BP神经网络在蛋白质二级结构预测中的应用

王菲露[1,2], 宋 杰[1], 宋 杨[1,2]

(1. 安徽大学 计算智能与信号处理教育部重点实验室,安徽 合肥 230039;
2. 安徽建筑工业学院 电子与信息工程学院,安徽 合肥 230601)

摘 要:蛋白质结构预测是生物信息学研究的重要问题,而蛋白质二级结构预测是蛋白质结构预测的关键步骤。文中通过BLAST工具得到Identity小于等于35%的46个蛋白质复合物的单链作为数据集,分别采用5位编码和Profile编码,通过不同大小的滑动窗口,对蛋白质二级结构进行预测。实验结果显示,富含"生物进化信息"的Profile编码有着明显的优势,各种精确度均得到了较好的结果,尤其是精确度QE明显高于5位编码的QE。

关键词:蛋白质二级结构;BP神经网络;氨基酸序列;5位编码;Profile编码
中图分类号:TP183 **文献标识码**:A **文章编号**:1673-629X(2009)05-0217-03

Application of BP Neural Network in Protein Secondary Structure Prediction

WANG Fei-lu[1,2], SONG Jie[1], SONG Yang[1,2]

(1. Ministry of Education Key Lab. of Computing Intelligence and Signal Processing, Anhui University, Hefei 230039, China;
2. School of Electronics & Information Engineering, Anhui University of Architecture, Hefei 230601, China)

Abstract: Prediction of protein structure plays an important role in the research of bioinformatics, and prediction of secondary structure is the key step to protein structure predicition. Using BLAST, and get 46 protein's single chains who's identities not more than 35% as the data set. With 5 encoding and Profile encoding, to prediction protein's secondary structure by different information windows. The experiment show that, Profile encoding method which is rich in "biological evolution information" gain the higher precision, the precision of QE is more higher than the precision QE of 5 encoding.

Key words: protein secondary structure; BP netual network; amino acid sequence; 5 encoding; Profile enconding

0 引 言

蛋白质结构预测,对于理解蛋白质空间结构与功能的关系,以及在此基础上进行蛋白质结构的药物设计、突变体等设计具有重要的意义。所谓蛋白质结构预测,是指直接从氨基酸序列推断某一蛋白质的功能位点或预测其三维结构,包括二级和三级结构预测,是目前分子生物学研究中迫切需要解决的问题,是最重要的课题之一[1]。蛋白质二级结构预测是生物信息学研究的热点问题,它不仅是联系蛋白质一级结构和三级结构的纽带,而且是从一级结构预测其三级结构的关键步骤[2]。蛋白质二级结构预测不仅有助于了解蛋白质的功能及其作用机制,对于正确预测蛋白质的空间结构更具有非常重要的意义。蛋白质二级结构预测是计算分子生物学中的分类问题或者说是数学中的多维空间非线性映射问题[3]。目前基于生物化学实验鉴定蛋白质二级结构的方法有多种,但是这种过程既费时又费力,而且存在一定的局限性。与生物化学实验方法相比,基于生物信息学的计算方法具有省时省力的特点,并且它的结果对实验工作者有着指导和辅助的作用。

蛋白质结构预测从模型建立的理论基础的角度上大致可分为四大类:以统计学为基础的模型、以生物化学为基础的模型、机器学习模型(包括最近邻居法、神经网络、支持向量机等)、混合模型[4]。而机器学习模型中的人工神经网络方法由于它的方法多样、适于计算机计算而得到广泛的研究与应用。

收稿日期:2008-09-07
基金项目:国家自然科学基金(60773114);安徽省自然科学基金(KJ2007B239);2009年度安徽省重大自然科学基金(ZD200906)
作者简介:王菲露(1981-),女,硕士研究生,研究方向为智能计算、生物信息学等;宋 杰,博士,副教授,硕士生导师,研究方向为智能计算、生物信息学、嵌入式系统等。

文中基于 BP 神经网络，通过不同的编码方式和不同大小的滑动窗口，选择 BFGS 准牛顿 BP 算法训练函数，对蛋白质二级结构进行预测，得到较好的结果。

1 基本理论

BP(Error Back Proragation)神经网络是一种单向传播的多层前馈神经网络，BP 学习算法是 Rmuehlart 等在 1986 年提出的。它是具有三层或三层以上的神经网络，包括输入层、中间层(隐层)和输出层。上下层之间实现全连接，而每层神经元之间无连接。当学习样本提供给网络后，神经元的激活值从输入层经各中间层向输出层传播，在输出层的各神经元获得网络的输出响应。接下来，按照减少目标输出与实际误差的方向，从输出层经过各个中间层逐层修正各个连接权值，最后回到输入层，这种算法称为“误差反向传播算法”，即 BP 算法。随着这种误差的传播修正不断进行，网络对输入模式相应的正确率也不断上升[5]。

BP 神经网络的传递函数要求必须是可微的，常用的传递函数有：

log－simgoid 型函数：

$$\text{logsig}(n) = 1/(1 + \exp(-n)) \quad (1)$$

tan－simgoid 型函数：

$$\text{tansig}(n) = 2/(1 + \exp(-2 * n)) - 1 \quad (2)$$

线性函数：

$$\text{purelin}(n) = n \quad (3)$$

本实验所用传递函数均为 tan－simgoid 型函数。

2 实现过程

2.1 数据集

数据集的选择对于预测的精确度起着举足轻重的作用。文中借助 BLAST[6]工具在 FariSelli 等[7]建立的数据集基础上进行筛选，排除 Identity 大于 35% 的蛋白质单链，最终选择出 46 个蛋白质复合物的多条单链作为进行蛋白质二级结构预测研究的实验数据集，其中 56 条单链作为训练数据集(见表 1)，另外 21 条单链构成测试数据集(见表 2)。文中选择的训练数据集中：H 螺旋结构的氨基酸有 5448 个，占氨基酸总数的 40.81%；E 折叠结构的氨基酸有 2869 个，占氨基酸总数的 21.49%；C 卷曲结构的氨基酸有 5033 个，占氨基酸总数的 37.70%。文中选择的测试数据集中：H 螺旋结构的氨基酸有 2054 个，占氨基酸总数的 40.64%；E 折叠结构的氨基酸有 1040 个，占氨基酸总数的 20.58%；C 卷曲结构的氨基酸有 1960 个，占氨基酸总数的 38.78%。

在上述数据集的基础上，提取有效的蛋白质特征值作为实验的研究数据。使用 DSSP[8]软件生成每个蛋白质单链的 dssp 文件，从 dssp 文件中提取出每个蛋白质单链的所有氨基酸序列信息和其对应的二级结构信息。并从 ftp://ftp.cmbi.ru.nl/pub/molbio/data/hssp 下载蛋白质的 hssp 文件，从中提取每条单链的 Profile 信息。

表 1 训练数据集

1ABY-A	1ABY-B	1ACY-L	1AD0-B	1AGB-A	1AGB-B	1AGR-A
1AGR-E	1AIS-A	1AIS-B	1ALL-A	1AOK-A	1AQD-A	1AQD-B
1ATN-A	1ATN-D	1AUI-A	1AUI-B	1AXI-A	1AXI-B	1BFV-H
1BFV-L	1BPL-A	1BPL-B	1BRL-A	1BRL-B	1CAU-A	1CAU-B
1EBD-A	1EBD-C	1EFU-A	1EFU-B	1EFV-A	1FIN-A	1FIN-B
1FRV-A	1FRV-B	1GLA-F	1GLA-G	1GUA-A	1GUA-B	1IBC-A
1IBC-B	1IGT-B	1IHF-A	1IHF-B	1LGB-A	1MEL-L	1MHL-A
1MHL-C	1MIO-A	1MIO-B	1NPO-C	1PHN-A	1PHN-B	1RBL-A

表 2 测试数据集

1RBL-M	1RLB-A	1RLB-E	1SCT-B	1SCU-A	1SCU-B	1SEB-D
1TCR-A	1TCR-B	1TMC-B	1TTP-A	1TTP-B	1VOL-A	1YRN-A
1YRN-B	1YUH-L	2BTF-P	2PCC-A	2PCC-B	2REQ-A	2REQ-B

2.2 编码方式

合适的编码方法是决定最终网络性能的重要因素之一，不同的氨基酸编码方式对蛋白质二级结构预测的准确率有着重要影响，用哪种氨基酸编码方式可以取得较高的预测率，还需做进一步研究[9]。目前常用的输入信息的编码表示形式有正交编码、5 位编码、Profile 编码方式等。

每个蛋白质都是由 20 种氨基酸组成的，正交编码是将每种氨基酸用 20 位二进制向量表示，并且满足不同氨基酸的编码向量值的内积为 0。正交编码表示形式的优点在于不引入任何单体间的代数相关。例如，氨基酸 A 的正交编码为：10000000000000000000，氨基酸 Y 的正交编码为：00000000000000000001。

5 位编码方式，即把每个氨基酸对应的十进制整数转化为 5 位二进制编码表示，如氨基酸 A 对应的十进制数值为 1，转化后的 5 位二进制编码为(00001)，而氨基酸 Y 对应的十进制数值为 20，转化后的 5 位二进制编码为(10100)。

Profile 是一个二维数组，第一维对应于序列中的位置序号，每一行是一个 20 维的向量(对应 20 种氨基酸)，向量中的每个元素(对应于数组的第二维)分别代表 20 种氨基酸在这个位置出现的频率。Profile 编码富含“生物进化信息”。文中从蛋白质复合物 hssp 文件中分别提取相应单链的 Profile 编码，采用 5 位编码和 Profile 编码两种方式，对蛋白质二级结构进行预测。

2.3 BP 神经网络预测器的构建

蛋白质二级结构预测是基于已知的一级结构，采用预测方法和技术实现对二级结构的分类预测。对

BP 神经网络而言,输入的是已知蛋白质的一级序列,输出的是二级结构类型。文中使用一个 3 层 BP 神经网络(见图 1)作分类器,对蛋白质二级结构进行预测。设计时,将 BP 网络的输入层设计成一个沿着氨基酸序列滑动的窗口,窗的位置是对称的。该 BP 神经网络的输入层使用了大小分别为 3、5、9 的“滑动窗口”,这些“滑动窗口”都是沿着氨基酸序列“滑动”的窗口,每次都是对输入层“滑动窗口”中间位置的氨基酸的二级结构进行预测。如图 2 中输入层是 VRKKRWACD9 个氨基酸,该时刻就是对窗口中间的氨基酸“R”进行预测,下一时刻将对氨基酸“W”进行预测。

实验采用的 5 位二进制编码,对应的 BP 神经网络输入层“滑动窗口”的神经元个数分别为 3 * 5、5 * 5、9 * 5。采用的 Profile 编码,对应的 BP 神经网络输入层“滑动窗口”的神经元个数分别为 3 * 20、5 * 20、9 * 20。对应的隐含层神经元个数分别为 10、20、20,输出层均是由 3 个神经元组成的,对应三种蛋白质二级结构状态 H 螺旋、E 折叠和 C 卷曲。输出层的 3 个神经元分别编码成 3 维的二进制向量,即 H 螺旋(1 0 0)、E 折叠(0 1 0)、C 卷曲(0 0 1)。假设输出层 3 个神经元输出结果为(0.5 0.6 0.2),则根据“胜者为王”的原则,判定输入层中间的氨基酸残基为 E 折叠(0 1 0)结构。“滑动窗口”不停地移动到下一个氨基酸残基的位置,就可以逐一预测出所有蛋白质氨基酸序列的二级结构。

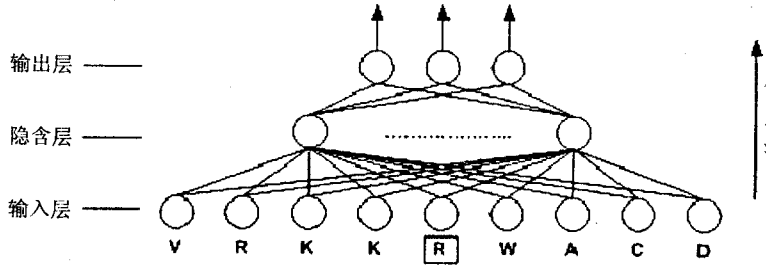

图 1　文中采用的 BP 神经网络

实验选取表 1 中的 56 条蛋白质单链作为训练数据集,分别采用 5 位编码和 Profile 编码方式作为输入层数据,选择 BFGS 准牛顿 BP 算法训练函数,训练 BP 分类器。最后用训练过的 BP 分类器,对表 2 所示测试集中的氨基酸序列进行二级结构预测。如果某个氨基酸残基预测结果为(0.6 0.3 0.2),则判定该氨基酸为 H 螺旋(1,0,0)结构。

3　实验结果评析

选用简单并广泛使用的评估公式来表示预测的精确度,具体公式如下:

$$Q_3 = \frac{P_H + P_E + P_C}{N} \tag{4}$$

Q_3 表示选取数据集的总的预测精度,P_H、P_E、P_C 分别表示数据集中正确预测的每种二级结构状态的氨基酸总数,N 表示数据集中实际氨基酸的总数。

$$Q_H = \frac{P_H}{N_H} \tag{5}$$

$$Q_E = \frac{P_E}{N_E} \tag{6}$$

$$Q_C = \frac{P_C}{N_C} \tag{7}$$

Q_H、Q_E、Q_C 分别表示选取的数据集中三种蛋白质二级结构状态的预测精度,N_H、N_E、N_C 分别表示数据集中每种二级结构实际的氨基酸总数。

本实验使用 BP 神经网络预测蛋白质二级结构,分别选取目标氨基酸残基和与之在序列上相邻的 2 个、4 个、8 个氨基酸,形成氨基酸信息窗大小为 3、5、9 的输入向量,经 BP 网络训练后,再进行预测,分别得出基于数据集的各种精确度率。并最终计算出基于 5 位编码和 Profile 编码的 3 窗口(Win3)、5 窗口(Win5)、9 窗口(Win9)各种预测精确度。具体预测结果见表 3。

从表 3 可看出:随着窗口大小的增加,5 位编码和 Profile 编码的预测精度 Q_3 明显呈上升趋势,其它精确度 Q_H、Q_E 基本也成上升趋势,只有精确度 Q_C 略有下降。5 位编码和 Profile 编码方式的预测精度 Q_3、Q_E 均在 9 窗口(Win9)时获得了最大值。并且 Profile 编码的预测精度 Q_H 和 Q_C 分别达到 75.61%、74.10%,得到了较好的精确度,说明了本实验方法的有效性。

通过结果还看出,Profile 编码的各种预测精度都明显高于 5 位编码,这说明虽然 Profile 编码较为复杂,但由于它富含“生物进化信息”,因而相对 5 位编码方式更适合于一致性较低的蛋白质二级结构的预测。

表 3　实验预测结果

窗口 / 精确度(%)	Win3		Win5		Win9	
	5 位编码	Profile 编码	5 位编码	Profile 编码	5 位编码	Profile 编码
Q_H	62.40	71.48	65.71	75.61	66.55	74.59
Q_E	16.44	36.00	13.65	44.88	17.12	52.40
Q_C	68.99	74.10	67.62	71.05	66.22	71.73
Q_3	55.62	65.23	55.78	67.44	56.25	68.90

4　结束语

研究了蛋白质二级结构预测数据集的选取、参数

(下转第 223 页)

3 岁年龄组小儿发生意外伤害的主要危险因素是坠落 47.6%、跌伤 23.3%、车祸 23.53%,与某市疾病控制中心所提供的研究结论一致[13]。3～6 岁、6～16 岁两个年龄组发生意外伤害的最主要因素均为车祸,分别占 80.77%、77.8%,坠落伤两组分别占 15.4%、2.14%,扎、砸伤两组分别占 3.85%、7.41%。车祸于文献报导明显增高,分析原因为我国机动车增多。砸伤、坠落伤两年龄组较文献报导明显降低,分析原因与家庭、学校、社会高度重视、加强管理、强化自身安全教育有关。

3　结束语

利用关联规则挖掘方法分析出了隐藏在儿童意外伤害的海量数据背后的有效信息,进而从中获取有意义的部分,并从中挖掘出内在规律,以达到防患未然、防微杜渐的目的。然而关联规则挖掘技术在针对儿童意外伤害中的应用不只是文中提到的这几个方面,还包括如何针对这几种常见的意外形式做出有效的保障和预防工作。数据挖掘技术是具有广阔前景的数据处理与分析技术,它将在有大量信息的数据处理工作中发挥不可估量的作用。

参考文献:

[1] Han Jiawei, Kamber M. Data Mining: Concepts and Techniques[M]. 范　明,孟小峰,等译. 北京:机械工业出版社,2001.

[2] 李　锋,孙根年. 旅游目的地灾害事件的影响机理研究[J]. 灾害学,2007,22(3):134－137.

[3] 陈　安,陈　宁. 数据挖掘技术及应用[M]. 北京:科学出版社,2006:50－110.

[4] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases[C]// Proceedings of the ACM SIGMOD International Conference Management of Date. Washington: [s.n.],1993:207－216.

[5] Konias S, Chouvarda I, Vlahavas I, et al. A novel approach for incremental uncertainty rule generation from databases with missing values handling: application to dynamic medical databases[J]. Med Inform Internet Med,2005,30(3):211－215.

[6] 秦亮曦,史忠植. 关联规则研究综述[J]. 广西大学学报:自然科学版,2005(4):48－55.

[7] Agrawal R, Skirant R. Fast Algorithms for Mining Association Rules in Large Databases[J]. California: IBM Almaden Research Center,1994.

[8] 谈恒贵,王文杰,李克双. 频繁项集挖掘算法综述[J]. 计算机仿真,2005(11):1－4.

[9] 梁　循. 数据挖掘:建模、算法、应用和系统[J]. 计算机技术与发展,2006,16(1):1－4.

[10] Wolff R, Schuster A. Association rule mining in peer－to－peer systems[J]. IEEE Trans Syst Man Cybern B Cybern, 2004,34(6):24－26.

[11] 姚卫新,黄丽华. 智能数据分析在医学领域的应用综述[J]. 计算机工程,2004(7):3－5.

[12] 霍　焱,张永文,张莉琴. 儿童意外伤害的健康教育研究[J]. 护理研究,2003(19):130－131.

[13] 尚　磊,徐勇勇,江　逊. 儿童意外伤害住院的情况调查[J]. 疾病控制杂志,2001(3):233－235.

(上接第 219 页)

的设置等,采用不同编码方式和不同大小的滑动窗口,使用 BP 神经网络分类器对蛋白质二级结构进行预测,得到了较好的结果,尤其是 Profile 编码方式获得了较好的实验结果,证明了实验方法的可行性和有效性。为进一步提高预测的精确度,可采用串联 BP 神经网络或其它多种神经网络结合的方式,调整隐含层神经元个数,采用富含生物进化信息的其它编码方式作为输入层数据等方法进行改进,这也是今后需要做的工作。

参考文献:

[1] 马栋萃. 基于改进 BP 神经网络预测蛋白质二级结构[D]. 北京:北京工业大学,2005.

[2] Wang Zhixin. The current situation and prospect of protein structure prediction[J]. Chemistry of Life, 1998,18(6):19－22.

[3] 郑婷婷,毛军军. Profile 覆盖算法在蛋白质二级结构预测中的应用[J]. 计算机技术与发展,2007,17(9):171－172.

[4] 王勇献. 蛋白质二级结构预测的模型与方法研究[D]. 长沙:国防科学技术大学,2004.

[5] 葛哲学,孙志强. 神经网络理论与 MATLAE7 实现[M]. 北京:电子工业出版社,2007.

[6] McGinnis S, Madden T L. BLAST: at the core of a powerful and diverse set of sequence analysis tools[J]. Nucleic Acids Res,2004,32:20－25.

[7] Fariselli P, Pazos F, Valencia A, et al. Predication of protein－protein interaction sites in heterocomplexes with neural networks[J]. Eur. J. Biochem, 2002,269:1356－1361.

[8] Kabsch W, Sander C. Dictionary of Portein seconday structure: patten Recognition of hydrogen－bonded and geometrical features[J]. Biopolymers,1983,22:2577－2637.

[9] 阮晓钢,孙海军. 编码方式对蛋白质二级结构预测精确度的影响[J]. 北京工业大学学报,2005,31(3):229－231.