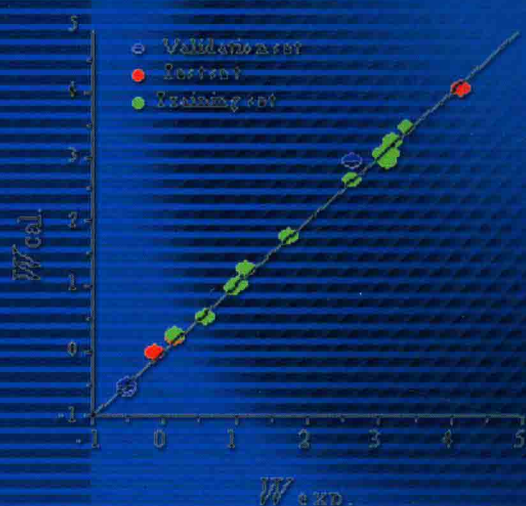


徐州工程学院学术著作出版基金资助

物质构效学应用

冯长君 著



中国矿业大学出版社

著作出版基金资助

物质构效学应用

冯长君 著

中国矿业大学出版社

内 容 提 要

本书对物质构效学的定义、基本原理、基本方法等做了简要阐述,重点讨论了量子化学参数与二维拓扑指数建模(含人工神经网络模型)以及三维构效关系;在二维描述符基础上提出了新型手性拓扑指数,并研究了原子级拓扑指数。本书图文并茂,实例丰富,既可作为高等学校相关专业的教学用书,也可作为从事物质构效学研究的科研人员、教师以及学生的参考用书。

图书在版编目(CIP)数据

物质构效学应用/冯长君著. —徐州:中国矿业大学出版社,2017. 11

ISBN 978 - 7 - 5646 - 3807 - 8

I. ①物… II. ①冯… III. ①生物化学—高等学校—教学参考资料 IV. ①Q5

中国版本图书馆 CIP 数据核字(2017)第 297840 号

书 名 物质构效学应用

著 者 冯长君

责任编辑 孙 浩 齐 畅

责任校对 王慧颖

出版发行 中国矿业大学出版社有限责任公司
(江苏省徐州市解放南路 邮编 221008)

营销热线 (0516)83885307 83884995

出版服务 (0516)83885767 83884920

网 址 <http://www.cumtp.com> E-mail: cumtpvip@cumtp.com

印 刷 江苏凤凰数码印务有限公司

开 本 787×1092 1/16 印张 25 字数 624 千字

版次印次 2017 年 11 月第 1 版 2017 年 11 月第 1 次印刷

定 价 68.80 元

(图书出现印装质量问题,本社负责调换)

前 言

自从 1998 年诺贝尔化学奖授予从事理论化学计算的两位化学家之后,理论与实验并列成为化学的两根支柱。徐光宪院士在 20 世纪末提出的“结构与性能之间关系”是 21 世纪化学的四个“世纪难题”之一,这是涉及化学本质的问题。因此,出版本书的理论价值就是初步构建化学(部分)的分子结构与性能的数学图像,即以数学模型来实现:① 描述结构与性能之间的数量关系;② 探讨影响性能的主要结构基团;③ 推断发挥性能的微观历程。例如:探讨生物作用机理(如抑菌机理、杀虫机理、抗癌机理等)、化学反应机理以及物理变化机理(如色谱分离机理、疏水性作用)等。

通过实验可以获得物质的结构与性能(包括化学性质、物理性质以及生物活性等)的有关物理量。但由于测定耗材费时费力,且化合物数量非常庞大(截至 2014 年底美国化学文摘 CA 登记化合物为 9 108 万种,1900 年仅为 55 万种),不可能实现对每种化合物的所有性能都予以检测。因此,出版本书的应用价值是:① 对未进入数学模型的化合物,可预测其性能,从而避免了繁重的实验和高昂的费用,并减少受试动物的数量及环境污染;② 依据 QSAR 模型推断未知化合物的结构,则是其逆向应用。目前,此逆向应用尚少。

作者相信该部著作的问世,将会进一步完善物质构效学学科内涵,充实学科研究内容,增强学科知识点之间的逻辑性,有力地推动学科方法学发展,也将引起更多学者关注,使之尽快得到公认,早日成为一门真正独立的学科。

作者从事物质构效学研究 20 余载,首先在《化学通报》上发文提出“物质构效学”学科名词。在此领域主持或参与国家自然科学基金、国家重点实验室开放基金、江苏省高校自然科学基金、徐州市科技局计划项目、江苏师范大学科研基金和徐州工程学院科研基金等 20 多项,于国内发表核心期刊 100 余篇,被 SCI、EI 收录论文 43 篇(第一作者或通讯作者)。作者对上述科研成果进行归纳整理,完成理论飞跃,初步形成“物质构效学”理论体系。考虑作者理论水平尚未达到该学科的理论高度,故将此书定名为:《物质构效学应用》。本书分为五编十章。第一编 3D-QSAR 模型与人工神经网络模型:第 1 章 3D-QSAR 模型,第 2 章人工神经网络 QSAR 模型;第二编量化参数 QSAR 模型:第 3 章量化参

数 QSAR 模型;第三编有机物生物活性 QSAR 模型:第 4 章生物富集因子、生物降解性 QSAR 模型,第 5 章抑菌、抑制发芽率等 QSAR 模型,第 6 章其他生物活性 QSAR 模型;第四编有机物 QSPR 模型及手性 QSRR 模型:第 7 章有机物的 QSPR 模型,第 8 章手性有机物 QSRR 模型;第五编无机物 QSPR 模型:第 9 章无机物原子级 QSPR 模型,第 10 章无机物分子级 QSPR 模型(以上五编内容均来自作者已发表的论文,限于篇幅,不再列出参考文献,敬请谅解)。具体涉及二维与三维的构效关系,在二维描述符中还提出新型手性拓扑指数,并研究了原子级拓扑指数。在“绪论”部分对物质构效学的定义、基本原理、基本方法等做了简要阐述。

本书的出版,得益于徐州工程学院学术著作出版基金、徐州工程学院化学化工学院科研基金资助。

出版此书,希望有益于从事物质构效学研究的科研人员、教师以及学生(研究生、本科生)理解该学科的原理和方法,并在实践中综合、创新应用,发挥启示与借鉴作用。然而,作者在编写过程中对学科内涵、方法学的理解深感学识之匮乏,因此,书中之不妥和疏漏在所难免,诚请读者起笔斧正!

冯长君

徐州工程学院化学化工学院

2017 年 9 月 10 日

目 录

0 绪论:物质构效学基础	1
0.1 物质构效学定义	1
0.2 物质构效学基本原理	1
0.3 分子结构特征的数值描述	2
0.4 QSER 模型的类型	8
0.5 2D-QSER 模型建立方法	10
0.6 3D-QSER 模型建立方法	13
0.7 QSER 模型的质量要求	15
0.8 建立多元线性 QSER 模型的一般步骤	16

第一编 3D-QSAR 模型与人工神经网络模型

1 3D-QSAR 模型	21
1.1 N,N-二甲基-2-溴苯乙胺类衍生物对大鼠阻断活性的 CoMFA 研究	21
1.2 喹唑啉衍生物抗癌活性的三维构效关系研究	26
1.3 基于比较分子力场分析方法研究取代喹啉类化合物抗菌活性及分子设计	31
1.4 基于 3D-QSAR 研究硝基芳烃对呆鲦鱼急性毒性	36
1.5 硝基芳烃类含能材料撞击钝度的 CoMFA 模型	40
2 人工神经网络 QSAR 模型	45
2.1 基于多元回归与人工神经网络研究有机物的生物富集因子	45
2.2 用新型连接性指数建立取代苯酚和取代苯甲酸生物降解性的神经网络模型	54
2.3 有机污染物生物富集因子的神经网络模型	60
2.4 部分农药在茶叶上生物降解性的电性距离矢量模型	66
2.5 基于 MOPAC 方法研究新三取代嘧啶苯磺酰脲衍生物的除草活性	69

2.6	基于新连接性指数研究取代芳烃生物降解性	74
2.7	基于人工神经网络 BP 算法研究胍基单胺氧化酶抑制剂抑制活性	78
2.8	QSAR Study of the Action Strength of D_{OM} of Phenyl-isopropyl-amine Dopes Using MLR and BP-ANN	83

第二编 量化参数 QSAR 模型

3	量化参数 QSAR 模型	95
3.1	取代巯基-三唑衍生物抑菌活性的量化模型和结构修饰	95
3.2	QSAR Models of Bioconcentration Factors of Chloroanilines in Fish by Density Functional Theory	102
3.3	基于密度泛函理论研究苯并噻(噁)唑酮衍生物杀虫活性	107
3.4	呋虫胺衍生物杀虫活性的量化模型	115
3.5	采用 MOPAC-AM1 方法研究取代巯基三唑衍生物抑菌活性	119
3.6	马来酸酯类化合物抑菌活性的密度泛函研究	124
3.7	基于密度泛函理论研究卤代苯的生物富集因子	128
3.8	运用密度泛函理论研究新烟碱类杀虫剂生物活性与结构修饰	134
3.9	FQ 均三唑稠杂环类衍生物抗肿瘤活性的理论研究	143

第三编 有机物生物活性 QSAR 模型

4	生物富集因子、生物降解性 QSAR 模型	151
4.1	有机物生物富集因子的原子类型电性拓扑状态指数模型	151
4.2	基于拓扑指数研究有机污染物的生物富集因子	157
4.3	有机物生物富集因子的苯环因子与分子形状指数方程	161
4.4	有机污染物生物富集因子的电性距离矢量方程	171
4.5	Estimation and Prediction of Bioconcentration Factors of Nonionic Organic Chemicals in Fish by Electrotological State Indices and Structural Parameter	176
5	抑菌、抑制发芽率等 QSAR 模型	193
5.1	基于原子类型电性拓扑指数研究新型苯并噻(噁)唑酮衍生物 抑制黄瓜炭疽菌活性	193

5.2	苯甲酰硫脲类化合物抑菌活性的电性拓扑状态模型和结构修饰	198
5.3	运用 QSAR 方法研究取代苯甲酸对植物生长的调节活性	203
5.4	基于拓扑方法研究取代芳烃抑制黄瓜种子发芽率	206
5.5	新颖三唑并嘧啶衍生物抑菌活性的拓扑模型	211
5.6	嘧啶苯磺酰脲衍生物对小麦纹枯病菌体外抑菌活性的 QSAR 研究 与分子设计	215
6	其他生物活性 QSAR 模型	220
6.1	运用拓扑指数研究苯砷基环烷酸酯类对发光菌急性毒性	220
6.2	应用自相关拓扑研究取代芳烃的急性毒性及理化性质	225
6.3	苯酚类化合物对大型蚤急性毒性的形状指数方程	228
6.4	电拓扑方法研究多酚类化合物抗氧化活性	232
6.5	运用 QSAR 方法研究食品香味化合物克拉克值	237
 第四编 有机物的 QSPR 模型及手性 QSRR 模型		
7	有机物的 QSPR 模型	243
7.1	链烷烃沸点和热力学性质的 QSPR 模型	243
7.2	卤代烷标准熵与分子树拓扑指数的数学方程	252
7.3	边价拓扑指数与链烃标准熵	256
7.4	调和拓扑研究链烷烃热力学性质	266
7.5	分子树拓扑指数与羧酸化合物 pKa 值的 QSPR 模型	273
7.6	自相关拓扑研究脂肪族胺醇醚卤代烷的第一电离能	281
7.7	烷烃及其衍生物沸点与距序拓扑指数的数学模型	286
7.8	链烷烃、脂肪族醛酮沸点的边价连接性指数方程	291
7.9	分子拓扑研究醇类化合物疏水性参数	300
8	手性有机物 QSRR 模型	305
8.1	手性有机酸保留指数的手性拓扑指数研究	305
8.2	基于手性拓扑指数研究手性三唑类化合物的容量因子	310
8.3	手性有机酸色谱保留的手性连接性指数表征	317
8.4	手性有机酸保留指数的手性电性距离矢量方程	322
8.5	手性有机酸保留指数的手性 Kier 指数表征	328

第五编 无机物 QSPR 模型

9 无机物原子级 QSPR 模型	335
9.1 镧系元素理化性质的原子连接性指数模型	335
9.2 镧系离子理化性质的离子特征拓扑指数方程	337
9.3 基于配分函数研究阳离子标准熵、水合焓	339
9.4 有效主量子数连接性指数及其应用	344
9.5 运用原子参数探讨固体化合物中阳离子的标准熵	346
9.6 g_i , mG 及其应用	349
10 无机物分子级 QSPR 模型	354
10.1 价电子能级连接性指数及其应用	354
10.2 价电子平均能级连接性指数及其应用	360
10.3 基于拓扑方法研究碱金属卤化物的 F 心能带、晶格能、标准熵	363
10.4 价电子密度连接性指数及其应用	367
10.5 基于磁性拓扑研究碱金属化合物磁化率	370
10.6 无机物晶格能及其他理化性质的 QSPR 模型	375
10.7 运用拓扑方法研究铍小团簇的基态能量	386
参考文献	389

0 绪论:物质构效学基础

0.1 物质构效学定义

著者在中国知网上全文搜索“物质构效学”,发现最早是著者于 2003 年在《化学通报》上撰文提出的。著者给出物质构效学的管见(暂作基本定义):它是基于图论及量子化学衍生物结构的定量描述符,运用数学、统计学理论建立物质效能(包含理化性质、生物活性及代谢等)与描述符之间的数量依存关系,借以估算、预测物质性能及推断作用机理的科学。因此,物质构效学应包含物质定量结构-性质相关性(Quantitative Structure-Property Relationship, QSPR)、定量结构-活性相关性(Quantitative Structure-Activity Relationship, QSAR)、定量结构-代谢相关性(Quantitative Structure-Metabolism Relationship, QSMR)等研究内容。顺便提及 QSAR、QSPR、QSMR 作为术语首先被提及分别是在 1967 年、1985 年和 1991 年的 ISI 数据库。据此,建议物质构效学的英文名称:Quantitative Structure-Efficacy Relationship, 缩写为:QSER(注:本书只在“绪论”中以 QSER 代替 QSAR、QSPR 及 QSMR)。它是化学与数学、统计学、计算机的交叉科学,其主要作用:呈现结构与生物活性或理化性质之间的量变规律;经统计诊断的规律可预测新化合物的活性或性质,揭示影响效能的结构因素,探讨生物作用机理(如抑菌机理、杀虫机理、抗癌机理等)、化学反应机理以及物理变化机理(如色谱分离机理、疏水性作用等)。

0.2 物质构效学基本原理

物质构效学的基本假定:① 抽象的分子结构是可以数值化的;② 结构决定效能,效能只是结构的反映。

基本思想:描述符与效能都源自分子结构,描述符(可有多个数值)与效能之间一定存在联系。对于多种分子,与其对应的描述符就是变量集(即含多个变量),以“ X_i ”表示,其中 i 是变量的序号;这些分子的同—效能也是一个变量,以“ Y ”表示。根据第二条基本原理, Y 作为因变量,描述符 X_i 为自变量,它们之间是因果关系,必然存在某种数学关系,即 QSER 模型。物质构效学的基本功能就是建立 QSER 模型以及模型解释。

0.3 分子结构特征的数值描述

0.3.1 分子结构描述符

对化合物结构的准确描述是物质构效学研究的基础。因此,分子结构的数值化表征是进行物质构效学[主要是定量结构-性质/活性相关性(QSPR/QSAR)研究]研究的关键步骤。

构造化合物参数(分子描述符)应具备条件:① 良好的选择性,结构不同,则参数值不同;② 与化合物活性/性质有良好的相关性,才可用于 QSER 建模。目前,用于 QSER 研究的主要参数类型有:拓扑类参数、量子化学类参数、几何类参数和物理化学类参数。

表征分子结构的参数(即分子描述符,是反映分子结构差异的变量)常用的两种类型是:① 量子化学参数。此类参数虽然可以获得较好的预测结果且具有较明确的物理意义,但计算过程往往较复杂,只能计算小分子。② 在数学图论基础上提出分子拓扑指数(topological index)。拓扑指数的最大优点在于其计算简单,不依赖实验,易于运用。

因此,结合量子化学参数与拓扑指数的优点,依据化合物分子结构直接提取拓扑-量子参数建立了简单、适应范围广泛的 QSER 预测模型,这些指数能较好地反映出分子结构特征且易于计算。

0.3.2 拓扑指数的定义

拓扑指数是分子结构数值化的一种表征方式,它源于分子隐氢图的各种数学矩阵(如距离矩阵、邻接矩阵、关联矩阵等),经一定的数学运算所衍生出的揭示分子结构某种特征的描述符。它是一种图的不变量,直接产生于分子结构。构造拓扑指数有 4 个基本的要求:① 拓扑指数应为一个或一组数据,即分子图拓扑不变特征的数值化;② 能充分反映分子图的连接信息和化学环境;③ 能充分反映分子结构差异,良好结构选择性:选择性指数=拓扑指数个数/化合物数;④ 能有效地表达化合物结构与性质/活性的数学关系。

拓扑指数是基于图论的二维参数,能反映出分子大小、分支、柔韧性及分子整体形状,可直接从化合物的分子结构计算得到且对结构各异的各类化合物都适应。

拓扑指数所研究的是几何图形的一些性质,它们在图形被弯曲、拉大、缩小或任意的变形下保持不变,只要在变形过程中不使原来不同的点重合为同一个点,又不产生新点。换句话说,这种变换的条件是:在原来图形的点与变换了图形的点之间存在着——对应的关系,并且邻近的点还是邻近的点,这样的变换叫作拓扑变换。因此,拓扑有一个形象说法——橡皮几何学。描述图形的不变量即拓扑指数。

自 Wiener 指数以来,已有 400 多种新指数问世并得到广泛应用。目前较为常用的拓扑指数有分子连接性指数、电性距离矢量、原子类型电性拓扑指数及分子形状指数。

0.3.3 分子连接性指数

分子连接性指数是 1975 年 Randic 提出、后经 Kier、Hall 等人扩展的一类分子拓扑指

数(${}^mX_i^v$)。该指数是在分子隐氢图的邻接矩阵基础上衍生出的分子描述符,用以揭示分子的大小、分支及刚性等结构特征。其核心概念之一是原子点价(δ_i^v),表征原子结构信息,定义式为:

$$\delta_i^v = \frac{m_i - h_i}{M_i - m_i - 1} \quad (0-1)$$

式中, M_i, m_i 为非氢原子 i 的电子总数及价电子数, h_i 为与非氢原子 i 直接键合的氢原子数。

Kier 等人在分子隐氢图的邻接矩阵基础上,定义分子的价连接性指数(${}^mX_i^v$):

$${}^mX_i^v = \sum (\delta_i^v \delta_j^v \dots)^{-0.5} \quad (0-2)$$

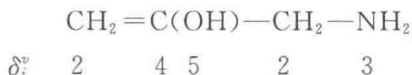
式中: ${}^mX_i^v$ 是一个指数体系,由 $(m+1)$ 个指数组成; m 表示相应指数的阶数, $m=0,1,2,\dots,t$ 代表子图的类型,常用4种子图路径(链)、簇(星)、簇-径、环状子图,依次对应 t 为 p, c, pc 及 ch 。这些拓扑指数不因图形形状变化而变。理论上 m 可以无限大,即有无穷多个连接性指数。利用MATLAB软件自编程序可以给出12个分子连接性指数,即每个分子有12个数值予以描述。即 $chi0pw', chi1pw', chi2pw', chi3pw', chi3cw', chi4pw', chi4cw', chi5pw', chi4pcw', chi5pcw', chi5cw', chihw'$ 。对应: ${}^0X_p^v, {}^1X_p^v, {}^2X_p^v, {}^3X_p^v, {}^3X_c^v, {}^4X_p^v, {}^4X_c^v, {}^5X_p^v, {}^4X_{pc}^v, {}^5X_{pc}^v, {}^5X_c^v, X_{ch}^v$ 。

其中, ${}^0X_p^v, {}^1X_p^v$ 的计算公式为:

$${}^0X_p^v = \sum (\delta_i^v)^{-0.5} \quad (0-3)$$

$${}^1X_p^v = \sum (\delta_i^v \delta_j^v)^{-0.5} \quad (0-4)$$

式(0-3)中的“ $\sum$ ”是对分子中全部非氢原子求和,式(0-4)中的“ $\sum$ ”是对分子中化学键求和。例如烯丙胺的分子图及 δ_i^v 值为:



其 ${}^0X_p^v, {}^1X_p^v$ 的值为:

$${}^0X_p^v = (2)^{-0.5} \times 2 + 3^{-0.5} + 4^{-0.5} + 5^{-0.5} = 2.9388$$

$${}^1X_p^v = (2 \times 4)^{-0.5} \times 2 + (5 \times 4)^{-0.5} + (3 \times 2)^{-0.5} = 1.3390$$

它们的物理意义:① 均能反映分子的大小。② 二阶指数能够有效区分同分异构体。③ 三阶指数能够反映分子的柔韧性和折叠程度。④ ${}^4X_{pc}^v$ 含有苯环及其取代基、取代基三维空间构型、取代基长度和杂原子取代基等信息。通常取代基相邻程度越高或者越拥挤,其值越大。当然,取代基长度大,其值亦大。另外还能区分取代基中杂原子位置。⑤ ${}^5X_c^v$ 包含分子的空间效应。⑥ 通常高级指数包含同类低级指数的结构信息,但高级指数与因变量的相关性不一定高。

实际上键长、键角、键能以及各原子的电负性等对化合物的理化性质及生物活性都会有一定的影响。可以对这些结点原子进行染色,即创新。因此,连接性指数的创新:① 修改 δ_i^v ;② 修改次方数;③ 修改公式等。

0.3.4 电性拓扑态指数

原子类型电性拓扑态指数(electrotopological state indices, E-state indices, E_n)是分子

连接性指数创始人 Kier 和 Hall 于 1990 年提出的基于原子水平的二维分子描述子。表征分子中每种非氢原子类型的结构参数,由两部分构成:其一是非氢原子类型本身的原子结构及局部拓扑环境,由此形成原子 n 的固有状态值(即本征值),以“ I_n ”表示;其二是反映该原子受分子中其他非氢原子扰动程度的本征值增量,以“ ΔI_n ”表示。将原子的本征值与扰动值相加,便得到原子的 E - 状态指数。将相同类型原子的 E - 状态指数相加,即得到原子类型电性拓扑状态指数。具体计算过程:

(1) 分子中非氢原子 n 的固有状态值(I_n) 定义为:

$$I_n = \frac{4\delta_n^v + N_n^2}{\delta_n \cdot N_n^2} \quad (0-5)$$

式中: N_n 为原子 n 的电子层数; $\delta_n = \sigma_n - h_n$, δ_n 、 σ_n 为原子 n 的支化度及形成 σ 键键数, h_n 为与原子 n 直接键合的氢原子数。 δ_n^v 定义为:

$$\delta_n^v = \sigma_n + \pi_n + k_n - h_n \quad (0-6)$$

式中: π_n 、 k_n 为原子 n 所成的 π 键键数及孤对电子数,例如硝基中氧原子,按式(0-5)、式(0-6) 计算得 $I_O = 7.0$ 。

(2) 原子 n 受到其他非氢原子的扰动量:成键原子 n 不是孤立存在的,它会受到分子中其他原子 j 的电性作用,而使原子 n 的电性发生改变,此改变量即“ ΔI_n ”。显然,原子 n 与 j 的固有状态值相差越大、两原子间相距越小,原子 j 对原子 n 所产生的扰动量(ΔI_n) 越大,因此, ΔI_n 定义为:

$$\Delta I_n = \sum_j \Delta I_{nj} = \sum (I_n - I_j) \cdot r_{nj}^{-2} \quad (0-7)$$

式中: r_{nj} 为原子 n 与 j 相隔最短的边数(d_{nj} , 可用分子拓扑图的邻接矩阵得到) 加 1, 即 $r_{nj} = d_{nj} + 1$ 。“ $\sum$ ” 是对分子中除原子 n 外其他非氢原子 j 的求和。可见, ΔI_n 包含拓扑与电性两种扰动。

(3) 原子类型 n 的电性拓扑状态指数 E_n 可按下式计算:

$$E_n = \sum_t (I_n + \Delta I_n)_t \quad (0-8)$$

式中: t 为分子所含原子类型 n 的个数,如硝基中的两个氧原子是同一原子类型,其 $t=2$ 。

通常有机分子中包含 78 种原子类型,即有 78 种原子类型电性拓扑指数(E_n)。著者所用电脑程序中只给出最为常用的 46 种原子基团的电性拓扑指数。

注:程序中“1”代表 1 个单键;“11”代表 2 个单键;“111”代表 3 个单键;“1111”代表 4 个单键;“11111”代表 5 个单键;“2”代表 1 个双键;“22”代表 2 个双键;“3”代表 1 个三键。另外,“char==0”代表非共轭体系,“char==1”代表共轭体系(多为芳香性)。

0.3.5 Kappa 形状指数

Kappa 形状指数也是 Kier 和 Hall 二人所创建的。表征分子形状的参数常用分子体积(V)、表面积(A),但其本能只是反映分子的大小。对于实验测定两个分子的 V 或 A 相同,但其形态未必相同。目前建立描述分子形状参数侧重于量子化学法与拓扑指数法,后者以

其简便性及在构效关系中常可获得良好的相关性,更受世人青睐。表征分子形状的拓扑指数常用 Kier 的 Kappa 形状指数(mK)。

对于烷烃及其环烷烃的 Kappa 形状指数(mK)定义式:

$${}^mK = [n_j - (m - 1)] \cdot (n_j - m)^2 \cdot {}^mP_j^{-2} \quad (0-9)$$

式中: m 是 Kappa 形状指数的阶数; n_j 是分子 j 中非氢原子数; mP_j 为分子 j 中具有路径数为 m 的个数。含多重键及杂原子分子的 Kappa 形状指数(mK)定义式:

$${}^mK = [n_j + a_j - (m - 1)] (n_j + a_j - 1)^2 \cdot ({}^1P_j + \alpha_j)^{-2} \quad (0-10)$$

式中引入分子 j 中原子 i 的修正值 α_i 定义为:

$$\alpha_i = r_i / r - 1 \quad (0-11)$$

式中: r 为 sp^3 杂化碳原子的原子半径($r = 77 \text{ nm}$); r_i 为非 sp^3 杂化碳原子或其他杂原子的原子半径。对于烷烃, $\alpha_i = 0$,无须校正。分子 j 的修正值 α_j 则为所有的非 sp^3 杂化碳原子及杂原子的 α_i 之代数和: $\alpha_j = \sum \alpha_i$ 。

表 0-1 常见原子的杂化状态及其修正值 α_i

No.	原子状态	原子半径/nm	α_i	No.	原子状态	原子半径/nm	α_i
1	C, sp^3	77	0	5	N, sp^3	74	-0.04
2	C, sp^2	67	-0.13	6	N, sp^2	62	-0.20
3	O, sp^3	74	-0.04	7	Cl, sp^3	99	0.29
4	O, sp^2	62	-0.20	8	Br, sp^3	114	0.48

各种有机物分子的 1K 及 2K 所反映分子结构信息:

(1) 1K 所显示的结构信息主要为分子结构的复杂度,确切地说是反映分子的环性。著者提出分子环比(β_j) 予以表征:分子中所含的环数(g_j) 与其原子数之比:

$$\beta_j = g_j n_j^{-2} \quad (0-12)$$

如苯酚、萘酚的 β_j 分别为 $1/7$ 、 $2/11$,其 1K 依次为 1.967 7, 2.954 1。

(2) 2K 主要揭示分子中原子的空间密度。因为在原子数 n_j 相同下,分子越似球体,其 2P 值越大,该分子的体积越小,相应分子的原子空间密度越大,据式(0-10),该分子的 2K 越小。

例如苯酚的修正值:6 个 sp^2 杂化碳原子及羟基氧原子。

$$\alpha_j = 6 \times (-0.13) + 1 \times (-0.04) = -0.82$$

苯酚:

$${}^1K = (7 - 0.82) \times (7 - 0.82 - 1)^2 \times (7 - 0.82)^{-2} = 0.838$$

$${}^2K = (7 - 0.82 - 1) \times (7 - 0.82 - 2)^2 \times (8 - 0.82)^{-2} = 1.756$$

$${}^3K = (7 - 0.82 - 1) \times (7 - 0.82 - 3)^2 \times (5 - 0.82)^{-2} = 2.998$$

0.3.6 分子电性距离矢量

刘树深等考察了多种著名拓扑指数的局限,提出能够较为全面反映分子的拓扑、几何及电性特征的分子电性距离矢量描述子(M_K)。其计算主要包括以下 3 个步骤:第一,定义分子

中非氢原子的原子类型(K_i)与原子属性(I_i),电性距离矢量理论给出 13 种原子类型及 43 种原子属性。分子中非氢原子 i 的固有属性(I_i) 定义为:

$$I_i = \sqrt{\frac{v_i}{4}} \cdot \frac{(2/n_i)^2 \cdot \delta_i^v + 1}{\delta_i} \quad (0-13)$$

式中: n_i 、 δ_i 、 δ_i^v 、 v_i 依次称为非氢原子 i 的电子层数、支化度、原子点价及价电子数。 δ_i 及 δ_i^v 的定义为:

$$\delta_i = \sigma_i - h_i, \delta_i^v = \sigma_i + \pi_i - h_i \quad (0-14)$$

式中: σ_i 、 π_i 为非氢原子 i 成 σ 键、 π 键所用的电子数, h_i 为与其直接键合的氢原子数。如 Cl 原子的 $n = 3$ 、 $\sigma = 1$ 、 $\pi = 0$ 、 $h = 0$, 则:

$$I_{\text{Cl}} = \sqrt{\frac{7}{4}} \times \frac{(2/3)^2 \times 1 + 1}{1} = 1.9108$$

由于非氢原子 i 在分子中并不是孤立存在的,还要受其他非氢原子(j) 的影响,将其扰动值以“ ΔI_i ”表示:

$$\Delta I_i = \sum_{i \neq j} \frac{I_i - I_j}{d_{ij}^2} \quad (0-15)$$

式中: d_{ij} 为非氢原子 i 、 j 之间的最短拓扑距离,即其间直接相连的边的数目。如 1,2-二氯乙烷中 $\Delta I_{\text{Cl}} (I_{\text{C}} = 1.5000)$:

$$\Delta I_{\text{Cl}} = \frac{1.9108 - 1.5000}{1^2} + \frac{1.9108 - 1.5000}{2^2} = 0.5135$$

第二,把非氢原子的相对电性(q_i) 定义为式(0-16):

$$q_i = I_i + \Delta I_i \quad (0-16)$$

则氯原子的相对电性为: $q_{\text{Cl}} = I_{\text{Cl}} + \Delta I_{\text{Cl}} = 2.4243$ 。

第三,为了进一步考虑不同原子类型的非氢原子之间的电性及拓扑作用,定义了电性距离矢量(M_K):

$$M_{kl} = \sum \frac{q_i \cdot q_j}{d_{ij}^2} \quad (0-17)$$

式中: k 、 l 分别为非氢原子 i 、 j 所属的原子类型号,其定义式:

$$K_i = 4(v_i - 4) + J_i \quad (0-18)$$

式中: J_i 表示非氢原子 i 在分子中所连接的其他非氢原子的数目。如氯原子的 $v_{\text{Cl}} = 7$, $J_{\text{Cl}} = 1$, 其 $K_{\text{Cl}} = 13$ 。 M_K 给出 13 种原子类型。这 13 种原子类型两两相互作用,便构成 91 个 M_K 。例如 1,2-二氯乙烷中,其 C 原子的 $k = 2$, Cl 原子的 $l = 13$, 它们的两两组合共有 3 种 M_K : M_{14} (即 M_{2-2})、 M_{25} (即 M_{2-13}) 及 M_{91} (即 M_{13-13}):

$$M_{14} = \frac{0.9865 \times 0.9865}{1^2} = 0.9732$$

$$M_{25} = 2 \times \left(\frac{0.9865 \times 2.4243}{1^2} + \frac{0.9865 \times 2.4243}{2^2} \right) = 5.9789$$

$$M_{91} = \frac{2.4243 \times 2.4243}{3^2} = 0.6530$$

例如 methionine 的 M_5 及 M_{15} 为 0.414 3、-1.865 6。

0.3.7 手性电性距离矢量的建立方法

(1) 构建手性指数。

依据手性碳原子上所连取代基的次序规程,定义其取代基中非氢原子的次序特征值(t_i):

$$t_i = \frac{Z_i - h_i/3}{(n_i - 1)^{0.5}} \quad (0-19)$$

按式(0-19)可求得—OH、=O、—NH₂、—S—等基团中非氢原子的 t_i 依次为 7.666 7、8.000 0、6.333 3、11.313 7。对于伯、仲、叔、季碳原子的 t_i 依次为 5.000 0、5.333 3、5.666 7、6.000 0;对于含有 π 键电子的叔、季碳原子的 t_i 亦为 5.666 7、6.000 0。对于含有多个非氢原子的取代基,其特征值为这些非氢原子特征值的平均值,如异丙基的特征值为 5.222 2,氢原子的特征值为 0。由此可见, t_i 基本反映了取代基的次序规则。分子 j 的手性指数(w_j)定义为:

$$w_j = 1 + \sum \frac{q_k}{2p_j} \left(1 - \frac{y_k}{y_{\max}^3}\right) \quad (0-20)$$

式中: q_k 为手性碳原子 k 的手性因子,对于 R 型、S 型碳原子的 q_k 值为 -1、1;而非手性碳原子的 q_k 值为 0。 p_j 为分子 j 中所含手性碳原子的个数,相应的“ $\sum$ ”是对所有手性碳原子求和;若 $p_j > 1$,则 w_j 为分子 j 的平均手性指数。手性碳原子 k 所连 4 个取代基的大小顺序为: $a < b < c < d$, y_k 定义为:

$$y_k = (t_d^2 - t_a^2)(t_c^2 - t_a^2)(t_b^2 - t_a^2) \quad (0-21)$$

$y_{\max}$ 则为所有分子中最大的 t_d^2 与最小的 t_a^2 之差,若取 $y_{\max} = (7.666 7^2 - 0.000 0^2) = 58.778 3$ 。

例如 R-methionine 分子的 $q_k = -1$ 、 $p_j = 1$, $y_k = (t_d^2 - t_a^2)(t_b^2 - t_a^2) = (6.745 1^2 - 0.000 0^2) \times (6.333 3^2 - 0.000 0^2) = 1\,824.890 9$

对于手性氨基酸(羧基酸)中,如只有一个手性碳原子,且均连接一个羧基,故计算 y_k 时均不包含此羧基。相应式(0-21)中“ $y_{\max}^3$ ”应为“ $y_{\max}^2$ ”。代入式(0-20)得: $w_{j,R} = 0.764 10$;相应得 S-methionine 分子的 $w_{j,S} = 1.235 90$ 。

(2) 手性电性距离矢量(cM_K)定义为:

$${}^cM_K = M_K \times w_j \quad (0-22)$$

对于 R-methionine、S-methionine 的 cM_5 分别为 0.316 6、0.512 0, ${}^cM_{15}$ 分别为 -1.425 5、-2.305 7。

0.3.8 量子化学参数

虽然在计算机相当发达的今天,对于量子力学的奠基定律——Schördinger 方程的精确求解仍然是极其困难的。成功克服此困难的一个理论飞跃就是电子密度泛函理论(Density functional theory, DFT)。电子密度泛函理论是在 Thomas-Fermi 理论的基础上发展起来

量子理论的一种表述方式。传统的量子理论将波函数作为体系的基本物理量,而密度泛函理论则通过粒子密度来描述体系基态的物理性质,此即体系的基态唯一的决定于电子密度的分布(Hohenberg-Kohn 定理)。因为粒子密度只是空间坐标的函数,这使得密度泛函理论将 3N 维波函数问题简化为三维粒子密度问题,十分简单直观。即可通过KS-SCF自洽迭代求解单电子多体薛定谔方程来获得电子密度分布。另外,粒子密度通常是可以实验直接观测的物理量。粒子密度的这些优良特性,使得密度泛函理论具有诱人的应用前景。密度泛函理论也是一种完全基于量子力学的从头算(abinitio)理论,但是为了与其他的量子化学从头算方法区分,人们通常把基于密度泛函理论的计算叫作第一性原理(first-principles)计算。

基本功能:① 计算特定分子结构的能量。和能量相关的一些性质也可以通过某些方法来计算。② 完成构型优化,确定全局或局域最小点。构型优化主要取决于能量的梯度,也就是能量对原子坐标的一阶导。构型优化是化学计算的基础,任何性质的计算都是在已优化好的分子结构上进行的。一个分子结构对应一个波函,一个波函就可以解出一个能量值。在自然条件下,体系倾向于以能量最低的形式存在(也就是我们常说的最稳定的构型)。所给的分子结构(不管是手绘的还是来自晶体结构的)在转换为高斯输入的文件时,由于画得不准确(如立体的画成平面的等)和具体化学环境的变化(在不选择溶剂时,默认是在气相中优化,这与晶体环境是完全不同的)而存在一些不合理的地方。所以首先要进行优化,才能进行其他类型的性质计算。③ 计算由于分子内原子间运动所引发的分子振动频率。频率是能量对原子坐标的二阶导。频率计算还可以确定其他与二阶导有关的性质。不是所有的化学计算方法都能计算频率。

Gaussian 是一个功能强大的量子化学综合软件包,可以完成 DFT 的各种物理量的计算。Gaussian 具体功能:分子能量和结构、过渡态能量和结构、键和反应能量、分子轨道、多重矩、原子电荷和电势、振动频率、红外和拉曼光谱、核磁性质、极化率和超极化率、热力学性质、反应路径。上述计算可以对体系的基态或激发态执行。可以预测周期体系的能量、结构和分子轨道。因此,Gaussian 可以作为功能强大的工具,用于研究许多化学领域的课题,例如取代基的影响、化学反应机理、势能曲面和激发能等。执行程序可在不同型号的大型计算机、超级计算机、工作站和个人计算机上运行,并相应有不同的版本。

根据 Gaussian 03 程序计算的各种量子化学参数作为自变量集,相应生物活性为因变量,输入统计软件包 SPSS 17.0 对相关数据进行多元线性回归分析,主要是逐步回归(stepwise regression)与最佳子集变量回归(leaps-and-bounds regression)等建模方法,以建立相应的 QSAR 模型。依据进入最终 QSAR 模型中的量化参数可以揭示影响生物活性的主要分子结构基团,探讨化合物分子发挥生物活性作用的微观历程。采用量化参数建立模型的最大优点是物理意义明确,利于解释发挥生物作用的分子机理。

0.4 QSAR 模型的类型

对 QSAR 模型的分类标准有多种,本书只介绍以下两种分类标准。