DOI:10.13656/j.cnki.gxkx.2013.03.016

广西科学 Guangxi Sciences 2013,20(3):234～238,243

氨基酸和蛋白质的组合特征与秀丽隐杆线虫蛋白质的结晶倾向的相关分析*

Correlation of Combined Features of Amino Acid and Protein with Crystallization Propensity of Proteins from *Caenorhabditis elegans*

严少敏,王何健,吴　光**
YAN Shao-min,WANG He-jian,WU Guang

(广西科学院非粮生物质酶解国家重点实验室,国家非粮生物质能源工程技术研究中心,广西生物炼制重点实验室,广西南宁　530007)
(State Key Laboratory of Non-food Biomass and Enzyme Technology,National Engineering Research Center for Non-food Biorefinery,Guangxi Key Laboratory of Biorefinery,Guangxi Academy of Sciences,Nanning,Guangxi,530007,China)

摘要:通过逻辑回归和神经网络模型,分别研究3种单个氨基酸和整个蛋白质的组合特征与秀丽隐杆线虫(*Caenorhabditis elegans*)蛋白质的结晶倾向的相关性,并以535种单个氨基酸特征为基准,进行蛋白质结晶倾向的相关比较。结果显示,组合特征与秀丽隐杆线虫蛋白质的结晶倾向具有相关性,可用于预测蛋白质的结晶倾向。

关键词:秀丽隐杆线虫　逻辑回归　神经网络　预测　蛋白质结晶
中图法分类号:Q51　　**文献标识码**:A　　**文章编号**:1005-9164(2013)03-0234-05

Abstract:By means of logistic regression and neural network,each of three combined features of individual amino acids and a whole protein is correlated with crystallization propensity of proteins from *Caenorhabditis elegans* while each of 535 features of individual amino acids is also correlated with crystallization propensity of the proteins to serve as benchmark. The results show that the combined features have the relationship with crystallization propensity of proteins from *Caenorhabditis elegans*. This study provides the information that the combined features can be used for predicting crystallization propensity of protein.

Key words: *Caenorhabditis elegans*, logistic regression, neural network, prediction, protein crystallization

尽管蛋白质结晶的研究有了巨大进展,蛋白质是否能够结晶的问题仍然没有完全解决[1]。X射线晶体学理论和NMR是测定蛋白质3D结构的有效工具,但是这些方法非常耗时,而成本又高。因此,科研人员进行了许多单个氨基酸或者整个蛋白质特征与蛋白质结晶倾向的相关性研究,开发可以准确预测某种蛋白质是否可以结晶的模型[2~8],利用单个氨基酸或整个蛋白质的特征来预测蛋白质的结晶倾向。实际上此类研究还用在结晶前的其它过程,例如,蛋白质表达、蛋白质提纯等。

迄今为止,几乎所有已知的单个氨基酸和整个蛋白质的特征都已被用于预测,即AAIdex中540多种单个氨基酸的特征[9]。但不可否认的是,每种特征都具有局限性。如分子量可以作为氨基酸的一个特征,对于同种类型的氨基酸而言,其分子量是恒定不变的,因此,不论这种氨基酸位于蛋白质中何处、还是它

收稿日期:2013-05-23
作者简介:严少敏(1958-),女,博士,研究员,主要从事计算变异学和模型研究。
*广西科学基金项目(11107021-5-2、12237022、13-051-08、13-051-50和2013GXNSFDA019007)和八桂学者建设工程专项经费资助。
**通讯作者:吴　光(1956-),男,博士,研究员,主要从事模型分析研究。E-mail:hongguanglishibahao@yahoo.com。

与哪些氨基酸相邻，它的分子量都不会变。也就是说，单个氨基酸的特征并不能反映整个蛋白质中氨基酸的特性。另一方面，整体蛋白质的特征，如蛋白质长度，显得过于简单，因为它们并不包含单个氨基酸的特征。由此可见，研发一种有效的方法用于预测特定蛋白质的结晶倾向极为有用，本研究目的就是分析并解决这一重要问题。在过去的十多年里，我们研发出了3个组合特征，它们结合了单个氨基酸和整个蛋白质的特性[10~13]。

秀丽隐杆线虫（*Caenorhabditis elegans*）是一种较为原始的小蠕虫[14]，全身透明，共有959个细胞，不需染色即可进行显微镜下研究；其整个生命周期仅3天，细胞分裂和组织形成具有高度的程序性；秀丽隐杆线虫的染色体数很少，基因组也很小，而且非重复序列很高（达到83%），这些特点使它成为现代发育生物学、遗传学和基因组学研究的重要模式材料[15]。该线虫器官发育和“程序性细胞死亡”过程中基因规则的发现获得了2002年诺贝尔生理学或医学奖。目前，秀丽隐杆线虫已成为用于研究多种疾病的机理和治疗新方法的经典模型[16~19]。因此，我们选用秀丽隐杆线虫的蛋白质作为研究对象，以期在预测其结晶倾向方面获得一些更为深入的认识。在此研究中，我们分别以530多种单个氨基酸特征为基准，确定组合特征是否与秀丽隐杆线虫蛋白质的结晶倾向相关。

1 材料与方法

1.1 数据

2011年以前，数据库TargetDB[20]中在纯化栏目下纪录了454个秀丽隐杆线虫蛋白质，其中有117个蛋白质已被结晶[5]。

组合特征和基准特征的区别在于组合特征需要对每个蛋白质逐一计算，而基准特征则是常量。

1.2 第1个组合特征的计算

氨基酸的分布概率需要使用 $\frac{n!}{q_0! \times q_1! \times \cdots \times q_k!} \times \frac{r!}{r_1! \times r_2! \times \cdots \times r_n! \times n^{-r}}$[21]来计算蛋白质中的每种氨基酸的分布概率（该值并非常量），其中！表示阶乘，r是某种氨基酸的数量，q是具有相同数量氨基酸的组分数，n是某种类型的氨基酸在蛋白质中的组分数。在线计算可以从http://www.nerc-nfb.ac.cn/calculation/dp.htm获取。

1.3 第2个组合特征的计算

以64个RNA密码子和20个被翻译的氨基酸之间的关系为基础计算氨基酸的未来组成[22~24]，这种关系会因翻译概率的不同导致其不成比例。例如，蛋氨酸与1个RNA密码子（AUG）相对应，苯丙氨酸与2个RNA密码子（UUC和UUU）相对应，而亮氨酸却与6个RNA密码子（CUA，CUC，CUG，CUU，UUA和UUG）相对应。可通过在线服务器（http://www.nerc-nfb.ac.cn/calculation/fc.htm）算出这一特征。

1.4 第3个组合特征的计算

以排列为基础计算氨基酸对的可预测性[10~13]。例如，G_YK4166蛋白质由245个氨基酸组成，其中有22个精氨酸（R）、25个谷氨酸（E）和19个天门冬氨酸（D）。根据排列组合，氨基酸对RE将出现2次（22/245×25/244×244＝2.24），而这个蛋白质中的确有2对RE，因此氨基酸对RE是可预测的。然而，氨基酸对DD本应出现1次（19/245×18/244×244＝1.40），但事实上它却出现了6次，所以氨基酸对DD是不可预测的。通过这种计算可将所有的氨基酸对分为可预测的和不可预测的两类，氨基酸对的可预测部分和不可预测部分均可成为量化蛋白质的一种指标。对于G_YK4166蛋白质而言，其可预测和不可预测的氨基酸对数量所占百分比分别是58.25%和41.75%。蛋白质的可预测部分这一组合特征可登录网站查询（http://www.nerc-nfb.ac.cn/calculation/pp.htm）。

1.5 建模分析

组合特征是否与蛋白质结晶倾向相关联这一问题需通过建模来决定，因为没有一种实验可以排除单个氨基酸特征或整个蛋白质特征的影响。我们采用逻辑回归模型与20-1前馈反向传播神经网络模型预测结晶倾向。分别以20种氨基酸的每一种特征性的数值编码作为模型的输入，秀丽隐杆线虫蛋白质的结晶成功率以“是”或“否”的形式作为模型输出。结果分为真阳性、真阴性、假阳性和假阴性，按常规方法计算预测的准确性、敏感性和特异性。使用MatLab进行逻辑回归和神经网络的运算。用Mann-Whitney U-检验来进行统计比较。

2 结果与分析

表1列出两个组合特征与JOND750101的不同之处。JOND750101是一个描述疏水性的氨基酸特征，用20个不同的值代表20种不同的氨基酸（第4列和第5列），尽管两种蛋白质中氨基酸的组成结构不同（第2列和第3列），但是这些值都是恒定的。为了结合整个蛋白质中的信息，我们可以通过增加氨基

表 1　比较 JOND750101 与组合特征对蛋白质的量化结果

Table 1　Difference between combined features and JOND750101 in two exampled proteins

氨基酸 Amino acid	数量 Number		疏水性 JOND750101		疏水性×数量 JOND750101×Number		分布概率 Distribution probability		未来组成 Future composition(%)	
	P_1	P_2	P_1	P_2	P_1	P_2	P_1	P_2	P_1	P_2
A	11	17	0.87	0.87	9.57	14.79	0.2020	0.1098	6.21	6.58
R	22	11	0.85	0.85	18.70	9.35	0.0171	0.1010	8.53	7.38
N	14	10	0.09	0.09	1.26	0.90	0.1178	0.1143	3.76	4.84
D	19	11	0.66	0.66	12.54	7.26	0.0005	0.2020	5.28	4.10
C	3	3	1.52	1.52	4.56	4.56	0.1111	0.2222	2.01	2.07
E	25	18	0	0	0	0	0.0189	0.0831	5.06	4.10
Q	18	7	0.67	0.67	12.06	4.69	0.0831	0.1071	4.14	3.89
G	14	8	0.10	0.10	1.40	0.80	0.0079	0.2523	6.53	4.85
H	8	9	0.87	0.87	6.96	7.83	0.2243	0.1967	4.73	3.08
I	6	16	3.15	3.15	18.90	50.40	0.3472	0.1362	4.27	6.02
L	24	20	2.17	2.17	52.08	43.40	0.0396	0.0023	8.47	9.24
K	11	24	1.64	1.64	18.04	39.36	0.0040	0.0352	4.60	4.26
M	5	8	1.67	1.67	8.35	13.36	0.3840	0.2523	1.32	2.00
F	5	12	2.87	2.87	14.35	34.44	0.3840	0.0709	2.45	2.84
P	12	17	2.77	2.77	33.24	47.09	0.0621	0.0005	5.61	5.93
S	15	15	0.07	0.07	1.05	1.05	0.0392	0.0981	6.71	7.36
T	10	16	0.07	0.07	0.70	1.12	0.1524	0.1362	5.05	7.20
W	3	4	3.77	3.77	11.31	15.08	0.2222	0.0938	0.92	0.66
Y	6	5	2.67	2.67	16.02	13.35	0.1543	0.0384	2.72	2.49
V	14	14	1.87	1.87	26.18	26.18	0.0589	0.1178	6.85	6.89

P_1：蛋白 1(G_YK4166)；P_2：蛋白 2(Y39E4B.11)；JOND750101：氨基酸疏水性特征指标。

The accession numbers are G_YK4166 for protein 1(P_1) and Y39E4B.11 for protein 2(P_2). JOND750101 is an amino acid feature to describe the hydrophobicity.

酸的组成成分的方式来权衡在第 4 列和第 5 列的常量，就像在第 6 列和第 7 列中展示的那样。然而，氨基酸的分布概率和未来组成这两个组合特征，会随着蛋白质和氨基酸种类的不同而出现不同的值(表 1 中的后 4 列)。虽然半胱氨酸(C)、丝氨酸(S)和缬氨酸(V)的数量在这两个蛋白质中相同，但它们的分布概率和未来组成却各不相同，充分体现了组合特征的动态特点。

通过逻辑回归模型对结晶涉及到的组合特征与单个氨基酸的其他特征进行比较，结果如图 1 所示。图中的横轴为分组特征，纵轴分别为预测的精确度(上图)、敏感性(中图)和特异性(下图)，灰柱上方的数字表示氨基酸特征数。图 1 可以这样解读：每一条灰柱代表有多少特征得到了相同或相似的精确度、敏感性以及特异性。例如，单个氨基酸的 3 个特征(EISD860103、EISD860102 和 ZIMJ680103)有相同的精确度 0.731(图 1 上图左数第 1 柱)，而 11 个特征因其在精度方面的相似性可组成一组(0.742±0.001，图 1 上图左数第 5 柱)。结果显示单个氨基酸特征和两个组合特征都与蛋白质结晶倾向相关联。事实上，图 1 体现了各种特征在蛋白质结晶中所扮演角色的程度这一问题，表明两个组合特征确实与秀丽

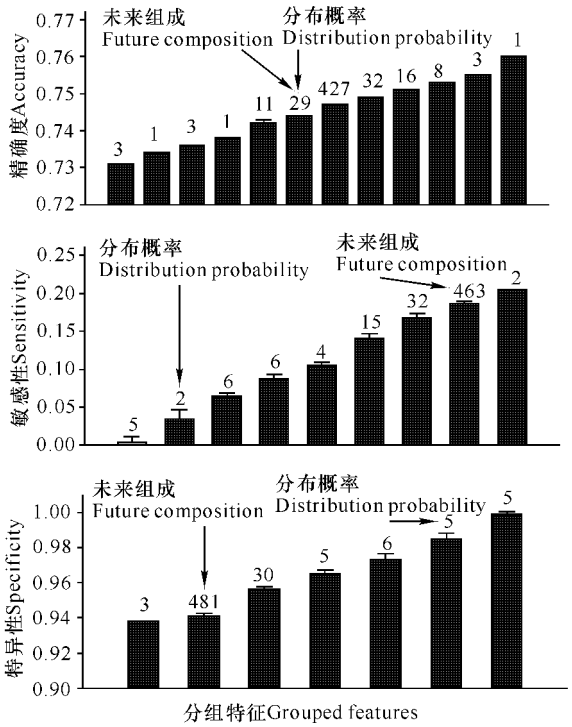

图 1　用逻辑回归模型预测秀丽隐杆线虫蛋白质结晶倾向的精确度、敏感性和特异性

Fig. 1　Accuracy, sensitivity and specificity for predicting crystallization propensity of *Caenorhabditis elegans* proteins obtained from logistic regression

隐杆线虫蛋白质的结晶具有相关性。在模型分析中，

我们每次仅用 1 个特征，以便在不同特征之间比较分类结果，这不同于其他研究[1,6]，他们同时采用几乎所有 500 多个氨基酸特征用于结晶成功率的建模。

通常逻辑回归的形式都很简单，因为其关系式 $P(y)=\frac{1}{1+e^{b_0+b_1x_1+\cdots+b_{20}x_{20}}}$，式中 x_i 是 20 种氨基酸各自的特征，y 为结晶成功率，b_i 是模型参数。因此，我们引入了神经网络模型，因为原则上它能解释各种隐性与显性关系[25,26]。图 2 用 20-1 神经网络模型拟合秀丽隐杆线虫蛋白质结晶倾向的精确度、敏感性和特异性，其表述方式与图 1 一样。显然，神经网络模型与逻辑回归模型都得到了相似的分组特征结果，而氨基酸分布概率这个组合特征得到了最高的结晶预测精确度和敏感性。

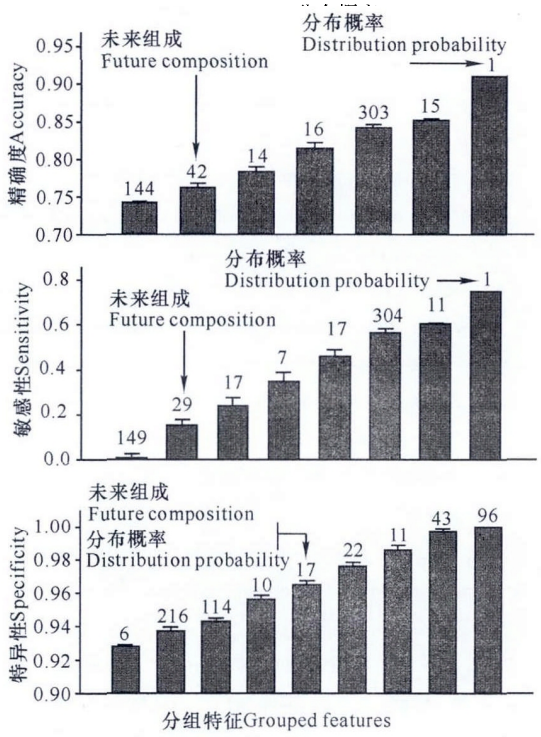

图 2　用 20-1 神经网络模型拟合秀丽隐杆线虫蛋白质结晶倾向的精确度、敏感性和特异性

Fig. 2 Accuracy, sensitivity and specificity for fitting crystallization propensity of *Caenorhabditis elegans* proteins obtained from the 20-1 feedforward backpropagation neural network

一旦确定了组合特征在结晶过程的作用，我们需要对预测结果进行验证。图 3 使用删除 1 个蛋白质的折刀法验证秀丽隐杆线虫蛋白质结晶倾向的结果，其表示方法与图 1 和图 2 相同，唯一的区别在于将秀丽隐杆线虫的 454 个蛋白质分成两组，前一组用来生成模型参数，后一组则用于预测，每次用删除 1 个蛋白质的折刀法进行验证。可以看出，组合特征有相对较好的预测性。

图 1～3 显示对秀丽隐杆线虫蛋白质结晶的预测效果，特异性远远高于敏感性，即对非结晶蛋白质的预测效果优于结晶蛋白质，其统计差异非常显著（$P<0.001$，图 4）。另外，从图 1、图 2 和图 3 中得到的上述结果涉及了两种组合特征，即氨基酸的分布概率和未来组成。而图 4 的右侧数据显示了另一个组合特征即氨基酸对的可预测部分的结果，这个可预测部

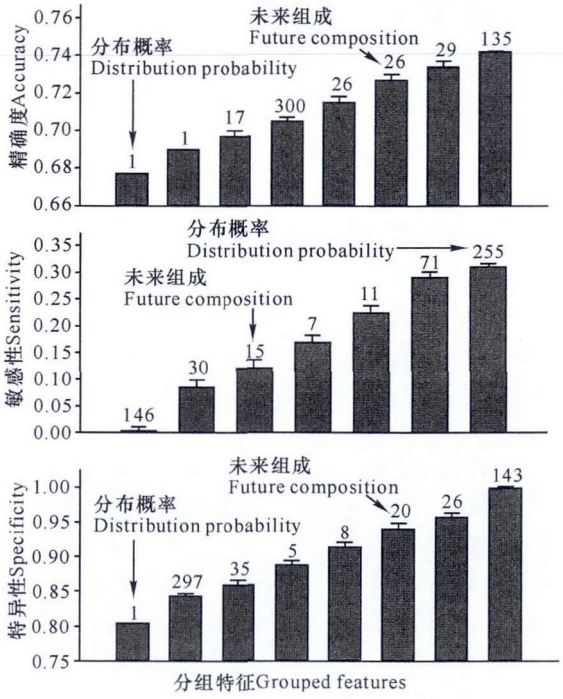

图 3　用删除 1 个蛋白质的折刀法验证秀丽隐杆线虫蛋白质结晶倾向的精确度、敏感性和特异性

Fig. 3　Accuracy, sensitivity and specificity for validation of crystallization propensity of *Caenorhabditis elegans* proteins using delete-1 jackknife validation

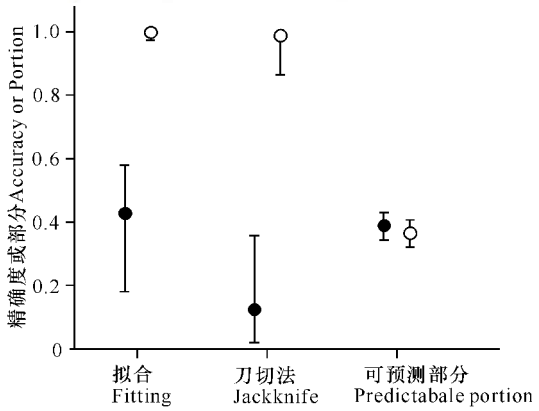

图 4　结晶与非结晶秀丽隐杆线虫蛋白质的预测精确度和氨基酸对可预测部分的统计比较

Fig. 4　Statistical comparison of prediction accuracy and predictable portion of amino acid pairs in crystallized and non-crystallized *Caenorhabditis elegans* proteins

数据以中值和 25%～75%区间表示。The data were presented as median with inter-quartiles. $P<0.001$ (the Mann-Whitney U-test).

●：结晶，○：非结晶。●: Crystallized, ○: Noncrystallized.

分专门适用于以一个数值来描述整个蛋白质。结果显示结晶蛋白质的可预测部分大于非结晶蛋白质（$P<0.001$），提示蛋白质结构的随机性越大越容易结晶。

3 结论

在本研究中，我们使用了3种氨基酸和蛋白质的结合特征来预测秀丽隐杆线虫蛋白质的结晶倾向，并与530多种氨基酸特征进行比较。所得结果与我们之前的研究[27~29]相一致，表明组合特征不仅与结晶过程有关，还可用于预测蛋白质的结晶倾向，有助于深入理解蛋白质的结晶过程。

参考文献：

[1] Kurgan L, Mizianty M J. Sequence-based protein crystallization propensity prediction for structural genomics: review and comparative analysis[J]. Natural Sci, 2009, 1:93106.

[2] Smialowski P, Schmidt T, Cox J, et al. Will my protein crystallize? A sequence-based predictor[J]. Proteins, 2006, 62:343-355.

[3] Overton I M, Barton G J. A normalised scale for structural genomics target ranking: the OB-Score[J]. FEBS Letters, 2006, 580:4005-4009.

[4] Slabinski L, Jaroszewski L, Rodrigues A P C, et al. The challenge of protein structure determination - lessons from structural genomics[J]. Protein Science, 2007, 16: 2472-2482.

[5] Slabinski L, Jaroszewski L, Rychlewski L, et al. XtalPred: a web server for prediction of protein crystallizability [J]. Bioinformatics, 2007, 23:3403-3405.

[6] Chen K, Kurgan L, Rahbari M. Prediction of protein crystallization using collocation of amino acid pairs[J]. Biochemical and Biophysical Research Communications, 2007, 355:764-769.

[7] Overton I M, Padovani G, Girolami M A, et al. ParCrys: a Parzen window density estimation approach to protein crystallization propensity prediction[J]. Bioinformatics, 2008, 24:901-907.

[8] Kurgan L, Razib A A, Aghakhani S, et al. CRYSTALP2: sequence-based protein crystallization propensity prediction[J]. BMC Structural Biology, 2009, 9:50.

[9] Kawashima S, Pokarowski P, Pokarowska M, et al. AAindex: amino acid index database, progress report 2008[J]. Nucleic Acids Res, 2008, 36:D202-D205.

[10] Wu G, Yan S. Randomness in the primary structure of protein: methods and implications[J]. Mol Biol Today, 2002, 3:55-69.

[11] Wu G, Yan S. Mutation trend of hemagglutinin of influenza A virus: a review from computational mutation viewpoint[J]. Acta Pharmacol Sin, 2006, 27:513-526.

[12] Wu G, Yan S. Lecture notes on computational mutation [M]. New York: Nova Sciences Publishers, 2008.

[13] Yan S M, Wu G. Creation and application of computational mutation[J]. J Guangxi Acad Sci, 2010, 26:130-139.

[14] Zhang S, Kuhn J R. Cell isolation and culture[J]. WormBook, 2013, 21:1-39.

[15] Hanna M, Wang L, Audhya A. Worming our way in and out of the *Caenorhabditis elegans* germline and developing embryo[J]. Traffic, 2013, 14:471-478.

[16] Frooninckx L, Van Rompay L, Temmerman L, et al. Neuropeptide GPCRs in *C. elegans* [J]. Front Endocrinol (Lausanne), 2012, 3:167.

[17] Hashmi S, Wang Y, Parhar R S, et al. A *C. elegans* model to study human metabolic regulation[J]. Nutr Metab (Lond), 2013, 10:31.

[18] Yang S, Chen Y, Ahmadie R, et al. Advancements in the field of intravaginal siRNA delivery[J]. J Control Release, 2013, 167:29-39.

[19] Vistbakka J, VanDuyn N, Wong G, et al. *C. elegans* as a genetic model system to identify Parkinson's disease-associated therapeutic targets [J]. CNS Neurol Disord Drug Targets, 2012, 11:957-964.

[20] Chen L, Oughtred R, Berman H M, et al. TargetDB: a target registration database for structural genomics projects[J]. Bioinformatics, 2004, 20:2860-2862.

[21] Feller W. An introduction to probability theory and its applications [M]. 3rd ed. New York: Wiley, Vol I, 1968.

[22] Wu G, Yan S. Determination of mutation trend in proteins by means of translation probability between RNA codes and mutated amino acids [J]. Biochem Biophys Res Commun, 2005, 337:692-700.

[23] Wu G, Yan S. Determination of mutation trend in hemagglutinins by means of translation probability between RNA codons and mutated amino acids [J]. Protein Pept Lett, 2006, 13:601-609.

[24] Wu G, Yan S. Translation probability between RNA codons and translated amino acids, and its applications to protein mutations[M]//Ostrovskiy M H. Leading-Edge Messenger RNA Research Communications. New York: Nova Science Publishers, 2007:47-65.

[25] Demuth H, Beale M. Neural network toolbox for use with MatLab[S]. User's guide, version 4, 2001.

（下转第243页 Continue on page 243）

多种序列分析比较结果揭示了 PdhE-1 蛋白与其它预苯酸脱氢酶的氨基酸序列一致性不高，但是与催化功能密切相关的氨基酸残基却十分保守。系统发育进化树分析结果表明 PdhE-1 与 TyrA 蛋白家族中的预苯酸脱氢酶处于同一分支，但是与家族中其他成员之间的进化距离非常远。新型预苯酸脱氢酶基因 *pdhE-1* 的克隆和生物信息学分析研究为进一步完成酶基因的功能鉴定奠定了基础。

参考文献：

[1] Carol B, Terrence D, Kaitlyn H. Cohesion group approach for evolutionary analysis of tyra, a protein family with wide-ranging substrate specificities[J]. Microbiol Mol Biol, 2008, 72(1): 13-53.

[2] Song J, Bonner C, Wolinsky M, et al. The TyrA family of aromatic-pathway dehydrogenases in phylogenetic context[J]. BMC Biol, 2005, 3: 13.

[3] Ku H K, Park S R, Yang I, et al. Expression and functional characterization of prephenate dehydrogenase from *Streptococcus mutans* [J]. Process Biochem, 2010, 45(4): 607-612.

[4] O'Brien C, Mahoney C, Tharion WJ, et al. Dietary tyrosine benefits cognitive and psychomotor performance during body cooling[J]. Physiol Behav, 2007, 90(2-3): 301-307.

[5] Neri D F, Wiegmann D, Stanny R R, et al. The effects of tyrosine on cognitive performance during extended wakefulness[J]. Aviat Space Environ Med, 1995, 66(4): 313-319.

[6] Bonuccelli U, Del Dotto P. New pharmacologic horizons in the treatment of Parkinson disease[J]. Neurology, 2006, 67(7 Suppl 2): 30-38.

[7] Cowan D A, Arslanoglu A, Burton S G, et al. Metagenomics, gene discovery and the ideal biocatalyst[J]. Biochem Soc Trans, 2004, 32(Pt 2): 298-302.

[8] Lorenz P, Eck J. Metagenomics and industrial applications[J]. Nat Rev Microbiol, 2005, 3(6): 510-516.

[9] Jiang C, Wu B. Molecular cloning and functional characterization of a novel decarboxylase from uncultured microorganisms[J]. Biochem Biophys Res Commun, 2007, 357(2): 421-426.

[10] Lee S W, Won K, Lim H K. Screening for novel lipolytic enzymes from uncultured soil microorganisms[J]. Appl Microbiol Biotechnol, 2004, 65(6): 720-726.

[11] Roh C, Villatte F, Kim B G, et al. Screening and purification for novel cytochrome b(5) from uncultured environmental micro-organisms[J]. Lett Appl Microbiol. 2007, 44(5): 475-480.

（责任编辑：陈小玲）

（上接第 238 页 Continue from page 238）

[26] MathWorks Inc. MatLab-The language of technical computing (version 6. 1. 0. 450, release 12. 1, 1984-2001) [CP]. 2001.

[27] Yan S, Wu G. Correlating dynamic amino acid properties with success rate of crystallization of proteins from *Bacteroides Vulgatus* [J]. Cryst Res Tech, 2012, 47: 511-516.

[28] Yan S, Wu G. Possible random mechanism in crystallization evidenced in proteins from *Plasmodium Falciparum* [J]. Cryst Growth Des, 2011, 11: 4198-4204.

[29] Yan S, Wu G. Randomness in crystallization of proteins from *Staphylococcus Aureus* [J]. Protein Pept Lett, 2012, 19: 784-789.

（责任编辑：陈小玲）