不同编码方法对蛋白质二级结构预测精度的影响

马栋萍[1,2]，阮晓钢[1]

（1 北京工业大学电子信息与控制工程学院，北京 100022，Email:moleeyan@sina.com；2 北京联合大学生物化学工程学院，北京 100023）

摘要： 蛋白质二级结构预测是蛋白质结构预测的关键步骤。考虑氨基酸编码方式对预测精度的影响，提出基于一种改进的 BP 算法，采用 3 种不同氨基酸序列编码方式进行蛋白质二级结构的预测。在不同编码方式下，运用 MATLAB 语言实现改进 BP 神经网络的初始化和训练，并分析比较了各编码方式对蛋白质二级结构预测精度的影响。实验表明，采用此 BP 算法时，氨基酸序列的正交编码方式可获得较高的预测精度。

关键词： 蛋白质二级结构；编码方式；改进 BP 网络；预测精度

Different Encodings Influence the Precision of Protein Secondary Structure Prediction

MA Dong-ping[1,2]，*RUAN Xiao-gang*[1]

(1 Electronic Information and Control Engineering School, Beijing University of Technology, Beijing 100022, Email: moleeyan@sina.com; 2 Biochemistry Engineering School, Beijing Union University, Beijing 100023)

Abstract: Prediction of secondary structure is the pivotal step of the prediction of protein sequence. In view of the influence of the encoding methods, based on an improved BP algorithm, this paper discusses 3 encoding techniques which are used in protein secondary structure prediction. Under different encoding techniques, BP neural network is initialized and trained in MATLAB, and the precisions of prediction are analyzed. The experimental results show that orthogonal encoding technique can get highest precision of prediction by using improved BP neural network.

Keywords: Protein Secondary Structure, Encoding Technique, Improved BP Neural Network, Prediction Of Prediction

1 前 言

蛋白质结构预测在蛋白质工程学中占有重要地位。各种生命活动都是通过蛋白质来实现的，对其研究可为疾病的早期诊断，药物精确筛选等提供方案。而蛋白质的作用取决于其空间结构，如果掌握了蛋白质的空间结构，就可以预测并解决人类疾病。但现有的实验测定方法局限性较大，且测定速度很慢，因此蛋白质的结构预测是目前分子生物学研究中迫切需要解决的问题。

通过对已知空间结构的蛋白质分子的研究和分析，人们发现尽管一条多肽链可能采取的构象数目是相当大的，但是在蛋白质分子中，由二级结构组装而形成一定的空间结构的方式却是有限的。一般认为，如果二级结构的预测准确率能达到 80%，那么便可以基本准确地预测一个蛋白质分子的三维空间结构，因此需要进一步提高蛋白质二级结构预测的准确度。

人工神经网络在非同源性蛋白质二级结构预测中是最为成功的一种方法。

2 蛋白质二级结构

蛋白质二级结构是指由氨基酸序列通过氢键联结成的 α 螺旋、β 折叠片和无规卷曲等规则的蛋白

基金项目： 国家自然科学基金资助项目(No:60234020)。

作者简介： 马栋萍(1977-),女,在职硕士生,助教,从事专业:神经网络研究；阮晓钢(1958-),男,教授,博士生导师,从事专业:人工智能与神经网络研究。

质局部结构元件。蛋白质分子中多个肽键平面通过氨基酸上α碳原子的旋转，相互紧密盘曲成稳定的α螺旋构象;而β折叠则是由蛋白质分子中被称为β链的几段区域多肽链形成，肽链的伸展使肽键平面之间一般折叠成锯齿状。除α螺旋和β折叠以外的构象我们都归为无规卷曲。如图 1。

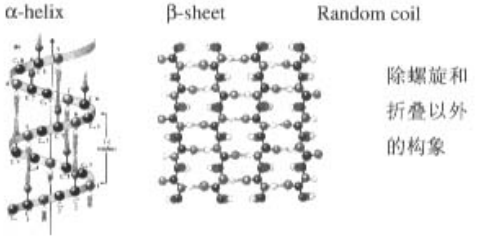

图 1 蛋白质二级结构的类型

Fig.1 The types of protein secondary structure

3 编码规则

用神经网络方法预测蛋白质二级结构所用的氨基酸编码方式有很多种不同的形式，本文将对常用的 3 种编码方式进行分析和比较，研究其对蛋白质二级结构预测准确度的影响。

一般来说，对于某一特定氨基酸，其前后的 8 个氨基酸残基与它具有统计相关性，会影响到该氨基酸的二级结构的形式，因此通常选取神经网络窗口的长度为 17，即同一时刻允许 17 个连续氨基酸序列的编码作为神经网络的输入信息。

3.1 正交编码方式

常见的氨基酸有 20 种，但当网络的输入窗口移动到蛋白质的氨基或者羧基端的时候，窗口有些位置不对应任何氨基酸，这时候我们不输入任何东西，即产生一个空输入。对此 21 种状态可采用正交编码，即 21 位编码方式。一个残基的 21 个神经元中只有代表该氨基酸残基处于激发态的神经元其编码为 1，其他神经元输出均为 0。即氨基酸 A 的编码：100000000000000000000；B 的编码：010000000000000000000；……；Y 的编码：000000000000000000010；氨基或者羧基端的编码：000000000000000000001。

采用此种编码方式，神经网络的输入层神经元个数为 17×21 个，选择隐含层神经元个数为 30 个。

3.2 5 位编码方式

由于输入窗口的状态可能有 21 种，而 $2^4<21<2^5$，因此用 5 位二进制数可以唯一表示它们。

氨基酸 A 的编码：00001；B 的编码：00010；……；Y 的编码：10010；氨基或者羧基端的编码：10011。

采用此种编码方式，神经网络的输入层神经元个数为 17×5 个，选择隐含层神经元个数为 15 个。

3.3 归一化编码方式

由于氨基酸分子不同的理化性质使得各氨基酸替换的频率大不一样，其中理化性质相同的氨基酸的替换倾向于保存蛋白质的结构与功能，这种中性替换在进化中会不断积累，出现的频率也就越大。因此，可以将氨基酸用其疏水性和极性来唯一表示。将 20 种氨基酸的极性值作归一化处理，将极性交量范围划分为 8 个个区间，每一区间对应变量编码用 8 维二进制向量表示，第一个区间用 10000000；第二个区间用 01000000；……；第八个区间用 10000000。

再考虑各种氨基酸的疏水特性，若某一氨基酸为疏水性则在前面所得向量之前加编码 1；如果是亲水性则加编码 0。因而得到 9 位二进制向量。

采用此种编码方式，神经网络的输入层神经元个数为 17×9 个，选择隐含层神经元个数为 20 个。

4 改进 BP 神经网络

4.1 改进的 BP 算法

本文采用一种改进的 BP 算法，即动量法和学习率自适应调整的策略，通过 Matlab 语言实现对蛋白质二级结构的预测。因此避免了由于标准 BP 算法的主要缺点：1 收敛速度慢；2 易出现局部极小值，从而影响到蛋白质二级结构的预测准确度。

首先加入动量项：

$$W(k+1)=W(k)+lr[(1-mc)D(k)+mcD(k-1)]$$

式中 W(k)表示权值向量；D(k)为 k 时刻的负梯度；D(k-1)为 k-1 时刻的负梯度；lr 为学习率，lr>0；mc 为动量因子，0≤mc<1。

动量法实质上相当于加入了阻尼项，降低了网络对于误差曲面局部细节的敏感性，减小了学习过程的振荡趋势，有效地抑制网络陷于局部极小，从而改善了收敛性。

另外在 BP 算法中，网络权值的调整在于学习速率和梯度。在标准 BP 算法中收敛速度慢的一个重要原因是学习率选择不当，因而可以采用自适应学习率调整法。在 Matlab 工具箱中，给出了一种自适应学习速率 lr 的调整公式:

$$lr(k+1)=\begin{cases} 1.05lr(k); & mse(k+1)<mse(k) \\ 0.7lr(k); & mse(k+1)>1.04mse(k) \\ lr(k); & 其它 \end{cases}$$

mse 为均方差。因而可得到比标准 BP 算法更快的收敛速度，有利于缩短学习时间。

因此，将附加动量法和自适应学习率调整法两种策略结合起来，既可有效地抑制网络陷入局部极小，又有利于缩短学习时间。

4.2 神经网络设计

采用的 BP 神经网络为 3 层前馈网络，即输入层、隐含层和输出层之间的前向连接。其结构见图 2。

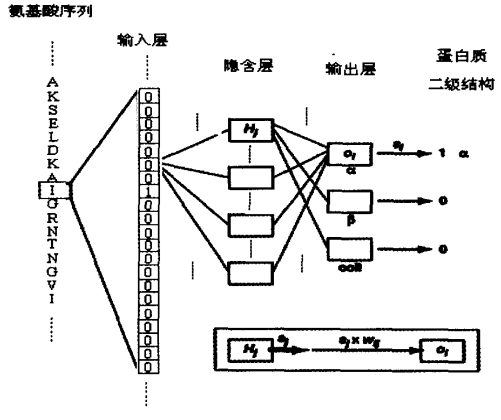

图 2 神经网络拓扑图

Fig.2 Structure of neural network

该网络输入层是一个沿着蛋白质的氨基酸序列滑动的窗口，预测是对窗口中间位置的氨基酸进行的。设氨基酸序列编码的位数为 M，则：输入层包含 17×M 个神经元，隐含层设计神经元个数依据输入层神经元的个数而定，输出层包含 3 个神经元，采用正交编码，即：001 为 α 螺旋；010 为 β 折叠；100 为无规卷曲。网络结构确定后，便可考虑神经元的变换函数选取；网络的初始化(连接权值和阈值的初始化)；训练参数设置；样本数据导入方式等，然后利用 Matlab 语言编制仿真程序。

首先选择输入层到隐含层的变换函数为 tansig，隐含层到输出层的变换函数为 purelin，算法采用动量法和学习率自适应 BP 算法。设定性能目标值为 0.01，动量系数为 0.95，学习率初始值为 0.02。

5 结论及评价

运用以上各种编码方式实验时，共选用了 40 个蛋白质作为训练集，12 个蛋白质作为测试集，所用的蛋白质均来源于 Brookhaven Protein Data Bank。评价二级结构预测算法好坏的方法有很多，最常用的是简单考虑预测总体正确率，即是所有正确预测三种二级结构的残基的百分比，其计算公式如下:

$$Q=(p_\alpha+p_\beta+p_c)/N$$

其中 N 为蛋白质残基的个数，p 分别为正确预测某类二级结构(螺旋、折叠、卷曲)的残基个数。

在不同编码方式下，运用此网络对测试集中的 12 个氨基酸序列进行仿真预测。发现，采用正交编码方式时，由于网络结构较大，因而运算速度较慢，收敛所需时间长，但预测准确度较高，最高可达 72.13%；采用归一化编码方法时，因为考虑到蛋白质自身的特性，且网络结构较小，因此收敛速度快，预测准确度也较好，可达到 69.25%；采用 5 位编码方式时，网络结构最小，但由于编码方式不能很好的反映各个氨基酸的特性，因此收敛速度较慢且预测准确度也低，仅为 63.74%。

由实验结果看出，不同的编码方式对预测精度的影响是比较大的，因此神经网络中的编码问题是蛋白质二级结构预测中的重点，应基于此寻求更好的编码技术，以便在应用神经网络进行蛋白质二级结构预测时，能够获得较小的网络模型，较快的运算速度以及高的预测精度。

参考文献

[1] 梅启鹏. 蛋白质二级结构预测中的简化编码技术[J]. 2004，科技情报开发与经济，14 卷 5 期:133-134.

[2] 刘晋钢. BP 神经网络改进算法的应用[J]. 华北工学院学报，2002，23 卷 6 期:449-451.

[3] Qin Hong-shan. Prediction of the Helix/Sheet Content of Proteins from Their Primary Sequences by Neural Net work Method[J]. ransactions of Tianjin University, Dec. 2002, 8, 4:303-307.

[4] Silvio Tosatto. From Sequence to Secondary Structure and Conformation in Proteins[Z]. 2003.

[5] Fetrow Burg and Miller. Artificial Neural Networks for Secondary Structure Prediction[R]. 2004.

References:

[1] Mei Qi-peng. Simplified Encoding of Secondary Structure Prediction [J]. SCI/TECH Information Development & Economy, 2004, Vol.14 No.5:133-134.

[2] Liu Jin-gang. Application of an Improved Arithmetic on BP Neural Networks[J]. Journal of North China Institute of Technology，2002，23, 6:449-451.

[3] Qin Hong-shan. Prediction of the Helix/Sheet Content of Proteins from Their Primary Sequences by Neural Net work Method[J]. Transactions of Tianjin University, Dec. 2002, 8, 4:303-307.

[4] Silvio Tosatto. From Sequence to Secondary Structure and Conformation in Proteins[Z]. 2003.

[5] Fetrow Burg and Miller. Artificial Neural Networks for Secondary Structure Prediction[R]. 2004.

不同编码方法对蛋白质二级结构预测精度的影响

作者： 马栋萍， 阮晓钢
作者单位： 马栋萍(北京工业大学电子信息与控制工程学院,北京,100022;北京联合大学生物化学工程学院,北京,100023)， 阮晓钢(北京工业大学电子信息与控制工程学院,北京,100022)