2004 年第 62 卷
第 17 期, 1585～1589

化 学 学 报
ACTA CHIMICA SINICA

Vol. 62, 2004
No. 17, 1585～1589

R-基团相似性的比较

朱 倩　姚建华*　李 丰　陈海峰　袁身刚*

(中国科学院上海有机化学研究所计算机化学实验室　上海 200032)

摘要　介绍了对化学取代基(R-基团)进行相似性比较的工作. 每个 R-基团主要依赖一些对反应影响较为明显,同时又便于计算的结构描述符来进行描述,例如电负性、氢键受体、氢键给体等参数. 由于与反应核心相距过远的环境对反应的影响较小,R-基团相似性比较对象,就由每个 R-基团截取从它与母体连接的位点出发向外扩展 6 层的子结构组成. 在确立了对 R-基团描述和距离限制的基础上,提出了用一个 45 维向量表示一个 R-基团,并由化学结构的比较,转化为向量比较来实现 R-基团相似性比较的方法. 采用这一方法,成功地对大量的 R-基团进行了相似性的区分,并实现了对未知化合物进行相似性预测的目标.

关键词　相似性,描述符,归一化,重心

Comparison of the R-Group Similarity

ZHU, Qian　YAO, Jian-Hua*　LI, Feng　CHEN, Hai-Feng　YUAN, Shen-Gang

(Laboratory of Computer Chemistry, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai 200032)

Abstract　This work is mainly concerned with the comparison of the similarities of the R-groups. Each R-group is based on several descriptors, such as electronegativity, hydrogen-bond acceptor, hydrogen-bond donor *etc*., which will affect reactions obviously and can be calculated conveniently. Since the environment with very long distance from the reaction core has relatively small effect on the reaction, the comparison partners of R-group can be regarded as some substructures that have through-bond of up to six from the point of substitution on the reaction core. This work proposed a method based on the description of R-group and the constraints of distance to compare the R-group similarities, which represented an R-group with a 45 dimensional vector and converted the comparison of chemical structures into the corresponding multiple vectors. Employing this method, we successfully distinguished a large number of R-group similarities, and realized the goal of predicating the similarity of unknown chemical compounds.

Keywords　similarity, descriptor, normalization, barycenter

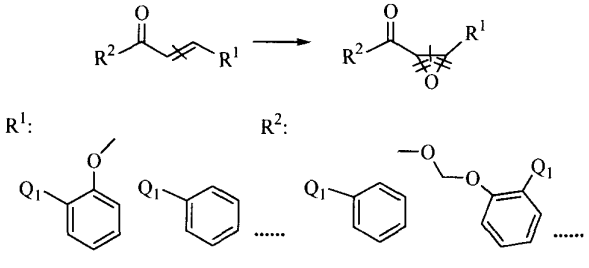
一个化学数据库检索系统,如果没有相似检索的功能,该系统的功能就不可能很完全. 因为不管该数据库系统所涵盖的数据有多丰富,实验室所合成的新化合物就很可能不在其中,因而对与之相近或相似的化合物检索是很重要的[1]. 因此,为了能让化学家通过数据库的检索,最终搜索到尽可能多的与提问结构(或反应)相同或相似的信息,只有在数据库中实现结构相似性比较的功能. 近年来科学家们为解决这一问题做了相当多的努力[2,3]. 我们课题组也开展了这方面的研究,该工作主要建立在一个同类反应知识库的基础上[4],每个反应的知识都是由反应核心和 R-基团构成,如图式 1.

如图式 1 所示,要使系统在建成后具有相似性检索的功能,在反应核心(除去 R 原子)精确匹配成功的同时,R-基团相似性比较的工作是必不可少的.

* E-mail: yaojh@mail.sioc.ac.cn; Tel: +86-021-64163300-2888; Fax: +86-21-64166128.

Received February 9, 2004; revised April 20, 2004; accepted May 10, 2004.

本工作由 2003CB114401,2002AA231011,02DJ14013,CNRS/CAS14916 和"基于现代理论和技术的复方中药系统研究"五个项目所组成,它们分别得到中国科学技术部,上海市科学技术委员会和中国科学院的资助.



图式 1 Sharpless 环氧化同类反应的一个实例

Scheme 1 One of the Sharpless epoxidation generic reactions

Holliday 等[3]最近应用 R-基团描述符研究了结构 - 生物活性关系中取代基之间的相似性. 他们选取了 7 种原子性质作为 R-基团的描述符,同时以基团中连接母体(环)的原子为起点 6 根键以内的原子依此分为 6 层,每 1 层内所有原子的性质之和作为该层的描述符,没有原子的层则用零填充. 这样每个基团被用最多可达 42 维的矢量所描述. 基于这样定义的描述符可以方便地计算 R-基团间的相似性. 他们的这一工作对如何研究取代基间的相似性很有启发. 但是,由于他们的研究兴趣主要是取代基团的生物等排和非等排的相似性,与我们的研究课题有很大的不同. 应该说取代基团对反应的影响要比生物等排性更复杂,因为这种影响通常用各种效应来解释,最常用的有:电子效应、立体效应、诱导效应、共振效应、极化度、超共轭、氢键等. 因此,我们的工作借鉴 Holliday 等[3]的方法,定义一套适用于描述对反应影响的基团描述符,以期得到中肯的 R-基团相似性度量.

1 结构描述符的确定

由于 R-基团中离连接反应中心处过远的环境对反应影响较小,因此为了使模型更简洁,我们仅将从 R-基团连接点出发向外扩展 6 层拓扑距离内的结构考虑为对反应有影响的部分,称为活泼部分,并用于相似性的计算.

我们从 9.2×10^5 个 MDL 的反应数据[5]中随机地选取了 8058 个反应,按照一定的规则[4]截取到了 4341 个互不相同的 R-基团,组成了研究数据集. 通过分析数据集中基团间的异同点,并综合考虑了有机反应所共有的特性,我们主要选择了一些结构片断中对反应影响较为明显同时又便于计算的结构参数作为 R-基团描述符. 例如,原子质量(含氢原子)、电负性[6]、环数、芳香性、氢键受体、氢键给体、价键形式. 我们获得的 R-基团是从每个原始反应中截去核心部分后剩余的基团[4],所以这些 R-基团可能存在一个或多个的连接位点,也可能为环型和非环型两种形式,而且其连接点也可能呈现出不同的杂化形式,如图 1 所示. 这些不同虽然在我们确立的结构参数中已能得以区分,但由于它们的不同将直接导致 R-基团所具有化学意义的不同. 为了更着重地体现出它们之间的差别,我们添加了 3 个针对连接位点(Q)的描述符:Q 的个数、Q 成环与否和 Q 的杂化度对此加以区别.

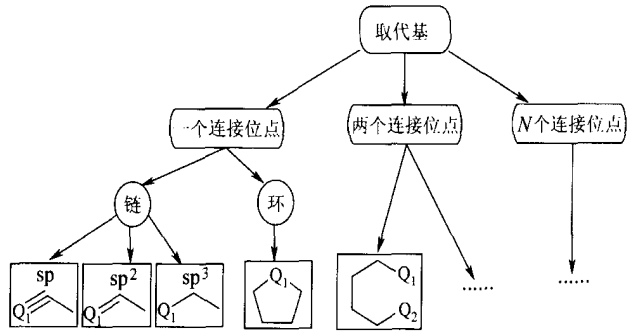


图 1 R-基团的分类

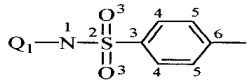
Figure 1 Classification of the R-groups

另外,我们还规定了 R-基团中每一层原子对应特定性质的描述,为该层所有原子被该性质描述的累加.

1.1 结构描述符的表述

数据集中的每个 R-基团都用前面定义好的 6 层环境和 10 个描述符去进行描述,由于针对连接位点的描述符是独立于 6 层环境存在的,所以最后每个 R-基团就被表示成 45 维($7\times6+3$)空间中的一个点. 那么相似性的比较就转化成 45 维空间中两点间距离的比较了,如果其中两个点间的距离相比其他的距离小,即可认为这两个点所代表的两个 R-基团较其他的基团更相似.

首先用定义好的结构描述符来分别描述数据集中的所有 R-基团,继而进行 R-基团相似性的比较. 以图式 2 中所示的 R-基团为例,对描述符的表示加以说明.



图式 2 一个 R-基团例子

Scheme 2 One of the R-groups

图式 2 中 R-基团上相应的数字分别表示了从连接位点(Q_1)出发,以简单的拓扑距离(1 根化学键)扩展得到的 1～6 层化学环境. 现在只需对每一层环境中的原子用描述符进行描述,表 1 是 10 个描述符描述该 R-基团的最终结果.

1.2 描述符的归一化

在比较相似性之前,应尽可能减小分子大小的影响,使不同 R-基团描述符对基团的贡献趋于均衡,我们对此进行了描述符的归一化(normalization)操作.

所有需要进行相似性比较的 R-基团都存在于由 10 个描述符和 6 层环境组成的 45 维空间里,组成空间的每一维向量的数值变化范围都不相同,因此为了使它们对 R-基团的相似性比较的贡献基本相同,而对它们进行归一化处理,使每一维的标准离差等于 1,平均值等于 0. 于是我们将整个数据集中的每一维向量进行了归一化.

表 1　描述 R-基团的结构描述符[a]

Table 1　Structure descriptors to describe the R-groups

描述符		层次					
		1	2	3	4	5	6
原子质量	每个原子在周期表中查得	15.01	32.07	44.01	26.03	26.04	12.01
电负性	查表获得[6]	3.05	2.6	8.7	5.2	5.2	2.6
环数	被描述原子所在环数	0	0	6	12	12	6
芳香性	被描述原子在芳环上返回 1,否则为 0	0	0	1	2	2	1
氢键受体	O, N, S,卤素等	1	1	2	0	0	0
氢键给体	O, N, S等上含有活泼氢	1	0	0	0	0	0
价键形式	包括单、双、叁和芳香键	1	2	8	8	8	4
Q 的个数	R-基团连接位点的个数	1					
Q 成环否	R-基团连接位点在环上,N 元环返回 N ($N \geq 3$)	0					
Q 杂化度	R-基团连接位点的杂化度(sp^3, sp^2, sp)	1					

[a] 价键形式中的单、双、叁、芳香键分别用 1, 2, 3, 4 表示. 优先级别为:芳香键 > 叁键 > 双键 > 单键,各原子的价键以其所在价键中优先级别最高的价键为准.

1.3　聚类分析法验证

为了验证结构描述符和归一化方法选择的合理,我们将这 4341 个结构进行聚类分析. 在聚类时,我们将每个结构用 45 (6 ×7 + 3)维向量来进行描述,每维向量在聚类前先进行归一化. 45 维是一个维数很高的空间,由于维数高带来的信息不一定多,为了以后的有效处理,本方法采用主成分分析法(PCA)对原始描述符向量进行特征抽提,由抽提所得的特征组成一个维数较低的特征向量空间,简称特征空间. 这一步主成分数的确定十分重要,具体细节参阅参考文献[7]. 我们采用了比较简单的方法,选取当累积百分数已超过 95 %时的主成分数.

从聚类的结果看,基本上达到我们预期的目的 ——相似的取代基被聚在一类里,如图 2 所示. 这是描述符经过归一化处理后的聚类结果,我们选取了前21个主成分,它们的累

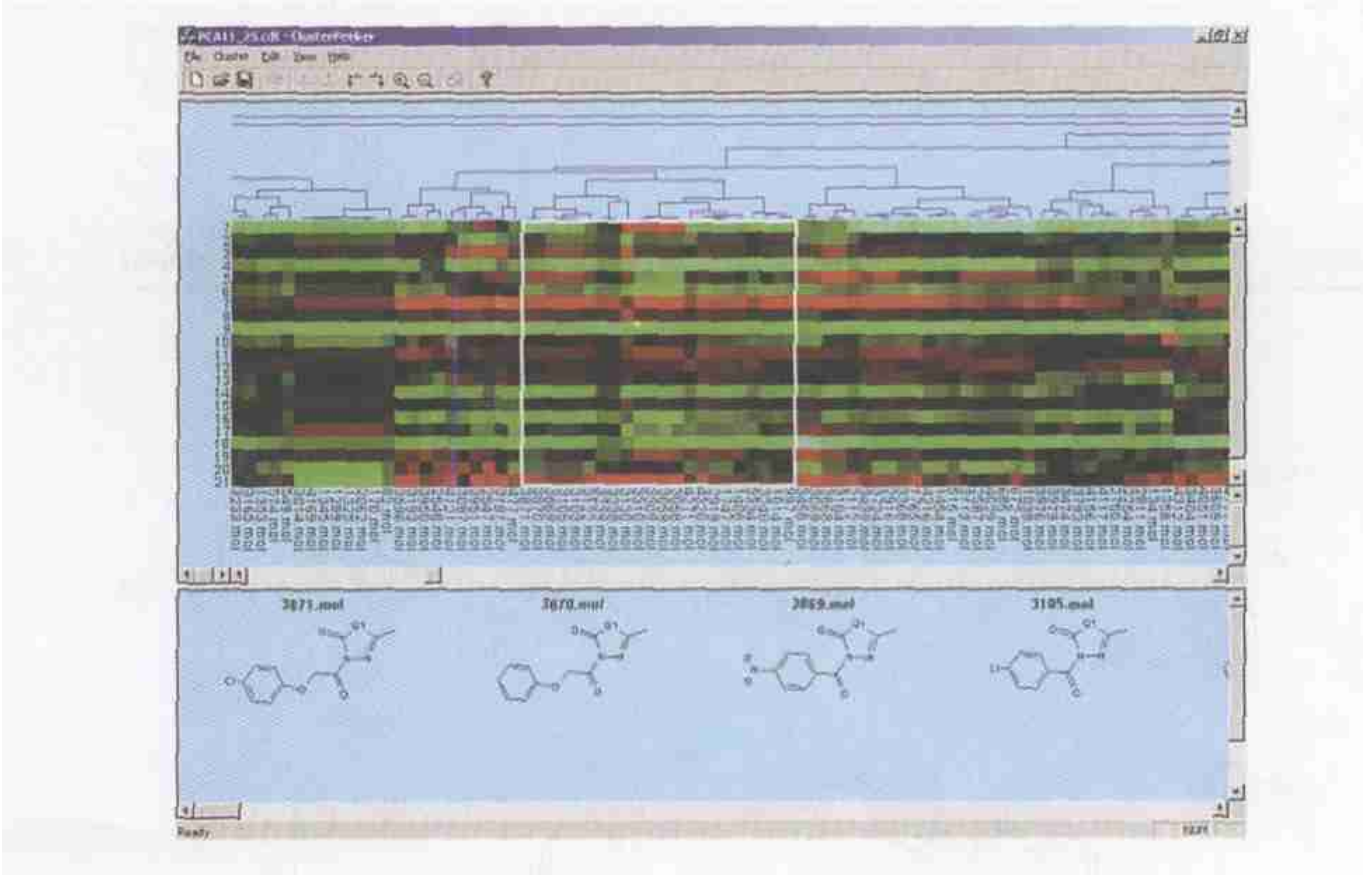


图 2　归一化后的聚类结果

Figure 2　Clustering results after normalization

积百分数为 95.119 %,所用聚类分析程序用我们自己开发的 ReacAnalys[8]. 图中显示了 R-基团的聚类情况,界面上面窗口显示了反应聚类树状图,彩色带表示了被进行聚类的 R-基团,每 1 行表示了主成分分析得到的每 1 主成分,每 1 列表示 1 个 R-基团. 第 i 行和第 j 列方块的颜色表示了第 j 个反应第 i 个主成分上的坐标值. 当用鼠标在彩色带中任意圈定一组反应,它们所对应的产物结构就显示在下面. 图中圈出了聚类树的 1 个小分支(限于篇幅的原因,无法将所有的分支聚类结果都标注出来). 整个聚类结果基本已无不同类别的基团交叉混淆比较的情况出现. 另外还发现每个分支中的基团区分效果非常好,完全达到了由基团结构不同进行区分的目的,其他的分支聚类情况也完全如此,这里就不再赘述了.

1.4 距离计算验证

我们将 4341 个 R-基团表述为 45 维空间中的 4341 个点,分别计算它们两点间的距离,距离计算公式为:

$$\text{Distance} = \left[\sum_{i=1}^{10} \sum_{j=1}^{6} (d_{Aj}^{i} - d_{Bj}^{i})^2 \right] \quad (1)$$

(i:10 个描述符; j:6 层; d_{Aj}^{i}: A 基团中第 j 层原子用第 i 个描述符描述的返回累加值; d_{Bj}^{i}:同理.)

从距离的大小可以看出 2 个结构的相似程度,如表 2 所示. 从表 2 中可以看出,经过归一化计算求得的两结构间的距离,已体现出结构间的相似程度,越相似,距离越小,反之越大. 表中第 1 和第 6 对间差别较大,因为它们是属于不同类别的基团,结构上差异较大;反之,第 2, 3, 4 和 5 之间差别较小,反映出它们之间结构比较相似. 由此可证明基团大分支(类别)的区分在进行完归一化后是得以实现了的.

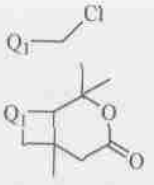
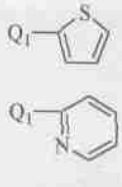
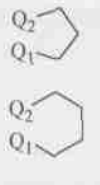
2 R-基团相似性检索

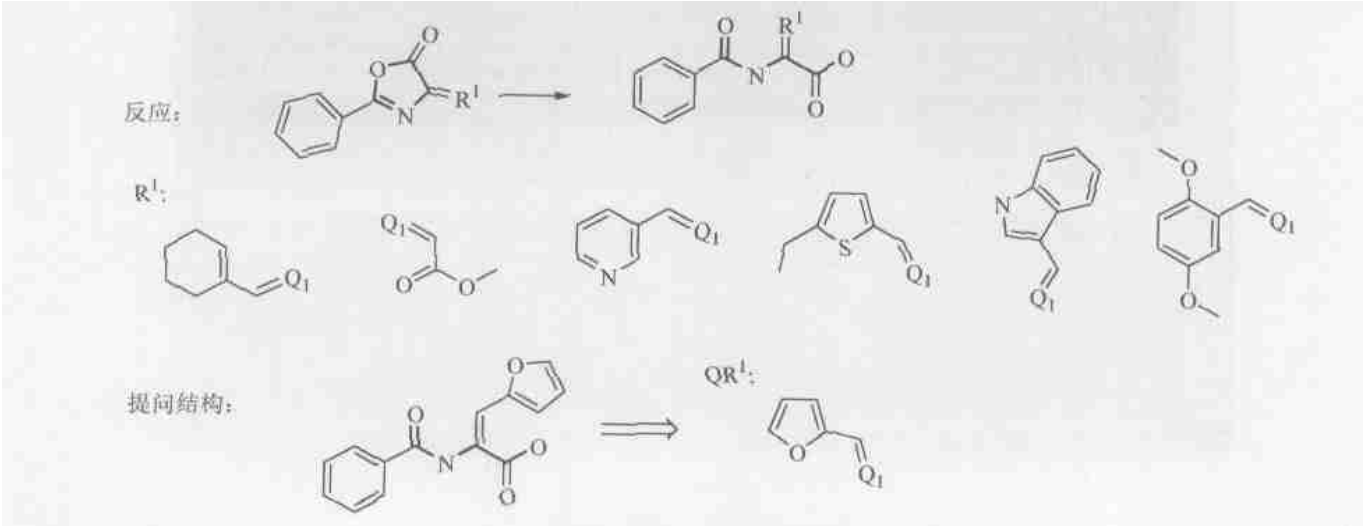
我们希望用这些描述符(表 1)和归一化方法对未知化合物进行相似性预测进而判断反应可发生性,达到相似性检索的目的. 例如,存在图式 3 所示的一同类反应(由反应骨架和对应的取代基构成)和一提问结构.

现在希望预测提问结构是否能按图式 3 中的反应进行合成. 提问结构具有与反应产物完全相同的核心骨架(除产物中的 R-基团),即是需要判断反应产物中 R^1 改变成 QR^1 后,反应是否仍可进行? 要判断反应能否进行,就需进行 R-基团相似性的比较.

表 2 两结构间的距离比较

Table 2 Comparison of the distance between two structures

	1	2	3	4	5	6
结构						
距离	0.13455	0.03405	0.01577	0.04644	0.02213	0.07462



图式 3 待预测的反应和提问结构(黑粗线表示反应中心)

Scheme 3 Forecasted reaction and the query structure (the bold lines show the reaction centers)

表 3 R-基团与重心的距离

Table 3 Distances between one of the R-groups and the weight center

	R-Group1	R-Group2	R-Group3	R-Group4	R-Group5	R-Group6	QR^1-Core
Normalized	1.71920	2.83502	1.18189	1.51774	1.91567	2.59146	1.49712

我们设想用 R^1 集合的重心来代表整个集合,使 QR^1 与集合的比较变为与单个基团的比较,而进一步使整个比较的过程变得简洁和快速了.

2.1 重心的计算

每个 R-基团用 45 维向量来表示,一个 R^1 集合就表示为 45 维空间中多个点的集合,重心可在这样一个 45 维空间中按公式(2)计算获得(n:空间中点的个数;m:各层对应的不同描述符):

$$(X)_m = (x_1 + x_2 + \ldots + x_n)/n \quad (1 < m < 45) \quad (2)$$

2.2 相似性的预测

图式 3 中 R^1 的每个 R-基团及 QR^1 与重心计算获得的距离如表 3 所示. 我们认为如果 R^1 中的 R-基团与重心计算获得的最大距离(Max-distance),最小距离(Min-distance),及 QR^1 与重心的距离(QR^1-core)满足:0 < QR^1-core < Max-distance,那么 QR^1 在一定程度上与 R-基团具有相似性;否则可近似认为 QR^1 与 R-基团不具有相似性,进而能够判断未知化合物对已知反应的可发生性. 从表 3 中的数据可看到经过归一化处理后计算得到的距离满足 Min-distance < QR^1-core < Max-distance 的条件,即可认为提问结构能按图式 3 中的关环反应方式进行合成,文献中也确实有该合成反应的报道[9,10],这样也就证明了比较的正确性.

3 结论

本文介绍了对化学取代基(R-基团)进行相似性比较的工作. 选择了一些对反应影响较为明显,同时又便于计算的结构描述符来对每个 R-基团进行描述,使对化学结构相似性的比较转化为向量的比较. 此外,由于与反应核心相距过远的环境对反应的影响较小,R-基团相似性比较的对象就由每个 R-基团截取从它与母体连接的位点出发向外扩展六层的子结构构成. 于是 10 个选出的描述符和 6 层环境就构成了一个 45 维向量组成的超空间,每个 R-基团就被表示为方便 45 维空间中的一个点,进而将相似性的比较再次转化为 45 维空间中两点间距离的比较了. 这样使一个复杂的化学问题转化成了一个简单的数学问题,使整个比较过程变得更简捷和快速. 采用这一方法,成功地对大量的 R-基团进行了相似性的区分,并实现了对未知化合物进行相似性预测的目标.

References and notes

1 Xu, L.; Hu, C.-Y. *Applied Chemistry Graph Theory*, Science Press, Beijing, **2000** (in Chinese).
(许禄, 胡昌玉, 应用化学图论, 科学出版社, 北京, **2000**.)

2 *Concepts and Applications of Molecular Similarity*, Eds.: Johnson, M. A.; Maggiora, G. M., Wiley, New York, **1990**.

3 Holliday, J. D.; Jelfs, S. P.; Willett, P.; Gedeck, P. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 406.

4 Zhu, Q.; Yao, J.-H.; Li, F.; Chen, H.-F.; Yuan, S.-G. *Acta Chim. Sinica* **2004**, *62*, 112 (in Chinese).
(朱倩, 姚建华, 李丰, 陈海峰, 袁身刚, 化学学报, **2004**, *62*, 112.)

5 MDL Information Systems, Inc. MDL Series Database, http://www.mdli.com.

6 Dean, J. A. *Lange's Handbook of Chemistry*, 13th ed., McGraw-Hill, New York, **1985**.

7 Malinowski, E. R.; Howery, D. G. *Factor Analysis in Chemistry*, Wiley-Interscience, New York, **1980**.

8 ReactAnalys is a program developed by the Key Laboratory of Computer Chemistry of Chinese Academy of Sciences. It is used for cluster analysis of chemical structures and reactions. For more information about this program, you can contact with the authors.

9 Tripathy, P. K.; Mukerjee, A. K. *Synthesis* **1984**, *5*, 418.

10 Gaset, A.; Gorrichon, J. P. *Synth. Commun.* **1982**, *12*, 711.

(A0402092 SHEN, H.; LING, J.)