生物工程学报
journals.im.ac.cn
cjb@im.ac.cn

Chin J Biotech 2009, October 25; 25(10): 1508-1515
Chinese Journal of Biotechnology ISSN 1000-3061

系统生物技术

基于二级结构氨基酸组成识别酸性、中性及碱性酶

张光亚, 高嘉强, 方柏山

华侨大学生物工程与技术系, 厦门 361021

摘　要: 本研究系统分析了酸性、碱性和中性酶在二级结构氨基酸组成上的差异。结果发现在形成特定二级结构过程中, 酸性酶和碱性酶有着不同的氨基酸使用偏向; 同时, 在酸性和碱性酶中, 中性氨基酸和侧链微小的氨基酸含量明显较高, 这可能是它们适应极端 pH 的普遍机制。基于此, 提出了一种提取蛋白质序列特征值的新方法, 其 10 倍交叉验证的精度可达 80.3%。与其他常见特征值提取方法相比, 其精度提高了 9.4%到 18.7%不等; 而随机森林算法比其他机器学习算法识别精度也高出 2.7%到 21.8%不等。

关键词: 二级结构, 氨基酸组成, 酸性酶, 碱性酶, 稳定性机制, 特征提取, 随机森林

Discriminating acidic, neutral and alkaline enzymes with secondary structure amino acid composition

Guangya Zhang, Jiaqiang Gao, and Baishan Fang

Department of Bioengineering and Biotechnology, *Huaqiao University*, *Xiamen* 361021, *China*

Abstract: In this work, we systematically analyzed the secondary structure amino acid compositions of acidic and alkaline enzymes and compared them with neutral ones. We found that the propensity of the individual residues to participate in secondary structures and the consistently higher composition of neutral and tiny residues might be the general stability mechanisms for their adaptation to pH extremes. Based on this, we presented a secondary structure amino acid composition method for extracting useful features from sequence. The overall prediction accuracy evaluated by the 10-fold cross-validation reached 80.3%. Comparing our method with other feature extraction methods, the improvement of the overall prediction accuracy ranged from 9.4% to 18.7%. The random forests algorithm also outperformed other machine learning techniques with an improvement ranging from 2.7% to 21.8%.

Keywords: secondary structure, amino acid composition, acidic enzyme, alkaline enzyme, mechanism of stability, feature extraction, random forests

嗜酸和嗜碱菌分别指最适生长 pH 很低(通常在 pH 2~5)或很高(通常在 pH 9~11)的微生物[1-2]。过去 20 年中, 对嗜酸和嗜碱菌的生理学及分子遗传学进行了大量研究以了解其嗜酸和嗜碱机理[3], 同时, 对嗜酸和嗜碱菌的工业应用也取得了一些进展[4-5]。可见, 对其进行深入研究无论在学术上还是工业应用上都具有重要意义。然而, 嗜酸和嗜碱菌虽生长在强酸和强碱环境中, 但其细胞内部的 pH 却接近

Received: April 2, 2009; **Accepted:** July 30, 2009
Supported by: National Program on Key Basic Research Project (973 Program) (No. 2007CB707804), National Natural Science Foundation of China (No. 20806031).
Corresponding author: Baishan Fang. Tel: +86-595-22691095; E-mail: fangbs@hqu.edu.cn
国家重点基础研究发展规划(973 计划) (No. 2007CB707804), 国家自然科学基金(No. 20806031)资助。

中性, 细胞内酶反应和生化代谢过程也与中性菌相似, 只有它们产生的胞外酶因处在酸性或碱性环境而成为酸性酶和碱性酶[1,3]。在极端 pH 条件下, 多数酶变得不稳定而成为限制其应用的瓶颈, 故探讨酸性和碱性酶稳定性机制也一直是关注的焦点。有研究者分析了 *Nocardiopsis alba* 中酸性蛋白酶 A(NAPase)的结构及去折叠(Unfolding)动力学特性并与其中性同源蛋白进行了比较, 结果表明二者虽然包含的酸性氨基酸数量相近, 但 NAPase 对酸去折叠自由能势垒高度(Height of the unfolding free energy barrier)却没有其同源蛋白那么敏感[6]。Dubnovtsky AP 等[7]研究了来自于专性嗜碱微生物 *Bacillus alcalophilus* 的磷酸丝氨酸氨基转移酶(Phosphoserine aminotransferase, PSAT)的晶体结构并和来自于大肠杆菌的中性同源蛋白进行了比较, 发现碱性酶拥有一些独特的结构特点。此外, 酸性和碱性酶对极端 pH 条件的适应也体现在氨基酸水平上[8-10]。尽管如此, 到目前为止, 鲜有对酸性、中性和碱性酶进行理论预测的报道, 而这在嗜热酶中却比较常见[11-12]。

究其原因主要是获取和收集嗜酸和嗜碱酶的序列和结构信息比较困难。如前所述, 虽然一些嗜酸菌(如 *Ferroplasma acidarmanus*)和嗜碱菌(如 *Bacillus halodurans*)基因组测序已经完成[13], 但其蛋白质组信息并不能直接用于分析嗜酸和嗜碱酶稳定性机制; 而且, 一些非嗜酸和嗜碱菌也能产生酸性和碱性酶[14-15]。由于大多研究仅探讨某一种酸性或碱性酶, 这往往会得到一些互相矛盾的结果[7]。因此, 迫切需要进行较多样本的酸性和碱性酶研究和探讨, 以期获得一些酶蛋白适应极端 pH 环境的普遍机制。在此过程中, 要将其和中性酶进行一级结构和高级结构某些特征的比较。

本研究获得了序列一致性(Identity)小于 25%的酸性、中性和碱性酶序列共计 523 条, 通过预测其二级结构, 系统分析了三者在不同二级结构中氨基酸组成的差异, 深入探讨了酸碱稳定性机制, 在此基础上提出了一种提取蛋白质序列特征值的新方法, 用随机森林算法对 3 种酶进行了识别, 效果优于其他序列特征值提取方法。

1 材料和方法

1.1 数据来源

数据样本按照以下步骤构建: 1)从 BRENDA 数据库中获取所有最适 pH 小于 5 的酸性酶、最适 pH 在 6.5~7.5 的中性酶以及最适 pH 大于 9 的碱性酶信息(最适 pH 均为实验所得数据), 其网址为: http://www.brenda-enzymes.info/[16], 酶蛋白序列也来源于该数据库, 它们源于 UniProt/Swiss-Prot。2)序列长度小于 50 个氨基酸的酶蛋白序列被剔除, 因为它们可能是部分长度或者片段。3)剔除了包含连续 3 个或以上未知氨基酸的序列(例如: “XXX”, “XXXX”等)。4)为了避免同源序列带来的偏差, 使用 Blastclust 程序[17]剔除了序列一致性(Identity)大于 25%的序列。最后共得到 105 条酸性酶, 307 条中性酶和 111 条碱性酶序列。上述 523 条序列 ID 号、FASTA 格式的序列以及蛋白质长度等信息保存在一个基于 Microsoft Access 的数据库中。

1.2 二级结构氨基酸组成(ssAAC)

上述 523 条序列二级结构的预测由 Predator 程序完成[18], 与一般二级结构预测工具不同的是它可以对大批量序列进行预测, 而且对序列很长的蛋白也可, 目前得到较广泛的使用[19-20]。3 种二级结构类型: α-螺旋、β-折叠和无规则卷曲中氨基酸含量定义如下:

$$Comp(i,j)=\sum\frac{n_{i,j}}{N_j} \qquad (1)$$

式中, i 表示 20 种氨基酸, j 表示α-螺旋、β -折叠和无规则卷曲, $n_{i,j}$ 表示 i 在 j 中的数量, N_j 表示 j 中 20 种氨基酸的总数量。上述所有计算由自行用 C^{++}开发的软件完成, 经测试无误, 可向作者免费索取用于学术目的。

1.3 随机森林(Random forests, RF)算法

随机森林[21]是一种新组合分类方法, 其基本思路是通过在单棵决策树中引入随机性, 以减少其输出结果之间的相关程度, 即减少单棵决策树由于训练方法、训练数据等因素所引起的某种偏置, 使得犯同样错误的概率减少, 最终的判决性能提高。其性能取决于 2 个方面: 1)组成随机森林的单棵决策树

的性能越好, 所得到的随机森林的性能也越好。2) 组成随机森林的单棵决策树之间的相关性越小, 相似性越低, 则随机森林的性能越好。随机森林算法目前在生物信息学领域得到广泛应用[22-23], 具体运算过程见文献[21]。

1.4 有效性检验

在评估模型优劣过程中, 经常采用 3 种方法: 独立样本测试、交叉验证和 jackknife 测试。其中 jackknife 测试最为客观[24-25]。而在实际操作过程中, 该法运算速度较慢而且消耗计算机资源庞大, 因此, 交叉验证被越来越多的研究者采用[26-28], 它实际上是 jackknife 测试的一个特例。本研究采用 10 倍交叉验证(10-fold cross-validation, 10-CV), 具体做法是: 将训练数据随机分为 10 组, 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。

1.5 识别效果评估

模型最终表现通过以下 2 个参数进行描述, 预测正确率(Success rate)和受试者操作特性曲线下面积(Area under the receiver operation characteristic curve, AUC)。一般而言, 分类器的 AUC 大于 0.9, 则被认为优秀。

本研究中实现随机森林算法及其他机器学习算法的软件均来自于 Weka(Waikato environment for knowledge analysis), 它是基于 JAVA 虚拟机开发的[29]。使用的 PC 为 DELL precision^{TM}490 工作站。

2 结果与分析

2.1 酸性和碱性酶与中性酶氨基酸组成差异

与中性酶相比, 酸性和碱性酶氨基酸组成存在较明显差异。表 1 列出了酸性酶和中性酶有较大差异($\left|Comp_{AC,i}-Comp_{NE,i}\right|>1$)的氨基酸。可见, 酸性酶中存在较多的 Gly、Ser 和 Thr, 较少的 Glu、Lys、Leu 和 Arg。Thr 和 Ser 由于含有羟基, 非常容易和水分子发生相互作用, 形成氢键, 这可能有利于酸性酶分子在低 pH 条件下维持其结构的稳定性[30]; 而 Lys 和 Arg 属于碱性氨基酸, 在酸性条件下会带较多的正电荷, 而酶分子在酸性条件下由于羧基被中和, 分子中正电荷会大量积累, 而较多的 Lys 和 Arg 会进一步加剧这种状况, 对其结构产生不良影响, 使许多由羧基介导的“盐桥”断裂而变得不稳定[31], 故其含量较少。而 Glu 为酸性氨基酸, 有研究认为一些酸性酶通过减少分子中羧基的数量以降低 pH 依赖型效应的量级(The magnitude of the pH-dependent effect), 从而维系分子稳定性[32]。

而在不同二级结构中, 两种酶对氨基酸的使用依然差异明显。在α-螺旋中, 酸性酶 Leu、Ala、Thr、Gly、Ser 和 Val 出现频率较高, 而中性酶 Lys、Glu、Gln、Asp 和 Arg 则较高。众所周知, Leu、Ala 和 Glu 是强烈的α-螺旋形成子, 而 Val 和 Gln 是α-螺旋螺旋形成子[33], 可见, 酸性和中性酶在形成α-螺旋过程中偏向于使用不同的氨基酸; 在β-折叠中, 酸性酶 Thr、 Gln、Trp、Ala、Lys 和 Ser 含量较高, 而中性酶 Ile、 Val、Leu、Cys 和 Gly 较高, 同样, Thr、Gln、Trp、Cys 和 Leu 是的 β-折叠形成子[33], 它们分别分布于酸性和中性酶中, 可见, 在形成β-折叠过程中, 依然存在不同氨基酸使用偏向; 而在无规则卷曲中, 酸性酶与中性酶的差异不如在α-螺旋和β-折叠中那么大, 酸性酶中 Ser、Thr 和 Gly 较多, 而 Lys、Leu 和 Glu 较少。研究表明[34] Gly、Asn、Asp、Pro、His 和 Ser 具有较强的无规则卷曲形成趋向。而酸性酶中 Gly 和 Ser 在无规则卷曲中含量高于中性酶, 可见在无规则卷曲中酸性酶和中性酶在氨基酸使用上依然存在差异。

表 1 酸性酶中特征性氨基酸
Table 1 Significant amino acids in acidic enzymes

	$Comp_{AC}-Comp_{NE}>1$/number	$Comp_{AC}-Comp_{NE}<-1$/number
Overall	Gly(1.4), Ser (2.14), Thr(1.82)/3	Glu(−2.22), Lys(−1.5), Leu(−1.21), Arg(−1.63)/4
Helix	Ala(3.81), Gly(2.62), Leu(5.87), Ser(1.1), Thr(3.72), Val(1.02)/6	Asp(−1.77), Glu(−3.6), Lys(−4.52), Gln(−2.08), Arg(−1.65)/5
Sheet	Ala(1.79), Lys(1.35), Gln(2.61), Ser(1.12), Thr(3.61), Trp(2.05)/6	Cys(−1.71), Gly(−1.7), Ile(−3.28), Leu(−2.07), Val(−2.18)/5
Coil	Gly(2.02), Ser(3.36), Thr(2.34)/3	Glu(−1.77), Lys(−2.86), Leu (−2.38)/3

表 2 列出了碱性酶和中性酶有较大差异($\left|Comp_{AK,i}-Comp_{Ne,i}\right|>1$)的氨基酸。总体而言, 其差异不如酸性酶明显。碱性酶中 Ala 整体含量较高。Ala 的侧链只有一个甲基, 其空间位阻较小, 为非极性氨基酸, 可以增加分子中疏水相互作用, 从而对维系酶分子高级结构具有积极作用。在不同二级结构中, 两种酶对氨基酸的使用依然存在一些差异。在α-螺旋中, 碱性酶 Ala 出现频率较高, 而中性酶 Lys 则较高。如上所述, Ala 是强烈的α-螺旋形成子, 而 Lys 则不易形成α-螺旋。可见, 碱性和中性酶在形成α-螺旋过程中也存在偏向; β-折叠中, 碱性酶 Ser, Ala、Thr、Asp 和 Glu 含量较高, 而中性酶 Val 和 Phe 较高, 同样, Thr 和 Val 比较容易形成 β-折叠[34], 它们分别分布于碱性和中性酶中。这说明, 在形成 β-折叠过程中, 依然存在这种氨基酸使用的偏向; 而无规则卷曲中, 碱性酶中 Gly 和 Pro 较多。这两种氨基酸具有较强的无规则卷曲形成趋向, 故而, 在无规则卷曲形成过程中两种酶依然存在差异。综上可知, 酸性酶和碱性酶在极端 pH 条件下可能采用不同的氨基酸形成特定的二级结构, 这很可能是一种它们适应极端 pH 的普遍机制。

为了进一步了解其差异, 参考相关文献[35]把氨基酸分成若干类型, 包括: 带电的、脂肪族、芳香族、极性的、中性的、疏水性的、带正电、带负电、微小的、小的、大的、含硫的以及酰胺。比较了它们在酸性(碱性)酶与中性酶中的差异, 见图 1。由图 1A 可知, 酸性酶和中性酶各类型氨基酸组成的差异在α-螺旋中最为明显。酸性酶中性氨基酸(AGHPSTY)和微小的氨基酸(ACDGST)明显较高, 尤其在 α-螺旋和无规则卷曲中, 而带电的氨基酸(DEKHR)和极性氨基酸(DERKQN)则明显较低, 尤其在 α-螺旋中差异更为明显, 而较小的氨基酸(EHILKMNPQV)在无

表 2 碱性酶中特征性氨基酸
Table 2 Significant amino acids in alkaline enzymes

	$Comp_{AK}-Comp_{NE}>1$/number	$Comp_{AK}-Comp_{NE}<-1$/number
Overall	Ala(1.26)/1	0
Helix	Ala(1.89)/1	Lys(−1.18)/1
Sheet	Ala(1.28), Asp(1.14), Glu(1.01), Ser(1.39), Thr(1.16)/5	Phe(−1.09), Val(−2.34)/2
Coil	Gly(2.05), Pro(1.55)/2	0

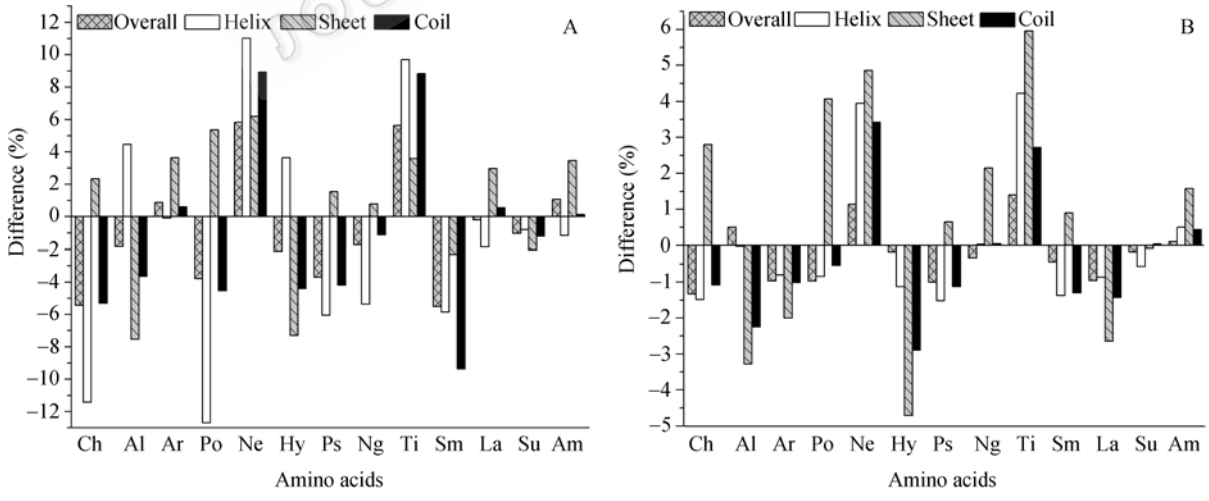

图 1 不同类型氨基酸组成的差异
Fig. 1 Compositional differences of different kinds of amino acids. (A) Acidic subtract neutral enzymes. (B) Alkaline subtract neutral enzymes. The upper half shows the dominance of residues in acidic (or alkaline) enzymes. Ch: charged (DEKHR); Al: aliphatic (ILV); Ar: aromatic (FHWY); Po: polar (DERKQN); Ne: neutral (AGHPSTY); Hy: hydrophobic (CVLIMFW); Ps: positively charged (HKR); Ng: negatively charged (DE); Ti: tiny (ACDGST); Sm: small (EHILKMNPQV); La: large (FRWY); Su: sulfur (CM); Am: amide (NQ) residue.

规则卷曲中更少。由图 1B 可知，碱性酶与中性酶各类型氨基酸组成的差异不如酸性酶大，而且差异比较明显的主要在 β-折叠中。类似的是，碱性酶中性氨基酸(AGHPSTY)和微小的氨基酸(ACDGST)明显较高，尤其在α-螺旋和 β-折叠中，而疏水性氨基酸(CVLIMFW)和脂肪族氨基酸(ILV)则明显较低，而在 β-折叠和无规则卷曲差异尤为明显。由以上分析可以看出，相对于中性酶而言，酸性和碱性酶均明显含有更多的中性(AGHPSTY)和微小的氨基酸(ACDGST)。研究表明，中性氨基酸在极端 pH 条件下很容易形成疏水相互作用，而侧链更小的氨基酸在蛋白质内核组装(Core packing)过程中更容易占据蛋白质分子中不同空间(Voids)，这对维持酶蛋白分子的稳定性非常重要[36]。而这也似乎是酸性和碱性酶适应极端 pH 的一个普遍机制。

2.2 随机森林算法与其他机器学习算法的比较

随机森林算法和其他机器学习算法识别效果列于表 3。此过程中，以二级结构氨基酸组成作为特征值提取方法，所包含的变量数目为 20×3=60 个。经 10 倍交叉验证后，该算法成功识别出 76 条酸性酶、55 条碱性酶和 289 条中性酶，正确率分别为 72.4%、49.5%和 94.1%，平均正确率为 80.3%，识别效果较理想，其 AUC 值为 0.919。可见，此法对碱性酶识别效果不好。尽管如此，随机森林识别效果依然是所有供试算法中最好的。它比基于 Desion stump 的 Adaboost 方法高出 21.8%，比其他算法也高出 2.7%到 15.5%不等。而相比于已成为生物信息学标准算法的支持向量机(SVM)而言，其识别精度仍有 3.4%到 6.5%的提升。可见，在本识别过程中，随机森林算法表现最好，而且其运算速度较快，对计算机资源的消耗较少。

2.3 ssAAC 与其他特征值提取方法的比较

为了检验二级结构氨基酸组成这种特征值提取方法的有效性，与其他蛋白质序列特征值提取方法进行了比较，这些方法包括氨基酸组成、二肽组成、伪氨基酸组成[37]、不同标度的伪氨基酸组成[38]、两性伪氨基酸组成[39](在 http://www.csbio.sjtu.edu.cn/bioinf/PseAA/完成运算)、标准化的 Moreau-Broto 自相关指数 [40]、Moran 自相关指数 [41]、Geary 自相关指数[42]、组成、转变及分布(CTD)[43]以及序列顺序耦合数和部分序列顺序(SQ)[44]，后 5 种提取方法由 PROFEAT(Protein Feature Server)服务器计算完成[45]，网址: http://jing.cz3.nus.edu.sg/cgi-bin/prof/prof.cgi。经 10 倍交叉验证，结果见表 4。

表 3 各种机器学习算法的比较

Table 3 Comparisons of machine learning techniques

Machine learning techniques		ACC	AUC
Bagging	Decision stump	64.8	0.751
	Decision table	69.2	0.795
	REP tree	72.5	0.858
	J4.8	75.0	0.867
Adaboost	Decision stump	58.5	0.747
	Decision table	69.2	0.795
	REP tree	70.6	0.869
	J4.8	75.5	0.879
Logitboost	Decision stump	74.8	0.885
	REP tree	77.6	0.881
Decision tree J4.8		67.3	0.719
NBTree		67.9	0.786
k-nearest neighbour		69.6	0.723
naïve Bayes		70.2	0.876
Bayes net		70.7	0.866
SVM (Puk)		73.8	0.74
BPNN		74.2	0.874
SVM (linear kernel)		75.0	0.784
SVM (polynomial kernel E=2)		76.9	0.832
Random forest		80.3	0.916

Puk: the Pearson VII function-based universal kernel; BPNN: back propagation neural network; ACC: accuracy; AUC: area under ROC curve.

由表 4 可知，以随机森林算法为例，上述 15 种特征值提取方法，ssAAC 效果最好，其识别精度比 Geary 自相关指数高出 18.7%，比其他特征值提取方法识别精度也高出 9.4%到 17.6%不等，其 AUC 值也比其他方法高出 0.088 到 0.26 不等。为了进一步证实 ssAAC 有效性，同样比较了在另外两种分类效果较好的算法(Logitboost 和 SVM)中各种特征值提取方法的表现(见表 4)。从中可以看出：在这两种算法中，ssAAC 依然表现最好，分别比 Geary 自相关指数高出 21.0%和 32.7%，比其他特征提取方法也至少高出 10.6%和 8.3%；除二肽组成在 SVM 中识别精度有 4.2%的提高外，其他特征值提取方法在 Logitboost 和 SVM 中表现均不如随机森林。

表 4 各种特征值提取方法的比较

Table 4 Comparisons of feature extraction methods

Feature extraction methods	RF		Logitboost		SVM		NF
	ACC	AUC	ACC	AUC	ACC	AUC	
Secondary structure amino acid composition	80.3	0.919	77.6	0.881	76.9	0.832	60
Composition, transition and distribution	70.9	0.813	65.8	0.749	60.8	0.699	147
SQ	70.9	0.799	62.5	0.683	61.8	0.679	240
Pseudo-amino acid composition (*Z*-scales)	70.7	0.789	64.6	0.722	65.0	0.726	60
Pseudo-amino acid composition (*K*-scales)	70.1	0.782	63.4	0.679	68.4	0.733	60
Amphiphilic pseudo-amino acid composition	69.8	0.783	62.9	0.711	67.1	0.746	60
Pseudo-amino acid composition (*F*-scales)	69.7	0.770	67.0	0.684	66.3	0.732	60
Amino acid composition distribution	69.7	0.768	65.1	0.713	67.8	0.726	60
Pseudo-amino acid composition (*T*-scales)	69.5	0.767	63.2	0.660	66.1	0.712	60
Pseudo-amino acid composition (Chou)	68.5	0.768	62.3	0.666	65.8	0.726	60
Amino acid composition	68.1	0.764	63.5	0.678	68.6	0.686	20
Normalized moreau-broto autocorrelation	65.2	0.715	57.6	0.606	52.0	0.613	240
Moran autocorrelation	62.9	0.678	58.7	0.613	45.5	0.566	240
Dipeptide composition	62.7	0.741	58.5	0.558	66.9	0.715	400
Geary autocorrelation	61.6	0.659	56.6	0.539	44.2	0.577	240

SQ: sequence-order-coupling number & Quasi-sequence-order descriptors; NF: number of features.

因上述 15 种特征值提取方法所产生的变量数目各异, 而ssAAC仅60个变量, 会产生这样的疑问: ssAAC 较高的识别精度是否是由于其较少的变量数目而引起的？对此, 可从以下几个方面进行解释:1)氨基酸组成只有 20 个变量, 少于 ssAAC, 但在 3 种机器学习算法中, 其识别精度分别比后者低 12.2%、14.1%和 8.3%。2)伪氨基酸组成(λ=40)、不同标度的伪氨基酸组成(λ=40)以及两性伪氨基酸组成(λ=20), 其变量数目与 ssAAC 等同, 但其识别效果至少分别比后者低 9.6%、10.6%和 8.5%。3)SQ 与 Geary 自相关指数变量数目相等, 而识别精度却相差 9.3%(以随机森林为例), 同时 SQ 也比标准化的 Moreau-Broto 自相关指数和 Moran 自相关指数识别精度高, 尽管它们变量数目也相等。4)二肽组成变量数目最多, 但在随机森林算法中并不是识别效果最差的, 而且在SVM中其识别精度优于不少其他特征值提取方法。5)研究[46]表明随机森林和 SVM 均对变量数目不敏感(Robust to large feature sets)。

3 讨论

相比于嗜热酶、嗜冷酶等极端酶, 对酸性酶和碱性酶的研究较少, 原因如前所述。尽管如此, 仍有不少研究者对此进行了有意义的探讨[1-10]。然而, 其中大部分仅比较了某一种酸性酶(或碱性酶)及其中性同源蛋白, 故导致部分研究结果互相矛盾。虽然每种酶可能都有其特殊适应极端 pH 的机制, 但或许也存在某些共同的适应机制。本研究的一个主要目标就是试图找出其中一些共有规律。从上面结果可以看出, 酸性酶和碱性酶在形成特定二级结构的过程中偏向于使用不同的氨基酸, 这种现象已在一些嗜盐蛋白中发现[19]。此外, 在其二级结构中, 中性氨基酸和侧链微小的氨基酸含量普遍较高, 或许也是酸性(碱性)酶适应极端 pH 的另一个机制, 这对酶的生物学改造与酶结构和功能强化的理性设计具有非常重要的理论意义。

本研究的另一个目标是提出一种提取蛋白质序列特征值的新方法。众所周知, 特征值提取对构建一个预测体系至关重要, 它需要将原始蛋白序列转换成合适的数据特征, 同时应尽可能减少信息损失(Information loss)。氨基酸和二肽组成是两种最常见的基于组成(Composition based)的特征值提取方法, 此外, Chou 提出的伪氨基酸组成[32]以及两性伪氨基酸组成[34]等方法也被广泛应用于蛋白质序列特征值提取。本研究提出了一种基于组成的蛋白质序列特

征值提取新方法并用于识别酸性、中性和碱性酶。从其预测结果可见，至少在本次识别过程中它优于目前一些常见的特征值提取方法。可以预见，ssAAC 将会成为一个特征提取新方法，以用于一些其他生物信息学问题，例如：蛋白质亚细胞定位、蛋白质结构类型、蛋白质-蛋白质相互作用和膜蛋白类型等，这或许对生物信息学方法的发展也有一定推动作用。

此外，本方法识别效果较好，达到 80.3%，虽然未能达到预期的 90%以上。但相比对这 3 种类型酶进行随机猜测的几率(1/3=33.3%)而言，其效果已经有了明显提高。这说明，酶对不同 pH 的适应机制与其二级结构氨基酸组成密切相关。此外，研究过程中也发现，此法对碱性酶识别效果较差，而且从其预测结果来看，错误主要集中在把碱性酶预测为中性酶(其中 53 个碱性酶被预测成中性酶，占总数量的 47.7%)，这可能与中性酶二级结构氨基酸组成与中性酶差别较小有关。因此，如何提高碱性酶识别效果将是后续研究重点。一旦建立了较高精度的识别方法，就可以建立一个完全基于序列对酸性、中性和碱性酶进行预测的服务系统，让研究者在获得酶序列信息(甚至其编码的基因序列)之后通过提取其序列特征值可以用此方法判断其酸碱性。

REFERENCES

[1] Zhang HX, Hao CB, Bai ZH. Advance in acidophile. *J Microbiol*, 2006, **26**(2): 68–72.
张洪勋，郝春博，白志辉. 嗜酸菌研究进展. 微生物学杂志, 2006, **26**(2): 68–72.

[2] Ma YH. Alkaliphiles. *Microbiol*, 1999, **26**(4): 309.
马延和. 嗜碱微生物. 微生物通报, 1999, **26**(4): 309.

[3] Takami H, Horikoshi K. Analysis of the genome of an *alkaliphilic Bacillus* strain from an industial point of view. *Extremophiles*, 2000, **4**: 99–108.

[4] Horikoshi K. Alkaliphiles: some applications of their products for biotechnology. *Microbiol Mol Biol Rev*, 1999, **63**: 735–750.

[5] Baker-Austin C, Dopson M. Life in acid: pH homeostasis in acidophiles. *Trends Microbiol*, 2007, **15**: 165–171.

[6] Kelch BA, Eagen KP, Erciyas FP, *et al*. Structural and mechanistic exploration of acid resistance: kinetic stability facilitates evolution of extremophilic behavior. *J Mol Biol*, 2007, **368**, 870–883.

[7] Dubnovitsky AP, Kapetaniou EG, Papageorgiou AC. Enzyme adaptation to alkaline pH: atomic resolution (1.08 Å) structure of phosphoserine aminotransferase from *Bacillus alcalophilus*. *Protein Sci*, 2005, **14**: 97–110.

[8] Geierstanger B, Jamin M, Volkman BF, *et al*. Protonation behavior of histidine 24 and histidine 119 in forming the pH 4 folding intermediate of apomyoglobin. *Biochemistry*, 1998, **37**: 4254–4265.

[9] Settembre EC, Chittuluru JR, Mill CP, *et al*. Acidophilic adaptations in the structure of *Acetobacter aceti* N5-carboxyaminoimidazole ribonucleotide mutase (PurE). *Acta Crystallogr Sect D Biol Crystallogr*, 2004, **60**: 1753–1760.

[10] Shirai T, Suzuki A, Yamane T, *et al*. High resolution crystal structure of M-protease: phylogeny aided analysis of the high-alkaline adaptation mechanism. *Protein Eng*, 1999, **10**: 627–634.

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

[12] Gromiha MM, Suresh MX. Discrimination of mesophilic and thermophilic proteins using machine learning algorithms. *Proteins*, 2008, **70**: 1274–1279.

[13] Schneider KL, Pollard KS, Baertsch R, *et al*. The UCSC archaeal genome browser. *Nucleic Acids Res*, 2006, **34**: D407–D410.

[14] Zhu J, Wu MC. Studies on the characterization of acidic β-mannanase from *Aspergillus niger*. *J Food Sci Biotechnol*, 2007, **26**(2): 21–25.
朱劼，邬敏辰. 黑曲霉酸性 β-甘露聚糖酶的酶学特性. 食品与生物技术学报, 2007, **26**(2): 21–25.

[15] Chen XW, Shi ZY. cDNA sequence analysis and tertiary structure prediction of alkaline phosphatase from *Paralichthys olivaceus*. *Chin J Biochem Mol Biol*, 2007, **23**(6): 442–449.
陈晓武，施志仪. 牙鲆碱性磷酸酶 cDNA 序列分析与蛋白质高级结构预测. 中国生物化学与分子生物学报, 2007, **23**(6): 442–449.

[16] Barthelmes J, Ebeling C, Chang A, *et al*. BRENDA, AMENDA and FRENDA: the enzyme information system in 2007. *Nucleic Acids Res*, 2007, **12**(35): D511–D514.

[17] 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.

[18] Frishman D, Argos P. Seventy-five percent accuracy in protein secondary structure prediction. *Proteins*, 1997, **27**, 329–335.

[19] Paul S, Bag` SK, Das S, *et al*. Molecular signature of hypersaline adaptation: insights from genome and proteome composition of halophilic prokaryotes. *Genome Biol*, 2008, **9**: R70.

[20] Wang L, Brown SJ. Prediction of DNA-binding residues from sequence features. *J Bioinform Comput Biol*, 2006, **4**: 1141–1158.

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

[22] Qi Y, Joseph ZB, Judith KS. Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins*, 2006, **63**: 490–500.
[23] Wu JS, Liu HD, Duan XY *et al.* Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature. *Bioinformatics*, 2009, **25**: 30–35.
[24] Chou KC, Shen HB. Cell-PLoc: a package of web-servers for predicting subcellular localization of proteins in various organisms. *Nat Prot*, 2008, **3**: 153–162.
[25] Chou KC, Shen HB. Recent progresses in protein subcellular location prediction. *Anal Biochem*, 2007, **370**: 1–16.
[26] Wang T, Yang J, Shen HB, *et al*. Predicting membrane protein types by the LLDA algorithm. *Protein Peptide Lett*, 2008, **15**: 915–921.
[27] Li FM, Li QZ. Predicting protein subcellular location using Chou's pseudo amino acid composition and improved hybrid approach. *Protein Peptide Lett*, 2008, **15**: 612–616.
[28] Lin H. The modified mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. *J Theor Biol*, 2008, **252**: 350–356.
[29] Inamdar NM, Ehrlich KC, Ehrlich M, *et al*. Data mining in bioinformatics using Weka. *Bioinformatics*, 2004, **20**: 2479–2481.
[30] Kumar S, Tsai CJ, Nussinov R. Factors enhancing protein thermostability. *Protein Eng*, 2000, **13**: 179–191.
[31] Ionescu RM, Eftink MR. Global analysis of the acid-induced and urea-induced unfolding of staphylococcal nuclease and two of its variants. *Biochemistry*, 1997, **36**: 1129–1140.
[32] Schafer K, Magnusson U, Scheffel F, *et al.* X-ray structures of the maltose-maltodextrin-binding protein of the thermoacidophilic bacterium *Alicyclobacillus acidocaldarius* provide insight into acid stability of proteins. *J Mol Biol*, 2004, **335**: 261–374.
[33] Chou PY, Fasman GD. Conformational parameters for amino acids in helical, beta-sheet and random coil regions calculated from proteins. *Biochem*, 1974, **13**: 211–222.
[34] Costantini S, Colonna G, Facchiano AM. Amino acid propensities for secondary structures are influenced by the protein structural class. *Biochem Biophys Res Commun*, 2006, **342**: 441–451.
[35] Yu X, Cao J, Cai Y, *et al*. Predicting rRNA-, RNA-, and DNA-binding proteins from primary structure with support vector machines. *J Theor Biol*, 2006, **240**: 175–184.
[36] Britton KL, Baker PJ, Borges KMM, *et al.* Insights into thermal stability from a comparison of the glutamate dehydrogenases from *Pyrococcus furiosus* and *Thermococcus litoralis*. *Eur J Biochem*, 1995, **229**: 688–695.
[37] Chou KC. Prediction of protein cellular attributes using pseudo-amino-acid composition. *Proteins*, 2001, **43**: 246–255.
[38] Zhang GY, Li HC, Fang BS. Predicting the cofactors of oxidoreductases by the modified pseudo-amino acid composition. *Chin J Biotech*, 2008, **24**(8): 1439–1445.
张光亚, 李红春, 方柏山. 基于修正的伪氨基酸组成预测氧化还原酶辅酶类型的研究. 生物工程学报, 2008, **24**(8): 1439–1445.
[39] Chou KC. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics*, 2005, **21**: 10–19.
[40] Feng ZP. Zhang CT. Prediction of membrane protein types based on the hydrophobic index of amino acids. *J Protein Chem*, 2000, **19**: 269–275.
[41] Horne DS. Prediction of protein helix content from an autocorrelation analysis of sequence hydrophobicities. *Biopolymers*, 1988, **27**: 451–477.
[42] Sokal RR, Thomson BA. Population structure inferred by local spatial autocorrelation: an example from an Amerindian tribal population. *Am J Phys Anthropol*, 2006, **129**: 121–131.
[43] Cui J, Han LY, Lin HH, *et al.* Prediction of MHC-binding peptides of flexible lengths from sequence-derived structural and physicochemical properties. *Mol Immunol*, 2007, **44**: 866–877.
[44] Chou KC. Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. *Biochem Biophy Res Comm*, 2000, **278**: 477–483.
[45] Li ZR, Lin HH, Han LY, *et al*. PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res*, 2006, **34**: 32–37.
[46] Wu B, Abbott T, Fishman D, *et al.* Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*, 2003, **19**: 1636–1642.