文章编号：1000-5013(2014)02-0169-06 **doi**：10.11830/ISSN.1000-5013.2014.02.0169

不同嗜盐机制微生物蛋白质组特性及其识别

葛慧华，黄可君，张光亚

（华侨大学 化工学院，福建 厦门 361021）

摘要： 选取两种不同嗜盐机制微生物蛋白质组，并将其同非嗜盐微生物蛋白质组进行比较．研究结果发现：积累无机盐的蛋白质氨基酸组成与非嗜盐相差明显，而积累细胞相容性物质的嗜盐蛋白则差异较小，后者分子中 His 和分子量较小的氨基酸均显著多于非嗜盐蛋白，而 Ala 则相反；两种类型蛋白质中酸性氨基酸和碱性氨基酸的差值均显著高于非嗜盐蛋白．基于此，使用一种新型 Person 通用核函数的支持向量机对 3 种类型蛋白进行识别，其精度可达 84.1%，优于其他核函数的支持向量机及其他机器学习算法．

关键词： 嗜盐微生物；非嗜盐微生物；蛋白质组；氨基酸；支持向量机；识别

中图分类号： Q 554.903；Q 811.212 **文献标志码**： A

来源于嗜盐菌[1]的嗜盐酶（尤其是水解酶）能适应高浓度盐，也能忍受很宽的 pH 值，在医药、食品、纺织及化工产业等领域有广泛应用[2]．嗜盐微生物的嗜盐机理一直是研究者关注的焦点，目前已发现微生物主要通过两种方式对抗胞外的高盐环境[3-5]：1）在细胞内积累高浓度的 K^+ 以对抗高渗环境（salt-in）；2）在细胞内积累细胞相容性溶质来抗衡外界盐环境的负面影响（salt-out）．近年来，随着嗜盐微生物基因组和蛋白质组计划的完成，对第一种嗜盐机制微生物蛋白质组的分析日益增加[6]，而对后一种适应机制的微生物蛋白质组的研究则很少．由于细胞相容性物质都有带电荷的基团（如氨基酸来源的相容性物质），因此蛋白质为了在高浓度的细胞相容性物质中保持溶解性和稳定性，必须减少其分子表面非极性的面积[7-8]．这种微生物中蛋白质与其同源蛋白同样存在与嗜盐机制有关的差异，不过其差异程度相对较小[7]．生物信息学和比较蛋白质学的发展为研究嗜盐菌的机理提供了新思路，所得结果对设计新的嗜盐蛋白具有积极指导价值[9-10]．然而，多数方法并未对两种不同嗜盐机制的蛋白进行区分．本文选取了两株不同嗜盐机制的微生物及一株非嗜盐微生物的全蛋白质组序列，探讨了不同嗜盐蛋白稳定性机制，并使用一种新型核函数的支持向量机方法对 3 种蛋白进行了识别．

1 材料和方法

1.1 菌株选取及序列数据构建

根据以下 3 点规则从数据库选取微生物：1）必须是嗜盐微生物，且基因组注释已经完成，可提供大量蛋白质序列；2）所选取微生物最适生长温度接近，减少了温度对其氨基酸使用偏好的影响；3）微生物基因组的 G+C 摩尔分数非常接近，最大程度减少了 GC 摩尔分数对氨基酸使用偏好的影响．所选取的嗜盐菌分别为 *Halobacterium* sp. NRC-1 和 *Halomonas elongata*，前一个为细胞内积累 KCl（salt-in）[11]，后一个为积累细胞相容性物质（salt-out）[12]，而非嗜盐菌为 *Caulobacter crescentus* CB15[13]．这样，它们在氨基酸使用上的差异就主要是由嗜盐机制不同造成的．

使用 Blastclust 程序[14]共得到 1 701，2 382 及 2 703 条序列，其所占比例分别为 25.1%，35.1% 和 39.8%．上述 6 786 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．

收稿日期： 2013-03-28

通信作者： 葛慧华（1979-），女，实验师，主要从事酶工程和分子动力学模拟的研究．E-mail：zhgyghh@hqu.edu.cn.

基金项目： 福建省高校新世纪优秀人才支持计划项目（07176C02）

1.2 氨基酸组成差异

考虑到蛋白质序列中20种氨基酸出现的频率存在较大差异,因此,在比较不同氨基酸组成差异的时候需要考虑这个因素,以使结果更能反映真实差异[15]. 为此,统计了Uniprot数据库中所有蛋白序列氨基酸组成,并计算3种微生物中各蛋白质氨基酸组成. 两种嗜盐蛋白 *Halobacterium* sp. NRC-1 和 *Halomonas elongata* 分别表示为HIP和HOP,其与非嗜盐蛋白(表示为NP)氨基酸组成的差异为

$$D_{j,\text{I-N}} = N_{j,\text{I}} - N_{j,\text{N}} = \frac{C_{j,\text{I}} - C_{j,\text{N}}}{C_{j,\text{av}}}, \tag{1}$$

$$D_{j,\text{O-N}} = N_{j,\text{O}} - N_{j,\text{N}} = \frac{C_{j,\text{O}} - C_{j,\text{N}}}{C_{j,\text{av}}}. \tag{2}$$

式(1)~(2)中:$N_{j,\text{I}}$, $N_{j,\text{O}}$和$N_{j,\text{N}}$分别表示积累盐离子(salt-in)、细胞相容性物质(salt-out)及非嗜盐蛋白标准化的氨基酸组成; j 表示20种氨基酸;$C_{j,\text{I}}$,$C_{j,\text{O}}$和$C_{j,\text{N}}$分别表示这3种蛋白中氨基酸组成; $C_{j,\text{av}}$表示Uniprot数据库所有序列氨基酸组成平均值. 经计算,3种蛋白质组累计统计的氨基酸数量分别为525 159,805 826和896 556.

1.3 有效性检验

在评估模型优劣过程中,经常采用独立样本测试、交叉验证和Jackknife测试3种方法[16-17]. 在实际操作过程中,该法运算速度较慢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[18-20],而它实际上是Jackknife测试的一个特例. 文中采用10倍交叉验证(10-CV).

1.4 识别效果评估

模型最终表现通过以下2个参数进行描述,预测准确率(γ)和受试者操作特性曲线下面积(A). 一般而言,分类器的A值大于0.9,则被认为优秀. 文中实现所有算法的软件均来自于怀卡托智能分析环境(Weka 3-6-8)[21],使用DELL precision™ 490工作站,所有运行参数均采用默认值.

2 结果与分析

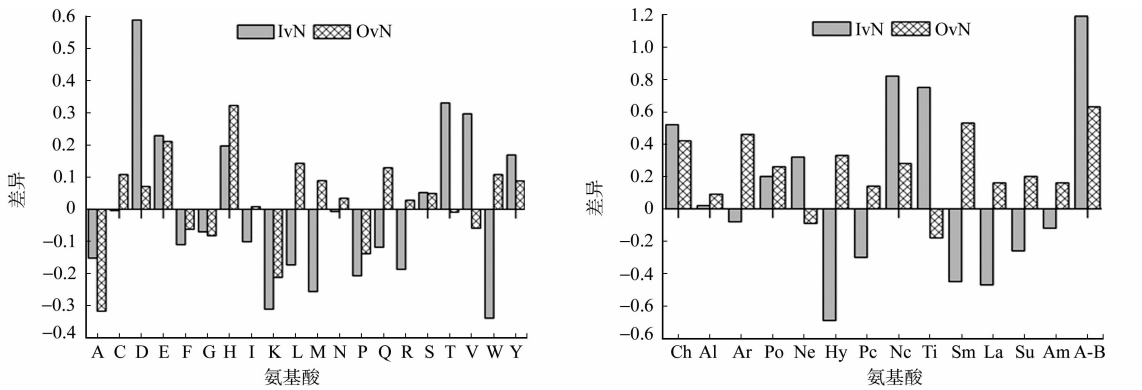
2.1 两种嗜盐蛋白与非嗜盐蛋白氨基酸组成差异

经与非嗜盐蛋白比较,两种嗜盐与非嗜盐蛋白氨基酸组成存在较明显差,如图1(a)所示. 为此,特定义在HIP或HOP中,若$|D_{j,\text{I-N}}|>0.25$或$|D_{j,\text{O-N}}|>0.25$,则氨基酸j视为显著性氨基酸. 可见,在HIP中存在较多的Asp,Thr和Val,较少的Lys,Trp和Met;而HOP中这种显著性氨基酸则明显较少,只有较多的His和较少的Ala. Asp作为一种酸性氨基酸在嗜盐蛋白(salt-in)中大量存在,这已得到广泛证实. Asp主要存在于蛋白分子表面,与阳离子(如K^+)相互作用,从而增加嗜盐蛋白的稳定性. 此外,其分子中碱性氨基酸(如Lys)的含量则显著减少[22]. Thr由于侧链带有羟基,非常容易和环境中水分子形成氢键,有助于蛋白在高盐浓度中保持可溶性及结构和功能. Val具有一定疏水性,且分子较小,有利于保持嗜盐蛋白更紧凑的疏水核心,从而增加其稳定性[23]. Met是一种疏水性较强的氨基酸,而研究表明嗜盐蛋白稳定性与其较低的疏水性有关[24],因此,Met在嗜盐蛋白中含量较低. Trp属于芳香族且疏水性较强,研究表明芳香族氨基酸在嗜盐蛋白中含量很少[25].

对HOP(通过细胞相容性物质稳定的蛋白)而言,其与非嗜盐蛋白的氨基酸组成虽然也存在差异,但相比于HIP则明显较少. 其中His较多而Ala则较少. Costantini等[26]认为,Ala非常容易形成α-螺旋,而His则具有很强的无规则卷曲形成趋向. 众所周知,α-螺旋是一种较刚性的结构,而无规则卷曲则极具柔性,减少α-螺旋和增加无规则卷曲可增加蛋白质分子的柔性,而分子柔性与其功能密切相关[27]. 这很可能有助于蛋白在细胞相容性物质浓度较高的细胞液中保持稳定.

为了进一步了解其差异,参考相关文献[28]把氨基酸分成14种类型,包括带电的(Ch),脂肪族(Al)、芳香族(Ar)、极性的(Po)、中性的(Ne)、疏水性的(Hy)、带正电(Ps)、带负电(Ng)、微小的(Ti)、小的(Sm)、大的(La)、含硫的(Su)、酰胺(Am)及酸性与碱性氨基酸(A-B)的差值. 比较它们在HIP及HOP与NP中的差异,结果如图1(b)所示.

由图1(b)可知:HIP中性氨基酸、微小的氨基酸及带电氨基酸明显较高,疏水性氨基酸则明显较少;而HOP中仅有较小的氨基酸明显多于非嗜盐蛋白. 研究表明:中性氨基酸不与离子发生静电引力,



（a）氨基酸组　　（b）14 种类型的氨基酸

图 1　嗜盐与非嗜盐蛋白氨基酸组成的差异

Fig. 1　Compositional differences of amino acids between halophilic and non-halophilic proteins

其在高盐条件下很易形成疏水相互作用，而侧链微小或较小的氨基酸在蛋白质内核组装过程中更容易占据蛋白质分子中不同空间[29]. 这对维持蛋白稳定性非常重要.

此外，HIP 和 HOP 与 NP 中酸性氨基酸与碱性氨基酸差值均差异明显，尤其在 HIP 中. 这与大多数研究所发现的嗜盐蛋白酸性氨基酸含量明显超过碱性氨基酸含量的结果吻合. 然而，Bardavid 等[30]的研究发现，在 *Halanaerobiales* 属中一些在细胞内积累高浓度 KCl 的嗜盐蛋白并不存在这种差异.

综上所述，对于 HIP 而言，文中结果与大多数文献报道结果相吻合，但通过引入所有蛋白平均氨基酸组成使其结果更为明显；而对 HOP 而言，其氨基酸组成差异虽然没有 HIP 与 NP 这么明显，但它们可能通过增加分子柔性及更高效的内核组装，来保持蛋白质在高细胞相容性物质环境中的可溶性、稳定性及行使正确的生物学功能.

2.2　基于 PUKF 的支持向量机识别精度

支持向量机（SVM）是生物信息学领域最常用的分类工具[31]，其分类性能主要取决与核函数，而核函数的选取及其对应参数的优化非常耗时. 为此，采用一种通用核函数，通过参数调整，可适应各种数据[32-33]. 文中选取 Person 通用核函数（PUKF）[32]，它在生物学领域尚不多见[34]. 基于 PUKF 的支持向量机对两种嗜盐蛋白和两种嗜盐蛋白与非嗜盐蛋白的识别效果，如表 1 所示.

表 1　不同算法的预测精度

Tab. 1　Performances of different algorithms

算法	HIP-HOP-NP		HIP-HOP		算法	HIP-HOP-NP		HIP-HOP	
	γ/%	A	γ/%	A		γ/%	A	γ/%	A
SVM(PUKF)	84.1	0.895	92.5	0.921	Bagging(决策树桩)	52.1	0.644	77.2	0.746
SVM(多项式核函数，E=2)	82.4	0.882	91.5	0.911	Bagging(决策表)	69.7	0.853	86.7	0.930
SVM(多项式核函数，E=3)	83.4	0.890	92.2	0.917	Bagging(REP 树)	76.8	0.908	88.7	0.951
SVM(多项式核函数，E=4)	83.4	0.891	92.3	0.919	Bagging(J4.8 决策树)	77.5	0.905	89.2	0.956
SVM(多项式核函数，E=5)	83.6	0.892	92.4	0.921	Adaboost(决策树桩)	52.2	0.635	84.9	0.919
SVM（线性核函数）	78.2	0.855	88.3	0.876	Adaboost(决策表)	74.7	0.890	87.9	0.941
SVM（RBF 核函数）	75.6	0.837	85.4	0.832	Adaboost(REP 树)	75.9	0.898	88.3	0.945
RBF 神经网络	74.1	0.884	87.8	0.938	Adaboost(J4.8 决策树)	77.6	0.905	89.6	0.958
贝叶斯神经网络	76.3	0.901	87.8	0.948	Logitboost(决策树桩)	75.1	0.893	86.1	0.929
k-近邻	73.9	0.796	89.4	0.889	Logitboost(REP 树)	77.8	0.914	88.5	0.953
BP 神经网络	80.6	0.916	90.2	0.955					

对两种嗜盐蛋白而言，10-倍交叉验证的结果表明：基于 PUKF 的支持向量机识别精度最佳，可达 92.5%. 其 A 值为 0.921，大于 0.9，说明该分类器的识别效果优秀. 相比而言，其识别精度比单一分类器的径向基核函数（RBF）的支持向量机高 7.1%，比组合分类器的 Bagging（基础分类器为 Decision Stump）高 15.3%. 此外，同其他几种核函数的支持向量机相比，其准确率分别高于 RBF 和线性核函数的 7.1%和 4.2%，与多项式核函数的支持向量机比较接近. 然而，相对于后者，PUKF 的支持向量运算

所需时间明显较少,如完成本次运算,前者所需时间约 6 min,而同样条件下的多项式支持向量机($E=5$)所需时间约 14 min,是前者的 2 倍多.由此可见,PUKF 的支持向量机能兼顾运算效率与运算精度.

同样,采用 PUKF 的支持向量对两种嗜盐蛋白与非嗜盐蛋白进行识别.对于这三种类型的蛋白,该方法的 10-倍交叉验证验证的精度达到 84.1%,其 A 值达到 0.895,接近 0.9,说明其识别精度依然令人满意.相对与其他单一分类器而言,其精度提高 0.5%至 10.2%不等;相比组合分类器,其识别精度有 6.3%至 32%的提高;相比其他几种常见核函数的支持向量机,其精度依然最佳,比 RBF 和线性核函数的支持向量机分别高出 5.9%和 8.5%;相比多项式核函数的支持向量机,也有 0.5%至 1.7%左右的提升,识别效果基本相当,但其运算效率则提升明显.因此,在本识别过程中,PUKF 的支持向量机算法表现最好,而且其运算精度高、速度快,对计算机资源的消耗较少,在大规模数据分析方面更具优势.

此外,文中方法对两种嗜盐及非嗜盐蛋白识别精度为 84.1%,虽然未能达到预期的 90%以上,但相比对这 3 种类型蛋白进行随机猜测的几率 33.3%而言,其效果已明显提高了 53.8%;而对两种嗜盐的识别精度达 92.5%,相比于随机猜测的几率 50%而言,其精度提高了 42.5%.由此看来,对 3 种不同类型蛋白识别的精度还是令人满意的.

2.3 不同长度序列预测精度的差异

当使用氨基酸组成作为序列特征值时,识别精度随序列长度的变化而出现差异.这种现象在之前的相关研究中已有报道.如 Grominha[35] 在识别球状蛋白和外膜蛋白过程中发现,对少于 300 个氨基酸的蛋白而言,其识别精度为 86%,对氨基酸数量在 300～800 之间的蛋白,其识别精度高达 98%,而对大于 800 个氨基酸的蛋白,其精度为 100%.Zhang 等[36] 对嗜热和常温蛋白的预测结果也表明,对小分子蛋白(少于 200 个氨基酸),其识别精度为仅为 79%,而对大分子蛋白(大于 800 个氨基酸)的识别精度则达 100%.然而,研究者并未对此现象进行过多的解释,其可能的原因也未作进一步探讨.

按照序列长度(L)将蛋白质序列分为 4 个类型,并进行自一致性检验,其平均识别精度为 89.5%,不同长度蛋白质序列的识别精度分析结果,如表 2 所示.表 2 中:L 为序列长度;n 为序列数量;n_c 为正确预测的序列数量;φ 为不同长度蛋白质序列数量的百分比;η 为预测的准确率.从表 2 可知:随着蛋白质序列长度的增加,识别精度逐渐上升,对较小的蛋白分子,其识别精度为 86.2%,比平均值低 3.3%;对大分子蛋白,其识别精度达 97.1%,比平均值高出 7.6%;而对中等大小(200～800)的识别精度也均高于平均值.

表 2 不同长度蛋白质序列的预测结果

Tab.2 Prediction performances of different sequence lengths

项目	$100\leqslant L<200$	$200\leqslant L<500$	$500\leqslant L<800$	$L\geqslant 800$	合计
n	2 013	3 723	809	241	6 787
φ/%	29.7	54.9	11.9	3.6	100
n_c	1 736	3 357	745	234	6 072
η/%	86.2	90.2	92.1	97.1	89.5

3 结论

文中严格区分两种不同嗜盐机制的蛋白,并分别将其同非嗜盐微生物蛋白质组进行了比较.结果表明:对 HIP 而言,其结果与报道结果吻合,而对 HOP 而言,其分子中含有更多柔性的二级结构,同时分子中较小的氨基酸占多数,这在之前相关文献中未见报道.这对认知两种不同嗜盐机制蛋白稳定性的机制及对结构和功能强化的理性设计具有重要指导意义.

此外,从蛋白质类型而言,出现预测错误主要是 HOP 与 NP 之间,而从蛋白质大小而言,主要是对分子量小的蛋白预测精度偏低.因此,如何有效解决上述两个问题以提高识精度将是后续研究重点.

参考文献:

[1] EICHLER J. Biotechnological uses of archaeal extremozymes[J]. Biotechnol Adv,2001,19(4):261-278.

[2] DELGADO-GARCÍA M,VALDIVIA-URDIALES B,AGUILAR-GONZÁLEZ C N,et al. Halophilic hydrolases as a

new tool for the biotechnological industries[J]. J Sci Food Agric,2012,92(13):2575-2580.

[3] ROBERTS M F. Organic compatible solutes of halotolerant and halophilic microorganisms[J]. Saline Systems,2005,1:5.

[4] RHODES M E,FITZ-GIBBON S T,OREN A,et al. Amino acid signatures of salinity on an environmental scale with a focus on the Dead Sea[J]. Environ Microbiol,2010,12(9):2613-2623.

[5] OREN A. Microbial life at high salt concentrations: Phylogenetic and metabolic diversity[J]. Saline Systems,2008,4:2.

[6] COQUELLE N,TALON R,JUERS D H,et al. Gradual adaptive changes of a protein facing high salt concentrations [J]. J Mol Biol,2010,404(3):493-505.

[7] SIGLIOCCOLO A, PAIARDINI A, PISCITELLI M, et al. Structural adaptation of extreme halophilic proteins through decrease of conserved hydrophobic contact surface[J]. BMC Struct Biol,2011,11:50.

[8] STREET T O,BOLEN D W,ROSE G D. A molecular mechanism for osmolyte-induced protein stability[J]. Proc Natl Acad Sci USA,2006,103(38):13997-14002.

[9] EBRAHIMIE E,EBRAHIMI M,SARVESTANI N R,et al. Protein attributes contribute to halo-stability, bioinformatics approach[J]. Saline Systems,2011,7(1):1.

[10] HAYES R J,BENTZIEN J,ARY M L,et al. Combining computational and experimental screening for rapid optimization of protein properties[J]. Proc Natl Acad Sci USA,2002,99(25):15926-15931.

[11] COKER J A,DASSARMA P,KUMAR J,et al. Transcriptional profiling of the model Archaeon *Halobacterium* sp. NRC-1: Responses to changes in salinity and temperature[J]. Saline Systems,2007,25(3):6.

[12] SCHWIBBERT K,MARIN-SANGUINO A,BAGYAN I,et al. A blueprint of ectoine metabolism from the genome of the industrial producer *Halomonas elongata* DSM 2581(T)[J]. Environ Microbiol,2011,13(8):1973-1994.

[13] NIERMAN W C,FELDBLYUM T V,LAUB M T,et al. Complete genome sequence of *Caulobacter crescentus*[J]. Proc Natl Acad Sci USA,2001,98(7):4136-4141.

[14] ALTSCHUL S F,MADDEN T L,SCHAFFER A A,et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs[J]. Nucleic Acids Res,1997,25(17):3389-3402.

[15] DING Yan-rui,CAI Yu-jie,ZHANG Ge-xin,et al. The influence of dipeptide composition on protein thermostability [J]. FEBS Lett,2004,569(1/2/3):284-288.

[16] CHOU Kuo-chen,SHEN Hong-bin. Cell-PLoc: A package of web-servers for predicting subcellular localization of proteins in various organisms[J]. Nat Prot,2008,3(2):153-162.

[17] CHOU Kuo-chen,SHEN Hong-bin. Recent progresses in protein subcellular location prediction[J]. Anal Biochem,2007,370(1):1-16.

[18] WANG Tong,YANG Jie,SHEN Hong-bin,et al. Predicting membrane protein types by the LLDA algorithm[J]. Protein & Peptide Lett,2008,15(9):915-921.

[19] LI Feng-min,LI Qian-zhong Z. Predicting protein subcellular location using Chou's pseudo amino acid composition and improved hybrid approach[J]. Protein & Peptide Lett,2008,15(6):612-616.

[20] LIN Hao. The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition[J]. J Theor Biol,2008,252(2):350-356.

[21] FRANK E,HALL M,TRIGG L,et al. Data mining in bioinformatics using Weka[J]. Bioinformatics,2004,20(15):2479-2481.

[22] KASTRITIS P L,PAPANDREOU N C,HAMODRAKAS S J. Haloadaptation: Insights from comparative modeling studies of halophilic archaeal DHFRs[J]. Int J Biol Macromol,2007,41(4):447-453.

[23] PAUL S,BAG S K,DAS S,et al. Molecular signature of hypersaline adaptation: Insights from genome and proteome composition of halophilic prokaryotes[J]. Genome Biol,2008,9:R70.

[24] WRIGHT D B,BANKS D D,LOHMAN J R,et al. The effect of salts on the activity and stability of *Escherichia coli* and *Haloferax volcanii* dihydrofolate reductases[J]. J Mol Biol,2002,323(2):327-344.

[25] ARAKAWA T,TOKUNAGA M. Electrostatic and hydrophobic interactions play a major role in the stability and refolding of halophilic proteins[J]. Protein Pept Lett,2004,11(2):125-132.

[26] COSTANTINI S,COLONNA G,FACCHIANO A M. Amino acid propensities for secondary structures are influenced by the protein structural class[J]. Biochem Biophys Res Commun,2006,342(2):441-451.

[27] RADIVOJAC P,OBRADOVIC Z,SMITH D K,et al. Protein flexibility and intrinsic disorder[J]. Protein Sci,2004,13(1):71-80.

[28] BETTS M J,RUSSELL R B. Amino acid properties and consequences of substitutions[M]. Chichester: Bioinformatics for Geneticists Wiley,2003:289-316.

[29] BRITTON K L,BAKER P J, BORGES K M M, et al. Insights into thermal stability from a comparison of the glutamate dehydrogenases from *Pyrococcus furiosus* and *Thermococcus litoralis*[J]. Eur J Biochem,1995,229(3):688-695.

[30] BARDAVID R E,OREN A. The amino acid composition of proteins from anaerobic halophilic bacteria of the order Halanaerobiales[J]. Extremophiles,2012,16(3):567-572.

[31] WARD J J,MCGUFFIN L J,BUXTION B F,et al. Secondary structure prediction with support vector machines [J]. Bioinformatics,2003,19(13):1650-1655.

[32] UESTUEN B,MELSSEN W J,BUYDENS L M C. Facilitating the application of support vector regression by using a universal Pearson Ⅶ function based kernel[J]. Chemometrics and Intelligent Laboratory Systems,2006,81(1):29-40.

[33] 郑启富,陈德钊,刘化章. 基于PersonⅦ核函数的支持向量机及其在化学模式分类中的应用[J]. 分析化学,2007,35(8):1142-1146.

[34] SABDERS W S,JOHNSTON C I,BRIDGES S M,et al. Prediction of cell penetrating peptides by support vector machines[J]. PLOS Comput Biol,2011,7(7):e1002101.

[35] GROMIHA M M. Motifs in outer membrane protein sequences: Applications for discrimination[J]. Biophy Chem,2005,117(1):65-71.

[36] ZHANG Guang-ya,FANG Bai-shan. LogitBoost classifier for discriminating thermophilic and mesophilic proteins [J]. J Biotechnol,2007,127(3):417-424.

Amino Acid Signatures of Different Hypersanline Adaptation Proteomes and Their Classification

GE Hui-hua, HUANG Ke-jun, ZHANG Guang-ya

(College of Chemical Engineering, Huaqiao University, Xiamen 361021, China)

Abstract: We selected two halophilic proteomes with different halophilic mechanism, and compared with a non-halophilic one. The results showed the difference between the halophilic (salt-in) and the non-halophilic proteome was obvious than that of halophilic (salt-out) and the non-halophilic proteome. In the halophilic (salt-out) proteome, the His and the small residues were significantly higher than those of non-halophilic proteome, while the Ala was significantly lower. However, both halophilic proteomes showed a large excess of acidic over basic amino acids. Based on these results, we introduced a novel Person Universal Kernel Function based support vector machine to classify the three kinds of proteins and the overall prediction accuracy could reach 84.1%. This method outperformed support vector machines based on other usually used kernels and other machine learning algorithms.

Keywords: halophile; non-halophile; proteome; amino acid; support vector machine; discrimination

(**责任编辑**: 黄晓楠 **英文审校**: 刘源岗)