研究简报

香茶菜属植物二萜化合物核磁共振碳谱模拟

仝建波[1]，张生万[1,2]，马云霞[1,3]，李改仙[3]
([1] 山西大学化学化工学院，山西 太原 030006；[2] 山西大学生命科学与技术学院，山西 太原 030006；
[3] 晋中学院化学化工系，山西 晋中 030600)

关键词：香茶菜属植物二萜化合物；定量结构波谱相关；^{13}C NMR 化学位移；原子电性作用矢量；原子杂化状态指数
中图分类号：O 641.1 **文献标识码**：A **文章编号**：0438－1157（2004）04－0975－05

Spectroscopic simulation of ^{13}C nuclear magnetic resonance of diterpenoids of isodon species

TONG Jianbo[1], ZHANG Shengwan[1,2], MA Yunxia [1,3], LI Gaixian[3]
([1] *School of Chemistry and Chemical Engineering, Shanxi University, Taiyuan* 030006, *Shanxi, China*;
[2] *School of Life Science and Technology, Shanxi University, Taiyuan* 030006, *Shanxi, China*;
[3] *Department of Chemistry and Chemical Engineering, Jinzhong College, Jinzhong* 030600, *Shanxi, China*)

Abstract: Atomic electronegativity interaction vector (AEIV) and atomic hybridization state index (AHSI) were used for establishing the quantitative structure-spectroscopy relationship (QSSR) model of ^{13}C NMR chemical shifts of isodon diterpenoid compounds. Multiple linear regression (MLR) and computational neural network (CNN) were used to create the models, and the estimation stability and generalization ability of the models were strictly analyzed by both internal and external validations. The established MLR and CNN models were correlated with experimental values and the correlation coefficients of model estimation, leave-one-out (LOO) cross-validation (CV), and predicted values of external samples were R_{cum} = 0.9724, R_{CV} = 0.9723, Q_{ext} = 0.9738 (MLR); R_{cum} = 0.9957, Q_{ext} = 0.9956 (CNN), respectively. The results indicated that CNN gave significantly better prediction of ^{13}C NMR chemical shifts for isodon diterpenoids than MLR. Satisfactory results showed that AEIV and AHSI were obviously good for modeling ^{13}C NMR chemical shifts of isodon diterpenoid compounds.

Key words: diterpenoids of isodon species; quantitative structure-spectrum relationship; ^{13}C NMR chemical shift; atomic electronegativity interaction vector; atomic hybridization state index

引 言

核磁共振（NMR）技术在化合物结构鉴定、构型、构象、反应机理研究中起着极其重要的作用[1-3]。碳原子作为有机化合物的骨架，其 NMR 谱被广泛地应用于对有机化合物结构鉴定等的研究[4-5]。因此，通过化合物结构参数与其^{13}C NMR 化学位移的定量关系来定量预测未知化合物的^{13}C NMR 化学位移，即碳谱模拟[6-8]，可为鉴定化合物结构、探讨反应机理、揭示^{13}C NMR 化学位移

2006－05－24 收到初稿，2006－07－05 收到修改稿。
联系人：张生万。**第一作者**：仝建波（1975—），男，博士研究生。
基金项目：山西省工业攻关项目基金（2006031204）；山西省研究生创新基金项目。
Received date: 2006－05－24.
Corresponding author: ZHANG Shengwan. **E－mail**: zswan@sxu.edu.cn
Foundation item: supported by Industry Innovation Foundation of Shanxi Province (2006031204) and the Graduate Student Innovation Foundation of Shanxi Province.

随结构的变化规律提供理论依据。香茶菜属(Isodon)植物属唇形科(Labiatae)-罗勒亚科(Ocimoideae),种类很多,植物资源非常丰富,具有清热解毒、活血化淤、抗菌消炎、抗肿瘤、治疗各种肝炎等功效,对各种癌症患者有缓解症状的作用[9-10]。本文从分子二维结构出发,利用不同种类原子对目标原子作用效果建立的原子电性作用矢量(AEIV)[11]来描述等价碳原子所处化学微环境特征;并利用原子杂化状态指数(AHSI)描述原子杂化状态。以此建立起表征不同有机物等价共振原子所处化学环境和自身状态的定量结构波谱(QSSR)模型,对 350 个香茶菜属植物二萜化合物中 7000 个碳原子进行碳谱(^{13}C NMR)模拟。在对模型的检验过程中,采用内部及外部双重验证的办法对所得模型稳定性能进行深入分析和检验,均取得了令人满意的结果。

1 原理及方法

1.1 原子电性作用矢量

众所周知,有机化合物分子中原子的核磁共振谱化学位移受到很多因素的影响,其中原子所处的局部化学微环境以及自身的杂化状态对其化学位移影响最大。因此,在模拟化合物中不同等价共振碳原子的化学位移时必须考虑这两方面的因素。

首先,研究影响原子化学微环境因素的表达方式,基于分子中原子之间存在着相互作用,各个相连的原子都对等价碳施加影响。因此构建化合物的分子结构化方法,探寻分子中各等价碳原子的化学位移变化规律,就要充分考虑碳原子周围的化学环境。由于分子中各等价原子的化学位移大小与其周围电子云分布有关,该分布与周围各键合原子的电负性及相隔距离相关[12],并且具有不同化学性质的原子对目标原子的作用效果不尽相同。本文采用原子电性作用矢量(AEIV)表征原子的这一局部环境特征。

原子电性作用矢量将有机化合物分子中的原子按其所在元素周期表的主族进行分类,即将有机物分子中常见原子分为 5 类,结果见表 1。原子电性作用矢量考虑的是分子中某一指定原子受到的来自其他各类原子的作用,且具有不同化学性质的原子对目标原子的作用效果不尽相同,其具体运算公式为

$$v_{i,k}=\sum_{j\in k,j\neq i}^{\mathrm{all}(j)}\frac{\chi_j}{d_{i,j}^6}\quad(1\leqslant k\leqslant 5)\tag{1}$$

式中 k 为原子类型;i 为目标原子;j 为分子中属于第 k 种类型的所有原子($j\neq i$);χ 为原子相对电性大小,即以碳原子的电负性为基准得到其他原子与其的比值大小,均采用鲍林电负性标度,例如氧的相对电负性为:3.44/2.55=1.349;d_{ij} 表示第 i 个原子到第 j 个原子之间的距离,是从原子 i 通过一个或多个化学键连接到原子 j 的所有路径中各个相对键长加合的最小值。对于键长则取化学键相对于碳碳单键的键长大小,即C—C单键的相对键长为 1,则C—O、C═C、C═O的相对键长分别为 d_{C-O} = 0.143 nm/0.154 nm = 0.927,$d_{C=C}$=0.134 nm/0.154 nm=0.870,$d_{C=O}$=0.122 nm/0.154 nm=0.792。分别用 v_H、v_C、v_N、v_O、v_X 表示各类原子对中心碳原子的作用项。

表 1 有机化合物中常见原子类型划分

Table 1 Division of atomic type of atoms in organic compounds

Type of atoms	Families of periodic table	Atoms
1	IA	H
2	IVA	C
3	VA	N, P
4	VIA	O, S, Se
5	VIIA	F, Cl, Br, I

1.2 原子杂化状态指数

为描述原子自身杂化状态对其化学位移的影响,引入原子杂化状态指数(atomic hybridization state index, AHSI)[11],用于表征原子自身的杂化状态。计算方法为

$$\mathrm{AHSI}=\sqrt{\nu/4}\left[(2/n)^2\delta_{\sigma+\pi}+1\right]/\delta_\sigma\tag{2}$$

式中 ν 是原子价层电子数;n 为该原子价电子层主量子数;$\delta_{\sigma+\pi}$是 σ 和 π 键总电子数;δ_σ为成 σ 键电子数。表 2 列出了碳和氧两种原子不同杂化类型的 AHSI 值。

表 2 碳、氧两种原子不同杂化状态的 AHSI 值

Table 2 AHSI of different hybridization state of carbon and oxygen atom

Hybridization state of atom	AHSI
C_{sp3}	1.2500
C_{sp2}	1.6667
C_{sp}	2.5000
O_{sp3}	1.8371
O_{sp2}	3.6742

2　结果与讨论

2.1　数据集选取及划分

所选 350 个香茶菜属植物二萜化合物中碳原子^{13}C NMR 化学位移数据取自文献[13]。将 350 个萜类化合物分子中的母体碳原子编号，第 1 号分子的 1～20 个碳原子为第 1～20 号原子，第 2 号分子的 1～20 个碳原子为第 21～40 号原子，依次类推，第 350 号分子的 1～20 个碳原子为第 6981～7000 号原子。

为深入研究 AEIV、AHSI 与甾族化合物^{13}C NMR化学位移内在联系，用多元线性回归(MLR)、神经网络（CNN）这两种典型线性及非线性方法进行建模。另外对所建模型的外部预测能力和真实有效性进行验证是定量构效关系中非常重要的一个部分，其中留一法（LOO）交叉检验(CV)[14]复相关系数 R_{CV}是目前较为广泛使用的一种模型验证方法，然而 Tropsha 等[15-17]最近研究结果表明，R_{CV}的大小与模型预测能力并无明显相关关系，对模型预测能力的评价只能通过外部样本集即测试集来进行，模型外部预测能力可用 Q_{ext}(external Q)[17]来衡量。

$$Q_{ext} = \sqrt{1 - \frac{\sum_{i=1}^{test}(y_i - \hat{y}i)^2}{\sum_{i=1}^{test}(y_i - \bar{y}_{tr})^2}} \qquad (3)$$

式中　y_i和 $\hat{y}_i$ 分别为测试集中样本的实验值和预测值；$\bar{y}_{tr}$为训练集样本实验的平均值。鉴于此，从 350 个化合物中每隔 5 个化合物抽取一个组成测试集（test set），共 70 个；剩余 280 个化合物作为训练集（training set）。

2.2　模型建立

MLR 是一种经典的建模方法，它对自变量和因变量加以线性拟合以得到最小二乘（LS）意义下的最佳结果。将各分子结构原子编号并将原子数目、类型及连接关系输入计算机，由利用 C 语言自编应用程序 AEIV.exe 进行识别、找寻最短路径并计算 AEIV 描述子。对训练集中 280 个化合物 5600 个碳原子的 AEIV、AHSI 与其^{13}C NMR 化学位移建模。

$$CS = -206.6203 - 1.4878v_H - 0.6824v_C + 11.0302v_O + 209.4488AHSI$$

$$n = 5600,\ m = 4,\ R_{cum} = 0.9724,$$

$$SD = 11.18,\ F = 24299.99 \qquad (4)$$

CV 建模

$$n = 5600,\ m = 4,\ R_{CV} = 0.9723,$$

$$SD_{CV} = 11.20,\ F_{CV} = 24213.42$$

式中　n 为样本数；m 为变量数；R 为复相关系数；SD 是标准偏差；F 是 F 检验统计量。

另外使用式（3）对 350 个香茶菜属植物二萜化合物 7000 样本进行拟合与预测，并将计算结果与实验值相关情况绘于图 1 中，其相关统计参数列于表 3 中。可以看到 AEIV 及 AHSI 与香茶菜属植物二萜化合物^{13}C NMR 化学位移有较好的相关性，具体表现为图 1 中大部分样本分布于过原点 45°直线周围，且绝大多数残差点均在二倍标准偏差（SD）以内。但值得提出的是图 1 中出现了“条带”现象，图 1 中部分样本的计算结果误差较大。可能有以下几种原因所致：(1) AEIV 和 AHSI 描述子解释香茶菜属植物二萜化合物的结构特征不充分；(2) 样本自身结构的特殊性；(3) 建模方法没有充分表达结构参数与性质之间的联系。经分析发现误差较大的样本有羰基中的碳原子以及与多个氧原子相连的碳原子，这可能是由于上述描述子没有表达出电负性较大的氧原子对碳原子强的去屏蔽效应所致；另外本文所用描述子 AEIV 和 AHSI 与香茶菜属植物二萜化合物中^{13}C NMR 化学位移间可能存在非线性关系，而利用 MLR 建模体现不出该信息。

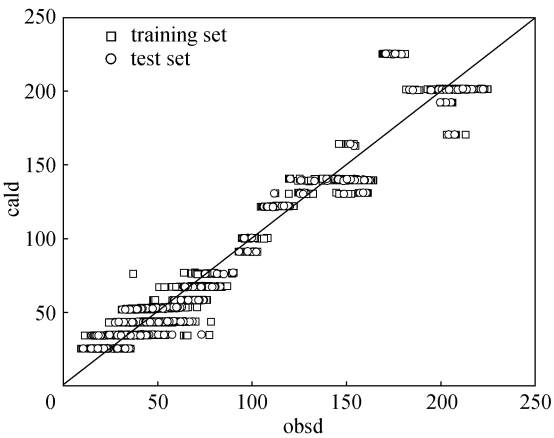

图 1　多元线性回归模型对 5600 个训练集样本估计值及对 1400 个测试集样本预测值与实验观测值相关情况

Fig. 1　Plot of estimated values of 5600 samples in training set as well as predicted values of 1400 samples in test set *versus* observed values (MLR model)

为深入研究上述 4 个描述子与香茶菜属植物二萜化合物中^{13}C NMR 谱化学位移之间的隐含关系，使用误差反传（BP）算法训练前馈型多层感知机（FMLP）来实现该类目的（CNN 模型使用 NeuroSolutions for Matlab 神经网络工具包基于 Matlab7.0 环境实现）[18-19]。所采用带有偏置（bias）节点 CNN 神经网络相关参数为：网络层数：3；输入向量维数：4＋1bias（4 个描述子）；隐含层神经元数目：40＋1bias；输出层神经元数目：1；隐含层传递函数：Sigmoid；输出层传递函数：Linear；网络权值初始化方法：Nguyen-Widrow 法；训练规则：带动量项及自适应学习速率的梯度递减法；初始学习速率 η：0～1 之间随机赋值；初始动量项 δ：0～1 之间随机赋值；数据预处理：自定标化。为防止网络出现过拟合，以 1400 个测试集样本作为监控集，并以训练过程中监控集均方根误差平方（MSE）达最小来确定网络权值。由此最终获得 CNN 模型对 5600 个训练集样本拟合结果及对 1400 测试集样本预测值相关统计参数，见表 3、图 2。可看到 CNN 建模结果明显优于 MLR 线性模型，这可能是由于所选描述子与香茶菜属植物二萜化合物中^{13}C NMR谱化学位移存在一定的非线性关系，CNN 拟合使计算精度进一步提高所致。

从上述结果可看出，利用 AEIV 与 AHSI 描述子所建回归模型对香茶菜属植物二萜化合物^{13}C NMR化学位移值模拟结果具有较高精度；且用 CNN 所建模型的稳定性明显优于 MLR 模型。

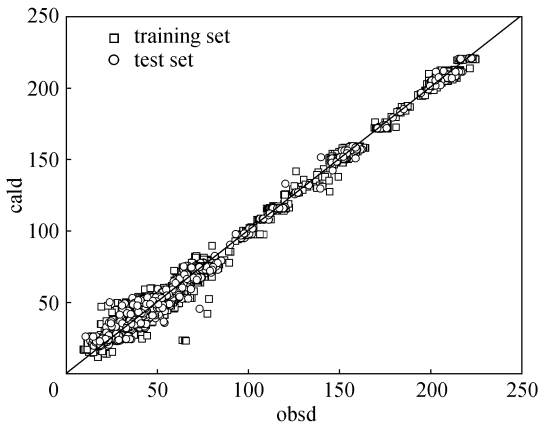

图 2 人工神经网络回归模型对 5600 个训练集样本估计值及对 1400 个测试集样本预测值与实验观测值相关情况

Fig. 2 Plot of estimated values of 5600 samples in training set as well as predicted values of 1400 samples in test set *versus* observed values (CNN model)

表 3 不同回归模型的统计参量比较

Table 3 Statistical data of fitting result by models

Models	Training set	Test set	R_{cum}	R_{CV}	Q_{ext}	SD	SD_{CV}
MLR	5600	1400	0.9724	0.9723	0.9738	11.18	11.20
CNN	5600	1400	0.9957	—	0.9956	4.69	—

Note: R_{cum}—cumulative multiple correlation coefficient of training set; R_{CV}—cumulative cross-validated R_{cum} of training set; Q_{ext}—external Q of test set; SD—standard deviation of training set; SD_{CV}—cross-validated standard deviation of training set.

3 结 论

分子中各个原子之间存在着相互作用，各个相连的原子都对等价碳施加影响，同时原子自身状态对其化学位移也有影响。因此，要构建分子中各等价碳原子的化学位移变化规律，就要充分考虑碳原子本身及其周围的化学环境。为表征原子局部环境特征，用 AEIV 与 AHSI 描述子对香茶菜属植物二萜化合物^{13}C NMR 谱化学位移进行模拟，得到令人满意的结果。所建模型不仅在一定程度上阐明了香茶菜属植物二萜化合物^{13}C NMR 谱化学位移与其分子结构信息之间的关系，同时也为模拟有机化合物分子 NMR 谱化学位移提供了一种新方法。值得提出的是本文所用描述子是基于二维空间提出的，因而不能分辨诸如顺反异构、手性等三维空间问题，对此还需做进一步的研究。

References

[1] William S P. Protein association studied by NMR diffusometry. *Curr. Opin. Colloid. In.*, 2006, **11**: 19

[2] Witkowski S, Maciejewska D, Wawer I. ^{13}C NMR studies of conformational dynamics in 2, 2, 5, 7, 8-pentamethylchroman-6-ol derivatives in solution and the solid state. *J. Chem. Soc., Perkin Trans.* 2, 2000 (7): 1471

[3] Neuvonen H, Neuvonen K. Correlation analysis of carbonyl carbon ^{13}C NMR chemical shifts, IR absorption frequencies and rate coefficients of nucleophilic acyl substitutions. A novel explanation for the substituent dependence or reactivity. *J. Chem. Soc., Perkin Trans.* 2, 1999 (7): 1497

[4] Beger R D, Bolton P H. Protein ϕ and ψ dihedrals restraints determined from multidimensional hypersurface correlations of backbone chemical shifts and their use in the determination of protein tertiary structures. *J. Biomol.*

NMR, 1997, **10**: 129

[5] Wishart D S, Sykes B D. Chemical shifts as a tool for structure determination. *Methods Enzymol.*, 1994, **239**: 363

[6] Kvasnicka V. An application of neural networks in chemistry. Prediction of ^{13}C NMR chemical shifts. *J. Math. Chem.*, 1991, **6**: 63

[7] Grant D M, Paul E G. Carbon-13 nuclear magnetic resonance (Ⅱ): Chemical shifts data for the alkanes. *J. Am. Chem. Soc.*, 1964, **86**: 2984

[8] Lindeman L P, Adams J Q. Carbon-13 nuclear magnetic shifts for the resonance spectrometry. Chemical shifts for the paraffins through C_9. *Anal. Chem.*, 1971, **43**: 1245

[9] Zhang Y, Liu J W, Jia W, Zhao A H, Li T. Distinct immunosuppressive effect by *Isodon serra* extracts. *Int. Immunopharmacol.*, 2005, **5**: 1957

[10] Ulbelen A. Cardioactive and antibacterial terpenoids from salvia species. *Phytochemistry*, 2003, **64**: 395

[11] Zhou Peng (周鹏), Mei Hu (梅虎), Zhou Yuan (周原), Tian Feifei (田菲菲), Li Zhiliang (李志良). *Chin. J. Anal. Chem.* (分析化学), 2006, **34** (2): 200

[12] Liu S S, Liu H, Yu B M, Li Z. Investigation on quantitative relationship between chemical shift of carbon-13 nuclear magnetic resonance spectra and molecular topological structure based on a novel Atomic Distance-Edge Vector (ADEV). *J. Chemometr.*, 2001, **15** (5): 427

[13] Sun Handong (孙汉董), Xu Yunlong (许云龙), Jiang Bei (姜北). Diterpenoids from Isodon Species (香茶菜属植物二萜化合物). Beijing: Science Press, 2001

[14] Wold S. Cross-validation estimation of the number of components in factor and principal components models. *Technometrics*, 1978, **20**: 897

[15] Golbraikh A, Tropsha A. Beware of q^2! *J. Mol. Graphics Mod.*, 2002, **20**: 269

[16] Gramatica P, Pilutti P, Papa E. Validated QSAR prediction of OH tropospheric degradation of VOCs: splitting into training-test sets and consensus modeling. *J. Chem. Inf. Comput. Sci.*, 2004, **44**: 1794

[17] Tropsha A, Gramatica P, Gombar V K. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci.*, 2003, **22**: 69

[18] Peng Qianrong (彭黔荣), Yang Min (杨敏), Shi Yanfu (石炎福), Yu Huarui (余华瑞), Liu Zhongxiang (刘钟祥). Artificial neural network based on hybrid genetic algorithm and prediction of melting points of organic compounds. *Journal of Chemical Industry and Engineering* (*China*) (化工学报), 2005, **56** (10): 1922

[19] Jiang Kaiyu (姜开宇), Su Tongyi (苏同义), Wang Minjie (王敏杰), Yu Tongmin (于同敏). Simulation on rule of shrinkage of large CPUE based on neural network. *Journal of Chemical Industry and Engineering* (*China*) (化工学报), 2005, **56** (8): 1520