生物工程学报
journals.im.ac.cn
cjb@im.ac.cn

Chin J Biotech 2008, February 25; 24(2): 302-308
Chinese Journal of Biotechnology ISSN 1000-3061

研究报告

基于氨基酸组成分布的嗜热和嗜冷蛋白随机森林分类模型

张光亚, 方柏山

华侨大学工业生物技术福建省高校重点实验室, 泉州 362021

摘　要: 文献报道采用氨基酸组成分布提取特征值能有效提高预测分类精度, 本文采用该方法提取特征值, 使用一种新的组合分类器——随机森林, 从蛋白质一级结构对嗜热和嗜冷蛋白进行分类。通过 10 倍交叉验证和独立样本测试两种方法检测, 结果表明: 当分段数量为 1 时, 其精度最优, 分别为 92.9%和 90.2%, 暗示使用基于氨基酸组成分布提取特征值在该算法中并不能有效提高识别精度, 这与报道结果不符, 而该提取方法在 SVM 中却能适当提高识别精度; 当引入 6 个新变量后, 其精度分别提高到 93.2%和 92.2%, ROC 曲线下面积分别为 0.9771 和 0.9696, 优于其它组合分类器。

关键词: 随机森林, 氨基酸组成分布, 嗜热和嗜冷蛋白, ROC 曲线

Random Forest for Classification of Thermophilic and Psychrophilic Proteins Based on Amino Acid Composition Distribution

Guangya Zhang, Baishan Fang

Key Laboratory of Industrial Biotechnology, Huaqiao University, Quanzhou 362021, *China*

Abstract: We used amino acid composition distribution (AACD) to discriminate thermophilic and psychrophilic proteins. We used 10-fold cross-validation and independent testing with other dataset to evaluate the models. The results showed that when the segment was 1, the overall accuracy reached 92.9% and 90.2%, respectively. The AACD method improved the prediction accuracy when support vector machine was used as the classifier. When six new features were introduced, the overall accuracy of random forest improved to 93.2% and 92.2%, the areas under the receiver operation characteristic curve were 0.9771 and 0.9696, which was better than other ensemble classifiers and comparable with that of SVM.

Keywords: Random forest, amino acid composition distribution, thermophilic and psychrophilic protein, ROC curve

嗜热和嗜冷微生物是两种重要的极端微生物, 存在于其中的嗜热和嗜冷酶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极端环境的生物催化剂[1]。随着第一个极端嗜热微生物*Methanococcus jannaschii*基因组的公布[2], 研究者通过比较基因组(蛋白质组)的方法对其稳定性机制进行了深入的探讨, 近年来, 不少嗜冷微生物的基因组测序工作陆

Received: May 28, 2007; **Accepted:** September 18, 2007
Supported by: the National "973" Project (No. 2007CB707804) and the natural science foundation of Fujian province (No. 2007J0360).
Corresponding author: Baishan Fang. Tel: +86-595-22691560; E-mail: fangbs@hqu.edu.cn
"973 计划"(No.2007CB707804)和福建省自然科学基金资助项目(No.2007J0360)资助项目。

续完成[3–5]，使得对嗜热和嗜冷蛋白稳定性机理的研究不断深入。尽管研究者对上述极端蛋白稳定性机理的探讨较多[6–8]，但利用蛋白质序列信息对其嗜热和嗜冷特性的理论预测却很少[9]。从蛋白质序列出发对其高级结构及特性进行理论预测所面临的一个重要课题是如何有效提取蛋白质序列特征，氨基酸组成是最常用的一种方法，此外，利用二肽组成[10]和伪氨基酸组成[11]在一些情况下也取得了较好效果。

随机森林算法是Leo Breiman于 2001 年提出的一种新型分类和预测模型[12]，它具有需要调整的参数较少、不必担心过度拟合、分类速度很快，能高效处理大样本数据、能估计那个特征在分类中更重要以及较强的抗噪音能力等特点，因此，在基因芯片数据挖掘、代谢途径分析及药物筛选等生物学领域得到应用并取得了较好的效果，国内报道的随机森林算法在生物学领域的应用很少[13]。

本文采用新近报道的氨基酸组成分布方法提取蛋白序列特征值，并运用随机森林这种新的组合分类算法作为分类器，取得了良好的识别效果，但同时发现使用随机森林作为分类器时，氨基酸组成分布似乎不能有效提高分类精度，但却能提高 SVM 的分类精度，与文献报道存在一定差异，说明使用该方法提取特征值可能需要考虑所使用的分类器算法。

1 材料和方法

1.1 数据来源

样本包含训练和测试数据。训练数据分别来源于 12 种嗜热微生物和 6 种嗜冷微生物蛋白质组序列(见表 1)，其中嗜热数据来源于以前研究[7]，嗜冷蛋白序列为重新从Swiss-Prot下载，为了减少信息冗余，剔除了所有长度小于 100 个氨基酸且注释为推测的(putative)、可能的(probable)、假设的(hypothe- tical)、部分的(partial)和片断(fragment)的蛋白质序列，最后分别得到 3551 条嗜热和 2498 条嗜冷蛋白序列，共计 5049 条，测试数据来源于超嗜热细菌*Aquifex aeolicus*和嗜冷细菌*Colwellia psychrerythraea*，共获得 621 条序列。为了消除上述数据中高度同源

表 1 数据的来源

Table1 Sources of the dataset

	Strain name	Kingdom	OGT	G+C/%	BBC	ABC
Training data	*Pyrococcus furiosus*	A	98	41.2	266	155
	Aeropyrum perni	A	90	57.5	235	195
	Pyrococcus abyssi	A	97	45.1	298	147
	Thermotoga maritime	B	80	46.4	367	332
	Methanopyrus kandleri	A	98	61.2	331	293
	Pyrobaculum aerophilum	A	98	51.9	201	173
	Sulfolobus tokodaii	A	80	33.6	233	181
	Methanococcus jannaschii	A	85	31.3	354	346
	Archaeoglobus fulgidus	A	82	49.4	245	243
	Sulfolobus solfataricus	A	78	36.5	287	234
	Thermus thermophilus	B	75	69.6	440	298
	Pyrococcus horikoshii	A	95	42.3	294	204
	Desulfotalea psychrophila	B	7	46.8	227	227
	Psychrobacter cryohalolentis	B	-10	42.0	119	117
	Psychrobacter arcticum	B	0	42.0	219	198
	Pseudoalteromonas haloplanktis	B	<10	40.1	187	182
	Photobacterium profundum	B	2	41.7	418	410
	Psychromonas ingrahamii	B	-1	40.1	1328	957
Testing data	*Aquifex aeolicus*	B	90	43.5	382	380
	Colwellia psychrerythraea	B	8	37.9	239	232

OGT: optimal growth temperature; BBC: before Blastclust; ABC: after Blastclust; A:archaea; B: bacteria

的蛋白序列对识别效果的影响，使用BLASTCLUST程序[14]剔除了样本所有相似性 (sequence similarity) 大于 30%的序列，使任意两条序列之间的相似性均小于 30%，从而提高识别效果的可靠性。最后共得到 4892 条训练序列和 612 条测试序列，所有序列可通过电子邮件向作者获取。

1.2 随机森林算法基本原理[15]

随机森林是通过一种新的自助法(bootstrap)重采样技术生成很多个树分类器，其步骤如下：

从原始训练数据中生成 k 个自助样本集，每个自助样本集是每棵分类树的全部训练数据。

每个自助样本集生长为单棵分类树。在树的每个节点处从 M 个特征中随机挑选 m 个特征 ($m<<M$)，按照节点不纯度最小的原则从这个 m 特征中选出一个特征进行分支生长。这棵分类树进行充分生长，使每个节点的不纯度达到最小，不进行通常的剪枝操作。

根据生成的多个树分类器对新的数据进行预测，分类结果按每个树分类器的投票多少而定。每次抽样生成自助样本集，全体样本中不在自助样本中的剩余样本称为袋外数据(out-of-bag, OOB)，OOB 数据被用来预测分类正确率，每次的预测结果进行汇总来得到错误率的 OOB 估计，用于评估组合分类器的正确率。该算法运算过程请见文献[10]。

1.3 氨基酸组成分布

按照文献[16]的报道，氨基酸组成分布(amino acid composition distribution, AACD)考虑了氨基酸之间的顺序和总体分布信息，其基本运算过程为：对于一个长度为 L 的蛋白质，将其均分为 n 段，再分别统计这 n 段的氨基酸组成，这样，每个蛋白质序列就可表示成 $20n$ 个特征向量。其中，氨基酸组成成分是是氨基酸组成分布的一种特例，即 n=1。相比多肽组成成分，氨基酸组成分布具有较低的计算阶数，随着 n 的增大，数据计算量只是线型增大，且提供了一种从粗到细多尺度氨基酸组成分析手段，而且可以了解某种氨基酸在蛋白质序列中的概率密度分布。

1.4 有效性检验

模型的稳定性及泛化能力采用以下两种方法进行检验：(1) 交叉验证(cross-validation)：根据相关文献[19]采用了 10 倍交叉验证(10-fold cross-validation, 10-CV)，具体做法是：将训练的 2801 条嗜热和 2091 条嗜冷蛋白随机分为 10 组(每组约包含 280 个嗜热和 209 嗜冷蛋白)，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 9 组作为训练数据，这样轮流进行 10 次，使得每组数据都能作为测试数据进行预测；(2) 独立测试(independent testing, IT)：利用训练数据产生的模型对测试样本进行预测，以进一步检验模型的稳定性及泛化能力。

1.5 识别效果评估

模型最终表现通过以下 5 个参数进行描述：敏感性(sensitivity, SE)，特异性(specificity, SP)，准确率(accuracy, ACC)、Matthew 相关系数(Matthew's Correlation coefficient, MCC)和 ROC 曲线。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，指嗜热蛋白预测为嗜热蛋白；FN 为假阴性，指嗜热蛋白预测为嗜冷蛋白；TN 为真阴性，指嗜冷蛋白预测为嗜冷蛋白；FP 为假阳性，指嗜冷蛋白预测为嗜热蛋白。

受试者操作特性(receiver operation characteristic, ROC)曲线能兼顾灵敏度和特异性要求以综合评价分类器的识别性能，ROC 曲线下面积作为量化指标可以直观有效的比较不同分类器的性能优劣。分析时按不同的“截断点”，可得到反映系统灵敏度的真阳性率 (TPR)和反映系 统特异性的假阳性率(FPR)，然后以 FPR 为横坐标，TPR 作为纵坐标画出 ROC 曲线。曲线越凸说明判别模型诊断价值越高，并可通过计算 ROC 曲线下面积 ($0.5 \le A \le 1$)这一综合统计量作定量分析，任何一个随机猜测的模型其 A 值为 0.5；一个完美的分类器其 A 值为 1，一般而言，A 越接近 1，预测效果越好，不同 ROC 曲线下面积的比较可以作为评价分类器性能的量化指标。文中，TPR 和 FPR 分别定义为：

$$TPR = \frac{\text{实际阳性数目中检出为阳性的数目}}{\text{实际阳性数目}} \quad (5)$$

$$FPR=\frac{\text{非阳性数目中被检测为阳性的数目}}{\text{实际非阳性数目}} \quad (6)$$

其原理和计算过程请见文献[17]。

文中实现所有算法的软件均来自于Weka (Waikato environment for knowledge analysis), 该程序包是基于JAVA虚拟机开发的[18], 在生物信息学领域有非常广泛应用, 所有算法的运行参数均采用其默认值。使用的PC为DELL precision^{TM}490 工　作站。

2　结果与分析

2.1　基于氨基酸组成分布的嗜热和嗜冷蛋白的分类

基于氨基酸组成分布的随机森林分类模型分类效果如图 1 及表 2 所示。10 倍叉验证(10-CV)的结果表明, 当 n=1 时, 该模型分别正确识别出了 2689 个嗜热蛋白和 1854 个嗜冷蛋白, 整体识别精度达到 92.9%, 在所有分段数目中效果最佳; 而在独立样本测试(IT)过程中依然存在类似的现象, 当分段数

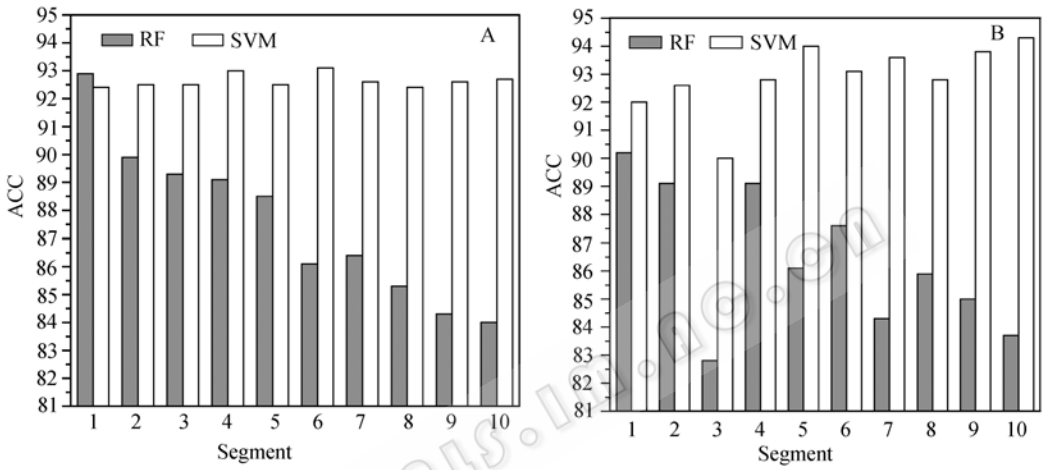

图 1　不同分度数量对识别精度的影响

Fig 1　Influence of different segment on prediction accuracy

A：based on 10-fold cross-validation; B：based on independent test

表 2　不同分段数量对识别效果的影响

Table 2　Influence of different segment on classification results

Segment	10-fold cross-validation								Independent test							
	Random forest				Support vector machine				Random forest				Support vector machine			
	SE	SP	ACC	MCC	SE	SP	ACC	MCC	SE	SP	ACC	MCC	SE	SP	ACC	MCC
1	96.0	88.7	**92.9**	0.85	93.9	90.3	92.4	0.84	94.2	83.6	**90.2**	0.79	95.3	86.6	92.0	0.83
2	95.4	82.6	89.9	0.80	94.1	90.3	92.5	0.85	94.7	79.7	89.1	0.77	96.3	86.6	92.6	0.84
3	95.6	80.9	89.3	0.78	94.0	90.6	92.5	0.85	90.5	70.3	82.8	0.63	94.5	82.8	90.0	0.79
4	95.8	80.1	89.1	0.78	94.5	91.1	93.0	0.86	94.7	79.7	89.1	0.77	95.8	87.9	92.8	0.85
5	95.4	79.2	88.5	0.77	93.5	91.1	92.5	0.85	94.2	72.8	86.1	0.70	96.3	90.1	94.0	0.87
6	94.3	75.2	86.1	0.72	94.3	91.4	**93.1**	0.86	94.7	75.9	87.6	0.73	96.6	87.5	93.1	0.85
7	95.3	74.5	86.4	0.73	93.8	90.9	92.6	0.85	93.4	69.4	84.3	0.66	96.3	89.2	93.6	0.86
8	94.5	73.0	85.3	0.70	93.2	91.3	92.4	0.85	94.5	72.0	85.9	0.70	95.8	87.9	92.8	0.85
9	94.5	70.5	84.3	0.68	93.8	91.0	92.6	0.85	92.1	73.3	85.0	0.68	96.1	90.1	93.8	0.87
10	94.9	69.5	84.0	0.68	93.8	91.3	92.7	0.85	92.1	69.8	83.7	0.65	96.6	90.5	**94.3**	0.88
26 FEA	95.9	89.7	***93.2***	0.86	93.9	90.7	***92.5***	0.85	95.3	87.1	***92.2***	0.83	95.0	86.6	***91.8***	0.83

FEA: features; SE: sensitivity; SP: specificity; ACC: accuracy; MCC: Matthew's Correlation coefficient

目为 1 时，其效果最佳，识别的精度达到 90.2%，为所有分段数目中的最佳。与此同时，在 10-CV 和 IT 过程中，随着分段数目的增加，其识别精度呈现逐步下降的趋势，而造成精度下降的主要原因来源于模型特异性的明显下降，在 10-CV 过程中，SP 值由分段数目为 1 时的 88.7%下降到分段数目为 10 时的 69.5%，降幅达 19.2%；在 IT 过程中，SP 值也由 83.6%下降到 69.8%，降幅达 13.8%。而在此过程中，敏感性下降幅度较小，约在 2%左右。这暗示在使用随机森林作为分类器时，采用基于氨基酸组成分布的方法提取蛋白质序列的特征值似乎并不能有效提高分类精度，这与文献报道中氨基酸组成分布的方法能有效提高分类精度的结果不符。

考虑到随机森林为一种较新的分类算法，而且文献中是以支持向量机(SVM)作为成员分类器，本文考察了 SVM 的分类效果(见图 1 和表 2)，10-CV 的结果表明当 n=6 时，该模型分别正确识别出了 2641 个嗜热蛋白和 1912 个嗜冷蛋白，整体识别精度达到 93.1%，在所有分段数目中效果最佳；而在 IT 过程中，当 n=10 时，模型正确预测出 367 个嗜热蛋白和 210 个嗜冷蛋白的类型，整体精度达 94.3%，为所有分段数目中的最佳。与此同时，随着氨基酸分段数目的增加，其识别精度都出现一定程度的增加，且所有识别效果均优于 n=1 时的结果，这说明用 SVM 作为成员分类器，采用氨基酸组成分布的方法提取蛋白质序列特征值，能取得比采用氨基酸组成更好的结果，尽管其提高的幅度不大(最大为 2.3%)。这与文献报道的结果基本吻合。

2.2 随机森林分类模型识别精度的进一步提升

虽然，使用氨基酸组成分布在随机森林模型中并不能提高识别的精度，但当引入 6 个新变量，分别为CvP-bias、(E+K)/(H+Q)、Class II AA、E/(K+R)、疏水性指数和脂肪族氨基酸指数。其中CvP-bias是带电氨基酸百分含量总和与不带电氨基酸百分含量总和的差值，class II AA为氨基酸G、T、A、P、H、S、D、N、K、F百分含量的总和，根据相关文献的报道，上述这些参数与蛋白质的热稳定性密切相关[19-22]。20 个氨基酸组成以及这 6 个变量共计 26 个变量，以此为特征值的随机森林识别模型效果见表 2 和图 2。

引入 6 个新变量后，在 10-CV 中，RF 分别正确识别出 2685 和 1876 个嗜热和嗜冷蛋白，敏感性和特异性分别为 95.9%和 89.7%，总体识别精度为 93.2%，提高了约 0.3%，ROC 曲线下面积达 0.9771；而在 IT 过程中，RF 分别对 362 和 202 个嗜热和嗜冷蛋白作出了正确预测，敏感性和特异性分别为 95.3%和 87.1%，总体识别精度为 92.2%，ROC 曲线下面积达 0.9696。可见，引入这 6 个新的变量，随机森林算法对独立样本的预测精度有较大提高，约 2%左右。在两种检验方法中，其 ROC 曲线下面积均超过 0.9，说明该算法有着非常优异的表现。

2.3 与其它分类器识别效果的对比

采用 26 个变量，利用 10-CV，比较了随机森林算法与其它几种组合分类器及单一分类器识别效

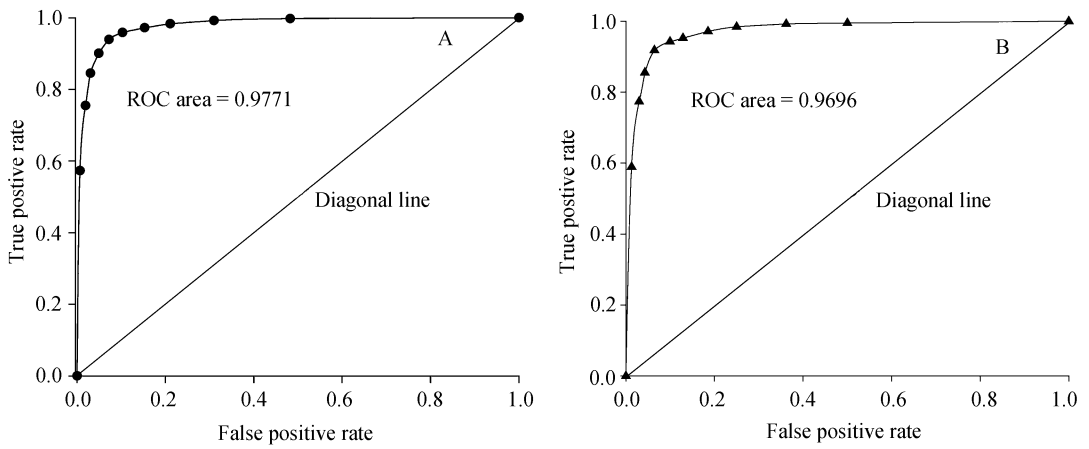

图 2 随机森林分类模型的 ROC 曲线

Fig 2 The ROC curve of RF based on 26 features

A：based on 10-fold cross-validation; B：based on independent test

果，如表 3 所示。Bagging 算法作为一种常见组合学习算法，当采用 4 种不同方法为基函数时，其整体正确率分别为 86.6%，91.2%，92.1%和 92.5%，其识别精度比随机森林低 6.6%、2.0%，1.1%和 0.8%；而在 Adaboost 中，除采用 J4.8 决策树为基函数的识别效果略高于(0.1%)随机森林外，其它均比随机森林低，识别精度最大相差 4.8%；而 Logitboost 集成分类器，其所能采用的两种基函数效果也都略逊于随机森林。

表 **3** 随机森林和其它分类器的比较

Table 3 Comparison of RF to other classifiers

Classifiers		SE	SP	ACC	MCC
Bagging	Decision stump	89.3	83.1	86.6	0.73
	REP tree	93.4	88.2	91.2	0.82
	AD tree	94.2	89.2	92.1	0.84
	J4.8	94.0	90.4	92.5	0.85
Adaboost	Decision stump	89.2	87.4	88.5	0.76
	REP tree	92.2	91.0	91.6	0.83
	AD tree	94.5	91.2	93.0	0.86
	J4.8	94.4	91.7	**93.3**	0.86
LB	Decision stump	92.3	87.9	90.4	0.80
	REP tree	94.1	91.2	92.8	0.85
SVM (linear kernel)		93.9	90.7	92.5	0.85
SVM (RBF kernel)		94.9	87.2	91.6	0.83
Random forest		95.9	89.7	**93.2**	0.86

SE: sensitivity; SP: specificity; ACC: accuracy; MCC: Matthew's Correlation coefficient; LB: Logitboost

SVM 作为单一分类器的典型代表，本文采用了两种常见的 SVM 模型——线性核函数 SVM 和径向基(RBF)核函数 SVM。对线性核函数 SVM，其识别精度为 92.5%(RF 为 93.2%)，ROC 曲线下面积为 0.9229(RF 为 0.9771)，其表现均劣于随机森林；而对 RBF 核函数 SVM，其识别精度为 91.6%，ROC 曲线下面积为 0.9104，可见，其效果也劣与随机森林。不仅如此，随机森林还存在其它优势，如在运算过程中，RBF 核函数 SVM 对计算机资源的消耗较大，10-CV 整个运算过程约需要 15 min，而随机森林算法仅需约 37 s，而另一个识别效果略好于随机森林的基于 Adaboost 的 J4.8 决策树分类器，其运算的时间为 188 s，在上述所有分类器中，只有线性核函数 SVM 的运算速度快于随机森林，约为 12 s。可见，在大量数据运算过程中随机森林优势明显，它能同时兼顾识别精度和运算速度。考虑到 SVM 在氨基酸组成分布中的最佳识别效果略好于随机森林，可以认为随机森林的效果与 SVM 相当。

3 小结

施建宇等[15]提出用氨基酸组成分布提取蛋白质的一级结构信息并对蛋白质同源寡聚体进行分类，取得了优于氨基酸组成和二肽组成的效果；而在本研究中却发现使用该方法，以随机森林算法作为分类器却无法达到预期目标；有趣的是，当采用SVM作为分类器时，基于氨基酸组成分布的效果优于氨基酸组成，与文献的报道基本吻合，这意味采用不同的方法提取蛋白质一级结构信息，其最终效果似乎与采用的分类器算法有一定关系，这是许多研究可能未曾考虑的问题，出现这种现象的原因可能与每种算法自身的特点有关，通常来说，SVM可以忍受较高的维数。使用氨基酸组成分布增加了输入的维数。因此，这可能意味着RF算法对高维数比较敏感，导致当分段数目 n 增大时，总体精度下降。本文所提的现象是否普遍存在，还需要进一步研究和探讨。无论如何，本文的结果都可提醒其他研究者，在使用不同方法提取特征值时，可能还要考虑与所采用分类器算法的匹配，只有找到二者最佳的匹配，才能得到最好的分类效果。

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，许多分类方法在生物信息学领域得到了应用，在这些方法中，没有一种总是优于其它分类方法。因此，研究者需要更多分类工具以适应不同的模式识别问题，本文所采用的随机森林算法作为一种新的组合学习算法，其良好的分类能力和快速的运算能力都得到了充分的体现。由于简单有效，随机森林算法会在生物信息学领域中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、转录起始点以及蛋白质同源寡聚体分类等。

REFERENCES

[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] Bult CJ, White O, Olsen GJ, *et al*. Complete genome sequence of the methanogenic archaea *Methanococcus jannaschii*. *Science*, 1996, **273**: 1058–1073.

[3] Barbara A. M, Karen EN, Jody W. D, *et al*. The psychrophilic lifestyle as revealed by the genome sequence of *Colwellia psychrerythraea* 34H through genomic and proteomic analyses. *PNAS*, 2005, **102**(31): 10913–10918.

[4] Claudine M, Evelyne K, Géraldine P, *et al*. Coping with cold: The genome of the versatile marine Antarctica bacterium *Pseudoalteromonas haloplanktis* TAC125. *Genome Res*. 2005, **15**: 1325–1335.

[5] Rabus R, Ruepp A, Frickey T, *et al*. The genome of *Desulfotalea psychrophila*, a sulfate reducing bacterium from permanently cold Arctic sediments, *Environ Microbiol* 2004, **6**: 887–902.

[6] Thierry L, Charles G, Georges F. Psychrophilic enzymes: revisiting the thermodynamic parameters of activation may explain local flexibility. *BBA-Protein Structure and Molecular Enzymology*, 2000, **1543** : (1): 1–10.

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

[8] Ding YR, Cai YJ, Zhang GX, *et al*. The influence of dipeptide composition on protein thermostability. *FEBS Lett*. 2004, **569**: 284–288.

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

[10] Zhang GY, Fang BS. Study on the discrimination of thermophilic and mesophilic proteins based on dipeptide composition. *Chinese Journal of Biotechnology*, 2006, **22**(2): 293–298.
张光亚, 方柏山. 基于二肽组成识别嗜热和常温蛋白的研究. 生物工程学报, 2006, **22**(2): 293–298.

[11] Chou KC, Cai YD. Predicting protein quaternary structure by pseudo amino scud composition. Proteins, 2003, **53**(2): 282–299.

[12] Breiman L, Random Forests, *Mach Learn*. 2001, **40**:5–32.

[13] Xu HL, Zhang WT, Zhao NQ, Jiang QW. Hypervariable sites screening on HA sequence that affect the virulence of influenza A H5N1 for mammalian species. *Fudan Univ J Med Sci*, 2006, **33**(5): 642–646.
许慧琳, 张文彤, 赵耐青, 姜庆五. 影响 HS N1 甲型流感病毒对哺乳动物毒力变异的 HA 序列关键位点研究. 复旦学报(医学版), 2006, **33**(5): 642–646.

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

[15] Jia FC, Li H. Multi-spectral magnetic resonance image segmentation using random forests. *Computer Engineering*, 2005, **31**(10): 159–161.
贾富仓, 李华. 基于随机森林的多谱磁共振图像分割. 计算机工程, 2005,**31**(10): 159–161.

[16] Shi JY, Pan Q, Zhang SW, Chen YM.. Classification of protein homo-oligomers using amino acid composition distribution. *Acta Biophisica Sinica*, 2006, **22**(1): 49–56.
施建宇, 潘泉, 张绍武, 程咏梅.基于氨基酸组成分布的蛋白质同源寡聚体分类研究.生物物理学报, 2006, **22**(1): 49–56.

[17] Tom Fawcett. An introduction to ROC analysis. *Pattern Recogn Lett*. 2006, 27:861–874.

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

[19] Karsten Suhre and Jean-Michel C. Genomic Correlates of Hyperthermostability, an Update. *J Biol Chem*, 2003, **278**(19), 17198–17202.

[20] Sávio TF, Maria Christina MB. Preferred amino acids and thermostability. *Genet Mol Res*. 2003, **2** (4): 383–393.

[21] Liron K, Ilya S, Boris T, Mark S. Optimal growth temperature of prokaryotes correlates with class II amino acid composition. *FEBS Lett*, 2006, **580**: 1672–1676.

[22] Fredj T, Edouard Y, Bernard D. Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis. *Gene*, 2002, **297**: 51–60.

本 期 广 告 索 引