# 基于SVM的蛋白质二级结构预测

吴琳琳[1],徐硕[2*]

(1. 滨州医学院烟台校区基础学院,滨州 264003;2. 中国科学技术信息研究所信息技术支持中心,北京 100038)

**摘要**:蛋白质结构预测是现代计算生物领域最重要的问题之一,而蛋白质二级结构预测是蛋白质高级结构预测的基础。目前蛋白质二级结构的预测方法较多,其中SVM方法取得了较高的预测精度。重在阐述使用SVM用于蛋白质二级结构预测的步骤,以及与其他方法进行比较时应该注意的事项,为下一步的研究提供参考及启发。

**关键词**:支持向量机;蛋白质二级结构预测;非典型肺炎

**中图分类号**:Q71 **文献标识码**:A **文章编号**:1672-5565(2010)-03-187-04

## Protein secondary structure prediction based on SVM

WU Lin-lin[1], XU Shuo[2*]

(*1. College of Basic Sciences, Binzhou Medical University (Yantai Campus), Yantai 264003, China;*

*2. Information Technology Supporting Centre, Institute of Scientific and Technical Information of China, Beijing 100038, China*)

**Abstract**: Protein structure prediction is one of most important problems in modern computational biology, but protein secondary structure prediction is the foundation for high-level structure prediction. At the present time, there are many methods for protein secondary structure prediction, where one can obtain higher precision of prediction with SVM. The emphasis is put on how to apply SVM to protein secondary structure prediction, and some notes when comparing with other approaches, thus providing reference and inspiration for further study.

**Key Words**: Support Vector Machine (SVM); Protein Secondary Structure Prediction; Severe Acute Respiratory Syndrome (SARS)

蛋白质是执行生物功能的大分子,在生命活动中起着极其重要的作用。它的功能是由其空间结构决定的,结构与功能的这种一致性极大地促进了对蛋白质结构的研究。1973年Anfinsen的实验证明了蛋白质的氨基酸序列在失性后可自发恢复天然构象,这说明氨基酸序列决定了特定的空间结构[1]。通常,蛋白质结构包括4个层次[2]:一级结构即氨基酸的排列顺序;二级结构主要是由氢键维持的$\alpha$-螺旋和$\beta$-折叠;三级结构是完全折叠的蛋白质的空间结构残基的立体排列模式;四级结构是多个蛋白质亚基组成的蛋白质复合体的结构(即蛋白质之间的交互作用)。蛋白质二级结构预测为三级结构模型构建的起点,是三、四级结构预测的基础。

随着后基因组时代的来临,蛋白质结构预测的任务越来越迫切,开发有效的结构预测方法势在必行。目前蛋白质结构的预测方法较多,有些方法已取得了一些不错的效果。比较著名的算法包括:基于单残基构象性统计的Chou-Fasman方法[3-6],借用信息论方法并建立在统计基础之上的GOR方法[7-8],基于知识的人工神经网络(Artificial Neural Network,ANN)[9-12]和支持向量机(Support Vector Machine,SVM)[13-18]等。与其他方法相比,SVM方法取得了较高的预测精度。本文着重阐述使用SVM预测蛋白质二级结构的步骤,以及与其他方法比较时应注意的事项,为下一步的研究提供参考及启发。

## 1 支持向量机

支持向量机最初是由Vapnik等人[19-20]针对两类分类问题而提出的,它的基本思想是用少数支持向量代表整个样本集,本质上是通过某一事先选择好的非线性函数$\Phi(\bullet)$将输入空间的数据映射到一个高维(可能为无穷维)特征空间,然后在这个空间内按照结构风险最小化的原则构造一个最优分类面。

研究表明[19]-[20],输入空间的向量总是以特征空间像的点积形式出现,与特征空间的维数无关。

**收稿日期**:2009-07-11;修回日期:2009-11-23.

**作者简介**:吴琳琳,女,山东烟台人,讲师,硕士,研究方向:数据挖掘,生物信息。E-mail:linlinqer@gmail.com.

*通讯作者:徐硕,男,山东菏泽人,助理研究员,博士,研究方向:数据挖掘,生物信息。E-mail:xush@istic.ac.cn.

也就是说，只要能够找到某个函数 $K(\bullet,\bullet)$ 可以在输入空间完成相应特征空间的点积运算就可以了，即 $K(x_i, x_j) = <\Phi(x_i), \Phi(x_j)>$，其中 $x_i$、$x_j$ 为输入空间中的向量。已经证明满足 Mercer 定理的函数都具有这种功能[19]-[20]，这类函数有很多，被通称为核函数。目前比较常用的有四种：

(1)线性核：$K(x_i,x_j) = x_i^T x_j$；

(2)多项式核：$K(x_i,x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0$；

(3)径向基核(RBF 核)：$K(x_i,x_j) = \exp(-\gamma \| x_i - x_j \|^2), \gamma > 0$；

(4)Sigmoid 核：$K(x_i,x_j) = \tanh(\gamma x_i^T x_j + \lambda)$

本文选用 RBF 核，原因有三方面，(1)线性核是 RBF 核的一个特例[21]；(2)Sigmoid 核不是正定核，而且对于一定的参数取值，Sigmoid 核和 RBF 的性能相当[22]；(3)相对来说，多项式的参数比较多，而且它在运算时可能会溢出。

## 2 氨基酸编码方式

SVM 是一种数值型机器学习方法，为了将其用于蛋白质二级结构预测，需要事先将组成蛋白质的各氨基酸进行数值编码。当然编码方式不尽相同，如5 位编码法[23]、20 位编码法[9]、基本型密码子(codon)编码法[23]、扩展型密码子编码法[23]、构象倾向性打分编码法[17]以及 Profile 编码法[11][12][18]等。文献[23]讨论了常用的几种编码方法的优缺点，结果表明，用富含“生物进化信息”的 Profile 编码方法可以得到较高的预测结果，同时也表明，充分利用生物本身所具有的生物信息对提高蛋白质二级结构预测精度是非常重要的。

为描述方便起见，本文采用了一种简单的20 位编码(也称作正交编码)机制。但考虑到滑口窗口(滑动窗口的大小一般设为 5 - 15 间的奇数)滑到残基序列 N 端或 C 端时，会有位于 N 端或 C 端之外的空残基位于滑动窗口之中，故本文采用 21 位(20 种氨基酸残基 + 1 种空残基)编码机制，这也是许多蛋白质二级结构预测软件中常采用的处理方式。

## 3 二级结构分类方法

蛋白质二级结构的分类方法有很多种，包括根据蛋白质三维结构坐标中氢键的重复模式划分的 DSSP[24]、根据氢键和二面角统计分布划分的 STRIDE[25]以及根据结构坐标中原子间距离和标准的二级结构中原子距离的差来确定的 DEFINE[26]等方法，不同的分类方法对蛋白质二级结构的预测精度将造成不同程度的影响[27]。其中传统的 DSSP 将蛋白质的二级结构分为 8 种状态：H($\alpha$ - helix)、G(3 - helix or $3_{10}$ helix)、I(5 - helix or $\pi$ - helix)、E(extended strand participates in $\beta$ - ladder)、B(residue in isolated $\beta$ - bridge)、T(Hydrogen bond turn)、S(bend) 和 -(其他)。而这种 8 状态的二级结构是 3 状态(H($\alpha$ - 螺旋)、E($\beta$ - 折叠)、C(卷曲))的一种细化，将 8 状态的二级结构转化为 3 状态的二级结构一般采用如下几种形式[16]-[17]：(1)H、G、I→H，E→E，其他→C；(2)H、G→H，B、E→E，其他→C；(3)H、G→H，E→E，其他→C；(4)H→H，B、E→E，其他→C；(5)H→H，E→E，其他→C；

不难看出，在其他条件相同的条件下，采用简化方式(5)将会使蛋白质二级结构的预测精度偏高，而简化方式(2)则会使预测精度偏低。目前大多采用简化方式(1)或简化方式(2)，本文采用简化方式(1)。这样，蛋白质二级结构预测问题就可以看作是一个多类(三类)分类问题，本文采用了两种多类分类方法，分别为成对分类法(记为 OvO)，以及基于多类后验概率估计的分类方法(记为 WLW)[28]。

## 4 实验数据

实验数据采用 Severe Acute Respiratory Syndrome(SARS)数据集[29]，有关 SARS 数据集的信息摘要见表 1。其中列 H 含量及 E 含量中的括号分别代表各序列具有 $\alpha$ - 螺旋、$\beta$ - 折叠的残基数，列缺失结构残基数中的括号的三个数分别表示位于序列首部、中间和尾部的缺少结构信息的残基数，并且本文将结构信息缺失或不全的残基的构象都归为 C(卷曲)状态。

**表 1 SARS 数据集信息摘要**

Table 1 Information summary on SARS dataset

| 标号 | PDB ID | 序列长度 | H 含量/% | E 含量/% | 支链 | 缺失结构残基数 |
|---|---|---|---|---|---|---|
| 1 | 1P4X | 250 | 58.40(146) | 10.24(21) | A | 0 |
| 2 | 1P9S | 300 | 23.00(69) | 26.33(79) | A | 0 |
| 3 | 1Q2W | 308 | 24.35(75) | 25.97(80) | A | 11(4 + 4 + 5) |
| 4 | 1T4Y | 105 | 37.14(39) | 22.86(24) | A | 0 |
| 5 | 1UW7 | 143 | 9.79(14) | 35.66(51) | A | 21(21 + 0 + 0) |
| 6 | 1UJ1 | 306 | 26.14(80) | 27.45(84) | A | 5(2 + 0 + 3) |
| 7 | 1T4Z | 105 | 32.38(34) | 22.86(24) | A | 0 |
| 8 | 1XAK | 83 | 0.00(0) | 48.19(40) | A | 15(1 + 0 + 14) |

需要说明的是，与文献[29]中的信息摘要有点出入，经仔细分析发现，文献[29]所考虑的结构信息完全取自 PDB 数据文件中所记录的结构信息，而这些结构信息是数据提供者提供的，数据提供者给

出的结构信息往往不完整，有的误差较大、不准确[2,30]。

将前五条用做训练数据（残基个数：1106，H 含量：31.01%，E 含量：23.06%），后三条用做测试数据（残基个数：494，H 含量：23.08%，E 含量：29.96%）。

## 5 结果及分析

### 5.1 评价指标

评价蛋白质二级结构的预测性能，本文选用了国际上通用的两个指标如下[31]：

（1）三态准确率

$$Q_i = \frac{TP_i}{TP_i + FP_i}, i \in \{H, E, C\} \quad (1)$$

其中 $TP_i$ 表示被正确预测为 $i$ 状态的残基个数，$FP_i$ 表示被错误预测为 $i$ 状态的残基个数。

（2）整体准确率

$$Q_3 = \frac{TP_H + TP_E + TP_C}{T} \quad (2)$$

其中 $TP_i(i \in \{H, E, C\})$ 分别表示被正确预测出三状态的残基个数，$T$ 代表残基总数。

### 5.2 参数选择

（1）最优窗口大小

可以将窗口大小分别设为 5，7，9，11，13，15；然后通过训练集的 $k$ - 折交叉验证精度来确定，不过本文将窗口大小设为典型值 13。

（2）核参数及惩罚参数 $C$

到目前为止，SVM 的参数选择问题仍然没有从理论上得到很好的解决，但在实际应用中，一般都采取试探法，格网搜索法（Grid Search）或启发式的选参方法等。本文采用格网搜索选参法，参数 $\gamma$ 和 $C$ 分别以指数增长、减少的序列进行搜索，即 $\gamma$ 和 $C$ 分别取下列各组值：

$\gamma = 2^{-15}, 2^{-13}, \mathrm{L}, 2^3, C = 2^{-5}, 2^{-3}, \mathrm{L}, 2^{15}$

根据整体准确率 $Q_3$ 最大的原则采用 5 - 折交叉验证来确定合适的参数值，如表 2 所示。

**表 2　采用格网搜索法选得的最优参数**

Table 2 Optimal parameters selected following grid search

| 多类分类方法 | C | $\gamma$ | $Q_3$ |
|---|---|---|---|
| OvO | $2^3$ | $2^{-3}$ | 62.12 % |
| WLW | 2 | $2^{-3}$ | 63.56 % |

### 5.3 结果分析

根据上面选得的参数，可训练得到一个 SVM 多类分类模型（本文采用的 SVM 软件为：LibSVM - 2.82[32]），然后对测试集进行二级结构预测，实验结果如表 3 所示：

**表 3　两种多类分类方法的实验结果**

Table 3 Experimental results of two multi-classification classifiers

| 多类分类方法 | $Q_3$ | $Q_H$ | $Q_E$ | $Q_C$ | nSV |
|---|---|---|---|---|---|
| OvO | 89.27 % (441/494) | 94.74 % (108/114) | 75.68 % (112/148) | 95.26 % (221/232) | 1106 |
| WLW | 89.47 % (442/494) | 94.74 % (108/114) | 81.08% (120/148) | 92.24 % (214/232) | 1106 |

从表 3 不难看出，无论是三态准确率还是整体准确率均高于文献[29]，但这并不能说明本文得到 SVM 模型较好，因为它的 5 - 折交叉验证的精度较低（见表 2）。之所以出现这种情况，本文认为原因可能有两个方面：（1）测试数据太少，难分的残基样本可能很少；（2）参数选择对 SVM 的预测精度有很大影响，而以往的研究大都是凭经验选参，难以选到最优或近似最优的参数。

WLW 的分类效果要略好于 OvO，这与文献[33]的观测是一致的。WLW 牺牲了 C 态的预测精度，但提高了 E 态的预测精度。本文认为这是值得的。因为长程效应的影响，使得 E 态的预测精度一直难以提高，相对来说 C 态的预测精度提高更容易一些。至于 WLW 是否真能提高 E 态的预测精度，尚需在大规模数据上加以验证。

从表 3 还可以看出，支持向量的个数（nSV）等于训练样本数，这表明 SVM 模型并未起到信息压缩的作用，同时也表明蛋白质二级结构预测是一个比较难的问题，需要增加训练数据的数量。

## 6 结论

在对各种蛋白质二级结构预测方法进行比较时，不应将注意力只放在质量评估指标上。要注意比较的公平性，也就是说在其他条件相同的条件下才具有可比性。这些条件包括：二级结构分类方法、8 状态到 3 状态的简化方法、数据集等等。本文没有列出其他预测方法所得结果，主要原因是本文对其他工作的先决条件并不完全了解，需进一步调研。

在各种氨基酸残基的编码方法之中，基于序列比对的 Profile 编码法是目前最有效的方法之一。本文希望引入这种编码法的同时，考虑更多有关各残基的物理 - 化学属性。根据上文的分析，本文认为 SARS 数据集太小，可能不足以学习到蛋白质二级结构的本质，所以下一步将考虑利用数据集 RS126[10] 及 CB513[34]。

研究表明[34]，组合多类预测方法可以提高蛋白

质二级结构的预测精度,它包括同质或异质预测方式的组合。如果要组合多种 SVM 预测模型,那么最多需要多少个 SVM 模型尚需进一步研究。也就是说 SVM 模型的个数有没有一个上界,使得超过这个上界后最后得到的预测精度将会下降。如果存在这么一个上界,那么它是多少;如果不存在,那么如何同时考虑无穷多个这样的模型。

目前,几乎所有的基于知识的预测方法都是利用滑动窗口中的残基特征预测中间的残基构象,如果能同时预测中间几个残基(比如三个)的构象可能会好些。文献[35]利用 ANN 做了这方面的尝试,验证了结论的正确性。

**参考文献**(References):


[1] ZOU C L. The Second Genetic Code[J]. Science Bulletin, 2000, 45(16): 1681 - 1687.

[2] 阎隆飞,孙之荣. 蛋白质分子结构[M]. 北京: 清华大学出版社, 1999.

[3] CHOU P Y, FASMAN G D. Conformational Parameters for Amino Acids in Helical, Beta - sheet, and Random Coil Regions Calculated from Proteins[J]. Biochemistry, 1974, 13(2): 211 - 222.

[4] CHOU P Y, FASMAN G D. Prediction of Protein Conformation [J]. Biochemistry, 1974, 13(2): 222 - 245.

[5] CHOU P Y, FASMAN G D. Prediction of the Secondary Structure of Proteins from their Amino Acid Sequence[J]. Advances in Enzymology and related Areas of Molecular Biology, 1978, 47: 45 - 148.

[6] CHOU P Y, FASMAN G D. Empirical Predictions of Protein Conformation[J]. Annuals Review of Biochemistry, 1978, 47: 251 - 276.

[7] GARNIER J, OSGUTHORPE D J, ROBSON B. Analysis of the Accuracy and Implications of Simple Methods for Predicting the Secondary Structure of Globular Proteins[J]. Journal of Molecular Biology, 1978, 120(1): 97 - 120.

[8] GIBRAT J F, GARNIER J, ROBSON B. Further Developments of Protein Secondary Structure Prediction using Information Theory: New Parameters and Consideration of Residue Pairs[J]. Journal of Molecular Biology, 1987, 198(3): 425 - 443.

[9] QIAN N, SEJNOWSKI T J. Predicting the Secondary Structure of Globular Proteins using Neural Network Models[J]. Journal of Molecular Biology, 1988, 202: 865 - 884.

[10] ROST B, SANDER C. Prediction of Protein Secondary Structure at Better than 70% Accuracy[J]. Journal of Molecular Biology, 1993, 232: 584 - 599.

[11] RIIS S K, KROGH A. Improving Prediction of Protein Secondary Structure using Structured Neural Networks and Multiple Sequence Alignments[J]. Journal of Computational Biology, 1996, 3: 163 - 183.

[12] JONES D T. Protein Secondary Structure Prediction based on Position - specific Scoring Matrices[J]. Journal of Molecular Biology, 1999, 292(2): 195 - 202.

[13] HUAN S J, SUN Z R. A Novel Method of Protein Secondary Structure Prediction with High Segment Overlap Measure: Support Vector Machine Approach. Journal of Molecular Biology, 2001, 308(2): 397 - 407.

[14] WANG J Y. Application of Support Vector Machines in Bioinformatics[D]. Master's thesis, Department of Computer Science and Information Engineering, National Taiwan University. 2002.

[15] NGUYEN M N, RAJAPKSE J C. Multi - class Support Vector Machines for Protein Secondary Structure Prediction[J]. Genome Informatics, 2003, 14: 218 - 227.

[16] KIM H, PARK H. Protein Secondary Structure Prediction based on an Improved Support Vector Machines Approach[J]. Protein Engineering, 2003, 16(8): 553 - 560.

[17] WANG L H, LIU J, LI Y F, et al. Predicting Protein Secondary Structure by a Support Vector Machine based on a new Coding Scheme[J]. Genome Informatics, 2004, 15(2): 181 - 190.

[18] GUO J, CHEN H, SUN Z R, et al. A Novel Method for Protein Secondary Structure Prediction using Dual - Layer SVM and Profiles[J]. Proteins: Structure, Function, and Bioinformatics, 2004, 54: 738 - 743.

[19] VAPNIK V N. The Nature of Statistical Learning Theory[M]. 2nd Edition. New York: Springer Verlag, 1999.

[20] VAPNIK V N. Statistical Learning Theory[M]. New York: Wiley, 1998. 许建华,张学工 译. 统计学习理论[M]. 北京: 电子工业出版社, 2004.

[21] KEERTHI S S, LIN C J. Asymptotic Behaviors of Support Vector Machines with Gaussian Kernel[J]. Neural Computation, 2003, 15(7): 1667 - 1689.

[22] LIN H T, LIN C J. A Study on Sigmoid Kernels for SVM and the Training of non - PSD Kernels by SMO - type Methods[R]. Technical Report, Department of Computer Science, National Taiwan University, 2003.

[23] 阮晓钢,孙海军. 编码方式对蛋白质二级结构预测精度的影响[J]. 北京工业大学学报,2005, 31(3): 230 - 235.

[24] KABSCH W, SANDER C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen - based and Geometrical Features[J]. Biopolymers, 1983, 22(12): 2577 - 2637.

[25] FRISHMAN D, ARGOS P. Knowledge - based protein secondary structure assignment[J]. Proteins: Structure, Function, and Genetics, 1995, 23(4): 566 - 579.

[26] RICHARDS F M, KUNDROT C E. Identification of Structure Motifs from Protein Coordinate Data: Secondary Structure and First - level Supersecondary Structure[J]. Proteins, 1988, 3(2): 71 - 84.

[27] CUFF J A, BARTON G J. Evaluation and Improvement of Multiple Sequence Methods for Protein Secondary Structure Prediction [J]. Proteins: Structure, Function, and Genetics, 1999, 34: 508 - 519.

[28] WU T F, LIN C J, WENG R C. Probability Estimates for Multi - class Classification by Pairwise Coupling[J]. Journal of Machine Learning Research, 2004, 5: 975 - 1005.

[29] 李元乐,陶兰. 基于小波核支持向量机的蛋白质二级结构预测[J]. 深圳大学学报: 理工版, 2006, 23(2): 117 - 121.

[30] PDB Format. Available [online]: http://www.rcsb.org/pdb/file_formats/pdb/pdbguide2.2/guide2.2_frame.html.

[31] 张海霞,唐焕文,张立震,等. 蛋白质二级结构预测方法的评价[J]. 计算机与应用化学, 2003, 20(6): 735 - 740.

[32] LibSVM - 2.82. Available [online]: http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[33] DUAN K, KEERTHI S S. Which is the Best Multiclass SVM Method? An Empirical Study[R]. Technical Report CD - 03 - 12, Control Division, Department of Mechanical Engineering, National University of Singapore, 2003.

[34] CUFF J A, BARTON G J. Evaluation and Improvement of Multiple Sequence Methods for Protein Secondary Structure Prediction [J]. Proteins: Structure, Function, and Genetics, 1999, 34: 508 - 519.

[35] PETERSEN T N, LUNDEGAARD C, NIELSEN M, et al. Prediction of Protein Secondary Structure at 80% Accuracy[J]. Proteins: Structure, Function, and Geneties, 2000, 41: 17 - 20.

# 基于SVM的蛋白质二级结构预测



作者： 吴琳琳， 徐硕， WU Lin-lin， XU Shuo
作者单位： 吴琳琳,WU Lin-lin(滨州医学院烟台校区基础学院,滨州,264003)， 徐硕,XU Shuo(中国科学技术信息研究所信息技术支持中心,北京,100038)

刊名： 生物信息学ISTIC
英文刊名： CHINA JOURNAL OF BIOINFORMATICS
年，卷(期)： 2010，08(3)



## 参考文献(36条)

1. LIN H T.LIN C J A Study on Sigmoid Kernels for SVM and the Training of non-PSD Kernels by SMO-type Methods 2003

2. 阮晓钢.孙海军 编码方式对蛋白质二级结构预测精度的影响 2005(3)

3. KABSCH W.SANDER C Dictionary of Protein Secondary Structure:Pattern Recognition of Hydrogen-based and Geometrical Features 1983(12)

4. FRISHMAN D.ARGOS P Knowledge-based protein secondary structure assignment 1995(4)

5. RICHARDS F M.KUNDROT C E Identification of Structure Motifs from Protein Coordinate Data:Secondary Structure and First-level Supersecondary Structure 1988(2)

6. CUFF J A.BARTON G J Evaluation and Improvement of Multiple Sequence Methods for Protein Secondary Structure Prediction 1999

7. WU T F.LIN C J.WENG R C Probability Estimates for Multi-class Classification by Pairwise Coupling 2004

8. 李元乐.陶兰 基于小波核支持向量机的蛋白质二级结构预测 2006(2)

9. PDB Format

10. 张海霞.唐焕文.张立震.靳利霞.唐一源 蛋白质二级结构预测方法的评价 2003(6)

11. LibSVM-2.82

12. DUAN K.KEERTHI S S Which is the Best Multiclass SVM Method? An Empirical Study[Technical Report CD-03-12] 2003

13. CUFF J A.BARTON G J Evaluation and Improvement of Multiple Sequence Methods for Protein Secondary Structure Prediction 1999

14. PETERSEN T N.LUNDEGAARD C.NIELSEN M Prediction of Protein Secondary Structure at 80% Accuracy 2000

15. ZOU C L The Second Genetic Code 2000(16)

16. 阎隆飞.孙之荣 蛋白质分子结构 1999

17. CHOU P Y.FASMAN G D Conformational Parameters for Amino Acids in Helical,Beta-sheet,and Random Coil Regions Calculated from Proteins 1974(2)

18. CHOU P Y.FASMAN G D Prediction of Protein Conformation 1974(2)

19. CHOU P Y.FASMAN G D Prediction of the Secondary Structure of Proteins from their Amino Acid Sequence 1978

20. CHOU P Y.FASMAN G D Empirical Predictions of Protein Conformation 1978

21. GARNIER J. OSGUTHORPE D J. ROBSON B Analysis of the Accuracy and Implications of Simple Methods for Predicting the Secondary Structure of Globular Proteins 1978(1)

22. GIBRAT J F. GARNIER J. ROBSON B Further Developments of Protein Secondary Structure Prediction using Information Theory:New Parameters and Consideration of Residue Pairs 1987(3)

23. QIAN N. SEJNOWSKI T J Predicting the Secondary Structure of Globular Proteins using Neural Network Models 1988

24. ROST B. SANDER C Prediction of Protein Secondary Structure at Better than 70% Accuracy 1993

25. RIIS S K. KROGH A Improving Prediction of Protein Secondary Structure using Structured Neural Networks and Multiple Sequence Alignments 1996

26. JONES D T Protein Secondary Structure Prediction based on Position-specific Scoring Matrices 1999(2)

27. HUAN S J. SUN Z R A Novel Method of Protein Secondary Structure Prediction with High Segment Overlap Measure:Support Vector Machine Approach 2001(2)

28. WANG J Y Application of Support Vector Machines in Bioinformatics 2002

29. NGUYEN M N. RAJAPKSE J C Multi-class Support Vector Machines for Protein Secondary Structure Prediction 2003

30. KIM H. PARK H Protein Secondary Structure Prediction based on an Improved Support Vector Machines Approach 2003(8)

31. WANG L H. LIU J. LI Y F Predicting Protein Secondary Structure by a Support Vector Machine based on a new Coding Scheme 2004(2)

32. GUO J. CHEN H. SUN Z R A Novel Method for Protein Secondary Structure Prediction using Dual-Layer SVM and Profiles 2004

33. VAPNIK V N The Nature of Statistical Learning Theory 1999

34. VAPNIK V N Statistical Learning Theory 1998

35. 许建华. 张学工 统计学习理论 2004

36. KEERTHI S S. LIN C J Asymptotic Behaviors of Support Vector Machines with Gaussian Kernel 2003(7)

## 相似文献(10条)

1. 期刊论文 李元乐. 陶兰. LI Yuan-le. TAO Lan 基于小波核支持向量机的蛋白质二级结构预测 -深圳大学学报（理工版） 2006, 23(2)

提出一种基于小波核支持向量机分类模型,将其用于SARS蛋白质二级结构预测.实验表明,该模型与其他同类方法相比,提高蛋白质二级结构预测的准确度达到1%～2%.

2. 学位论文 李元乐 基于SVM的蛋白质二级结构预测研究 2006

为了达到对蛋白质功能的深入了解，现代生物学迫切需要快速的方法来获得蛋白质空间结构，如二级结构、三级结构等信息。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，是蛋白质三、四级结构预测的基础。因此，蛋白质二级结构预测是生物信息学研究的重要课题之一。预测蛋白质分子的三维空间结构，需要提高蛋白质二级结构预测的准确度。

本文研究了基于知识的蛋白质二级结构预测中的两种重要方法人工神经网络方法和支持向量机方法。对使用这两种方法预测蛋白质二级结构进行了分析，从编码方式、输入处理、参数调节等方面进行了比较，最后比较了两种方法预测的结果。

根据支持向量机和小波的特点，本文提出一种基于小波核的支持向量机分类模型，并使用该模型对蛋白质二级结构进行预测。该方法充分利用蛋白质序列编码后稀疏变化特性，通过小波的多分辨率特点，得到序列的不同尺度信息。实验表明，该方法可以提高二级结构预测的准确率，取得较好的效果。

通过进一步分析蛋白质二级结构预测的知识，发现二级结构和序列之间存在一定的关系。序列的比对可以体现蛋白质序列的进化信息，以及特定残基替换模式和长程信息。本文提出一种基于BLOSUM62矩阵进行序列比对，然后采用最近邻方法修剪学习样本集的方法。使用支持向量机对修剪后的样本集进行学习，将其用于蛋白质二级结构预测。实验结果表明，使用这种方法预测蛋白质二级结构的准确率平均值为78.58%，分类时间明显减少。

3. 学位论文 张海霞 蛋白质二级结构预测方法研究 2004

随着人类基因组计划的完成,人们已经获得了大量生物的遗传信息,数以万计的蛋白质序列也已经被测出,到2004年4月13日为止SWISS-PROT数据库中

总共收集了148516条已被测序的蛋白质序列.然而一条蛋白质序列必须折叠成一定的空间结构时才能发挥它特定的生物功能,人们在对蛋白质序列测序完成之后更希望的是得到这些蛋白质的空间结构,以便发现结构与功能之间的联系.因此,蛋白质结构和功能的研究就成为了后基因组时代生命科学领域人们研究的主要任务和目的.目前,通过实验的方法获得的蛋白质结构序列只有两万多条(2004年4月20日,PDB数据库中共收集了25176条),远远落后于蛋白质序列的测序速度,因此理论预测蛋白质结构势在必行.然而,直接从蛋白质一级序列预测其三维空间结构时人们又遇到了诸多困难.在对蛋白质分子的仔细研究和分析后发现由二级结构组装而成的空间结构是有限的.因此,如果能从蛋白质一级序列先预测出二级结构,再由二级结构预测三级结构便成为一条有效的途径.这里,蛋白质二级结构预测不仅成为联系蛋白质一级序列和三级结构的纽带,而且也是从一级序列预测其三维空间结构的关键步骤.该文的主要工作是蛋白质二级结构预测方法的研究.

4.期刊论文 黄振 支持向量机用于蛋白质二级结构预测 -世界华商经济年鉴·高校教育研究2008,""(5)

本文提出一种蛋白质二级结构预测的方法.该方法首先对数据集中的氨基酸序列采用疏水性编码,然后采用支持向量机算法,利用滑动窗口方法对二级结构进行预测.采用分类器集成,将所有的样本进行训练,对蛋白质的三种二级结构(H/E/C)进行识别.

5.学位论文 刘欣阳 基于支持向量机的蛋白质二级结构预测 2004

支持向量机对于小数据集的训练预测效果比较好,而且支持向量机适合解决线性不可分的情况.因此用支持向量机进行蛋白质二级结构预测是一个比较有发展的方法.蛋白质二级结构预测中比较重要的一个技术就是编码.神经网络预测系统PHD中提出采用中心编码技术.但是PHD方法的实际预测结构可能出现比较明显的错误,也就是现实世界中不存在的结构,即单残基螺旋片断.出现这种情况的原因是由于中心编码只是考虑单残基结构.针对这一问题,我们提出了一种新的编码方法——片断编码.片断编码不是以中心残基作为编码对象,而是以整个二级结构片断作为编码对象.根据我们采用的片断编码方法,我们又提出了一种片断积分预测方法,实验结果表明片断积分方法得到的预测结果并不理想,但是片断积分的方法得到的得分矩阵是一个很有用的结构信息.因此我们提出了将片断编码和中心编码相结合的后处理方法.我们提出的后处理方法的主要思想是将片断编码预测的得分矩阵作为结构倾向因子.主要的预测工作由中心编码完成,我们用结构倾向因子来纠正预测中出现的显著错误.实验表明这种方法可以提高预测的正确率.

6.期刊论文 王宝文.王水星.刘文远.于家新.WANG bao-wen.WANG shui-xing.LIU wen-yuan.YU jia-xin 结合支持向量机和贝叶斯方法进行蛋白质二级结构预测 -生物信息学2010,8(1)

组建一个分两个阶段的分类器来进行蛋白质二级结构预测.第一阶段由支持向量机分类器组成,在第二阶段中使用第一阶段已预测的结果来进行贝叶斯判别.预测性能的改进表明了结合支持向量机和贝叶斯方法预测性能优越于单独使用支持向量机的预测性能.同时也证明残基在形成二级结构时是相互影响的.

7.学位论文 闫蓬勃 蛋白质二级结构预测准确率影响因素探讨 2009

从蛋白质的一级序列得到其对应的三维结构是目前生物信息学领域重要的课题之一。计算机预测方法被广泛应用于蛋白质二级结构的研究，其发展过程大体分为两个阶段：第一个阶段以数理统计作为出发点，基于单个氨基酸信息，如Chou-Fasman和GOR（Garnier-Osguthorpe-Robson）方法；第二个阶段基于进化信息，主要利用BLAST等工具在序列数据库中对搜索序列进行多重比对以取得同源信息PSSM（特异位点打分矩阵）利用PSI-BLAST取得相应的进化信息PSSM。本实验致力于氨基酸特性对基于PSSM预测方法的改进和预测准确率的提高。<br>
以SVM（支持向量机）作为实现手段，在PSSM基础上分别添加疏水因子和HEC（螺旋、折叠、无规则卷曲）倾向性两种理化因子作为单个氨基酸的特征值对蛋白质二级结构进行预测。本实验还同时设计对SVM使用进行改进方法实现双层SVM，即通过理化因子和双层SVM工具两种方法共同达到提高蛋白质二级结构预测准确率的目的。实验结果经相关系数分析表明，添加的疏水因子和HEC倾向性对Q3微弱正相关，与SOV值显著正相关。它证明氨基酸的疏水性与HEC倾向性对蛋白质二级结构的形成起到一定作用。通过双层SVM实验，无论是准确率的绝对值还是相关系数分析，双层网络都在二级结构预测的准确率上占有优势，改进的SVM对其预测过程起到明显的优化作用。预测的准确率的Q3值和SOV比目前国际常用的PSSM方法分别提高了2.76%和1.25%。

8.学位论文 李琳 基于RBF核的SVM学习算法优化及其在蛋白质二级结构预测中的应用 2006

SVM方法的核函数及其参数的选择，仍没有形成一个统一的模式。针对此现状，本文分析了现有的网格搜索法(GridSearchMethod，GSM)和双线性搜索法(BilinearSearchMethod，BSM)，并对GSM和BSM进行了改进，提出了一种新的基于RBF核的SVM学习算法的参数优化方法——双线性网格搜索法(BilinearGridSearchMethod，BGSM)。BGSM方法结合了BSM和GSM方法的优点，设计出了一种更为有效的学习算法，该方法在保持分类精确率不降低的情况下，大幅度减少了SVM的训练量。

BGSM方法用于UCI数据的验证，实验结果表明，BGSM方法比相关的方法具有更好的学习效率和较高的学习精确率。BGSM方法应用到蛋白质二级结构预测领域也得到了比相关算法更好的学习精确率。采用了滑动窗口技术和多重序列比对方法对氨基酸残基进行编码。通过比对未知蛋白质序列和已知蛋白质序列的相似程度，以判定二者之间是否具有同源性，从而获得蛋白质进化过程的信息。

本文通过改进GSM和BSM方法，设计和实现了一种以RBF为核的SVM优化算法，通过对学习方法和学习策略等方面的改进，使得BGSM具有比相关方法更好的学习性能和较高的学习精确率。

9.期刊论文 孙向东.韦柳静.黄日波 蛋白质二级结构预测的支持向量机模型研究 -广西农业生物科学2004,23(1)

蛋白质序列种类繁多,形态、结构和功能极其丰富多样,而已知三维结构的蛋白质数量却非常有限.为了能够利用有限的已知结构模拟和预测未知蛋白的结构,小样本统计学技术非常重要.本文主要探讨统计学习理论应用于蛋白质二级结构预测的模型,说明支持向量机(SVMs)进行蛋白质二级结构模式识别的主要步骤,讨论利用SVMs预测蛋白质二级结构优化过程中样本空间的确定方法,研究氨基酸序列向量化途径以及分析如何选择映射输入空间到特征空间的核函数等问题.

10.学位论文 邹东升 蛋白质超二级结构预测研究 2009

后基因组时代生命科学中最重大的研究课题之一是蛋白质组研究，蛋白质结构预测正是蛋白质组研究中一个富有挑战性的研究课题，其研究不仅对于理解蛋白质空间折叠机制与蛋白质功能具有理论价值，更对生物制药、农业生物科技等应用领域具有直接的指导作用。蛋白质的三维空间结构与其功能紧密相关，而超二级结构正是构成三维结构的基本单元。从蛋白质一级结构直接预测三维空间结构非常困难，蛋白质二级结构及超二级结构正是两者直接之间的重要桥梁，因此超二级结构的预测有着重要的研究意义。<br>
现有的许多研究主要是针对蛋白质二级结构预测，超二级结构预测的相关研究还比较少。作为超二结构预测的基础和前提，二级结构预测是不可缺少的环节。氨基酸的编码方式对蛋白质二级结构预测精度有重要影响，因此有必要对氨基酸编码方式进行分析比较，为二级结构预测编码方式的选择提供直接依据；目前较少的超二级结构预测研究在特征表达上有缺陷，仅仅考虑氨基酸基本组成成份，特征信息表达不完整；同时在超二级结构分类方法上也有待进一步探索。<br>
本文应用机器学习技术对蛋白质超二级结构预测问题进行深入研究：本文首先对二级结构预测的氨基酸编码问题进行研究；然后对蛋白质结构中频繁出现的一种特殊超二级结构(β发夹)进行预测研究；最后将特殊超二级结构研究进一步推广到一般超二级结构的预测研究。论文取得的主要成果与创新工作概括如下：<br>
①研究分析了不同的氨基酸编码方式对使用支持向量机进行蛋白质二级结构预测精度的影响。蛋白质二级结构预测采用何种氨基酸编码方式会对预测精度有很大影响。选择具有较好的分类能力的支持向量机进行蛋白质二级结构预测。建立二级结构预测模型，分析比较正交编码、5位编码、Codon编码(基本)、Codon编码(扩展)和Profile编码等5种氨基酸编码方案以及不同的支持向量机核函数对二级结构预测精度的影响。实验数据表明：使用支持向量机进行蛋白质二级结构预测时，经过多重序列比对、包含更多生物进化信息的Profile编码方式的预测精度明显优于其他4种编码方式。<br>
②提出一种新的β发夹特征表达方法。用离散量及离散增量表征蛋白质β-β模体的信息。用氨基酸基本组成成份，二肽成份以及氨基酸组成分布三种方式表达β-β模体特征。每个β-β模体表达成一个18维的特征向量，用作分类器的输入。实验数据集选择ArchDB40数据库(3088个蛋白质)、Kumar数据库(2088个蛋白质)、CASP6数据集(63个蛋白质)。将支持向量机用于β发夹的预测分类器，取得了较高的预测精度。<br>
③使用提出的β发夹特征表达方法，首次将离散增量结合二次判别分析方法用于β发夹的预测。在ArchDB40数据集、Kumar数据集、CASP6数据集上均取得较高的的预测精度。上述工作充分说明：本文提出的新的β发夹特征表达方法是有效的。<br>
④将特殊超二级结构特征表达策略进一步推广到一般超二级结构特征表达。用离散量及离散增量表达一般超二级结构序列的表征信息。用氨基酸基本组

成成份，二肽成份以及氨基酸组成分布三种方式表达一般超二级结构特征。每个超二级结构序列表达成一个36维的特征向量，用作分类器的输入。实验数据集选择ArchDB40数据库中9180个β－β模体、5737个β－α模体、6378个α－β模体、4176个α－α模体。将支持向量机用于超二级结构的预测，在训练集及独立测试集均取得较高的预测精度。<br>
⑤首次将二次判别分析方法用于一般超二级结构的预测。使用相同的数据集，在训练集及独立测试集上均获得较高的预测精度。上述工作充分说明：特殊超二级结构特征表达策略进一步推广到一般超二级结构特征表达是有效的。