山东农业大学学报(自然科学版),2014,45(增刊):27-32 VOL.45 S 2014
Journal of Shandong Agricultural University (Natural Science Edition) doi:10.3969/j.issn.1000-2324.2014.z.006

基于词频统计编码和 SVM 的蛋白质二级结构预测方法

石陆魁*, 刘倩倩, 王靖鑫, 张 军
河北工业大学计算机科学与软件学院, 天津 300401

摘 要: 在蛋白质二级结构预测中，常用的氨基酸序列编码方法产生的编码除了具有较高的维数外，也没有利用氨基酸序列片段中的统计信息。为此，提出了一种新的氨基酸序列编码方法——基于词频统计的编码方法，该方法统计每个氨基酸在氨基酸序列片段中出现的频率，利用该编码方法对氨基酸序列片段编码后得到一个 20 维的向量。与其它编码方法相比不但具有较低的维数，而且也充分利用了氨基酸片段内部所有氨基酸对目标氨基酸的影响。在实验中比较了四种编码方法结合支持向量机和 BP 神经网络的预测结果，实验结果表明，通过结合词频统计编码和支持向量机来预测蛋白质二级结构极大地提高了预测精度，远优于其它方法的预测结果。

关键词: 词频统计编码;支持向量机;蛋白质二级结构预测;滑动窗口法
中图法分类号： TP 181 **文献标识码：** A **文章编号：** 1000-2324(2014)S-0027-06

Prediction Method for Protein Secondary Structure Based on Word Frequency Statistics Coding and SVM

SHI Lu-kui, LIU Qian-qian, WANG Jing-xin, ZHANG Jun
School of Computer Science and Engineering, Hebei University of Technology, Tianjin 300401, *China*

Abstract: In protein secondary structure prediction, the codes from the existing amino acid coding methods have higher dimension. And these coding methods don't also use the statistic information in the amino acid sequence. To do that, a new coding method based on word frequency statistics was presented, which counted the frequency of each amino acid emerging in amino acids sequence. A 20 dimensional vector was obtained after coding the amino acid sequence with the new coding method. In contrast to other the coding methods, the codes from the new coding method have lower dimension and fully utilize all information in the amino acid sequence. In experiments, we compared the methods combing different coding methods and SVM with BP neural network. Experiment results show that the method combing word frequency statistics coding method and SVM greatly improve the prediction accuracy of protein secondary structure and is superior to other methods.

Keywords: Word frequency statistic coding;support vector machine; protein secondary structure prediction;sliding window method

1 引 言

蛋白质二级结构的预测有助于了解蛋白质的一些基本性质、功能及其作用机制，对于理解蛋白质结构与功能的关系具有重要的意义。生物学家们认为对蛋白质二级结构预测研究能够达到 80%的准确率，就基本可以确定这个蛋白质的三维空间构象，从而能够更加准确地分析蛋白质的物化特性[1]。同时从二级结构的预测到蛋白质结构的确定又是进行蛋白质复制、突变体设计以及基于蛋白质结构的药物设计的基础。因此，蛋白质二级结构预测不仅是蛋白质一级结构和三级构象的重要枢纽，更是从一级结构预测其三维空间构象的至关重要的一步。蛋白质二级结构预测已经成为后基因组时代的一个重要研究课题。

目前，国外涌现出了许多蛋白质二级结构预测方法和软件，国内对于蛋白质结构和功能的分析与预测也在众多国外先进技术的引领下，取得了非常大的研究进展。蛋白质二级结构预测从一开始的单残基分析时期到以滑动窗口获取到的氨基酸片段分析，再到目前经常应用的多重序列比对，取得了很大的进展[2-13]。

统计建模方法是常用的一类蛋白质二级结构预测方法，经典的统计建模预测方法有 Chou & Fasman[14]、GOR[15]等。最近邻居法也属于一种统计建模方法，通过寻找与目标蛋白质结构相近的氨

基金项目: 河北省自然科学基金项目(F2013202104)
作者简介: 石陆魁(1974-),男, 副教授, 研究方向为机器学习、数据挖掘.E-mail:shilukui@scse.hebut.edu.cn
***通讯作者:** Author for correspondence. E-mail: shilukui@scse.hebut.edu.cn

基酸，确定其与目标蛋白质具有相同的结构。目前应用较多的人工神经网络，隐形马尔科夫模型（Hidden Markov Model, HMM）以及支持向量机（Support Vector Machine，SVM）等方法均属于统计建模方法的范畴[9]，由于利用统计法建模在算法设计上并不很复杂，因而所得到的预测精度并不高。对统计建模方法来说，氨基酸序列的编码方法是影响预测精度的重要因素之一。在统计建模法中，在编码时都是利用滑动窗口法将氨基酸序列分成以某个氨基酸为中心的氨基酸序列片段，然后对对该序列片段进行编码。目前常用的编码方法有 5 位编码法、21 位编码法、Profile 编码法等[2]，这些编码方法不但编码的位数较多，而且也没有考虑序列片段中的统计信息。本文提出了一种基于词频统计的编码方法，该编码方法具有较低的维数且充分利用了序列的统计信息，实验结果表明该编码方法结合支持向量机后极大地提高了预测精度。

2 滑动窗口技术与氨基酸序列编码方法

氨基酸序列的编码方法一直是统计建模方法进行蛋白质二级结构预测时面临的问题，通过对氨基酸序列进行有效地分析，之后选择适当的编码方式进行编码，使计算机能够识别生物学的氨基酸序列，才能得到更加准确的预测结果。在对氨基酸序列进行编码时，人们常常通过滑动窗口法来阐述目标氨基酸与其二级结构的对应关系，5 位编码法、21 位编码法和 Profile 编码法是目前较为常用的序列编码方法。

2.1 滑动窗口技术

滑动窗口法的主要依据是氨基酸序列的二级结构不仅和它本身有关，还与周围的氨基酸相关。因此，在使用滑动窗口法时，作为输入的数据不应当仅仅包括当前位置的氨基酸的数据，而是由窗口内部所有氨基酸编码组成。假设有一个长度为 n 的氨基酸链，设定滑动窗口的宽度 w，可以得到 n 个含有 w 个氨基酸的窗口，这里要求 w 是奇数，原因是整个窗口所对应的二级结构是窗口内最中间的氨基酸的二级结构，即第 $\left\lceil \frac{w}{2} \right\rceil$ 个氨基酸，并且考虑这个氨基酸的前 $\left\lfloor \frac{w}{2} \right\rfloor$ 个和后 $\left\lfloor \frac{w}{2} \right\rfloor$ 个氨基酸对它的影响。

2.2 5 位编码法

5 位编码法是针对组成蛋白质的氨基酸种类的个数采用二进制表示来编码，是一种纯数学编码。组成蛋白质的氨基酸共有 20 种，同时加上一种当遇到蛋白质的氨基或羧基或采用滑动窗口法时产生的无效位时，共有 21 种可能，而 $2^4<21<2^5$，因此采用 5 位二进制数表示这 21 种可能。编码的具体方法如下：A 可以编码为 00001，C 可以编码为 00010，……，Y 可以编码为 10100，无效位可以编码为 10101。当采用滑动窗口法时，选择窗口的宽度是 w，则样本集的维数可认为是 $5\times w$。5 位编码法的优点是样本集的维数较低，通过分类器进行分类的时间不会过长，但是在预测问题中大大增加了非线性因素，对于区分不同种类的氨基酸的差异性有所降低。

2.3 21 位编码法

21 位编码法又称为正交编码法，与 5 位编码法相同的是，它也是一种纯数学编码，同样也是针对组成蛋白质的氨基酸种类的个数进行的编码。不同的是，它的编码共需 21 位，不同的氨基酸种类加上一位无效位将会在不同的位置置为 1。因此，这种编码的 21 种情况是两两正交的，因而 21 位编码又称为正交编码。编码的具体方法如下：A 可以编码为 100000000000000000000，C 可以编码为 010000000000000000000，……，Y 可以编码为 000000000000000000010，无效位可以编码为 000000000000000000001。当采用滑动窗口法时，选择窗口的宽度是 w，则样本集的维数可认为是 $21\times w$。

21 位编码法的优点在于不引入任何单体间的代数关系，从而大大降低了非线性因素，然而，这种编码方式的无关特征过多，由于样本集的维数为 $21\times w$，而每个样本中仅有 w 个 1，过多的 0 作为

无关的特征对资源造成较大的浪费。同时，在样本个数相同的前提下，21 位编码的输入是 5 位编码的 4 倍还多，大大增加了数据处理的时间。

2.4 Profile 编码法

Profile 编码是一种基于蛋白质进化信息的编码，是目前认为预测效果比较成功的一种编码方法。与 5 位编码法和 21 位编码法不同，Profile 编码并不针对当前进行编码的某一条氨基酸序列进行编码，而是基于同一家族的蛋白质序列的对比产生，计算当前位置的氨基酸出现的相对概率，因此它被认为带有更多的生物信息而被生物学家广泛使用。Profile 编码的生成方式较为困难，可在网站 ftp://ftp.cmbi.ru.nl/pub/molbio/data/上下载。例如表 1 中列出了蛋白质 1acx 的前 10 个氨基酸的 Profile 编码。

表 1 Profile 编码示例（%）
Table 1 Profile code method example(%)

Seq	A_M	V	L	I	M	F	W	Y	G	A	P	S	T	C	H	R	K	Q	E	N	D
A	AAAAAAAASASASSAAAA	0	0	0	0	0	0	0	0	78	0	22	0	0	0	0	0	0	0	0	0
P	PPPPPPPPPPAPPAAAAA	0	0	0	0	0	0	0	0	32	68	0	0	0	0	0	0	0	0	0	0
A	AAAATTTAVAAGSAAAAA	5	0	0	0	0	0	0	5	68	0	5	16	0	0	0	0	0	0	0	0
F	FFFFAAAVVLVVIVVIII	32	5	21	0	26	0	0	0	16	0	0	0	0	0	0	0	0	0	0	0
S	SSSSTTTSTASTSSTSSS	0	0	0	0	0	0	0	0	5	0	63	32	0	0	0	0	0	0	0	0
V	VVVVVVVAVVVVAVVAAA	74	0	0	0	0	0	0	0	26	0	0	0	0	0	0	0	0	0	0	0
S	SSSSTTTTSSSTTSTAAA	0	0	0	0	0	0	0	0	16	0	47	37	0	0	0	0	0	0	0	0
P	PPPPPPPPPPPPPPPPPP	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0
A	AAAASSSAAGAASAASSS	0	0	0	0	0	0	0	5	58	0	37	0	0	0	0	0	0	0	0	0
S	SSSSSSSTTSTTTTTTTT	0	0	0	0	0	0	0	0	0	0	47	53	0	0	0	0	0	0	0	0

表 1 中 Seq 列是蛋白质的氨基酸序列，A_M 列是与对应 Seq 列对比过的相关序列。例如第一个氨基酸 A 所对应的行 A_M 为“AAAAAAAASASASSAAAA”，其中氨基酸 A 出现的概率是 14÷18≈0.78，S 出现的概率是 4÷18≈0.22，则氨基酸 A 的 Profile 编码就是 000000000.7800.22000000000。同理第八个氨基酸 P 所对应的行 A_M 为“PPPPPPPPPPPPPPPPPP”，其中只含有氨基酸 P，显然它出现的概率为 1，那么氨基酸 P 的 Profile 编码为 0000000001.00000000000，其余以此类推。当采用滑动窗口法时，选择窗口的宽度是 w，则样本集的维数可认为是 $20\times w$。

3 支持向量机

支持向量机是一种基于统计学的学习方法，其归根结底是一个分类问题，本质上应归为前向神经网络。支持向量机的基本思想是基于结构风险最小化准则，在尽可能分开两类样本的前提下，构造一个分类超平面，尽量使更多的样本点被无误地分开。支持向量机自提出之后就广泛应用于人脸识别、文本分类等模式识别领域，最近几年也逐渐应用于蛋白质二级结构预测。

对于线性可分问题，通常可以建立最优分割平面，使得样本在全局范围内得到划分。所谓的最优分割平面就是不但使样本能够得到正确地划分，同时使类间的距离达到最大。假设样本 x_i（i=1,2,…,N）对应的类别为 y_i，其中 $y_i\in\{-1,1\}$，建立的超平面方程为 $W\cdot X+b=0$，对于样本集 X 应满足如下公式：

$$y_i(W\cdot X + b) - 1 \geq 0 \tag{1}$$

使得分类间隔最大的约束应当使$\|w\|^2$最小，即目标优化函数为：

$$\varphi(W)=\frac{\|w\|^2}{2} \tag{2}$$

满足公式（1）并且使$\varphi(W)$最小的分类面就是最优分割平面。

对于近似线性可分问题，若想继续利用最优分割面来划分训练集，就不得不对最优分割面函数增加一个约束$\xi_i\geq 0$，将公式（1）变换为：

$$y_i(W\cdot X + b) - 1 + \xi_I \geq 0 \quad (3)$$

同时目标优化函数变为：

$$\varphi(W) = \frac{\| w \|^2}{2} + C\sum_{i=1}^{n}\xi_i \quad (4)$$

对于线性不可分的情况，SVM 引入核函数将数据映射到一个更高维的空间中，使得数据集在这个空间上是线性可分的。数据 x_i 经过变换Φ得到新的样本$\Phi(x_i)$，使得原始问题可以描述为：

$$y_i(W\cdot\Phi(X) + b) - 1 \geq 0 \quad (5)$$

由于上式中需要计算空间中向量对的点积 $K(x_j,x_j)=\Phi(x_i)\cdot\Phi(x_j)$，那么就要引入核函数来进行计算，常用的核函数有线性核函数、多项式核函数、径向基核函数、Sigmoid 核函数等。线性核函数计算简单，但是只适用于线性可分的情况；多项式核函数参数较多，函数值跨度较大，可能会出现溢出的情况；Sigmoid 核函数不是正定的核函数，在部分情况下通用性不强。因此，在后边的实验中使用只具有一个参数的径向基核函数。

4 基于词频统计的氨基酸序列编码方法

5 位编码法、21 位编码法和 Profile 编码法不但编码的维数较高，而且都没有利用氨基酸序列的统计信息，预测的精度都不高。为此，提出了一种基于词频统计的氨基酸序列编码方法。该方法把每条氨基酸序列看作一个文档，每个氨基酸看作一个词，20 个氨基酸看作 20 个特征词，统计每个氨基酸在序列中出现的频率。每个序列片段编码后都是 20 维的向量，每个元素表示相应氨基酸在该序列片段中出现的次数。如，蛋白质 1acx 的部分序列为“APAFSVSPASGASDGQSV……GHVALTFG”，长度为 n，表 2 中的 Num 列为 1~n，Seq 列为 Num 位置对应的氨基酸，根据滑动窗口法，选择窗口的长度为 13，可以得到表 2 中的 Window 列，具体的编码结果如表 2 的 A～Y 列所示。

表 2 中 Num=1 所在的行，即氨基酸 A 对应的窗口序列为“XXXXXAPAFSVS”，其中氨基酸 A、S 出现的次数为 2，P、F、V 出现的次数均为 1，将其各自出现的次数记在对应的位置作为编码，其编码为“20001000000010020100”，且各位的数值相加之和为窗口长度减去滑动窗口补位数，即 13-6=7。同理，Num=7 所对应的氨基酸 S 的窗口序列为“APAFSVSPASGAS”，其中氨基酸 A、S 出现的次数为 4，P 出现的次数为 2，F、G、V 出现的次数为 1，则其编码为“40001100000020040100”，且各位的数值相加之和为 13-0=13。由此可以看出，这种编码方式的结果是一个 20 维的向量，且各元素值相加之和的取值范围应当为 7~13。

表 2 基于词频统计的氨基酸编码

Table 2 Word frequency statistics coding

序号 Num	序列 Seq	当前窗口 Window	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
1	A	XXXXXXAPAFSVS	2	0	0	0	1	0	0	0	0	0	0	0	1	0	0	2	0	1	0	0
2	P	XXXXXAPAFSVSP	2	0	0	0	1	0	0	0	0	0	0	0	2	0	0	2	0	1	0	0
3	A	XXXXAPAFSVSPA	3	0	0	0	1	0	0	0	0	0	0	0	2	0	0	2	0	1	0	0
4	F	XXXAPAFSVSPAS	3	0	0	0	1	0	0	0	0	0	0	0	2	0	0	3	0	1	0	0
…	…	……											…	…								
n-1	F	GHVALTFGXXXXX	1	0	0	0	1	2	1	0	0	1	0	0	0	0	0	0	1	1	0	0
n	G	HVALTFGXXXXXX	1	0	0	0	1	1	1	0	0	1	0	0	0	0	0	0	1	1	0	0

在基于词频统计的编码方式下，编码是基于整个窗口内部各个氨基酸出现的次数做统计，因此，这种编码方法不仅考虑到目标氨基酸的结构与其相邻的氨基酸相关，更是考虑到窗口内部所有氨基酸对目标氨基酸的影响。而且，使用基于词频统计的编码方法对氨基酸序列编码后的维数为 20，不像 21 位编码法、5 位编码法及 Profile 编码法使用滑动窗口法后样本集维数增长至 20×w。这种编码方式使得样本的输入变得更加简洁，并且易于计算。

5 实验结果

为了验证基于词频统计的编码方法的有效性，在三个数据集上进行测试，如表 3 所示。在选取

数据集时考虑了样本集的大小和同源性对预测结果的影响。数据集 A 选自 SARS 病毒库，包括 8 个蛋白质，共 1560 个氨基酸，是同源蛋白质数据集；数据集 B 选自 HSSP 同源数据库 ftp://ftp.ebi.ac.uk/pub/databases/hssp，包括 36 个蛋白质，共 7232 个氨基酸；数据集 C 选自 RS126 非同源蛋白质数据库，共 41 个蛋白质，包含 6051 个氨基酸。通过数据集 A 和 B 比较样本集大小对预测结果的影响，通过数据集 A 和 C 比较蛋白质的同源性对预测结果的影响，并比较了 SVM 和 BP 神经网络的预测效果。

表 3 实验数据集

Table 3 Datasets for Examples

数据集 Database	蛋白质个数 Number of Proteins	氨基酸个数 Number of amino acids	同源性 Homeology
A	8	1560	同源
B	36	7232	同源
C	41	6051	非同源

在每个数据集上分别执行 SVM 和 BP 神经网络两种分类方法，四种编码方法每种编码方法都执行一次，结果如表 4 所示。在实验中，将每个数据集分成 8 组，选择其中一组作为测试集，其余 7 组作为训练集。对于 SVM，核函数采用径向基核函数，参数 g 和γ选择如表 5 所示。对氨基酸序列进行编码时，窗口的宽度均为 13。

对于蛋白二级结构类别的划分，本文采用 DSSP 划分方法，把蛋白质二级结构定义为 8 种形态（Q8）：H(α -helix)、I(π -helix)、G(3_{10} -helix)、B(isolated β -strand)、E(extended β -strand)、S(bend)、T(turns)以及其他的形态归为一类 C(coil)。在进行蛋白质二级结构预测时，又在此基础上精简为三类（Q3）：H(helix)、E(sheet)和 C(coil)，其中 H 类包含 H(α -helix)，E 类包含 E(extended β -strand)，其余均为 C 类。评价预测结果采用整体准确率，其计算公式如下：

$$Q_3 = \frac{P_H + P_E + P_C}{n} \qquad (6)$$

整体准确率反应的是蛋白质的整体预测效果，公式（6）中的 n 指的是被预测的目标蛋白质所含的氨基酸残基总数，P_H、P_E、P_C 分别代表的是正确预测出二级结构的 H、E、C 态的残基个数。

表 4 不同编码方式的比较（%）

Table 4 Comparison of different coding methods

编码方式 Coding methods	数据集 A Data set A		数据集 B Data set B		数据集 C Data set C	
	SVM	BP	SVM	BP	SVM	BP
21 位编码	76.41	66.67	60.70	46.90	59.52	45.05
5 位编码	73.85	61.03	53.26	49.67	51.69	49.01
Profile 编码	64.62	62.31	67.23	63.61	65.83	62.09
词频统计编码	91.28	79.49	82.13	67.08	82.35	67.94

表 5 SVM 中的参数

Table 5 Parameters in SVM

编码方式 Coding methods	数据集 A Data set A		数据集 B Data set B		数据集 C Data set C	
	g	γ	g	γ	g	γ
21 位编码	2	1	4	1.4142	4	1.4142
5 位编码	1.4142	1	2	1	2	1
Profile 编码	2.8284	1	8	1.4142	4	1.4142
词频统计编码	4	11.3137	8	11.3137	11 34 37	11.3137

由表 4 可知，不论采用那种编码方式，SVM 的预测结果都远优于 BP 神经网络的预测结果。用词频统计对三个不同的数据集进行编码后，用 SVM 和 BP 神经网络预测的结果都远优于其他三种编码方式的预测结果，表明词频统计编码在结合了窗口内部所有氨基酸和目标氨基酸周围因素影响后，对预测结果产生了较好的影响。同时词频统计编码在非同源数据集 C 上也取得了较好的结果。实验结果表明，结合词频统计编码和 SVM 后有效提高了蛋白质二级结构的预测精度。

6 结束语

蛋白质二级结构预测不仅是蛋白质一级结构和三级构象的重要枢纽，更是从一级结构预测其三维空间构象的至关重要的一步。经过几十年的发展，蛋白质二级结构预测方法取得了巨大的成果，但目前蛋白质二级结构预测的精度还有待提高。对统计建模方法来说，氨基酸序列的编码方法是影响预测精度的重要因素之一。目前常用的氨基酸序列编码方法不但编码后的数据维数较高，而且也没有考虑序列片段中的统计信息。本文提出了一种新的氨基酸序列编码方法——基于词频统计的编码方法，该方法统计每个氨基酸在氨基酸序列片段中出现的频率。与其它编码方法相比不但具有较低的维数，而且也充分利用了氨基酸片段内部所有氨基酸对目标氨基酸的影响。实验结果表明，结合词频统计编码和支持向量机预测蛋白质二级结构极大地提高了预测精度。下一步的工作是继续用更多的氨基酸序列来验证词频统计编码的有效性，并将其应用在实际的蛋白质二级结构预测中。

参考文献

[1] Yann G, Gianluca P, Andre E,*et al* .Combining protein secondary structure prediction models with ensemble methods of optimal complexity[J].Neurocomputing,2004(56):305-327

[2] Pollastri G, Przybylski D, Rost B, *et al*.Improving the prediction of protein secondary structure in three and eight classes using neural networks and profiles[J].Proteins: Structure, Function, and Bioinformatics,2002,47(2):228-235

[3] Armano G, Mancosu G, Milanesi L,*et al* .A hybrid genetic-neural system for predicting protein secondary structure [J].BMC Bioinformatics,2005,6(suppl3):1-7

[4] 王秦虎,邓 麟,韩翠芹,等.人工神经网络预测蛋白质二级结构的编码技术综述[J].陕西农业科学,2007,6:109-112

[5] Jahandideh S, Hoseini S, Jahandideh M, *et al*.A hybrid genetic-neural model for predicting protein structural classes [J].Biologia,2009,64(4): 649-654

[6] Yang B R, Hou W, Zhou Z Kaapro,*et al*.An approach of protein secondary structure prediction based on kdd* in the compound pyramid prediction model[J].Expert Systems with Application,2009,36(5):9000-9006

[7] 孟翔燕,孟 军,葛家麒,等.蛋白质二级结构预测方法的评价[J].生物信息学,2010,8(3):206-209

[8] 刘 君,熊忠阳,王银辉,等.蛋白质二级结构的协同训练预测方法[J].计算机应用研究,2011,28(5):1688-1691

[9] 隋海峰,曲 武,钱文彬,等.基于混合 SVM 方法的蛋白质二级结构预测算法[J].计算机科学,2011,38(10):169-174

[10] Bettella F, Rasinski D, Knapp E W,*et al*.Protein secondary structure prediction with SPARROW [J].Journal of Chemical Information Model,2012,52(2):545-556

[11] Madera M, Calmus R, Thiltgen G, *et al*.Improving protein secondary structure prediction using a simple k-mer model [J].Bioinformatics,2010,26(5):596-602

[12] 王菲露,宋 杨.基于广义回归神经网络的蛋白质二级结构预测[J].计算机仿真,2012,29(2):184-187

[13] 连云涓,熊惠霖.蛋白质二级结构预测的多核学习方法[J].计算机应用研究,2013,33(S1):43-45

[14] Chou P Y, Fasman G D.Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins [J].Biochemistry,1974,13:211-222

[15] Gamier J, Osguthorpe D J, Robso B, *et al*.Analysis of the accuracy and implications of simple methods for prediction the secondary structure of globular proteins [J].J Mol Biol,1978,120:97-120

基于词频统计编码和SVM的蛋白质二级结构预测方法

作者： 石陆魁， 刘倩倩， 王靖鑫， 张军， SHI Lu-kui， LIU Qian-qian， WANG Jing-xin， ZHANG Jun
作者单位： 河北工业大学计算机科学与软件学院,天津,300401
刊名： 山东农业大学学报（自然科学版） ISTIC PKU
英文刊名： Journal of Shandong Agricultural University (Natural Science Edition)
年，卷(期)： 2014(z1)

参考文献(15条)

1. Yann G, Gianluca P, Andre E,et al.Combining protein secondary structure prediction models with ensemble methods of optimal complexity[J].Neurocomputing,2004(56):305-327 2004

2. Pollastri G, Przybylski D, Rost B, et al.Improving the prediction of protein secondary structure in three and eight classes using neural networks and profiles[J].Proteins:Structure, Function, and Bioinformatics,2002,47(2):228-235 2002

3. Armano G, Mancosu G, Milanesi L,et al.A hybrid genetic-neural system for predicting protein secondary structure [J].BMC Bioinformatics,2005,6(suppl3):1-7 2005

4. 王秦虎,邓麟,韩翠芹 人工神经网络预测蛋白质二级结构的编码技术综述[期刊论文]-陕西农业科学 2007(06)

5. Jahandideh S, Hoseini S, Jahandideh M, et al.A hybrid genetic-neural model for predicting protein structural classes [J].Biologia,2009,64(4):649-654 2009

6. Yang B R, Hou W, Zhou Z Kaapro,et al.An approach of protein secondary structure prediction based on kdd*in the compound pyramid prediction model[J].Expert Systems with Application,2009,36(5):9000-9006 2009

7. 孟翔燕,孟军,葛家麒 蛋白质二级结构预测方法的评价[期刊论文]-生物信息学 2010(03)

8. 刘君,熊忠阳,王银辉 蛋白质二级结构的协同训练预测方法[期刊论文]-计算机应用研究 2011(05)

9. 隋海峰,曲武,钱文彬,杨炳儒 基于混合SVM方法的蛋白质二级结构预测算法[期刊论文]-计算机科学 2011(10)

10. Bettella F, Rasinski D, Knapp E W,et al.Protein secondary structure prediction with SPARROW [J].Journal of Chemical Information Model,2012,52(2):545-556 2012

11. Madera M, Calmus R, Thiltgen G, et al.Improving protein secondary structure prediction using a simple k-mer model [J].Bioinformatics,2010,26(5):596-602 2010

12. 王菲露,宋杨 基于广义回归神经网络的蛋白质二级结构预测[期刊论文]-计算机仿真 2012(02)

13. 连云涓,熊惠霖 蛋白质二级结构预测的多核学习方法[期刊论文]-计算机应用 2013(z1)

14. Chou P Y, Fasman G D.Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins [J].Biochemistry,1974,13:211-222 1974

15. Gamier J, Osguthorpe D J, Robso B, et al.Analysis of the accuracy and implications of simple methods for prediction the secondary structure of globular proteins [J].J Mol Biol,1978,120:97-120 1978