett 电性常数,但是对于分子的处理基本是二维结构,因而当遇到化合物构型构象时则难以得到满意的结果。

80年代初以来,在定量构效关系研究中,陆续出现了几种考虑生物活性分子与受体结合时三维结构的研究方法,统称为三维(three dimensional)定量构效关系(3D-QSAR)。3D-QSAR 与建立在超热力学参数基础上的 Hansch 途径最大的不同就在于它们考虑生物活性分子和三维构象性质,在 QSAR 中引入与生物活性分子三维结构信息有关的量作为变元,因而能更精确地反映生物活性分子与受体作用的真实因素,更深刻地揭示药物/毒物-受体间相互作用的机理。为此,世界上诸多软件公司推出他们的软件系统。这些公司如 Biosym(美国)、MSI(美国, Biosym 已被 MSI 收购)、Biodesign(美国)、Tripos(美国)、Oxford Molecule(英国)、MAD(法国)和分子(日本)等, Tripos 的 CoMFA(Comparati e Molecular Field Analysis)[14] 是目前几种典型的 3D-QSAR 方法之一。

定量构效关系研究涉及到有机化学、物理化学、物理有机化学、分析化学、药理学、药剂学、生理学、统计学、信息学及电子计算机等,因而要求多方面的基础和理论。归结起来,进行构效关系研究需解决如下问题:

(1) 变量的提取　由化合物结构来提取特征(即变量)是构造数学模型的关键环节。用于 QSAR 研究的主要变量类型为: ①拓扑的,如分子中原子类型、键的类型及计数、拓扑指数等。②几何的,如键长、键角、分子体积和形状等等。③电的,如电负性、局域电荷(partial charge),由量子理论所得的 HOMO(最高占有轨道)、LUMO(最低空轨道)等。④物理化学的,如超热力学参数(见前)等。提取何种特征为宜,则以研究对象而定。

(2) 变量的压缩　由一化合物可以同时提取多种和多个变量,这些变量对于构造数学模型的重要性不是等同的,所以需剔除非显著性变量。变量过多,不仅计算量大,且对构造出较稳定的数学模型不利。用于变量压缩的方法有多种,如主成分分析、模拟退火、最优子集选择法等等,此处不拟赘述。

(3) 数学模型的构造　QSAR 中数学模型的构造是建立在实验基础上的。即首先由实验测得一组化合物的某种性质(如抗癌活性),然后运用前述所得变量则可建立用于未知物预报的数学方程。建立数学模型的最常用方法为回归分析、主成分回归(PCR)和偏最小二乘(PLS)方法。近年来,广为应用的另一种方法是神经网络法。

(4) 结果的表示　此处侧重介绍的是 3D-QSAR 的结果表示,因为 2D-QSAR 的结果是简单的和显然的。

以 CoMFA 为例。CoMFA 方法的基本假设是药物与生物大分子的相互作用主要不是靠

经典的共价键, 而是靠空间的立体效应和空间的静电效应。这种作用可借助 PLS 建立数学模型, 并以系数等高图(coefficient contour map) 进行三维显示(图 2)。图 2 为 N-4 甲氧苯基苯甲酰胺类化合物的静电势等高图。图中, 在蓝色区域, 取代官能团的正电性越强, 则对其抗炎活性越有利。相反, 在红色区域, 取代基团负电性越强对其活性越有利。显然, 由这种类型的三维图形显示, 对分子的修饰及新化合物的创制是极为有益的。

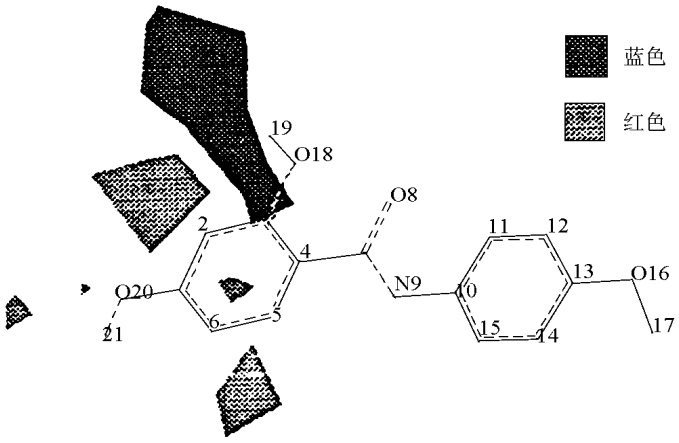


图 **2** 静电的三维图形显示

五、 化学计量学方法的研究及应用

化学计量学(Chemometrics) 是将数学、统计学应用于化学的边缘学科。它是数学与化学之间的一座桥梁。

数学是自然科学的语言, 它在化学中的地位和作用日益突出和重要。自 70 年代以来, 随着计算机技术的迅速普及, 数学和计算机科学在化学中应用日益广泛, 于是化学计量学的方法和内容得到充实和发展, 使化学计量学成为化学、生物化学、医学化学、环境化学及药物化学中信息处理的强有力手段。1974 年, 由美国的 Kowalski 和瑞典的 Wold 等发起, 在美国华盛顿大学成立了国际化学计量学学会, 开展了一系列学术交流活动, 推动了化学计量学的迅速发展。从 1982 年起, 在美国分析化学杂志(Anal. Chem.) 两年一度的评论中开辟了“ Chemometrics” 专题。一些年来, 国内国外都不断有化学计量学方面的专著问世[15]。

化学计量学是建立在多学科基础上的横向学科。反过来, 它在多种学科中的应用也在逐年迅速增加。1994 年的 Anal. Chem. 中化学计量学专题评论, 仅计算机检索(事实证明漏检很多), 有关文章已多达 20000 篇, 而 1996 年度又增至 25000 篇。化学计量学在化学学科的发展中起着越来越大的作用。化学计量学主要包括:

(1) 统计学(statistics): 在化学中统计学主要用于方法的评估、采样方案制订、检测限测试、误差分析、过程控制、实验室间结果比较及异常点确证等。

(2) 最优化(optimization): 所谓最优化即求取函数的极大值或极小值。常用方法如单纯形最优化、模拟退火、遗传算法及经典的梯度法等。其中模拟退火法及遗传算法, 方法较新, 尚有问题需要探讨; 运用这两种方法有可能达到全局最优。因而, 受到了化学家们的极大重视。

(3) 信号处理(signal processing): 属于该领域的研究主要为时间序列分析、数字滤波、平滑、退卷积、背景校正及图象分析等。信号处理的核心是如何增强有用信号, 因而数字滤波和平滑集中了该类课题中的大量研究工作。

(4) 分解(resolution): 所谓“ 分解” 是指重叠峰分解。在分析化学中, 重叠峰的存在是极为普遍的。在混合物情况下, 仅靠仪器本身的分辨率常常不够, 则必须借助化学计量学方法来辅助解析。为此, 近年来涌现出不少新的方法, 如移动窗口主成分分析(moving window PCA)、实时因子分析(real factor analysis)、渐进因子分析(evolving factor analysis) 及卡尔曼滤波等。

(5) 校正(calibration): 校正即为进行相关,即为构造数学模型。如前边所介绍的结构-性质相关分析中,数学模型的建立所用化学计量学方法即为校正。如稳健(robust)回归及经典的多元回归分析等。

(6) 参数测定(parameter estimation): 事实上,参数测定所用方法与"校正"极为相近,如多元回归、PLS、人工神经网等,但侧重面不一样。参数测定的很主要应用是光谱的曲线拟合。

(7) 模式识别(pattern recognition): 该类方法主要应用多元统计方法来揭示隐含于事物内部的规律性。在化学中,主要用于化合物的分类。经典的方法如聚类分析、PCA、KNN、SIMCA及逐步判别分析(SDA)等。目前,人工神经网作为模式识别器在诸多应用中均获良好结果。

另外,与其说化学计量学尚包括数据库检索,人工智能,结构/活性相关性研究,还不如说化学计量学(如上述内容)作为一种工具而渗透或服务于这些领域。

六、 结 论

本文介绍了计算机化学中的主要内容。作为数据库检索,由于起步较早,发展得已比较成熟。目前,在大型分析测试仪器(如 MS、IR、NMR 等)中,数据库已成为其重要的组成部分。结构解析的人工智能研究,尽管起步较早,但是鉴于问题的复杂性和难度,至今尚在发展中。目前研究的焦点集中在多维波谱的应用上。计算机辅助合成在国外已有许多商用系统,但国内开展还极少,亟需加强。关于分子设计,目前在国内外均集中了一大批科学家在从事这一课题的研究,今后,各国的投资还将会继续增加。关于化学计量学,目前主要是已有方法的大量应用,但也不断有新的方法在涌现、在发展,如小波变换就是一例。

参 考 文 献

1 许禄. 化学计量学方法. 科学出版社, 北京: 1995

2 Lindsay R K, Buchanan B G, FeigenHaum E A *et al*. Application of Artificial Intelligence For Organic Chemistry — The DENDRAL Project. McGraw-Hill, New York: 1990

3 Funatsu K, NiShizaki M, Sasaki S. *J Chem Inf Comput Sci*, 1994; 34: 745

4 Razinger M, Balasubramanian K, Perdih M *et al*. *J Chem Inf Comput Sci*, 1993; 33: 812

5 Tomellini S A, Ste enson J M, Woodruff H B, *Anal Chem*, 1984; 56: 67

6 胡昌玉, 许禄. 中国科学(B 辑), 1994; 24: 1014

7 Hu C Y, Xu L. *J Chem Inf Comput Sci*, 1994; 34: 840

8 Hu C Y, Xu L. *Anal Chim Acta*, 1994; 295: 127

9 Hu C Y, Xu L. *Anal Chim Acta*, 1994; 298: 75

10 许禄, 郭传杰. 计算机化学方法及应用. 化学工业出版社, 1990

11 Ugi I, Bauer J, Bley K *et al*. *Angew Chem Int Ed Engl*, 1993; 32: 201

12 Xu L, Wang H Y, Su Q. *Comput Chem*, 1992; 16: 187

13 Xu L, Yao Y Y, *J Chem Inf Comput Sci*, 1995; 35: 45

14 Gramer R E III, Patterson D E. *J Am Chem Soc*, 1988, 110: 5959

15 Massart D L, Vandeginste B G M, Deming S N *et al*. Chemometrics: a textbook. Elser ier Science Publisher B V, Amsterdam: 1988

中国科技大学首轮"绿色化学"课程教学顺利结束

在中科院院士朱清时副校长的倡导下,中国科技大学开设了"绿色化学"课程。主要内容有:环境生态化学、零排放有机合成、环境友好催化、超临界流体技术、绿色生态材料、生物技术、绿色能源等章节,共计 60 学时。对象为化学与材料工程学院 97 级研究生和本科生。该课程体现了多学科交叉和面向经济发展主战场的特点。教员在备课过程中参阅了大量国内外近年来的科研成果及有关书籍。讲授的资料丰富、内容新颖,同学们都感到收获很大。

(中国科技大学化学物理系 范崇正 供稿)