Hans Journal of Computational Biology 计算生物学, 2020, 10(4), 49-56
Published Online December 2020 in Hans. http://www.hanspub.org/journal/hjcb
https://doi.org/10.12677/hjcb.2020.104006

Hans汉斯

基于生成对抗和卷积神经网络的蛋白质二级结构预测

赵亚武，张华兰，刘毅慧

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: zhaoyawu9608@163.com, yxl@qlu.edu.cn

收稿日期：2020年11月9日；录用日期：2020年11月25日；发布日期：2020年12月2日

摘　要

在生物信息学领域，对于蛋白质二级结构预测是一项具有挑战性的任务，对于确定蛋白质的结构和功能有着极其重要的意义。本文融合了生成对抗网络和卷积神经网络模型进行蛋白质二级结构预测，首先利用生成对抗网络提取蛋白质特征，其次将生成对抗网络提取的特征结合PSSM矩阵作为卷积神经网络的输入，得到预测结果。在测试集CASP9，CASP10，CASP11，CASP12，CB513和PDB25获得了87.06%，87.24%，87.31%，87.39%，88.13%和88.93%，比单独使用卷积神经网络提高了3.88%，4.6%，7.97%，5.85%，5.78%，4.25%。实验结果表明，生成对抗网络特征提取能力是非常显著的。

关键词

生物信息学，生成对抗网络，卷积神经网络，蛋白质二级结构预测

Protein Secondary Structure Prediction Based on Generative Confrontation and Convolutional Neural Network

Yawu Zhao, Hualan Zhang, Yihui Liu

School of Computer Science and Technology, Qilu University of Technology (Shandong Academy of Sciences), Jinan Shandong
Email: zhaoyawu9608@163.com, yxl@qlu.edu.cn

Received: Nov. 9th, 2020; accepted: Nov. 25th, 2020; published: Dec. 2nd, 2020

Abstract

In the field of bioinformatics, the prediction of protein secondary structure is a challenging task, and it is extremely important for determining the structure and function of proteins. In this paper,

文章引用: 赵亚武, 张华兰, 刘毅慧. 基于生成对抗和卷积神经网络的蛋白质二级结构预测[J]. 计算生物学, 2020, 10(4): 49-56. DOI: 10.12677/hjcb.2020.104006

the generation of adversarial networks and convolutional neural network models are combined for protein secondary structure prediction. First, the anti-network is generated to extract protein features. Secondly, the extracted features of the anti-network are combined with the PSSM matrix as the input of the convolutional neural network to obtain the prediction results. In the test set CASP9, CASP10, CASP11, CASP12, CB513 and PDB25 obtained 87.06%, 87.24%, 87.31%, 87.39%, 88.13% and 88.93%, which is 3.88%, 4.6%, 7.97%, 5.85%, 5.78%, 4.25% higher than the convolutional neural network alone. The experimental results show that the feature extraction ability of generating adversarial networks is very significant.

Keywords

Bioinformatics, Generative Adversarial Network, Convolutional Neural Network, Protein Secondary Structure Prediction

1. 引言

蛋白质是重要的生物大分子之一，几乎所有的生命活动都离不开蛋白质。伴随着人类基因组计划的完成，科学家从未停止对蛋白质结构的研究。对于蛋白质结构的分类信息研究应通过蛋白质研究领域来解决，并且在生物信息学领域也是十分重要的[1] [2]。蛋白质二级结构预测是三级结构预测的关键一步，是了解和预测三级结构的前提，蛋白质二级结构预测准确率的提高，不仅可以使我们了解到蛋白质序列和蛋白质结构之间复杂的关系，还有助于对蛋白质功能进行分析和制造药物[3]，所以蛋白质二级结构预测是一项具有挑战性的任务并且具有重要意义。利用生物学的方法来测定蛋白质的结构是昂贵且费时的，因此，我们可以借助于计算机的方式进行蛋白质二级结构预测。

在生物信息学领域，已经有很多的计算方法用于蛋白质二级结构预测的问题，比如常见的机器学习算法包括有支持向量机[4] [5]，最近邻算法[6]和贝叶斯算法[7]等。但是机器学习的特征提取依赖于经验，使得对数据的特征提取有一定难度。伴随着科技的发展计算能力的加强，深度学习模型逐渐受到人们的重视，它能够从原始数据中学习特征，不依赖专家经验。蛋白质二级结构预测被应用到卷积神经网络(CNN) [8]和循环神经网络(RNN) [9]中来提高预测的精度。SPIDER3 [10]方法利用长短时记忆双向递归神经网络，能够捕捉到更长的氨基酸序列信息，使准确率达到了 80%以上。SPOT-1D [11]方法是目前较新的蛋白质二级结构预测方法，它是 SPIDER3 的改进，在 SPIDER3 方法的基础上结合了残差卷积网络，获得了较好的效果。Ma 等[12]提出了一种基于数据分割和半随机子空间方法，在 PDB25 和 CB513 数据集上测试 3 态准确率为 86.38%和 84.53%。MUFOLD [13]方法采用名为 Deep3I 的网络，由两个嵌套的可进行卷积操作的初始模块、卷积以及完全连通的致密层组成，有效地处理了氨基酸残基之间的局部和全局相互作用。

近年来，生成对抗网络[14] [15]作为一种较新的深度学习模型，在特征提取，图像去噪方面有着显著的效果。基于上述原因，本文融合了 GAN 和 CNN 神经网络，提出了基于生成对抗网络和卷积神经网络的蛋白质二级结构预测。生成对抗网络可以通过生成器和判别器之间相互博弈来提取氨基酸残基之间的特征，将提取的特征与原始蛋白质特征融合之后送入到卷积神经网络中进行 3 类蛋白质二级结构预测。

2. 蛋白质二级结构预测模型

蛋白质二级结构是根据蛋白质序列预测氨基酸残基对应的结构类型，基于 PSI-BLAST 的位置特异性评分矩阵(PSSM) [16]来表示蛋白质序列，且含有丰富的生物进化信息。PSI-BLAS 参数设置为阈值为 0.001 和 3 次迭代得到 20*M 的 PSSM 矩阵，其中 M 是氨基酸序列的长度，20 代表氨基酸类型的数目。蛋白质结构定义 DSSP [17]中包含有八种结构类型，分别为 H(α 螺旋)、B(β 转角)、E(折叠)、G(3-螺旋)、I(5-螺旋)、T(转角)、S(卷曲)和 L(环)。本文实验将采用 G、H、I 替换为 H，B、E 替换为 E，其他都采用 C 的划分方式。

本文采用生成对抗网络和卷积神经网络预测蛋白质二级结构，首先对数据预处理，按照滑动窗口为 13 和 19 的方式对 PSSM 矩阵进行分割，得到网络的输入数据。其预测模型如见下方图 1 所示。

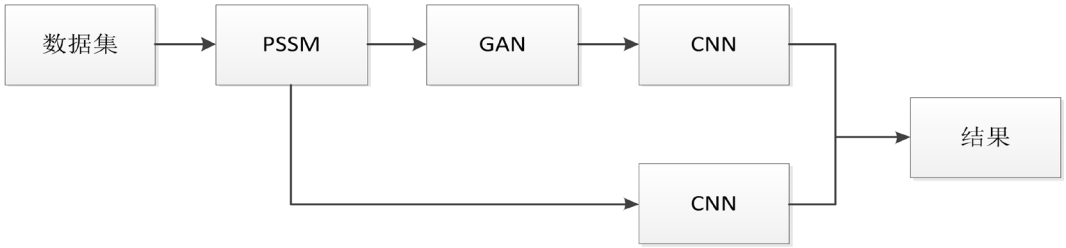

Figure 1. Protein secondary structure prediction model structure (PSSM stands for protein sequence, GAN stands for generative countermeasure network, CNN stands for convolution neural network)
图 1. 蛋白质二级结构预测模型结构(PSSM 代表蛋白质序列，GAN 代表生成对抗网络，CNN 代表卷积神经网络)

2.1. 数据集

本文采用 ASTRAL [18]和 CullPDB [19]数据集作为该模型的训练集，去除 ASTRAL 和 CullPDB 数据集中的重复蛋白质一共有 15696 条蛋白质。测试集采用 CASP [20] [21] [22]类数据集，包括 CASP9，CASP10，CASP11 和 CASP12。除此之外，CB513 [23]和 PDB25 [24]数据集也作为该模型的测试集，测试集的蛋白质序列数目如表 1 所示。

Table 1. Number of protein sequences in test set
表 1. 测试集蛋白质序列数目

数据集	蛋白质序列数目
CASP9	122
CASP10	99
CASP11	81
CASP12	19
CB513	513
PDB25	1672

2.2. 生成对抗网络

2014 年 Ian Goodfellow 提出生成对抗网络(Generative Adversarial Networks) [25]，文献[14]和文献[15]利用生成对抗网络进行图像去噪和特征提取，证明了生成对抗网络具有良好的特性。生成对抗网络包括

两个部分：生成器和判别器。生成器可以学习真实数据的分布特征，为了生成和真实蛋白质数据相似的数据，然而判别器是判断数据是生成器生成的还是真实的数据，实际上是一个二分类问题。从博弈论的角度来看，生成器为了提高自己的生成能力，判别器为了提高自己的判别能力，都是需要不断去优化的，但是两者最终会达到纳什均衡(Nash equilibrium)。生成器和判别器可以分别用 G 和 D 来表示，生成对抗网络模型如图 2 所示。

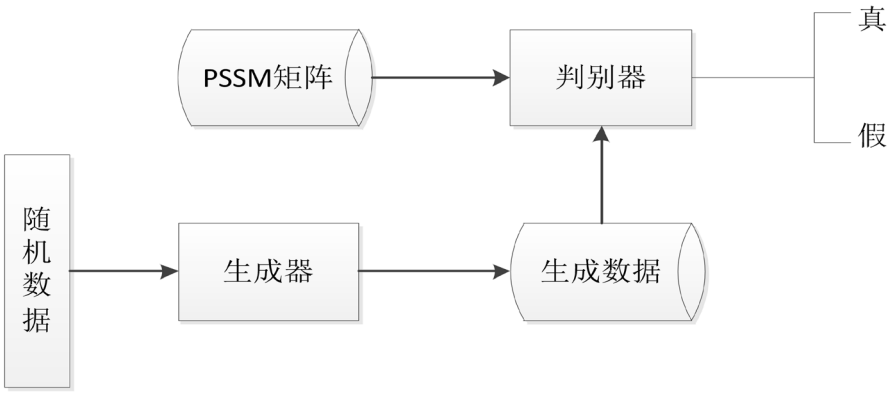

Figure 2. Generative confrontation network model
图 2. 生成对抗网络模型

GAN 的学习过程就是 D 和 G 对抗的过程，由 D 对输入的蛋白质矩阵 PSSM 进行分类，D 可以判别生成数据和真实数据，判别生成数据为假，则 $D\left(G\left(z\right)\right)=0$，对真实数据判别为真，则 $D\left(x\right)=1$。当出现这种情况时，G 就需要不断的调整优化自身参数，使得生成的数据更加接近真实的数据，使得 D 无法判断数据是真实的还是由 G 生成的，即 $D\left(G\left(z\right)\right)=1$。$G$ 与 D 的对抗过程被称为极大极小博弈，它的损失函数定义如下。

$$\min_{G}\max_{D}V\left(D,G\right)=E_{x\sim p_{data}(x)}\left[\log D\left(x\right)\right]+E_{z\sim p_{z}(x)}\left[\log\left(1-D\left(G\left(z\right)\right)\right)\right] \tag{1}$$

式中，x 表示真实的蛋白质数据，z 代表输入到 G 的随机数据，$G\left(z\right)$ 表示 G 网络生成的假的数据，$D\left(x\right)$ 表示 D 网络判断真实数据是否真实的概率，对于 D 来说，这个值越接近 1 越好。而 $D\left(G\left(z\right)\right)$ 是 D 网络判断 G 生成蛋白质数据是否真实的概率，生成器希望自己生成的数据更加接近真实数据，所以 G 希望 D(G(z))尽可能的大，这时 $V\left(D,G\right)$ 会变小，因此我们看到式(1)的最前面的记号是 $\min_{G}$。判别器的能力越强，$D\left(x\right)$ 应该越大，$D\left(G\left(x\right)\right)$ 越小，这时 $V\left(D,G\right)$ 会变大，所以式(1)对于 D 来说是求最大值。

在本文生成对抗模型中，将卷积网络引入到 G 和 D 网络中，目的是为了提高生成对抗网络的特征提取能力，用以提高蛋白质二级结构预测精度。G 网络中使用反卷积进行上采样，激活函数采用 ReLU 函数，D 网络采用步长为 1 的卷积层，激活函数采用 ReLU 函数。将生成对抗网络提取到的特征与 PSSM 矩阵结合并利用深度卷积神经网络进行蛋白质二级结构预测。

2.3. 卷积神经网络

近年来，卷积神经网络作为流行的深度学习算法，被应用到图像处理[26]和计算机视觉[27]等领域。基于卷积神经网络[8] [13]的方法已经应用到蛋白质二级结构预测当中，已经取得显著的效果，它与传统的神经网络相比，它具有权值共享和局部感知的特点，可以减少网络参数加快计算速度，卷积神经网络模型结构图如图 3 所示。

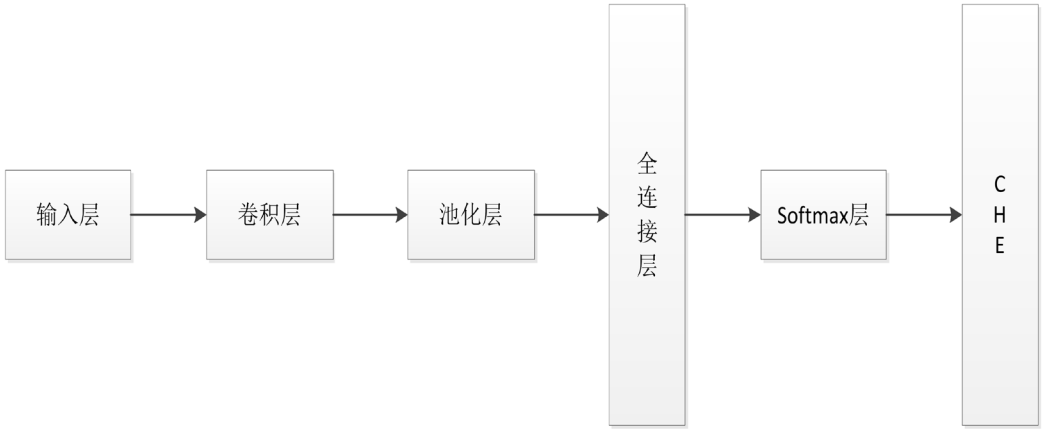

Figure 3. Convolutional neural network model
图 3. 卷积神经网络模型

卷积层通过卷积核对输入的蛋白质数据进行特征提取，卷积的过程就是按照卷积核的大小在输入的蛋白质矩阵上做运算，产生和卷积核数目相同的特征图。特征图由输入矩阵和权重相乘再加上偏置得到，令：

$$F_k^i = f\left(\sum_h P_h^{i-1} * W_k^i + b\right) \tag{2}$$

式中，f 为激活函数 ReLU，P_h^{i-1} 表示输入数据与上一层的卷积核得到的特征图，W_k^i 是第 i 层的一个卷积核，k 表示卷积核的数量，i 表示卷积层数，b 代表偏置参数。

池化层并不执行任何的学习，它通常也被称为一种非线性的下采样形式。池化层处理的结果是使特征维度下降、参数减少来减少计算量，提高计算速度，并且还能有效的减少过拟合，同时还有平移不变的特性，增加了鲁棒性。为了调整权重以进行训练，使用了使用梯度下降算法的反向传播算法。

全连接层和 softmax 层作为卷积神经网络的输出层，全连接层的每一个神经元都要和前一层的神经元相连，输出三类蛋白质二级结构。Softmax 函数层使用激活功能来解决 3 类蛋白质结构的分类问题，其函数定义为：

$$P\left(t_r/x\right) = \frac{P\left(x/t_r\right)P\left(t_r\right)}{\sum_{j=1}^{D} P\left(x/t_j\right)P\left(t_j\right)} = \frac{e^{o_r}}{\sum_{j=1}^{D} e^{o_j}} \tag{3}$$

式中，$P\left(x/t_r\right)$ 是给定类别样本的条件概率，$P\left(t_r\right)$ 是蛋白质结构类别的先验概率。Softmax 函数被视为 logistic Sigmoid 函数的多类推广[28]。

3. 实验结果

本文的实验环境参数如下：处理器 Intel(R) Xeon(R) Glod 5118 CPU 2.30GHz，图形加速卡为 RTX2080Ti，操作系统为 Linux，采用 Keras2.3 版本构建模型。

为了评估本文模型的准确率，采用六种公开的测试集：CASP9、CASP10、CASP11、CASP12、CB513 和 PDB25 进行测试，为了验证生成对抗网络的有效性，针对 3 类蛋白质二级结构预测问题，本文设置了两个不同的实验。实验一是采用卷积神经网络模型用于蛋白质二级结构预测，实验二是首先采用生成对抗网络对蛋白质数据进行特征提取再结合卷积神经网络进行蛋白质二级结构预测。本文采用滑动窗口分别为 13 和 19，卷积层的卷积核大小和尺寸分别为 11 × 11 × 270，11 × 11 × 160 (13 窗口下)和 19 × 19 × 290，16 × 16 × 170 (19 窗口下)。

为了验证迭代次数对生成对抗网络提取特征的影响，迭代次数的单位为万次。由于滑动窗口的不同导致包含的蛋白质特征信息也不同，因此本文分别对 13 和 19 窗口下的蛋白质数据进行验证，实验结果如表 2 和表 3 所示。

Table 2. The impact of the number of iterations under 13 windows on accuracy
表 2. 13 窗口下迭代次数对准确率的影响

Iterations	CASP9	CASP10	CASP11	CASP12	CB513	PDB25
10	84.4	84.71	84.25	85.04	86.23	86.66
20	84.52	85.26	84.09	84.33	86.24	86.82
30	82.76	85.27	84.58	85.01	86.21	86.72
40	84.4	85.46	84.13	84.33	86.19	86.84
50	84.38	84.60	84.09	83.81	86.30	85.52
60	84.29	84.90	84.36	84.05	86.15	86.80

Table 3. The impact of the number of iterations under 19 windows on accuracy
表 3. 19 窗口下迭代次数对准确率的影响

Iterations	CASP9	CASP10	CASP11	CASP12	CB513	PDB25
10	87.06	87.24	87.31	87.39	88.13	88.93
20	86.13	87.05	86.94	86.49	87.75	88.04
30	86.53	86.59	87.24	86.42	87.93	88.61
40	86.63	87.05	87.14	86.70	88.07	88.62
50	86.46	86.75	86.59	86.52	87.94	88.41
60	86.45	87.12	87.27	87.03	87.71	88.49

从表 2 和表 3 可以看出，在滑动窗口为 19 的时候准确率较高，因为在 19 窗口下能够包含更多的蛋白质特征信息，并且在伴随着生成对抗网络中生成器与判别器之间迭代次数的增加，准确率呈现下降趋势，在迭代次数为 10 万的时候取得较好的结果。

Table 4. Convolutional neural network prediction accuracy
表 4. 卷积神经网络预测准确率

滑动窗口	CASP9	CASP10	CASP11	CASP12	CB513	PDB25
13	82.97	82.15	79.23	79.56	81.21	83.73
19	83.18	82.64	79.34	81.54	82.35	84.68

通过表 2，表 3 和表 4 进行对比可以发现，生成对抗网络提取到的特征与 PSSM 矩阵融合，进行 3 类蛋白质二级结构预测，与单独的使用卷积神经网络相比，准确率有了较大的提高。通过本文实验可以看出，生成对抗网络的特征提取是非常有效的，在 CAS P9，CASP10，CASP11，CASP12，CB513，PDB25 数据集上分别提高了 3.88%，4.6%，7.97%，5.85%，5.78%，4.25%。证明了生成对抗网络特征提取能力的优越性。

4. 结论

蛋白质二级结构预测是生物信息学领域一项具有重大意义的工作，对全面了解蛋白质的功能和结构是必要的。本文融合了生成对抗网络和卷积神经网络模型进行蛋白质二级结构预测，由生成对抗网络提

取蛋白质序列特征，再结合 PSSM 矩阵作为卷积神经网络的输入，进行蛋白质二级结构分类预测。通过与只用卷积神经网络预测结果对比，生成对抗网络的特征提取能力是较强的，能够取得非常显著的效果，有较好的可扩展性。

基金项目

国家自然科学基金(No. 61375013)，山东省自然科学基金(No. ZR2013FM020)。

参考文献

[1] Chou, K.C. (2005) Progress in Protein Structural Class Prediction and Its Impact to Bioinformatics and Proteomics. *Current Protein & Peptide Science*, **6**, 423-436. https://doi.org/10.2174/138920305774329368

[2] Costantini, S. and Facchiano, A.M. (2009) Prediction of the Protein Structural Class by Specific Peptide Frequencies. *Biochimie*, **91**, 226-229. https://doi.org/10.1016/j.biochi.2008.09.005

[3] Yang, Y., Gao, J., Wang, J., Heffernan, R., Hanson, J., Paliwal, K. and Zhou, Y. (2018) Sixty-Five Years of the Long March in Protein Secondary Structure Prediction: The Final Stretch? *Briefings in Bioinformatics*, **19**, 482-494.

[4] Hu, H., Pan, Y., Harrison, R., *et al*. (2004) Improved Protein Secondary Structure Prediction Using Support Vector Machine with a New Encoding Scheme and an Advanced Tertiary Classifier. *IEEE Transactions on Nanobioscience*, **3**, 265. https://doi.org/10.1109/TNB.2004.837906

[5] Wang, Y., Cheng, J., Liu, Y., *et al*. (2016) Prediction of Protein Secondary Structure Using Support Vector Machine with PSSM Profiles. *Information Technology*, *Networking*, *Electronic & Automation Control Conference*, Chongqing, 20-22 May 2016. https://doi.org/10.1109/ITNEC.2016.7560411

[6] Zheng, X., Li, C. and Wang, J. (2010) An Information-Theoretic Approach to the Prediction of Protein Structural Class. *Journal of Computational Chemistry*, **31**, 1201-1206.

[7] Robles, V., Larrañaga, P., Peña, J., *et al*. (2004) Bayesian Network Multi-Classifiers for Protein Secondary Structure Prediction. *Artificial Intelligence in Medicine*, **31**, 117-136. https://doi.org/10.1016/j.artmed.2004.01.009

[8] Liu, Y. and Cheng, J. (2016) Protein Secondary Structure Prediction Based on Wavelets and 2D Convolutional Neural Network. *International Conference on Computational Systems Biology & Bioinformatics*, 53-57. https://doi.org/10.1145/3029375.3029382

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

[10] Heffernan, R., Yang, Y., Paliwal, K., *et al*. (2017) Capturing Non-Local Interactions by Long Short Term Memory Bi-directional Recurrent Neural Networks for Improving Prediction of Protein Secondary Structure, Backbone Angles, Contact Numbers, and Solvent Accessibility. *Bioinformatics*, **33**, 2842-2849. https://doi.org/10.1093/bioinformatics/btx218

[11] Hanson, J., Paliwal, K., Litfin, T., *et al*. (2018) Improving Prediction of Protein Secondary Structure, Backbone Angles, Solvent Accessibility, and Contact Numbers by Using Predicted Contact Maps and an Ensemble of Recurrent and Residual Convolutional Neural Networks. *Bioinformatics*, **35**, 2403-2410. https://doi.org/10.1093/bioinformatics/bty1006

[12] Ma, Y., Liu, Y. and Cheng, J. (2018) Protein Secondary Structure Prediction Based on Data Partition and Semi-Random Subspace Method. *Scientific Reports*, **8**, Article No. 9856. https://doi.org/10.1038/s41598-018-28084-8

[13] Chao, F., Yi, S. and Dong, X. (2018) MUFOLD-SS: New Deep Inception-inside-Inception Networks for Protein Secondary Structure Prediction. *Proteins*: *Structure*, *Function*, *and Bioinformatics*, **86**, 592-598. https://doi.org/10.1002/prot.25487

[14] 陈人和, 赖振意, 钱育蓉. 改进的生成对抗网络图像去噪算法[J/OL]. 计算机工程与应用, 1-8. https://kns.cnki.net/kcms/detail/11.2127.TP.20200529.1705.020.html, 2020-11-26.

[15] 秦月红, 王敏. 基于生成对抗网络的跨视角步态特征提取[J]. 计算机系统应用, 2020, 29(1): 164-170.

[16] Jones, D.T. (1999) Protein Secondary Structure Prediction Based on Position-Specific Scoring Matrices. *Journal of Molecular Biology*, **292**, 195-202. https://doi.org/10.1006/jmbi.1999.3091

[17] Kabsch, W. and Sander, C. (2010) Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers*, **22**, 2577-2637. https://doi.org/10.1002/bip.360221211

[18] Fox, N., Brenner, S. and Chandonia, J. (2014) SCOPe: Structural Classification of Proteins-Extended, Integrating SCOP and ASTRAL Data and Classification of New Structures. *Nucleic Acids Research*, **42**, 304-309. https://doi.org/10.1093/nar/gkt1240

[19] Wang, G. and Dunbrack, R. (2005) PISCES: Recent Improvements to a PDB Sequence Culling Server. *Nucleic Acids Research*, **33**, W94-W98. https://doi.org/10.1093/nar/gki402

[20] Moult, J., Fidelis, K., Kryshtafovych, A., *et al*. (2011) Critical Assessment of Methods of Protein Structure Prediction (CASP)—Round IX. *Proteins*, **79**, 1-5. https://doi.org/10.1002/prot.23200

[21] Moult, J., Fidelis, K., Kryshtafovych, A., *et al*. (2014) Critical Assessment of Methods of Protein Structure Prediction (CASP)—Round X. *Proteins*: *Structure*, *Function*, *and Bioinformatics*, **82**, 1-6. https://doi.org/10.1002/prot.24452

[22] Moult, J., Fidelis, K., Kryshtafovych, A., *et al*. (2012) Critical Assessment of Methods of Protein Structure Prediction (CASP)—Round X. *Proteins*: *Structure*, *Function*, *and Bioinformatics*.

[23] Cuff, J., Barton, G., *et al*. (1999) Evaluation and Improvement of Multiple Sequence Methods for Protein Secondary Structure Prediction. *Proteins*: *Structure*, *Function and Genetics*, **34**, 508-519. https://doi.org/10.1002/(SICI)1097-0134(19990301)34:4<508::AID-PROT10>3.0.CO;2-4

[24] Kedarisetti, K., Kurgan, L., Dick, S., *et al*. (2006) Classifier Ensembles for Protein Structural Class Prediction with Varying Homology. *Biochemical and Biophysical Research Communications*, **348**, 981-988. https://doi.org/10.1016/j.bbrc.2006.07.141

[25] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., *et al*. (2014) Generative Adversarial Nets. In: *International Conference on Neural Information Processing Systems*, MIT Press, Cambridge, 2672-2680.

[26] Chartrand, G., Cheng, P.M., Vorontsov, E., *et al*. (2017) Deep Learning: A Primer for Radiologists. *Radiographics*, **37**, 2113-2131. https://doi.org/10.1148/rg.2017170077

[27] 卢宏涛, 张秦川. 深度卷积神经网络在计算机视觉中的应用研究综述[J]. 数据采集与处理, 2016, 31(1): 1-17.

[28] Bishop, C. (2006) Pattern Recognition and Machine Learning. Springer, New York.