药学学报 Acta Pharmaceutica Sinica 2009, 44 (5): 486−490

基于启发式和基因表达式编程法预测磺胺类药物的 pK_a 值

李玉琴[1*], 司宏宗[2], 肖玉良[1], 刘彩红[1], 夏成才[1], 李 珂[1], 齐永秀[1]

(1. 泰山医学院药学院, 山东 泰安 271016; 2. 青岛大学计算科学与工程技术研究中心, 山东 青岛 266071)

摘要: 应用启发式算法（HM）和基因表达式编程方法（GEP）建立了 31 种磺胺类药物 pK_a 值的定量构效关系模型。用 ChemOffice2004 软件进行化合物的结构输入，利用半经验方法进行分子结构优化，在 CODDESA 软件中计算出组成、拓扑、几何、电子和量子化学参数，并用启发式方法筛选出 4 个相关参数，在此基础上运用多元线性回归和基因表达式编程方法建立 QSPR 模型。两种方法均得到了较好的结果，HM 和 GEP 的的相关系数分别为 0.90 和 0.95。两种 QSPR 模型在新药研究中可以预测化合物的 pK_a 值，将为新药研究提供理论指导。

关键词: 磺胺类药物；pK_a；定量构效关系；启发式算法；基因表达式编程方法

中图分类号: R916 **文献标识码**: A **文章编号**: 0513-4870 (2009) 05-0486-05

Quantitative structure activity relationship models based on heuristic method and gene expression programming for the prediction of the pK_a values of sulfa drugs

LI Yu-qin[1*], SI Hong-zong[2], XIAO Yu-liang[1], LIU Cai-hong[1], XIA Cheng-cai[1], LI Ke[1], QI Yong-xiu[1]

(*1. School of Pharmaceutical Sciences, Taishan Medical College, Tai'an 271016, China;*
2. Institute for Computational Science and Engineering, Qingdao University, Qingdao 266071, China)

Abstract: Quantitative structure-property relationships (QSPR) were developed to predict the pK_a values of sulfa drugs via heuristic method (HM) and gene expression programming (GEP). The descriptors of 31 sulfa drugs were calculated by the software CODESSA, which can calculate constitutional, topological, geometrical, electrostatic, and quantum chemical descriptors. HM was also used for the preselection of 4 appropriate molecular descriptors. Linear and nonlinear QSPR models were developed based on the HM and GEP separately and two prediction models lead to a good correlation coefficient (R) of 0.90 and 0.95. The two QSPR models are useful in predicting pK_a during the discovery of new drugs and providing theory information for studying the new drugs.

Key words: sulfa drug; pK_a; quantitative structure-property relationship; heuristic method; gene expression programming

药物在发挥作用前至少要通过一个生物膜, 这个过程是被动的或需要由某个中间体携带来完成。许多药物都包含离子基团, 而且它们都有特定的pK_a值。药物通常在特定的pK_a值下通过主动运输机制穿透细胞膜或通过毛细孔[1]。磺胺是PABA的抗代谢 物, pK_a直接影响其与蝶酸合成酶的结合。因此, pK_a值是影响药物作用的一个重要因素。通常采用实验 得到pK_a值的方法比较烦琐，发展一种应用方便和精确预测新化合物pK_a值的方法在药物研究中十分重要。

化学和生物效应与分子性质密切相关, 而这些性质可以用多种方法计算或预测得到[2−4]。近年来, 定量结构-性质关系（QSPR, QSRR）及其变化形式已成为一种潜在的有效预测药物活性参数的方法[5−8]。QSPR 方法的优越性在于一旦建立了模型就可以仅通过化合物的结构就可以预测化合物的性质。此种方

收稿日期: 2008-10-29.
基金项目: 泰山医学院博士启动基金资助项目 (2006-742).
*通讯作者 Tel: 86-538-6229753-8004, E-mail: lqq29@163.com

法扩展了合理筛选药物的范围，有助于寻找药物的作用机制。

基因表达式编程方法（GEP）[9]是一种基于自然群体遗传演化机制的高效探索算法, 它摒弃了传统的搜索方式，模拟自然界生物进化过程，采用等长线性符号编码，具有极强的函数发现能力和很高的效率。GEP 在面对复杂问题的求解方面有极大的优越性。目前，GEP 在公式发现、函数挖掘、关联规则发现、因子分解和太阳黑子预测的研究中取得了良好结果[2, 10, 11]。

本文利用启发式算法（HM）和基因表达式编程方法（GEP）建立磺胺类药物 pK_a 值的定量结构关系模型，预测了 31 个磺胺类药物的 pK_a 值，所用的描述符通过 CODESSA 软件计算得到。HM 也用来挑选合适的分子描述符。通过本研究，探讨了建立准确预测 pK_a 值的 QSPR 模型的可能性并比较了两种方法的优劣，同时讨论了影响 pK_a 值的结构因素。

数据集和分子描述符的产生

1 数据集

采用的 31 个磺胺类药物的名称及其相应的 pK_a 值来自于文献[10, 11]，列于表 1。在启发式方法（HM）和基因表达式编程方法（GEP）研究中, 数据集被随机分为两个子集：训练集包含 21 个化合物，用来建立模型；测试集包含 10 个化合物，用于评价所建模型的稳定性和预测能力。

2 分子描述符的产生

分子的二位结构用 ChemOffice2004 软件画出。所有化合物在 Hyperchem7.0 的软件中首先采用分子力学方法 MM+进行初步优化，在此基础上用半经验 PM3 方法进行几何优化，获得能量最低构象。对优化完的分子结构在 MOPAC 6.0 程序中进行计算，将 MOPAC 的结果文件转入到 CODESSA 程序中, 计算 5 类描述符：构成描述符、拓扑描述符、几何描述符、静电描述符和量子化学描述符，共得到 568 个描述符。

计算方法

1 启发式方法（HM）

CODESSA 软件中的 HM 可对大量的分子描述符进行完全搜索，从而建立最佳的线性回归方程。该方法首先对分子描述符进行共线性控制，如任意两个相关系数大于 0.8 的描述符不会同时包含在同一个模型中, 并采用启发式算法对参数进行快速筛选建立最佳模型，而不是考察所有可能的参数组合。HM 根据以下 4 条规则排除一些描述符：① 不是每个化合物都共有的参数；② 对所有化合物来说，数值变化较小的描述符；③ 在一个参数相关方程中，F 检验值小于 1.0 的参数；④ t 检验值小于某一定义值的描述符。模型的好坏由相关系数（R^2）、检验值（F）以及标准偏差（S）等来检验。模型的稳定性用留一法 (Leave-One-Out, LOO) 交互检验的相关系数 R^2_{cv} 来检验。本研究中, 启发式回归结果用误差 (root-mean-square, RMS) 来表示。

Table 1 The name, experimented and predicted pK_a of the compounds by HM and GEP

No.	Name	Exp. pK_a	Calculated			
			HM		GEP	
			Pred[a]	Residue[b]	Pred.	Residue
1	Sulfinpyrazone	2.80	1.90	0.90	2.80	0.00
2	Sulfamethizole	5.40	5.66	−0.26	5.41	−0.01
3	Sulfamethopyrazine	6.10	6.26	−0.16	6.19	−0.09
4	Sulfacetamide	5.40	6.31	−0.91	5.30	0.10
5	Sulfanilic acid	3.20	3.01	0.19	3.34	−0.14
6*	Sulfaethidole	5.40	6.16	−0.76	5.60	−0.20
7*	Sulfamethoxazole	5.60	6.89	−1.29	5.37	0.23
8	Sulfametrole	4.80	5.91	−1.11	5.13	−0.33
9	Sulfasymazine	5.50	6.01	−0.51	5.16	0.34
10	Sulfaguanidine	7.40	6.16	1.24	7.06	0.34
11	Sulfisoxazole	5.00	5.77	−0.77	5.35	−0.35
12*	Sulfadoxine	6.10	5.96	0.14	6.47	−0.37
13*	Sulfacarbamide	5.40	6.05	−0.65	5.02	0.38
14	Sulfamethoxydiazine	7.00	6.18	0.82	7.44	−0.44
15	Sulfadiazine	6.52	5.87	0.65	6.07	0.45
16	Sulfamoxole	6.80	6.89	−0.09	7.28	−0.48
17	Sulfadimethoxine	6.80	6.50	0.30	6.30	0.50
18*	Sulfaquinoxaline	5.50	6.57	−1.07	6.02	−0.52
19	Sulfamerazine	7.10	6.43	0.76	6.67	0.43
20	Dapsone	2.41	3.36	−0.95	3.01	−0.60
21*	Sulfachloropyridazine	6.00	6.15	−0.15	6.65	−0.65
22	Sulfamethazine	7.40	7.20	0.20	6.73	0.67
23*	Sulfamonomethoxine	6.05	5.56	0.49	5.37	0.68
24	Sulfapyridine	8.40	6.53	1.87	7.70	0.70
25	Succinylsulfathiazole	4.50	5.47	−0.97	5.21	−0.71
26	Sulfasalazine	9.70	9.40	0.30	8.96	0.74
27	Sulfaphenazole	6.10	6.10	0.00	6.90	−0.80
28*	Sulfathiazole	7.10	6.68	0.42	6.07	1.03
29*	Sulfamethoxypyridazine	7.20	8.16	−0.96	8.47	−1.27
30	Sulfameter	6.98	6.34	0.64	8.36	−1.38
31*	Sulfisomidine	7.37	6.21	1.16	9.88	−2.51

The star '*' is test set. "a" is the predicted pK_a. "b" = Exp − Pred

2 基因表达式编程（GEP）

GEP 是葡萄牙科学家 Candida Ferreira 于 1999 年发明的一种基于基因组（genome, GA）和表现型 (phoneme, GP) 的新的遗传算法。GEP 主要包括两个方面的内容：染色体和表达树（ETs）。ET 主要用来表达染色体的遗传编码信息。在 GEP 中，有两种语言使用：基因和 ETs 语言。GEP 的实现技术主要包括编码方式、K 表达式、选择算子、变异算子、插串操作、基因倒置、重组算子、多基因染色体及连接函数、基于频繁函数集的标准函数集和用户自定义函数、适应度函数选择等（表 2），经典 GEP 算法有 3 种适应度计算函数，本文采用基于绝对误差的适应度函数：

$$f_i = \sum_{j=1}^{C_j} (M - \left|C_{(i,j)} - T_{(j)}\right|)$$

Table 2 The all parameters and selection of GEP

Parameter	Selection
Division	/
Addition	+
Multiplication	*
Subtraction	−
10^x	Pow10
Cosine	Cos
Logistic	Log
Chromosomes	100
Genes	9
Head size	8
Gene size	26
Linking function	+
Generations without change	200
Number of tries	3
Max. complexity	9
Fitness-function	MSE
Mutation rate	0.044
Inversion rate	0.1
IS transposition rate	0.1
RIS transposition rate	0.1
One-point recombination rate	0.3
Two-point recombination rate	0.3
Gene recombination rate	0.1
Gene transposition rate	0.1
Constants per gene	10
Lower bound	−10
Upper bound	10
RNC mutation	0.01
Dc mutation	0.044
Dc inversion	0.1

结果

1 HM 计算结果

所有 31 个化合物通过 CODESSA 软件计算后总共获得 568 个描述符，用所有计算出的描述符建立预测 pK_a 值的线性模型 (表 3)。为了确定合适的描述符个数, 研究了不同的描述符子集。当添加另外一个描述符对模型的统计性能没有明显改进时，就表明达到了合适的描述符个数。为了避免模型的“过参数化”，R^2 增加小于 0.02 或 R^2_{cv} 降低时被选做极限标准。本研究中最终选取了其中与 pK_a 密切相关的 4 个描述符，4 个描述符的相关矩阵见表 4。从表中可以看出每两个描述符之间的相关系数都小于 0.80，说明了它们是相互独立的[12]。

Table 3 Descriptors, physical-chemical meanings, coefficient, error and *t*-test in HM

No.	Descriptor	Physical-chemical meaning	Coefficient	Error	*t*-Test
0		Intercept	−5.22E+00	3.22E+00	−1.620 8
1	MENA	Max e-e repulsion for a N atom	1.28E−01	2.54E−02	5.054 4
2	ARICA	Avg 1-electron react index for a C atom	7.82E+02	1.58E+02	4.94 6
3	GAMMAP	(1/6)X Gamma polarizability (DIP)	1.35E−04	2.68E−05	5.047 5
4	KHI	Kier&Hall index (order 3)	−1.04E+00	2.33E−01	−4.477 6

Table 4 Crrelation matrix of four descriptors

Descriptor	KHI	GAMMAP	ARICA	MENA
KHI	1.000	0.399	0.294	−0.247
GAMMAP		1.000	0.001	0.029
ARICA			1.000	0.024
MENA				1.000

用这些参数建立的线性模型如下：

$$pK_a = -5.22+1.28\times10^{-1}MENA+7.28\times10^{2}ARICA+1.35\times10^{-4}GAMMAP-1.04KHI$$

训练集: $R = 0.90$, $R_{cv} = 0.66$, F =21.08, RMS = 0.75

测试集: $R = 0.71$, $R_{cv} = 0.58$, F =26.66, RMS = 1.22

图 1 为多元线性回归模型的预测值和实验值的相关图，它包括训练集和测试集共 31 个化合物。这些化合物的 pK_a 预测值见表 1。

2 GEP 计算结果

在建立了线性模型之后，相同的描述符作为 GEP 的变量建立了非线性模型。为了得到满意的结果，优化了影响 GEP 的参数。GEP 所用软件包（automatic problem solver, APS）容易控制，进化模型能够用测试集测试，在进化过程中，主要是对函数进行了很好

的选择，选取了加、减、乘、除、指数、对数和 cos 共 7 个函数，拟合函数是 MSE。通过拟合，对所选的 4 个描述符建立了最佳 QSPR 模型, 其预测值和残差见表 1 和图 2、3。所建模型的统计结果为：训练集: R = 0.95, s = 0.56；测试集中：R = 0.80, s = 1.02。

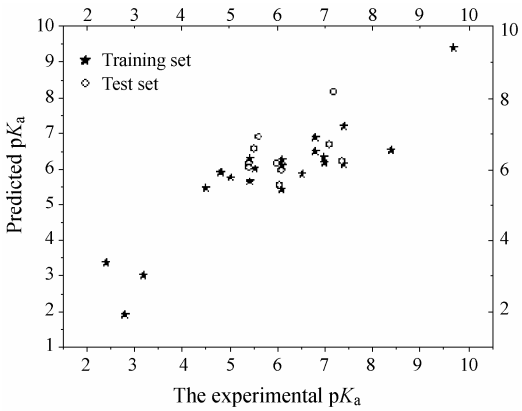

Figure 1 Predicted *vs* experimental pK_a by HM

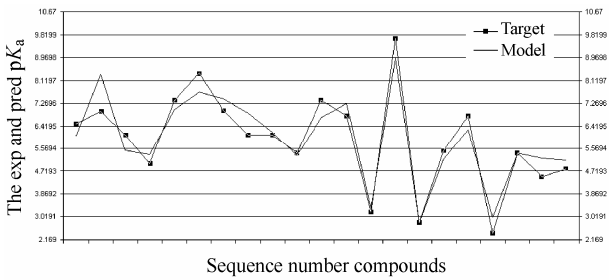

Figure 2 Fitting curve of training set by GEP

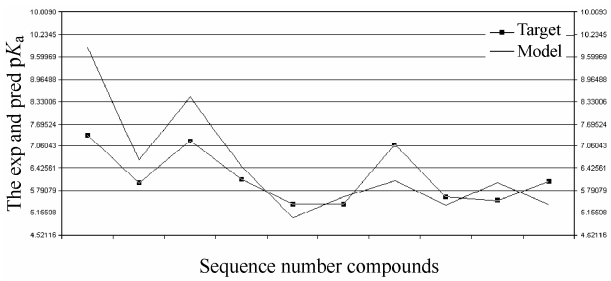

Figure 3 Fitting curve of test set by GEP

用 GEP 算法建立的模型可用下列方程来表示：

```
dblTemp = cos((d[1] + (d[2]*pow(10,pow(10,d[1])))));
dblTemp += cos(((pow(10,d[0])*d[1])/d[0]));
dblTemp += pow(10,(((d[1] + 2.15)/d[3]) −
           cos((2.15/d[3]))));
dblTemp += pow(10,(d[0]*((cos(d[0]) − (d[2] −
           d[1])) − (d[2] + d[3]))));
dblTemp += ((apsLogi((d[3] − d[0])) +
           cos(d[0]))/(4.09*(d[1]*d[3])));
dblTemp += apsLogi(cos((d[2]/pow(10,((d[1]/d[3]) +
           9.99)))));
dblTemp += cos(pow(10,pow(10,d[0])));
dblTemp += cos((d[0]*(((d[2]/d[3]) + 9.67) + 9.99)));
dblTemp += apsLogi(cos((d[2]/pow(10,(d[0] −
           d[3])))));
```

d(0) 到 d(3) 分别是 KHI, GAMMAP, ARICA 和 MENA

3 模型中相关描述符的讨论

通过解释模型中的描述符可以找出影响这类化合物 pK_a 值的结构特征。在模型所选的参数中, MENA、ARICA 和 GAMMAP 为量子化学描述符, KHI 是一个拓扑描述符。MENA 描述了对 N 原子的最大电子-电子排斥能，它对分子中的电子云密度有影响，它的值越大，分子中的电子云密度越大，这势必造成分子中的正负电荷产生的极性越大。由于它在模型中的系数是正的，因此它的增大导致了 pK_a 值的增大。

ARICA 表示 C 原子的平均一电子反应指数，与分子结构中电荷的分布有关。它估计了分子中 C 原子的相对反应能力，而且它与反应的活化能有关。模型中正的相关系数意味着增大这个描述符的值，将使 pK_a 值上升。

GAMMAP 表示分子的 γ 极化度。分子的电荷分布是影响分子极化度的重要因素，而极化度又对质子的脱离产生影响，模型中 GAMMAP 正的系数表明此参数对 pK_a 值是正的贡献。

KHI 三阶指数, 代表了分子的大小、形状和分支度，在一定程度上体现了分子间的色散力。分子的体积越大，分子的色散力越大。由于它在模型中的系数是负的，因此它的增大导致了 pK_a 值的减小，这就意味着色散力有利于从 N 原子上脱离氢原子。

但化合物 **29**、**30**、**31** 的预测值和实验值之间偏差较大, 而且都是负值, 预测值远大于实验值。这可能是由于这 3 个化合物的分子结构较大, 且含有多个 N 原子和 S 原子，从而影响了 3 个化合物的 MENA、ARICA 和 GAMMAP，使其理论 pK_a 值大于实测值。

通过以上结果分析可得出，提出的模型能够正确反映这类化合物的结构性质，而且仅通过结构计算出来的 MENA、ARICA、GAMMAP 和 KHI 4 个分子描述符能够说明影响 pK_a 值的结构特征。

结论

本研究建立了预测磺胺类药物 pK_a 值的 QSPR 模型。提出的线性模型可以识别和提供对这些化合物

的 pK_a 值起作用的描述符。另外，用同样的分子描述符建立的非线性的 GEP 模型体现了更强的预测能力。用所建立的模型可以预测磺胺类药物的 pK_a 值，所以本方法在药物早期研究中具有很好的指导作用。

References

[1] Liu GQ. Pharmacology (药理学) [M]. 2nd ed Beijing: China Medico-Pharmaceutical Science and Technology Publishing House, 2006: 32−45.

[2] Si HZ, Yuan SP, Zhang KJ, et al. Quantitative structure activity relationship study on EC_{50} of anti-HIV drugs [J]. Chemometr Intell Lab Syst, 2008, 90: 15−24.

[3] Si HZ, Wang T, Zhang KJ, et al. Quantitative structure activity relationship model for predicting the depletion percentage of skin allergic chemical substances of glutathione [J]. Anal Chim Acta, 2007, 591: 255−264.

[4] Si HZ, Wang T, Zhang KJ, et al. QSAR study of 1,4-dihydropyridine calcium channel antagonists based on gene expression programming [J]. Bioorg Med Chem, 2006, 14: 4834−4841.

[5] Mercader AG, Duchowicz PR, Fernández FM, et al. Modified and enhanced replacement method for the selection of molecular descriptors in QSAR and QSPR theories [J]. Chemometr Intell Lab Syst, 2008, 92: 138−144.

[6] Xu J, Liang H, Chen B, et al. Linear and nonlinear QSPR models to predict refractive indices of polymers from cyclic dimer structures [J]. Chemometr Intell Lab Syst, 2008, 92: 152−156.

[7] Pan Y, Jiang JC, Wang R, et al. Advantages of support vector machine in QSPR studies for predicting auto-ignition temperatures of organic compounds [J]. Chemometr Intell Lab Syst, 2008, 92: 169−178.

[8] Porto LC, Souza ES, Junkes B da S, et al. Semi-empirical topological index: development of QSPR/QSRR and optimization for alkylbenzenes [J]. Talanta, 2008, 76: 407−412.

[9] Koza JR. Genetic Programming II: Automatic Discovery of Reusable Programs [M]. Cambridge: The MIT Press, 1994.

[10] Yu JJ, Qin BT, Jiang XD. Time series prediction of grid host load based on gene expression programming [J]. Comput Eng Sci (计算机工程与科学), 2008, 30: 105−107.

[11] Zeng X, Hu JH, Duan L. A prediction method of telephone traffic based on gene expression programming [J]. Comput Simul (计算机仿真), 2008, 25: 170−173.