生物工程学报
journals.im.ac.cn
cjb@im.ac.cn

Chin J Biotech 2008, November 25; 24(11): 1968-1974
Chinese Journal of Biotechnology ISSN 1000-3061

研究报告

基于不同标度伪氨基酸组成预测脂肪酶的类型

张光亚, 李红春, 高嘉强, 方柏山

华侨大学生物工程与技术系, 厦门 361021

摘　要: 从序列出发预测某蛋白质是否为脂肪酶以及属于哪种脂肪酶具有重要的理论和应用价值。提出了基于 Z 标度和 T 标度的伪氨基酸组成方法提取序列特征值, 采用了 *k*-近邻算法回答上述问题。经参数选择后, 三种方法在各自最优运行参数下, 其 10 倍交叉验证的结果为: 对脂肪酶和非脂肪酶预测精度分别为 92.8%、91.4%和 91.3%; 对脂肪酶类型预测的精度分别为 92.3%、90.3%和 89.7%。其中基于 Z 标度伪氨基酸组成效果最佳, 基于 T 标度的次之, 但均明显优于其他 6 种常见的特征值提取方法, 并对其可能的原因进行了探讨。

关键词: 脂肪酶, Z-标度, T-标度, 伪氨基酸组成, *k*-近邻

Prediction of Lipases Types by Different Scale Pseudo-amino Acid Composition

Guangya Zhang, Hongchun Li, Jiaqiang Gao, and Baishan Fang

Department of Biotechnolog and Bioengineering, Huaqiao University, Xiamen 361021, *China*

Abstract: Lipases are widely used enzymes in biotechnology. Although they catalyze the same reaction, their sequences vary. Therefore, it is highly desired to develop a fast and reliable method to identify the types of lipases according to their sequences, or even just to confirm whether they are lipases or not. By proposing two scales based pseudo amino acid composition approaches to extract the features of the sequences, a powerful predictor based on *k*-nearest neighbor was introduced to address the problems. The overall success rates thus obtained by the 10-fold cross-validation test were shown as below: for predicting lipases and nonlipase, the success rates were 92.8%, 91.4% and 91.3%, respectively. For lipase types, the success rates were 92.3%, 90.3% and 89.7%, respectively. Among them, the Z scales based pseudo amino acid composition was the best, T scales was the second. They outperformed significantly than 6 other frequently used sequence feature extraction methods. The high success rates yielded for such a stringent dataset indicate predicting the types of lipases is feasible and the different scales pseudo amino acid composition might be a useful tool for extracting the features of protein sequences, or at lease can play a complementary role to many of the other existing approaches.

Keywords: lipase, Z-scales, T-scales, pseudo-amino acid composition, *k*-nearest neighbor

脂肪酶(EC 3.1.1.3)能在油水界面或水不溶性系统中催化多种化学反应, 在医药、食品、洗涤剂化学合成及油脂等工业具有广泛用途[1]。不仅如此, 脂肪酶活性高低以及其基因多态性与某些疾病密切相

Received: March 11, 2008; **Accepted**: June 23, 2008
Supported by: the Research Fund for the Doctoral Program of Higher Education (No. 20070385001) and the Nature Science Foundation of Fujian Province (No. 2007J0360).
Corresponding author: Baishan Fang. Tel: +86-595-22691095; E-mail: fangbs@hqu.edu.cn
高等学校博士学科点专项科研基金项目(No. 20070385001), 福建省自然科学基金项目(No. 2007J0360)资助。

关, 是胰腺疾病和冠心病诊断的一种重要依据[2]; 故而, 脂肪酶常成为一些药物设计的潜在靶蛋白。尽管脂肪酶催化的化学反应类似, 但在序列上的差异却非常明显[3]。

后基因组时代的来临使蛋白质数据库序列数量急剧增长, 为此, 需要回答以下 2 个非常基础、同时又具有重大实践意义的问题: (1) 能否从新发现的蛋白质序列中快速鉴别它是否为脂肪酶? (2) 如果是, 能否仅从其序列出发预知其类型? 尽管通过传统的生物化学实验可回答上述问题, 但此法费时耗力, 难以满足需要; 另一方面, 一些生物信息学手段, 如序列比对、聚类等也被采用[4,5], 但如果在数据库中不存在新发现蛋白质的同源序列, 或者其同源序列具有其他不同功能情的况下, 上述方法的有效性就大为降低[6]。例如: 一些完成基因组测序的病毒基因所编码的蛋白中 20%~100%无法通过 PSI-BLAST 在 Swiss-Prot 数据库中找到同源序列[7,8]。由此可见, 需要发展一种快速、可靠的方法, 能够仅从蛋白序列出发来回答上述问题, 尤其在目前蛋白序列急剧膨胀的情况下, 显得尤为迫切。

从蛋白质序列中提取特征值是首先面临的问题。目前大多采用 20 种氨基酸组成表示蛋白质序列的特征值, 这主要是由于蛋白质序列长度差异很大, 而且即使长度相同, 其 20 种氨基酸排列方式也变化多样。但是, 该方法无法将序列中氨基酸顺序的影响考虑进去, 因此, 丧失了序列中非常重要的特征信息。为了解决这一问题, 研究者提出了伪氨基酸组成的方法[9], 这不仅能明确表征每个蛋白质序列而且同时考虑了序列中氨基酸排列顺序的影响, 其中, 每种氨基酸采用其亲水性、疏水性和侧链分子量进行描述, 众所周知, 这 3 个因素在蛋白质折叠、蛋白质与环境或其他分子的互作等过程中发挥重要作用, 对蛋白质的结构和功能产生重要影响。因此, 在预测诸如蛋白质亚细胞定位[10]、膜蛋白类型[9]、蛋白质四级结构类型[11]、酶的类型以及亚类型[12]等方面取得了很好效果。

氨基酸有众多理化性质, 对其进行主成分分析(PCA)并用于多肽定量构效关系的研究[13]是目前常用方法。如: 在 Z 标度中[14], 对氨基酸的 29 个理化性质进行 PCA 分析后, 得到了 3 个主成分 Z_1, Z_2 和 Z_3; 而在 T 标度中[15], 对氨基酸的 67 种结构和拓扑参数进行 PCA 分析, 得到了 5 个主成分, 包含了原有信息的 91.14%, 二者在研究多肽定量构效关系中效果良好。

本研究发展了几种基于不同标度的伪氨基酸组成方法来提取蛋白质序列特征值, 采用 *k*-近邻方法预测了脂肪酶和非脂肪酶以及脂肪酶的类型, 取得了令人满意的结果, 报道如下。

1 材料和方法

1.1 数据来源

数据样本按照以下步骤构建: (1) 脂肪酶的类型按照 LED 数据库(Lipase Engineering Database, release 2.3), 网址为: http://www.led.uni-stuttgart.de[3], 酶蛋白序列也来源于该数据库, 它们来源于 UniProt/Swiss-Prot; (2) 序列长度小于 100 个氨基酸的酶蛋白序列被剔除, 因为它们可能是部分长度或者片段; (3) 剔除了包含了连续 3 个或以上未知氨基酸的序列(例如: ”XXX”, “XXXX”等); (4) 为了避免由于同源序列带来的偏差, 使用 Blastclust 程序[16]剔除了序列相同性(Identity)大于 25%的序列[17,18]。最后一共得到了 2740 条脂肪酶序列, 其中 GGGx 类型 777 条, Gx 类型 1651 条, Y 类型 312 条。与此同时, 按照上述步骤, 从 UniProt/Swiss-Prot 中随机挑选了 2329 条非脂肪酶蛋白序列。上述 5069 条序列 ID 号, FASTA 格式的序列以及蛋白质长度等信息保存在一个基于 Microsoft Access 数据库中, 可向作者免费索取用于学术目的。

1.2 不同标度的伪氨基酸组成

本研究采用不同标度对伪氨基酸组成(PseAA)进行部分修正。对于一个长度为 L 的蛋白质 P, 其序列为 $R_1R_2R_3R_4R_5R_6R_7\cdots R_L$, 其序列顺序的影响可通过一系列的序列顺序相关因子(order-correlation factor)进行表述:

$$\begin{cases}\tau_1=\dfrac{1}{L-1}\displaystyle\sum_{i=1}^{L-1}J_{i,i+1}\\ \tau_2=\dfrac{1}{L-2}\displaystyle\sum_{i=1}^{L-2}J_{i,i+2}\\ \tau_3=\dfrac{1}{L-3}\displaystyle\sum_{i=1}^{L-3}J_{i,i+3}\\ \tau_\lambda=\dfrac{1}{L-\lambda}\displaystyle\sum_{i=1}^{L-\lambda}J_{i,i+\lambda}\end{cases}\quad(\lambda<L)\qquad(1)$$

式中，τ_1 表示第一等级(First-tier)相关因子，表示蛋白质序列中相邻氨基酸相关性，τ_2 表示第二等级相关因子，表示蛋白质序列中每间隔 1 个氨基酸之间的相关性，其他依此类推；偶联因子(Coupling factor)$J_{i,i}$ 定义为:

$$J_{i,i}=\frac{1}{k}\begin{Bmatrix}[z_1(R_i)-Z_1(R_j)]^2+[z_2(R_i)-z_2(R_j)]^2+\cdots\\+[z_k(R_i)-z_k(R_j)]^2\end{Bmatrix} \quad (2)$$

式中，$Z_1(R_i)$，$Z_2(R_i)$和 $Z_k(R_i)$分别表示氨基酸 R_i 的第 1 个标度值，第 2 个标度值和第 k 个标度值。$z_1(R_j)$, $z_2(R_j)$和$z_k(R_j)$分别表示氨基酸R_j的第1个标度值，第 2 个标度值和第 k 个标度值，不同类型的标度 k 取值不同。将氨基酸第 1 个，第 2 个和第 k 个标度值带入公式 2 之前，将它们按照公式 3 进行标准化转换:

$$\begin{cases}z_1(R_i)=\dfrac{z_1^0(R_i)-\left\langle z_1^0\right\rangle}{SD(z_1^0)}\\ z_2(R_i)=\dfrac{z_2^0(R_i)-\left\langle z_2^0\right\rangle}{SD(z_2^0)}\\ z_k(R_i)=\dfrac{z_k^0(R_i)-\left\langle z_k^0\right\rangle}{SD(z_k^0)}\end{cases} \quad (3)$$

式中，$z_1^0(R_i)$ 表示第 i 个氨基酸第 1 个原始标度值，$z_2^0(R_i)$ 表示第 i 个氨基酸第 2 个原始标度值，$z_k^0(R_i)$ 表示第 i 个氨基酸第 k 个原始标度值，$\left\langle z_1^0\right\rangle$ 表示 20 种氨基酸第 1 个原始标度值的平均值，其他依此类推。然后，将公式(1)计算的结果与 20 种氨基酸组成进行离散化处理。这样，1 个蛋白质 P 序列就可通过 20+λ 维的数组表示(公式 4), (4)式中

$$P=\begin{bmatrix}p_1\\ \vdots\\ p_{20}\\ p_{20+1}\\ \vdots\\ p_{20+\lambda}\end{bmatrix} \quad (4)$$

$$p_u=\begin{cases}\dfrac{f_u}{\sum_{i=1}^{20}f_i+w\sum_{j=1}^{\lambda}\tau_j}, & (1\leqslant u\leqslant 20)\\ \dfrac{w\tau_u}{\sum_{i=1}^{20}f_i+w\sum_{j=1}^{\lambda}\tau_j}, & (20+1\leqslant u\leqslant 20+\lambda)\end{cases} \quad (5)$$

公式(5)中，f_i表示蛋白质X经过归一化处理后的20种氨基酸组成，τ_j表示第j等级序列相关因子(由公式 1 计算)，w 为权重因子。在计算过程中，λ和 w 需要进行选择，以达到最佳识别效果。

表 1　20 种氨基酸的两种标度值

Table 1　Two scales of the 20 amino acids

Amino acids	Z-scales			T-scales				
A	0.07	–1.73	0.09	–9.11	–1.63	0.63	1.04	2.26
C	0.71	–0.97	4.13	–7.35	–0.86	–0.33	0.80	0.98
D	3.64	1.13	2.36	–4.65	0.75	1.39	–0.40	1.05
E	3.08	0.39	–0.07	–3.03	1.82	0.51	–0.58	0.43
F	–4.92	1.30	0.45	0.49	–0.94	–0.63	–1.27	–0.44
G	2.23	–5.36	0.30	–10.61	–1.21	–0.12	0.75	3.25
H	2.41	1.74	1.11	–1.01	–1.31	0.01	–1.81	–0.21
I	–4.44	–1.68	–1.03	–4.25	–0.28	–0.15	1.40	–0.21
K	2.84	1.41	–3.14	–2.59	2.34	–1.69	0.41	–0.21
L	–4.19	–1.03	–0.98	–4.38	0.28	–0.49	1.45	0.02
M	–2.49	–0.27	–0.41	–4.08	0.98	–2.34	1.64	–0.79
N	3.22	1.45	0.84	–4.62	0.66	1.16	–0.22	0.93
P	–1.22	0.88	2.23	–5.11	–3.54	–0.53	–0.36	–0.29
Q	2.18	0.53	–1.14	–3.00	1.72	0.28	–0.39	0.33
R	2.88	2.52	–3.44	0.23	3.89	–1.16	–0.39	–0.06
S	1.96	–1.63	0.57	–7.44	–0.65	0.68	–0.17	1.58
T	0.92	–2.09	–1.40	–5.97	–0.62	1.11	0.31	0.95
V	–2.69	–2.53	–1.29	–5.87	–0.94	0.28	1.10	0.48
W	–4.75	3.65	0.85	5.73	–2.67	–0.07	–1.96	–0.54
Y	–1.39	2.32	0.01	2.08	–0.47	0.07	–1.67	–0.35

本文采用了 2 种标度值: Z 标度和 T-标度。当采用 Z 标度时, 公式(2)中 k=3; 当采用 T-标度时, 公式(2)中 k=5。20 种氨基酸的两种标度值见表 1。

1.3 有效性检验

模型的稳定性及泛化能力, 采用了较为客观和严格的交叉验证(Cross-validation)方法, 采用了 10 倍交叉验证(10-fold cross-validation, 10-CV)。

文中构建的分类器基于 k-近邻算法, 实现该算法的软件为 Weka, 它是基于 JAVA 虚拟机开发的[19]。

k-近邻的基本思路是: 在给定新样本后, 考虑在训练样本集中与该新样本距离最近(最相似)的 K 个样本, 根据这 K 个样本所属的类别判定新样本所属的类别, 其详细内容可参见文献[20], 文中 k=1, 采用了 Euclidean Distance(欧氏距离)来衡量。使用 PC 为 DELL precision^{TM}490 工作站。

2 结果与分析

2.1 不同标度伪氨基酸组成运行参数的选择

如上所述, λ和 w 是伪氨基酸组成的 2 个重要参数, 本研究考察了这 2 个参数对识别精度的影响。从图 1A 中可以看出, 在 10 倍交叉验证过程中, w 一定的条件下, 随着λ的增大, 识别精度也随之而增大, 当达到最高识别精度后略微下降; 对基于 Z 标度的伪氨基酸组成而言, 当λ =80 时, 其识别精度最高; 同样的情况也存在于基于 T 标度的伪氨基酸组成中; 而对伪氨基酸组成而言, 当λ =95 时, 其识别精度最高; 在选择好合适的λ以后, 考察了 w 对识别精度的影响, 如图 1B 所示, 由图可知, 当λ一定时, 随着 w 的增大, 识别精度表现出不同的变化趋势, 但相比λ而言, 它对识别精度整体影响不大。这与报道结果相吻合[21]。综上可知, 当 w=0.5, λ=95 时, 伪氨基酸组成识别效果最好, 当 w=0.45, λ=80 时, 基于 T 标度的伪氨基酸组成识别效果最好, 而当 w=0.35, λ=80 时, 基于 Z 标度的伪氨基酸组成识别效果最佳, 对脂肪酶类型预测精度分别为 89.7%、90.3%和 92.3%, 此时输入分类器的数组维数分别为 115、100 和 100。

由于在各自最优条件下输入分类器的数组维数各异, 因此, 会产生这样的疑问: 究竟这种识别精度的提高是由序列特征提取方法本身带来的还是由数组维数的减少带来的? 为了回答这一问题, 比较了在相同输入数组维数(分别为 100 维和 115 维)的情况下 3 种方法的识别效果, 结果如图 2 所示, 从图中可知, 3 种方法中, 基于 Z 标度的伪氨基酸组成效果组好, 基于 T 标度的伪氨基酸组成次之, 伪氨基酸组成再次之, 但后两者之间差异不明显。同样的情况也可从图 1A 中看出, 可见, 这种预测精度的提高不是由于输入数组维数减少导致的, 而是序列特征值提取方法本身所带来的。

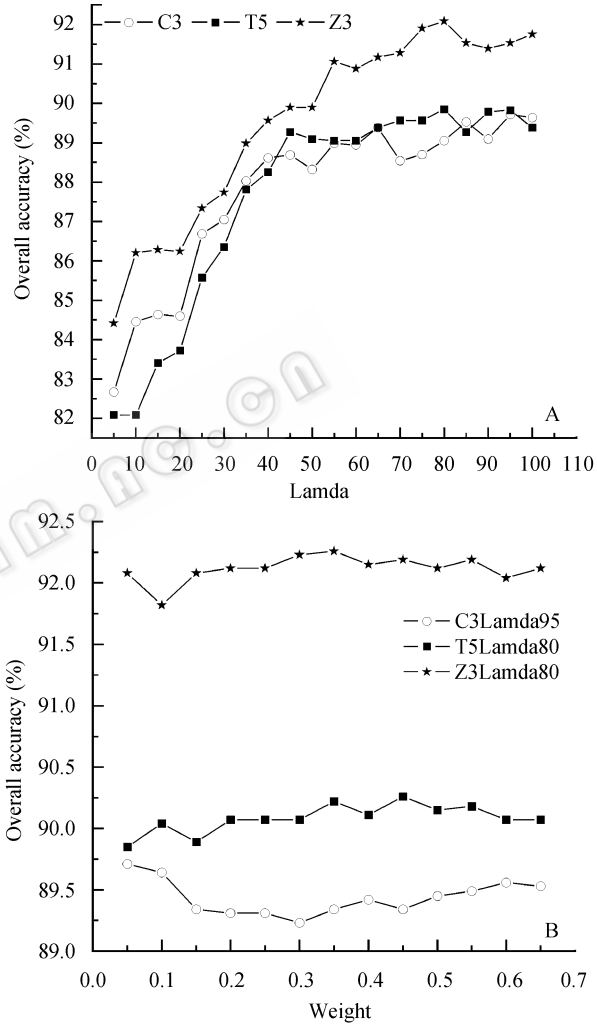

图 1 运行参数的选择

Fig. 1 Selection for the running parameters

2.2 与其他序列特征值提取方法的比较

为了进一步考察 2 种不同标度的伪氨基酸组成有效性, 利用同样的数据以及算法, 在相同条件下, 采用了其他几种序列特征值提取方法, 这些特征值提取方法包括: 氨基酸组成、标准化的 Moreau-Broto 自相关指数(Normalized Moreau-Broto autocorrelation descriptors)[22], Moran 自相关指数(Moran autocorrelation

descriptoers)[23], Geary 自相关指数(Geary autocorrelation descriptoers)[24], 组成、转变及分布(CTD)[25]以及序列顺序耦合数和部分序列顺序(SQ)[26], 上述计算由PROFEAT(Protein Feature Server)服务器计算完成[27]。网址: http://jing.cz3.nus.edu.sg/cgi-bin/prof/prof. cgi。经10 倍交叉验证, 其结果见表 2 和 3。

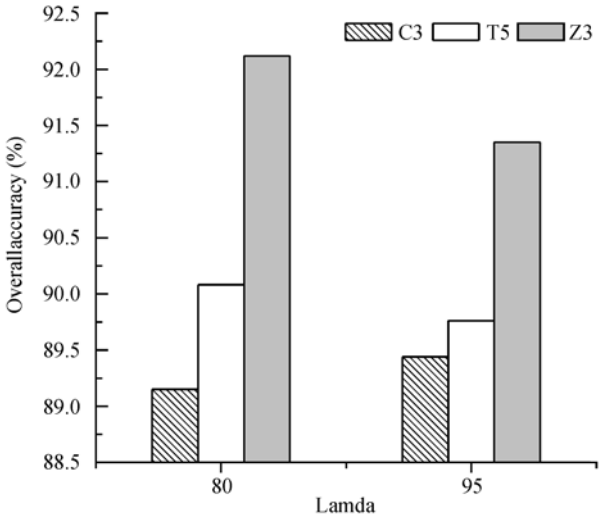

图 2 相同参数下 3 种方法的比较

Fig. 2 Comparison of three methods based on the same parameters

由表 2 可知, 在对脂肪酶和非脂肪酶识别过程中, 9 种特征值提取方法, 其中基于 Z 标度的伪氨基酸组成识别精度最高, 基于 T 标度的伪氨基酸组成次之, 伪氨基酸组成再次之, 但三者的预测精度均超过 90%, 而基于其他特征值提取方法的识别精度均在 90%以下。对脂肪酶而言, 基于 Z 标度的伪氨基酸组成比 CTD 预测精度高出了 13.7%; 对非脂肪酶而言, 基于Z标度的伪氨基酸组成比SQ预测精度提高了 12.6%; 整体预测精度比 CTD 提高了 11.9%。其 ROC 曲线下面积为 0.928, 说明分类器表现优秀。

由表 3 可知, 在对脂肪酶不同类型预测过程中, 9 种特征值提取方法, 其中基于 Z 标度的伪氨基酸组成识别精度依然最高, 基于 T 标度的伪氨基酸组成次之, 伪氨基酸组成再次之, 前两者预测精度超过 90%, 后者为 89.7%, 而基于其他特征值提取方法的识别精度较低, 均在 85%以下。对 GGGX 类型脂肪酶酶而言, 基于 Z 标度的伪氨基酸组成比 CTD 预测精度高出了 26.9%; 对 GX 类型脂肪酶酶而言, 基于 Z 标度的伪氨基酸组成比氨基酸组成预测精度高出了 13.7%; 对 Y 类型脂肪酶而言, 基于 Z 标度的伪氨基酸组成比 CTD 预测精度提高了 47.4%; 其整体预测精度比 CTD 提高了 20.1%。其 ROC 曲线下面积达到 0.915, 说明分类器表现优秀。

表 2 不同方法预测脂肪酶和非脂肪酶的精度

Table 2 Success rates of different methods for predicting lipase and nonlipase

Feature extraction methods	Lipase	Nonlipase	Overall
Pseudo-amino acid composition (Z-scales)	92.3	93.4	92.8
Pseudo-amino acid composition(T-scales)	90.8	92.1	91.4
Pseudo-amino acid composition (Chou)	90.8	91.9	91.3
Amino acid composition	88.5	89.9	89.1
Geary autocorrelation descriptors	84.0	85.2	84.5
Moran autocorrelation descriptors	84.3	84.3	84.3
Normalized Moreau-Broto autocorrelation descriptors	82.2	85.8	83.8
Sequence-order-coupling number & quasi-sequence-order	83.3	80.8	82.1
Composition, transition and distribution	78.6	83.6	80.9

表 3 不同方法预测脂肪酶类型精度

Table 3 Success rates of different methods for predicting lipase types

Feature extraction methods	Lipase types			Overall
	GGGX	GX	Y	
Pseudo-amino acid composition (Z-scales)	88.7	93.6	93.9	92.3
Pseudo-amino acid composition(T-scales)	87.3	91.2	92.6	90.3
Pseudo-amino acid composition (Chou)	86.0	90.9	92.9	89.7
Moran autocorrelation descriptors	74.8	88.9	53.2	80.8
Geary autocorrelation descriptors	75.3	88.6	53.2	80.8
Amino acid composition	79.4	79.9	86.5	80.5
Normalized Moreau-Broto autocorrelation descriptors	69.4	85.2	63.5	78.2
Sequence-order-coupling number & quasi-sequence-order	66.5	85.0	47.8	75.5
Composition, transition and distribution	61.8	82.0	46.5	72.2

3 讨论

如前所述, 氨基酸的亲水性、疏水性和侧链分子大小在蛋白质折叠、蛋白质与环境或其他分子的互作等过程中发挥重要作用。例如: 蛋白质分子中的螺旋结构多为两性, 即亲水性和疏水性氨基酸按照某种方式沿螺旋链分布[28]。然而, 氨基酸具有很多结构以及理化特性, 正是这些因素的共同作用决定了蛋白质的结构和功能。由此看来, 伪氨基酸组成考虑了亲水性、疏水性和侧链分子量, 虽然很好但却不够。然而, 实际应用中也不能把氨基酸所有的理化特性都考虑进去, 因为这样会极大提高运算的复杂性和难度。为了解决这个问题, 选择的标准就是: 能否找到包含更多信息量的定量描述常数而几乎不增加运算的难度? 对上述理化特性进行主成分分析(PCA)是一个非常不错的选择。众所周知, PCA 能显著降低数据的维数而不至于丢失较多的信息[29]。在 Z 标度中, 3 个主成分 z_1, z_2 和 z_3, 第 1 个主成分与亲水性显著相关, 第 2 个主成分和分子量, lH NMR 以及亲水性/疏水性相关, 第 3 个主成分则与 pKa, pI 和 1H NMR 相关。显然, z_1, z_2 和 z_3 包含了更多信息, 因此, 其识别效果的提高也在情理之中了。而 T 标度虽然在描述短肽定量构效关系中效果良好, 但并未考虑亲水性/疏水性等在折叠过程中非常重要的因素。故其结果仅略好于伪氨基酸组成。基于上述分析, 可以预见, 本研究给出的基于不同标度值的伪氨基酸组成可望成为一种新的特征值提取方法, 用于预测蛋白质其他特性的研究。

最近, 有研究者[17,18]利用功能域组成(Functional domain composition)和伪氨基酸组成相结合的方法预测了蛋白酶和非蛋白酶以及蛋白酶的类型, 其交叉验证的精度分别为 91.82%和 94.75%。他们认为要区分蛋白酶以及它们的类型, 最重要的任务就是抓住与其功能密切相关的核心特征, 而这可通过功能域的组成来实现[30], 但目前由于定义功能域的数据库不够完善。因此, 采用伪氨基酸组成进一步提取序列特征值, 使用该方法不仅可以明确地定义每个蛋白序列, 而且其序列顺序的影响也能得到很大程度的体现, 故而, 获得很高的预测精度。众所周知, 蛋白质序列决定其结构及功能, 同样, 蛋白质功能也会限制其序列变化的范围, 尤其是与功能密切相关的序列区域[31]。因此, 具有类似功能的蛋白在序列上必然具有相似的特征, 对其进行识别和分类的关键就是如何最大限度的表征这些特性。本研究计算了脂肪酶中从相邻氨基酸之间的互作到相距 80 个氨基酸之间的互作(λ=80), 即: 充分考虑了其氨基酸排列的顺序; 同时采用包含了更多信息量的常数来描述每个氨基酸, 这在很大程度上表征了脂肪酶在序列上的特性。因此, 也取得了较高的分类精度。

4 小结

简而言之, 本研究给出了基于不同标度值的伪氨基酸组成方法用于提取蛋白质序列信息, 并成功发展了一种基于 *k*-近邻的方法预测脂肪酶的类型, 其令人满意的预测精度说明了这种方法的有效性, 这有利于填补“已知序列的蛋白质”和“已知功能的蛋白质”之间的巨大缺口。相关软件已经测试完成, 可向作者免费索取, 用于学术目的。

REFERENCES

[1] Shu ZY, Wang AL, Yang JK, *et al*. Methods for improving production, activity and stability of microbial lipases. *Ind Microbiol*, 2007, **37**(3): 52−57.
舒正玉, 王爱玲, 杨江科, 等. 提高微生物脂肪酶产量、活性和稳定性的方法研究. 工业微生物, 2007, **37**(3): 52−57.

[2] Hu M, Shao JG, Zhu Y, *et al*.The-514C/T polymorphism of hepatic lipase and the relation to coronary heart disease. *Chin Pharmacol Bull*, 2006. **22**(9): 1118−112l.
胡敏, 邵建国, 朱毅, 等. 肝脂肪酶基因-514C/T 多态性与冠心病相关性研究. 中国药理学通报, 2006, **22**(9): 1118−112l.

[3] Fischer M, Pleiss J. The lipase engineering database: a navigation and analysis tool for protein families. *Nucl Acids Res*, 2003, **31**: 319−321.

[4] Bork P, Koonin EV. Predicting functions from protein sequences—where are the bottlenecks? *Nature Genet*, 1998, **18**: 313−318.

[5] Enright AJ, Quzounis CA. GeneRAGE: a robust algorithm for sequence clustering and domain detection. *Bioinformatics*, 2000, **16**: 451−457.

[6] Whisstock JC, Lesk AM. Prediction of protein functions from protein sequence and structure. *Q Rev Biophys*, 2003, **36**: 307−340.

[7] Rustici G, Milne RG, Accotto GP. Nucleotide sequence, genome organization and phylogenetic analysis of Indian citrus ring spot virus. *Arch Virol*, 2002, **147**: 2215−2224.

[8] Sabanadzovic S, Ghanem-Sabanadzovic NA, Saldarelli P, *et al*. Complete nucleotide sequence and genome organization of *Grapevine fleck* virus. *J Gen Virol*, 2001, **82**: 2009–2015.

[9] Chou KC. Prediction of protein cellular attributes using pseudo amino acid composition. *Proteins*: *Struct Funct Genet*, 2001, **43**: 246–255.

[10] Chou KC, Cai YD. Prediction and classification of protein subcellular location: sequence-order effect and pseudo amino acid composition. *J Cell Biochem*, 2003, **90**: 1250–1260.

[11] Chou KC, Cai YD. Predicting protein quaternary structure by pseudo amino acid composition. *Proteins: Struct Funct Genet*, 2003, **53**: 282–289.

[12] Chou KC, Cai YD. Predicting enzyme family class in a hybridization space. *Protein Sci*, 2004, **13**: 2857–2863.

[13] Andersson PM, Sjostrom M, Lundstedt T. Preprocessing peptide sequences for multivariate sequence-property analysis. *Chemometr Intell Lab Sys*, 1998, **42**: 41–50.

[14] Hellberg S, Sjostrom M, Skagerberg B, *et al*. Peptide quantitative structure-activity relationships, a multivariate approach. *J Med Chem*, 1987, **30**: 1126–1135.

[15] Tian FF, Zhou P, Li ZL. T-scale as a novel vector of topological descriptors for amino acid and its application in QSARs of peptides. *J Mol Struct*, 2007, **830**: 106–115.

[16] Altschul SF, Madden TL, Schaffer AA, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl Acids Res*, 1997, **25**: 3389–3402.

[17] Chou KC, Cai YD. Prediction of protease types in a hybridization space. *Biochem Biophys Res Commun*, 2006, **339**: 1015–1020.

[18] Zhou GP, Cai YD. Prediciting protease types by hybridizing gene ontology and pseudo amino acid composition. *Proteins*: *Struct Funct Genet*, 2006, **63**: 681–684.

[19] Inamdar NM, Ehrlich KC, Ehrlich M, *et al*. Data mining in bioinformatics using *Weka*. *Bioinformatics*, 2004, **20**: 2479–2481.

[20] Chou KC, Shen HB. Review: Recent progresses in protein subcellular location prediction. *Anal Biochem*, 2007, **370**: 1–16.

[21] Chou KC. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics*, 2005, **21**: 10–19.

[22] Feng ZP, Zhang CT. Prediction of membrane protein types based on the hydrophobic index of amino acids. *J Protein Chem*, 2000, **19**: 269–275.

[23] Horne DS. Prediction of protein helix content from an autocorrelation analysis of sequence hydrophobicities. *Biopolymers,* 1988, **27**: 451–477.

[24] Sokal RR, Thomson BA. Population structure inferred by local spatial autocorrelation: an example from an Amerindian tribal population. *Am J Phys Anthropol*, 2006, **129**: 121–131.

[25] Cui J, Han LY, Lin HH, *et al*. Prediction of MHC-binding peptides of flexible lengths from sequence-derived structural and physicochemical properties. *Mol Immunol*, 2007, **44**: 866–877.

[26] Chou KC. Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. *Biochem Biophy Res Comm*, 2000, **278**: 477–483.

[27] Li ZR, Lin HH, Han LY, *et al*. Chen, PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucl Acids Res*, 2006, **34**: 32–37.

[28] Chou KC, Zhang CT, Maggiora MG. Disposition of amphiphilic helices in heteropolar environments. *Proteins*, 1997, **28**: 99–108.

[29] Duttaa D, Mohanty AK, Choudhury RK, *et al*. Pattern recognition of particle tracks using principal component analysis and artificial neural network. *Nucl Instrum Methods Phys Res A*, 1998, **404**: 445–454.

[30] Chou KC, Cai YD. Using functional domain composition and support vector machines for prediction of protein subcellular location. *J Biol Chem*, 2002, **277**: 45765–45769.

[31] Minshull J, Ness JE, Gustafsson C, *et al*. Predicting enzyme function from protein sequence. *Curr Opio Chem Biol*, 2005, **9**: 202–209.