计算机化学化工丛书

Computer Chemistry and Chemical Engineering Series

应用化学图论

许　禄　胡昌玉　著

科学出版社

内 容 简 介

本书为《计算机化学化工丛书》之一.

本书强调化学图论的实用性，尽可能对介绍的方法给出实例，同时以不同形式给出它们的可运行程序和源程序. 前三章简要介绍图论的基本概念、结构编码和拓扑指数；第四章和第五章介绍惟一性连接表及拓扑等价性算法，并详细介绍结构产生器；后三章用较大篇幅介绍了结构-性质相关性研究的成果，主要讨论化合物结构图特征的提取及压缩和构效关系，介绍分子形状的 2D 和 3D 表征方法，并给出了化合物相似度的概念及计算，以用于先导化合物的筛选.

本书可供从事计算机化学、生物化学、药物化学、医学化学、毒物化学、环境科学、化工、数据库等方面的研究人员及大学师生阅读参考.

图书在版编目（CIP）数据

应用化学图论/许禄、胡昌玉著. -北京：科学出版社. 2000
（计算机化学化工丛书/许志宏主编）
ISBN 7-03-007914-0

Ⅰ. 化… Ⅱ. ①许… ②胡… Ⅲ. 化学工业-应用数学-图论
Ⅳ. TQ011

中国版本图书馆 CIP 数据核字（1999）第 61957 号

科学出版社 出版
北京东黄城根北街 16 号
邮政编码：100717
源海印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2000 年 6 月第 一 版 开本：850×1168 1/32
2000 年 6 月第一次印刷 印张：11
印数：1—2500 字数：281 000

定价：26.00 元

（如有印装质量问题，我社负责调换〈北燕〉）

《计算机化学化工丛书》
编　委　会

《计算机化学化工丛书》序

化学化工是信息量特别大的一门学科．已发现的化合物超过2300万种，它们的各种性质，包括二元、三元、多元以及结构的性质，可以说是一个无边无际的数据海洋．于是，化学数据库的建设就成为20世纪后20年国际上的一件大事．中国科学院从1979年开始建设化学数据库，迄今已经整整20年．其间，多次得到国家和中国科学院的奖励．

长期以来，人们不仅希望能定性地掌握而且希望能定量地了解化学化工学科的规律，而计算机的高速发展，帮助人们一步一步地实现着这个愿望．从量子化学的计算、分子轨道的计算、分子动力学的计算、多种化合物的谱图解析、化学计量学、化工单元模拟、热力学和动力学的复杂计算、化工系统流程模拟计算、动态过程的模拟等，都已经利用计算机的帮助逐步得以实现．

上述基础工作为化学化工领域的工作者增加了很大的自由度：可以用计算机帮助进行药物分子设计，帮助对化合物的谱图解析，帮助选择合成路线，可以进行新过程、新技术的开发，可以进行大型工业装置的设计，可以对工厂的生产过程进行优化……

在改革开放的20年里，我们的计算机化学从无到有，已经形成了一个很大的用户和读者群体．到21世纪，这个群体更需要有能力利用计算机帮助自己的工作，所以本套丛书中也包含一些计算机化学化工的教材，以利于化学化工本科生和研究生的培养和工程人员自学．

所以，我们希望通过本套丛书介绍一些解决问题的方法，帮助读者在遇到问题时，知道如何去解决问题．为此，要求作者在自己的著作中，要给出软件、数据的出处、网络地址或光盘．时

代发展得很快，仅做到这一点显然还是不够的．我们特别注意到近年来，Internet 网络的高速发展已经给我们的时代带来了巨大变化．首先它是连接 4000 多万个节点的一个网络系统，它上面有无限多的各种信息的存储．这些信息一个最大的长处是时间的滞后最少，易于通过计算机帮助搜索需要的信息．

如果能将科学数据库在网上的功能，由数据的存取扩大到运算、绘图、模拟等多方面，必将极大地推动科学数据库工作的发展和广泛使用．在 21 世纪，将逐步可以做到，人们可以在用户端将数据从库中取出，在服务器端程序系统上计算，结果以图形或多媒体方式输出到用户端．据了解，我们有些作者通过不同途径，已能达到要求．能够在网上实现数据查询、计算、绘图、三维图形显示等．

进入 21 世纪，Internet 网络系统的应用将更加普及．Internet 网络的客户/服务器的应用，将进入千家万户，进入教室和办公室的各个角落．所以，如果能将科学数据库和计算程序库搬到网络的服务器上，那么它的普及应用将会随着计算机网络的推广而推广．

如果有的丛书作者，目前仅能给出单机版本的软件，也欢迎他们能再作一点工作，很快能达到上网服务的目的．相信进入 21 世纪不久，在用户端上，人们就有可能逐步享受到多种媒体的全方位的科学信息的服务．

这套丛书是我国多位化学化工学科的专家、教授、学者多年辛勤劳动的成果，也是科学出版社、国家自然科学基金委员会优秀研究成果专著出版基金和中国科学院科学出版基金大力支持的产物，希望其能促进我国 21 世纪计算机化学化工学科的发展，并促进相关学科发展．

《计算机化学化工丛书》编委会

2000 年 1 月

前　言

图论是离散数学的一个分支，它与拓扑学和组合数学密切相关．将图论用于化学，则为化学图论．本书侧重于阐述其实用性，故命之为《应用化学图论》．

图论已广泛应用于诸多学科领域，如建筑学、通讯、遗传学、工业管理、几何学、心理学及社会学，但是重要的应用领域应首推化学学科．在化学中，已有的应用涉及合成化学、结构自动解析、聚合化学、量子化学、金属有机化学、热化学、化学动力学、统计力学、相平衡、波谱学及化学信息的存储和检索等．近年来，最大量的应用集中在定量结构－活性/性质相关性（QSAR/QSPR）的研究方面．

当今，在生物化学、药物学、毒物学和环境科学中一个很重要的发展趋势是将分子生物活性/性质的定量预测用于分子设计．有人估计，约从10000个化合物中才能够筛选出一个作为有效药物投入临床使用．以往凭经验由母体化合物衍生同源化合物的效率很低，因而促使人们去寻找更优的方法来创制新药．与此类似，在工业和环境科学中，每年要测试（毒性、致突变、致癌等）和处理的化合物多达数十万种，若仅靠实验将耗费极大的人力物力，所以必须有新的方法才能满足客观需要．QSAR/QSPR方法为我们提供了一条可行的途径．目前已涌现出了诸多QSAR/QSPR方法，其中图论方法有其独特优点，因为这类方法仅依赖于分子结构，即由结构图可以直接衍生结构特征，因而近年来人们作出许多努力来开拓这类方法．

全书共八章．作为基础，第一章对于图的基本知识作了扼要介绍．第二章概略叙述了化合物结构的编码及计算机描述．第三章为图论指数，即拓扑指数，侧重于高选择性拓扑指数．关于拓

扑指数与化合物性质的相关性在第六章给出.

进行高选择性拓扑指数研究的必要和先决条件是要有结构的生成程序. 事实上，该类程序是有机化合物结构自动解析中的关键环节，人们常称之为“结构产生器”. 当给定一分子式，结构产生器应能穷举地、无冗余地而且是高效地将所有同分异构体生成出来. 而拓扑等价性是结构穷举生成过程中的重要概念. 同时，拓扑等价性的概念还可用于化合物结构的显示等许多方面. 因而我们分两章阐述，即第四章介绍惟一性连接表和拓扑等价性，第五章介绍结构产生器.

本书很重要的一个方面是 QSAR/QSPR 研究，其涉及第六、七、八章. 第六章讨论化合物结构图的表征和构效关系，较详细地介绍了变量的提取和压缩. 第七章介绍分子形状的表征方法，相关的方法不仅可直接用于三维定量构效关系研究（3D QSAR），而且分子形状亦可作为参数用于 2D QSAR 数学模型的构造. 第八章给出化合物相似度的概念及计算，并用于先导化合物的筛选.

考虑到本书的实用性，一般每介绍一种方法，都尽可能给出应用实例，同时还以不同形式给出本书中主要方法的可运行程序及源程序.

中国科学院长春应用化学研究所的齐玉华、郝军峰、章文军、胡建强等同志为本书的编写、定稿做了大量的工作，在此一并表示感谢.

在本书出版之际，还要特别感谢中国科学院科学出版基金和国家自然科学基金委员会优秀研究成果专著出版基金给予的出版资助，以及科学出版社刘俊来编辑所付出的辛勤劳动.

由于作者水平所限，书中缺点和错误在所难免，敬请读者不吝赐教.

作　者

1999 年 6 月于中国科学院长春应用化学研究所

目　录

第一章　图论的基本概念

1.1　图的定义

图论中的中心概念是图.为说明图论的概念,我们首先引进简单图.

一简单图定义为一有序对$[V(G),E(G)]$,此处$V=V(G)$,是一非空集.其元素称为图G的顶点(或点);$E=E(G)$,是边(或线)的集合,它是$V(G)$中元素的无向对.$V(G)$和$E(G)$是图的顶点集和边集.N和M分别为顶点数和边数.

图论的重要特点是图的可视性.因为顶点可用小的圆圈或点表示,而边可用直线或曲线表示.

一简单图示于图1.1,此图是具有标号的简单图,其顶点集$V(G)$为$\{V_1,V_2,V_3,V_4\}$,或简单表示为$\{1,2,3,4\}$;它的边集$E(G)$为$\{V_1,V_2\}$,$\{V_2,V_3\}$,$\{V_2,V_4\}$和$\{V_3,V_4\}$,或简单表示为$\{1,2\}$,$\{2,3\}$,$\{2,4\}$和$\{3,4\}$.边集也可表示为$\{e_1,e_2,\cdots,e_m\}$.

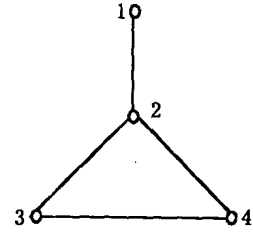

图1.1　一简单图G

图论中的另一概念是序列(family).序列允许其中的元素多次重复.如$\{1,2,3,4\}$为一集合,而$\{1,1,2,2,2,3\}$为一序列.

在图论中,两个顶点允许多条边与之相连.同时,一条边可以连接同一个顶点,此时称之为圈(loop).一个广义图允许有多重边及圈.图1.2中G_1为多重图,G_2为具有圈的多重图.

有一类特殊图为有向图.一有向图D定义为一有序对$[V(D),A(D)]$,此处$V(D)$为顶点集,$A(D)$为弧序列.图1.3

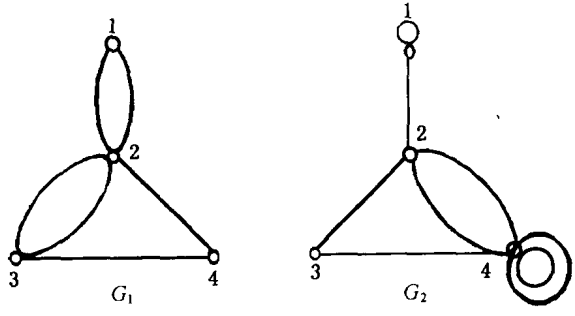

图 1.2　多重图 G_1 及具有圈的多重图 G_2

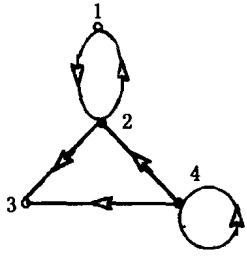

图 1.3　具有标号的有向图 D

为一有向图，其中 $V(D)=\{1,2,3,4\}$，$A(D)=\{(1,2),(2,1)(2,3),(4,2),(4,3),(4,4)\}$.

本书中，如不特别注明，均认为图为无向图，而且没有多重边及圈.

1.2　邻接及关联

我们说一个图的两个顶点 V_i 和 V_j 是邻接的，则有一条边连接这两个顶点，此时称顶点 V_i 和 V_j 对于此边是关联(incidence)的. 相类似，若图 G 的两个不同的边 e_i，e_j 是邻接的，则它们至少有一个顶点是共享的. 如图 1.4 所示，顶点 V_1 和 V_3 相邻接，而 V_2 和 V_4 不邻接. 同样，e_1 和 e_2 是邻接的，而 e_1 和 e_5 是不邻接的.

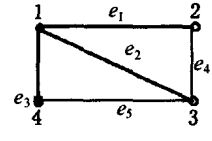

图 1.4　邻接和关联的概念

1.3　图的同构

同构是指两个图的顶点一一对应. 两个图同构，即两图相同，

其不同的是两图在画法上有差异，如图 1.5 所示. 对于简单的同构图 G_1 和 G_2 易于识别，而对于比较复杂的 G_3 和 G_4 则不易于识别. 在图论中，对于同构图的识别是一 NP 问题，因为两图影射时有 $N!$ 种可能性.

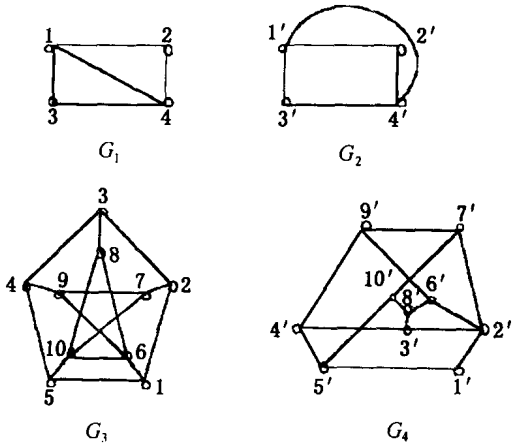

图 1.5　两对同构图

图论中有一非常重要的概念是图的不变量. 所谓图的不变量是指某量对于图 G 的任何同构图都是相同的. 由此，图的顶点和边都是图的不变量.

1.4　步程、行迹、路径、距离及价

图 G 的某一步程(walk)是点和边的交替序列 $e_0, v_0, e_1, v_1, e_2, v_2, \cdots, e_i, v_i$，且起点与终点均为顶点，在此序列中，边的前后二邻接点与此边相关联，这样的序列也可表示为 $v_0, v_1, v_2, \cdots, v_i$(边不显性地表示)，而步程长度是步程中的边数. 封闭步程为 $v_i - v_i$，即一步程开始并终止于同一顶点，否则，称为开放步程. 所有边均不相同的步程为行迹(trail)，而所有顶点均不相同的步程为路径(path).

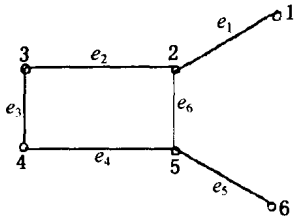

图 1.6 用于步程、行迹、路径概念的图

如图 1.6 所示，顶点序列 $v_1\ v_2\ v_3\ v_2 v_5$ 是一步程，它也可表示为 12325. 长度为 4 的一路径是 $v_1\ v_2\ v_3\ v_4\ v_5$，同样它可表示为 12345. 步程 12521 和 121 是封闭的，而步程 345、123256 及 12543 是开放的. 步程 345 和 12543 分别为长度是 2 和 4 的路径. 步程 234521 为长度是 5 的行迹.

两顶点 v_i 和 v_j 间最短的路径为距离，记为 $d(v_i, v_j)$ 或者 $d(i,j)$. d 为非负量且均为整型，其具有如下性质：

$d(i,j)=0$，当且仅当 $i=j$ 时；

$d(i,j)=d(j,i)$；

$d(i,j)+d(j,k)\geqslant d(i,k)$；

$d(i,j)=1$，当且仅当 $(i,j)\in E(G)$ 时，即 (i,j) 为某一边.

在图 G 中，若任一对顶点由路径相连接，则称 G 为连通图，否则，为非连通图，$d(i,j)=\infty$，此二点分属于图 G 中的不同部分. 图 G 所属部分将记为 $K=K(G)$，如图 1.7 中 G_1、G_2 和 G_3 分别由一、二、三部分组成.

由于我们已经引进“距离”的概念，则顶点的价或度则易于定义. 距离为 1 的顶点（即邻接顶点）称为第一层近邻，距离为 2 的顶点称为第二层近邻等等. 顶点 v_i 的第一层近邻的顶点数称为 v_i 的价和度，记为 $D(i)$，它是入射到此顶点的边数. 价为 0 的顶点称之为游离顶点. 价为 1 的顶点称之为终端顶点.

在图 G 中，所有顶点价的加和为边数的二倍，因为在加和中每一条边计数两次.

$$\sum_{i=1}^{N} D(i) = 2M$$

另外，顶点的价 1、2 和 3，分别以 F、S 和 T 表示，则

$$F + 2S + 3T + \cdots = 2M$$

$$F + S + T + \cdots = N$$