Hans Journal of Biomedicine 生物医学, 2021, 11(4), 214-219
Published Online October 2021 in Hans. http://www.hanspub.org/journal/hjbm
https://doi.org/10.12677/hjbm.2021.114027

Hans汉斯

酶蛋白质中8类二级结构的识别

高苏娟

内蒙古工业大学理学院，内蒙古 呼和浩特

收稿日期：2021年8月13日；录用日期：2021年9月3日；发布日期：2021年9月15日

摘　要

酶是一种具有催化功能的蛋白质，研究酶蛋白质中的二级结构对研究酶的结构及功能有重要作用。本文从酶蛋白质序列出发，以位点氨基酸和20种氨基酸*n*-gap 2肽组分为参数，首次将矩阵打分的方法用于酶蛋白质中8类二级结构的识别，预测总精度Q8最高达到61.4%。

关键词

酶蛋白质，蛋白质二级结构，矩阵打分

Indentification of 8-State Secondary Structure in Enzymes Protein

Sujuan Gao

College of Sciences, Inner Mongolia University of Technology, Huhhot Inner Mongolia

Received: Aug. 13th, 2021; accepted: Sep. 3rd, 2021; published: Sep. 15th, 2021

Abstract

Enzymes are a kind of protein that has catalytic function. The study of secondary structures in enzymes plays an important role in the structure and function of enzymes. Based on enzyme protein sequence information, amino acids of sites and *n*-gap dipeptide composition of twenty amino acids were selected as parameters. Scoring matrix method was first applied to the identification of 8-state secondary structure in enzymes protein. The prediction accuracy of Q8 reached 61.4%.

Keywords

Enzyme Protein, Protein Secondary Structure, Scoring Matrix

文章引用: 高苏娟. 酶蛋白质中 8 类二级结构的识别[J]. 生物医学, 2021, 11(4): 214-219.
DOI: 10.12677/hjbm.2021.114027

1. 引言

蛋白质的二级结构是指多肽链中主链原子的局部空间排布，是不涉及侧链部分的构象。它们是完整肽链构象(三级结构)的结构单元，是蛋白质复杂的空间构象的基础。蛋白质二级结构预测通常作为蛋白质空间结构预测的第一步，是了解蛋白质的折叠模式和三级结构的基础，并为研究蛋白质的功能以及它们之间的相互作用模式提供结构基础，同时还可以为新药研发提供帮助。因此，对蛋白质二级结构的预测具有重要的理论意义和应用价值。

以往对蛋白质二级结构的预测研究大多集中在 3 态(H, E, C) [1]-[7]，近年来，有一些研究已经从 3 态拓展到 8 态(G, H, I, E, B, T, S, C)，8 态二级结构能够比 3 态二级结构提供更加细致的结构信息，在很多应用中特别重要，但是对 8 态二级结构的预测仍然相对较少。2002 年 Pollastri [8]等人首次用服务器 SSPRO8 预测了 8 态蛋白质二级结构，平均预测精度 Q8 在 62%~63%之间；2011 年王[9]等人用有条件的神经域模型(CNFs)预测 8 态蛋白质二级结构，平均预测精度 Q8 达到 67.9%，但是，其中 G，I，B，S 的预测精度非常低，主要原因是在蛋白质数据库中出现的频率较低；2013 年从[10]等人基于结构的位置特异性打分矩阵(SPSSM8)预测 8 态蛋白质二级结构，预测精度更高，Q8 达到 71.7%，G，I，B，S 的预测精度也有所提高并且各类预测精度相对平衡；2014 年 Yaseen [11]等人利用结构信息和环境特性建立结构模板(C8-SCORPION)对 8 类蛋白质二级结构预测，预测精度进一步提高，Q8 达到 78.85%，但是 I 的预测精度为零，主要由于 I 太少，常常被错误的归类为 H。

本文对酶蛋白质中8类二级结构进行预测，还未见相关报道。酶是一类特殊的蛋白质，是生命中必需和通用的大分子，研究酶类中二级结构对研究酶催化作用的分子机制[12]，酶活性[13]等方面有重要意义。

2. 材料和方法

2.1. 数据库

首先从SCOP (http://scop.mrc-lmb.cam.ac.uk/scop/)数据库中整理出序列相似性 < 25%，分辨率 < 3.0 Å的蛋白质4442个，从中按照酶的EC编号[14]挑选出2261个酶蛋白质。为了构建更大的数据库，又选取了另外一个数据库EVA，来自http://cubic.bioc.columbia.edu/eva/res/weeks.html#unique (on November 25, 2002)，包括2878个蛋白质，序列相似性小于33%，去掉和SCOP中重复的蛋白质632个后，剩余蛋白质2246个，再按照酶的EC编号[14]挑选出841个酶蛋白质，所以最后得到的酶蛋白质数量是3102个。依据DSSP [15]定义，蛋白质二级结构分成8类，H (alpha-helix)，G (3_{10} helix)，I (π-helix)，E (extended beta-strand)，B (isolated beta-strand)，T (turns)，S (bend) and others (C)即H，E，B，G，I，S，T，C 8个字符。

在3102个酶蛋白质中，每个氨基酸序列，每次移动步长为1个残基，分别截取21残基长度的片段981,470个，然后依据其中心残基的二级结构类型分成8个子库，即H，E，B，G，I，S，T，C库。中心残基为H的有324,461个，中心残基为E的有324,461个，中心残基为B的有12,236个，中心残基为G的有39,193个，中心残基为I的有241个，中心残基为S的有89,792个，中心残基为T的有114,909个，中心残基为C的有200,245个。去掉非标准氨基酸残基后，中心残基为H的有300,045个，中心残基为E的有182,930个，中心残基为B的有10,957个，中心残基为G的有35,770个，中心残基为I的有222个，中心残基为S的有81,612个，

中心残基为T的有104,981个，中心残基为C的有182,218个，见表1。

Table 1. The numbers of the 8-state secondary structure
表 1. 8 类二级结构数量

H (个)	E (个)	B (个)	G (个)	I (个)	S (个)	T (个)	C (个)
300,045	182,930	10,957	35,770	222	81,612	104,981	182,218

2.2. 计算方法

2.2.1. 矩阵打分算法(PCSF)

矩阵打分方法在转录因子结合位点预测，超二级结构预测方面取得较好结果[16] [17] [18] [19]。本文以位点氨基酸和20种氨基酸n-gap 2肽组分(n = 1，即紧邻，n = 2，次紧邻，n = 3，次次紧邻)作为参数，将酶蛋白质中的8类二级结构用矩阵打分的方法分类。

1) 位置权重矩阵(PWM)

考虑到氨基酸频率计数时的标准偏差的影响，我们引入了伪计数概率[19]来计算二级结构的位点位置概率，公式如下：

$$p_{ij} = \frac{n_{ij} + \frac{\sqrt{N_i}}{l}}{N_i + \sqrt{N_i}} \tag{1}$$

这里，以位点氨基酸为参数时，l = 20，j表示20种氨基酸，N_i表示第i个位置上所有氨基酸出现的总数，n_{ij}表示第i个位置上第j种氨基酸出现的频数；以20种氨基酸n-gap 2肽组分为参数时，l = 400，N_i表示第i个位置上所有氨基酸2肽组分出现的总数，n_{ij}表示第i个位置上第j种氨基酸2肽组分出现的频数。

利用位点位置概率，构建位置权重矩阵W。位置权重矩阵的矩阵元定义为：

$$w_{ij} = \log \frac{p_{ij}}{p_{0j}} \tag{2}$$

其中，p_{0j}表示氨基酸j出现的背景概率。以位点氨基酸为参数的矩阵是20行L列；以氨基酸2肽组分为参数的矩阵是400行$L-n$列，L为选取的酶蛋白质二级结构序列模式的片断长度。

2) 打分函数

为识别待测序列的二级结构类型， 我们在训练集中建立了8种二级结构相应的位置权重矩阵{W_H}，{W_E}，{W_B}，{W_G}，{W_I}，{W_S}，{W_T}，{W_C}。对于任意一个待测序列，应用位置权重矩阵中每一位置上与所给序列对应氨基酸的矩阵元之和，即打分函数：

$$s = \sum_{i=1}^{L} w_{ij} \tag{3}$$

这样对于同一待测序列，通过和8种二级结构序列构建的位置权重矩阵{W_H}，{W_E}，{W_B}，{W_G}，{W_I}，{W_S}，{W_T}，{W_C}比对打分， 得到8个不同的分值S_H，S_E，S_B，S_G，S_I，S_S，S_T，S_C，分值越大，与位置权重矩阵描述的二级结构类型越相似。我们比较8个分值，谁的得分最大，待测序列的结构就被预测为该类。

2.2.2. 系统检验

本文对分类结果的评价使用10交叉检验的方法，随机将8类数据集共898,735个21残基片段(H：300,045个，E：182,930个，B：10,957个，G：35,770个，I：222个，S：81,612个，T：104,981个，C：182,218

个)分为10个子集，依次取出1个子集作测试集，而其余9个子集作为训练集，此过程循环10次。

2.2.3. 精确度评价指标

本文用2个指标来衡量预测的精度，第一个是蛋白质二级结构8态预测的总精度：

$$Q_8 = \frac{\sum_{i=1}^{8} c_i}{N} \times 100\% \tag{4}$$

其中，c_i 表示被正确预测的8态总样本数(i = H, E, B, G, I, S, T, C)，N表示蛋白质二级结构8态(H, E, B, G, I, S, T, C)的总样本数。另外一个指标是8种二级结构的正确预测率：

$$Q_i = \frac{A_{ii}}{a_i} \times 100\% \tag{5}$$

其中，A_{ii} 表示二级结构为i被正确预测的样本数，a_i 表示该结构中总样本数。

3. 结果与讨论

在3102个酶蛋白质中，分别截取全长21残基的片段，移动步长为1个残基，将得到的所有21残基片段根据其中心残基的二级结构类型(H, E, B, G, I, S, T, C)分成8个集合。依据公式(1-3)统计21残基片段中21个位点上20种氨基酸出现的频率，建立位置权重矩阵，是个20 × 21维的矩阵；同样，我们也可以统计20种氨基酸n-gap 2肽组分出现的频率，建立位置权重矩阵，当n = 1时是个400 × 20维的矩阵，当n = 2时是个400 × 19维的矩阵，当n = 3时是个400 × 18维的矩阵。这样基于8个集合，我们可以分别建立8个位置权重矩阵即{W_H}，{W_E}，{W_B}，{W_G}，{W_I}，{W_S}，{W_T}，{W_C}。对于一个中心残基待测的21残基片段，通过打分函数计算出S_H，S_E，S_B，S_G，S_I，S_S，S_T，S_C 8个打分值，找出最大的分值，从而得到预测的结果。采用10交叉检验，计算结果见表2。

Table 2. The predicting results of using scoring matrix
表 2. 使用矩阵打分的预测结果

	Q_H (%)	Q_E (%)	Q_B (%)	Q_G (%)	Q_I (%)	Q_S (%)	Q_T (%)	Q_C (%)	Q_8 (%)
位点氨基酸	71.7	60.2	0.21	10.2	0.01	9.8	39.5	60.0	55,7
n = 1	79.6	70.6	0.34	12.8	0.08	15.9	50.3	60.6	61.4
n = 2	74.5	69.7	0.12	13.1	0.03	13.9	43.5	58.3	58.2
n = 3	75.0	63.2	0.25	10.8	0.02	13.2	45.4	56.0	57.6

从表2中可以看出，当n = 1时，即以紧邻关联为参数，预测效果最好，8态预测总精度达到61.4%。我们发现，无论以位点氨基酸还是20种氨基酸n-gap 2肽组分为参数，都是H的预测精度比较好，n = 1时达到79.6%，其次是E、C，此外，I的预测精度几乎为零，原因是由于I太少，常常被错误的归类为H，这也和文献[11]是一致的。

本文尝试预测酶蛋白质中8类二级结构，是前人所没有研究过的。参考前人对各类蛋白质中二级结构的预测结果，我们的预测精度虽不及前人，但是我们的数据集更大，是我们后续研究工作的有利基础。而且本文首次将矩阵打分的方法用于酶蛋白质中8类二级结构的预测，计算简单，操作方便。

4. 结论

本文选取了3102个酶蛋白质，分别截取21个氨基酸残基片段，统计位点氨基酸以及20种氨基酸n-gap

2肽组分在8种二级结构序列中各个位点的位置权重矩阵，然后利用打分函数来预测，取得了比较好的预测效果。但是位点氨基酸及其*n*-gap 2肽组分还不足以提供足够的二级结构信息，因此，预测精度还有待提升。近几年的研究工作[20] [21] [22] [23] [24]表明：除了氨基酸序列信息外，影响其二级结构的形成还取决于其它因素，如残基的亲疏水性和当地环境，接触数，溶剂易访问性的残留物，蛋白质结构类，甚至受到不同物种的影响，所以今后的工作中也可以考虑整合这些信息，进一步提高酶蛋白质8类二级结构预测的精度。

基金项目

内蒙古工业大学科学研究项目(X201517)。

参考文献

[1] Chandonia, J.M. and Karplus, M. (1996) The Importance of Larger Datasets for Protein Secondary Structure Prediction with Neural Network. *Protein Science*, **5**, 768-774. https://doi.org/10.1002/pro.5560050422

[2] Anders, K. and Lareo, R. (1994) Hidden Markov Models in Computational Biology Applications to Protein Modeling. *Journal of Molecular Biology*, **235**, 1501-1531. https://doi.org/10.1006/jmbi.1994.1104

[3] Asai, K., Hayamizu, S. and Hands, K. (1993) Prediction of Protein Secondary Structure by the Hidden Markov Model. *Computer Applications in the Biosciences*, **9**, 141-146. https://doi.org/10.1093/bioinformatics/9.2.141

[4] Goldman, N., Thorne, J.L. and Jones, D.T. (1996) Using Evolutionary Trees in Protein Secondary Structure Prediction and Other Comparative Sequence Analyses. *Journal of Molecular Biology*, **263**, 196-208. https://doi.org/10.1006/jmbi.1996.0569

[5] Rost, B. and Sander, C. (1994) Combining Evolutionary Information and Neural Networks to Predict Protein Secondary Structure. *Proteins*, **19**, 55-72. https://doi.org/10.1002/prot.340190108

[6] Dor, O. and Zhou, Y.Q. (2007) Achieving 80% Ten-Fold Cross-Validated Accuracy for Secondary Structure Prediction by Large-Scale Training. *Protein*, **66**, 838-845. https://doi.org/10.1002/prot.21298

[7] Pollastri, G. and Mclysaght, A. (2005) Porter: A New, Accurate Server for Protein Secondary Structure Prediction. *Bioinformatics*, **21**, 2. https://doi.org/10.1093/bioinformatics/bti203

[8] Pollastri, G., Przybylski, D., Rost, B. and Baldi, P. (2002) Improving the Prediction of Protein Secondary Structure in Three and Eight Classes Using Recurrent Neural Networks and Profiles. *Proteins-Structure Function and Genetics*, **47**, 228-235. https://doi.org/10.1002/prot.10082

[9] Wang, Z.Y., Zhao, F., Peng, J. and Xu, J.B. (2011) Protein 8-Class Secondary Structure Prediction Using Conditional Neural Fields. *Proteomics*, **11**, 3786-3792. https://doi.org/10.1002/pmic.201100196

[10] Cong, P.S., Li, D.P., Wang, Z.H., Tang, S.N. and Li, T.H. (2013) SPSSM8: An Accurate Approach for Predicting Eight-State Secondary Structures of Proteins. *Biochimie*, **95**, 2460-2464. https://doi.org/10.1016/j.biochi.2013.09.007

[11] Yaseen, A. and Li, Y.H. (2014) Template-Based C8-SCORPION: A Protein 8-State Secondary Structure Prediction Method Using Structural Information and Context-Based Features. *BMC Bioinformatics*, **15**, 204-218. https://doi.org/10.1186/1471-2105-15-S8-S3

[12] 王志强, 董彩华, 王延枝. 温度和底物对大豆液泡膜 H^+-ATPase 二级结构的影响[J]. 生物物理学报, 2000, 16(3): 489-493.

[13] 王玮, 葛毅强, 陈颖, 徐幸莲, 周光宏. 超高压对高铁肌红蛋白还原酶活性及二级结构的影响[J]. 中国食品学报, 2015, 15(10): 134-140.

[14] Webb, E.C. (1992) Enzyme Nomenclature. Academic Press, San Diego.

[15] Kabsch, W. and Sander, C. (1983) Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen Bonded and Geometrical Features. *Biopolymers*, **22**, 2577-2637. https://doi.org/10.1002/bip.360221211

[16] Cartharius, K., Frech, K., Grote, K., *et al*. (2005) Mat Inspector and Beyond: Promoter Analysis Based on Transcription Factor Binding Sites. *Bioinformatics*, **21**, 2933-2942. https://doi.org/10.1093/bioinformatics/bti473

[17] Quandt, K., Frech, K., Karas, H., *et al*. (1995) MatIand and Mat Inspector: New Fast and Versatile Tools for Deteciono Consensus Matches in Nucleotide Sequence Data. *Nucleic Acids Research*, **23**, 4878-4884. https://doi.org/10.1093/nar/23.23.4878

[18] Kel, A.E., GoBling, E., Reuter, I., *et al*. (2003) MATCHTM: A Tool for Searching Transcription Factor Binding Sites

in DNA Sequences. *Nucleic Acids Research*, **31**, 3576-3579. https://doi.org/10.1093/nar/gkg585

[19] Wasserman, W.W. and Sandelin, A. (2004) Applied Bioinformatics for the Identification of Regulatory Elements. *Nature Reviews Genetics*, **5**, 276-287. https://doi.org/10.1038/nrg1315

[20] Zhong, L. and Johnson, W.C. (1992) Environment Affects Amino Acid Preference for Secondary Structure. *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 4462-4465. https://doi.org/10.1073/pnas.89.10.4462

[21] Lakizadeh, A. and Marashi, S.A. (2009) Addition of Contact Number Information Can Improve Protein Secondary Structure Prediction by Neural Networks. *EXCLI Journal*, **8**, 66-73.

[22] Macdonald, J.R. and Johnson, W.C. (2001) Environmental Features Are Important in Determining Protein Secondary Structure. *Protein Science*, **10**, 1172-1177. https://doi.org/10.1110/ps.420101

[23] Costantini, S., Colonna, G. and Facchiano, A.M. (2006) Amino Acid Propensities for Secondary Structures Are Influenced by the Protein Structural Class. *Biochemical and Biophysical Research Communication*, **342**, 441-451. https://doi.org/10.1016/j.bbrc.2006.01.159

[24] Marash, S.A., Behrouzi, R. and Pezehk, H. (2007) Adaptation of Proteins to Different Environments: A Comparison of Proteome Structural Properties in *Bacillus subtilis* and *Escherichia coli*. *Journal of Theoretical Biology*, **244**, 127-132. https://doi.org/10.1016/j.jtbi.2006.07.021