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 $\{e_1,e_2,\cdots,e_m\}$.

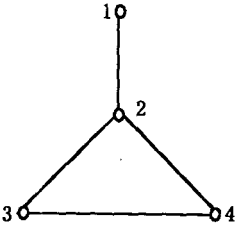


图 1.1　一简单图 G

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，$A(D)$为孤序列. 图 1.3

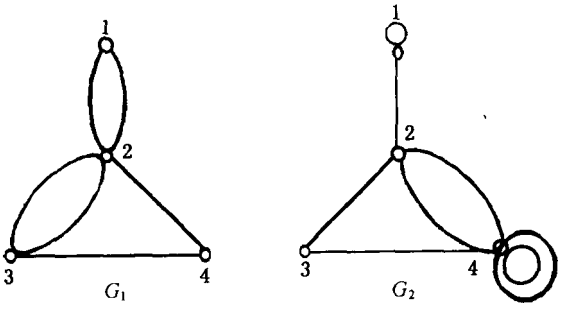


图 1.2　多重图 G_1 及具有圈的多重图 G_2

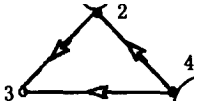


图 1.3　具有标号的有向图 D

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．

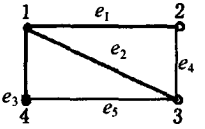


图 1.4　邻接和关联的概念

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。

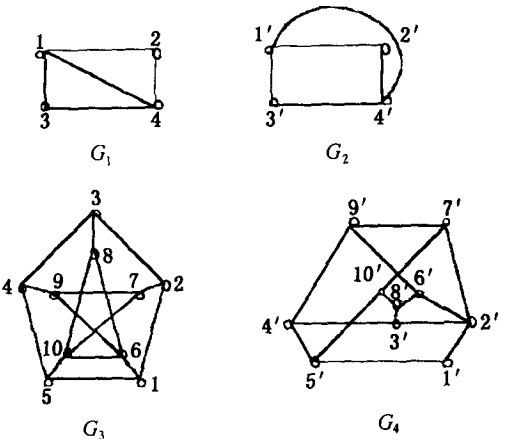


图 1.5　两对同构图

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）.

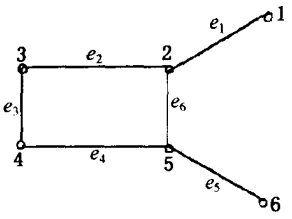


图 1.6 用于步程、行迹、路径概念的图

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

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

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

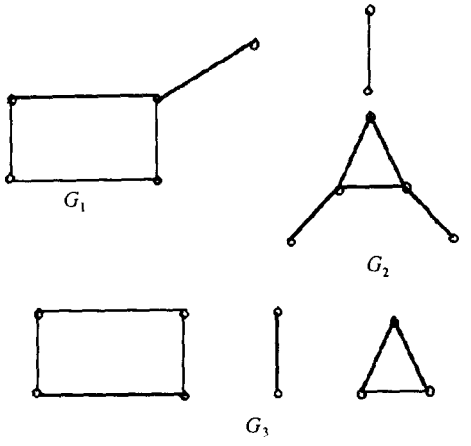


图 1.7 图 G 的组成

$1[K(G_1)=1],2[K(G_2)=2],3[K(G_3)=3]$

式中 M 和 N 分别为边数及顶点数.

1.5 子 图

若图 G' 是图 G 的子图,则 $V(G')$ 是 $V(G)$ 的子集,$E(G')$ 是 $E(G)$ 的子集:

$$V(G') \subseteq V(G)$$

$$E(G') \subseteq E(G)$$

若由图 G 中仅删去一条边 e,则子图 $G-e$ 含图 G 的全部顶点和 $M-1$ 条边若从图 G 中删除一个顶点 v,则子图 $G-v$ 含图 G 的 $N-1$ 个顶点及 $M-D(v)$ 条边.子图 $G-(e)$ 是指删除图 G 的一条边及其相连的两个顶点 u 和 v,则此子图含图 G 的 $N-2$ 个顶点和 $M-[D'(u)+D'(v)]+1$ 条边. 支撑子图是含原图 G 全部顶点的子图子图的例子示于图 1.8.

图 G 的子图 $G-v_i(i=1,2,\cdots,N,N\geqslant 3)$ 的完全集,与另一图 G' 的子图的完全集相同,Ulam[1]猜测图 G 与 G' 同构这一猜

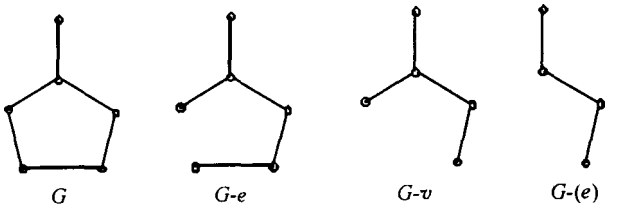


图 1.8 一套属于图 G 的子图

测由树图[2]及某些特殊类别的图[3]所证明,但对于其他任意的图尚待证明.

1.6 正 规 图

每一顶点价均相同的图称为正规图（regular graph).若顶点的度(价)为 D,则此图称作度为 D 的正规图.

若在一连通图中的全部顶点的度均为 2,则此图称为环或圈.环记为 $C_N(N\geqslant 3)$. 若 N 为偶数,则环为偶数,否则为奇数.

对于度为 D 的正规图，则满足：

$$M = \frac{1}{2}ND$$

其意为对于一正规图,仅可能 N 或 D 其一为偶数.

若 $D=1$,则仅有一连通(正规)图,此图有两个顶点并有一条边相连,记为 K_2,一正规图若有 N 个顶点,且度 $D=N-1$,则称之为完全图,记为 K_N.在此类图中，每一对顶点均相连接，其边数为：

$$M = \left(\frac{N}{2}\right) = \frac{N(N-1)}{2}$$

图 1.9 所示为正规图及完全图.

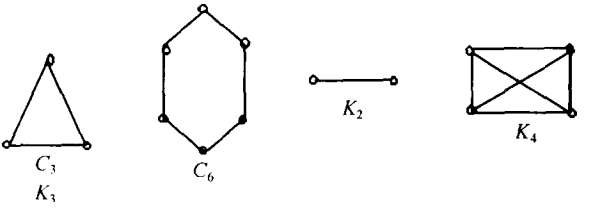


图 1.9　正规图和完全图的例子

1.7　树

没有环的图为非环图. 树是连通图的非环图. 若一树含 N 个顶点,则其边数为 $N-1$. 有根的树是一种树，在此树中有一顶点以某种方式区分于其他顶点. 图 1.10示出树及有根树的例子.

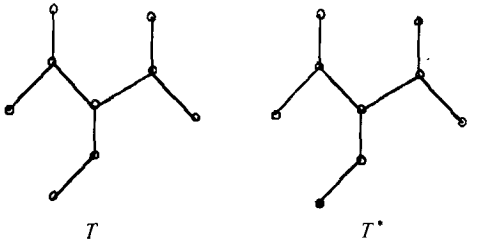


图 1.10　树 T和有根树 T^*(黑点表示根顶点)

树必须具有度为 1 的顶点，称之为终端顶点. 具有极少终端顶点(2个)的树称为链(chain);具有极大终端顶点 ($N-1$ 个)的树称为星(star). 链和星的例子示于图 1.11.

树的重要特征是支化，但支化是直观的概念而并非严格的定义. 支化顶点的度大于或等于 3. 除链外所有的树均有支化顶点，在一星中极大支化顶点为中心顶点.

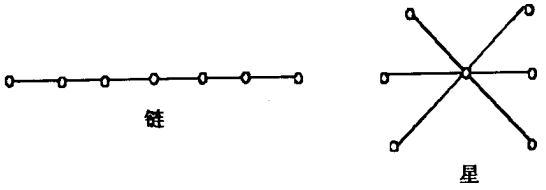


图 1.11　链和星的例子

1.8　平　面　图

一个图画在平面上，而没有任意两条边相交，则此图为平面图. 图 1.12 中 G 为平面图,而 K_5 为非平面图,是一度数为 5 的完全图,此图很有名,由 Kuratowski[4]在他的“论平面图”一文中给出. 图 G 似乎不是平面图，若将其变换成另一种画法，如图 1.12 中的 G'就一清二楚了,因为 G'即 G.

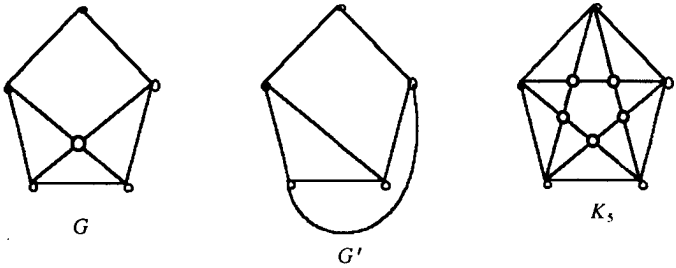


图 1.12　平面图(G, G')和非平面图(K_5)

平面图是由 Euler[5]在他的多面体研究中所引入的. 对于多面体的顶点(V)、边(E)和面(F)符合如下关系式：

$$V - E + F = 2$$

此式称为 Euler 公式. Euler 公式同样适用于平面图. 但需指出,若平面图要分成两部分，一部分是一无限面，另一部分含一个或多个面，对于平面图公式则变为：

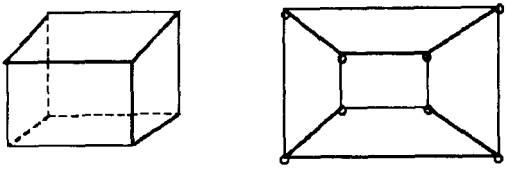
$$V(G) - E(G) + F(G) + 1 = 2$$

即

$$V(G) - E(G) + F(G) = 1$$

其中“1”表示一个无限面.

图 1.13 示出一立体(cube)图和一平面图 G.“cube”属于多面体图，它的平面图即为 G.



立体图

$V=8$

$-E(G)=-12$

$F=6$

2

G

$V(G)=8$

$-E(G)=-12$

$F(G)=5$

1

图 1.13 立体(cube)图及相应的多面体 G

Rudel[6]、Levin[7] 及 Mindel[8]等发现相律与 Euler 公式很相似：

$$P - C + F = 2$$

其中 P、C和 F 分别为平衡体系中不同相的相数、相中成分数及自由度.

1.9 化 学 图

在化学中,图可以描述不同的内容,如分子、反应、晶体、聚合物、簇等等.其共同特征是点及其点间的连接.点可以是原子、分子、电子、分子片断、原子团及轨道等,点间的连接可以是键、键及非键作用、反应的某些步、重排、van der Waals 力等等.

化学图中的一类为分子图，即结构图.在这类图中顶点为原子,边为键.为简单起见,一般将氢原子略去,此时结构图称为分子

骨架或隐氢图. 分子图中一般不考虑几何的、立体及手性的因素即便如此，分子图仍可较好地用于化合物物理化学性质的预测，这是化学图论得以发展的重要原因.

另外，对于 Hückel 图和 Kekulé 图，尽管我们并不打算作详细介绍，但是由于在阅读文献时经常会碰到这两个概念，所以在此处顺便给出扼要的说明.

Hückel 图是一隐氢分子图，而此分子为共轭分子，它是一无向平面图的连通图. 这种图首先由 Hückel 所应用，所以常称为 Hückel 图. 这种图用于描述共轭系统中 π 电子间与其第一层紧邻的相互作用. 如苯及其 Hückel 图示于图 1.14.

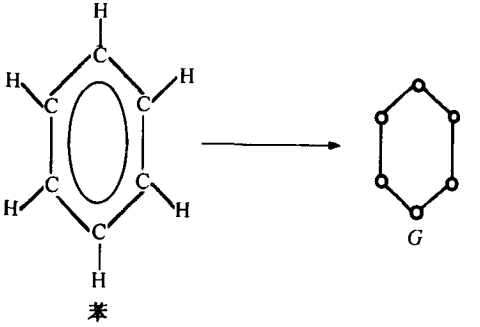


图 1.14 Hückel 图 G 及所表征的苯

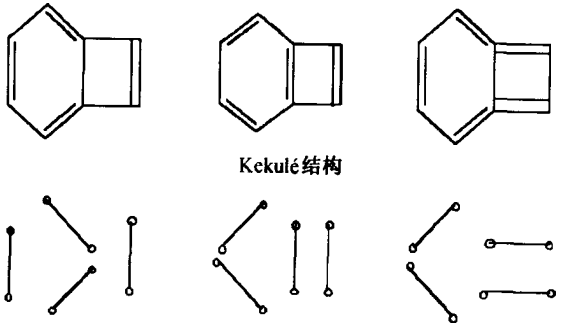


图 1.15 苯环丁烯的 Kekulé 结构和图

Kekulé 图是一非连通图，它由一个或多个 K_2 成分组成．在一个 Kekulé 结构中 K_2 数等于双键数．图 1.15所示为苯环二丁烯的 Kekulé 结构及其相应的 Kekulé 图．

1.10 图论矩阵

图论中最常用的是邻接矩阵和距离矩阵．其中，邻接矩阵又分为顶点邻接矩阵和边邻接矩阵．

对于具有 N 个顶点的连通图 G，顶点邻接矩阵 $A(G)$是一个 $N\times N$ 的对称矩阵，矩阵中元素为：

$$(A)_{ij}=\begin{cases}1 & \text{顶点 } V_i \text{ 和 } V_j \text{ 是邻接的}\\ 0 & \text{其他}\end{cases}$$

$$(A)_{ii}=0$$

图 1.16给出多种顶点邻接矩阵．

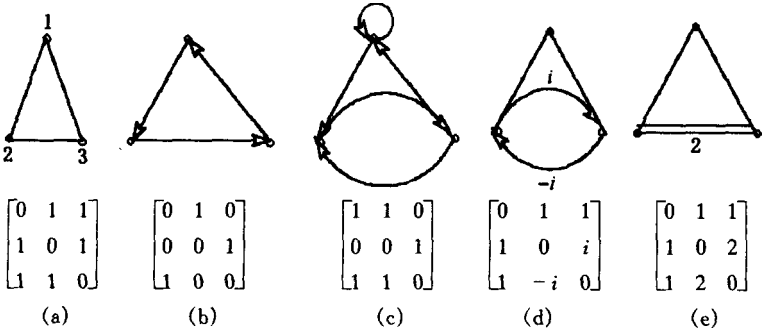
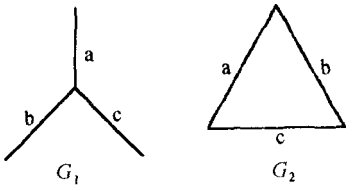


图 1.16 多种图的顶点邻接矩阵

(a)无向图；(b)有向图；(c)带有圈的有向图；(d)边权重的有向图；(e)具有多重边的无向图

边邻接矩阵的定义与顶点邻接矩阵类似，但顶点邻接矩阵惟一地对应一图，而边邻接矩阵则不一定惟一对应，因为二非同构图可以为同一边邻接矩阵（见图 1.17）．

一个图 G 的距离矩阵 $D(G)$定义为：

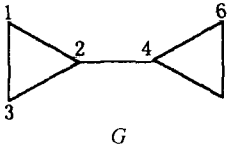


G_1　　G_2

图 1.17　具有相同边邻接矩阵的图 G_1 和 G_2

$$(D)_{ij}=\begin{cases}p_{ij} & \text{若 } i\neq j\\ 0 & \text{若 } i=j\end{cases}$$

其中 p_{ij} 是顶点 V_i 和 V_j 间最短的路径(path). 图 1.18 给出距离矩阵的例子. 将距离矩阵 $D(G_1)$ 中大于"1"的元素删除则得顶点邻接矩阵 $A(G)$.



G

$$D(G)=\begin{bmatrix}0&1&1&2&3&3\\1&0&1&1&2&2\\1&1&0&2&3&3\\2&1&2&0&1&1\\3&2&3&1&0&1\\3&2&3&1&1&0\end{bmatrix}$$

图 1.18　距离矩阵的例子

参 考 文 献

[1] S. M. Ulam, A Collection of Mathematical Problems, John Wiley & Sons, New York, 1960.

[2] P. J. Kelly, Pac. J. Math., 1957, 7, 961.

[3] B. Manvel, In Proof Techniques in Graph Theory, E. Harary Ed., Academic Press, New York, 1969, 103.

[4] H. S. M. Kuratowski, Regular Polytopes, 3rd ed., Dover, New York, 1973.

[5] I. Euler, Comment ACAD. Sci. Imp. Petropolitanae, 1736, 8, 128.

[6] O. Rudel, Z. Electrochem., 1929, 35, 54.

[7] I. Levin, J. Chem. Educ., 1946, 23, 183.

[8] J. Mindel, J. Chem. Educ., 1962, 39, 512.

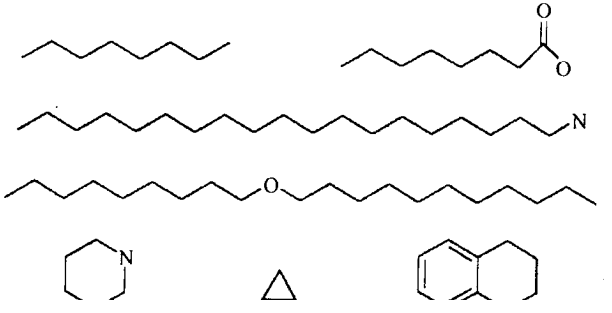
第二章　结构编码

2.1 引　　言

化学信息学中一个最基本的问题就是化合物结构在计算机中的描述. 数据库的登录，化合物结构的存储、管理和检索，结构解析专家系统，有机合成线路方案选择，计算机辅助分子设计等诸多方面的研究都和结构的计算机表达即结构编码密切相关. 在介绍结构编码之前，有必要先介绍一下图及化合物结构的基本概念.

化合物结构是一个意义非常广泛的概念.从化合物种类来看，可分为有机化合物、无机化合物、高分子化合物、生物大分子等.从成键类型看,可分共价结构、晶体结构、离子结构、配合结构等；依据结构中是否含有不确定基团，可分为确定结构与族性结构，即 Markush 结构.

本书中所提到的结构，若无特殊说明，皆是指原子间以共价键相连的有机化合物的结构，而且一般指二维结构，即拓扑结构，它只反映原子之间的连接关系，如：



拓扑结构在计算机化学中是一种非常重要且有用的概念. 图

论中有关连通性、对称性、同构与同态及子图等概念和理论都可直接应用于化合物的拓扑结构.

除了二维结构外还有空间结构，即三维结构. 三维结构有构型(configuration)和构象(conformation)两种. 构型包括顺反结构、椅式构型和手性中心. 构象则包括构型及各原子在三维空间中的相对位置. 各种结构类型间的层次关系为：

二维结构（拓扑结构）→构型→构象

有机化合物的结构异构体是指分子式相同而各原子之间连接关系不同，即拓扑结构不同的异构体. 构型异构体是指拓扑结构相同，而构型不同的异构体. 构象异构则是指构型相同，而原子在三维空间中的相对位置不同的异构体.

2.2 结构编码

一种较好的用于计算机处理的结构编码方案应满足以下四个基本要求：

(1) 要使结构输入简单易行，即要能很容易地把化学家所常用的标准结构图转化为一种无二义的计算机码.

(2) 要便于结构信息处理，如结构信息提取、子结构抽提等. 同时要能很容易地与其他方案进行相互转化.

(3) 计算机编码应易于还原为结构图形式.

(4)应具有惟一性,单义性.

有机化合物结构在计算机中的存储形式是多种多样的，常用的方法有线性编码、邻接矩阵、连接表、连接堆栈、拓扑指数编码和碎片编码等. 其中最常用的是二维连接表，而邻接矩阵和连接堆栈则是连接表的不同表达形式.

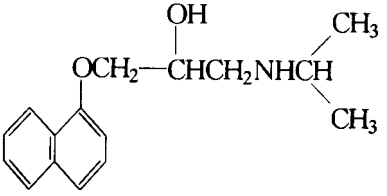
在进行化合物结构的计算机存储、管理与检索时，总是要对库中的每一化合物结构都赋以一个数值编码作为惟一识别码. 如美国化学文摘的 CA 登录号，而 CA 登录号是在 CA 库建立过程中任意赋予的，故与化合物自身的结构或性质毫无关系，对任一

化合物，从其结构本身无法知道它的 CA 登录号.

2.2.1 线性编码

线性编码方案主要有 WLN[1~3]、IUPAC[4]、Hayward[5] 等. Lynch 等人曾对各种线性编码方案作过详细的比较[6].

WLN(Wiswesser Line-formula Notation)由 Wiswesser 于 1949 年首次提出，在 1968 和 1976 年 Smith 对此作了两次修订[2,3]，80 年代又出现了新版本 AWLN(Advanced WLN)及其他改进方案. 它以一个字符串表示一个化合物的结构，如结构 St2-1 及其 WLN 码如下：



St2-1

WLN: L66JBO-1YQ1My

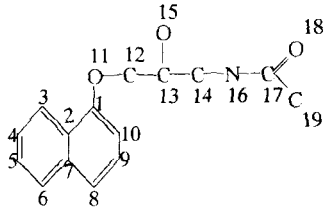
其中：L66 表示连在一起的两个 6 元碳环；J 表示闭环；B 表示原子在环上位置；BO 表示 B 位上有一个氧原子；1Y、Q1 表示 —CH_2CH—(OH)CH_2—；MY 表示异丙氨基.

由于(1)规则复杂，线性编码与其他编码及结构图之间很难实现快速、准确转换，(2)从线性编码抽提结构信息及处理都较为困难，因而线性编码在使用上受到很大限制.

2.2.2 邻接矩阵及二维连接表[7,8]

图可用一个邻接矩阵表示，而化学结构就是一个拓扑图，因而化合物结构也可用一个邻接矩阵表示. 表示化合物结构的邻接矩阵与图的邻接矩阵不同，在我们的 ESESOC 系统中所应用的化合物结构的邻接矩阵形式为：(1)对角元不是 0，而是元素；(2)非 0 元素可以是 1,2,3,9 等分别表示单键、双键、三键、芳香键. 如

结构 St2－2 及其邻接矩阵示意于图 2.1



1	C	9	0	0	0	0	0	0	9	1	0	0	0	0	0	0	0	0
2	9	C	9	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0
3	0	9	C	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	9	C	9	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	9	C	9	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	9	C	9	0	0	0	0	0	0	0	0	0	0	0
7	0	9	0	0	0	9	C	9	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	9	C	9	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	9	C	9	0	0	0	0	0	0	0	0
10	9	0	0	0	0	0	0	0	9	C	0	0	0	0	0	0	0	0
11	1	0	0	0	0	0	0	0	0	0	O	1	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	1	C	1	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	1	C	1	1	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	1	C	0	1	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	1	0	O	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	N	1	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	C	2	1
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	O	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	C

图 2.1　邻接矩阵示意图

把邻接矩阵中的 0 元素略去，重新整理一下即可得结构图的二维连接表. 二维连接表有多种形式，如 CAS Register III 形式[9]、Ray and Kirsch 形式[10]、DENDRAL 形式[11]等. 表 2.1 给出了结构 St2－2 在 CIAC－^{13}C NMR 谱图数据库系统中的二维连接表[12].

由于二维连接表中把每个连接键都描述了两次，因而浪费计算机系统资源，若剔除冗余描述，即得到二维连接表的紧凑形式，见表 2.1.

表 2.1 结构 St2-2 的各种二维连接表

CIAC-^{13}C NMR 数据库中的连接表							紧凑连接表		
序号	元素	电荷	氢原子数	连接关系			连接关系		
1	C	0	0	2(9)	10(9)	11(1)	2(9)	10(9)	11(1)
2	C	0	0	1(9)	3(9)	7(9)	3(9)	7(9)	
3	C	0	1	2(9)	4(9)		4(9)		
4	C	0	1	3(9)	5(9)		5(9)		
5	C	0	1	4(9)	6(9)		6(9)		
6	C	0	1	5(9)	7(9)		7(9)		
7	C	0	0	2(9)	6(9)	8(9)	8(9)		
8	C	0	1	7(9)	9(9)		9(9)		
9	C	0	1	8(9)	10(9)		10(9)		
10	C	0	1	1(9)	9(9)				
11	O	0	0	1(1)	12(1)		12(1)		
12	C	0	2	11(1)	13(1)		13(1)		
13	C	0	1	12(1)	14(1)	15(1)	14(1)	15(1)	
14	C	0	2	13(1)	16(1)		16(1)		
15	O	0	1	13(1)					
16	C	0	1	14(1)	17(1)		17(1)		
17	C	0	0	16(1)	18(2)	19(1)	18(2)	19(1)	
18	O	0	0	17(2)					
19	C	0	3	17(1)					

ESESOC 系统的连接表

序号	原子	邻接原子				连接键				连接度
1	12	2	10	11	0	9	9	1	0	3
2	13	1	3	7	0	9	9	9	0	3
3	11	2	4	0	0	9	9	0	0	2
4	11	3	5	0	0	9	9	0	0	2
5	11	4	6	0	0	9	9	0	0	2
6	11	5	7	0	0	9	9	0	0	2
7	13	2	6	8	0	9	9	0	0	3
8	11	7	9	0	0	9	9	0	0	2
9	11	8	10	0	0	9	9	0	0	2
10	11	1	9	0	0	9	9	0	0	2
11	15	1	12	0	0	1	1	0	0	2
12	2	11	13	0	0	1	1	0	0	2
13	4	12	14	15	0	1	1	1	0	3
14	2	13	16	0	0	1	1	0	0	2
15	14	13	0	0	0	1	0	0	0	1
16	19	14	17	0	0	1	1	0	0	2
17	8	16	18	19	0	1	2	1	0	3
18	16	17	0	0	0	2	0	0	0	1
19	1	17	0	0	0	1	0	0	0	1

紧凑连接表只给出连接关系，而其他信息与 CIAC-^{13}CNMR 数据库中的连接表形式一样，故略去 .