第 12 卷 第 9 期 物 理 化 学 学 报 Vol.12, No.9
1996 年 9 月 ACTA PHYSICO-CHIMICA SINICA Sep., 1996

# 分子相似性和取代苯酚 $pK_a$ 值的预测 *

张向东
(辽宁大学化学系，沈阳 110036)

关键词： 分子相似性， 神经网络， 取代苯酚

$pK_a$ 值是有机酸碱强度的标度. 在给定 pH 值的情况下，体系中酸或碱解离部分的量与未解离部分量的比值将取决于化合物的 $pK_a$ 值. 这个比值与化合物的溶解、吸附、生物富集、毒性等性质密切相关，因此， $pK_a$ 是环境化学，药物设计等学科中重要物性数据. $pK_a$ 值的估算一般用线性自由能关系 (LFER) 法. 其估算的平均误差为 ±15%. 误差的根本来源是由于 LFER 假定取代基常数 ($\sigma$) 和反应常数 ($\rho$) 是可以分离的. $\sigma$ 绝对值大的取代基，对假定的偏离大误差也大，多官能团取代的化合物误差也大. 近年来 Dixon 提出通过计算原子电荷估算 $pK_a$ 值的方法 [1]. 对 $pK_a$ 变化范围为 1–19 的含氧有机酸，估算误差在 0.5 单位以内，但这种方法需计算原子电荷比较麻烦. 本文提出用相似系数 - 神经网络法 (SC-ANN) 预测多取代酚类化合物的 $pK_a$ 值.

## 1 相似系数

相似系数的计算有多种方法，有代表性如 Euclidian 距离和 Tanimoto 系数等，本文提出以公式 (1) 计算化合物间相似系数.

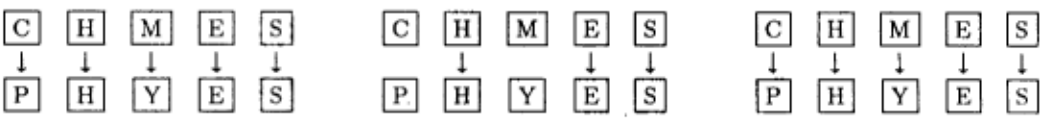
$$SC_{ij} = \frac{\sum A_{ki}}{\sum W_{mi} + \sum W_{mj} - \sum A_{ki}} \tag{1}$$

其中 $SC_{ij}$ 为样本 $i$ 相对于参照样本 $j$ 的相似系数； $\sum W_{mi}$ 、$W_{mj}$ 为样本 $i$ 和 $j$ 在其全部特征码位格上的数值求和；其中 $A_{ki} = 2W_{ki}\sqrt{(W_{ki} \times W_{kj})}/(W_{ki} + W_{kj})$ 为同位格码的几何平均值与算术平均值之比乘以同位格样本 $i$ 的数值. 式中对所有同位格 $A$ 值求和. 显然 $SC_{ij}$ 小于或大于 1.0 的偏离程度表示样本 $i$ 和 $j$ 的相似 / 不相似程度， $SC_{ij}$ 越接近于 1.0, $i$ 越相似于 $j$. 相似系数 $SC$ 与 Euclidian 距离和 Tanimoto 系数比较，均有相同和不同之处 (见图 1). Euclidian 距离是系统 $i$ 与系统 $j$ 按序逐个匹配比较. Tanimoto 系数是系统 $i$ 与系统 $j$ 相比较其共性占的比重. 本文所提出的相似系数 $SC$ 既是逐一匹配比较又量度共性占的比重. Tanimoto 系数要求 $i$ 与 $j$ 相同的位格上数值是相同的， $SC$ 系数 $i$ 与 $j$ 相匹配的位格可以有不同的数值. 例如取代苯酚有 5 个苯环上不同位置，可能有 13 种取代基，若用 $SC$ 系数仅用 5 个码位，而 Tanimoto 系数就需要 5×13 个码位.

## 2 相似系数 - 神经网络法 SC-ANN

1996-01-24 收到初稿， 1996-04-15 收到修改稿. * 沈阳市科委科技基金资助项目

SC-ANN 法以相似系数作为 BP 网络的输入参数经训练预测物性，其基本步骤为四步．①将样本集的每个样本编码：抽取样本的结构特征，确定位格的数目并排序，赋以每个位格以一定数值． $n$ 个样本和 $m$ 个位格，组成 $n \times m$ 原始数据矩阵．②在样本集内选择若干样本作为参照样本，以参照样本的全体能覆盖总样本的结构特征和目标函数变化范围为原则来挑选．③根据式 (1)，计算样本集的每个样本相对于参照样本的相似系数， $n$ 个样本， $p$ 个参照样本，得到 $n \times p$ 相似系数矩阵．④以样本的相似系数为变量．若有 $p$ 个参照样本，则每个样本有 $p$ 个变量作为 BP 网络的输入参数，物性数据作为目标函数输出，选取样本集中 1/2 左右的样本组成训练集 (其中包含全体参照样本)．全部样本组成预测集，选定合适的网络结构和网络参数，进行训练和预报．




$$D = \sum \sqrt{(W_i - W_j)^2} \quad ST = \sum W_k / (\sum W_i + \sum W_j - \sum W_k) \qquad \text{式 (1)}$$

**图 1 三种相似系数的比较**

**Fig.1 Comparison of three similarity**

**表 1 取代基的 $F$, $R$ 值**

Table 1 The $F$ and $R$ constants for substituents

| Substituent | $F$ | $R$ | Substituent | $F$ | $R$ |
|---|---|---|---|---|---|
| F | 0.45 | 0.432 | OEt | 0.26 | 0.316 |
| Cl | 0.42 | 0.642 | $NO_2$ | 0.65 | 0.979 |
| Br | 0.45 | 0.611 | CN | 0.51 | 1.000 |
| I | 0.42 | 0.589 | CHO | 0.33 | 0.937 |
| OMe | 0.29 | 0.253 | H | 0.03 | 0.842 |
| Me | 0.01 | 0.653 | OH | 0.33 | 0.105 |
| Ph | 0.12 | 0.705 | | | |

**表 2 四种化合物的分子编码矩阵 (示例)**

Table 2 Molecular coding matrix for four compounds

| No. | Compound | $F2$ | $F3$ | $F4$ | $F5$ | $F6$ | $R2$ | $R3$ | $R4$ | $R5$ | $R6$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 4 | phenol | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 | 0.842 | 0.842 | 0.842 | 0.842 | 0.842 |
| 9 | 4-Br | 0.03 | 0.03 | 0.45 | 0.03 | 0.03 | 0.842 | 0.842 | 0.611 | 0.842 | 0.842 |
| 37 | 2-$NO_2$ | 0.65 | 0.03 | 0.03 | 0.03 | 0.03 | 0.979 | 0.842 | 0.842 | 0.842 | 0.842 |
| 51 | 2-$NO_2$,4,6-Br2 | 0.65 | 0.03 | 0.45 | 0.03 | 0.45 | 0.979 | 0.842 | 0.611 | 0.842 | 0.611 |

**表 3 四种化合物的相似系数矩阵**

Table 3 Similarity coefficient matrix of four compounds

| No. | 1 | 4 | 12 | 15 | 20 | 24 | 28 | 31 | 36 | 51 | 58 | 59 | 61 | 64 | 66 | 69 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4 | 1.01 | 1.00 | 0.97 | 0.96 | 0.87 | 0.84 | 0.95 | 0.90 | 0.84 | 0.77 | 0.79 | 1.05 | 0.95 | 0.78 | 0.95 | 0.68 |
| 9 | 1.09 | 0.94 | 0.91 | 0.89 | 0.82 | 0.79 | 0.96 | 0.95 | 0.91 | 0.81 | 0.82 | 0.97 | 0.88 | 0.73 | 0.89 | 0.65 |
| 37 | 1.01 | 0.99 | 0.97 | 0.96 | 0.87 | 0.85 | 1.11 | 1.05 | 1.00 | 0.91 | 0.92 | 1.04 | 0.95 | 0.91 | 1.10 | 0.79 |
| 51 | 1.02 | 0.88 | 0.86 | 0.85 | 0.78 | 0.76 | 1.06 | 1.04 | 1.02 | 1.00 | 1.00 | 0.92 | 0.84 | 0.89 | 0.95 | 0.72 |

表 4 取代苯酚类化合物的 $pK_a$ 值预测

Table 4 Prediction of $pK_a$ for substituted phenols using SC-ANN(training set)

| No. | X | $pK_a$ | | No. | X | $pK_a$ | |
|---|---|---|---|---|---|---|---|
| | | exp. | SC-ANN | | | exp. | SC-ANN |
| 1 | 4-OMe | 10.21 | 10.19 | 39 | 2-$NO_2$,5-OMe | 7.09 | 6.68 |
| 4 | Unsubstitute | 10.0 | 10.1 | 41 | 2-$NO_2$,5-ph | 6.73 | 7.11 |
| 6 | 4-F | 9.89 | 9.98 | 43 | 2-Cl,4-Cl,6-Cl | 6.23 | 6.15 |
| 8 | 3-OEt, 5-OEt | 9.37 | 9.53 | 47 | 2-$NO_2$,6-Cl | 5.48 | 5.43 |
| 12 | 5-F | 9.29 | 9.42 | 49 | 2-Cl,6-Cl | 6.79 | 6.93 |
| 15 | 3-I | 9.06 | 9.04 | 51 | 2-$NO_2$,4-Br,6-Br | 4.71 | 4.77 |
| 17 | 3-Cl | 8.79 | 9.02 | 52 | 2-Cl,4-$NO_2$,6-Cl | 3.55 | 3.85 |
| 20 | 3-CN | 8.61 | 8.74 | 57 | 2-$NO_2$,4-Cl | 6.46 | 6.27 |
| 22 | 2-I | 8.51 | 8.44 | 58 | 2-Cl,4-CN,6-Cl | 4.38 | 4.14 |
| 24 | 3-$NO_2$ | 8.36 | 8.19 | 59 | 3-Me | 10.00 | 10.00 |
| 25 | 3-Cl,5-Cl | 8.19 | 8.09 | 61 | 3-OH | 9.81 | 9.67 |
| 28 | 2-Cl,4-ph | 8.07 | 8.53 | 63 | 3-OH,5-OH | 9.35 | 9.27 |
| 30 | 4-CN | 7.97 | 7.99 | 64 | 2-CHO,6-$NO_2$ | 6.00 | 5.89 |
| 31 | 2-Cl,4-Cl | 7.89 | 7.63 | 66 | 2-CHO,6-$CH_3$ | 10.40 | 9.56 |
| 34 | 2-Cl,3-Cl | 7.69 | 7.72 | 67 | 2-CHO,6-Cl | 7.80 | 7.91 |
| 36 | 2-$NO_2$,4-OMe | 7.31 | 7.20 | 69 | 2-CHO,3-$NO_2$,5-$NO_2$ | 2.60 | 2.66 |

表 5 取代苯酚类化合物的 $pK_a$ 值预测

Table 5 Prediction of $pK_a$ for substituted phenols using SC-ANN(prediction set)

| No. | X | $pK_a$ | | No. | X | $pK_a$ | |
|---|---|---|---|---|---|---|---|
| | | exp. | SC-ANN | | | exp. | SC-ANN |
| 2 | 4-OEt | 10.13 | 10.24 | 35 | 2-Cl,5-Cl | 7.51 | 7.22 |
| 3 | 2-OEt | 10.11 | 10.29 | 37 | 2-$NO_2$ | 7.23 | 7.18 |
| 5 | 2-OMe | 9.98 | 9.92 | 38 | 4-$NO_2$ | 7.16 | 7.73 |
| 7 | 3-OMe | 9.65 | 9.53 | 40 | 2-$NO_2$,4-ph | 6.74 | 6.64 |
| 9 | 4-Br | 9.36 | 9.77 | 42 | 2-Br,6-Br | 6.67 | 6.67 |
| 10 | 3-OMe,4-OMe | 9.35 | 9.85 | 44 | 2-Cl,4-Br,6-Cl | 6.21 | 6.18 |
| 11 | 4-I | 9.30 | 9.42 | 45 | 2-$NO_2$,6-F | 6.07 | 5.86 |
| 13 | 4-Cl | 9.20 | 9.25 | 46 | 2-$NO_2$,5-Cl | 6.05 | 5.55 |
| 14 | 4-OMe,5-Br | 9.09 | 9.62 | 48 | 2-CN | 6.86 | 7.19 |
| 16 | 3-Br | 9.03 | 8.93 | 50 | 2-Cl,4-$NO_2$ | 5.45 | 5.27 |
| 18 | 2-F | 8.73 | 8.20 | 53 | 2-Br,4-$NO_2$,6-Br | 3.39 | 3.25 |
| 19 | 2-Br,4-$OCH_3$ | 8.63 | 8.74 | 54 | 2-I,4-$NO_2$,6-I | 3.32 | 3.29 |
| 21 | 2-Cl,4-Cl | 8.58 | 8.44 | 55 | 3-OEt | 9.66 | 9.61 |
| 23 | 2-Br | 8.44 | 8.24 | 56 | 2-Cl,4-Br | 7.64 | 7.62 |
| 26 | 3-I,5-I | 8.10 | 8.35 | 60 | 4-$CH_3$ | 10.20 | 10.35 |
| 27 | 2-Cl | 8.10 | 8.49 | 62 | 4-OH | 10.35 | 10.05 |
| 29 | 3-Br,5-Br | 8.06 | 7.97 | 65 | 2-CHO,3-$CH_3$,5-$CH_3$ | 10.40 | 9.22 |
| 32 | 3-Cl,4-Cl,5-Cl | 7.84 | 7.95 | 68 | 2-CHO,5-$NO_2$ | 7.40 | 6.97 |
| 33 | 2-Br,4-Br | 7.79 | 7.34 | | | | |

## 3 取代苯酚化合物 $pK_a$ 值的预测

影响酚类化合物 $pK_a$ 值的主要因素是取代基的位置及取代基的诱导和共轭效应[1]. 因此, 本文选用场 / 诱导 (field/inductive) 参数 $F$ 和共振效应 (resonance effect) 参数 $R$ 作特征码

位格上的数值，数据取自文献[2]. 由于取代基的可能位置是苯酚环上 2 至 6 的五个位置，因此，每个化合物共 10 个位格码，依次顺序为 $F2, F3, F4, F5, F6, R2, R3, R4, R5, R6$. 因式 (1) 要求 $W_{ki}$ 取正值，计算时首先将参数 $F$ 和 $R$ 数据作范围标度化处理. 标度化后的 $F$, $R$ 值列于表 1. 表 2 中列出 4 种化合物的分子编码矩阵 (全体样本组成 69 × 10 的矩阵). 根据分子编码矩阵，选择 16 个样本化合物 (见表 3) 作为参照样本，经式 (1) 计算得到 69×16 的相似系数矩阵. 部分示例见表 3. 相似系数矩阵的行矢量代表一个样本，以 16 个相似系数作为误差反向传播神经网络的输入变量，$pK_a$ 值作为网络的输出. 网络结构为 16-8-1, 网络参数：学习速率 $\alpha$=0.1, 动量项 $\eta$=0.9. 训练集样本数 32 个 (表 4), 预测集由全部 37 个未知样本构成 (表 5). 训练集中实验值与预测值拟合的相关系数为 $r$=0.9927, 标准偏差 $SD$=0.241, $F$=2096($n$=32); 预测集未知样本的拟合结果为 $r$=0.9813, $SD$=0.344, $F$=913.9($n$=37).

## 参 考 文 献


1 Dixon S L, Jurs P C. *J. Comput. Chem.*, **1993, 14**(12):1463
2 Hansch C, Leo A, Taft R W. *Chem. Rev.*, **1991, 91:**165


## Molecular Similarity and Prediction of p$K_a$ for Substituted Phenols


Zhang Xiangdong
*(Department of Chemistry, Liaoning University, Shenyang, 110036)*



**Abstract** A procedure is presented for the prediction of physical property of organic compound and QSAR/QSPR analysis based on the similarity indices using a back-propagation neural network. The similarity indices were calculated on a chosen set of structural descriptors by equation 1 and used to quantify the similarity or dissimilarity of organic compound. The similarity indices were also used as an input parameter of neural network. The p$K_a$ values of the 69 substituted phenols were predicted by using 32 compounds as a training set and all 69 compounds as a predicting set. The results obtained were satisfying.



**Keywords:** Molecular similarity, Phenols, Neural network