Journal of Computer Applications ISSN 1001-9081 -【-
计算机应用, 2016, 【(【): 【-【 CODEN JYIIDU http://www.joca.cn

文章编号:1001-9081(2016)【-0000-00 doi:10.11772/j.issn.1001-9081.2016.【.0000

基于深度学习的八类蛋白质二级结构预测算法

张蕾，李征，郑逢斌，杨伟*

（河南大学 计算机与信息工程学院 空间信息处理实验室，开封 475004）

(*通信作者电子邮箱 yang0sun@gmail.com)

摘 要: 蛋白质二级结构预测是结构生物学中的一个重要问题。针对八类蛋白质二级结构预测，提出了一种基于递归神经网络和前馈神经网络的深度学习预测算法。该算法通过双向递归神经网络建模氨基酸间的局部和长程相互作用，递归神经网络的隐层输出进一步送入到三层的前馈神经网络以便进行八类蛋白质二级结构预测。实验结果表明，提出的算法在 CB513 数据集上达到了 67.9%的 Q_8 预测精度，显著地优于 SSpro8 和 SC-GSN。

关键词: 深度学习；递归神经网络；前馈神经网络；蛋白质二级结构预测

中图分类号: TP183 人工神经网络与计算 **文献标志码:** A

Prediction of Eight-class Protein Secondary Structure based on Deep Learning

ZHANG Lei, LI Zheng, ZHENG Fengbin, YANG Wei*

(Laboratory of Spatial Information Processing, School of Computer and Information Engineering, Henan University, Kaifeng 475004, China)

Abstract: Predicting protein secondary structure is an important problem in structural biology. Aiming at the prediction of eight-class protein secondary structure, a novel deep learning prediction algorithm was proposed by combining recurrent neural network and feed-forward neural network. The method utilizes a bidirectional recurrent neural network to model local and long-range interaction between amino acid residues in protein. In order to predict the eight-class protein secondary structure, the outputs of the hidden layer in the bidirectional recurrent neural network are further fed to the three-layer feed-forward neural network. Experimental results show that the proposed method achieves Q_8 accuracy of 67.9% on the CB513 dataset, which is significantly better than SSpro8 and SC-GSN.

Keywords: deep learning; recurrent neural network; feed-forward neural network; protein secondary structure prediction

0 引言

理解蛋白质的氨基酸序列与其结构之间的潜在关系是结构生物信息学中的一个基本问题[1]。蛋白质二级结构是氨基酸残基在蛋白质多肽链中的局部空间构象，其具有 8 种类型[2]，分别是 α-螺旋（H）、β-桥（B）、折叠（E）、螺旋-3（G）、螺旋-5（I）、转角（T）、卷曲（S）和环（L）。每个二级结构类型都是由其所在蛋白质链中氨基酸残基间的局部和长程相互作用共同决定的。蛋白质二级结构预测任务就是把一个由 20 种氨基酸 A、C、D、E、F、G、H、I、K、L、M、N、P、Q、R、S、T、V、W 和 Y 组成的氨基酸序列映射为对应的二级结构序列。与蛋白质二级结构预测相关的问题有蛋白质折叠预测[3]和蛋白质三级结构预测。特别地，蛋白质的二级结构预测信息不仅有助于确定蛋白质的三维结构，而且能够用于蛋白质的功能预测与互作预测[4]。

早期的蛋白质二级结构预测研究主要集中在粗粒度的三类二级结构预测，即把八类二级结构粗略地归类为螺旋、折叠和卷曲三类。代表的算法有 PHD[5]、PSIPRED[6]和 JPred[7]等。相对于三类二级结构预测，细粒度的八类二级结构预测能够提供更为丰富的信息，并且更具有挑战性。针对八类蛋白质二级结构预测，文献[8]基于双向朴素递归神经网络提出了预测算法 SSpro8。考虑到 SSpro8 不能建模相邻残基二级结构类型之间的依赖关系，文献[9]提出采用条件神经场构建二级结构预测模型。此外，基于结构模板，文献[10]提出了八类二级结构预测算法算法 C8-SCORPION。通过采用带卷积的生成随机网络（Generative Stochastic Network），文献[11]获得了 66.4%的 Q_8 预测精度。

收稿日期:2016-10-28; 修回日期:2016-12-01。

基金项目: 国家自然科学基金面上项目(41571417)。

作者简介: 张蕾（1983-），女，河南周口，助教，硕士，主要研究方向：生物信息学；李征（1985-），女，河南驻马店，讲师，博士，主要研究方向：软件工程；郑逢斌（1963-），男，河南信阳，教授，博士，主要研究方向：空间信息处理、自然语言处理；杨伟（1983-），男，河南信阳，讲师，博士，主要研究方向：机器学习、深度学习。

深度学习是一种通过多个非线性处理层来建模数据中抽象信息的机器学习技术。由于深度学习具有较强的建模能力并且能够基于数据自适应地进行参数学习，因此其已成功地应用于图像分类[12]、物体检测[13]、人脸识别[14]、人类行为识别[15] 、图像识别[16]和图像描述生成[17]等多个领域。为此，本文提出了一种基于深度模型的八类蛋白质二级结构预测算法，并与其他八类蛋白质二级结构算法进行了比较。实验结果表明，提出的算法能够获得较好的预测精度。

1 氨基酸的编码

为了基于氨基酸序列预测蛋白质二级结构，需要为序列中的每个氨基酸构造数值编码。当前常用的编码是正交编码和位置特异性计分矩阵(Position-Specific Scoring Matrices, PSSM)谱编码。图 1 给出了氨基酸的 21 位正交编码。图中的前 20 个字母 A、C、E、D、G、F、I、H、K、M、L、N、Q、P、S、R、T、W、V 和 Y 是构成蛋白质链的基本氨基酸，冒号后面 0 和 1 组成的序列是对应的正交编码。显然，任意两个氨基酸编码向量的内积都为 0。除了上述 20 个字母以外，在蛋白质序列数据库中额外地引入了字母 X 表示氨基酸的具体类型未知，这是由于实验方法有时不能确定一个氨基酸的具体类型。图 1 也给出了 X 对应的 21 位正交编码。

```
A:10000000000000000000  L:00000000010000000000
C:01000000000000000000  N:00000000001000000000
E:00100000000000000000  Q:00000000000100000000
D:00010000000000000000  P:00000000000010000000
G:00001000000000000000  S:00000000000001000000
F:00000100000000000000  R:00000000000000100000
I:00000010000000000000  T:00000000000000010000
H:00000001000000000000  W:00000000000000001000
K:00000000100000000000  V:00000000000000000100
M:00000000010000000000  Y:00000000000000000010
                X:00000000000000000001
```

图 1 氨基酸的正交编码

Fig.1 The orthogonal encoding of amino acids

PSSM 谱编码是通过把目标蛋白质链与蛋白质序列数据库中的蛋白质链进行多序列比对获得的。为了生成目标蛋白质链的 PSSM 谱编码，需要把 NCBI nr(ftp://ftp.uniprot.org/pub/databases/uniprot/uniref/) 数据库下载到本机。在 NCBI nr 数据库中，有基于序列一致性阈值 100%、90%和 50%分别构造的三种非冗余蛋白质数据库，本文采用序列一致性为 50%的数据库 uniref50.fasta。对于 uniref50.fasta，需要首先调用 pfilt 程序对其蛋白质序列数据进行过滤，以便删除卷曲-卷曲片段、跨膜区域和低复杂性区域。然后调用 makeblastdb 程序对其进行格式化生成便于多序列比对的数据和索引文件。在处理完 uniref50.fasta 数据库后，可调用 3 次迭代的 psiblast 程序把目标蛋白质链的氨基酸序列与数据库中的蛋白质链进行多序列比对。在迭代完成后，psiblast 程序会生成目标蛋白质链的 PSSM 谱编码文件。图 2 给出了蛋白质 4Y3F 的部分氨基酸的 PSSM 谱编码。在图中，最左边的数字是氨基酸在蛋白质链中的编号，字母表示氨基酸，字母右边的 20 个数是对其的数值编码。比如，对于第 12 个氨基酸 S，图中虚线长方形中的 20 个数值组成的向量就是它的 PSSM 谱编码。此外，在使用 PSSM 谱编码之前，我们进一步采用 Sigmoid 函数把编码中的整数值映射到 0 到 1 之间。

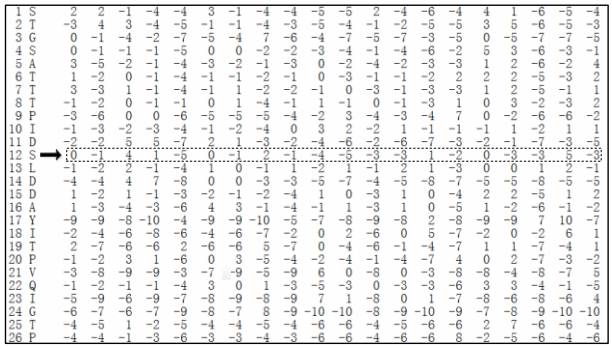

图 2 蛋白质 4Y3F 的前 26 个氨基酸的 PSSM 谱编码

Fig.2 The PSSM profile encoding of the first 26 amono acid residues of the protein 4Y3F

2 蛋白质二级结构的深度预测模型

蛋白质二级结构预测是氨基酸序列到二级结构序列的映射问题。为了按照序列方式预测蛋白质二级结构，本文通过组合递归神经网络和前馈神经网络构造深度神经网络预测模型。图 3 给出了具体的深度预测模型。特别地，预测模型的具体数据处理流程如下：首先氨基酸序列中的每个氨基酸通过氨基酸编码形式化为数值向量送入到递归神经网络——双向长短时记忆模型(Long Short-Term Memory, LSTM)中，然后组合双向 LSTM 的前向和后向隐层输出送入到前馈神经网络的输出层中，最后根据前馈神经网络输出层的结果确定预测的蛋白质二级结构序列并输出。

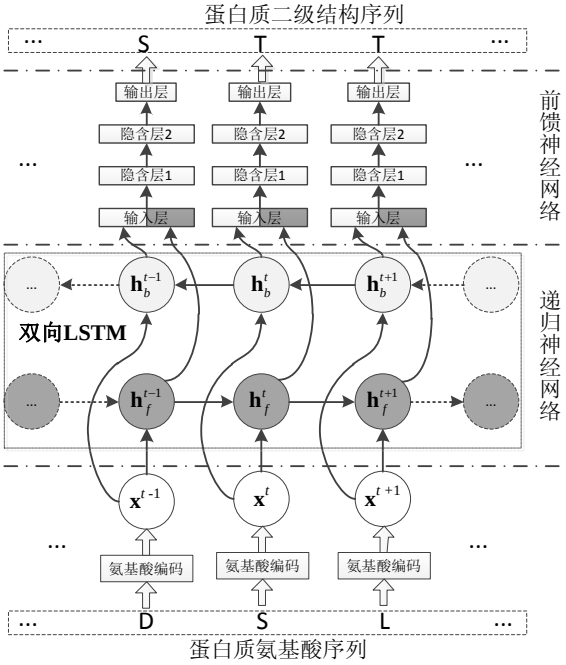

图 3 蛋白质二级结构的深度预测模型

Fig.3 The deep prediction model of protein secondary structure

对于氨基酸序列，本文采用组合的 PSSM 谱编码和 21 位正交编码表示氨基酸，因此每个氨基酸可由长度为 41 的特征向量表示。当用 $\boldsymbol{x}^t$ 表示蛋白质链中第 t 个位置氨基酸的特征向量时，则长度为 τ 的蛋白质可形式化为序列数据 $\{\boldsymbol{x}^1,\ldots,\boldsymbol{x}^t,\ldots,\boldsymbol{x}^\tau\}$。此外，对应的目标二级结构序列可表示为 $\{\boldsymbol{y}^1,\ldots,\boldsymbol{y}^t,\ldots,\boldsymbol{y}^\tau\}$，其中 $\boldsymbol{y}^t$ 是处于第 t 个位置的二级结构类型的 8 位正交编码。

由于单个氨基酸的二级结构类型是由蛋白质链中的整个序列信息决定的，因此我们采用双向 LSTM 处理数值化的氨基酸序列信息。对于双向 LSTM，前向 LSTM 在 t 时刻可以获取目标 $\boldsymbol{x}^t$ 以前的序列信息 $\boldsymbol{x}^1,\ldots,\boldsymbol{x}^t$，后向 LSTM 对应的获取目标 $\boldsymbol{x}^t$ 后面的序列信息 $\boldsymbol{x}^1,\ldots,\boldsymbol{x}^t$。通过组合双向 LSTM 在 t 时刻的两个隐层输出向量 $\boldsymbol{h}_f^t$ 和 $\boldsymbol{h}_b^t$，可以获取整个氨基酸序列的信息。特别地，LSTM 的基本单元由下面的方程定义:

$$\begin{aligned}
\boldsymbol{i}^t &= \sigma\left(\boldsymbol{U}_i\boldsymbol{x}^t + \boldsymbol{V}_i\boldsymbol{h}^{t-1} + \boldsymbol{b}_i\right)\\
\boldsymbol{o}^t &= \sigma\left(\boldsymbol{U}_o\boldsymbol{x}^t + \boldsymbol{V}_o\boldsymbol{h}^{t-1} + \boldsymbol{b}_o\right)\\
\boldsymbol{f}^t &= \sigma\left(\boldsymbol{U}_f\boldsymbol{x}^t + \boldsymbol{V}_f\boldsymbol{h}^{t-1} + \boldsymbol{b}_f\right)\\
\boldsymbol{g}^t &= \mathrm{Tanh}\left(\boldsymbol{U}_g\boldsymbol{x}^t + \boldsymbol{V}_g\boldsymbol{h}^{t-1} + \boldsymbol{b}_g\right)\\
\boldsymbol{s}^t &= \boldsymbol{s}^{t-1} \odot \boldsymbol{f}^t + \boldsymbol{i}^t \odot \boldsymbol{g}^t\\
\boldsymbol{h}^t &= \mathrm{Tanh}(\boldsymbol{s}^t) \odot \boldsymbol{o}^t
\end{aligned} \tag{1}$$

其中，符号 $\odot$ 表示两个向量按元素依次相乘，$\boldsymbol{i}^t$、$\boldsymbol{o}^t$、$\boldsymbol{f}^t$、$\boldsymbol{g}^t$、和 $\boldsymbol{s}^t$ 分别是 t 时刻的输入网关向量、输出网关向量、遗忘网关向量、候选结点向量和内部记忆向量，$\boldsymbol{U}_i$、$\boldsymbol{U}_o$、$\boldsymbol{U}_f$、$\boldsymbol{U}_g$、$\boldsymbol{V}_i$、$\boldsymbol{V}_o$、$\boldsymbol{V}_f$ 和 $\boldsymbol{V}_g$ 是需要学习的权值矩阵，$\boldsymbol{b}_i$、$\boldsymbol{b}_o$、$\boldsymbol{b}_f$ 和 $\boldsymbol{b}_g$ 是偏置向量。通过引入网关和内部记忆机制，LSTM 不仅可以克服消失的梯度问题，而且能够显式地建模序列数据之间的长距离依赖关系。特别地，通过学习模型参数，LSTM 可以针对不同的任务自适应地学习采用何种记忆机制以便建模序列数据的长距离依赖关系。

对于前馈神经网络，本文采用三层的多层感知机(MultiLayer Perceptron)。多层感知机的输入层是由双向 LSTM 的前向和后向隐层输出向量 $\boldsymbol{h}_f^t$ 和 $\boldsymbol{h}_b^t$ 组成的特征向量，两个隐含层都采用校正激活单元 ReLU 作为激活函数，输出层通过采用 Softmax 函数可以计算预测的 8 类二级结构概率分布。给定输入序列数据 $\boldsymbol{x}=\{\boldsymbol{x}^1,\ldots,\boldsymbol{x}^t,\ldots,\boldsymbol{x}^\tau\}$，则深度预测模型将输出对应的序列数据 $\hat{\boldsymbol{y}}=\{\hat{\boldsymbol{y}}^1,\ldots,\hat{\boldsymbol{y}}^t,\ldots,\hat{\boldsymbol{y}}^\tau\}$，其中 $\hat{\boldsymbol{y}}^t$ 是处于第 t 个位置氨基酸的 8 类二级结构预测概率分布。结合目标输出序列 $\boldsymbol{y}=\{\boldsymbol{y}^1,\ldots,\boldsymbol{y}^t,\ldots,\boldsymbol{y}^\tau\}$，本文基于交叉熵定义单个序列对的损失函数为：

$$\xi(\hat{\boldsymbol{y}},\boldsymbol{y}) = -\sum_{t=1}^{\tau}\sum_{k=1}^{8} y_k^t \ln \hat{y}_k^t \tag{2}$$

当训练数据包括 n 个序列对时，深度预测模型的目标损失函数就是 n 个序列对对应损失函数的均值。通过采用基于时间的反向传播算法(Backpropagation Through Time, BPTT)最小化深度预测模型的目标函数，可以自适应地学习双向 LSTM 和多层感知机的参数。

3 实验

3.1 数据集

本文采用文献[11]使用的非冗余蛋白质二级结构数据集 PISCES CullPDB 训练深度预测模型。该数据集包含 6128 个蛋白质链，由 PISCES CullPDB 服务器按照蛋白质分辨率不大于 2.5 和蛋白质序列一致性小于 30%的约束条件生成。同时，常用的 CB513 数据集被当作测试集，以便评估提出算法的分类性能。为了消除训练数据与测试数据之间的序列相似性，我们进一步从 PISCES CullPDB 数据集中删除与 CB513 中的蛋白质链存在序列一致性大于 25%的蛋白质链。通过删除操作，PISCES CullPDB 数据集最后剩下 5534 条蛋白质链。对于剩下的蛋白质链，本文选取 5278 个蛋白质链作为深度预测模型的训练数据集，余下的 256 条蛋白质链作为验证集。引入验证集的目的是为了采用早期停止方法防止过训练。也就是说，当训练模型在验证集上的验证误差不再降低时停止参数训练。

3.2 实验结果与分析

为了进行实验比较，需要对提出的深度预测模型进行超参数设置。对于双向 LSTM，隐层单元个数设置为 96，初始隐层状态设为零向量。对于多层感知机，两个隐层结点的个数都设置为 192，激活函数采用校正线性单元(Rectified Linear Units, ReLU)，输出层采用输出类别为 8 类的 Softmax 函数。此外，本文采用带冲量的随机梯度下降算法训练网络参数，其中冲量系数设为 0.9，学习率为 0.01，minibatch 的大小设置为 128，梯度剪切的阈值设置为 5。当网络在验证集上的预测精度不再增加时停止训练。

表 1 给出了本文算法在 CB513 数据集上的混淆矩阵，其中粗体字标出了 8 种蛋白质二级结构类型 α-螺旋（H）、β-桥（B）、折叠（E）、螺旋-3（G）、螺旋-5（I）、转角（T）、卷曲（S）和环（L）的预测正确率。从表中可以看出，α-螺旋和折叠正确率最高，分别达到了 91.98%和 81.92%；β-桥和螺旋-5 的正确率最低，分别为 1.19%和 0.00%，这主要是由于它们在蛋白质结构数据库所占的比率极低的缘故。特别地，

螺旋-5 出现的比率只有 0.0002。因此，八类蛋白质二级结构预测是不平衡分类问题。

表 2 给出了本文算法与四个算法 SC-GSN[11]、RaptorX-SS8[9]、SSpro8[8]和 LSTM large[18]在 CB513 数据集上的 Q_8 预测精度。对于所比较的四个算法，SC-GSN 采用的是带卷积的生成随机网络，RaptorX-SS8 使用的预测模型是条件神经场，SSpro8 采用的是双向朴素递归神经网络，LSTM large 使用双向 LSTM。从表中可以看出，本文算法获得了最高的预测精度，并且明显地优于 SSpro8 和 SC-GSN。

表 1 本文算法在 CB513 测试集上的混淆矩阵

Tab. 1 The confusion matrix of the proposed algorithm on the CB513 dataset

	H_{pred}	B_{pred}	E_{pred}	G_{pred}	I_{pred}	T_{pred}	S_{pred}	L_{pred}
H	**91.98%**	0.00%	1.48%	0.77%	0.00%	1.99%	0.34%	3.45%
B	8.21%	**1.19%**	25.57%	0.76%	0.00%	7.20%	5.00 %	52.07%
E	2.46 %	0.02%	**81.92%**	0.21%	0.00%	1.71%	1.37%	12.32%
G	29.15%	0.03%	7.63%	**19.89%**	0.00%	18.30%	3.22%	21.78%
I	66.67%	0.00%	10.00%	0.00%	**0.00%**	16.67%	0.00%	6.67%
T	18.89%	0.01 %	6.26%	2.70%	0.00%	**48.08%**	5.24 %	18.83%
S	8.57%	0.06%	12.22%	1.08%	0.00%	16.76%	**20.88%**	40.43%
L	5.76%	0.03%	17.74%	0.61%	0.00%	6.60%	4.85%	**64.40%**

表 2 CB513 数据集上的分类性能比较

Tab. 2 Performance comparision on the CB513 dataset

算法	Q_8(%)
SC-GSN	66.4
RaptorX-SS8	64.9
SSpro8	63.4
LSTM large	67.4
本文算法	67.9

4 结语

针对八类蛋白质二级结构预测，本文通过组合递归神经网络和前馈神经网络提出了一种新的基于深度学习的预测算法。该方法首先采用双向 LSTM 模型处理氨基酸序列数据，以便建模氨基酸之间的长距离依赖关系。然后，双向 LSTM 的两个隐层输出被组合成单个特征向量进一步送入到三层的多层感知机。最后，采用交叉熵作为目标函数以便训练深度网络参数。特别地，通过采用递归神经网络，提出的算法能够直接进行氨基酸序列到蛋白质二级结构序列的预测。CB513 数据集上的实验结果表明，本文算法能够获得较好的预测精度，并且明显地优于 SSpro8 和 SC-GSN。

参考文献

[1] Cheng J, Tegge A N, Baldi P. Machine learning methods for protein structure prediction[J]. IEEE Reviews in Biomedical Engineering, 2008, 1: 41-49.

[2] Touw W G, Baakman C, Black J, te Beek T A, Krieger E, Joosten R P, Vriend G. A series of PDB-related databanks for everyday needs[J]. Nucleic acids research, 2015: D364-D368.

[3] Lin C, Zou Y, Qin J, Liu X, Jiang Y, Ke C, Zou Q. Hierarchical Classification of Protein Folds Using a Novel Ensemble Classifier[J]. PLOS ONE, 2013, 8(2): e56499.

[4] Rost B. Review: protein secondary structure prediction continues to rise[J]. Journal of structural biology, 2001, 134(2): 204-218.

[5] Rost B,Sander C. Prediction of protein secondary structure at better than 70% accuracy[J]. Journal of molecular biology, 1993, 232(2): 584-599.

[6] Buchan D W, Ward S, Lobley A E, Nugent T, Bryson K, Jones D T. Protein annotation and modelling servers at University College London[J]. Nucleic acids research, 2010: gkq427.

[7] Drozdetskiy A, Cole C, Procter J, Barton G J. JPred4: a protein secondary structure prediction server[J]. Nucleic Acids Research, 2015.

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

[9] Wang Z, Zhao F, Peng J, Xu J. Protein 8 - class secondary structure prediction using conditional neural fields[J]. Proteomics, 2011, 11(19): 3786-3792.

[10] Yaseen A,Li Y. Template-based C8-SCORPION: a protein 8-state secondary structure prediction method using structural information and context-based features[J]. BMC Bioinformatics, 2014, 15: S3.

[11] Zhou J,Troyanskaya O G. Deep supervised and convolutional generative stochastic network for protein secondary structure prediction: the 31st International Conference on Machine Learning[C]. Beijing,China: 2014.

[12] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks: Advances in neural information processing systems[C]. 2012: 1097-1105.

[13] Gupta S, Girshick R, Arbeláez P, Malik J. Learning rich features from RGB-D images for object detection and segmentation: The 13th European Conference on Computer Vision[C]. Springer, 2014: 345-360.

[14] 张雯,王文伟. 基于局部二值模式和深度学习的人脸识别[J]. 计算机应用, 2015(05): 1474-1478. (ZHANG W, WANG W. Face recognition based on local binary pattern and deep learning [J]. Journal of Computer Applications, 2015(05): 1474-1478.)

[15] Shuiwang J, Wei X, Ming Y, Kai Y. 3D Convolutional Neural Networks for Human Action Recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1): 221-231.

[16] 康晓东，王昊，郭军，于文勇. 无监督深度学习彩色图像识别方法[J]. 计算机应用, 2015(09): 2636-2639. (KANG X, WANG H, GUO J, YU W. Unsupervised deep learning method for color image recognition [J]. Journal of Computer Applications, 2015(09): 2636-2639.)

[17] Karpathy A,Fei-Fei L. Deep visual-semantic alignments for generating image descriptions: The IEEE Conference on Computer Vision and Pattern Recognition[C]. 2015.

[18] Sønderby S K,Winther O. Protein secondary structure prediction with long short term memory networks[J]. arXiv preprint arXiv:1412.7828, 2014.

基金项目：

国家自然科学基金资助项目(41571417)。

Background

This work is partially supported by the National Natural Science Foundation of China (41571417).

作者简介：

张蕾（1983-），女，河南周口，助教，硕士，主要研究方向：生物信息学；

李征（1985-），女，河南驻马店，讲师，博士，主要研究方向：软件工程；

郑逢斌（1963-），男，河南信阳，教授，博士，主要研究方向：空间信息处理、自然语言处理；

杨伟(1983-)，男，河南信阳，讲师，博士，主要研究方向：机器学习、深度学习。

ZHANG Lei, born in 1983, M. S.. His research interests include Bioinformatics.

LI Zheng, born in 1985, Ph. D.. His research interests include software engineering.

ZHENG Fengbin, born in 1963, Ph. D., professor. His research interests include spatial information processing, natural language processing.

YANG Wei, born in 1991, Ph. D.. Her research interests include machine learning and deep learning.