第 15 卷 第 5 期 物 理 化 学 学 报 Vol. 15, No. 5
1999 年 5 月 ACTA PHYSICO-CHIMICA SINICA May, 1999

用氨基酸加和法计算多肽的脂水分配系数*

陶 鹏 王任小 来鲁华
(北京大学物理化学研究所,北京 100871)

关键词: 分配系数, 疏水性, 多肽, 进化验证, 氨基酸加和法
学科代码: B031002

在药物设计中, 化合物的疏水性是值得考虑的一种重要的性质. 目前常使用化合物在正辛醇和水两相间的分配系数的对数值 ($\log P$) 来度量其疏水性. 仅从化合物的结构出发来预测其脂水分配系数具有重要的意义,已有多种计算方法见诸报导[1,2]. 对于普通的有机化合物,它们能给出较好的结果.

多肽是一类具有重要生物功能的化合物. 在药物设计中,多肽也是常用的物系. 目前预测一般有机化合物脂水分配系数的方法对于多肽尚不能给出满意的结果. 鉴于多肽类化合物特殊的重要性, 专门发展一种方法来预测多肽的脂水分配系数也是十分必要的. 在这方面, 以 Akamatsu 等人所做的一系列工作最具代表性[3-6]. 他们合成了许多小肽, 并测定其脂水分配系数值,然后使用线性回归分析来得出一个经验模型. 但是他们的模型较为繁琐,且通用性不好. 从实用的角度来考虑,仍需大力改进.

在本文中, 我们提出了一种利用氨基酸贡献的简单加和来得出多肽的脂水分配系数的新方法. 各种氨基酸的贡献大小由对训练集的回归分析得出. 该方法简单清晰, 预测精度较高, 对于研究多肽的疏水性可能具有重要的价值.

1 计算方法与结果

1.1 训练集的构建

我们从文献[3-8]中共搜集到 219 个多肽疏水常数, 构建回归所用的训练集. 其中包含了从二肽到五肽大小不同的分子, 含有 21 种常见的天然氨基酸. 这些多肽既有自由形式 (unblocked),又有端基保护形式(blocked). 自由形式的肽端为氨基($-NH_2$)与羧基($-COOH$). 对于保护形式的肽,氮端为乙酰基保护($-NHCOCH_3$),羧端为氨基保护($-CONH_2$).

多肽的疏水常数包括 $\log P$ 与 $\log D$ 值. 分配系数 (P) 为多肽分子在两相分配平衡时其中性形式的浓度比, 分配比 (D) 为多肽分子在分配平衡时两相各种形式 (含中性及电离形式) 的总浓度比.

1998-07-01 收到初稿,1998-08-25 收到修改稿. 联系人:来鲁华. * 863"高技术资助项目

1.2 具体模型

我们认为,多肽的脂水分配系数可以由组成多肽的各氨基酸的贡献加和得到. 基于这一思想,我们提出的氨基酸加和模型共选用了二十三个回归参数,其中包括二十一种基本氨基酸类型. 余下的两个参数 B 与 U 用于标明多肽的形式:对于自由的肽, B 取 0, U 取 1,对于保护形式的肽, B 取 1,而 U 取 0. 具体计算式为

$$\log P = \sum a_n A_n^p + B_p + U_p \qquad (1)$$

式中, $A^p{}_n$ 为第 n 个氨基酸对 $\log P$ 的贡献值, a_n 为此氨基酸的出现次数. B_p 为标志变量 B 对 $\log P$ 的贡献值, U_p 为标志变量 U 对 $\log P$ 的贡献值.

与此类似, $\log D$ 的计算式为

$$\log D = \sum a_n A_n^D + B_D + U_D \qquad (2)$$

1.3 回归分析结果

对以上两个模型分别进行回归分析,结果如下:

对于模型 1($\log P$) $N = 219$, $r = 0.978$, $s = 0.21$, $F = 189.8$

对于模型 2($\log D$) $N = 216$, $r = 0.974$, $s = 0.21$, $F = 155.8$

我们用计算值与实验值进行了拟合,并绘出相应的图形 (图 1 与图 2). 由图可见,计算值和实验值吻合得很好. 所得各参数的基本贡献值见表 1(表中置信区间的置信度为 95%).

我们对回归结果进行了交叉验证 (leaving-one-out cross validation), 对于模型 1 交叉验证的结果为 $N = 219$, $r = 0.974$, $s = 0.23$; 对于模型 2 的结果为 $N = 216$, $r = 0.969$, $s = 0.23$. 由这些结果可见,我们的方法具有较高的预测能力.

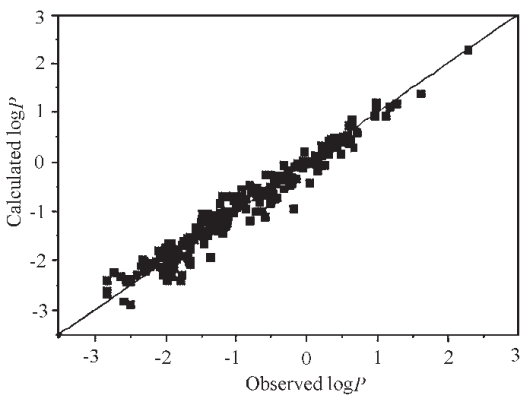

图 1 多肽 log P 计算值与实验值拟和图

Fig. 1 Correlation between the calculated and the experimental log P values of 219 peptides

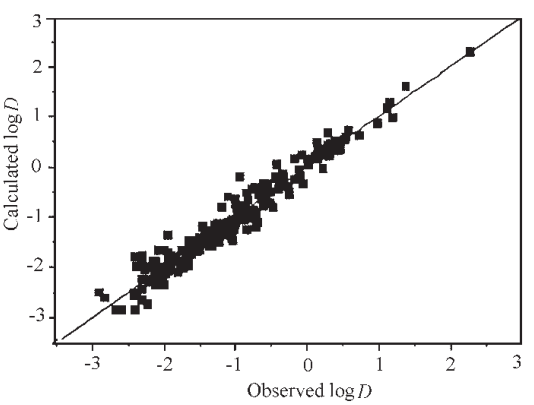

图 2 多肽 log D 计算值与实验值拟和图

Fig. 2 Correlation between the calculated and experimental log D values of 216 peptides

2 讨论

2.1 与先前工作的对比

Akamatsu 等人经过分析 124 个不含极性侧链的二肽至五肽的 $\log P$ 值,提出了计算式[5]

$$\log D = \underset{(0.064)}{0.942}\sum \pi - \underset{(0.096)}{0.582} I_{pep} + \underset{(0.089)}{0.546} E_s'^{c}(R_N) + \underset{(0.071)}{0.295}\left[\sum E_s'^{c}(R_M) + E_s'^{c}(R_c)\right] +$$
$$\underset{(0.172)}{0.516} I_{turn} + \underset{(0.211)}{0.764}\log f_{i+2} + \underset{(0.089)}{0.144} I_Y + \underset{(0.106)}{0.378} I_W + \underset{(0.199)}{1.158}(I_S + I_T) - \underset{(0.225)}{0.807} I_{P(N)} -$$
$$\underset{(0.118)}{0.346} I_{p(pep)} - \underset{(0.190)}{3.866}$$

$N = 124$, $r = 0.967$, $s = 0.209$, $F = 134$

其中, $\Sigma\pi$ 为处于保护形式的单个氨基酸以甘氨酸为基准的 log D 值之和，I_{pep} 表征了多肽的长度．参数 $E_s'^c$ 则表征了各个残基侧链的空间效应，R_N、R_C、R_M，分别为 N 端残基、C 端残基与除 N 端与 C 端外其它残基的校正因子 (对于二肽，R_M 取值为零)．I_{turn} 与多肽形成 β-turn 的能力有关．f 值是各个残基在 β-turn 出现的相对频率．I_Y、I_W、I_M、I_S 与 I_T 则是酪、色、蛋、丝与苏氨酸在多肽中出现与否的标志变量．$I_{P(N)}$ 为脯氨酸是否在 N 端出现的标志变量，$I_{p(pep)}$ 则是与有脯氨酸出现的多肽的残基数目相关．

表 1　氨基酸加和法的参数贡献值

Table 1　Hydrophobicity contributions of 21 natural amino acids

Amino acids	log P contribution	log D contribution
Ala	−0.27(±0.06)	−0.27(±0.07)
Arg	−0.79(±0.21)	−1.65(±0.22)
Asn	−0.98(±0.22)	−0.98(±0.22)
Asp	−0.28(±0.21)	−2.06(±0.22)
Cys	0.83(±0.33)	0.82(±0.34)
Gln	−1.00(±0.21)	−1.00(±0.22)
Glu	−0.34(±0.21)	−2.19(±0.22)
Gly	−0.22(±0.06)	−0.22(±0.06)
His	−0.31(±0.19)	−0.44(±0.20)
Ile	0.70(±0.06)	0.69(±0.06)
Leu	0.80(±0.06)	0.80(±0.06)
Lys	0.17(±0.19)	−2.27(±0.20)
Met	0.51(±0.14)	0.51(±0.14)
Phe	1.16(±0.06)	1.16(±0.06)
Pro	0.15(±0.13)	0.15(±0.13)
Ser	−0.45(±0.15)	−0.45(±0.15)
Thr	−0.26(±0.14)	−0.26(±0.14)
Trp	1.46(±0.11)	1.46(±0.11)
Tyr	0.55(±0.09)	0.55(±0.09)
Val	0.32(±0.06)	0.32(±0.06)
Orn	−0.29(±0.21)	−2.17(±0.27)
Blocked	−1.19(±0.12)	−1.18(±0.12)
Unblocked	−3.25(±0.15)	−3.25(±0.15)

他们又考察了九十个包含有可离子化侧链的被保护的二肽与三肽，并提出了一个类似的式子[6]

$$\log D_{(pH=7)} = \underset{(0.045)}{1.045}\sum\pi - \underset{(0.057)}{0.584}I_{tri} + \underset{(0.045)}{0.246}\sum E_s' + \underset{(0.086)}{0.068}I_Y + \underset{(0.080)}{0.242}I_W + \underset{(0.111)}{1.475}(I_S + I_T)$$
$$+ \underset{(0.155)}{1.749}I_N + \underset{(0.126)}{1.154}I_Q + \underset{(0.092)}{0.607}I_M - \underset{(0.105)}{1.091}I_O - \underset{(0.078)}{1.755}I_K + \underset{(0.105)}{0.274}I_R + \underset{(0.084)}{0.316}I_H -$$
$$\underset{(0.106)}{0.519}I_D - \underset{(0.089)}{1.222}I_E - \underset{(0.077)}{2.348}$$
$$N = 87, \quad r = 0.996, \quad s = 0.085, \quad F = 619$$

其中，I_{tri} 是标志变量：二肽取值为 0，三肽取值 1. I_Y、I_W、I_S、I_T、I_N、I_Q、I_M、I_O、I_K、I_R、I_H、I_D 与 I_E 为各个氨基酸的标志变量 .

由上可见，他们的工作细致入微，但考虑的因素过于庞杂，有些项缺乏明确可信的物理意义 . 提出的方程虽相关性较高，但是是对不同种的肽分别得出的，略显繁琐，且通用性不高 .

比较之下，我们提出的模型简单明了，预测结果准确度较高 . 这说明氨基酸加和模型基本符合实际情况 . 并且回归所得的各氨基酸的疏水贡献值与通常人们所认识的氨基酸的疏水性基本相符，有希望成为药物设计中的一种实用方法 .

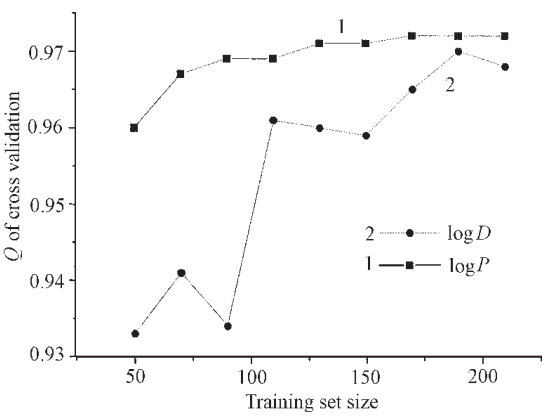

图 3　进化验证 Q 随训练集大小的变化趋势

Fig. 3　Q value *vs* training set size

2.2　一种新的验证方法

我们在本文中所提出的氨基酸加和法本质上是一种经验方法 (empirical method)，其做法是通过分析一组已知实验数据的样本 (即训练集) 得出回归模型，进而达到预测的目的 . 所以，训练集的选取对最终的回归模型起着举足轻重的作用 . 一般认为，训练集中应当包含有 "足够多" 的样本才能保证回归结果的可靠 . 但是，对于 "足够多" 的看法，即假设训练集中的样本数为 N，回归模型中的因子数为 M 时，N 与 M 合适的比率，目前还没有统一、明确的看法 .

为了解决这一问题，我们提出了一种能够检验回归模型稳定性的检验方法——进化验证 (evolution test)，目的是探讨训练集的大小对回归模型预测能力的影响 .

具体做法如下述 .

设已有一个训练集 (母集)，含有 N 个数据 (N 足够大) . 先取一个合适的数 n_1 ($n_1 < N$) . 从母集中随机抽取 n_1 个数据构成一子集 . 以此子集为训练集做回归分析，得出一套参数值 . 用此次回归的结果对此子集的补集中的所有数据进行预测，并比较计算值与实验值 . 可随机抽取若干次，做相同的分析 . 求出本轮分析各项指标的平均值，这些平均值即可认为代表了此模型与训练集在 n_1 水平下所具有的预测能力 . 将 n_1 加一适当步长，得 n_2. 取 n_2 个数据做类似的分析 . 重复上一步，得 n_3、n_4……直至 N 的分析结果 .

整个工作完成之后，可将每轮的结果对抽取的训练集的大小 n 作图 . 通过观察其变化趋势，来判断模型的预测能力，及训练集大小对回归结果的影响 . 可以猜想对于一个良好的、稳定的模型，其预测能力应当是随着训练集的逐步增大而稳步提高的 .

我们对本文中所提出的氨基酸加和模型进行了这样的进化验证．所用的母集包括 219 个样本，取 n_1 为 50，步长为 20，一共做了 9 轮检验，每轮做 20 次回归．图 3 是各轮用回归结果预测多肽疏水常数的 Q 值随 n 大小的变化图．由图可以看出，该模型的预测能力随着训练集的增大而增大，这一点与预计相符．同时我们也注意到，当训练集的大小超过 100 时，即 $N/M>5$ 时，模型预测能力的变化趋于平缓．这一现象说明，对于我们这个模型而言，训练集中的样本为 100～200 之间便已足够．刻意追求更大的训练集并无必要．这样的信息是传统的检验方法(如交叉验证)所无法给出的，显示了进化验证所具有的独特意义．

参 考 文 献

1 Leo A J. *Chemical Reviews,* **1993, 4:** 1281
2 Renxiao Wang, Ying Fu, Luhua Lai. *J. Chem. Inf. Comput. Sci.,* **1997, 37:** 615
3 Akamatsu M, Yoshida Y, Nakamura H, *et al. Quant. Struct. -Act. Relat.,* **1989, 8:** 195
4 Akamatsu M, Okutani S, Nakao K, *et al. Quant. Struct. -Act. Relat.,* **1990, 9:** 189
5 Akamatsu M, Fujita T. *J. Pharm. Sci.,* **1992, 81:** 164
6 Akamatsu M, Katayama T, Kishimoto D, *et al. J. Pharm. Sci.,* **1994, 83:** 1026
7 Buchwald P, Bordor N. *Proteins: Struct. Funct. Genet.,* **1998, 30:** 86
8 Sotomatsu Niwa T, Ogino A. *J. Mol. Struct.,* **1997, 392:** 43

Calculation of Peptide's Partition Coefficients by Amino Acid Addition Model

Tao Peng　Wang Renxiao　Lai Luhua
(Institute of Physical Chemistry, Peking University, Beijing　100871)

Abstract　Based on amino acid addition model, a set of hydrophobicity contributions of amino acids was obtained from the multivariate linear regression analysis of peptides' octanol/water partition coefficients. Multivariate regression was performed on a training set of 219 peptides including dipeptides to pentapeptides which we compose 21 natural amino acids. The correlation coefficients for the whole set fitting are 0.978 and 0.974, for $\log P$ and $\log D$ respectively. In addition, a new test method ~ evolution test ~ for regression analysis was discussed. The result of evolution test for amino acid addition model shows the advantage of this new test method.

Keywords: Partition coefficient, Hydrophobicity, Amino acid addition model, Peptides, Evolution test

Received 1998-07-01, revised 1998-08-25. Correspondent: Lai Luhua.