第 24 卷第 6 期　　四川理工学院学报(自然科学版)　　Vol. 24　No. 6
2011 年 12 月　　Journal of Sichuan University of Science & Engineering(Natural Science Edition)　　Dec. 2011
文章编号:1673-1549(2011)06-0666-05

基于多模集成神经网络的蛋白质二级结构预测

李林江,伍勇强 ,曾黄麟 ,张金山

(四川理工学院理学院,四川 自贡 643000)

摘　要:为提高蛋白质二级结构预测的精度,提出了一个由 5 个子网络集成的多模神经网络模型,预测结果由 5 个子网络综合得到。对于每个子网络采用神经网络分级思想分为二级网络,一级网络采用含进化信息的 profile 编码蛋白质序列作为输入,二级结构作为输出。二级网络编码一级网络输出结果作为输入,并将蛋白质序列用改进正交编码方式作为另一输入来提高二级网络的预测精度,输出仍为二级结构。采用子网络差异方式进行单独训练,结果表明该方法最终的预测精度达到 71.3%,较大提高了蛋白质二级结构的预测精度。

关键词:神经网络;二级结构预测;改进正交编码;预测精度
中图分类号:TP183　　文献标识码:A

引 言

蛋白质二级结构的预测是生物信息学中一个重要的课题,进行二级结构预测对于理解蛋白质结构与功能的关系,以及分子设计、生物制药领域有很重要的现实意义。1988 年 Qian 和 Sejnowski 提出用神经网络方法预测蛋白质二级结构[1],随即引起了神经网络预测蛋白质二级结构的高潮,Zhu 用多模神经网络平均预测精度为 66%[2],王艳春等用级联神经网络平均预测精度为 69.61%[3]。

本文采用了由 5 个二级神经网络集成的多模神经网络模型,使用了两种蛋白质序列编码手段对二级结构预测。利用文[2]中的 36 个非同源蛋白质样本对提出的网络模型进行验证,在 matlab 上仿真得出平均预测精度为 71.3%,提高了预测精度。

1　编码方式

在将神经网络应用于预测蛋白质二级结构之前,有必要讨论一下序列输入这一重要的问题。不同的编码方案必然会对输入窗所处空间的复杂度、神经网络的结构和学习的难易程度产生影响[4],需要综合考虑各方面因素选择编码方式。在二级结构预测中正交编码和 profile 编码是两种最常用的编码方式。

正交编码用 20 位二值节点编码,依次使一位为 1,其余位为 0,即可表示 20 种氨基酸,如代码为 H 的组氨酸编码为 $[00000000000000000001]^T$,代码为 V 的缬氨酸编码为 $[01000000000000000000]^T$。这种编码方式优点在于不同氨基酸编码向量值的内积为 0,不会引入任何单体间的代数相关。

而 profile 编码被认为携带了丰富的生物进化信息,一般认为生物进化信息对于提高蛋白质二级结构预测准确率是非常重要的。profile 是指在蛋白质序列的每一个位置上氨基酸在进化过程中出现的相对概率,可以在 HSSP 数据(ftp://ftp.ebi.ac.uk/pub/databases/hssp/)中得到。下表 1 中给出了 Profile 编码的一般格式,其中 Seq No 代表序列位置,AA 列代表该蛋白质相应位置的氨基酸类别,后面 20 列代表氨基酸在同源蛋白质中出现在此位置的次数,即表示蛋白质在进化过程中该位置氨基酸替换的情况,体现出了进化信息。

使用 profile 编码时需要计算该位置出现某氨基酸的概率,如对与表中的第二号氨基酸 Y,其 20 位编码为 $[0.01\ 0.02\ 0.03\ 0\ 0.15\ 0.02\ 0.75\ 0\ 0\ 0\ 0\ 0\ 0\ 0.01\ 0.01\ 0\ 0\ 0\ 0]^T$。

收稿日期:2011-09-20
基金项目:四川省技术厅应用基础研究专项课题(2011JY0051);四川省白酒及生物技术重点实验室重点专项课题(NJ2010-01)
作者简介:李林江(1989-),男,四川武胜人,主要从事生物信息学及计算数学方面的研究,(E-mial)li_linjiang@126.com

表 1　Profile 编码的一般格式

SeqNo	AA	V	L	I	M	F	W	Y	G	A	P	S	T	C	H	R	K	Q	E	N	D
1	T	0	0	0	0	0	0	0	0	0	0	55	19	0	8	0	0	0	19	0	0
2	Y	1	2	3	0	15	2	75	0	0	0	0	0	0	1	1	0	0	0	0	0
3	T	2	1	0	0	0	0	0	3	2	0	28	32	0	4	9	1	7	8	2	0
4	T	2	6	4	0	2	0	0	0	2	3	7	46	0	0	3	1	0	0	3	0
5	R	12	2	4	0	0	0	0	0	3	0	1	2	0	2	33	9	3	26	0	1

2　改进的多模集成神经网络

2.1　多模神经网络

通过单个神经网络(single neural network, SNN) 预测精度不高[2],可以考虑通过训练多个神经网络将结果进行综合来提高预测精度。每个神经网络采用不同的训练集,提高神经网络系统的泛化能力。本文采用 5 个子网络集成的多模神经网络(multi - modal neural networks, MNN) ,网络拓扑结构图 1。

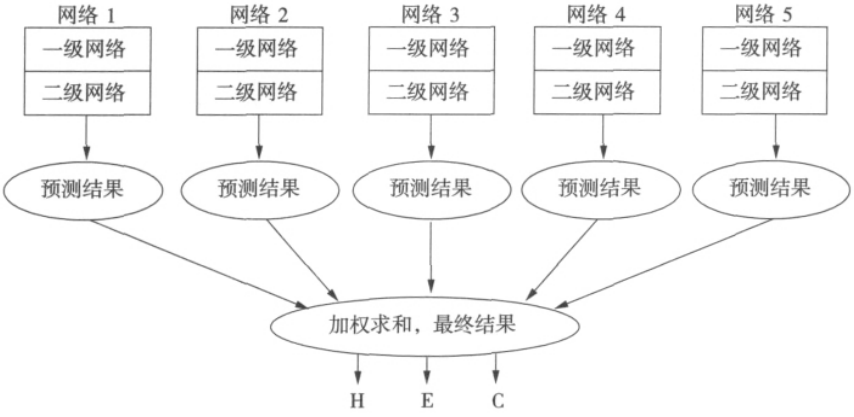

图 1　蛋白质二级结构预测多模集成神经网络

对于这样一个集成网络来说,每个子网络结构完全相同,是一个分级的二级 BP 神经网络。将 5 个子网络集成在一起共同预测同一个蛋白质得到 5 个不同的结果,本文采用均权求和得到最终结果避免了文献[5]出现票数相同的情况。

2.2　一级网络

如图 1 所示,每个子网络采用神经网络分级思想分为二级网络。在其学习过程中,根据输入的一级序列和二级结构的关系不断调整各节点的权重,最终目标是找到一种理想的输入与输出映射,一级网络结构如图 2。

一级网络输入层对应蛋白质的一级序列,采用 20 位 profile 编码作为输入。单个氨基酸所包含的二级结构信息量还很少,因为结构的形成是相邻氨基酸的共同作用,因此在每个网络的输入层引入“滑动窗口”技术,将残基按连续片段输入。在对整条序列进行预测时,窗口沿着氨基酸序列从第一个残基位置开始滑动至最后一个残基位置结束,对窗口中间位置的氨基酸结构进行预测,窗口一般取奇数。

在文[2]对不同窗口长度对精度的影响进行了探讨,

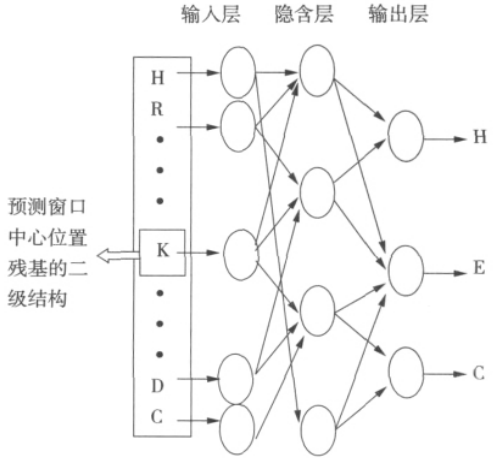

图 2　一级网络结构

本文窗口长度 N 统一选取为 13。所以一级网络输出层神经元个数为 20 × 13 = 260(个) ,对于序列两端会出现窗口不满的情形,用全为 0 的“空窗口”将其补满。

隐层节点数与求解问题的规模,输入输出单元数多少都有直接关系。通过多次训练测试,本文选取 26 个隐含层结点数。传输函数采用 log*sig* 函数,函数表达式为:

$$logsig(n) = \frac{1}{1+e^{-n}} \quad (1)$$

神经网络的输出端是蛋白质的二级结构,其编码是 3 维向量 H [0 0 1]^T, E [0 1 0]^T, C [1 0 0]^T,故输出层神经元设定为 3 个。

2.3 二级网络

在蛋白质结构预测中,单级神经网络模型预测法都存在一个缺点,如结构 *HHHHHHHH* 经常被预测为 *HHHHHCHH* 等。问题便在于单级网络没有完全"掌握"二级结构片段的平均长度,而二级结构本身的构成是一段连续的氨基酸残基。为了对一级网络的结果精炼,在一级网络的基础上串联一个二级网络,它是以一级网络的输出信号作为输入信号。

我们在这里引入了可靠性分数指标 RS (reliability score)[3],设第一级网络的 3 个输出值是 x_1, x_2, x_3,则

$$RS = \max\{x_1, x_2, x_3\} - second\max\{x_1, x_2, x_3\}$$

其中 $\max\{x_1, x_2, x_3\}$ 是 x_1, x_2, x_3 中最大的数 $second\max\{x_1, x_2, x_3\}$ 为 x_1, x_2, x_3 中第二大的数。

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。使用二级网络对一级网络的结果进行精练,就是利用 RS 值大的去影响邻近那些 RS 值较小的,从而得到正确的结果。Rost 和 Sander 曾经指出[7],加入二级网络只是能够使片段的长度更符合实际情况,但是对整体预测精度的提高作用不明显。说明精度的提高关键在于一级网络,其原因是一级网络是从序列到结构的训练,二级网络只是结构到结构的训练,那么如能在二级网络中加入序列的信息,就能够对预测精度提高起较为明显作用。因此本文在二级网络中另外加入了一个新的输入分量——蛋白质序列的信息,网络结构如图 3。

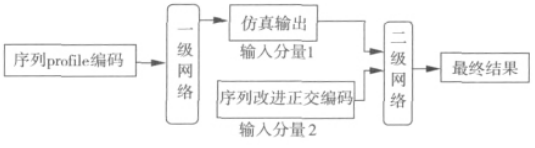

图 3 二级网络的框架

从图 3 看出二级网络具有两个输入分量,第一个分量是一级网络仿真的输出信号,第二个是蛋白质一级序列信息。二级网络输入较为复杂,采用两个隐层 BP 神经网络,以保证可以剔除冗余信息,两个隐含层节点数分别选取 40 个和 30 个,传输函数选取原理,输出层节点数都与一级网络相同。

2.4 二级网络输入优化编码

在二级网络第一个输入分量编码时加入 RS 值,比如一级网络输出为 [0.6 0.4 0.1]^T 表示此残基结构为 C 类,先将其转化编码 [1 0 0]^T,其 RS = 0.6 − 0.4 = 0.2,然后编码为 [1 0 0 0.2]^T,类似于一级网络我们依旧采用滑动窗口技术,以使得相邻残基互相影响,RS 值大的纠正 RS 值小的残基,所以二级网络第一个输入分量所占的神经元的个数就为 $4N$ 个。

对于二级网络的第二个输入分量和一级网络的输入一样,都是相同蛋白质的序列信息,可以直接利用一级网络的输入作为此处的输入。在训练的过程中,发现收敛速度很慢,收敛效果非常不理想,如图 4。

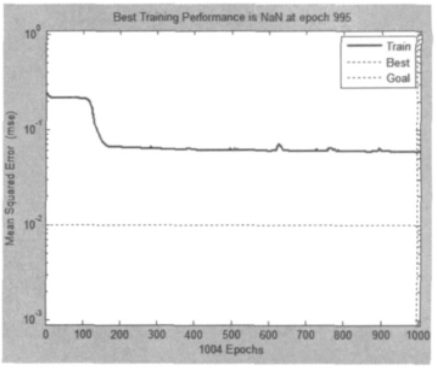

图 4 二级网络用 profile 编码序列收敛图

仔细分析原因,可能是由于一级网络的 profile 编码方式对于两个输入分量的二级网络不适用,加大了二级网络的输入窗空间复度和网络的学习难度。本文在二级网络的序列编码中,放弃 profile 编码尝试采用正交编码方式进行训练。正交编码是 0,1 二值节点的编码,任意两个不同氨基酸编码是正交的,而 profile 编码取值是 [0,1] 之间的任意值,故认为正交编码简化了网络的复杂度和非线性因素。这个简化思想在二级网络第一输入分量编码的时候就已经使用到,比如一级结构输出是 [0.6 0.4 0.1]^T 要先简化为 [1 0 0]^T 表示,然后再加上 RS 值 0.2,也要比直接用 [0.6 0.4 0.1]^T 作为输入要简单。采用正交编码后,收敛效果要好很多,如图 5。

而蛋白质二级结构预测中的一个难题就是怎样考

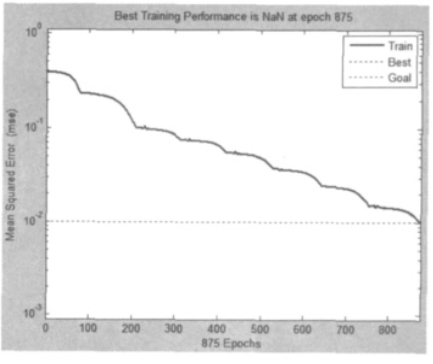

图 5 二级网络用正交编码序列收敛图

虑蛋白质内部相距较远的氨基酸残基在折叠过程中可能产生的远程作用力[8]，所以在输入窗中尽可能多的加入蛋白质的全局信息会对预测结果有所提高。

本文改进正交编码，在 20 位正交编码后加一位表示这种氨基酸在这条蛋白质链所有残基中所占的比例，比如某个蛋白质序列中，H 所占比例为 0.1，V 所占比例为 0.08，则 H 与 V 的编码为

H:　[0000000000000000000 10.1]^T。

V:　[0100000000000000000 0.08]^T。

不同的蛋白质样本中氨基酸所占比例不一样，最后一位编码也不同。这样第二个输入分量所占的神经元个数为 $21N$。调试结果显示加入氨基酸成分信息的改进正交编码要比原来的正交编码精度提高 1.5% 左右，且对网络的复杂程度影响不大。

这样二级网络两个输入分量共有 $21N + 4N = 325$ 个神经元，其结构较为复杂，尤其要注意选择编码方式降低网络的输入复杂度。

3　蛋白质二级结构预测的评价

本文预测精度采用蛋最常见的蛋白质二级结构预测准确率来衡量，三态(H、E、C) 单个准确率计算公式为:

$$Q_i = \frac{q_i}{n_i}, (i = H, E, C) \tag{2}$$

其中: Q_i 表示 i 类二级结构预测的准确率，q_i 表示准确预测 i 类二级结构的残基数目，n_i 表示 i 类二级结构结构的残基总数。

最终的预测精度是指整体预测准确率 Q_3 为:

$$Q_3 = \frac{(q_H + q_E + q_w)}{N} \tag{3}$$

其中: N 表示蛋白质残基总数。

4　数据的训练与仿真

HSSP 数据库二级结构分为 8 类，文献[9] 将其中 H、G、I 结构归为螺旋类表示为 H，B、E 结构归为折叠类表示为 E，其余归为卷曲类表示为 C，本文亦采用这种归类方法将二级结构归为 3 类。并在 CB513 中选取了 36 个非同源蛋白质作为样本集如图 6。

残基数量为 6868，H、E、C 所占比例为 28%，22%，49%，接近自然比例 3∶2∶5，文献[4] 提到样本集中二级结构各占比例越接近自然比例，在二级结构预测时效果也就越好。为了方便训练，将训练集均分为六组，每组残基数量为 1148. 对已经分组的样本，每个子网络选择互不相同的一组作为训练集共 5 组，剩下一组作为共同

1azu	1bbp	1cbh	1cc5	1cdt
1cse	1eca	1fc2	1fdl	1ow0
1r09	1tgs	2aat	2ccy	2gbp
2mev	2mhu	2rsp	2stw	2tgp
2wrq	3ebx	3hmg	3rnt	4cpv
51dh	6acn	6cpp	6hir	7rsa
1crn	1pyp	2hmz	2utg	5er2
9wga				

图 6　蛋白质样本

的测试集，避免了将所有样本在一个网络上训练出现收敛效果不好时间长等问题。每个子网络采用了不同的训练集，这种有差异的训练方式提高了多模神经网络的泛化能力和整体的预测精度，见图 7。

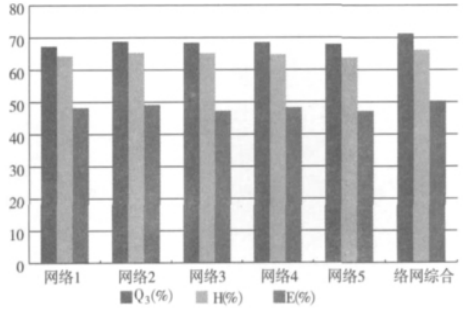

图 7　子网络与多模网络精度比较(无单位)

从图 7 可以看出，多模神经网络的预测结果要比单个网络预测精度要高得多。

采用交叉验证对模型进行验证，具体做法是: 随机选取一组作为测试集，其余 5 组作为样本集分别分给 5 个网络，进行训练得出测试精度。依次重复六次，保证每次作为测试的样本不同，得到六次预测精度取其平均值为模型的预测精度，见表 2。

表 2　交叉验证精度

	Q3(%)	H(%)	E(%)
测试 1 组	71.72	63.81	49.46
测试 2 组	70.94	65.53	46.56
测试 3 组	71.42	67.44	48.35
测试 4 组	71.54	68.53	47.52
测试 5 组	71.15	66.46	45.86
测试 6 组	71.03	66.62	47.11
平均	71.3	66.39	47.47

相比文献[3] 用交叉验证得到的 69.3% 的平均预测精度，本文预测精度取得了较大提高，分析其原因是我们一级采用了富含进化信息的 profile 编码，二级网络也输入了蛋白质序列信息并加入了成分信息在里面，进化信息和全局信息都是有明确生物学意义，可以较大的提高蛋白质二级结构预测的精度。

5 结 论

本文借助多模神经网络对蛋白质二级结构预测进行研究,为提高预测精度一级网络采用了富含进化信息的 profile 编码,并在二级网络加入了一级序列信息,这个改进改变了二级网络不能够较大提高预测精度的观点[7]。本文在二级网络中对序列的编码是改进的正交编码,正交编码比起 profile 编码的优点在于它是单体间无代数相关的二值节点编码可以简化网络的非线性复杂性。而且改进正交编码加入了每种氨基酸在整个肽链中所占比例,考虑到了蛋白质二级结构形成远程作用的问题[8]这种有明确生物学意义的全局信息对蛋白质二级结构精度提高有较大的帮助。

本文对编码方式和网络结构研究改进得到了较高的蛋白质二级结构预测精度,在以后的研究中可以考虑选取不同二级结构比例的样本集以及其他统计方法与神经网络方法相结合来提高蛋白质二级结构预测精度。

参 考 文 献:

[1] Qian Ning, Sejnowski T J. Predicting the Secondary Structure of Globular Proteins Using Network Modals [J]. Journal of Molecular Biology, 1988, 202, 865-884.

[2] Hanxi Zhu, Ikuo Yoshihhara, Kunihito Yamamori, Prediction of Protein Secondary Structure by Multi-Modal Neural Networks [C]. Proc. of International Joint Conference on Neural Networks. [S. 1.]: IEEE Press, 2002, 280-285.

[3] 王艳春,何东健,王守志. 基于级联神经网络的蛋白质二级结构预测 [J]. 计算机工程,2010,36(4):22-24.

[4] Pierre Baldi, Soren Brunak, 著. 张东晖,译. 生物信息学—机器学习方法 [M]. 北京: 中信出版社,2003.

[5] 韩 敏,林丽玉. 基于神经网络集成的蛋白质二级结构预测模型 [J]. 计算机与应用化学,2006,23(10).

[6] 林丽玉. 基于神经网络的蛋白质二级结构预测的研究 [D]. 大连: 大连理工大学博士论文,2005.

[7] B Rost, C Sander. Secondary structure prediction of all-helical proteins in two states [J]. Protein Engineering, 1993, 6(8): 831-836.

[8] 张阳德. 生物信息学 [M]. 北京: 科学出版社,2009.

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

Prediction of Protein Secondary Structure Based on a Multi-modal Integrated Neural Network

LI Lin-jiang, WU Yong-qiang, ZENG Hang-lin, ZHANG Jin-shan

(School of Science, Sichuan University of Science & Engineering, Zigong 643000, China)

Abstract: In order to improve the prediction accuracy of protein secondary structure, a model of multi-modal neural network integrated of five sub-networks is presented. Prediction comprehensive result of protein secondary structure from 5 sub-networks is got. Each sub-network using neural network classification is divided into two-level network. The protein sequences are encoded as the inputs to the first-level networks by using profile encoding which is thought to carry more evolution information. The outputs of first-level networks are secondary structure. Then, the outputs of the first-level networks are encoded as the inputs of the second-level networks. Meanwhile, the sequences are encoded by using improved orthogonal encoding as the other part of inputs of second-level networks. The outputs of second-level networks are also secondary structure. Different sub-networks are trained solely. The result demonstrates that high prediction accuracy of protein secondary structure can be got by our proposed method at 71.3%.

Key words: neural network; secondary structure prediction; improved orthogonal encoding; prediction accuracy