文章编号:1009-038X(2005)01-0084-05

基于神经网络的多聚脯氨酸二型结构预测

陆克中， 黄可望， 须文波
(江南大学 信息工程学院,江苏 无锡 214036)

摘 要: 使用"机器学习方法"对稀有的多聚脯氨酸二型(PPII)的二级结构进行预测,在预处理蛋白质序列的基础上,使用生物信息学中常用的 BP 神经网络预测 PPII 二级结构. 通过对不同输入窗口长度与不同隐层节点数的神经网络进行训练和测试,得出在输入窗口长度为 13 个氨基酸残基和隐节点数为 15 时预测效果最好,此时的预测精度可达 73.8%.

关键词: 神经网络;BP-模型;蛋白质二级结构;多聚脯氨酸二型结构预测;多聚脯氨酸二型结构
中图分类号:Q 61 **文献标识码**: A

Prediction of Polyproline Type II Secondary Structures by Artificial Neural Network

LU Ke-zhong, HUANG Ke-wang, XU Wen-bo
(School of Information Technology, Southern Yangtze University, Wuxi 214036, China)

Abstract: So far few works have been conducted to predict polyproline type II(PPII) secondary structure with machine learning approaches. On the base of preprocessing protein sequences, this paper predicts PPII secondary structure with BP type's neural network model that is most frequently used in bioinformatics. By training and testing neural networks with different window lengths and different hidden nodes, it could be concluded that, the best predicting result was obtained when the window length was 13 and the the hidden node was 15, the optimal predicting accuracy(Q) reached 73.8 percent.

Key words: neural network; BP-model; protein secondary structure; prediction of polyproline type II; polyproline type II

蛋白质二级结构,除了 α 螺旋、β 折叠外还有其它确定的二级结构类型,多聚脯氨酸二型(polyproline type II,缩写为 PPII)就是其中之一. 它是一种最少残基为 4 的三角形螺旋结构. 近年来 PPII 结构越来越受到重视,因为人们已逐渐了解到这种结构有着特殊的生物特性. PPII 结构在多种重要生化过程中扮演着重要角色,包括信号传导、转录、细胞运动,以及免疫反应等[1]. 例如,PPII 结构是 SH_3, WW 和 MHC-II 等结构域的识别部位,这些结构域都与某些疑难疾病相关. 另外,PPII 结构是胶原蛋白类和植物细胞壁蛋白的主要结构特征. 目前,在蛋白质二级结构预测中,专门针对 PPII 二级结构预测还比较少. 这主要是由于 PPII 结构非常稀少而难以预测所致.

收稿日期:2004-06-02; **修回日期**:2004-10-05.

作者简介: 陆克中(1976-),男,安徽枞阳人,自动化控制专业硕士研究生.

作者在预处理 PPII 二级结构的基础上,使用神经网络技术[2]预测 PPII 二级结构. 神经网络结构见图 1. 神经网络技术在生物信息学中已得到广泛应用,例如应用于神经网络用于序列编码分析、蛋白质结构和功能预测、蛋白质家族分类、单肽及其切割位点预测、遗传密码的结构和起源分析、真核生物基因寻找和内含子剪接位点预测等,已被证明是一种有效的处理工具.

1　蛋白质实体选取及 PPII 结构提取

1.1　蛋白质实体选取

蛋白质实体取自 PISCES 服务器[3]. PISCES 服务器根据序列的同一性和结构质量标准,从蛋白质数据库(PDB)中精选蛋白质实体. 蛋白质实体的选取条件为:序列的同一性不大于 30%,分辨率不低于 0.25 nm,R 值不大于 0.2,序列长度不小于 80,不含用非 X 衍射或仅用 CA 获得的蛋白质实体. 结果获得 2 303 个蛋白质实体.

1.2　PPII 二级结构定义与提取

使用常用的 DSSP 程序,从上面选取的蛋白质实体中获取蛋白质序列以及几何特性等信息. 使用这些信息并根据参考文献[4,5]定义 PPII 二级结构. 简要概括如下:

PPII 二级结构一般出现在 $\alpha = -110°$,$\phi = -75°$ 和 $\varphi = 105°$ 附近,这里 α 为虚拟角,是通过二面角 ϕ 和 φ 根据公式 $\alpha = 180° + \varphi_i + \phi_{i+1} + 20°(\sin\phi i + \sin\varphi_{i+1})$ 得到(见图 1).

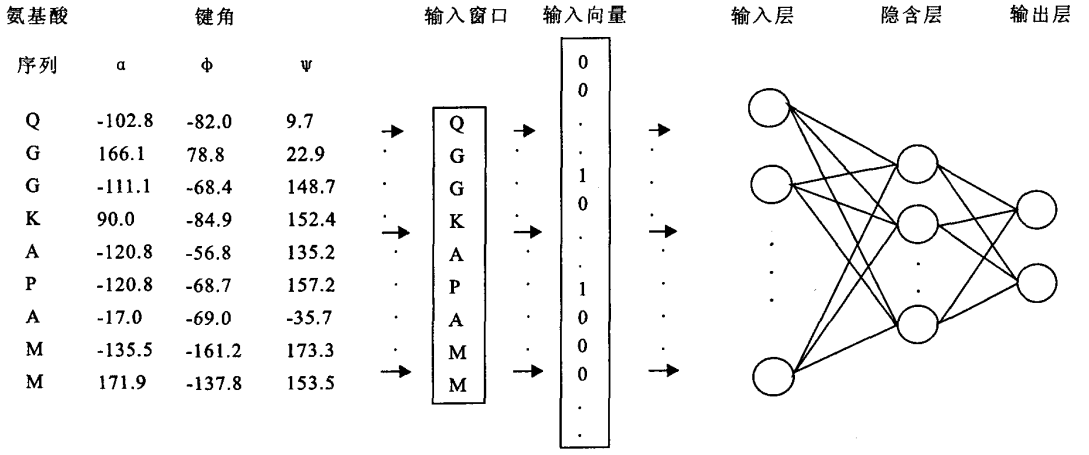


图 1　用于神经网络的序列编码

Fig. 1　Encoding a sequence for a neural network

通过使用公式 $D = \dfrac{\sum_{k=1}^{n-1} d_{k,k+1}}{n}$,利用 ϕ 角和 φ 角计算结构的规则度 D. 这里

$$d_{k-1,k} = \sqrt{(\Psi_{i-1} - \Psi_i)^2 + (\phi_i - \phi_{i+1})^2}.$$

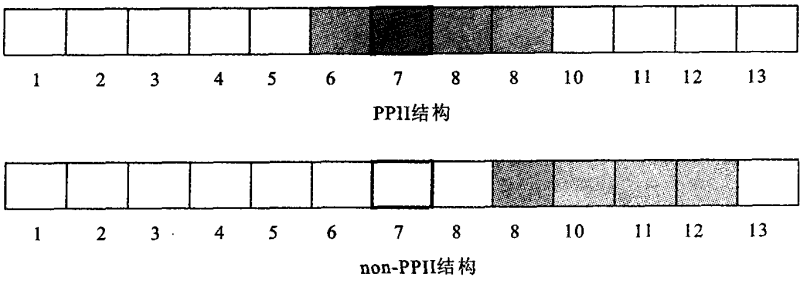
规则度是连续 ϕ 角与 φ 角的一种平均距离形式. *PPII* 二级结构的二面角要求为:$-145° < a < -70°$;　$-180° < \phi < -160°$ 或 $90° < \varphi < 180°$.

对于 PPII 二级结构,其二面角不仅要求在上述范围之内,而且还要求 PPII 二级结构至少有 3 个连续的以上范围的二面角,同时要求每一部分的规则度 D 不大于 50. 寻找 PPII 二级结构的算法流程如下:

```
while 有蛋白质实体
    while 氨基酸序列不为空
    if  -145<α<-70
        考虑下一个氨基酸
        n=1
        while -145<α<-70
            考虑下一个氨基酸
            n=n+1
        end while
        if n>3 then
            计算此结构的规则度 D
            if D<50 & 此种结构氨基酸个数>3 then
                计算此结构中每一部分的规则度 Di
                if 所有 Di<50 then
                    标记此部分结构为 PPII 二级结构
                end if
```

```
            end if
        end if
    end if
        考虑下一个氨基酸
    end while
    考虑下一个蛋白质
end while
```

用 C++程序实现上面的算法，并使用生物信息学中常用的窗口技术判断蛋白质序列中哪一部分是 PPII 结构或不是. 窗口技术用于选取序列作为神经网络的输入样本. 判断结构的依据以窗口中间氨基酸结构类型来确定，见图 2. 这里的窗口长度为 13，图中阴影部为 PPII 二级结构，其它部分为 non-PPII 二级结构. 窗口在氨基酸序列上滑动，当窗口的中间氨基酸二级结构类型为 PPII 类型时，则认为窗口所选取的这段连续结构类型为 PPII 二级结构类型；反之，则为 non-PPII 二级结构类型.



图中阴影部分为 PPII 结构

图 2 长度为 13 个氨基酸的窗口技术

Fig. 2 Windowing technique as exemplified by the window length of 13 amino acids

为了使用于神经网络训练的 PPII 结构数据和 non-PPII 结构数据彼此惟一，对这两种类型数据进行了两两比对，比对之后得到在不同窗口长度时的 PPII 结构数目见表 1. 由于 non-PPII 结构数据较多，故难以全部用来进行测试. 另外，由于训练的目的是识别出 PPII 和 non-PPII 这两种二级结构特征. 因此在满足训练数据量的情况下，采取裁减的办法，即随机选择 non-PPII 结构使其与相同窗口长度的 PPII 结构数目大致相等. 以这些数据作为数据集.

表 1 不同窗口长度下的 PPII 结构数目

Tab. 1 The number of PPII structures with the different window lengths

窗口长度	PPII 结构数目
5	7152
7	7203
9	7173
11	7145
13	7118
15	7089
17	7052

2 输入编码与测量精度

2.1 输入编码

对蛋白质序列采用经典的局部编码方案[6]，即利用上述的窗口技术判断神经网络的输入结构是 PPII 二级结构还是 non-PPII 二级结构. 对于每个残基，使用正交的二值编码，如(1,0,…,0),(0,1,…,0)等. 向量维数为 20. 如果窗口长度为 l，那么输入层的节点数即为 20 l.

2.2 预测精度评价

交叉确认对减少由于训练集或测试集的特定选取而导致的结果偏见性是必需的. 采用 7 折叠(fold)交叉确认将数据集分成大小相近的 7 等份，每份里面的 PPII 结构和 non-PPII 结构仍然保持 1∶1组合. 依次选 6 份作为训练集，剩下的 1 份作为测试集. 以下所得结果都是通过交叉确认获得.

结果的预测精度通常采用敏感度、特异度及总精度(Q)来评价[7]. 定义如下：

敏感度= TP/(TP+ FN)，

特异度= TP/(TP+ FP)，

总精度=(TP+ TN)/(TP+ TN+FP+FN)

其中：TP(真阳性)，意指 PPII 二级结构被正确预测出来的数目；TN(真阴性)，意指 non-PPII 二级结构被正确预测出来的数目；FP(假阳性)，意指 non-PPII 二级结构没有被正确预测出来的数目；FN(假阴性)，意指 PPII 二级结构没有被正确预测出来的数目.

文中采用了以上评价方法.

3　结果与讨论

在分类实验中,网络的输入节点数为 20 l(l 为窗口长度),输出节点数为 2. 对于只有一个隐层的 BP 算法,隐层节点的数目分别选为 4,8,15,30 和 40. BP 算法调用的是 Matlab 中的 newff 函数来建立网络并进行训练. 其中,训练函数为 traingdm,隐层和输出层的传递函数均为 sigmoid 函数,学习率 lr=0.6,动量系数 mc=0.9,目标误差 goal=0.01,最大跌代次数 epochs=10 000. 对于有 2 个隐层的 BP 算法,隐层与隐层之间的传递函数仍为 sigmoid 函数,其它参数同只有一个隐层的 BP 算法. 另外,实验中采用的径向机函数网络调用的是 matlab 中的 newrb 函数来建立网络并进行训练. 其中,扩展常数 sp=10,目标误差 goal=0.01. 实验结果见表 2. 表 2 中的结果都是使用 7-折叠交叉确认得到的.

表 2　在不同窗口长度和不同隐层节点数目下的神经网络测试结果

Tab. 2　The test results for different hidden nodes with the different window lengths

窗口长度	隐层节点数	敏感度/%	特异度/%	总精度/%
5	4	67.0	72.0	70.5
	8	67.3	72.5	70.9
	15	68.1	72.3	71.3
	30	67.1	71.7	70.3
	40	66.7	71.5	70.1
7	4	67.8	73.6	71.7
	8	68.1	74.2	72.2
	15	69.2	73.6	72.1
	30	67.8	73.9	71.9
	40	67.2	73.5	71.5
9	4	68.1	73.8	71.9
	8	68.4	73.9	72.1
	15	70.4	74.4	73.1
	30	67.6	73.6	71.7
	40	67.5	73.2	71.4
11	4	69.4	74.1	72.6
	8	70.2	74.5	73.1
	15	71.6	74.7	73.7
	30	70.7	73.8	72.8
	40	70.0	73.1	72.1
13	4	69.7	73.4	72.2
	8	69.4	73.7	72.3
	15	72.1	74.7	73.8
	30	71.7	72.3	71.9
	40	70.7	72.0	71.6
15	4	70.7	73.3	72.5
	8	71.6	73.9	73.2
	15	71.9	73.0	72.7
	30	70.3	73.7	72.6
	40	70.0	73.2	72.2
17	4	69.1	72.7	71.6
	8	70.1	72.7	71.9
	15	69.4	72.6	71.7
	30	69.0	72.2	71.2
	40	68.7	71.4	70.6

从表 2 可以看出,神经网络输入窗口长度为 13 时的测试结果优于其它长度时的窗口输入. 在窗口长度为 13,隐层节点数为 15 时,所得的最优结果为:敏感度 72.1%,特异度 74.7%,总精度 73.8%. 此时的训练集误差变化见图 3.

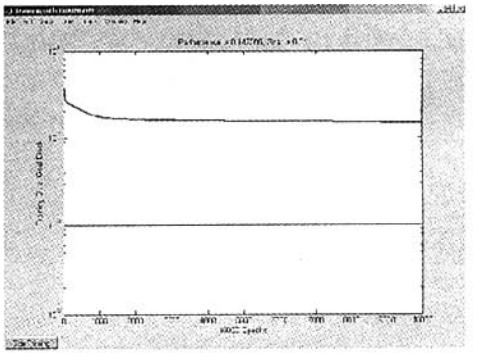


图 3　用在带有 15 个隐节点最优网络结构的一个训练集的误差变化

Fig. 3　The error profile of one of the training sets for the optimal network topology with 15 hidden nodes

将在不同窗口长度下的最优精度罗列在表 3 中,从表 3 可以看出,当窗口长度为 11,13,15 时,相邻窗口长度的敏感度、特异度和总精度之间差距较小,最大为 0.6%,表明此时窗口长度对预测精度不是很敏感. 窗口长度为 5,7,17 时,精度明显变小,特别是敏感度,最大差距达 4 个百分点. 原因是窗口长度过小,造成残基片段的信息丢失,致使结果精度偏小;若窗口长度太长,尽管可以提取较多的

残基信息，但同时也增加了信噪比，致使残基片段也不是越长越好. 理论上讲，增加感知器网络的隐层节点数，可以得到较好的分类决策面，但这需要以大量的学习数据集为前提，而这里所得的数据是有限的. 作者也采用了具有两个隐层的感知器网络与径向机函数(RBF)网络来测试数据集，见表 4. 当两个隐层的感知器网络的隐层分别为 20 和 10 时，所获得的最好预测精度为 72.6%，对应的敏感度为 70.4%；而使用径向机函数网络获得的最好预测精度为 72.1%，但其对应的敏感度只有 67.8%. 敏感度反映的是 PPII 二级结构能否被成功预测出来的能力，也就是说在同样条件下，使用两个隐层的感知器网络和径向机函数网络所能识别的 PPII 二级结构数目都要比使用只有一个隐层的感知器少. 因此相比较而言只有一个隐层的感知器比较好，它对 PPII 二级结构特征比较敏感.

表 3 不同窗口长度下的最优结果

Tab. 3 The optimal results with different window lengths

窗口长度	敏感度/%	特异度/%	总精度/%
5	68.1	72.7	71.3
7	68.1	74.2	72.2
9	70.4	74.4	73.1
11	71.6	74.7	73.7
13	72.1	74.7	73.8
15	71.6	73.9	73.2
17	70.1	72.7	71.9

表 4 不同分类器的最优结果

Tab. 4 The optimal results with different classifier

网络类型	最优窗口长度	敏感度/%	特异度/%	总精度/%
感知器(一个隐层 15)	13	72.1	74.7	73.8
感知器(两个隐层 20-10)	13	70.4	73.7	72.6
径向机函数	11	67.8	74.1	72.1

4 结 论

测试结果表明，使用前馈神经网络解决稀有的 PPII 结构分类问题已取得了较好的预测结果. 在输入窗口长度为 13 个氨基酸残基、隐层节点数为 15 时，使用 BP 神经网络得到的预测总精度达 73.8%. 但无法像预测 α-螺旋、β-折叠等常规结构那样数据集采用自然分布的形式. 如果那样，将会有太多的 non-PPII 结构不能被预测出来. 因为 PPII 结构很少，只占整个氨基酸残基的 1.2%. 为了提高预测精度，在今后的工作中，如果以利用多序列比对算法得到的序列谱作为神经网络的输入，可能会使预测精度有所提高. 序列谱的获得可以得用 HSSP 程序[8] 或 PSI-BLAST 程序[9] 得到，后者更好. 另外还可以结合其它预测方法，如 PHD、PSIPRED 等[10,11]，这样可能比使用单个预测方法获得更好的结果.

参考文献：

[1] Kelly M, Chellgre B, Rucker A, *et al*. Host-guest study of left-handed polyproline II helix formation[J]. **Biochemistry**, 2001, 40: 14376－14383.

[2] Simon Havkin. 神经网络原理[M]. 叶世伟，史忠植译. 北京：机械工业出版社，2004.

[3] Wang G, Dunbrack R L. PISCES: a protein sequence culling server[J]. **Bioinformatics**, 2003, 19(12): 1589－1591.

[4] Adzhubei A, Sternberg M. Left handed polyproline II helices commonly occur in globular proteins[J]. **J Mol Biol**, 1993, 229: 472－493.

[5] Siermala M, Juhola M, Vihinen M. On preprocessing of protein sequences for neural network prediction of polyproline type II secondary structures[J]. **Computers in Biology and Medicine**, 2001, 31: 385－398.

[6] Qian N, Sejnowski TJ. Predicting the secondary structure of globular proteins using neural network models[J]. **J Mol Biol**, 1988, 202: 865－884.

[7] Pierre B, Soren B. 生物信息学[M]. 李衍达，朱宗涵译. 北京：中信出版社，2003.

[8] Schneider R, Daruvar A, Sander C. The HSSP database of protein structure- sequence alignments[J]. **Nucleic Acids Res**, 1997, 25: 226－230.

[9] Altschul S F, Madden T L. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs[J]. **Nucleic Acids Res**, 1997, 25: 3389－3402.

[10] Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy[J]. **J Mol Biol**, 1993, 232: 584－599.

[11] Mcguggin L, Bryson K, Jones J. The PSIPRED protein structure prediction server[J]. **Bioinformatics**, 2000, 16: 404－405.

(责任编辑：杨 萌，秦和平)