文章编号:1001-6880(2016)1-0090-06

柠檬皮香气成分结构表征与色谱保留指数预测

廖立敏[1,2],李建凤[2],雷光东[1,2]*

[1]"果类废弃物资源化"四川省高等学校重点实验室;[2] 内江师范学院化学化工学院,内江 641100

摘　要:通过有机化合物分子顶点及顶点原子之间的关系对柠檬皮香气成分中的67个化合物进行了结构表征,采用多元线性回归(MLR)和偏最小二乘回归(PLS)建立了化合物结构与色谱保留指数关系(QSRR)模型,模型稳定性及预测能力经内部及外部双重检验进行了评价。两模型的建模复相关系数(R)分别为0.951和0.938,留一法交互检验复相关系数(R_{CV})分别为0.925和0.900,外部预测的复相关系数(R_{test})分别为0.955和0.940。结果表明所采用的分子结构描述符具有较强的分子结构表达能力,两模型具有良好的估计能力、稳定性和外部预测能力。

关键词:柠檬皮;香气成分;结构描述符;定量结构保留相关

中图分类号:R629　　**文献标识码**:A　　DOI:10.16333/j.1001-6880.2016.1.016

Structural Characterization and Chromatographic Retention Index Prediction for Aroma Components of Lemon Peels

LIAO Li-min[1,2], LI Jian-feng[2], LEI Guang-dong[1,2]*

[1]*Key Laboratory of Fruit Waste Treatment and Resource Recycling of Sichuan Provincial College*;

[2]*College of Chemistry and Chemical Engineering, Neijiang Normal University, Neijiang 641100, China*

Abstract: Structures of 67 aroma compounds of lemon peel were characterized through the molecular vertexes and relationships between molecular vertexes. Multiple linear regression (MLR) and partial least square regression (PLS) were adopted to construct two quantitative structure-retention relationship (QSPR) models. The estimation stability and predictability of the two models were strictly analyzed by both internal and external validations. Modeling multiple correlation coefficients (R) of the two models were 0.951 and 0.938, leave-One-Out (LOO) cross-validation (CV) correlation coefficients (R_{CV}) were 0.925 and 0.900, external forecasts multiple correlation coefficient (R_{test}) were 0.955 and 0.940, respectively. The results showed that the structural descriptors were superior in molecular structural representation. The estimated capacity, stability and predictability of the two models were good.

Key words: lemon peel; aroma components; structural descriptor; quantitative structure-retention relationship

柠檬(*Citrus limon*)原产东南亚,现主要产地为美国、意大利、西班牙等,目前中国四川安岳也有广泛种植,安岳被誉为"中国柠檬之乡",内江也有一定的分布。尤力克(Eureka)为常见的品种之一,又称油力克、油利加,安岳及内江等栽培的多为此类。柠檬具有许多药用价值,可以利尿并缓解风湿和肠道疾病。从柠檬皮中提取的香精油,可用于生产高级化妆品及治疗结石病的药物。关于柠檬精油成分的研究已有一些报道[1-4],何朝飞等[4]采用顶空固相微萃取提取了尤力克柠檬、粗柠檬和北京柠檬精油,结合GC-MS技术研究了3个品种果皮的香气成分,其中从尤力克柠檬精油中分离并鉴定出67种化合物。

定量结构-色谱保留关系(QSRR)研究成为预测化合物色谱保留值、解释色谱保留机理、辅助确证化合物结构的重要手段之一,研究者们在QSAR/QSRR方面已经做过许多有意义的工作[5-7]。本文采用简易方法对尤力克柠檬皮精油中分离出的67种化合物进行结构表征,借助多元线性回归(MLR)、偏最小二乘回归(PLS)构建化合物分子结构和气相色谱保留指数之间的(QSRR)相关模型,以期为研究天然产物中的挥发性化合物的色谱行为提供有益参考。

收稿日期:2015-10-08　　接受日期:2015-12-03

基金项目:四川省教育厅青年基金(13ZB0003)

* 通讯作者　E-mail:leigdnjtc@126.com

1　实验方法

1.1　数据集的选取

本研究选用柠檬皮香气成分中的 67 个化合物为研究样本(列于表 1),其保留指数的实验值来自于文献[4],保留指数是以 DB-5MS 石英毛细管柱为色谱柱分离测得。顶空固相微萃取条件:40 ℃平衡 15 min;顶空吸附 40 min;解吸 5 min;色谱条件:色谱柱为 DB-5MS 石英毛细管柱(30 m × 0.25 mm, 0.25 μm);程序升温,35 ℃保持 5 min,以 3 ℃/min 升至 180 ℃保持 2 min,再以 5 ℃/min 升至 240 ℃,保持 2 min;进样口温度 250 ℃,不分流进样;载气为氦气,1 mL/min;质谱条件:离子化方式离子电离(electron ionization, EI),电子能量 70 eV;传输线温度 280 ℃;离子源温度 230 ℃;四极杆温度 150 ℃;质量扫描范围 m/z 35 ~ 400,检索图谱库(NIST 2008 和 Flavour 2.0)得各化合物结构及名称。选取尾号为"0"、"5"的共计 13 个样本(标注" * ")作为测试集,用于评价模型预测能力,剩余 54 个样本作为训练集,用于构建模型。

表 1　67 个化合物色谱保留指数(RI)实验值(Exp.)及计算值(Cal.)

Table 1　67 compounds and their chromatographic retention index (RI)

序号 No.	化合物 Compound	RI (Exp)	RI (Cal. MLR)	RI (Cal. PLS)
1	己醛	733	842	829
2	反式-2-己烯醛	804	873	882
3	顺式-3-己烯醇	826	842	836
4	庚醛	857	911	900
5 *	2,4-己二烯醛	863	904	931
6	α-水芹烯	876	1002	1030
7	α-蒎烯	881	990	985
8	莰烯	896	963	983
9	4-己烯-1-醇	912	842	833
10 *	β-蒎烯	923	968	977
11	庚醇	927	879	856
12	月桂烯	947	999	1004
13	1-侧柏烯	953	980	1035
14	辛醛	957	948	926
15 *	α-松油烯	966	1000	1025
16	柠檬烯	986	974	1011
17	顺式-罗勒烯	991	1004	1008
18	δ-3-蒈烯	1004	989	980
19	γ-松油烯	1014	1003	1022
20 *	3-异丙烯基-5-甲基-1-环己烯	1021	942	977
21	甲酸辛酯	1024	1168	1197
22	异松油烯	1037	997	1012
23	反式-4-癸烯	1040	953	993
24	5-甲基-1-癸烯	1042	1040	1023
25 *	芳樟醇	1052	1108	1107
26	壬醛	1055	1049	1041
27	邻异丙基甲苯	1057	1066	1094
28	exo-异莰酮	1086	1078	1084
29	水合樟烯	1088	963	983
30 *	香茅醛	1100	1112	1093

31	异龙脑	1107	1090	1118
32	反式-对-2,8-孟二烯-1-醇	1111	1121	1124
33	4-松油醇	1120	1125	1116
34	(*S*)-顺式-马鞭草烯醇	1130	1119	1085
35*	α-松油醇	1136	1071	1078
36	萜品醇	1140	1178	1173
37	癸醛	1151	1118	1112
38	顺式-香芹醇	1160	1112	1117
39	橙花醇	1174	1124	1098
40*	香茅醇	1176	1079	1048
41	橙花醛	1185	1151	1143
42	d-薄荷酮	1193	1119	1118
43	香叶醇	1201	1124	1098
44	香叶醛	1215	1151	1143
45*	麝香草酚	1237	1228	1204
46	薄荷酮氧化物	1245	1266	1297
47	十一醛	1249	1187	1183
48	甘香烯	1272	1402	1378
49	乙酸香茅酯	1293	1325	1365
50*	乙酸橙花酯	1303	1373	1417
51	乙酸香叶酯	1322	1373	1417
52	β-石竹烯	1348	1426	1390
53	红花醛	1354	1235	1216
54	α-柏木萜烯	1367	1440	1443
55*	α-石竹烯	1381	1434	1475
56	(-)-异丁香烯	1389	1426	1390
57	丙酸橙花酯	1407	1483	1496
58	β-金合欢烯	1412	1401	1381
59	巴伦西亚橘烯	1419	1388	1414
60*	γ-绿叶烯	1422	1467	1433
61	β-雪松烯	1433	1393	1418
62	β-红没药烯	1439	1391	1391
63	香橙烯	1445	1363	1293
64	α-依兰油烯	1450	1477	1384
65*	α-绿叶烯	1459	1420	1448
66	γ-红没药烯	1470	1419	1394
67	红没药醇	1605	1515	1559

1.2 原理与方法

1.2.1 分子结构描述符的产生

有机化合物的色谱保留行为与分子结构直接相关,分子母体结构、所含基团的大小、种类、数目等都会影响其与固定相之间的作用,从而影响其在色谱柱中的保留指数。在有机化合物分子的骨架结构图中,将每一个非氢原子视为分子顶点。认为分子顶点自身及分子顶点之间的关系对化合物色谱保留指数产生重要影响,处在不同微环境中的分子顶点及不同类型分子顶点间的关系对分子性质的贡献不同,处在相同微环境中的分子顶点及相同类型分子顶点的关系对分子性质的贡献具有加和性。借鉴文献[7-9]中的分类方法将分子内的顶点原子依据其所连接的其它顶点原子数分为 A1、A2、A3、A4 四种类

型,如仅与 1 个分子顶点相连的伯碳原子属于 A1。然后根据顶点原子在元素周期表所处位置及该原子在分子中的链接情况,通过考察文献[10-12]方法,采用式(1)计算分子顶点特征值 Z_i。

$$Z_i = [(n_i - 1)m_i - h_i]^{1/2} \quad (1)$$

式中 n_i 为顶点原子 i 的电子层数,m_i 为最外层电子数,h_i 为与其直接相连的氢原子数。

对于不同类型分子顶点自身对色谱保留指数的贡献,按式(2)计算:

$$x_k = \sum_{i \in k} Z_i (k = 1,2,3,4) \quad (2)$$

式中 k 为顶点原子 i 所属类型,Z_i 为顶点特征值[按式(1)计算]。分子中最多含 4 种原子类型,每个分子最终可得到 4 个顶点项,分别用 x_1、x_2、x_3 和 x_4 表示。

对于分子顶点间的关系对色谱保留指数的贡献。这里所说分子顶点间的关系并不是原子间某种具体的作用,而是要反映其密切程度与分子顶点特征值的改变趋势一致及与两者距离的改变趋势相反的两方面情况。通常倒数形函数可满足这一要求,采用式(3)表达。

$$x_r = m_{nl} = \sum_{i \in n, j \in l} \frac{Z_i \cdot Z_j}{r_{ij}^2} (n = 1,2,3,4; n \leqslant l \leqslant 4) \quad (3)$$

Z 为分子顶点特征值,按式(1)计算;r_{ij} 是分子顶点 i、j 的相对距离(即所经最短途径相对键长和,相对键长以键长值与碳碳单键键长值之比表示);n 和 l 为分子顶点所属类型。化合物分子中 4 类分子顶点可以产生出 10 种不同相关项:m_{11}、m_{12}、m_{13}、m_{14}、m_{22}、m_{23}、m_{24}、m_{33}、m_{34}、m_{44},其中 m_{11} 表示分子中第一类顶点与第一类顶点之间的相关值,同理 m_{12} 表示分子中第一类顶点与第二类顶点之间的相关值。10 种不同相关项分别记为 x_5、x_6、x_7、x_8、x_9、x_{10}、x_{11}、x_{12}、x_{13} 和 x_{14},这样对于所有的样本最多将产生 14 个变量(结构描述符)来描述分子结构信息,依据以上原理可得本研究样本的结构描述符值(如读者需要可向作者索取)。

1.2.2　模型评价

通常采用“留一法”对模型进行交叉检验以评价模型的稳定性及预测能力,其过程参阅文献[13],计算出交互检验预测值与实验值的复相关系数(R_{CV})及标准偏差(SD_{CV})等统计量,R_{CV} 愈接近 1、SD_{CV} 愈小,表明模型稳定性及预测能力愈好。但一些研究[14,15]认为仅使用训练集的 R_{CV} 及 SD_{CV} 评估模型实际预测能力是不够的,还需要使用外部样本进行验证。模型对外部数据集的预测能力,采用 R_{test} 和 SD_{test} 表示:

$$R_{test} = \sqrt{1 - \frac{\sum_{i=1}^{test} (y_i - y_i)^2}{\sum_{i=1}^{test} (y_i - y_i)^2}} \quad (4)$$

$$SD_{test} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{test} (y_i - y_i)^2} \quad (5)$$

式(4)和(5)中,y_i 和表示预测集的实验值和预测值,表示预测集实验值的平均值。

2　结果与讨论

2.1　多元线性回归模型

在建模之前对变量进行筛选,减少变量数以寻找最佳变量组合是有必要的。首先采用逐步回归(SMR)分析进行筛选变量,以偏 F 检验值对应的显著水平依次将变量引入模型,模型采用“留一法”进行交互检验。为控制模型变量之间的共线性,采用方差膨胀因子(VIF)[10]对模型进行诊断,VIF = $(1-r^2)^{-1}$,其中 r 为某自变量与其它自变量相关系数。一般认为 VIF 值小于 10,表示变量间共线性不明显,所建模型可接受;VIF 值大于 10,表示变量间共线性明显,所得模型不可靠。逐步回归的复相关系数(R/R_{CV})及标准偏差(SD/SD_{CV})随变量引入的变化情况见图 1 及图 2。

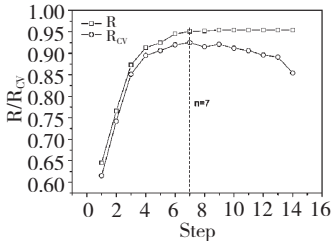

图 1　R 及 R_{CV} 随逐步回归的变化情况

Fig. 1　Changes of R and R_{CV} in SMR

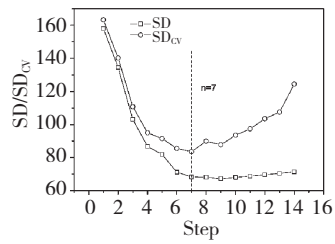

图 2　SD 及 SD_{CV} 随逐步回归的变化情况

Fig. 2　Changes of SD and SD_{CV} in SMR

图 1、图 2 显示逐步回归到第 7 步所得模型(n

=7)各项指标较为理想。此时,R 接近最大值,且 R_{CV}达到最大值;而 SD 接近最小值,且 SD_{CV}达到最小值。7 变量模型如下:

$$RI = 217.931 + 76.733 \times x_1 + 48.697 \times x_2 + 88.957 \times x_4 - 5.605 \times x_7 - 17.487 \times x_8 + 10.697 \times x_{10} + 16.849 \times x_{12} \quad (6)$$

模型拟合:$N=54$,$R=0.951$,$SD=68.298$,$F=61.711$;交互检验:$R_{CV}=0.925$,$SD_{CV}=83.544$,$F_{CV}=39.063$;外部预测:$R_{test}=0.955$,$SD_{test}=57.266$。N 为回归点数,R 为复相关系数,SD 为估计标准偏差,F 为 Fischer 检验值;CV 为交互检验、test 为外部预测。对模型进行变量共线性诊断,发现变量的 VIF 值中最大的为 9.426,说明变量间无明显共线性,模型质量良好。

2.2 偏最小二乘回归模型

将训练集样本结构描述符值作为 X 值,色谱保留指数作为 Y 值,采用 Simca-P 11.5 进行建模,同时对所得 PLS 模型进行检验。综合考虑复相关系数及标准偏差,最后得到由 3 个主成分(A)构建的 PLS 模型,模型的 R、R_{CV}和 R_{test}分别为 0.938、0.900 和 0.940;SD、SD_{CV}和 SD_{test}分别为 71.436、86.642 和 70.084。训练集样本在 PLS 前两个主成分得分散点绘于图 3,样本点全部落在 95% 置信圈内,统计结果表明结构描述符能够恰当表现分子结构特征,并在统计模型中作出正确反映。

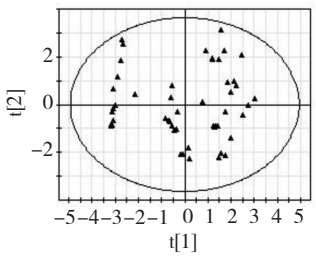

图 3 样本在前两个主成分的得分散点图

Fig. 3 The front two principal components' score distribution plots

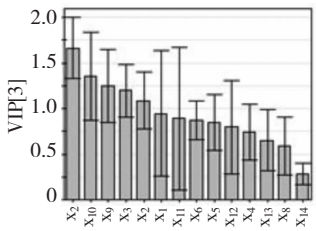

图 4 变量重要性投影

Fig. 4 Importance of variables

变量重要性投影(VIP,图 4)反映变量对 Y 的解释能力,VIP 值大于 1 的变量对 Y 的贡献较大。对于该样本集变量 x_2、x_{10}、x_9、x_3 和 x_7 的 VIP 值大于 1,表明它们对 Y 的解释能力较大。排在前面的 x_2 对应的是第二类分子顶点自身的贡献、x_{10}对应的是第二类和第三类分子顶点相关项的贡献、x_9 为第二类和第二类分子顶点相关项的贡献、x_3 对应的是第三类分子顶点自身的贡献,说明链越长、取代基越多,化合物的色谱保留指数就越大。

2.3 模型比较

上述模型的复相关系数 R、R_{CV}和 R_{test}均较为理想(≥0.9),说明两模型具有良好的估计能力、稳定性和外部预测能力。两模型对训练集样本色谱保留指数的估计值和对测试集样本的预测值分别列于表 1 的 MLR 栏及 PLS 栏,以模型计算值为纵坐标、实验值为横坐标绘图于图 5,可以看出绝大多数样本点落在斜率为 1 的对角线附近,说明模型的预测值较为准确;并且两模型样本点分布情况相似,说明两模型的预测准确性大体相当。误差分布见图 6,以 2 倍 MLR 模型标准偏差(SD)为限,大多数样本的计算误差都未超出此范围,对于 MLR 模型只有 1 个样本(21 号,不足 2%)的计算误差超出 ±2SD,对于 PLS 模型共有 4 个样本(6、21、53 和 63 号,不足 6%)的计算误差超出 ±2SD。个别样本预测值误差

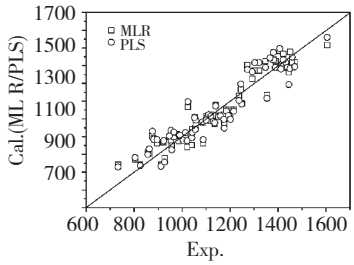

图 5 RI 实验值与计算值相关图

Fig. 5 Calculated vs. experimental values

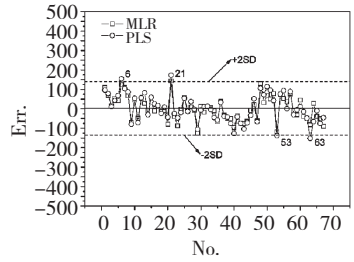

图 6 计算值残差分布

Fig. 6 Comparative residuals vs. compounds No.

稍大,可能是分子结构描述符不够完善,某些对色谱保留行为有影响的特殊结构信息未表达或表达不完全所致。

PLS 模型与 MLR 模型相比,PLS 模型计算误差超出 ±2*SD* 范围的样本数更多,从这点上讲 MLR 模型要略优于 PLS 模型。从复相关系数和标准偏差来看,MLR 模型的 R、R_{CV}、R_{test}均大于 PLS 模型的相应值,MLR 模型的 SD、SD_{CV}、SD_{test}均小于 PLS 模型的相应值,也说明 MLR 模型要略优于 PLS 模型。需特别说明的是,本文研究的样本体系中含烷烃、烯烃、醇、酚、醛、酮、酯等化合物,包含直链、苯环、五元环、六元环、七元环等结构,化合物种类多、分子结构跨度大,对于这样一个复杂的样本体系,两模型的结果应该是满意的。

3 结论

本文通过分子顶点自身和顶点之间的关系对柠檬皮香气成分中的 67 个化合物结构进行了参数化表达,采用多元线性回归和偏最小二乘回归建立了该类化合物结构与色谱保留指数之间的关系模型。分子结构描述符完全来自分子本身的计算,计算过程相对简单。构建的分子结构描述符能够恰当表现该类化合物的结构特征,所建模型能较准确地预测该类化合物在文献所述条件下的色谱保留指数,在一定程度上阐明化合物的色谱保留行为与其分子结构之间的关系。本文对于天然产物中的挥发性化合物的 QSRR 研究具有一定的参考价值。

参考文献

1 Lin H(林洪斌),Cao D(曹东),Chen Y(陈燕),*et al*. Microwave-assisted steam extraction process of lemon essential oil. *China Brew*(中国酿造),2015,34:76-79.

2 Zhang B(章斌),Hou XZ(侯小桢),Qin Y(秦轶),*et al*. Study on chemical composition,antioxidant and antibacterial activities of lemon peel essential oil. *Sci Technol Food Ind*(食品工业科技),2015,36:126-131.

3 Dugo P,Ragonese C,Russo M,*et al*. Sicilian lemon oil:composition of volatile and oxygen heterocyclic fractions and enantiomeric distribution of volatile components. *J Sep Sci*,2010,33:3374-3385.

4 He CF(何朝飞),Ran Y(冉玥),Zeng LF(曾林芳),*et al*. Analysis of aroma components from peels of different lemon varieties by GC-MS. *Food Sci*(食品科学),2013,34:175-179.

5 Qiu SS(邱松山),Jiang CC(姜翠翠),Zhou RJ(周如金),*et al*. A QSAR study on antibacterial activity of p-hydroxybenzoate esters. *Mod Food Sci Technol*(现代食品科技),2014,30:98-102.

6 Li H(李焕),Li MP(李美萍),Zhang SW(张生万). Study on chromatographic separation and quantitative structure-retention relationship of 37 fatty acid methyl esters. *Sci Technol Food Ind*(食品工业科技),2013,34:49-53.

7 Lan ZP(兰作平),Liao LM(廖立敏),Liu Y(刘宇),*et al*. Quantitative structure-retention relationship(QSRR) of volatile constituents of *Meconopsis integrifolia* Franch. *Nat Prod Res Dev*(天然产物研究与开发),2011,23:1091-1094.

8 Li JF,Liao LM. Structural characterization and acute toxicity prediction of substituted aromatic compounds by using molecular vertexes correlative index. *Chinese J Struct Chem*,2013,32:557-563.

9 Liao LM,Zhu J,Li JF,*et al*. QSRR study on components of *Styrax japonicus* sieb flowers using improved molecular electronegativity-distance vector(I-MEDV). *Chinese J Struct Chem*,2011,30:105-110.

10 Du XH(堵锡华). Study on a modified molecular connectivity index for QSAR/QSRR of chlorobenzenes,alcohols and esters. *J Instru Anal*(分析测试学报),2003,22:18-22.

11 Tang ZQ(唐自强),Du XH(堵锡华),Feng CJ(冯长君),*et al*. QSPR study on solubility and octanol-water partition coefficients of benzene halides and esters by molecular connectivity index. *J Shanghai Univ,Nat Sci*(上海大学学报,自科版),2003,9:266-270.

12 Qin S(覃松),Liao L M(廖立敏). Study on boiling point of aldehydes and ketones using a new structural descriptor. *Comput Appl Chem*(计算机与应用化学),2012,29:973-976.

13 Li B(李波),Zeng H(曾晖),Zhou P(周鹏),*et al*. QSAR studies of 7-substituted 20(S)-camptothec in analogues with antitumor activity using three-dimensional holographic vector of atomic interaction field(3D-HoVA IF). *Fine Chem*(精细化工),2006,23:41-46.

14 Tropsha A,Gramatica P,Gombar VK. The importance of being earnest:validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb Sci*,2003,22:69-77.

15 Gramatica P,Pilutti P,Papa E. A tool for the assessment of VOC degradability by tropospheric oxidants starting from chemical structure. *J Chem Inf Comput Sci*,2004,38:6167-6175.