第 36 卷 第 4 期 计 算 机 工 程 2010 年 2 月
Vol.36 No.4 Computer Engineering February 2010
·博士论文· 文章编号：1000—3428(2010)04—0022—03 文献标识码：A 中图分类号：TP183

基于级联神经网络的蛋白质二级结构预测

王艳春[1,2]，何东健[3]，王守志[4]
(1. 西北农林科技大学机械与电子工程学院，杨陵 712100；2. 青岛农业大学信息科学与工程学院，青岛 266109；
3. 西北农林科技大学信息工程学院，杨陵 712100；4. 威海职业学院机电工程系，威海 264210)

摘 要：为提高蛋白质二级结构预测的精度，提出一种由两层网络构成的级联神经网络模型。第 1 层网络采用具有差异度的 5 个子网构成的网络模型，对第 2 层网络的输入编码进行改进。对 PDBSelect25 中的 36 条蛋白质共 6 122 个残基进行测试，结果表明，该模型能有效预测蛋白质二级结构，其预测精度分别比 SNN, DSC, PREDSATOR 方法提高 5.31%, 1.21%和 0.92%，平均预测精度提高到 69.61%。
关键词：神经网络；蛋白质；二级结构预测

Protein Secondary Structure Prediction Based on Cascade Neural Networks

WANG Yan-chun[1,2], HE Dong-jian[3], WANG Shou-zhi[4]
(1. College of Mechanical and Electronic Engineering, Northwest A & F University, Yangling 712100; 2. College of Information Science and Engineering, Qingdao Agricultural University, Qingdao 266109; 3. College of Information Engineering, Northwest A & F University, Yangling 712100; 4. Department of Mechanical and Electronic Engineering, Weihai Vocational College, Weihai 264210)

【Abstract】In order to improve the prediction accuracy of protein secondary structure, a cascade neural networks composed of two-level network is presented. The first level is composed of five subnets with different structure, and the coding method of the second-level is studied and improved. The model is employed to predict 36 nonhomologous protein sequences with 6 122 residues in PDBSelect25. Results show that the proposed model can efficiently improve the prediction accuracy, increasing the prediction accuracy by 5.31%, 1.21% and 0.92% respectively compared with SNN, DSC and PREDSATOR method, improving the average prediction accuracy to 69.61%.
【Key words】neural networks; protein; secondary structure prediction

1 概述

蛋白质二级结构预测是后基因组时代的一项重要任务。人工神经网络方法被认为是目前应用最广、前景最乐观的方法之一。然而，采用局部信息为输入的单神经网络方法的预测精度不高，一般在 65%左右。近 10 年来，人们利用序列同源性信息将预测精度提高到 70%~80%。但这种算法不仅计算量大，而且对于那些低同源和无同源蛋白质的预测非常困难。在最坏的情况下，预测精度将下降 10 个百分点。使用单一序列的二级结构预测方法，虽然预测精度偏低，但不受同源模版的限制，具有快捷、普适的特点。因此，基于单序列的蛋白质二级结构预测仍然是蛋白质研究中的有力工具。

本文采用一种基于 2 层的复合神经网络模型，结合局部信息来预测蛋白质二级结构。利用 36 个非同源蛋白质序列共 6 122 个氨基酸残基，对提出的复合级联神经网络预测方法进行了验证，仿真结果表明该方法可以较好地预测蛋白质二级结构。

2 复合级联神经网络模型

本文提出如图 1 所示的复合级联神经网络模型来预测蛋白质二级结构。理论上早已证明[1]：只包含一个隐层的 3 层 BP 神经网络，能够以期望的精度逼近任意非线性函数。模型中的基本组成单元均采用包含一个隐含层的 3 层 BP 网。网络隐含层和输出层的传递函数为 Sigmoid 函数，采用附加动量法来避免网络陷入局部极小值。

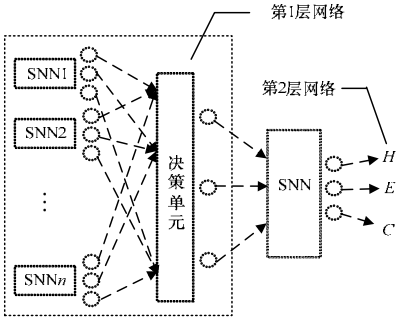

图 1 复合级联神经网络结构

由于蛋白质二级结构本身就是氨基酸之间相互作用的结果，从这个意义上讲，单个氨基酸所包含的二级结构信息量很少，还必须考虑相邻残基的信息，因此在每个 BP 网络的输入层引入“滑动窗口”技术。

预测时，窗口沿着氨基酸序列依次滑动，窗口的位置是对称的，每次预测都是对窗口中心的氨基酸进行的。根据文献[2]研究窗口宽度对预测精度影响所得的结论，本文将所有子网的窗口宽度取为 13。

基金项目：国家自然科学基金资助项目(30471138)
作者简介：王艳春(1972－)，女，讲师、博士研究生，主研方向：生物图像，计算机视觉技术；何东健，教授、博士、博士生导师；王守志，讲师、博士研究生
收稿日期：2009-09-20 **E-mail：**wychun99@sina.com

2.1 第 1 层网络结构

如图 1 所示，第 1 层网络由多个 BP 网和一个决策单元组成。一个单隐层 BP 网可以看作是实现从输入状态到输出状态的非线性映射器，并且映射形式由一组可调节的连接权所控制。通常情况下，网络的初始连接权都是随机获取的，采用误差函数梯度下降法在训练中逐步调整，最终得到一个较好的权值分布。这使网络不可避免地产生易陷入局部极小值的问题，且具体的极值位置又与初始权值密切相关。因此，初始权值的随机性影响着网络的性能，是网络泛化能力差的一个主要原因。

为了减少这种随机性带来的影响，同时考虑模型的复杂度，本文采用 5 个 BP 网络组成复合网络结构。

复合神经网络模型的泛化能力由各个子网的泛化能力以及它们之间的差异度共同决定，而差异度是由使用的训练集和网络结构等随机性产生的。早期研究表明[2-3]，网络隐层节点数对网络性能并不起决定性的作用，可以在较大范围内取值。在本模型中，通过改变各子网隐层单元数来增加差异性。5 个子网的隐层单元数根据仿真确定为 20~40 之间，隔 5 个取一次。

根据“相对多数”原则(当且仅当输出结果为某分类的神经网络数目最多，该分类才成为最终输出结果)，用一个决策单元对各子网的输出结果进行整合。如果有 2 个相等的输出结果，取二级结构所占比例小的二级结构状态为输出结果。

2.2 第 2 层网络结构

一个蛋白质二级结构是具有相同结构态的连续氨基酸残基的集合。也就是说，如果一个氨基酸处于某种结构时，绝大多数情况下，它邻近(左或右)的部分氨基酸也应该处于这种结构，这就是相邻残基之间的相关性。单层网络预测精度偏低，主要是缺乏对这种相关性的“认识”。为了更多地参考序列中相邻残基之间的相关性，在上述网络模型的基础上引入第 2 层网络，其主要作用是对第 1 层网络的输出结果进行精炼。仿真实验表明第 2 层网络是有效的，是对第 1 层网络有益的补充。在第 2 层网络中仅包含一个单隐层 BP 网，结构见图 2。

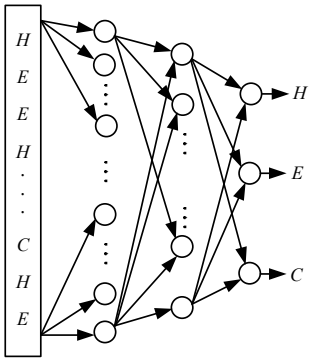

图 2　子神经网络结构

在该层网络中，隐层单元根据仿真设定为 20；输出层由 3 个单元组成。网络的最终结果由最大的输出类别节点按照“胜者通吃”原则来决定。

3 编码方法

3.1 第 1 层网络的编码

在蛋白质二级结构预测研究领域中，正交编码是一种较为常用的编码方式。它是用 20 位二进制数来唯一标识某一种氨基酸，并且满足不同氨基酸的编码向量值的内积为 0。由于这种编码方法未引入任何单体之间的代数相关，因此被研究者广泛使用。

本文采用 21 位正交编码方法对该层 5 个子网的输入进行编码。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，其他神经元的编码均为 0，“空窗口”由前 20 位为 0，最后一位为 1 表示。

输出层由 3 个神经元组成，以对应蛋白质 3 种二级结构状态。这 3 个神经元用三维二进制数编码来表示，即 α 螺旋对应 100，β 折叠对应 010，无规则卷曲 C 对应 001。

3.2 第 2 层网络的编码

在已有的 2 层网络结构模型中，通常只是把第 1 层网络的输出简单编码为 100, 010 和 001 作为第 2 层网络的输入。这种编码方法既没有充分利用第 1 层网络的预测结果也没有提供给第 2 层网络较多的信息，整体预测精度没有多大的提高。鉴于此，本文引入了可靠性分数指标[4](Reliability Score, RS)，其值由第 1 层网络的输出来确定。假定第 1 层网络的 3 个输出值为 $x1, x2, x3$，则 RS 为

$$RS = X - X' \tag{1}$$

其中，

$$X = \max\{x_1, x_2, x_3\} \tag{2}$$

$$X' = \text{second}\max\{x_1, x_2, x_3\} \tag{3}$$

文献[4]从统计学角度证实 RS 值能很好地反映每个残基位置的预测可信度，并且随着 RS 值的增大，预测精度呈上升趋势。

因此，第 2 层网络的输入层编码由两部分组成，可以用一个行向量(S, C)来表示。其中，S 表示第 1 层网络的预测结果，用三维向量来表示，α 螺旋编码为 100，β 折叠编码为 010，卷曲编码为 001；C 表示残基的 RS 值，用一维向量表示，则输入层的向量维数为 4×13(窗口的宽度)。假设对第 1 层的某个残基预测的 3 个输出为 H(0.87), E(0.46), C(0)，则 RS 的值为 0.87-0.46=0.41，因此，该残基对应的第 2 层网络的输入编码为(1 0 0 0.41)。

第 2 层网络的输出仍是窗口中心残基的结构类别，由 3 个神经元组成，编码方法同第 1 层网络。

4 测试数据与仿真结果

4.1 训练与测试数据集

文献[5]将 PDB 数据库中已知结构划分为 Helices, Sheets 和 Coils 这 3 类。其中，G, H, I 属于 Helices，记作 H；B, E 属于 Sheets，记作 E；S, T, C 属于 Coils，记作 C。采用的蛋白质数据均来源于 EMBL 的非冗余 PDB 子集 PDBSelect25。该数据集中的数据都是 PDB 数据库中同源性<25%的蛋白质数据，共 1 771 条序列，总残基数为 297 372。从中随机提取 36 个蛋白质，分别为 1JQLB, 1EYPA, 1E0CA, 1YTFD, 1PDNC, 3YGSP, 1KB5B, 1B6BA, 1JB0D, 1DTP_, 1JEJA, 1MSPB, 1HUP_, 2OCCG, 2OCCF, 1QO3D, 8PRN_, 1BFEA, 1G7RA, 1P32B, 1EVSA, 1TUPC, 1DYNA, 1AW8E, 1LLIA, 1OTGA, 1HZIA, 1FZRA, 1HQZ7, 1G13A, 1A1X_, 1EL6A, 1D6JA, 1AMX_, 1I9YA, 1VID_。每个蛋白质链的长度大于 80，共 6 122 个残基组成本研究数据集。其中，H 占 30%，E 占 29%，C 占 41%。

4.2 评价方法

选用最常用的 Q_3 正确率法，即所有正确预测 3 种二级结构残基的百分比，其计算公式如下：

$$Q_3 = (P_H + P_E + P_C)/N \tag{4}$$

其中，P_H, P_E, P_C 分别表示被正确预测出的三态{H, E, C}的残基个数；N 为残基总数。

三态准确率给出了具体的 3 种二级结构的预测精度：

$$Q_i = P_i/N_i \tag{5}$$

其中，$i \in \{H,E,C\}$；N_i为训练集或检验集里$\{H,E,C\}$的总数目。

4.3 仿真结果

采用 7 重交叉验证方法对模型进行验证，以 7 次预测结果的平均值作为最终预测结果。网络初始化时，隐层和输出层节点的权值在-1 和 1 之间随机赋值，BP 算法的迭代次数设定为 1 000。程序用 C 语言实现，实验平台为 1.6 GHz P4 处理器，768 MB 内存，Windows XP Professional。

表 1 给出了第 1 层网络中各子网及该层网络的平均预测精度。从表 1 可以看出，无论是从整体预测准确率还是从三态准确率来看，复合网络的预测精度都要高于各个子网。与预测精度最低的个体网络相比，复合网络的整体预测准确率要高出 8.5%，而且这个结果较文献[6]的预测结果高出 2.1%。

表 1 第 1 层网络的预测精度

类别	SNN1	SNN2	SNN3	SNN4	SNN5	复合网络
Q_3	0.608 1	0.597 1	0.635 6	0.624 3	0.613 7	0.682 4
Q_H	0.554 0	0.516 4	0.580 9	0.580 2	0.600 4	0.630 7
Q_E	0.516 4	0.490 5	0.548 0	0.490 5	0.517 0	0.615 7

为了对第 2 层编码中 RS 信息的有效性进行验证，比较了本模型的编码方法与传统编码方法。表 2 给出了 2 种编码方法对 6 个蛋白质进行预测的准确率。

表 2 6 个蛋白质的最终预测精度 (%)

类别	1QO3D	1TUPC	1FZRA	1AMX_	1G13A	1VID_
传统方法 Q_3	68.42	69.01	68.24	68.33	68.57	68.24
本模型 Q_3	69.65	70.13	69.41	69.46	69.68	69.37

从表 2 可以看出本模型预测精度要优于传统模型方法，其最优预测精度达到了 70.13%，比第 1 层网络的平均预测精度高出 1.8%，而传统方法的最好预测精度只有 69.01%。比较结果表明，在第 2 层网络中加入 RS 信息，可有效提高网络的预测精度，同时说明基于两层网络结构的蛋白质二级结构预测方法的预测精度要高于单层网络结构。

表 3 给出了本模型与其他模型的预测精度比较。

表 3 几种模型预测精度的比较

方法	Q_3/(%)
SNN	64.30
DSC	68.40
PREDATOR	68.69
PHD	72.20
本文模型	69.61

由表 3 可以看出，本文提出的预测方法具有较好的预测精度，平均预测准确率为 69.61%，较基于单层结构的 SNN 方法提高了 5.31%，较其他基于 2 层结构的 DSC, PREDATOR 方法分别提高了 1.21%和 0.92%，并且与基于同源信息的 PHD 方法的预测精度接近，说明本模型是有效的，能很好地提高预测的精度。

5 结束语

本模型通过增加第 1 层结构中各子网的差异性提高了复合网络的泛化能力，使用“相对多数”原则对各子网的结果进行整合，降低了预测结果的不确定性，提高了结构预测的精度。基于相邻残基之间的相关性，在第 1 层网络的基础上引入第 2 层网络，并且把表征第 1 层网络预测结果的可靠性指标(RS)信息融合到第 2 层网络的编码中，进一步提高了预测精度。

本文提出的网络模型可以很好地预测蛋白质二级结构。尽管目前该方法还不如 PHD 方法的预测精度高，但是本模型只是基于单序列进行结构预测，没有引入可以将预测精度提高 5%~10%的同源信息。以一种全新的方法启发人们从另外一个角度来考虑蛋白质二级结构预测问题，从而将预测精度提高到一个新的水平。

参考文献

[1] Hornik K M, Stinchcombe M, White H. Multilayer Feed Forward Networks Are Universal Approximators[J]. Neural Networks, 1989, 2(2): 359-366.

[2] Qian Ning, Sejnowski T J. Predicting the Secondary Structure of Globular Proteins Using Neural Network Modals[J]. Journal of Molecular Biology, 1988, 202(4): 865-884.

[3] Holley L H, Karplus M. Protein Secondary Structure Prediction with a Neural Network[J]. Proceedings of the National Academy of Sciences, 1989, 86(1): 152-156.

[4] Xin Huang, Huang De-Shuang, Zhang Guang-Zheng, et al. Prediction of Protein Secondary Structure Using Improved Two-level Neural Network Architecture[J]. Protein & Peptide Letter, 2005, 12(8): 805-811.

[5] Kabsch W, Sander C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen Bonded and Geometrical Features[J]. Biopolymers, 1983, 22(12): 2577-2637.

[6] Zhu Hanxi, Yoshihara I, Yamamori K. Prediction of Protein Secondary Structure by Multi-modal Neural Networks[C]//Proc. of International Joint Conference on Neural Networks. [S. l.]: IEEE Press, 2002: 280-285.

编辑 顾逸斐

(上接第 21 页)

桩信息传输进行了优化后的程序。实验结果如表 1 所示。

表 1 程序运行时间对比 ms

实验对象类型	第 1 组运行时间	第 2 组运行时间	第 3 组运行时间
无插桩	175 401.7	139 124.6	156 219.1
传统插桩	184 626.7	146 259.3	164 331.2
优化插桩	179 553.2	142 525.7	160 082.3

从上述实验结果可知，使用本文所提出的优化方法对桩信息的通信进行优化测试的程序运行时间与用传统插桩方法没有进行优化的程序进行仿真测试的时间相比，每组实验由于插桩导致的运行时间的额外开销分别降低了 55.0%, 52.3%和 2.5%，并且 2 种方法得出的覆盖率是相同的，没有出现覆盖率丢失的现象。这说明了本文所提出的插桩方法在降低对程序实时性影响的方面是有效的。

6 结束语

本文针对覆盖测试在 Host-Target 仿真测试模式下遇到的问题，即程序插桩对源程序造成的时间性能方面的影响，针对桩信息在目标机与宿主机之间的传输提出了一种优化方法。在没有丢失覆盖信息的情况下降低了桩信息传输导致的通信时间的增加，有效地提高了仿真测试的执行效率。

参考文献

[1] 丁 旭, 崔吉岗, 刘春裕. 军用嵌入式软件结构覆盖测试技术[J]. 指挥控制与仿真, 2008, 30(3): 120-122.

[2] 王学东, 汪文勇. 汇编程序覆盖测试中虚拟插桩的实现[J]. 计算机工程, 2007, 33(7): 87-88.

[3] 刘慧梅, 徐华宇. 软件测试中代码分析与插装技术的研究[J]. 计算机工程, 2007, 33(1): 86-88.

[4] 孙昌爱, 金茂忠. 基于程序插装的动态测试技术的实现[J]. 小型微型计算机系统, 2001, 22(12): 1475-1479.

[5] 杜 延, 刘从越. 嵌入式实时系统软件测试实践[J]. 微计算机信息, 2007, 23(11): 86-89.

编辑 顾逸斐