基于 Boosting 机制的决策树集成分类器识别嗜热和常温蛋白

Application of Boosting-based Decision Tree Ensemble Classifiers for Discrimination of Thermophilic and Mesophilic Proteins

张光亚,方柏山*

ZHANG Guang-Ya and FANG Bai-Shan*

华侨大学工业生物技术研究所,泉州 362021

Institute of Industrial Biotechnology, Huaqiao University, Quanzhou 362021, *China*

摘 要 采用 Boosting 机制的决策树集成分类器对嗜热和常温蛋白进行模式识别。通过自一致性检验、交叉验证和独立样本测试三种方法检测,其中作为 Boosting 算法中新的 Logitboost 算法表现更好,其识别的精度分别为 100%、88.4%和 89.5%,优于神经网络的识别效果。同时探讨了蛋白质分子大小对识别效果的影响。结果表明,将 Boosting 算法与其它单一分类器有效结合,有望提高研究者对生物分子相关特性的识别能力。

关键词 Boosting,决策树,集成分类器,模式识别,嗜热蛋白

中图分类号 Q51 **文献标识码** A **文章编号** 1000-3061(2006)06-1026-06

Abstract In this paper, the Boosting-based decision tree ensemble classifiers were applied to discriminate thermophilic and mesophilic proteins. Three methods, namely, self-consistency test, 5-fold cross-validation and independent testing with other dataset, were used to evaluate the performance and robust of the models. Logitboost, as a novel classifier in Boosting algorithm, performed better than Adaboost. The overall accuracy of the three methods was 100%、88.4% and 89.5%, respectively. It was demonstrated that LogitBoost performed comparably or even better than that of neural network, a very powerful classifier widely used in biological literatures. The influence of protein size on discrimination was addressed. It is anticipated that the power in predicting many bio-macromolecular attributes will be further strengthened if the Boosting and some other existing algorithms can be effectively complemented with each other.

Key words Boosting, decision tree, ensemble classifier, pattern recognition, thermophilic protein

了解蛋白质热稳定性一直是基础研究和工业应用的热点,它有助于认知蛋白质折叠、蛋白质结构和功能的关系以及设计用于高温环境的生物催化剂[1]。蛋白质热稳定性与其三维结构密切相关,但众所周知,人们对蛋白质三维结构的了解远少于对其序列的了解,这主要是由于获取蛋白质三维结构的信息远比获取其序列信息困难。许多研究[2~4]证实一些重要因素,如:氨基酸组成,较多的 α-螺旋及静电相互作用等都有助于提高蛋白质热稳定性。其中部分因素计算非常困难且要求获取相应蛋白的晶

Received: May 8, 2006; Accepted: July 12, 2006.

This work was supported by the grants from the Science Foundation of Overseas Chinese Affairs Office of the State Council of China (No.05Q0018), and the Key Science and Technology Foundation of Fujian, China (No. 2003I020).

* Corresponding author. Tel: 86-595-22691560, E-mail: fangbs@hqu.edu.cn

国务院侨办科研基金项目(No.05Q0018)和福建省科技计划项目重点项目基金(No.2003I020)资助。

体结构,因而在应用中受到限制[5]。研究者迫切需要一种能直接从序列出发来预测蛋白质热稳定性的工具。近年来,尽管研究者对蛋白质热稳定性机理的探讨较多,但对于蛋白质热稳定性的理论预测却很少[6]。最近,Cheng 等[7]利用蛋白质序列和结构的信息对发生单个氨基酸突变后的蛋白质热稳定性进行了预测,它们采用支持向量机的方法且预测的精度达到了 84%。我们以前也发展了一种基于二肽组成的统计学方法对嗜热和常温蛋白进行了识别,通过测试样本的检验,其识别的精度达到 89%[8],但该方法计算比较繁琐。

集成学习(ensemble learning)是一种新的组合学习器方法(学习算法),它通过某种组合方式把一些学习器组合起来,使得组合后的学习器能够表现出比单个学习器更好的性能。集成学习是机器学习的一个重要分支。最近几年,在机器学习、统计学等领域的众多研究者都投入到集成学习的研究中,使该领域成为一个相当活跃的研究热点,并位于当前机器学习领域四大研究方向之首[9]。现已有多种集成学习算法,Boosting 算法就是其中最有影响和广泛应用的一个。该算法由 Freund 和 Schapire 于 1990 年提出[10],是提高学习系统预测能力的有效工具之一,目前 Boosting 已成功应用于文本分类[11],人脸识别[12]以及手写体数字识别[13],但在生物学领域的应用非常少[14]。

本文采用 Boosting 算法中最典型的 Adaboost 和较新颖的 Logitboost,并组合决策树算法形成集成分类器,通过 8416 组数据训练,取得了良好的识别效果,并将该算法与另一种流行的集成学习算法——Bagging 集成分类器进行比较,同时与传统的 BP 神经网络以及径向基函数(RBF)神经网络进行了比较,并分析了蛋白质分子大小对识别效果的影响。

1 材料和方法

1.1 数据来源

训练数据分别来源于 9 种常温微生物和 15 种嗜热微生物(见表 1)。序列来源于 Swiss-Prot,这是一个非冗余专家库,为了进一步减少信息冗余,剔除了所有注释为推测的(putative)、可能的(probable)、假设的(hypothetical)、部分的(partial)和片段(fragment)的蛋白质序列,最后分别得到 3521 条嗜热和 4895 条常温蛋白序列,共计 8416 组。所有序列可通过 http://iib.hqu.edu.cn/prbi/index.htm 获取。

表 1 训练数据的来源

Table 1 Sources of training dataset

	Strain names	NOS	Adaboost	
			NOCI	ACC (%)
Thermophilic proteins	*Aeropyrum pernix*	235	229	97.45
	Archaeoglobus fulgidus	245	238	97.14
	Methanobacterium thermoautotrophicum	46	44	95.65
	Methanococcus jannaschii	354	345	97.46
	Methanopyrus kandleri	331	322	97.28
	Pyrobaculum aerophilum	201	193	96.02
	Pyrococcus abyssi	298	292	97.99
	Pyrococcus furiosus	266	263	98.87
	Sulfolobus acidocaldarius	**90**	**82**	**91.11**
	Sulfolobus solfataricus	287	271	94.43
	Sulfolobus tokodaii	233	220	94.42
	Thermoplasma acidophilum	216	201	93.06
	Thermoplasma volcanium	182	174	95.60
	Thermotoga maritima	367	356	97.00
	Thermus thermophilus	170	168	98.82
Mesophilic proteins	*Bacillus halodurans*	511	483	94.52
	Chlamydia trachomatis	343	332	96.79
	Deinococcus radiodurans	367	364	99.18
	Lactococcus lactis	585	564	96.41
	Mycoplasma genitalium	241	234	97.10
	Rickettsia prowazekii	**346**	**327**	**94.51**
	Shigella flexneri	1157	1136	98.18
	Synechocystis sp.	640	632	98.75
	Yersinia pestis	705	696	98.72
	Total	8416	8166	97.03

NOS: number of sequences; NOCI: number of correctly identified; ACC: accuracy

1.2 Boosting 算法

Boosting 用某个学习算法生成一系列的基分类器,每个基分类器的训练依赖于在其之前产生的分类器的分类结果,基分类器在训练集上的错误率用于调整训练实例的概率分布,最终分类器通过单个基分类器的加权投票建立起来。其基本思想是,找出若干个精度比随机预测略高的弱规则,再将这些弱规则组合成一个高精度的强规则。

1.2.1 Adaboost 算法:Adaboost 算法最早由 Freund 和 Schapire 于 1997 年提出,是 Boosting 方法中最常用的一种,Adaboost 算法采用加和对数回归模型(additive logistic regression model) $F(x)=\sum_{t=1}^{T}\alpha_t f_t(x)$ 进行拟合,采用的损失函数(loss function)为 $ELOSS(F)=E(e^{-yF(x)})$,该函数是单调、平滑函数,它会随着分类错误呈指数变化,从而使得 Adaboost 算法很容易受到样本中噪音的影响而产生过度拟合现象[15]。为此,Friedman 等于 2000 年提出了 Logitboost 算法。

1.2.2 Logitboost 算法:Logitboost 算法采用二项式对数似然(binomial log-likelihood)损失函数 $LLOSS(F)=E[-\log(1+e^{-yF(x)})]$,该函数随错误呈线性变化,因此对样本中噪音较不敏感,泛化(generalization)能力更强。有关 Logitboost 算法的运算过程请见文献[16]。

1.3 有效性检验

模型的稳定性及泛化能力采用以下三种方法进行检验:(1)自一致性检验(self-consistency test):用于训练的 3521 条嗜热和 4895 条常温蛋白的数据同时也被用来进行预测,以判断是否为嗜热或常温蛋白;(2)交叉验证(cross-validation):根据相关文献[17]采用了 5 倍交叉验证(5-fold cross-validation),具体做法是:将训练的 3521 条嗜热和 4895 条常温蛋白随机分为 5 组(每组约包含 704 个嗜热和 979 常温蛋白),然后采用'留一法(leave-one-out)'进行验证,每次留出 1 组作为测试数据,另 4 组作为训练数据,这样轮流进行 5 次,使得每组数据都能作为测试数据进行预测;(3)独立测试(independent testing):为了进一步验证模型的稳定性,另采用了 859 组数据进行预测,这些数据在上述 8416 组数据中从未出现。该 859 组数据来源于两部分:第一部分来源于嗜热微生物 *Aquifex aeolicus*(最适生长温度为 95℃)[18] 和常温微生物 *Xylella fastidiosa*(最适生长温度为 26℃)[19],同样剔除了带有上述注释的蛋白序列。最终分别得到 382 条嗜热和 325 条常温蛋白。第二部分包含 76 对嗜热和常温蛋白,这 152 组数据来源于文献[20]。

1.4 识别效果评估

各模型最终表现通过以下 4 个参数进行描述:敏感性(sensitivity, SE),特异性(specificity, SP),准确率(accuracy, ACC)和 Matthew 相关系数(Matthew's Correlation coefficient, MCC)。其计算方法见公式(1)~(4)。

$$SE = TP/(TP + FN) \quad (1)$$

$$SP = TN/(TN + FP) \quad (2)$$

$$ACC = (TP + TN)/(TP + FP + TN + FN) \quad (3)$$

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FN) * (TN + FP) * (TP + FP) * (TN + FN)}} \quad (4)$$

式中,TP 为真阳性,指嗜热蛋白被预测为嗜热蛋白;FN 为假阴性,指嗜热蛋白被预测为常温蛋白;TN 为真阴性,指常温蛋白被预测为常温蛋白;FP 为假阳性,指常温蛋白被预测为嗜热蛋白。

文中实现所有算法的软件均来自于 Weka(Waikato environment for knowledge analysis, http://www.cs.waikato.ac.nz/ml/weka/),该程序包是基于 JAVA 虚拟机开发的[21],在生物信息学领域有非常广泛的应用,所有算法的运行参数均采用该软件的默认值。使用的 PC 为 PentiumV 2.7GHz, 512MB RAM。

2 结果与分析

2.1 基于 Adaboost 机制的决策树集成分类器的识别

Adaboost 机制的决策树集成分类器识别效果如图 1a 所示。在自一致性检验中,所有训练数据均用作测试,结果表明该集成分类器的整体识别精度达到了 97.0%,说明经过足够的训练该分类器已掌握了氨基酸组成和蛋白质热稳定性之间的复杂关系,并取得了较好的识别效果,同时说明,尽管采取了一些措施减少信息冗余,仍有部分噪音存在于样本中,并影响了该分类器的识别效果。该分类器对各微生物蛋白质组识别的效果见表 1,可见对不同蛋白质组的识别效果存在一定的差异,从 91.1% 到 99.2% 不等。造成这种现象的原因将在后面进行探讨。

交叉验证的结果表明,该集成分类器分别正确识别出了 2965 个嗜热蛋白和 4323 个常温蛋白,整体识别精度为 86.6%,交叉验证精度的下降可能是由于每次参与训练的数据减少(约 1683 个)而导致

模型训练不充分。敏感性为 84.2%说明该分类器能识别约 84.2%的嗜热蛋白,特异性为 88.3%说明基于 Adaboost 的集成分类器能识别约 88.3%的常温蛋白。这意味着不借助蛋白质的结构信息而仅依赖其序列信息就可以达到一个较高的识别精度。

为了进一步验证该集成分类器的实用性及稳定性,对另外一组独立测试数据进行了测试。在此过程中,所有的 8416 组训练数据均参与训练,然后对 859 组测试数据进行预测,结果表明,该分类器从 458 个嗜热蛋白中成功识别出 400 个,从 401 个常温蛋白中成功识别出 364 个,正确率分别为 87.3%和 90.8%,总体识别精度为 88.9%。相比交叉验证,其精度提高了 2.3%,这可能是由于训练数据增加的缘故。

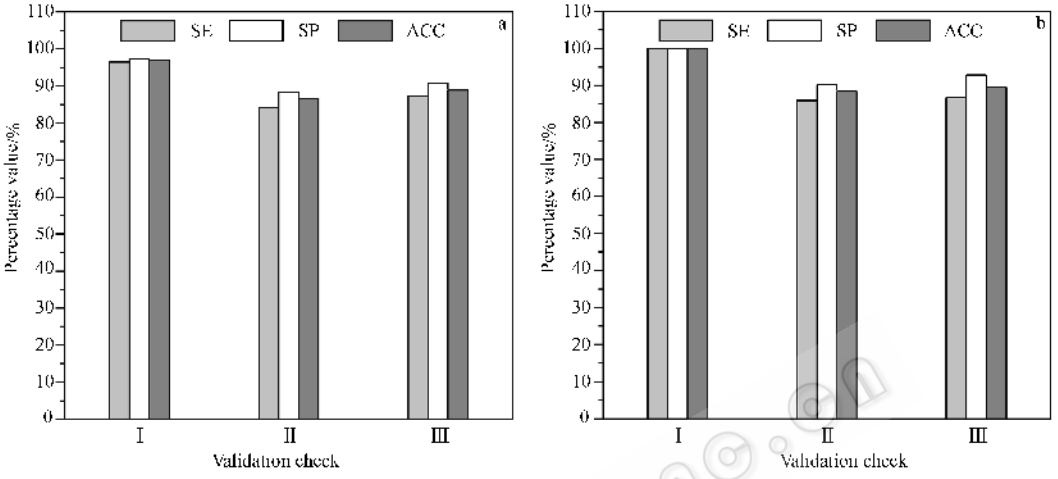

图 1 两种 boosting 集成分类器的识别效果

Fig. 1 Performance of the two boosting-based ensemble classifiers

Ⅰ: self-consistency test; Ⅱ: 5-fold cross-validation; Ⅲ: independent testing; a: Adaboost; b: Logitboost.

2.2 基于 Logitboost 机制的决策树集成分类器的识别

Logitboost 机制的决策树集成分类器识别效果如图 1b 所示。在自一致性检验中,所有参与训练的数据也用作测试,结果表明该集成分类器的整体识别精度达到了 100%,比基于 Adaboost 机制的决策树集成分类器提高了约 3%,这说明尽管在训练样本中存在一定的噪音,但由于 Logitboost 有较强的抗噪音能力,因此对该集成分类器识别效果并未造成影响。

交叉验证的结果表明,该集成分类器分别正确地识别出了 3023 个嗜热蛋白和 4414 个常温蛋白,整体识别精度为 88.4%,交叉验证精度的下降同样可能是由于在交叉验证过程中,由于每次参与训练的数据减少(约 1683 个)了,使模型训练不充分。敏感性为 85.9%说明该分类器能识别约 85.9%的嗜热蛋白,比 Adaboost 提高了约 1.7%;特异性为 90.2%说明基于 Logitboost 的集成分类器能识别约 90.2%的常温蛋白,比 Adaboost 提高了约 1.9%。

为了进一步验证该集成分类器的稳定性及泛化能力,对另外一组独立测试数据进行了测试。在此过程中,所有的 8416 组训练数据均参与训练,并对 859 组测试数据进行预测,结果表明,该分类器从 458 个嗜热蛋白中成功识别出 397 个,从 401 个常温蛋白中成功识别出 372 个,正确率分别为 86.7%和 92.8%,总体识别精度为 89.5%,略好于 Adaboost。虽然 Logitboost 在对嗜热蛋白识别准确率上略逊于 Adaboost,但对常温蛋白识别的精度却提高了 2%。上述三种检验方法均证实 Logitboost 优于 Adaboost。

2.3 与其它分类器识别效果的对比

基于 Boosting 机制的决策树集成分类器与其它 3 种分类器识别效果如表 2 所示。Bagging 算法作为另一种常见集成学习算法,它在三种检验方法中整体正确率分别为 94.6%,87.4%和 89.5%。在自一致性检验中,其识别精度比 Logitboost 低 5.4%,比 Adaboost 低 2.4%,说明它对训练样本中噪音的抗干扰能力弱于 Boosting 机制的集成分类器。在交叉验证以及独立样本测试中效果与 Adaboost 接近,略逊于 Logitboost。

神经网络作为单一分类器的代表,本文采用了两种常见的神经网络模型——RBF 神经网络和 BP 神经网络。对 RBF 神经网络,在三种检验方法中其精度分别为 85.6%,85.4%和 90.2%,除在独立样本测试中其效果略优于 Boosting 之外,在其它两种检验中,其表现均劣于 Boosting,尤其在自一致性检验

中,其效果明显较差,其识别精度分别比 Logitboost 和 Adaboost 低 14.4%和 11.4% ;而 BP 神经网络(隐含层节点数为 11,学习速率为 0.3,动态参数为 0.2,迭代次数为 1000 次,为 weka 软件的默认值)自一致性检验的效果也分别比前两者低 9.4%和 6.4%,这说明基于单一算法的分类器对训练样本中噪音的抗干扰能力较差。尽管 BP 神经网络在交叉验证和独立样本测试中效果略好于 Boosting 机制的组合分类器,但有文献认为[22]自一致性较弱的分类器不能被认为是好的分类器。而且在运算过程中,神经网络对计算机资源的消耗较大,其运算过程约需要 13min,而 Boosting 算法仅需约 5min,在大量数据运算过程中更具优势。综上所述,Boosting 机制的决策树组合分类器识别效果更优。可见,综合多个分类器形成一个集成分类器,能提高识别性能,产生优于单一分类器的识别效果。

表 2 Boosting 与其它算法的比较

Table 2 Compared with other algorithms

Methods	SE	SP	ACC	MCC
self-consistency check				
Adaboost	96.5	97.4	97.0	0.94
Logitboost	100.0	100.0	100.0	1.00
Bagging	94.3	94.7	94.5	0.89
RBF NN	84.0	86.7	85.6	0.70
BP NN	92.8	89.1	90.6	0.81
cross validation				
Adaboost	84.2	88.3	86.6	0.72
Logitboost	85.9	90.2	88.4	0.76
Bagging	84.9	89.2	87.4	0.74
RBF NN	84.0	86.4	85.4	0.70
BP NN	87.0	89.4	88.4	0.76
test with other datasets				
Adaboost	87.3	90.8	88.9	0.78
Logitboost	86.7	92.8	89.5	0.79
Bagging	87.6	91.8	89.5	0.79
RBF NN	88.4	92.3	90.2	0.81
BP NN	93.4	88.0	90.9	0.82

SE:sensitivity, SP:specificity, ACC:accuracy, MCC:Matthew's Correlation coefficient, NN:neural network.

2.4 蛋白分子大小对识别效果的影响

由于此法是完全基于蛋白质的序列信息,因此,识别效果与样本中蛋白质分子的大小可能存在某种联系,为此,将用于预测的 859 个样本按照分子大小分为四类,探讨了蛋白质分子大小对识别效果的影响,其结果如表 3 所示。对于较大的蛋白质分子($\geqslant$800个氨基酸),其识别的精度都很高,分别为 93.1%和 100% ;对于氨基酸数量在 500 到 800 之间的蛋白质分子,这两种方法从 112 个蛋白分子中分别正确识别出了 101 和 105 个,正确率分别为 90.2%和 93.8%,对于中等大小(氨基酸数量在 200 到 500 之间)的蛋白分子而言,其识别精度也较令人满意,分别达到了 91.1%和 90.9% ;而对较小(少于 200 个氨基酸)的蛋白分子,其识别效果较差,分别为 79.0%和 81.4%,比各自平均正确率分别低 9.9%和 8.1%。而 Bagging 算法中也存在类似的现象。

表 3 蛋白分子大小对识别效果的影响

Table 3 Influence of protein size on prediction accuracy

Protein size	Total number	Adaboost		Logitboost		Bagging	
		NOCI	ACC/%	NOCI	ACC/%	NOCI	ACC/%
L$\geqslant$800	29	29	100.0	27	93.1	29	100.0
500$\leqslant$L<800	112	101	90.2	105	93.8	105	93.8
200$\leqslant$L<500	551	502	91.1	501	90.9	506	91.9
L<200	167	132	79.0	136	81.4	129	77.3

NOCI:number of correctly identified, ACC:accuracy.

而对于 Adaboost 自一致性检验中出现的错误,如对 *S. acidocaldarius* 蛋白质组产生错误识别的 8 个蛋白中,有 4 个蛋白的氨基酸数量少于 200,属于较小蛋白分子;而对 *R. prowazekii* 蛋白质组产生错误识别的 19 个蛋白中,有 6 个(约占 31.6%)蛋白也属于较小蛋白质分子。从信息学的角度来理解,可认为较小蛋白分子所包含的信息量(information content)较少,从中提取的 20 个特征向量(指 20 种氨基酸组成)不足以反映其特性。如何提高对较小蛋白分子识别效果将是后续研究的重点。

3 小结

模式识别是生物信息学的重要组成部分,分类是模式识别和机器学习的基本问题,许多分类方法在实际领域得到广泛应用,如决策树、神经网络、贝叶斯方法等,在这些方法中,没有一种总是优于其它分类方法,因此,综合多个分类器的分类结果,形成一个集成分类器,无疑会提高分类性能。本文结果证实了该论断,基于 Boosting 机制的决策树集成分类器在整体识别精度上略好于神经网络的结果,尤其在自一致性检验中显示出了良好的抗干扰能力。

Boosting 方法的有效性使得在实际应用中,可以不再寻找通常很难获得的预测精度很高的强学习算法,只需要找出一个精度略好于随机预测的弱学习算法,就可以通过 Boosting 方法大幅度提高弱预测算法的准确率,从而促进机器学习成果的广泛应用。由于简单有效,Boosting 将会在生物信息学领域有更广泛的应用前景,例如:预测蛋白质的亚细胞定位、膜蛋白的类型、酶活性位点以及蛋白质四级结构类型等。

REFERENCES(参考文献)

[1] Marc Robinson R, Adam G. Structural genomics of *Thermotoga maritima* proteins shows that contact order is a major determinant of protein thermostability. *Structure*, 2005, **6**: 857 – 860

[2] Kumar S, Nussinov R. How do thermophilic proteins deal with heat? *Cell Mol Life Sci*, 2001, **58**: 1216 – 1233

[3] Mozo-Villarias A, Cedano J, Querol E. A simple electrostatic criterion for predicting the thermal stability of proteins. *Protein Eng*, 2003, **16**: 279 – 286

[4] Vieille C, Zeikus GJ. Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. *Microbiol Mol Biol Rev*, 2001, **65**: 1 – 43

[5] Gromiha MM, Oobatake M, Kono H *et al*. Relationship between amino acid properties and protein stability: buried mutations. *J Protein Chem*, 1999, **18**: 565 – 578

[6] Mozo-Villarias A, Querol E. Theoretical analysis and computational prediction of protein thermostability. *Curr Bioinf*, 2006, **1**: 25 – 31

[7] Cheng J, Randall A, Baldi P. Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins*, 2006, **62**: 1125 – 1132

[8] Zhang GY (张光亚), Fang BS (方柏山). A study on the discrimination of thermophilic and mesophilic proteins based on dipeptide composition. *Chinese Journal of Biotechnology* (生物工程学报), 2006, **22**(2): 293 – 298

[9] Dienerich TG. Machine learning research: Four current directions. *AI Magazine*, 1997, **18**: 97 – 136

[10] Schapire RE. The strength of weak 1earnability. *Mach Learn*, 1990, **5**: 197 – 227

[11] Cui L(崔林), Fu KM(付克明), Shi SS(石生树). Naïve Bayesian classifier using boosting mechanism. *Computer Engineering and Applications* (计算机工程与应用), 2005, **8**: 31 – 34

[12] Yang GL(杨国亮), Wang ZL(王志良), Ren JX(任金霞). Facial expression recognition based on Adaboost algorithm. *Computer Applications* (计算机应用), 2005, **25**(4): 946 – 948

[13] Zhao WP (赵万鹏), Gu LY (古乐野). Handwritten digit recognition based on Adaboost. *Computer Applications* (计算机应用), 2005, **25**(10): 2413 – 2415

[14] Feng KY, Cai YD, Chou KC. Boosting classifier for predicting protein domain structural class. *Biochem Biophy Res Co*, 2005, **334**: 213 – 217

[15] Breiman. Arcing classifiers. *Ann Stat*, 1998, **26**: 801 – 849

[16] Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting. *Ann Stat*, 2000, **28**: 337 – 407

[17] Park KJ, Kanehisa M. Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics*, 2003, **19**: 1656 – 1663

[18] Haruyuki A. Recent progress towards the application of hyperthermophiles and their enzymes. *Curr Opin Chem Biol*, 2005, **9**: 1 – 8

[19] David JL, Gregory AS, Donal AH. Synonymous codon usage is subject to selection in thermophilic bacteria. *Nucleic Acids Res*, 2002, **30**(19): 4272 – 4277

[20] Zhang GY, Fang BS. Discrimination of thermophilic and mesophilic proteins via pattern recognition methods. *Process Biochem*, 2006, **41**: 552 – 556

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

[22] Cai YD, Feng KY, Lu WC *et al*. Using LogitBoost classifier to predict protein structural classes. *J Theor Biol*, 2006, **238**: 172 – 176