基于随机森林和特征选择方法的蛋白质热稳定性影响因素预测

张力[1,2]，艾海新[1,2]，张吉宽[3]，胡桓[1]，刘宏生[1,2]，马树才[4]

（1.辽宁大学生命科学院，辽宁沈阳 110036）（2.辽宁省生物大分子计算模拟与信息处理工程技术研究中心，辽宁沈阳 110036）（3.辽宁大学信息学院，辽宁沈阳 110036）（4.辽宁大学经济学院，辽宁沈阳 110036）

摘要：酶的耐热性对其在食品工业中实现应用至关重要。本文以随机森林算法通过蛋白质序列预测酶的热稳定性，并对影响热稳定性的重要特征进行了分析。计算了从 Swiss-Prot 数据库获得的 1600 个包含热稳定性信息的酶的 430 个特征。采用重复欠抽样法处理数据不平衡问题，采用向后递归特征消去法优选出 30 个最重要的特征。通过交叉验证和独立测试比较以各特征子集构建的模型，发现仅以氨基酸组成为特征集构建的模型获得了最佳预测效果，模型的总体预测准确率为 85.83%、敏感性为 89.16%、特异性为 73.33%、精度为 77.00%、F1 度量为 74.87%。结果表明氨基酸组成对酶热稳定性的影响最大，嗜热酶中含有更多的谷氨酸、异亮氨酸和赖氨酸，而常温酶中含有更多的谷氨酰胺、丝氨酸和苏氨酸。研究为蛋白质工程改造食品工业用酶的热稳定性提供了一定的理论和方法。

关键词：酶热稳定性；随机森林；特征选择；氨基酸组成

文章篇号：1673-9078(2016)07-103-108 DOI: 10.13982/j.mfst.1673-9078.2016.7.017

Prediction of the Influencing Factors of Protein Thermal Stability using Random Forest and Feature Selection Techniques

ZHANG Li[1,2], AI Hai-xin[1,2], ZHANG Ji-kuan[3], HU Huan[1], LIU Hong-sheng[1,2], MA Shu-cai[4]

(1.School of Life Science, Liaoning University, Shenyang110036, China) (2.Research Center for Computer Simulating and Information Processing of Bio-macromolecules of Liaoning, Shenyang 110036, China) (3.School of Information, Liaoning University, Shenyang 110036, China) (4.School of Economics, Liaoning University, Shenyang 110036, China)

Abstract: Thermal stability is crucial for implementation of an enzyme in the food industry. The thermostability of enzymes were predicted through protein sequences, using a random forest algorithm and the important influencing factors on protein thermal stability were analyzed. Four hundred and thirty protein features were calculated for 1600 enzymes extracted from the Swiss-Prot database that contained thermal stability information. The data imbalance was solved by using repeated under-sampling methods, and the 30 most-important features were selected by backward recursive feature elimination (RFE). The classification performances of different random forest models built by different feature subsets were evaluated by cross-validation and independent testing. The results indicated that the model built by amino acid composition exhibited the best performance (accuracy = 85.83%, sensitivity = 89.16%, specificity = 73.33%, precision = 77.00%, and F-measure = 74.87%), suggesting that amino acid composition had the most significant impact on the thermal stability of an enzyme. Further, it was found that thermophilic enzymes contained relatively high contents of glutamic acid, isoleucine, and lysine, whereas mesophilic enzymes contained high contents of glutamine, serine, and threonine. The results in this study provided a theory and method for engineering proteins to improve enzyme thermostability for the food industry.

Key words: enzyme thermostability; random forest; feature selection; amino acid composition

收稿日期：2015-08-14

基金项目：辽宁省教育厅基金资助项目（L2014001）；辽宁省科技厅基金资助项目（2014001015；2013225086）；沈阳市科技局科技攻关专项（F14-154-9-00）；国家自然科学基金资助项目（31570160）

作者简介：张力（1988-），男，博士研究生，研究方向生物卫生统计学；艾海新为并列第一作者

通讯作者：刘宏生（1963-），男，教授，博士生导师，研究方向为生物信息学及功能食品学

酶在食品工业生产中有广泛的应用，如淀粉酶、纤维素酶、脂肪酶、果胶酶、蛋白酶、糖化酶等[1]，但是实际工业生产环境复杂，常常伴随高温高压等极端条件，然而大部分酶是常温酶，严重制约了它们在食品行业中的应用[2]。因此，采用蛋白质工程提高酶的热稳定性已成为食品工业微生物的重要研究方向，其中的关键技术突破点就是研究影响酶热稳定性的关键影响因素特征。随着 2013 年 10 月 9 日诺贝尔化学奖授予了“为复杂化学体系设计了多尺度模型”的计算化学，通过生物信息学方法预测酶热稳定性成为食品微生物学新的研究热点。

国内外已报道多种预测蛋白质热稳定性的方法[3~4]。蛋白质的一级结构决定蛋白质的功能，但是研究者普遍认为蛋白质的生化特征（如热稳定性、抗氧化性、耐酸碱特性等）是不能使用一级结构预测的。因此，研究者主要关注蛋白质的三级结构或四级结构特征与热稳定性之间的关系。最近有研究者利用蛋白质组学方法比较分析了嗜热微生物和常温微生物的蛋白质组，发现在氨基酸组成上有显著的差别[5~6]；也有报道利用蛋白质氨基酸组成和二肽组成区分嗜热蛋白和常温蛋白，都达到 80%以上的准确率[7~8]。然而，尚未见对影响蛋白质热稳定性的一级结构特征分析的报道。

随机森林（Random Forests, RF）[9]是由 Leo Breiman 提出的一种基于决策树理论的集成机器学习方法，具有很高的预测准确率，且不易过拟合，对存在噪声和缺失值的数据集有很好的容忍度，并且随机森林方法可以给出变量重要性的内在估计，更扩大了其应用范围。近年来随机森林算法已经在生物信息学和食品工业中得到了广泛的应用并取得了良好的效果[10~11]。

本研究提取了蛋白质的氨基酸组成、二肽组成和氨基酸类组成等一级结构特征，运用随机森林算法作为分类器对蛋白质的热稳定性进行分类预测，并对影响蛋白质热稳定性的关键氨基酸特征进行了讨论。本研究使用的方法对通过蛋白质工程改造食品工业微生物酶的热稳定性提高及其在工业生产中的应用具有理论和现实指导意义。

1 材料与方法

1.1 数据收集

从 Swiss-Prot 蛋白质序列数据库（2015.04 版本）中获得了 1852 条包含热稳定性信息的蛋白质序列。为了保证训练和测试数据的质量，删除了酶序列中属于前导肽和信号肽部分的序列、剔除了序列长度小于 50 个氨基酸、注释中完整性为片段（fragment）、以及不是酶的蛋白质序列。使用 BLAST 2.2.26 中的 blastclust 程序计算了各序列间的相似性，剔除序列相似性极高的（>95%）序列，从而减少数据冗余，提高预测模型的可靠性。最终剩余 1600 条酶序列。来自于嗜热菌和极端嗜热菌的酶的最适反应温度一般会大于 70 ℃，最适反应温度大于 70 ℃的嗜热酶也更有实际应用价值，因此根据酶的最适催化温度将这些序列分为两类：常温酶（最适温度<70 ℃，共 1135 条）和嗜热酶（最适温度≥70 ℃，共 465 条）。

1.2 特征计算

共计算了包括氨基酸组成（amino acid composititon）、二肽组成（dipeptide composition）、氨基酸类组成（amino acid class composition）在内的 430 个蛋白质序列特征。

氨基酸组成表示 20 种氨基酸在蛋白质序列中出现的频率，氨基酸组成特征由氨基酸的单字母形式表示，如 A 表示丙氨酸在序列中所占的比例。二肽组成表示两个氨基酸相连形成的二肽，二肽组成特征由两个氨基酸的单字母形式表示，如 AC 表示由丙氨酸和半胱氨酸组成的二肽在序列中所占的比例。氨基酸类组成：根据氨基酸侧链基团的极性可将氨基酸分为非极性（nonpolar）、极性（polar）、碱性极性（basic polar）、酸性极性（acidic polar）、以及所有极性氨基酸（包括极性、碱性极性和酸性极性，all polar）等五类类；根据氨基酸侧链基团带电情况可将氨基酸分为中性（neutral）、带电（charged）两类。根据亲水指数的不同可将氨基酸分为三类，分别为亲水指数>0 的疏水氨基酸（hydrophobic），亲水指数≤0 或≥-2 的亲水氨基酸（hydrophilic），和亲水指数<-3 的极亲水氨基酸（large hydrophilic）。分别计算这几类氨基酸在蛋白质序列中出现的频率。

1.3 特征选择

本研究共生成了 430 个蛋白质序列特征，为了找出对酶热稳定性影响最大的特征，并利用选择出的特征构建更准确的分类器，我们需要对特征进行筛选。特征选择使用了 R 语言“caret”软件包实现的递归特征消去（Recursive Feature Elimination, RFE）变量选择方法和 R 语言“randomForest”软件包中的随机森林算法，选择过程采用了 5×10 折交叉验证。首先，将数据集中所有序列的所有特征信息作为训练数据，使用随机森林算法计算出一个包含 2000 个决策树的随

机森林模型，使用此模型对特征的重要性进行排序。然后使用向后 RFE 法，在每一次迭代中依次消去重要性分值最低的特征，使用新的特征子集重新生成新的随机森林模型，不断重复此过程，直到消去所有的特征为止，依据模型的准确率评估该特征子集的分类能力，然后选择使随机森林模型获得最高准确率的特征子集。

1.4 随机森林分类模型训练

由于本研究数据集中嗜热酶与常温酶之间样本数目差别很大，产生了类不平衡问题。类不平衡将导致分类模型对少数类（嗜热酶）的预测准确度大大降低。本文尝试采用欠抽样的方法提高模型预测准确度，而欠抽样会损失多数类（常温酶）中所包含的信息，为了弥补这一缺陷，本文采用了重复欠抽样法，对多数类进行 10 次欠抽样，然后分别与少数类组合，训练 10 个随机森林模型，然后对这些模型进行综合，得到最终的分类模型。

本研究分别对所有特征、氨基酸组成、氨基酸类组成、二肽组成、RFE 方法选择的特征，以蛋白质的热稳定性为预测变量，使用 R 语言训练了随机森林模型，使用 5×10 折交叉验证法对各模型进行比较评价，并选出平均性能最好的模型作为最终的预测模型。对最终的预测模型进行了独立测试，将原始数据集分为训练集和测试集，测试集为随机从原始数据中抽取出 400 条数据，训练集为剩下的 1200 条数据。利用训练集的数据产生预测模型，使用此模型预测测试集中蛋白质的热稳定性。

1.5 模型评价方法

对于类不平衡数据通常使用准确率（Accuracy, ACC）、敏感性（Sensitivity, SE）、特异性（Specificity, SP）、精度（Precision, PR）和 F 度量（F-measure, F）作为模型的评价参数。其中准确率定义为分类器正确预测的嗜热酶和常温酶的数量占数据集中所有酶的比例；敏感性也叫召回率（recall）定义为分类器正确预测的嗜热酶的数量占数据集中所有嗜热酶的比例；特异性定义为分类器正确预测的常温酶的数量占数据集中所有常温酶的比例；精度定义为分类器正确预测的嗜热酶的数量占分类器预测得出的嗜热酶总数的比例；F 度量为精度和召回率的调和均值。计算公式如下：

$$ACC=\frac{TP+TN}{TP+TN+FN+FP}$$

$$SE=\frac{TP}{TP+FN}$$

$$SP=\frac{TN}{TN+FP}$$

$$PR=\frac{TP}{TP+FP}$$

$$F=\frac{2\times PR\times SE}{PR+SE}$$

其中 TP 为真阳性，表示嗜热酶被正确预测为嗜热酶；TN 为真阴性，表示常温酶被正确预测为常温酶；FN 为假阴性，表示嗜热酶被预测为常温酶；FP 为假阳性，表示常温酶被预测为嗜热酶。

1.6 数据统计分析

应用 R 3.1.3 和 SigmaPlot 12.5 对实验数据进行统计分析和作图，除文中注明实验重复方法外，所有实验都进行至少三次，数据取平均值，标准偏差使用误差棒在图中表示。

2 结果与讨论

2.1 随机森林参数确定

随机森林法通常对参数的设置不太敏感，参数的取值对模型的分类预测和特征选择结果往往影响很小。尽管如此，为了获得更精确的分类结果和更可靠的特征选择结果，对随机森林中的主要参数 ntree（分类器中决策树的数目）进行了调优。图 1 显示了在使用全部特征和记录（酶序列）的条件下 ntree 的值在 10~2000 范围内随机森林模型的准确率、敏感性、特异性、精度和 F 度量的变化情况，可以看出随着 ntree 的增加模型的准确性在上升，当 ntree 达到 500 后模型的准确性已经没有明显的增加了。因此，试验中 ntree 都设定为 500。

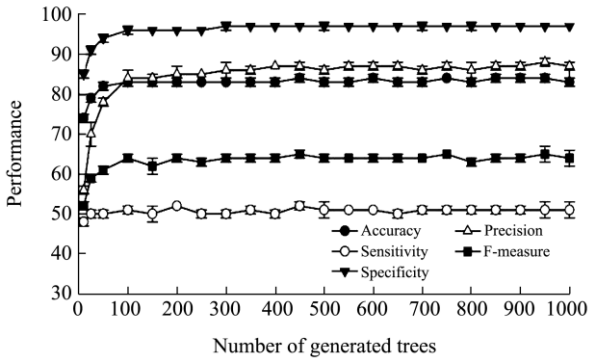

图 1 随机森林参数 ntree 的优化

Fig.1 Optimization of random forest parameter ntree

Note: Accuracy, sensitivity, specificity, precision, and F-measure at different ntree values were performed on all features and all records.

2.2 重复欠抽样法和递归特征消去法提升随机森林模型

图 1 的结果显示，当随机森林算法以全部特征和记录（酶序列）为数据集得出的模型的预测准确率、特异性和精度都大于 80%，而敏感性和 F 度量较低，表明模型预测嗜热蛋白的能力很弱，这可能是由数据不平衡和数据集的特征数较多引起的。

针对数据不平衡问题，采用重复欠抽样法对嗜热酶进行 10 次欠抽样，分别与常温酶组合，训练 10 个随机森林模型，然后对这些模型进行综合，得到最终的分类模型。最终模型的准确率为 83.79%，与欠抽样前比较没有明显变化，敏感性和 F 度量分别由欠抽样前的 50.47%和 63.77%升高到 67.31%和 70.71%，而特异性和精度分别由的欠抽样前的 96.80%和 86.60%降低为 90.54%和 74.46%。可以看出，重复欠抽样法可以在不影响模型总体准确率的前提下显著提升模型的敏感性和 F 度量，从而提高模型预测嗜热蛋白能力，而缺点是模型预测常温蛋白能力有所降低。

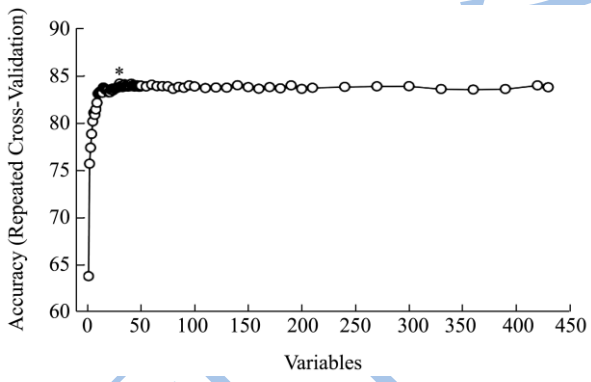

图 2 向后递归特征消去法特征选择结果

Fig.2 Result of feature selection using backwards recursive feature elimination

Note: The highest accuracy was achieved by the model trained based on the subset of top 30 features (marked with an asterisk).

特征选择方法可以从 430 个蛋白质特征中挑选出与酶热稳定性最相关的特征。在重复欠抽样的基础上使用随机森林算法和向后递归特征消去法（Recursive Feature Elimination, RFE）进行了特征选择，选择过程采用 5×10 折交叉验证，特征选择结果见图 2。可以看出，当纳入随机森林中的特征数目为 30 时，模型的准确率达到了最大值，因此选择 Q、E、polor、S、I、charged、neutral、C、K、H、EK、IK、EE、V、AD、acidic_polar、basic_polar、KV、T、KE、QA、II、RE、AA、GV、Y、EI、VI、VK 和 EG 等 30 个特征为最佳特征子集。同时可以看出当随机森林包含更多的特征时，模型准确率没有明显的下降，这也验证了随机森林算法对噪音有较好的容忍度。

使用重复欠抽样法以最佳特征子集和全部记录为数据集，构建随机森林模型，其准确率、敏感性、特异性、精度和 F 度量分别为 84.63%、71.40%、90.04%、74.61%和 72.97%，除特异性稍有降低外，其他评价指标均有所提高。

2.3 不同特征子集对随机森林模型的影响

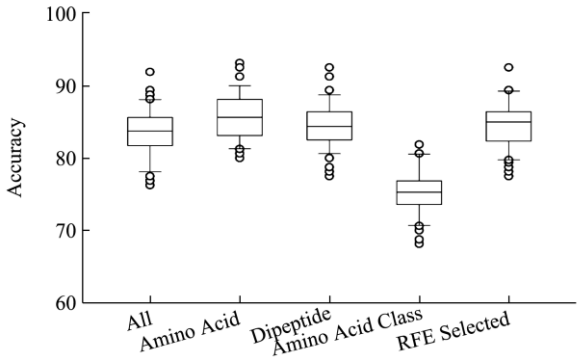

图 3 不同特征子集对随机森林模型预测准确率的影响

Fig.3 Influence of different feature subsets on prediction accuracy of random forest model based on 5×10 fold cross validation

表 1 不同特征子集对随机森林模型预测性能的影响

Table 1 Influence of different feature subsets on prediction performance of random forest model based on 5×10 fold cross validation.

Feature subset	All	Amino Acid	Dipeptide	Amino Acid Class	RFE Selected
Accuracy	83.51	85.83	84.44	75.25	84.52
Sensitivity	68.22	73.33	68.09	58.22	71.25
Specificity	87.26	89.16	87.38	82.75	88.33
Precision	73.28	77.00	76.14	57.33	74.46
F-measure	70.37	74.87	71.56	57.43	72.57

分别以全部特征（All）、氨基酸组成特征（Amino Acid）、二肽组成特征（Dipeptide）、氨基酸类组成特征（Amino Acid Class）和 RFE 法选择的特征（RFE Selected）结合重复欠抽样法构建随机森林模型，使用 5×10 折交叉验证法对各模型进行比较评价。所有模型交叉验证的结果见图 3 和表 1。可以看出，以氨基酸组成特征得到的随机森林模型的准确率最高，达到了 85.83%，稍高于 RFE 法选择的特征和二肽组成特征，使用氨基酸类组成特征的预测准确率最低。同时仅使用氨基酸组成特征得到的随机森林模型的敏感

性、特异性、精度和 F 度量的值都比其他特征子集高，特别是能综合评价不平衡数据集预测效果的 F 度量值达到了 74.87%，显著高于其他特征子集。可以看出氨基酸组成可能是影响蛋白质热稳定性最为关键的特征集。可以推测氨基酸组成对酶热稳定性的影响较大，近年发表的文献也发现了类似的现象[5,12]。

2.4 随机森林模型的独立测试

将原始数据集划分为训练集和测试集，测试集为随机从原始数据集中抽取出 400 条记录，训练集为剩下的 1200 条记录。用训练集中的记录以氨基酸组成作为特征集，采用重复欠抽样法训练随机森林模型，使用此模型预测测试集中蛋白质的热稳定性，对上述方法重复 10 次，结果取平均值。此预测模型在测试集上预测结果见表 2。预测模型的总体预测准确率为 85.52%、敏感性为 90.87%、特异性为 72.14%、精度为 76.01%、F1 度量为 74.02%，与上一节中使用交叉验证法的结果基本一致，说明此模型稳定性好，具有较好的泛化能力，可用于使用蛋白质序列预测酶的热稳定性。

表 2 氨基酸组成作为特征集的随机森林模型得到的混淆矩阵

Table 2 Confusion matrix obtained using the random forest model with amino acid composition as feature subset

		Predicted	
		mesophilic	thermophilic
Real	mesophilic	259.50 (7.67)	26.06 (4.66)
	thermophilic	31.88 (4.71)	82.56 (6.36)

Note: Standard deviations are given in parentheses.

2.5 影响酶热稳定性的氨基酸

明确各种氨基酸的含量对酶的热稳定性有何影响，对使用蛋白质工程改造酶具有重要意义。氨基酸含量对酶热稳定性影响的大小可以从其在随机森林模型中的重要性来推测。因此，计算上一节中仅使用了氨基酸组成特征的随机森林模型中各特征的重要性，结果见图 4。可以看到在 20 种氨基酸中谷氨酸（E）和谷氨酰胺（Q）含量的多少对酶热稳定性的影响最显著，其次是丝氨酸（S）、异亮氨酸（I）和赖氨酸（K），而亮氨酸（L）、脯氨酸（P）、苯丙氨酸（F）以及甘氨酸（G）对酶热稳定性的影响最小。图 5 使用盒图表示了 20 种氨酸的含量在常温酶和嗜热酶中的分布情况。从图中可以看出，嗜热酶倾向于含有更多的谷氨酸（E）、异亮氨酸（I）和赖氨酸（K），而常温蛋白质倾向于含有更多的谷氨酰胺（Q）、丝氨酸（S）和苏氨酸（T）。Ebrahimi [13]等人通过决策树发现谷氨酰胺（Q）的含量是区分嗜热酶和常温酶最关键的特征。Singer 和 Hickey [5]通过分析嗜热原核生物的基因组数据发现嗜热微生物以常温微生物的蛋白质组中某些氨基酸的频率有显著差异，嗜热微生物含有较少的谷氨酰胺、苏氨酸和组氨酸，较多的谷氨酸、异亮氨酸、赖氨酸和缬氨酸。Gromiha 和 Suresh [14]也通过机器学习算法发现一些带电荷的氨基酸（如赖氨酸、精氨酸和谷氨酸）和一些疏水氨基酸（如缬氨酸和亮氨酸）在嗜热酶中的出现概率更高。因此在蛋白质工程中，可以根据此规律增加蛋白质的 E、I 和 K，减少蛋白质中的 Q、S 和 T，从而增加酶的热稳定性。

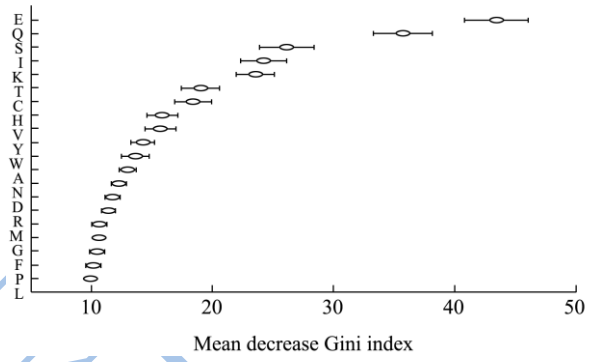

图 4 随机森林中各氨基酸组成特征的重要性

Fig.4 Importance of amino acid composition in the random forest model

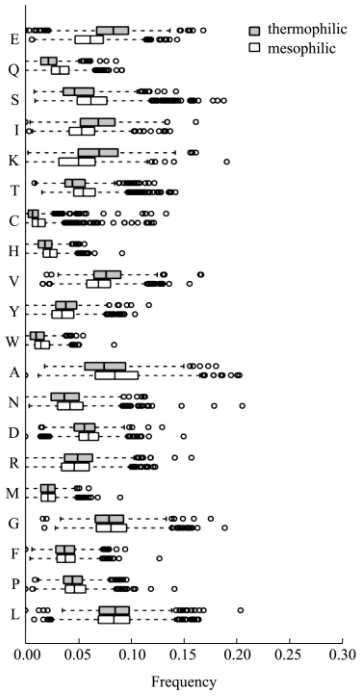

图 5 20 种氨基酸在常温酶和嗜热酶中含量的分布情况

Fig.5 Distribution of 20 amino acid contents of thermophilic and mesophilic enzymes estimated based on 465 mesophilic and 1135 thermophilic enzyme sequences

3 结论

使用随机森林算法构建嗜热酶和常温酶分类模型，在 Ntree 达到 500 时模型预测准确率达到稳定，且已具有较高的准确率、特异性和精度，但是敏感度和 F 度量的值较低，模型预测嗜热蛋白的能力很弱。采用重复欠抽样法和向后递归特征消去法优选出 30 个最重要的特征后，获得的分类模型除特异性稍有降低外，其他评价指标均有所提高，模型预测嗜热蛋白的能力显著提高。比较以各特征子集构建的分类模型，发现仅以氨基酸组成为特征集构建的模型的总体预测准确率为 85.83%、敏感性为 89.16%、特异性为 73.33%、精度为 77.00%、F1 度量为 74.87%，均高于其他特征子集构建的模型；进一步分析发现嗜热酶中含有更多的谷氨酸（E）、异亮氨酸（I）和赖氨酸（K），而常温酶中含有更多的谷氨酰胺（Q）、丝氨酸（S）和苏氨酸（T）。因此，推测氨基酸组成是影响酶热稳定性的最重要因素，尤其是谷氨酸（E）、异亮氨酸（I）和赖氨酸（K）具有更重要的热稳定性作用。本研究方法对食品工业微生物酶的热稳定性改造提供了关键性的技术支持，将加快酶在工业生产中的应用。

参考文献

[1] 姚轶俊.生物酶在食品工业中的应用[J].中国粮油学报,2011,26(1):124-128
YAO Yi-jun. Application of enzymes in food industry [J]. Journal of the Chinese Cereals and Oils Association, 2011, 26(1): 124-128

[2] Yeoman C J, Han Y, Dodd D, et al. Thermostable enzymes as biocatalysts in the biofuel industry [J]. Advances in applied microbiology, 2010, 70: 1-55

[3] Hoppe C, Schomburg D. Prediction of protein thermostability with a direction-and distance-dependent knowledge‐based potential [J]. Protein Science, 2005, 14(10): 2682-2692

[4] Wang L, Li C. Optimal subset selection of primary sequence features using the genetic algorithm for thermophilic proteins identification [J]. Biotechnology letters, 2014, 36(10): 1963-1969

[5] Singer G A, Hickey D A. Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content [J]. Gene, 2003, 317: 39-47

[6] Zhang G, Fang B. Application of amino acid distribution along the sequence for discriminating mesophilic and thermophilic proteins [J]. Process Biochemistry, 2006, 41(8): 1792-1798

[7] Nakariyakul S, Liu Z, Chen L. Detecting thermophilic proteins through selecting amino acid and dipeptide composition features [J]. Amino acids, 2012, 42(5): 1947-1953

[8] Zhang G. A simple statistical method for discrimination of thermophilic and mesophilic proteins based on amino acid composition [J]. International journal of bioinformatics research and applications, 2013, 9(1): 41-52

[9] Breiman L. Random forests [J]. Machine learning, 2001, 45(1): 5-32

[10] 蒋诗泉,刘中侠,蒋诗平,等.随机森林算法在红葡萄酒质量评价指标体系选择中的应用[J].食品工业科技,2014,35(7):264-267
JIANG Shi-quan, LIU Zhong-xia, JIANG Shi-ping, et al. Application of random forest algorithm on selecting evaluation index system of the quality of red wine [J]. Science and Technology Food Industry, 2014, 35(7):264-267

[11] Llorén-Rico V, Lluch-Senar M, Serrano L. Distinguishing between productive and abortive promoters using a random forest classifier in *Mycoplasma pneumoniae* [J]. Nucleic Acids Research, 2015, 43(7): 3442-3453

[12] Ebrahimi M, Lakizadeh A, Agha-Golzadeh P, et al. Prediction of thermostability from amino acid attributes by combination of clustering with attribute weighting: a new vista in engineering enzymes [J]. PLoS One, 2011, 6(8): e23146

[13] Ebrahimi M, Ebrahimie E, Ebrahimi M. Searching for patterns of thermostability in proteins and defining the main features contributing to enzyme thermostability through screening, clustering, and decision tree algorithms [J]. EXCLI Journal, 2009, 8: 218-233

[14] Gromiha M M, Suresh M X. Discrimination of mesophilic and thermophilic proteins using machine learning algorithms [J]. Proteins: Structure, Function, and Bioinformatics, 2008, 70(4): 1274-1279