第1期（总第21期） 南京药学院学报 No. 1 (General No. 21)
1983年5月 Journal of Nanjing College of Pharmacy May, 1983

信息群分与先导化合物的系列设计研究
— 103个脂肪取代基的信息群分法

王尔华 彭司勋
（药物化学研究室）

摘 要 本文根据取代基的5种理化参数Fr、H_A、H_D、MR和F，用信息群分法将103个脂肪取代基分成5群、10群和20群，并给出了分类树叉图，可供先导化合物系列设计时选择合成对象作参考。

关键词 信息群分，聚类分析，103个脂肪取代基

药物研究中，常需从先导化合物（Lead compound）入手进行同源物的系列设计。设计者总是希望合成的化合物为数较少而又具有较广泛的化学结构特征、从而用不太多的工作量获取最大的构效关系信息，进而建立初步的QSAR方程式或判别函数，以指导下一步的合成设计。但是，在系列设计中，可供选择的取代基很多，先导化合物结构改造的位置有多个，式（1）表明系列设计时可能组合的类似物数目有很多种[2]：

$$\text{类似物总数} = x^k \cdot \frac{n!}{k!(n-k)!} \qquad (1)$$

式中，x是可供选择的取代基数目；n是先导化合物非对称的可改造位置数目；k是一次引入先导化合物的取代基数目。以喹啉为例，它有7个非对称位置，若有10个取代基，每次引入2个则需合成2100个化合物，若有20个取代基，则需合成8400个化合物。如要这样做，显然要费浩繁的工作量，而且往往内含类似信息的甲、乙、丙、丁式的“me too”化合物[2]，人们不可能也没有必要采用“穷举法”去一一合成与试验（有些在理论上存在而事实上无法合成），总是从现有的药化、生化、药理等知识出发，精选少量进行探索性合成。不过，这样作出的判断常常是主观的，因而设计效果不是最佳的。因此，多年来，人们一直进行这方面的探索性研究，以寻找较合理的系列设计方法。Hansch[1,3]用聚类分析法（Cluster Analysis）将药物化学家惯用的取代基（脂肪的和芳香的），根据多种结构参数用欧氏距离作为相似性统计量进行归组分群，编制了脂肪和芳香两种取代基聚类群分表，以供药物研究参考。本文在此基础上，采用取代基向量空间的夹角余弦——$\cos\theta_{ij}$作为分类统计量，用取代基向量的信息量——I_i作为相似性矩阵简化的数学尺度，对103个脂肪取代基[2]进行了“信息群分”，所编信息群分表可供合成设计参考。

1982年9月20日收稿
1982年12月16日修改

方　法

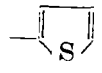
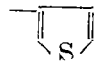
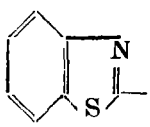
本文采用取代基的片断疏水常数（Fragment constants of hydrophobicity, Fr）、氢键供体虚潜参数（dummy parameters of hydrogen bond donors, H_D）、氢键受体虚潜参数（dummy parameters of hydrogen bond acceptors, H_A）、克分子折射（Molecular refraction, MR）和场效应常数（Swain & Lupton-type F constants, F）等作为取代基的疏水性、电性和空间效应的结构特征（详见表1）。以 r_{jk} 和 $\cos\theta_{il}$ 分别作为变量和取代基的分类统计量，进行R型逐次和Q型逐次信息群分。全部计算均用自编 ALGOL-60 语言程序，在国产709机上完成。

具体实施步骤如下：

表1　103 个脂肪取代基及其理化参数

编号	取代基	Fr	H_A	H_D	MR	F
1	Br	0.20	0.00	0.00	8.80	0.44
2	Cl	0.06	0.00	0.00	5.93	0.41
3	F	−0.38	0.00	0.00	1.05	0.43
4	I	0.59	0.00	0.00	13.76	0.40
5	NO_2	−1.16	1.00	0.00	6.71	0.67
6	H	0.23	0.00	0.00	1.03	0.00
7	OH	−1.64	1.00	1.00	2.55	0.29
8	SH	−0.23	0.00	1.00	8.76	0.28
9	NH_2	−1.54	1.00	1.00	4.37	0.02
10	CBr_3	2.03	0.00	0.00	28.81	0.27
11	CCl_3	1.61	0.00	0.00	20.12	0.31
12	CF_3	0.29	0.00	0.00	5.02	0.38
13	CN	−1.27	1.00	0.00	5.39	0.51
14	SCN	−0.48	1.00	0.00	13.40	0.36
15	CO_2^-	−5.19	1.00	0.00	5.15	−0.15
16	CO_2H	−1.11	1.00	1.00	6.03	0.33
17	CH_2Br	0.74	0.00	0.00	13.39	0.10
18	CH_2Cl	0.60	0.00	0.00	10.49	0.10
19	CH_2I	1.13	0.00	0.00	18.60	0.09
20	$CONH_2$	−2.18	1.00	1.00	9.81	0.24
21	$CH=NOH$	−1.02	1.00	1.00	10.28	0.25
22	CH_3	0.77	0.00	0.00	5.65	−0.04
23	$NHCONH_2$	−2.90	1.00	1.00	13.72	0.04
24	OCH_3	−1.54	1.00	0.00	7.33	0.26
25	CH_2OH	−1.10	1.00	1.00	7.19	0.00

编号	取代基	Fr	H_A	H_D	MR	F
26	$SOCH_3$	−2.24	1.00	0.00	13.70	0.52
27	OSO_2CH_3	−1.34	1.00	0.00	16.99	0.39
28	SCH_3	−0.02	0.00	0.00	13.33	0.20
29	$NHCH_3$	−1.38	1.00	1.00	9.11	−0.11
30	CF_2CF_3	1.34	0.00	0.00	9.23	0.44
31	C≡CH	0.01	0.00	1.00	8.25	0.19
32	CH_2CN	−0.73	1.00	0.00	10.11	0.21
33	$CH{=}CHNO_2$(反式)	−0.63	1.00	0.00	16.42	0.33
34	$CH{=}CH_2$	0.38	0.00	0.00	9.79	0.07
35	$COCH_3$	−1.13	1.00	0.00	10.29	0.32
36	$OCOCH_3$	−0.72	1.00	0.00	11.85	0.41
37	CO_2CH_3	−0.72	1.00	0.00	11.85	0.33
38	$NHCOCH_3$	−1.94	1.00	1.00	13.71	0.28
39	$C{=}O(NHCH_3)$	−1.94	1.00	1.00	13.39	0.34
40	CH_2CH_3	1.43	0.00	0.00	10.30	−0.05
41	OCH_2CH_3	−0.51	1.00	0.00	11.93	0.22
42	CH_2OCH_3	−0.23	1.00	0.00	12.07	0.01
43	SOC_2H_5	−1.70	1.00	0.00	18.35	0.52
44	SC_2H_5	0.52	0.00	0.00	17.93	0.23
45	$CH_2Si(CH_3)_3$	3.62	0.00	0.00	29.61	−0.15
46	NHC_2H_5	−0.84	1.00	1.00	13.76	−0.11
47	$N(CH_3)_2$	−0.64	1.00	0.00	14.11	0.10
48	$CH{=}CHCN$	−0.74	1.00	0.00	15.33	0.26
49	$c\text{-}C_3H_6$	1.49	0.00	0.00	13.53	−0.03
50	COC_2H_5	−0.59	1.00	0.00	14.65	0.32
51	$CO_2C_2H_5$	−0.18	1.00	0.00	16.76	0.33
52	$OCOC_2H_5$	−0.18	1.00	0.00	17.12	0.41
53	$CH_2CH_2CO_2H$	−0.03	1.00	1.00	16.52	−0.02
54	$NHCO_2C_2H_5$	−1.40	1.00	1.00	19.96	0.14
55	$CONHC_2H_5$	−1.40	1.00	1.00	18.04	0.34
56	$NHCOC_2H_5$	−1.40	1.00	1.00	18.36	0.28
57	$CH(CH_3)_2$	1.84	0.00	0.00	14.96	−0.05
58	C_3H_7	1.97	0.00	0.00	14.96	−0.06
59	$OCH(CH_3)_2$	−0.10	1.00	0.00	16.52	0.30
60	OC_3H_7	0.03	1.00	0.00	16.52	0.22
61	$CH_2OC_2H_5$	0.03	1.00	0.00	16.72	0.01
62	SOC_3H_7	−1.16	1.00	0.00	23.00	0.52

编号	取代基	Fr	H_A	H_D	MR	F
63	SC_3H_7	1.06	0.00	0.00	22.58	0.23
64	NHC_3H_7	−0.30	1.00	1.00	18.41	−0.11
65	$Si(CH_3)_3$	2.96	0.00	0.00	24.96	−0.04
66		1.58	0.00	0.00	24.04	0.10
67		1.58	0.00	0.00	24.04	0.04
68	$CH=CHCOCH_3$	−0.13	1.00	0.00	19.92	0.28
69	$CH=CHCO_2CH_3$	0.28	1.00	0.00	21.85	0.24
70	COC_3H7	−0.05	1.00	0.00	19.30	0.32
71	$OCOC_3H_7$	0.36	1.00	0.00	21.77	0.41
72	$CO_2C_3H_7$	0.36	1.00	0.00	21.46	0.33
73	$CH_2CH_2CH_2CO_2H$	0.51	1.00	1.00	21.17	−0.02
74	$NHCOC_3H_7$	−0.86	1.00	1.00	23.01	0.28
75	$CONHC_3H_7$	−0.86	1.00	1.00	22.69	0.34
76	C_4H_9	2.51	0.00	0.00	19.61	−0.06
77	$C(CH_3)_3$	2.22	0.00	0.00	19.62	0.07
78	OC_4H_9	0.57	1.00	0.00	21.12	0.25
79	$CH_2OC_3H_7$	0.57	1.00	0.00	21.37	0.01
80	NHC_4H_9	0.24	1.00	1.00	23.06	−0.28
81	$N(C_2H_5)_2$	0.16	1.00	0.00	23.44	0.01
82	$CH=CHCOC_2H_5$	0.41	1.00	0.00	24.57	0.28
83	$CH=CHCO_2C_2H_5$	0.82	1.00	0.00	26.03	0.24
84	C_5H_{11}	3.10	0.00	0.00	24.26	−0.06
85	$CH_2OC_4H_9$	1.11	1.00	0.00	26.02	0.01
86	C_6H_5	1.90	0.00	0.00	25.36	0.08
87	OC_6H_5	1.22	1.00	0.00	27.02	0.34
88	$SO_2C_6H_5$	−0.39	1.00	0.00	33.20	0.56
89	NHC_6H_5	0.75	1.00	1.00	28.50	−0.02
90		1.78	1.00	0.00	38.88	0.25
91	$CH=CHCOC_3H_7$	0.95	1.00	0.00	29.22	0.28
92	$CH=CHCO_2C_3H_7$	1.36	1.00	0.00	26.50	0.24
93	COC_6H_5	0.69	1.00	0.00	29.96	0.30
94	$CO_2C_6H_5$	0.60	1.00	0.00	31.60	0.33
95	$OCOC_6H_5$	1.22	1.00	0.00	32.33	0.23

编号	取代基	Fr	H_A	H_D	MR	F
96	$NHCOC_6H_5$	−0.03	1.00	1.00	34.28	0.09
97	$CH_2C_6H_5$	2.44	0.00	1.00	30.01	−0.08
98	$CH_2OC_6H_5$	1.71	1.00	0.00	31.77	0.02
99	$CH_2Si(C_2H_5)_3$	4.82	0.00	0.00	43.56	−0.15
100	$CH=CHC_6H_5$(反式)	2.72	0.00	0.00	32.97	0.06
101	$CH=CHCOC_6H_5$	1.81	1.00	0.00	39.05	0.22
102	2-茂络铁基	2.43	0.00	0.00	48.24	−0.15
103	$N(C_6H_5)_2$	2.43	1.00	0.00	53.25	0.07

一、原始数据矩阵的标准化

Fr, H_A, H_D, MR, 和F 5种取代基参数的含义不同，度量标准不同。例如表征取代基立体效应的克分子折射MR其数值较大，而反映亲脂性程度的Fr、电性参数F等绝对值较小（比如$CH_2OC_2H_5$的MR=16.72, Fr=0.03, F=0.01）。若直接用原始数据进行信息群分就会突出了MR的作用而压低了Fr和F等的作用，故需对原始数据进行标准化变换。

变换采用式(2)[7]：

$$Z_j = \frac{x_{ij} - \bar{x}_j}{S_j} \qquad (2)$$

其中

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \qquad S_j = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2\right]^{\frac{1}{2}}$$

式中，x_{ij} 为第i个取代基的第j个变量的观测值，而 Z_{ij} 为其标准化数值；$\bar{x}_j$ 为第j个变量的均值；S_j 为第j个变量的标准差。变换后的标准化数据矩阵为Z数据。经过标准化变换后，每个化学结构参数的均值都为0，标准差都为1，使各参数都处于同一量度（比如上面提到的 $CH_2OC_2H_5$，标准化后 Z_{MR}=-0.2061, Z_{Fr}=-0.2121, Z_F=0.5914）。

二、计算分类统计量

对变量进行R型信息群分时，采用变量间相关系数 r_{jk} 作为相似性分类统计量，计算公式如(3)所示[7,8]：

$$r_{jk} = \frac{1}{n-1}\sum_{i=1}^{n} Z_{ij} Z_{ik}, \quad (j,k = 1,2,\cdots,5) \qquad (3)$$

$-1 \leqslant r_{jk} \leqslant 1$，当 r_{jk} 愈接近1说明变量 x_j 和 x_k 关系愈密切，$r_{jk}=1$，说明 x_j 和 x_k 完全相关。将所有变量两两间相关系数都算出来后，便得到5×5实对称矩阵R：

$$R = \begin{pmatrix} 1.0000 & -0.5841 & -0.4289 & 0.5972 & -0.3821 \\ & 1.0000 & 0.2985 & 0.0566 & 0.2767 \\ & & 1.0000 & -0.1874 & -0.1636 \\ & & & 1.0000 & -0.2024 \\ & & & & 1.0000 \end{pmatrix} \qquad (4)$$

对取代基进行Q型信息群分时，把每个取代基看成五维空间中的一个向量。对于任何两个取代基 x_i 和 x_l 的相似程度可用这两个向量的夹角余弦（相似系数）$\cos\theta_{jl}$ 来表示[7-9]：

$$q_{il}=\cos\theta_{jl}=\frac{\sum_{j=1}^{5} Z_{ij}\,Z_{lj}}{\sqrt{\sum_{j=1}^{5} Z_{ij}^2\ \sum_{j=1}^{5} Z_{lj}^2}} \qquad (i,l=1,2,\cdots,103;\ i\neq l) \qquad (5)$$

$-1\leqslant\cos\theta_{jl}\leqslant 1$，当$\cos\theta_{jl}$愈接近1说明取代基 x_j 和 x_l愈相似，$\cos\theta_{lj}=1$，说明 x_j和x_l完全相似。将所有取代基两两间相似系数都算出来后，便得到103×103实对称矩阵Q（数据略）。

三、进行信息群分

现以R型逐次信息群分为例，概述其具体步骤：

1．从相关矩阵R中选出相关系数最大的$r_{14}=0.5972$，根据 Shannon[5,6] 方程式（6）计算Z矩阵第1列与第4列的信息量 I_1和 I_4。

$$I_i=-\sum_{l=1}^{m}\frac{S_l}{n}\log_2\frac{S_l}{n} \qquad (6)$$

式中，n 为取代基数目（103）；m为区间分组数，第 l 个小区间$\triangle l$ 为 $\left[a+\frac{(b-a)(l-1)}{m},\ a+\frac{(b-a)l}{m}\right]$，（$l=1,2,\cdots,m$）；a，b 分别是第 j 个和第 k 个结构参数的最小值和最大值；S_l 为第 l 组中出现的取代基的个数；S_l/n 为 n 个取代基在第 l 组中出现的几率。

标准化数据矩阵Z中，某列的信息量愈大，则表示103个取代基在该结构参数中数据分布愈均匀愈广泛，取代基跨度空间愈大；反之，某列信息量愈小，则数据分布愈密集，结构变异范围愈狭窄。所以简化矩阵时弃去信息量小的变量，保留信息量大的变量，再作下一步聚类。

经过计算，$I_1=5.44$，$I_4=5.16$，$I_1>I_4$，则去掉Z矩阵中信息量小的第4列元素，信息量大的第1列各元素按式（7）计算，接着去掉R阵中第4行第4列各元素，而R阵第1行第1列各元素按本节第3段重算。这样，R阵就由5×5实对称方阵简化成4×4实对称方阵R[10]，变量F_r和 MR 合并成一个新的变量 F_r'。

2．将 $Z=(Z_{jj})$（$i,j=1,2,\cdots,n$）中第 j 列与第k列按信息权重原则合并得：

$$Z_{ia}'=\frac{n_jZ_{ij}+n_kZ_{ik}}{n_j+n_k} \qquad (i=1,2,\cdots,103\quad j,k=1,2,\cdots,5) \qquad (7)$$

其中
$$a=\begin{cases} j & \text{当}\ I_j>I_k\ \text{时} \\ k & \text{当}\ I_k>I_j\ \text{时} \\ e & \text{当}\ I_j=I_k\ \text{时,}\end{cases} \qquad e=\min\{j,k\}$$

式中，Z_{ja}' 表示合并后的新变量标准化数值；n_j，n_k为权重系数，即分别表示变量 x_j和 x_k已经过组合过的次数。

3．重新计算组合后的新变量F_r'与其余变量（H_A，H_D，F）的相关系数，其余不变，得$R^{(1)}$阵。

4．从$R^{(1)}$阵中记下$r_{23}=\max\{r_{jk}\}$，按 Shannon 方程算得$I_2=0.92$，$I_3=0.78$，$I_2>I_3$，去掉Z阵中信息量小的第3列各元素，其后计算同前，如此循序重复，直至把所有的变量归成一个总体系为止。

5．用类似于R型信息群分的方法对103个取代基进行Q型逐次信息群分。所不同的只是采用相似系数$\cos\theta_{jl}$作为分类统计量对取代基分群。

结　果

一、R型逐次信息群分结果如表2所示。

表2　103个脂肪取代基的R型信息群分结果

群分步骤	连结化学结构参数		信息量	相关系数
1	F_r	MR	$I_1=5.44$ $I_4=5.16$	$r_{14}=0.5972$
2	H_A	H_D	$I_2=0.92$ $I_3=0.78$	$r_{23}=0.2985$
3	H_A，H_D	F	$I_2'=1.63$ $I_5=5.28$	$r^*_{1'5}=0.0566$
4	F_r，MR	H_A，H_D，F	* *	$r^*_{1'5'}=-0.2881$

* 加撇号表示组合后的新行（列）

** 最后一步不需要计算信息量

二、根据上述结果，画出R型逐次信息群分图（见图1）。

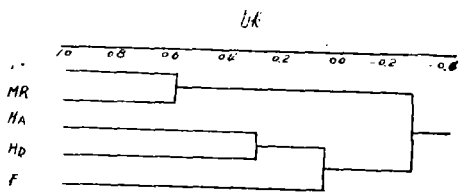


图1　103个脂肪取代基R型逐次信息群分树叉图

三、对取代基进行Q型逐次信息群分，结果见表3。计算信息量时需将式（6）中的n换成P，计算信息权重组合值时亦需对式（7）作相应变换，即将R型分析中的列换成Q型分析中的行。

$\cos\theta_{jl}$在0.410水平上可将103个取代基分成5个群；在0.696水平上可分成10个群；在0.858水平上可分成20群。

四、根据Q型逐次信息群分结果，画出103个取代基的群分树叉图（见图2）。

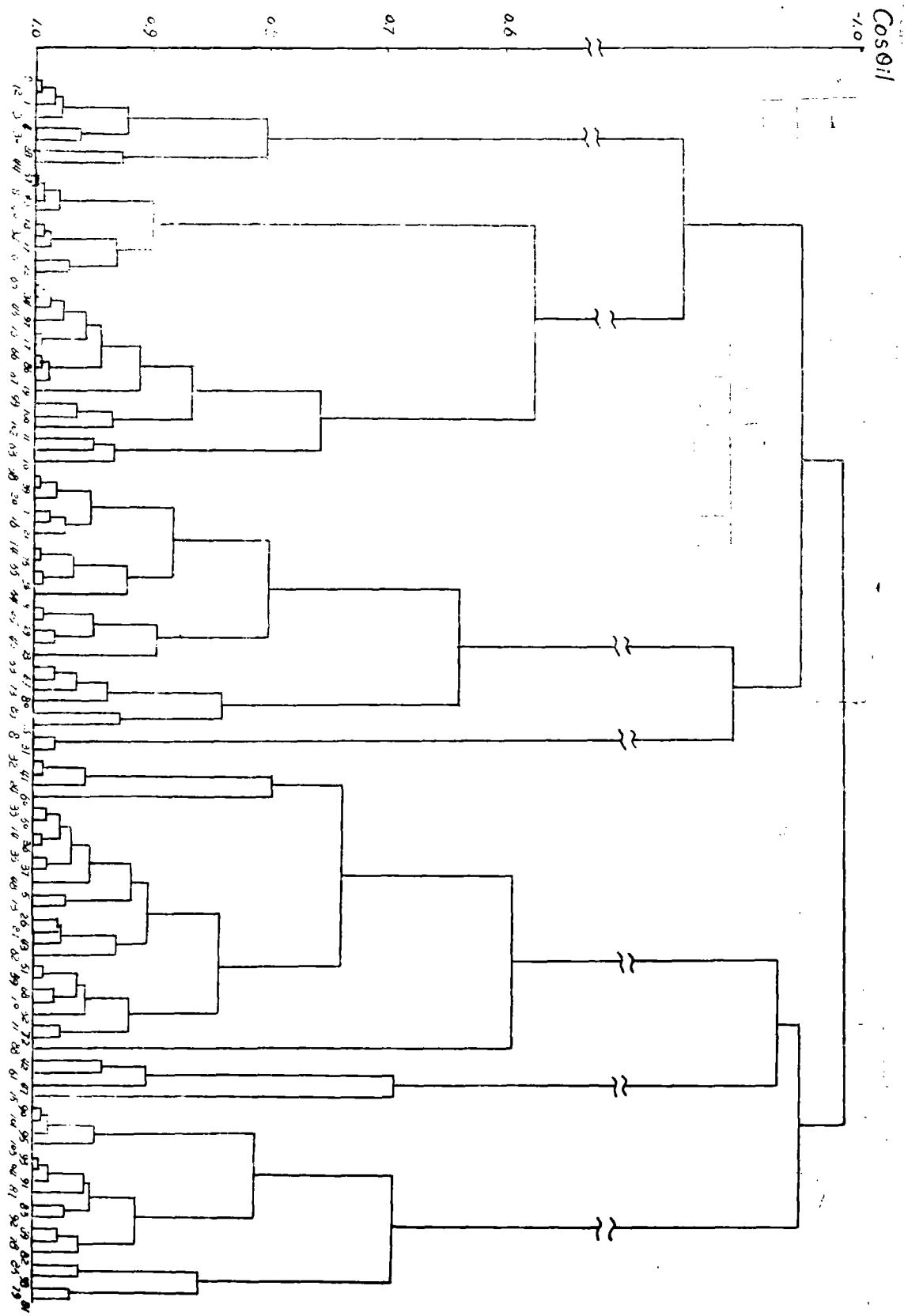
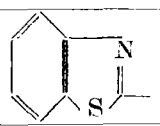


图 2　103个脂肪取代苯的Q型逐步聚合聚类分图

表 3　103 个脂肪取代基 Q 型逐次信息群分表

5 群法	10群法	20群法	取　代　基
1	1	1	Cl, CF_3, Br, F, I, CF_2CF_3
		2	SCH_3, FC_2H_5
	2	3	$CH(CH_3)_2$, C_3H_7, ▽, CH_2CH_3, CH_2Cl, $CH=CH_2$, CH_2Br, H, CH_3
	3	4	$Si(CH_3)_3$, C_5H_{11}, $CH_2Si(CH_3)_3$, $CH_2C_6H_5$, C_4H_9, $C(CH_3)_3$,  , C_6H_5,  , CH_2I, $CH_2Si(C_2H_5)_3$, $CH=CHC_6H_5$(反式), 2-茂络铁基
		5	CCl_3, SC_3H_7, CBr_3
2	4	6	$NHCOCH_3$, $C=O(NHCH_3)$, $CONH_2$, OH, CO_2H, $CH=NOH$, $NHCOC_3H_7$, $CCNHC_3H_7$, $CONHC_2H_5$, $NHCOC_2H_5$, $NHCO_2C_2H_5$
		7	NH_2, CH_2OH, $NHCH_3$, NHC_2H_5, $NHCONH_2$
	5	8	$CH_2CH_2CO_2H$, NHC_3H_7, $C(CH_2)_3CO_2H$, NHC_4H_9
		9	NHC_6H_5, $NHCOC_6H_5$
	6	10	SH, $C\equiv CH$
3	7	11	CH_2CN, OCH_2CH_3, OCH_3
		12	OC_3H_7
		13	$CH=CHNO_2$(反式), CCC_2H_5, SCN, $OCOCH_3$, $COCH_3$, CO_2CH_3, $CH=CHCN$, NO_2, CN, $SOCH_3$, OSO_2CH_3, SCC_2H_5, SCC_3H_7
		14	$CO_2C_2H_5$, $CCH(CH_3)_2$, $CH=CHCCCH_3$, COC_3H_7, $CCOC_2H_5$, $OCCC_3H_7$, $CO_2C_3H_7$
	8	15	$SO_2C_6H_5$
4	9	16	CH_2CCH_3, $CH_2OC_2H_5$, $N(CH_3)_2$
		17	CO_2^-
5	10	18	 , $CH=CHCCC_6H_5$, $OCCC_6H_5$, $N(C_6H_5)_2$
		19	COC_6H_5, $CO_2C_6H_5$, $CH=CHCCC_3H_7$, OC_6H_5, $CH=CHCO_2C_2H_5$ $CH=CHCO_2C_3H_7$, $CH=CHCO_2CH_3$, OC_4H_9, $CH=CHCOC_2H_5$
		20	$CH_2OC_4H_9$, $CH_2OC_6H_5$, $CH_2OC_3H_7$, $N(C_2H_5)_2$

讨　论

一、R型信息群分结果揭示了5种结构参数之间的内在关系。F_r和MR为首次组合参数，虽然它们的物理属性不同，但它们都具有“加合构成性”，其数值大小随基团类型、原子种类、连结方式不同而变化。因此F_r和MR间呈现一定的正变依赖关系。H_A和H_D同为一群，因为它们都表征了取代基的氢键特征，而这两个参数和场效应参数F又隶属于一个大群，显然H_A、H_D、F都属电性效应范畴，这和实际情况吻合。在进行系列设计时，选用的取代基要使它们化学结构参数间彼此相关性尽可能小。

二、Q型逐次信息群分表给出了3种群分结果，即5群法、10群法和20群法。事实上可以人为任意分成若干群。在对先导化合物进行系列设计时，可试行按群选择合成对象。通常应注意两点[2]：第一，所选取代基各参数应基本独立；第二，所选取代基要从药化、生化、药理学等角度出发，首先应是合成上可行的（包括工艺路线、原料、试剂等），其次还应结合体内生化机理、转运、代谢等。例如，$NHCOCH_3$，$CONH_2$，OH和COOH同为第6群，它们都具有形成氢键的能力，但是COOH在生理条件下就已经解离，而OH无此性质，$NHCOCH_3$在体内易受酰化酶的进攻而脱去乙酰基，生成的氨基一方面易被氧化，另一方面又易被可逆酰化。这样，不同药物类型以及不同受体部位即使理化综合因素相似(同一群)，其体内活性也可能有相当大的起落变化甚至完全相反。再如，NO_2，CN，$COCH_3$，SOC_2H_5等，按经典药化概念它们是很不相似的，然而却为同群（13群），而NO_2在体内比CN易于还原，其它取代基也无此性质。选择合成对象时就要全面分析，尤其要结合生化药理机理，最终结果是由实验加以说明的。

三、如果样本相同，所选的变量不同，则分群结果完全不同；即使变量也相同，但采用的相似性统计量或不同的群分方法，所得结果也是有差异的[9]。所以，如果影响活性的主要参数不是某群分表中的结构参数，应自行编制群分法。本文样本和变量均取自Hansch数据库，由于本文用$\cos\theta_{il}$为相似性统计量，而Hansch用欧氏距离；本文用Shannon公式算得的信息量作矩阵收缩的客观尺度，而Hansch用略去编号大的行列法，因此本文编制的信息群分表与Hansch的聚类表既有相同的地方（如本文的第1，7，10，19，20群与Hansch的第1，5，6，10，15群是相同的）又有相异的地方（其余一些群）。从矩阵收缩方法以及药化等概念出发，本文信息群分表似较合理。

四、信息群分是寻找客观分类的方法，而判别分析是在事先知道各类母体情况下寻找客观分类的判据，若各类母体情况不十分清楚，可先用本文方法进行聚类，建立判别函数，再对新样本进行判别[9]。对先导化合物进行系列设计时，信息群分的客观检验效果是精确的药理活性指标，因此将判别分析与它联合使用，将对合成工作提供借鉴。如将因子分析和主成分分析与信息群分联用，可使SAR建立在比较合理的基础上，其结果的预示性可能更强。聚类分析是数值分类学的一个新分支，它的历史不长，不论是聚类计量还是聚类方法本身还存在若干问题[7]，尽管它在气象、地质、农业、生物分类等方面应用中取得了不少成绩。但是它的不唯一性暴露出不少弱点，且建立不起来标准误或可信限，对任一特定群也无法进行显著性检验，因而对该法有褒有贬。本文信息群分法就是对聚类分析进行探讨的一次尝试。

安登魁副教授、程景才同志对本文提出许多宝贵意见，谨此致谢。

参 考 文 献

〔1〕 Hansch C, et al: J Med Chem **16**(11):1217, 1973

〔2〕 Hansch C, et al: 《Substituent Constants for Correlation Analysis in Chemistry and Biology》 p48, John Wiley & Sons, New York, 1979

〔3〕 Martin YC, et al: J Med Chem **22**(7):784, 1979

〔4〕 Topliss JG, et al: J Med Chem **22**(10):1238, 1979

〔5〕 中国科学院计算中心概率统计组译：数字计算机上用的数学方法 第三卷 220页，上海科学技术出版社，上海，1981

〔6〕 De Clercq H, et al: J Pharm Sci **66**(9):1269, 1977

〔7〕 丁士晟编著：多元分析方法及其应用 315页，362页
吉林人民出版社，长春，1981

〔8〕 成都地质学院概率论与数理统计编写小组：概率论与数理统计 296页，地质出版社，北京，1981

〔9〕 方开泰：数学的实践与认识 (1):64, 1977; (2):54, 1978

CLUSTER INFORMATION ANALYSIS OF 103 ALIPHATIC SUBSTITUENTS

Wang Erhua[1] and Peng Sixun[1]

Abstract

In design of lead compounds, the rational choice of a few substituents without intensive collinearities among the chemical structural parameters for an initial synthesis is very crucial for obtaining the maximum of information of SAR. 103 Common aliphatic substituents have been successively clustered into 5, 10 and 20 clusters with respect to various physicochemical parameters of substituents such as F_r, H_A, H_D, MR and F by cluster information analysis, the dendrogram classification being given. Thus, selection and preparation of derivative from each cluster based on bioorganic reaction mechanism, drug metabolism, or ease of synthesis will give a maximum range in parameters and help to establish SAR equation more rapidly. The information contents calculated from Shannon's equation for chemical structural parameters are used as the criterion of matrix contraction.

Key words Cluster information analysis, Cluster analysis, 103 Aliphatic substituents

1. Division of Medicinal Chemistry