文章编号: 1000-5013(2010)03-0317-05

采用伪氨基酸组成预测水解酶亚家族

李红春, 张光亚, 方柏山

(华侨大学 工业生物技术研究所, 福建 泉州 362021)

摘要: 利用伪氨基酸组成提取蛋白序列特征值, 考察参数 λ 和 w 对识别效果的影响, 以 k-近邻作为基础分类器, 用于预测水解酶的亚家族类型. 结果表明, 伪氨基酸组成特征提取法与单纯的 20 个氨基酸组成特征方法相比, 其识别精度有较大程度提高. 20AA 组成的平均预测精度为 72.3%, 而伪氨基酸组成特征提取的识别效果可达 82.7%. 在参数影响考察方面, 自相关性函数个数的选取对识别效果影响较大, 而权重因子 w 对识别效果影响则很小.

关键词: 水解酶亚家族; 特征值; 伪氨基酸; k-近邻

中图分类号: Q 556.03 文献标识码: A

蛋白质数据库中的数据量以爆炸性的速度增长, 使用这些在实验过程中得到的大量新序列数据前, 如何高效而准确地对它们进行分类, 是生物信息学的一项重要工作[1]. 分类是否准确不仅影响到序列在查询数据库时获取链接的改变, 同时也影响从数据库中查询所得序列信息的准确性, 以及对其生物学功能的判断与应用等. Swiss-Prot 数据库(ExPASy Web Service)中 6 大类酶总数有 11 9392 个, 其中第 3 类水解酶共有 33 165 个[2-3]. 水解酶在所有酶类中是数量最多的一类酶, 同时它的生物学作用也极其重要, 在临床医学[4]、造纸工业[5]、饲料食品[6]和环境领域均有极广泛用途. 因此, 研究高效的蛋白质序列特征提取方法, 从水解酶中提取其序列特征信息并对其亚家族进行准确分类有很大的生物学意义. 酶蛋白一级结构可以表述成由 20 种字母(表示 20 个氨基酸)组成的一段字符串, 因此需要从这段字符串中提取出适当的特征值, 从而提高预测精度. 目前, 主要从序列出发提取特征值的方法, 有氨基酸组成[7]、二肽组成[8]、伪氨基酸组成[9]、两性伪氨基酸组成[1]等. 本文采用伪氨基酸组成特征提取法, 研究两个参数自相关函数个数 λ 和归一化权重因子 w 对结果的影响.

1 材料与方法

1.1 数据来源

样本中所有水解酶来源于 ENZYME 数据库[10], 其序列来源于 Swiss-Prot[2], 数据库版本为 Release 54.4 (Sep 12 2007) of Swiss-Prot. 实验样本中所有序列根据其水解化学键的不同分如下 6 类. (1) 脂肪酶. 可用于水解酯键, 共计 8 541 条. (2) 糖苷酶. 可用于水解糖苷键, 共计 2 581 条. (3) 醚酶. 可用于水解醚键, 共计 157 条. (4) 肽酶. 可用于水解肽键, 共计 6 555 条. (5) 酰胺酶. 可用于水解酰胺键或脒键, 共计 3 898 条. (6) 酸酐酶. 可用于水解酸酐键, 共计 10 563 条.

首先, 剔除长度小于 100 个氨基酸残基的酶蛋白和片断酶蛋白, 并使用 BLASTCLUST 程序[11]剔除了其中相似性大于30%的序列. 最后, 得到6 类水解酶的样本量分别为5 065, 1 682, 90, 3 854, 2 336 和 6 236 条, 共计 19 263 条.

另外 EC3.7, EC3.8, EC3.10, EC3.11 和 EC3.13 的 5 类酶, 因为所得样本量太少, 缺乏统计意义而没有采用. 用统计软件 SPSS 统计所有序列的长度分布密度, 得到其长度分布正态曲线图, 如图 1 所示.

收稿日期: 2008-09-23

通信作者: 方柏山(1957-), 男, 教授, 主要从事工业生物技术与生物信息学的研究. E-mail: fangbs@hqu.edu.cn.

基金项目: 国家自然科学基金资助项目(30770059); 教育部博士点科研基金资助项目(20070685001)

酶蛋白序列平均长度约为 438 个氨基酸残基.

1.2 特征值提取方法

用以下两种方法提取水解酶序列特征值.

(1) 20 个氨基酸组成 $\boldsymbol{P}_{AA}$ 的计算式为

$$\boldsymbol{P}_{AA} = [f_1, f_2, \cdots, f_{20}]^T. \quad (1)$$

其中: f_i ($1 \leqslant i \leqslant 20$) 为某 AA 在序列中出现的频率.

(2) Chou 等[12] 提出的伪氨基酸组成($\boldsymbol{P}_{PseAA}$) 特征提取方法. 伪氨基酸组成分成两部分, 第 1 部分是 20 维向量 $\boldsymbol{P}_{AA}$, 计算同式(1); 另一部分是 λ 维向量 $\boldsymbol{P}_{corr}$, 由 20 个 AA 的理化参数(表 1) 替换长度为 L 的样本序列中相应字符, 得到的 3 组离散数值序列的 λ 个自相关性函数, 有

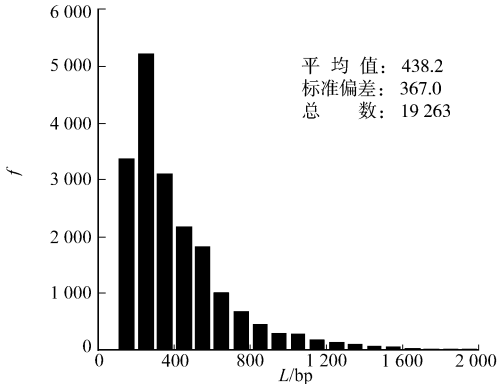


图 1 样本长度分布

Fig. 1 Length frequency of dataset

$$\boldsymbol{P}_{corr} = [\tau_1, \tau_2, \cdots, \tau_\lambda]^T, \quad (2)$$

$$\tau_j = \frac{1}{L-j} \cdot \sum_{i=1}^{L-j} \Theta(R_i, R_{i+j}), \quad (3)$$

$$\Theta(R_i, R_{i+j}) = \frac{1}{3}\{[\boldsymbol{H}_1(R_i) - \boldsymbol{H}_1(R_j)]^2 + [\boldsymbol{H}_2(R_i) - \boldsymbol{H}_2(R_j)]^2 + [\boldsymbol{M}(R_i) - \boldsymbol{M}(R_j)]^2\}. \quad (4)$$

式(4)中: $\boldsymbol{H}_1(i)$, $\boldsymbol{H}_2(i)$, $\boldsymbol{M}(i)$ 为表 1 中 3 组氨基酸参数标准化转换后所得数组向量. $\boldsymbol{P}_{corr}$ 与 $\boldsymbol{P}_{AA}$ 归一化后的正交和为 $\boldsymbol{P}_{PseAA}$. 即

$$\boldsymbol{P}_{PseAA} = \boldsymbol{P}_{AA} \oplus \boldsymbol{P}_{corr} = [f_1, f_2, \cdots, f_{20}, \tau_1, \tau_2, \cdots, \tau_\lambda]^T. \quad (5)$$

表 1 20 种氨基酸的参数

Tab. 1 Parameters of 20 amino acids

氨基酸	A	C	D	E	F	G	H	I	K	L
疏水性[13]	0.62	0.29	−0.90	−0.74	1.19	0.48	−0.40	1.38	−1.50	1.06
亲水性[14]	−0.50	−1.00	3.00	3.00	−2.50	0	−0.50	−1.80	3.00	−1.80
M_r/u	15	47	59	73	91	1	82	57	73	57
氨基酸	M	N	P	Q	R	S	T	V	W	Y
疏水性[13]	0.64	−0.78	0.12	−0.85	−2.53	−0.18	−0.05	1.08	0.81	0.26
亲水性[14]	−1.30	0.20	0	0.20	3.00	0.30	−0.40	−1.50	−3.40	−2.30
M_r/u	75	58	42	72	101	31	45	43	130	107

其中, 归一化公式为

$$f_i = \frac{f_i}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{\lambda}\tau_j}, \quad 1 \leqslant i \leqslant 20, \quad (6)$$

$$\tau_j = \frac{w\tau_j}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{\lambda}\tau_j}, \quad 1 \leqslant j \leqslant \lambda \quad (7)$$

由此得到了 20+ λ 维向量 $\boldsymbol{P}_{PseAA}$, 即 Chou Type 1 伪氨基酸组成.

整个算法过程如上, 具体计算可以提交到 Chou 的网站(http://chou.med.harvard.edu/bioinf/PseAA/) 进行计算. 该网站一次最多只能计算 50 条序列样本, 而采用 C^{++} 编程语言和 ACCESS 数据库结合, 把算法集成到软件 PseAA 中, 使单次计算过程和各参数评估的效率大幅提高.

1.3 有效性检验

1.3.1 *检验方法* 采用较为客观和严格的 10 倍交叉验证(10-CV). 即将训练数据随机分为 10 组, 然后采用"留一法"进行验证, 每次留出 1 组作为测试数据, 另 9 组作为训练数据, 循环 10 次.

1.3.2 *评估参数* 识别效果的优劣用准确率(r_{ACC}) 和受试者工作特征曲线(ROC) 下面积(S_{AUC}) 两个参数进行描述, 有

$$r_{ACC} = (TP + TN)/(TP + FP + TN + FN),$$

$$r_{SE} = TP/(TP + FN),$$

$$r_{SP} = TN/(TN + FP).$$

式中: TP 为真阳性; TN 为真阴性; FP 为假阳性; FN 为假阴性. 准确率(r_{ACC})是评价识别效果的常用指标; 受试者工作特征(ROC)曲线是评价识别效果敏感性(SE)和特异性(SP)连续变量的综合指标, 以假阳性率(r_{SP})为横坐标、真阳性率(r_{SE})为纵坐标绘制而成的曲线. S_{AUC}小于 0.5, 说明预测结果没有意义; S_{AUC}为 0.5~ 0.7, 说明识别效果较低; S_{AUC}为 0.7~ 0.9, 说明效果中等; 而 S_{AUC}大于 0.9, 则说明识别效果优秀.

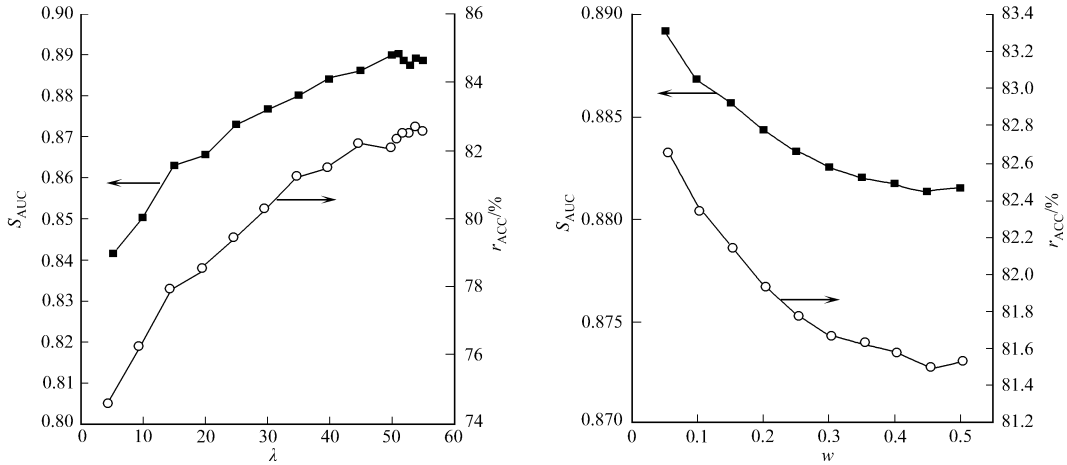
应用于模式识别算法的软件来自于 WEKA(Waikato Environment for Knowledge Analysis), 该程序包是基于 JAVA 虚拟机开发的[15]. 使用的算法为 k-近邻, 算法运行参数采用系统默认值. 使用的 PC 为 DELL Precision™ 490 工作站.

2　结果与分析

2.1　特征值提取方法参数的选择

伪氨基酸组成有两个运行参数 λ和 w 需要进行选择, 以达到最佳分类效果, 在此过程中使用的分类器为 k-近邻(k-NN)[16].

参数 λ与整体识别精度的关系, 如图 2(a)所示. 从图 2(a)可知, 随着 λ值的不断增大, 整体识别精度也不断提高, λ值增加到 50 附近时, 识别效果的提升趋于平缓, 最高识别精度时, λ取值为 54. 参数 w 与整体识别精度的关系, 如图 2(b)所示. 从图 2(b)可知, w= 0.05 时识别效果最佳, 随着 w 增大识别精度略为降低, 但降低幅度较小, w 的改变对 r_{ACC}影响最大不超过 1.2%.



(a) λ值　　(b) w 值

图 2　参数对识别精度的影响

Fig. 2　Influence of parameters on prediction accuracy

综上可知, 当 λ= 54, w= 0.05 时, 模型的识别效果最好, 此时, 其识别的正确率为 82.7%; ROC 曲线下面积(S_{AUC})为 0.889 2　这表明, 分类效果较好.

2.2　特征值提取方法的比较

当 λ= 54, w = 0.05 时, Chou Type 1 伪氨基酸组成识别精度最高. 实验样本中, 6 种水解辅亚家族类型识别正确率如表 2 所示.

表 2　两种预测精度的比较

Tab. 2　Comparison of the two prediction accuracy

亚家族	r_{ACC}		S_{AUC}	
	20AA	PseAA	20AA	PseAA
HY1	64.1	76.1	0.778	0.851
HY2	71.6	76.4	0.844	0.872
HT3	73.3	82.2	0.870	0.923
HT4	69.6	81.5	0.808	0.881
HT5	61.7	78.0	0.792	0.884
HT6	84.9	92.1	0.871	0.924
ALL	72.3	82.7	0.827	0.889

由表 2 可知, 各亚类的预测效果都比 20 个氨基酸百分组成高 10 个百分点左右, 整体识别精度(r_{ACC})为

82.7%,比20个氨基酸百分组成的识别精度72.3%提高了10.4%;而其S_{AUC}为0.889 2,识别效果接近优秀(大于0.9即可视为优秀),比20个氨基酸组成提高了0.072 6.

综上可见,实验中,伪氨基酸组成特征提取得到的预测精度比20个氨基酸百分组成的精度有明显提高,其变量比20个氨基酸组成多了λ个(20+λ).λ个相关函数表征出了各个氨基酸间相互关联等信息,在一定程度上体现了空间结构的信息.然而,蕴含较多信息量的同时,对计算资源的消耗也相对较大.k-近邻算法运行较快,识别效果也较好,在运算过程中的速度差异还不太明显.

3 讨论

从蛋白质序列出发对其生物学特性进行预测,是目前生物信息学研究的热点问题[17],也是探讨蛋白质结构和功能关系的一种重要研究手段.如何从序列信息中有效提取特征值是目前关注的焦点之一.如建立采用误差反传(BP)算法的多层感知机模型,对嗜热蛋白和常温蛋白进行模式识别[18].

与20个氨基酸组成相比,伪氨基酸组成引入了λ个自相关函数变量,该变量在一定程度上表征了各个氨基酸在空间结构上的相互作用关系.对不同的数据样本参数λ对结果的影响各异,不同样本的最佳λ值需要经实验计算得到.实验数据的最佳λ取值为54,同时权重因子w取值在0.05达到最佳.该参数下,预测精度达82.7%,比单纯20个氨基酸组成提高了10.4%,提升效果比较理想.

实验中引入了3组氨基酸的理化参数以表征各氨基酸的性质,分别是亲水性、疏水性和氨基酸残基侧链分子量(表1).在酶的生物催化过程中,亲水性和疏水性在各氨基酸间差别对催化进程的影响也是很重要的,蛋白质序列中某个亲水性残基变成疏水性残基就可能使其功能丧失.这3组参数在一定程度上反映了氨基酸的性质,但还不够充分.如果把20个氨基酸的其他理化性质也有效地引入到蛋白质特征提取算法中,以更大程度地提高识别效果,是一项值得再深入研究的工作.

另外,伪氨基酸组成的自相关函数计算,是基于各氨基酸理化性质引入到序列中,所形成的一组平稳随机离散数值(可视其为一组离散信号).有文献报道,在特征提取过程中,引入傅里叶变换[19]或小波变换处理随机离散数值,对信号进行去噪和频谱分析,可提高特征提取的效率[20].

综上所述,基于蛋白质序列一级结构的伪氨基酸组成特征提取方法,比传统的20个氨基酸组成的识别效果有了显著提高.同时,该算法在氨基酸理化性质参数选用时,其参数前处理及算法本身改良方面还有可再完善的空间.

参考文献:

[1] CHOU K C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes[J]. Bioinformatics, 2005, 21(1): 10-19.

[2] BAIROCH A, APWEILER R, WU C H, et al. The universal protein resource (uniprot)[J]. Nucleic Acids Res, 2005, 33: 154-159.

[3] BAIROCH A, APWEILER R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL[J]. Nucleic Acids Res, 2000, 25(1): 31-36.

[4] 段作纬,熊郁良,陈训如,等. 用胰蛋白酶和糜蛋白酶急救药盒救治毒蛇咬伤256例[J]. 中国危重病急救医学, 1998, 10(6): 336-338.

[5] VIIKARI L, PAUNA M, KANTELINEN A, et al. Bleaching with enzymes with enzymes[J]. In: Proceedings of the 3rd International Conference on Biotechnology in the Pulp and Paper Industry. Stockholm: Swedish Pulp and Paper Research Institute, 1986: 67-69.

[6] CLASSEN H L. Cereal grain starch and exogenous enzymes in poultry diets[J]. Animal Feed Science Technology, 1996, 62(1): 21-27.

[7] 丁彦蕊,蔡宇杰,须文波. 基于氨基酸组成预测蛋白质热稳定性的v-支持向量机方法[J]. 计算机与应用化学, 2005, 22(6): 51-57.

[8] GROMIHA M M, AHMAD S, SUWA M. Application of residue distribution along the sequence for discriminating outer membrane proteins[J]. Comput Biol Chem, 2005, 29(2): 135-142.

[9] CHOU K C. Prediction of protein cellular attributes using pseudo-amino acid composition[J]. Structure, Function,

and Genetics, 2001, 43(3): 246-255.
[10] BAIROCH A. The ENZYME database in 2000[J]. Nucleic Acids Res, 2000, 28(1): 304-305.
[11] ALTSCHUL S F, MADDEN T L, SCHAFFER A A, et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs[J]. Nucleic Acids Res, 1997, 25(17): 3389-3402.
[12] CHOU K C. Prediction of protein cellular attributes using pseudo-amino-acid-composition[J]. Proteins: Structure, Function, and Genetics, 2001, 43(3): 246-255.
[13] TANFORD C. Contribution of hydrophobic interactions to the stability of the globular conformation of proteins [J]. J Am Chem Soc, 1962, 84(22): 4240-4247.
[14] HOPP T P, WOODS K R. Prediction of protein antigenic determinants from amino acid sequences[J]. Proc Nat Acad Sci, 1981, 78(6): 3824-3828.
[15] INAMDAR N M, EHRLICH K C, EHRLICH M, et al. Data mining in bioinformatics using Weka[J]. Bioinformatics, 2004, 20(15): 2479-2481.
[16] COVER T M, HART P E. Nearest neighbor pattern classification[J]. IEEE Transactions on Information Theory, 1967, 13(1): 21-27.
[17] CUI J, HAN L Y, LIN H H, et al. Prediction of MHC-binding peptides of flexible lengths from sequence-derived structural and physicochemical properties[J]. Mol Immunol, 2007, 44(5): 866-877.
[18] 张光亚, 葛慧华, 方柏山. 采用 BP 算法的多层感知机模型的蛋白识别[J]. 华侨大学学报: 自然科学版, 2009, 30(2): 161-165.
[19] GUO Yan-zhi, LI Meng-long, ZOU Xiao-yong, et al. Fast fourier transform-based support vector machine for prediction of G-protein coupled receptor subfamilies[J]. Acta Biochimicaet Biophysica Sinica, 2005, 37(11): 759-766.
[20] QIU Jian-ding, LIANG Rui-ping, ZOU Xiao-yong, et al. Prediction of protein secondary structure based on continuous wavelet transform[J]. Talanta, 2003, 61(3): 285-293.

Using Pseudo Amino Acid Composition to Predict Hydrolase Subfamily

LI Hong-chun, ZHANG Guang-ya, FANG Bai-shan

(Institute of Industrial Biotechnology, Huaqiao University, Quanzhou 362021, China)

Abstract: Predicting the hydrolase subfamily is of great importance for designing a fast and reliable classification system. In this paper, the pseudo amino acid composition method was used to extract the features from protein sequencec, and the *k*-nearest neighbor algorithm was used as the classifier to predict the hydrolase subfamily. The influences of λ and ω on prediction accuracy were also studied. The results showed that the prediction accuracy of pseudo amino acid composition were much higher (about 10.4%) than that of amino acid composition, the prediction accuracy of amino acid was 72.3%, while the pseudo amino acid was 87.2%. The running parameter of λ had more influence on prediction accuracy when compared with ω.

Keywords: hydrolase subfamily; feature extraction; pseudo amino acid composition; *k*-nearest neighbor

(责任编辑: 黄仲一　　英文审校: 刘源岗)