Computer Science and Application 计算机科学与应用, 2019, 9(2), 450-457
Published Online February 2019 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2019.92051

Hans汉斯

Protein Secondary Structure Prediction Using Convolutional Neural Network and Softmax

Leilei Wang, Jinyong Cheng

School of Computer Science and Technology, Qilu University of Technology (Shandong Academy of Sciences), Ji'nan Shandong
Email: 1393517369@qq.com, cjy@qlu.edu.cn

Received: Feb. 9th, 2019; accepted: Feb. 21st, 2019; published: Feb. 28th, 2019

Abstract

Protein secondary structure prediction belongs to bioinformatics, and it's important in research area. In this paper, we propose a new prediction way of protein using convolutional neural networks and Softmax. First, the improved convolutional neural network is used to extract the characteristics of the protein amino acid sequence, and then the third convolved output in the convolutional neural network is used as input to the Softmax classifier, and these data are trained and predicted. The dataset is a typical 25PDB dataset for protein. In terms of accuracy, the method is the cross validation based on the 3-fold. The results demonstrate that the accuracy of protein secondary structure prediction is improved.

Keywords

Protein Secondary Structure, Convolutional Neural Networks, Softmax Classifier

基于卷积神经网络和Softmax的蛋白质二级结构预测

王蕾蕾，成金勇

齐鲁工业大学(山东省科学院)，计算机科学与技术学院，山东 济南
Email: 1393517369@qq.com, cjy@qlu.edu.cn

收稿日期：2019年2月9日；录用日期：2019年2月21日；发布日期：2019年2月28日

文章引用：王蕾蕾，成金勇．基于卷积神经网络和 Softmax 的蛋白质二级结构预测[J]．计算机科学与应用，2019，9(2): 450-457. DOI: 10.12677/csa.2019.92051

摘　要

蛋白质二级结构预测是生物信息学的重要组成部分，在生物信息学领域具有重要意义。本文提出了一种新的卷积神经网络结合Softmax分类器的算法预测蛋白质二级结构。首先用改进的卷积神经网络对蛋白质氨基酸序列进行特征提取，然后把卷积神经网络中第三次卷积后的输出作为Softmax分类器的输入并进行训练和预测。我们将本文提出的方法在25PDB数据集上做了3-折交叉验证，结果证明蛋白质二级结构预测的准确率有提高。

关键词

蛋白质二级结构，卷积神经网络，Softmax分类器

1. 引言

人类基因组计划(Human Genome Project, HGP) [1]是科学史上的三大伟大计划之一，生物信息学是为了应对计划中的基因测序问题而产生的一门学科。众所周知，蛋白质是生命系统中生命体进行生命活动的主宰者，是生命体不可或缺的一部分。目前，研究蛋白质的结构和功能已经成为生物信息学研究的一个重要的领域。此前为了得到蛋白质的结构主要是运用实验方法，如 X 射线晶体衍射和核磁共振的方法等[2]。但是，现实中最大的问题是我们生命系统中绝大部分的蛋白质的结构是不能用实验方法得到的，所以只能用人工智能预测其结构。所以，蛋白质二级结构预测课题应运而生。所谓蛋白质二级结构的预测[3]，其最重要的步骤是首先归纳我们已经知道结构的蛋白质序列，然后进行预测，预测时主要用到的是统计学方法和目前新兴的人工智能算法等。所以，在蛋白质二级结构发展的今天，已经出现了很多用来预测蛋白质二级结构的方法，如神经网络方法[4]、隐马尔可夫模型[5]方法等。

深度学习[6]是机器学习算法中人工神经网络研究衍生出的一个新的方法，其概念是在 2006 年被提出来的。其包含多种算法，如自动编码器[7]、深信度网络[8]及卷积神经网络等。卷积神经网络分类识别算法是由 Yann LeCun 等人[9]最早提出并应用在手写字体识别上的。卷积神经网络主要是用权值共享的思想降低网络学习的复杂度。本文提出了一种改进的卷积神经网络和 Softmax [10]相结合的方法，构建了一个 10 层的卷积神经网络，先用卷积神经网络对蛋白质二级结构的特征进行提取，把进入全连接层前的特征输入到 Softmax 分类器中，对提取到的特征进行分类预测实验。结果显示，本文的方法在 25PDB (Protein Data Bank，简称 PDB)数据集上的预测准确率有提高。

2. 基于卷积神经网络和 Softmax 的蛋白质二级结构预测

2.1. 卷积神经网络原理

近年来，随着深度学习算法的不断发展，卷积神经网络引起了广大研究者的关注。卷积神经网络(Convolutional Neural Networks, CNN) [11]的概念是上世纪 60 年代 Hubel 和 Wiesel [12]两位研究者在研究神经元的时候发现的一种独特的网络结构，其独特之处在于卷积神经网络可以有效地降低网络的复杂性。

卷积神经网络通过卷积和池化操作能很好地提取到输入数据的关键特征，所以目前卷积神经网络在模式分类领域中得到了广泛的关注和应用。CNN 卷积过程的数学表达式如下所示：

$$s(i,j)=(X*W)(i,j)+b=\sum_{k-1}^{n\_in}(X_k*W_k)(i,j)+b \tag{1}$$

公式中，n_in 是输入数据组成的矩阵个数，X_k 是第 k 个的输入矩阵。W_k 则是卷积过程中我们选取的卷积核里的第 k 个子卷积核矩阵。$s(i,j)$ 表示的是卷积核 W 对应位置的元素在其输出矩阵中的值。

卷积神经网络的基本结构包括卷积层、池化层和全连接层。卷积层主要执行的是卷积操作，其中用到的方法主要是局部连接和权值共享的方法，主要是在模拟大脑中有局部感受野的细胞，从而能够从获得的信息中提取出一些初级特征的过程。池化层主要执行的是下采样操作，包括最大值池化和平均池化等方法。输入的数据经过池化层的下采样操作后，输出的数据矩阵会变小，但是数量不变，所以池化层能够对从上一层的卷积层中输出的数据进行压缩，这样就能减小计算的复杂度、从而使学习参数数量减少，同时能有效地防止过拟合问题。在卷积神经网络模型中，最后的一层或是几层就是全连接层，全连接层的主要功能是对前面卷积和池化操作提取的特征进行加权求和，能够保证输入的数据在进行池化操作后保留下来的少量数据特征能够尽可能的重现原来的输入数据。

图 1 中，input 为数据以矩阵的形式输入卷积神经网络，输入的数据进行卷积操作，得到 C_1 层，C_1 层的数据进行下采样操作，得到 S_1 层，卷积神经网络完成第一次卷积、下采样操作。第一次卷积—下采样操作的输出作为下一次卷积操作的输入，进行第二次卷积、下采样操作，以此类推，直到最后一次卷积操作完成，把卷积、下采样后得到的特征进行全连接并作为卷积神经网络提取特征的输出，把输出的数据输入到分类器中进行分类，得到我们所需要的分类结果。

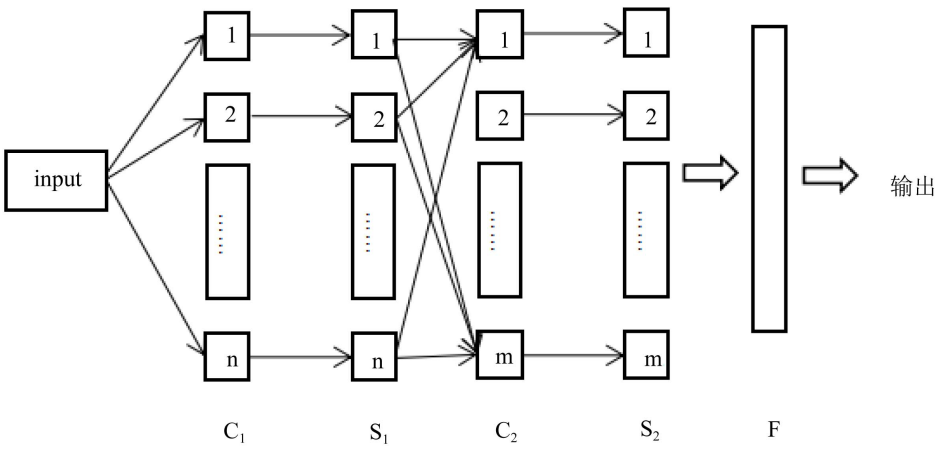

Figure 1. Convolutional neural network structure
图 1. 卷积神经网络结构图

2.2. Softmax 回归模型

传统的逻辑回归模型主要处理的是二分类的问题，如 Logistic 回归模型[13]。面对多分类问题时，传统的逻辑回归模型不能满足分类需要，继而衍生出了一种用于多分类问题的回归模型——Softmax 回归模型。传统的逻辑回归模型函数为：

$$h_\theta(x)=\frac{1}{1+\exp(-\theta^T x)} \tag{2}$$

其损失函数对应如下：

$$J(\theta)=-\frac{1}{m}\left[\sum_{i=1}^{m}y^{(i)}\log h_\theta\left(x^{(i)}\right)+\left(1-y^{(i)}\right)\log\left(1-h_\theta\left(x^{(i)}\right)\right)\right] \tag{3}$$

其中，$x^{(i)}$ 是输入的样本数据，$y^{(i)}$ 是对应的标签数据，θ 是训练的模型参数，m 是样本的总数，因为传统的逻辑回归应用于二分类问题，所以 m 的取值为 2。在多分类问题中，我们用到的是 Softmax 回归，其中 $y^{(i)}$ 可以取 $k(k>2)$ 个值，对应的 m 取值为 k。

对于给定的输入 x，针对每一个类别 j 估算出其概率值 $p=\left(y=j\middle|x\right)$。也就是说，针对每一种分类的结果估算其出现的概率。所以，对于 $y=k(k>2)$ 时回归模型函数的形式如下：

$$h_\theta\left(x^{(i)}\right)=\begin{bmatrix}p\left(y^{(i)}=1\middle|x^{(i)};\theta\right)\\p\left(y^{(i)}=2\middle|x^{(i)};\theta\right)\\\vdots\\p\left(y^{(i)}=k\middle|x^{(i)};\theta\right)\end{bmatrix}=\frac{1}{\sum_{j=1}^{k}e^{\theta_j^T x^{(i)}}}\begin{bmatrix}\theta_1^T x^{(i)}\\\theta_2^T x^{(i)}\\\vdots\\\theta_k^T x^{(i)}\end{bmatrix} \tag{4}$$

为了使公式看起来更加简便，用 θ 表示全部的模型参数，在 Softmax 回归中，把 $\theta_1,\theta_2,\cdots,\theta_k$ 按行排列组成一个矩阵 θ，如下所示：

$$\theta=\begin{bmatrix}—\theta_1^T—\\—\theta_2^T—\\\vdots\\—\theta_K^T—\end{bmatrix} \tag{5}$$

Softmax 回归所对应的损失函数如下所示：

$$J(\theta)=-\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{k}1\left\{y^{(i)}=j\right\}\log\frac{e^{\theta_j^T x^{(i)}}}{\sum_{l=1}^{k}e^{\theta_l^T x^{(i)}}}\right] \tag{6}$$

在上面的公式中，$1\{.\}$ 表示的是示性函数。

从以上可以推出，对于给定的输入数据 x，针对每一个类别 j 估算出的其概率值 $p=\left(y=j\middle|x\right)$ 如下所示：

$$P\left(y^{(i)}=j\middle|x^{(i)};\theta\right)=\frac{e^{\theta_j^T x^{(i)}}}{\sum_{l=1}^{k}e^{\theta_j^T x^{(i)}}} \tag{7}$$

2.3. 卷积神经网络和 Softmax 网络模型

为了简化计算量和解决反向传播时出现的梯度消失问题，在卷积层之后加入了 Relu 激活层。本文的网络结构包括输入层、卷积层、池化层、Relu 激活层和全连接层。一般的卷积神经网络是在全连接之后，对所提取的特征进行分类预测，本文是提取第三层卷积层之后的特征，输入到 Softmax 分类器中进行训练和预测。

主要的网络结构图和参数设置如图 2 所示。

3. 实验和分析

主要介绍蛋白质数据库以及本文的实验过程和结果分析。

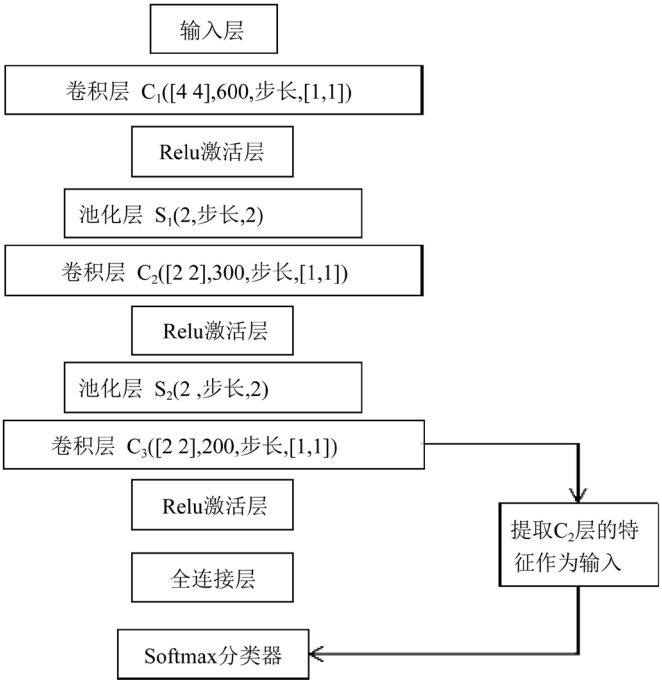

Figure 2. Convolutional neural network-Softmax network structure
图 2. 卷积神经网络－Softmax 网络结构图

3.1. 蛋白质数据库

在蛋白质二级结构预测的研究中，非同源蛋白质数据集 25PDB 数据集经常被用到。我们用本文提出的方法在 25PDB 数据集上做了交叉验证[14]，并把交叉验证的结果和其他方法做了比较。25PDB 数据集包含了 1673 条相似性不超过 25%的蛋白质序列。我们在准确率测试方面，是基于 3-折交叉验证(3-fold cross validation)。

通过位置特异性叠代 BLAST (Position-Specific Iterated BLAST, PSI-BLAST)程序进调用三次迭代并进行序列的对比就能获得位置特异性打分矩阵(Position-Specific Scoring Matrix, PSSM) [15]。PSSM 矩阵包含了物种的进化信息，同时其特有的滑动窗口技术保留了蛋白质序列中相邻氨基酸的关系。多序列比对[16]，从字面意思理解就是我们利用蛋白质序列的相似性对序列进行对比。实际操作中，我们很难知道每一个蛋白质序列的结构和组成，这就要求我们从 NCBT nr 数据库中搜索与其相关的同源序列的信息，预测我们所需蛋白质的结构。主要的操作是运行 PSI-BLAST 程序从 NCBT nr 数据库中搜索到 25PDB 相关蛋白质的同源序列信息之后，生成对应的 20 * 13 的 PSSM 矩阵。把 PSSM 矩阵用于蛋白质二级结构预测时，要选择一个滑动窗口，本文我们用到的滑动窗口大小为 13，沿着 PSSM 矩阵每滑动一次，就可以提取出一个 20 * 13 的特征向量，即生成一个 260 维的特征向量。

3.2. 实验过程

首先通过运行 PSI-BLAST 程序搜索 nr 数据库生成对应 PSSM 矩阵。用大小为 13 的滑动窗口沿着 PSSM 矩阵中的蛋白质序列滑动，得到了一个 260 维的数据。把数据作为卷积神经网络的输入，用 3 * 3 的卷积核对数据进行卷积操作，经过 3 次卷积操作，把卷积神经网络第三次卷积后提取到的特征输出作为 Softmax 分类器的输入，用 Softmax 对提取的特征进行训练和预测，得到预测结果也就是蛋白质二级结构的三种形式：C (卷曲)、E (链残基)和 H (卷曲)。主要的流程图如图 3 所示：

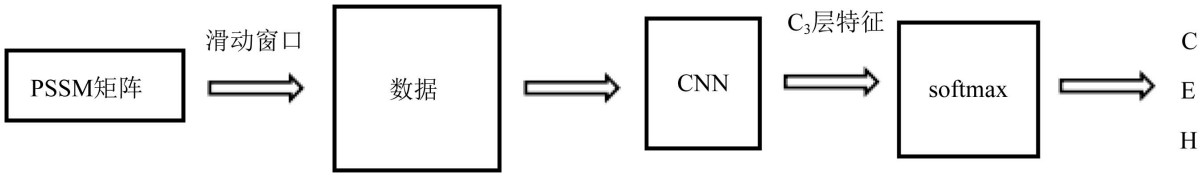

Figure 3. Experimental flow chart
图 3. 实验流程图

在数据处理的滑动窗口选择和在卷积神经网络提取特征时卷积核大小的选择上我们做了反复多次试验。在处理 PSSM 矩阵时，我们分别选取了 9、11、13、15 和 17 作为滑动窗口的大小，分别得到了数据维数为 180 维、220 维、260 维、300 维和 340 维的矩阵，经过实验得出，滑动窗口大小为 13 的时候，预测的效果是最好的。在卷积核大小的选择上，我们分别以 2 * 2，3 * 3，4 * 4 和 5 * 5 作为卷积核的大小进行实验，最后我们设置的卷积层分别是 600 个 4 * 4 的卷积核、300 个 2 * 2 的卷积核和 200 个 2 * 2 的卷积核。

3.3. 实验结果及分析

一个预测算法的好坏主要是通过衡量其预测精度来决定的。蛋白质二级结构的预测精度可以对正确预测的螺旋和折叠数量来进行计算。本文用到的衡量的方法是 Q_3 和片段重叠准确率(Segment Overlap Score, SOV)方法。

Q_3 是被用于残基上的，通过计算正确预测的蛋白质残基占已知蛋白质二级结构序列中总的残基数的比例计算出。Q_3 值的范围为 $[0,1]$，1 表示准确预测。Q_3 就可以表示为：

$$Q_3 = \frac{\text{正确预测的残基数}}{\text{残基总数}} \tag{8}$$

SOV 方法是由 Burkhard Rost [17]等人提出的一个基于重叠片段比值的测度。与 Q_3 正确率的计算方法不同的是，SOV 方法计算的是能够正确预测蛋白质二级结构的片段比例，从影响因素的角度来看，SOV 会忽略一些蛋白质二级结构元素末端的小错误。

本文把 25PDB 数据集分成三份并编号为 1、2 和 3，选择其中的一份作为训练集，剩余的两份作为测试集。我们依次以 1、2 和 3 作为测试集，用传统的卷积神经网络和本文的网络结构对其进行了训练和预测，得到了 Q_3 和 SOV 的平均值。

对于本文用到的 25PDB 数据集，只使用卷积神经网络对训练集和测试集进行分类预测的 Q_3 正确率和 SOV 如下表 1 所示：

Table 1. The results of convolutional neural network
表 1. 卷积神经网络预测结果

		1	2	3	平均
训练集	Q_3	77.41	76.50	76.97	76.96
	SOV	73.29	72.42	72.04	72.58
测试集	Q_3	76.34	75.52	75.85	75.90
	SOV	72.69	71.61	71.66	71.98

对于 25PDB 数据集，先使用改进的卷积神经网络进行特征提取，再提取第三层卷积层的特征输入到 softmax 分类器中进行分类预测的 Q_3 正确率和 SOV 如下表 2 所示：

Table 2. The results of convolutional neural network-Softmax
表 2. 卷积神经网络和 Softmax 预测结果

		1	2	3	平均
训练集	Q_3	78.41	78.41	78.31	78.37
	SOV	73.62	74.43	73.56	73.87
测试集	Q_3	77.01	76.99	77.18	77.06
	SOV	72.79	73.11	73.08	72.99

从上表预测结果可以看出，对于蛋白质数据集 25PDB，只使用传统的卷积神经网络对蛋白质数据进行分类预测，在训练集上的 Q_3 正确率是 76.96%，在测试集上的 Q_3 正确率是 75.90%，本文的方法是在传统的卷积神经网络中加入了 Relu 激活层，对蛋白质数据集进行特征提取，把卷积神经网络第三次卷积得到的特征作为 Softmax 分类器的输入对特征数据进行分类预测，在训练集上的 Q_3 正确率是 78.37%，在测试集上的 Q_3 正确率是 77.06%。本文的方法相较于经典的卷积神经网络方法在训练集和测试集上 Q_3 平均预测精度分别提高了 1.14%和 1.16%。因为自动编码器可以不断调整它的各层的参数，得到每一层的权重，因而能够捕捉可以代表输入数据的最重要的因素，是一种尽可能复现输入信号的神经网络，所以预测结果提高了。但是自动编码器没有全局优化，输入的重建可能不是学习通用表征的理想度量，所以预测正确率的提高不是很明显。

4. 总结

本文结合了改进了卷积神经网络方法，在传统的卷积神经网络方法中加入了 Relu 激活层，简化了计算量并优化了梯度消失问题，直接提取第三层卷积层的特征作为输出，在分类预测方面引入了 Softmax 分类函数。把改进的卷积神经网络对蛋白质数据集 25PDB 经过卷积层、Relu 激活层和池化层之后提取到的第三层卷积后的特征作为 Softmax 分类器的输入，用 Softmax 分类器对提取到的特征进行训练和预测。从表 1 和表 2 中可以看出只用卷积神经网络对 25PDB 数据集进行训练和预测的 Q_3 和 SOV 分别是 75.90%和 71.98%，在本文的方法中的 Q_3 和 SOV 分别是 77.06%和 72.99%，预测结果都有提高。卷积神经网络通过卷积和下采样操作能最大程度地提取到数据重要的信息，本文的方法在卷积神经网络中加入了 Relu 激活层，并把第三层卷积提取到的特征直接作为 Softmax 分类器的输入，最大程度地保留了原始信息，简化了计算量并解决了梯度消失问题。所以提高了预测结果的精度。

基金项目

本研究获得山东省自然科学基金(No. ZR2017LB024)项目资助。

参考文献

[1] Dulbecco, R. (1986) A Turning Point in Cancer Research: Sequencing the Human Genome. *Science*, **231**, 1055-1057. https://doi.org/10.1126/science.3945817

[2] Zvelebil, M.J. and Baum, J.O. (2007) Understanding Bioinformatics. Garland Science, USA. https://doi.org/10.1201/9780203852507

[3] 岳俊杰, 冯华, 梁龙. 蛋白质结构预测实验指南[M]. 北京: 化学工业出版社, 2010.

[4] Kaufman, L. and Rousseeuw, P.J. (2009) Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons, New York.

[5] Huang, Z. (1997) Clustering Large Data Sets with Mixed Numeric and Categorical Values. *Proceedings of the* 1*st Pacific-Asia Conference on Knowledge Discovery and Data Mining* (*PAKDD*), 21-34.

[6] LeCun, Y., Bengio, Y. and Hinton, G. (2015) Deep Learning. *Nature*, **521**, 436. https://doi.org/10.1038/nature14539

[7] 曲建岭, 杜辰飞, 邸亚洲, 等. 深度自动编码器的研究与展望[J]. 计算机与现代化, 2014(8): 128-134.

[8] 张阳, 刘伟铭, 吴义虎. 基于深信度网络分类算法的行人检测方法[J]. 计算机应用研究, 2016, 33(2): 594-597.

[9] LeCun, Y., Bottou, L., Bengio, Y., *et al.* (1998) Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, **86**, 2278-2324. https://doi.org/10.1109/5.726791

[10] Memisevic, R., Zach, C., Pollefeys, M., *et al*. (2010) Gated Softmax Classification. *Advances in Neural Information Processing Systems*, 1603-1611.

[11] Long, J., Shelhamer, E. and Darrell, T. (2017) Fully Convolutional Networks for Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**, 3431-3440.

[12] Hubel, D.H. and Wiesel, T.N. (1962) Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex. *The Journal of Physiology*, **160**, 106-154. https://doi.org/10.1113/jphysiol.1962.sp006837

[13] Hosmer Jr., D.W., Lemeshow, S. and Sturdivant, R.X. (2013) Applied Logistic Regression. John Wiley & Sons, New York. https://doi.org/10.1002/9781118548387

[14] 范永东. 模型选择中的交叉验证方法综述[D]: [硕士学位论文]. 太原: 山西大学, 2013.

[15] Altschul, S.F., Gish, W., Miller, W., *et al.* (1990) Basic Local Alignment Search Tool. *Journal of Molecular Biology*, **215**, 403-410. https://doi.org/10.1016/S0022-2836(05)80360-2

[16] 邹权, 郭茂祖, 韩英鹏, 等. 多序列比对算法的研究进展[J]. 生物信息学, 2010, 8(4): 311-315.

[17] Zemla, A., Venclovas, Č., Fidelis, K., *et al.* (1999) A Modified Definition of Sov, a Segment-Based Measure for Protein Secondary Structure Prediction Assessment. *Proteins*: *Structure*, *Function*, *and Bioinformatics*, **34**, 220-223. https://doi.org/10.1002/(SICI)1097-0134(19990201)34:2<220::AID-PROT7>3.0.CO;2-K