文章编号:0253-2778(2008)07-0765-05

基于进化信息和支持向量机的酶蛋白亚家族预测

冯焕清,张相华,许文龙

(中国科学技术大学电子科学与技术系,安徽合肥 230026)

摘要:提出一种使用PSI-BLAST得到的位置特异性打分矩阵中蕴含的进化信息作为酶蛋白的特征表示,结合支持向量机方法对酶蛋白的亚家族类别进行预测的方法.对包含16类亚家族的2 640条氧化还原酶数据集进行jacknife测试,总的预测精度达到92.12%,高于目前的任何其他预测方法.实验结果表明,进化信息是酶蛋白序列的有效表示,将其与支持向量机结合能够实现对酶蛋白亚家族的高精度预测.

关键词:酶蛋白亚家族预测;进化信息;支持向量机;位置特异性打分矩阵

中图分类号:R318;Q617 **文献标识码**:A

Prediction of enzyme subfamily classes using evolutionary information and support vector machines

FENG Huan-qing, ZHANG Xiang-hua, XU Wen-long

(*Department of Electronic Science and Technology, USTC, Hefei* 230026, *China*)

Abstract: A novel method was proposed to predict enzyme subfamily classes. It combined support vector machines (SVMs) and evolutionary information of amino acid sequences in the form of position-specific scoring matrix (PSSM) by PSI-BLAST. With a jackknife test on a widely used dataset that containing 2 640 oxidoreductase sequences classified into 16 subfamily classes, the proposed method achieved a high overall accuracy of 92.12%, which is much better than that of any previous method. The results indicate that evolutionary information has a strong correlation with enzyme types and the proposed method is a potential powerful tool for enzyme subfamily classification.

Key words: enzyme subfamily classification; evolutionary information; support vector machines; position-specific scoring matrix

0 引言

酶的高效、专一、多样和反应条件温和的性质使细胞内复杂的物质代谢过程能有条不紊地进行,人体若因遗传缺失或其他原因造成酶的活性异常,就会引发疾病,因此研究酶的功能及相关理论对生命科学的发展至关重要.根据催化反应的类型,可以把酶分成六大家族[1]:(Ⅰ)氧化还原酶,(Ⅱ)水解酶,(Ⅲ)转移酶,(Ⅳ)裂解酶,(Ⅴ)异构酶,(Ⅵ)合成酶.每一家族根据所催化的化学键特点和参加反应的不同基团,可以进一步分为不同的亚家族,如氧化还原酶分为20个亚家族,水解酶分为9个亚家族[2].

收稿日期:2008-05-15;**修回日期**:2008-05-28

基金项目:教育部研究生创新基金(C07-05),中国科技大学高水平大学建设重点科研基金资助.

作者简介:冯焕清(通讯作者),教授. E-mail: hqfeng@ustc.edu.cn

由于酶的类别归属与其功能、催化机理及特异性密切相关,对于一个新的酶蛋白可以通过确定其类别归属来辅助阐明其功能,这对酶的基础研究和相关的药物发现很有帮助. 传统的通过实验手段测定酶功能的方法,不仅非常费时,而且费用比较高. 因此,从酶蛋白序列信息中自动预测其类别的机器学习方法在近些年得到了快速的发展. 比如,Chou 等于 2003 年使用氨基酸组成和一种协变量判决算法(covariant discriminant algorithm)对一个包含 2 640条脱氧还原酶的数据集进行了分析,总体预测精度为 63.64%[3]. 由于氨基酸组成不包含序列的顺序信息,因此一些对分类有用的信息,在计算组成的过程中丢失. 随后,他们改用双极性伪氨基酸组成(amphiphilic pseudo-amino acid composition)方法,保留一些蛋白序列的结构信息,进一步将预测精度提高到 70.61%[4];Huang 等在 2006 年使用基于双极性伪氨基酸组成的模糊 k-近邻(fuzzy k nearest neighbors)方法,获得了 76.63%的精度[5];Zhou 等在 2007 年利用相同的特征表示方法,采用支持向量机进行预测,得到了目前最好的预测精度 80.17%[6]. 然而,基于氨基酸组成、双极性伪氨基酸组成的酶蛋白亚家族的预测表征方法仅保留了酶蛋白序列较少的局部信息,一些其他的重要信息,比如序列进化信息,并没有被考虑进去. 由于酶蛋白的功能具有很强的保守性,因此引入进化信息可以提高预测效果. 而使用 PSI-BLAST 程序[7] 所得到的 profile 中的位置特异性打分矩阵(position-specific scoring matrix, PSSM)包含有酶蛋白的进化信息,可以考虑将其作为特征进行预测.

本文将 PSSM 作为酶蛋白进化信息的表示,并以此作为分类输入特征,利用具有很好理论基础和泛化能力的支持向量机方法对酶蛋白的亚家族类别进行预测,并与其他预测方法比较. 结果表明,结合进化信息和支持向量机能够取得更好的预测效果.

1 数据集

为了便于和其他方法比较,实验选择了一个常用的公共数据集. 该数据集是由 Chou 等在 2001 年构建的[3],其中包括2 640条氧化还原酶蛋白,可分为 16 个亚家族,如表 1 所示. 为方便起见,将该数据集称为 Chou 数据集. 实验中,酶蛋白的氨基酸序列数据取自 SWISS-PROT 数据库 40.0 版[8].

表 1 Chou 数据集的数据构成

Tab. 1 The Chou's Dataset

subfamily	groups acted by the enzyme	No. of samples
1	CH-OH group	314
2	Aldehyd/oxo group	216
3	CH-CH group	194
4	CH-NH_2 group	130
5	CH-NH group	112
6	NADH/NADPH	305
7	other nitrogenous compounds	64
8	sulfur group	59
9	heme group	254
10	diphenols and related substances	94
11	peroxide	154
12	single donors	94
13	paired donors	257
14	superoxide radicals	155
15	$-CH_2$ group	84
16	reduced ferredoxin	154
total		2 640

2 方法

2.1 PSSM 矩阵的获取

从 PSI-BLAST 的 profile 中提取 PSSM 矩阵作为输入特征的思想,最早由 Jones 提出[9],现被广泛应用于生物信息学的研究,如蛋白质二级结构预测[9]、蛋白亚细胞定位[10]和 DNA 结合蛋白预测[11]等方面. 为了得到 PSSM 矩阵,首先使用 PSI-BLAST 程序对非冗余蛋白数据集进行搜索. 该数据集包含了多个数据库信息:GenBank translations,PDB,SWISS-PROT,PIR 和 PRF 等,共约430 000左右的蛋白质序列. 利用 pfilt 程序筛选去除数据集中所有跨膜区域,卷曲——卷曲段和低复杂度区域. 然后使用 PSI-BLAST 程序对 Chou 数据集中的每一条酶蛋白序列和筛选后的数据集进行比对搜索,得到其对应的 PSSM 矩阵. 其中用于 PSI-BLAST 的参数设置为:循环 3 次,E-value 阈值 0.001. PSSM 矩阵形式为

$$\mathrm{PSSM}=\begin{bmatrix} S_{1\to 1} & S_{1\to 2} & \cdots & S_{1\to 20} \\ \cdots & \cdots & \vdots & \cdots \\ S_{i\to 1} & S_{i\to 2} & \cdots & S_{i\to 20} \\ \cdots & \cdots & \vdots & \cdots \\ S_{L\to 1} & S_{L\to 2} & \cdots & S_{L\to 20} \end{bmatrix} \tag{1}$$

式中,L 为酶蛋白序列的长度,$S_{i\to j}$ 表示酶蛋白序列中第 i 个氨基酸被氨基酸 j 取代的可能性的分值大小. 将 PSSM 矩阵每列中具有相同氨基酸行标的数

据相加，分别除以酶蛋白链的长度得到 20×20（代表 20 种氨基酸）共 400 维的特征矢量，最后使用标准 Sigmoid 函数将该特征矢量中的每个值映射到[0,1]区间作为支持向量机的输入.

2.2 SVM 原理

支持向量机（support vector machine，SVM）是 Vapnik 等根据统计学习理论提出的一种新的机器学习方法[12]. SVM 基于结构风险最小化原则，具有很强的泛化能力，即由有限的训练集样本得到的小误差仍能够对独立的测试集保证小的误差. 目前，SVM 已在生物信息学的许多研究领域，如亚细胞定位[10,13]、微阵列数据分析[14]中得到了有效应用.

运用 SVM 进行分类的基本思想可简述为：假设有 N 个训练样本$(\boldsymbol{x}_i, y_i)$，$(i=1,2,\cdots,N)$，$\boldsymbol{x}_i\in\mathbb{R}^d$，$y_i\in\{+1,-1\}$是类别标记. 首先 SVM 将输入空间的样本通过某种非线性函数关系映射到一个高维特征空间，然后在此特征空间中构造一个最优分类超平面（optimal separating hyperplane，OSH）. 映射函数仅与低维输入向量和特征空间的点积有关，此映射函数点积可用一核函数 $k(x,x_i)$来替代，从而避免“维数灾难”，可解决高维特征问题. 其判别函数为

$$f(x)=\operatorname{sgn}\Big(\sum_{i=1}^{N}a_iy_ik(x,x_i)+b\Big),\ 0\leqslant a_i\leqslant C \tag{2}$$

式中，系数 a_i 可通过解相应的凸二次规划问题得到. 对于一个已知的数据集，SVM 模型仅有核函数和惩罚因子 C 来确定. 核函数的选择对 SVM 的分类性能有很重要的影响，选择不同的核函数可以得到不同的支持向量机. 常用的核函数有多项式核函数、径向基核函数（RBF）及 Sigmoid 核函数.

多项式核函数为

$$k(x_i,x_j)=(x_ix_j+1)^d \tag{3}$$

径向基核函数为

$$k(x_i,x_j)=\exp(-g\parallel x_i-x_j\parallel^2) \tag{4}$$

Sigmoid 核函数为

$$k(x_i,x_j)=\tanh(gx_ix_j+\alpha) \tag{5}$$

由于数据集中含有 16 个酶的亚家族，需要构建 SVM 多类分类器. 使用 SVM 处理多类分类问题，常见的有一对多和一对一两种方法. 对于 n 类待分数据，第一种方法是设计 n 个一对多的两类 SVM 分类器，即将其中每一类数据和剩余的所有数据分为两类，在此基础上设计两类 SVM 分类器. 该方法的最终预测结果采取多数投票的策略，即选择 n 个分类器输出中的最大值所代表的类别，作为测试数据的分类决策. 第二种方法考虑对任意两类数据，设计一个两类 SVM 分类器，对于 n 类数据，则共需设计 $n(n-1)/2$ 个两类分类器. 最终预测结果则采用投票表决的策略，即对所有二类分类器的分类结果进行统计，选择获得最多投票的类别作为测试数据的最终分类决策；如果有超过一个类别获得了最高数目的投票，则随机选择一个类别作为其最终分类决策. 同前者相比，采用一对一多类分类器设计方法，在类别数 n 较大时需要设计更多的二类分类器. 然而使用此方法，由于设计每一个分类器时所需求解优化问题的规模较小，因此其实际训练和预测速度与一对多方法相近. 同时，由于设计每一个分类器时，其对应两类数据的样本数量相对接近，从而避免了前者使用中常见的由于数据不平均造成的有偏分类问题. 因此本文在实验中采用一对一方法进行多类分类器设计.

2.3 评价指标

为了检验方法的有效性，采用子类精度和整体精度来评价和比较不同方法的预测效果. 对于第 i 类酶蛋白亚家族，其对应的子类预测精度定义为

$$\mathrm{Acc}(i)=\frac{\mathrm{TP}_i}{\mathrm{TP}_i+\mathrm{FN}_i}\times100\% \tag{6}$$

式中，TP_i，FN_i 分别表示测试后第 i 类酶亚家族中真阳性、伪阴性样本数目. 而整体预测精度定义为

$$\text{overall accuracy}=\frac{\sum_i\mathrm{TP}_i}{N}\times100\% \tag{7}$$

式中，N 是数据集中酶蛋白所有样本的数目.

3 结果与讨论

3.1 SVM 参数选取

核函数及训练参数的选择对支持向量机的性能有着至关重要的影响. 因此选择合适的核函数类型和参数对提高分类预测精确度有重要作用. 实验选用了多项式核函数、径向基核函数、Sigmoid 核函数进行测试工作. 为获得满意的预测效果，对三种核函数选择不同惩罚系数 C 及相应的核参数，采用 10 次交叉验证方法对 Chou 数据集进行训练和测试，以确定各核函数的最佳测试结果及相应的最优参数组合，结果如表 2 所示. 由表 2 可见，采用上述三种核函数的分类平均准确率均高于 85%. 其中，采用径向基核函数，参数选择 $C=32$，$g=0.031\ 25$时所获得的准确率最高，达到了 90.51%，优于多项式核

函数(88.08%)和 Sigmoid 核函数(85.12%).故在下面的计算中,选择径向基核函数进行运算.

表 2 采用不同核函数的预测性能比较

Tab. 2 Prediction accuracy with different type of kernel functions by 10-fold cross validation

10-fold CV	Polynomial	Sigmoid	RBF
1	90.49	89.73	90.11
2	85.98	87.50	92.80
3	86.31	83.27	87.45
4	84.47	84.47	90.91
5	86.69	84.79	93.92
6	89.02	84.09	89.02
7	87.07	83.65	89.73
8	89.77	85.23	88.64
9	87.07	82.51	89.73
10	93.94	88.26	92.80
average	88.08	85.12	90.51

3.2 实验结果比较

使用径向基核函数及参数 $C=32$, $g=0.031\ 25$,采用 jacknife 方法对 Chou 数据集进行测试,即每次从数据集中取出一条序列作测试,而其余序列用作训练.如此轮流,直至数据集中所有序列都被预测一次.与其他三种预测方法对该数据集的预测结果进行比较,如表 3 所示.由表 3 可以看出,本文提出的方法与 Chou 的方法[4]相比,在总的预测精度上有大幅度的提高,接近22%.同时在各子类的预测精度上也有很大的提升.比如,对于亚家族 8,提升了近 70%;对于亚家族 7,提升了约 40%.与 Huang 等的方法[5]相比,总的预测精度提高了近 16%.由于 Huang 等的工作仅给出整体精度,所以无法对各个子类的精度进行比较.

表 3 不同预测方法的性能比较

Tab. 3 Performances comparison for different prediction methods using the jackknife test

class	number of samples	Ref. [4] /%	Ref. [5] /%	Ref. [6] /%	the present work/%
1	314	72.61	—	72.93	92.99
2	216	66.20	—	71.30	87.96
3	194	65.46	—	77.84	89.18
4	130	62.31	—	70.77	86.92
5	112	47.32	—	51.79	84.82
6	305	77.70	—	84.59	94.75
7	64	45.31	—	68.75	84.38
8	59	23.73	—	62.71	93.22
9	254	82.28	—	92.91	94.49
10	94	63.83	—	77.66	87.23
11	154	81.17	—	92.86	97.40
12	94	51.06	—	71.28	85.11
13	257	78.99	—	89.49	93.39
14	155	92.90	—	97.42	99.35
15	84	59.52	—	92.86	98.81
16	154	73.38	—	87.00	92.21
overall	2 640	1 864/2 640 =70.61%	76.63%	2 135/2 640 =80.87%	2 432/2 640 =92.12%

Zhou 等用双极性伪氨基酸组成作为酶蛋白的特征表示,同本文一样,他们也使用支持向量机的方法对样本进行预测.基于进化信息的方法在所有亚家族的测试结果中均高于 Zhou 的方法[6].其中对于子类 5 和 8 的预测效果提升最多,高达 30%以上.同时,整体的预测精度提升了 11%.由此可见,与酶蛋白的双极性伪氨基酸特征表示方法相比,基于 PSSM 的进化信息表示方法由于蕴含更多与酶蛋白功能相关的信息,可以进一步提高分类预测的精度.

3.3 可信度指标

实验中采用 SVM 训练的概率模型,能够赋予每一个待测的酶蛋白样本归属于每一亚家族的隶属度,因此可以用测试样本隶属度的值为最终的预测结果提供一定的可信度度量.为此定义可信度指标(reliability index, RI)[15]为最大隶属度和第二大隶属度之间的差异,具体表示为

$$\mathrm{RI}=\mathrm{integer}(\mathrm{diff}*10)+1 \tag{8}$$

式中,integer( )是取整函数, diff 为最大隶属度和第二大隶属度之间的差值.利用式(8)计算得到的 RI 值可以为分类结果提供一定的置信度,RI 值越大,预测可靠程度越高.图 1 是使用 jacknife 方法对 Chou 数据进行预测时,具有不同 RI 值的序列占整个序列数量的比例以及按照可信度指标进行归类计算得到的不同可信度指标下的平均预测精度.如 RI

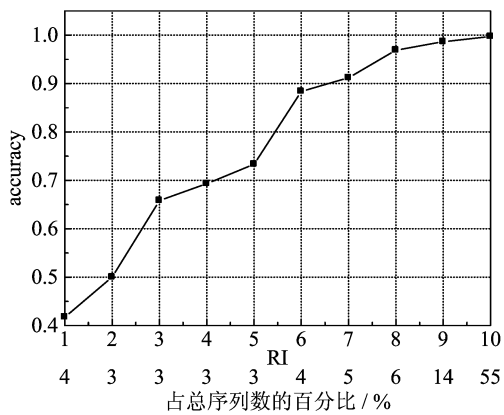

图 1 预测的可信度指标分布及不同可信度指标对应的预测精度

Fig. 1 The fraction of sequences and expected prediction accuracy according to reliability index

=9 的序列占总序列数的 14%，其对应序列的平均预测精度为 98.6%.

由图 1 可见，可信度比例较低的部分集中在 RI=1～6部分，所占整体序列的比例基本相同，约为 4%；RI>5 的序列达到整个序列总数的 84%；RI 为 10 的序列达到了总数的 55%. 这表明本文方法不仅整体置信度高，而且置信度高的部分比例也非常高. 同时易见，随着可信度指标增加，其平均预测精度也逐渐增高；当 RI 增加到 6 时，平均预测精度升至 88.4%，而当 RI 大于 7 时，平均预测精度接近 100%，说明使用隶属度差异定义的可信度指标可以较为客观地反映预测结果的可信度，从而为最终的分类决策提供更多的依据. 图 2 为不同可信度指标阈值下的平均预测精度. 由图 2 可见，87%的序列的 RI≥5，其中 97.5%的序列能够被本文的方法正确预测.

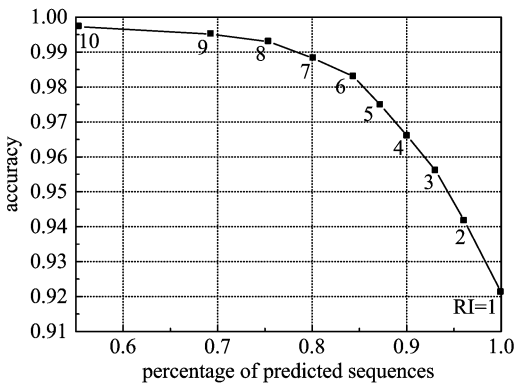

图 2　不同可信度指标阈值下的平均预测精度

Fig. 2　Average prediction accuracy with a reliability index above a given cut-off

4　结论

对酶蛋白的类型进行预测能够帮助人们了解其功能，这对与酶相关的疾病研究以及药物设计等都有非常大的帮助. 考虑到酶蛋白的功能具有保守性，本文采用酶蛋白序列中蕴含的进化信息作为其特征表示，结合支持向量机方法对酶蛋白的亚家族进行了预测，预测精度达到了 92.12%，是目前所有预测方法在该数据集上获得的最好结果. 在此基础上，该方法可以进一步改进，结合更多的表征酶蛋白的信息，如结构信息、序列 motif 信息等，从而提高最终的预测效果；同时，很容易将该方法扩展到酶蛋白与非酶蛋白的预测、酶蛋白家族的预测，进而构建一个自上而下的酶蛋白预测系统，这是我们下一步的工作方向.

参考文献(References)

[1] Webb E C. Enzyme Nomenclature [M]. San Diego, CA: Academic Press, 1992.

[2] Bairoch A. The ENZYME database in 2000[J]. Nucleic Acids Research, 2000, 28(1): 304-305.

[3] Chou K C, Elrod D W. Prediction of enzyme family classes [J]. Journal of Proteome Research, 2003, 2(2): 183-190.

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

[5] Huang W L, Chen H M, Hwang S F, et al. Accurate prediction of enzyme subfamily class using an adaptive fuzzy k-nearest neighbor method [J]. BioSystems, 2006, 90(2): 405-413.

[6] Zhou X B, Chen C, Li Z C, et al. Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes[J]. Journal of Theoretical Biology, 2007, 248 (3): 546-551.

[7] Altschul S F, Madden T L, Schäffer A A, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs [J]. Nucleic Acids Research, 1997, 25(17): 3 389-3 402.

[8] Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000 [J]. Nucleic Acids Research, 2000, 28(1): 45-48.

[9] Jones D T. Protein secondary structure prediction based on position-specific scoring matrices[J]. Journal of Molecular Biology, 1999, 292(2): 195-202.

[10] Xie D, Li A, Wang M H, et al. LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST [J]. Nucleic Acids Research, 2005, 33(S1): W105-110.

[11] Kumar M, Gromiha M M, Raghava G P. Identification of DNA-binding proteins using support vector machines and evolutionary profiles [J]. BMC Bioinformactics, 2007, 8: 463.

[12] Vapnik V N. Statistical Learning Theory [M], New York: John Wiley, 1998.

[13] Hua S, Sun Z. Support vector machine approach for protein subcellular localization prediction [J]. Bioinformatics, 2001, 17(8): 721-728.

[14] Brown M P S, Grundy W N, Lin D, et al. Knowledge-based analysis of microarray gene expression data using support vector machines[J]. Proceedings of National Academy of Sciences, 2000, 97(1): 262-267.

[15] Huang Y, Li Y. Prediction of protein subcellular locations using fuzzy k-NN method[J]. Bioinformatics, 2004, 20(1): 21-28.