神经网络方法预测蛋白质二级结构

闫化军 傅 彦 章 毅 李毅超
(电子科技大学计算机科学与工程学院 成都610054)

Neural Network Method for Protein Secondary Structure Prediction

YAN Hua-Jun FU Yan ZHANG Yi LI Yi—Chao
(College of Computer Science and Engineering, UEST, Chengdu 610054)

Abstract The explosive accumulation of protein sequences in the wake of large-scale sequencing projects is in shark contrast to the much slower experimental determination of protein structures. Neural Networks have been successfully applied into the prediction of protein structures, and the prediction accuracy continues to rise. This paper introduces the basic methods and technologies of the prediction of protein secondary structures using neural networks, especially expounds the two aspects: the improvement of neural network architecture and the adding of "evolutionary" information, which lead the ascent of prediction accuracy.

Keywords Protein secondary structure prediction, Secondary structure, Protein folding, Sequence analysis, Neural network, Bioinformatics

1 引言

蛋白质存在于所有的生物细胞中,是构成生物体最基本的结构物质和功能物质,它参与了几乎所有生命活动过程。蛋白质由氨基酸缩水聚合而成,自然界存在20种天然氨基酸,每种氨基酸可由一个大写英文字母表示。蛋白质的一级结构即构成该蛋白质的氨基酸代码串。蛋白质二级结构是指α螺旋和β折叠等规则的蛋白质局部结构元件,DSSP[1](Dictionary of Secondary Structure assignment of Proteins)把蛋白质二级结构分为八类: H(α-helix), G(3_{10}-helix), I(π-helix), E(extended strand), B(residue in b-bridge), T(turn), S(bend)和其他,在此基础上一般又把其划分为三类H、E和C。图1为蛋白质1azu一级序列和二级结构(从PDB[2]:www.rcsb.org获得)。

Sequence and secondary structure of Chain 1AZU:

```
1    AECSVDIQGN DQMQFNTNAI TVDKSCKQFT
     VNLSHPGNLP KNVMGHNWVL
     EEEEE SSS S SSE EE SSSSEEE EEEE SS TTTS B EE
51   STAADMQGVV TDGMASGLDK DYLKPDDSRV
     IAHTKLIGSG EKDSVTFDVS
     EETTTHHHHT HHHHHHTTTT TSS SS TT SEE B
     TT EEEEEESS
101  KLKEGEQYMF FCTFPGHSAL MKGTLTLK
     S SS EEE E STT TTT SEEEEEE
```

图1

神经网络方法预测蛋白质二级结构,就是指用神经网络的方法,以蛋白质一级序列信息为基础对蛋白质二级结构进行预测。

一般来说,蛋白质的一级序列决定了它的三维结构,也就决定了它的功能[3,4],这是进行蛋白质结构预测的理论依据。

蛋白质二级结构预测的重要意义:1.蛋白质二级结构可用来预测蛋白质的一些基本性质;2.蛋白质的功能由其三维结构决定,目前直接预测三维结构困难的情况下,二级结构预测是确定三维结构的关键步骤;3.人类基因组计划得到的大量基因序列还是"天书",基因编码区经过转录、翻译为蛋白质,基因功能依赖于所编码成的蛋白质,蛋白质结构的确定对解读生命奥妙意义重大;4.蛋白质序列信息爆炸性增长,而已知结构的蛋白质数量有限,现有的实验方法(x光衍射和核磁共振)经常需要几个月才能确定一个蛋白质结构,序列和结构之间日益扩大的鸿沟需要用智能计算方法来弥补。

1988年,Qian和Sejnowski[5]首次把人工神经网络用于蛋白质二级结构预测,即对非同源蛋白取得了Q_3=64.3%的平均预测精度,该精度高于先前其它方法所取得的结果(虽然训练测试集的不同导致一定程度上结果的不可比性),随即激起了一浪把人工神经网络用于蛋白质二级结构预测的热潮,预测精度不断攀升。

2 基本方法

神经网络预测蛋白质二级结构的研究中,单隐层全互连BP网络应用最为广泛。

2.1 单隐层BP网络结构

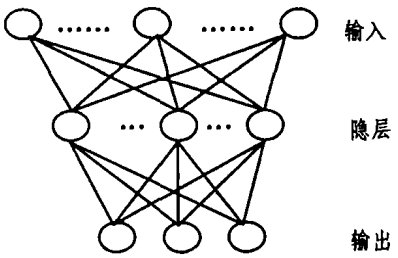


图2

2.2 数据表示

正交编码(orthogonal encoding)被普遍采用,20位二进制码,依次让一位为1、其余位为0即可表示20种天然氨基酸。用连续的2n+1个氨基酸作为输入,中间第n+1位氨基酸对应的二级结构代码(H,E,C)为网络的输出。这里的2n+1,即每次输入的连续氨基酸数目,称为窗口尺寸。由于序列两端会出现窗口"不满"的情形,故用20+1位表示一个氨基酸,"空窗口"用前20位为0、最后一位为1表示。输出层有三个节点,对应三种二级结构,可分别表示为:H-100,E-010,C-001。

2.3 预测精度衡量指标

1. Per-residue 精度

$Q_s = C_s/N_s$

其中，下标 s 取 H,E,C；C_s 表示正确预测 s 类二级结构的氨基酸数目；N_s 表示 s 类二级结构的氨基酸总的数目；Q_s 表示对 s 类二级结构预测的准确率。

$$Q_3 = \sum_{s=1}^{3} C_s/N$$

Q_3 是总体的预测准确率。Q_s、Q_3 简单易用，是最常使用的预测精度衡量系数。

2. Matthew 相关系数

$$C_s = (p_s n_s - u_s o_s)/[(n_s+u_s)(n_s+o_s)(p_s+u_s)(p_s+o_s)]^{0.5}$$

其中，下标 s 取 H,E,C；p_s 表示是 s 类结构且被预测为 s 类结构的氨基酸数目；n_s 表示不是 s 类结构且未被预测为 s 类结构的氨基酸数目；u_s 表示是 s 类结构但未被预测为 s 类结构的氨基酸数目；o_s 表示不是 s 类结构但被预测为 s 类结构的氨基酸数目。

该系数考虑了过预测的情形[6]。

3. 段折叠测量(Segment overlap measure，Sov)[7]

前两种衡量指标基于单个氨基酸预测的正确率，但 α 螺旋和 β 折叠由一段连续的氨基酸构成，单残基预测正确率高不一定能保证二级结构预测正确率高，所以在二级结构片段内预测的准确率就显得非常重要。

Sov 在所有的片段 i 上求和：

$$Sov = \sum_i [(minov(i)+\delta) * len(i)/maxov(i)]$$

其中：len(i)表示片段 i 的长度；minov(i)表示被预测序列与预测结果序列在 i 片段上的重叠部分；maxov(i)表示片段 i 在被预测序列与预测结果序列上的跨度；δ 是偏置量。

2.4 网络训练、测试的标准化

神经网络用于蛋白质结构预测研究的初期，很多方法被不断提出，但由于大家所采用的训练集、测试集、测试方法、二级结构分类方法等的不同，导致了预测精度相互之间在一定程度上的不可比性。随着 Rost & Sander 和 Cuff & Barton 分别提出了 RS126[8]和 CB513[9]训练、测试集，交叉测试等技术的引入，现在对不同预测方法已经能做出比较客观的比较。

1. 序列相似性(pair-wise similarity)要求　如果用于测试的某蛋白质一级序列与训练集中的某序列有较高的序列相似性，预测精度将偏高，过高反映了网络的概化能力。为了网络概化能力的可比性，训练和测试集中的一级序列要求较低的序列相似性，一般低于30%。RS126和 CB513中，任两序列间同源度低于25%。

2. 交叉测试(cross-validation)　方法：N 个训练和测试样本序列，把其大致等分为 m 个子集，依次用其中一个子集作为测试集、其余 m-1个子集作为训练集来测试网络的预测精度，取 m 次测试的平均结果作为对该网络概化能力的一个比较客观的评价。测试集的不同，可能导致预测精度变化很大，例如，Salzberg 和 Cost[10]指出：用最初选择的测试集时预测精度为71%的网络，引入多重交叉测试技术后，其预测精度(多个测试集上的平均值)降为65.1%，同一网络的两个不同测试集可导致预测精度相差6个百分点，可见，多重交叉测试是评估网络预测能力的一个很重要的技术细节。

3. 平衡训练(balanced training)　蛋白质按其分子外形对称程度可以分为球状蛋白质(globular protein)和纤维状蛋白质(fibrous protein)两大类。现在的蛋白质预测主要以球状蛋白质为研究对象，大多数蛋白质属于这一类。球状蛋白质中，三种二级结构分布是不均匀的：以 RS126为例，在总共的24395个残基中，H 占32%，E 占21%，C 占47%。通常，C 类预测得最好，H 类预测得较好，E 类预测得很差。预测结果的不平衡是由训练样本序列中三类二级结构分布不均衡所致。Hayward 和 Collins[11]用三类二级结构等比地训练网络，对最难预测的 E 类的预测精度取得较大改进。

4. 训练集和测试集　RS126和 CB513是目前两个标准的训练和测试集。RS126中包含126条蛋白质链，序列长度大于80个残基(蛋白质序列中的氨基酸也叫残基)，序列同源度低于25%；CB513中包含513条蛋白质链，其中16条序列长度≤30个残基，对同源性定义也更为严格。

5. 二级结构类的划分　不同的二级结构划分方法在一定程度上影响预测精度。DSSP 把八种二级结构进一步划分为三类时，有两种方法：① H,G 和 I 归入 H 类，E 类只包含 E，其余的都归入 C 类；②H 和 G 归入 H 类，E 和 B 归入 E 类，其余的归入 C 类。还有一些其他的类划分方法，Cuff & Barton[9]讨论了不同类划分方法对预测精度的影响。

神经网络方法预测蛋白质二级结构精度的不断提高，主要来自以下两个方面：网络结构的改进和“进化信息”的加入。

3 网络结构的改进

3.1 多模神经网络(MNN)

MNN 在网络结构方面做了改进，综合多个 SNN(单模神经网络)的输出，能显著提高预测精度。

1. 网络模型　如图3所示。

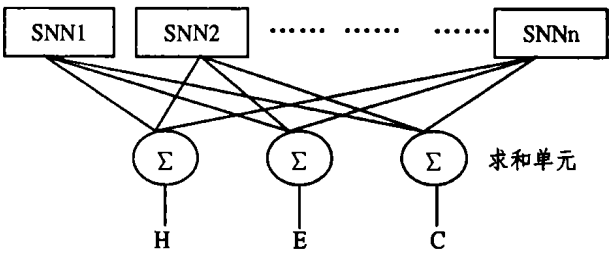


图3

2. 基本思想　n 个 SNN 并行预测，输出求加权和(如可简单地让每个 SNN 权为1/n)，三个求和单元对应三种二级结构，和值最大单元对应二级结构即为 MNN 输出。

3. 实验结果　引用文[12]的计算机实验结果，作者采用了5个 SNN，实验结果如下：

表1

	Q_3	Q_H	Q_E
SNN1	0.599022	0.569071	0.477099
SNN2	0.593149	0.568441	0.449727
SNN3	0.602121	0.588472	0.452932
SNN4	0.589885	0.582795	0.439071
SNN5	0.59217	0.583925	0.432668
MNN	0.661502	0.640499	0.489404

MNN 的预测精度比 SNN 高出约7个百分点。

文[12]中，作者 MNN 模型中的每一个 SNN 模型结构完全相同，如果采用其它构模方法构造 SNN，几种不同构模方法构造的、预测精度相近的 SNN 组成 MNN，应该对神经网络的概化能力能有所提高，从而提高模型的预测精度。

3.2 权值共享

权值共享(weight sharing)是神经网络中常用于加入先验知识的一项技术,同时,它能显著减少权值数目,从而避免了对样本的过拟合(overfitting)。

全互连网络预测蛋白质二级结构时,网络权值数目非常大,过拟合很难避免。文[13]中,采用权值共享对氨基酸进行自适应编码。正交编码中每个氨基酸用21位二进制数表示,最后一位表示"空窗口"情形,文[13]中省去了最后一位,"空窗口"用20位全为0表示;对于每个窗口位置,用相同的权值把20位输入与3个神经元相连,这样,每个氨基酸就用3个数字来表示了,从而导致了权值的骤减,避免了过分拟合。共享权值对于每个窗口的2n+1个氨基酸完全相同,对不同窗口又按 BP 算法对共享权值进行修改,文[13]认为:这样对氨基酸编码是自适应的,也应该是最优的。

3.3 按类建模

三类二级结构(H,E,C)有很大不同,可以依据每类结构的物理化学性质,为其设计一个 SNN,然后用 MNN 综合三个 SNN 的预测。

蛋白质数据库中的大多数 H 类结构是 α-helix,其结构特点是:一级序列中的每个氨基酸与序列中的前三个和后三个氨基酸氢键连接。文[13]把这一结构性质"内置"在网络结构中,如图4所示:窗口尺寸为13,每个小矩形代表一个氨基酸(文[13]中通过权值共享技术,用3个数字表示一个氨基酸,故每个小矩形中有三个数字),10个隐层节点,网络输出表示窗口中间位置残基是 H 类结构的概率。文[13]中,E 和 C 类网络中并未加入结构信息,还是采用全互连 BP,用 MNN 综合三个 SNN 后,取得了较好的预测结果。

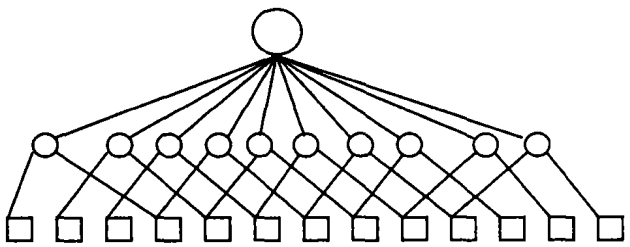


图4

3.4 二级网络

顾名思义,二级网络就是用两个神经网络预测蛋白质二级结构,第一个网络的输出为第二个网络的输入。第一个网络("sequence-to-structure" network)就是前面已经提的以一级序列为输入、三类二级结构为输出的网络,第二个网络("structure-to-structure"network)以第一个网络的预测结果为输入,以二级结构为输出。

为什么要引入二级网络呢?之前的预测方法存在一个共同的缺点,对于一个连续的二级结构片段,如 HHHHHHH,经常被预测为诸如 HHHEEHH 之类,也就是说,sequence-to-structure 网络没有"掌握"二级结构片段的平均长度,而二级结构本是由一段连续的氨基酸构成,故整个片段的预测准确率非常重要,这正是引入 structure-to-structure 网络的原因所在。

Rost 和 Sander[14]指出:加入 structure-to-structure 网络并没有明显提高总体预测精度 Q_3,但预测结果中的二级结构片段长度显然更接近实际情况。一个衡量预测结果二级结构片段长度的简单方法,就是对预测结果中的 α 螺旋和 β 折叠的平均长度和实际的加以比较。

3.5 支撑向量机(SVM)方法

清华大学的 Hua Sujun 和 Sun Zhirong 首次把 SVM 方法用于蛋白质二级结构预测[15],取得了当时基本是最高的预测精度:Sov=76.2%,Q_3=73.5%。

SVM 由 Vapnik 在1992年提出,是一种新的模式分类方法,已经被广泛运用到目标识别、语音鉴定、基因功能预测等多个方面。SVM 方法的基本思想是:构造一个高维超平面把两类样本分离,同时使两类样本相邻边界点(支撑向量)到该超平面距离最大化,故所得超平面是最优超平面。

SVM 方法有很多好的性质[15]:根源于统计学习理论,有很好的理论基础;在大多数情况下,SVM 方法的性能优于或相当于传统机器学习方法;SVM 方法可以有效避免过拟合;能处理大的特征空间,等等。

文[15]还是采用了传统的正交编码,SVM 的内积核采用径向基函数;网络结构上,先构建了6个二元分类器(H/~H,E/~E,C/~C,H/E,E/C,C/H),然后依据不同原则,综合6个训练好的二元分类器为几个三元分类器,最后,用一个会议机综合多个三元分类器的是输出为网络输出。

4 "进化信息"的加入

"进化信息(evolutionary information)"主要指蛋白质结构的一些物理化学性质和对已知蛋白质结构序列的一些统计信息。进化信息向神经网络提供了更多的蛋白质结构信息,有助于神经网络"掌握"蛋白质结构的规律性,从而提高预测精度。

前面已经提过,蛋白质一级序列决定其三维结构,这是我们进行蛋白质结构预测的前提。要让神经网络"掌握"一级序列是如何决定三维结构,充分的输入信息是必要的,但如果没有进化信息,网络得到的输入只是窗口尺寸范围的局部信息。进一步提高预测精度,向网络输入更宏观范围内残基相互关系信息非常必要。

甚至,文[16,17]认为,现在的蛋白质二级结构预测系统局限于:①需要加入更多的进化信息;②β 折叠中残基在大范围内相互作用,局部输入不能给出足够信息,预测精度低。

4.1 多序列比对信息和保守权

对于一级序列中的每个残基,通过序列比对可以计算出每种氨基酸在该位置的出现频率,用这样得到的20维向量来表示这个残基。同源度>30%的蛋白质有相同的三维折叠和近似的二级结构[18],所以多序列比对输入携带着更多的二级结构信息。Rost 和 Sander 指出:多序列比对输入可使 Q_3 增加6个百分点。

对于某个残基,如果其在该位置有较高的出现频率,则赋予较高的保守权。

多序列比对信息和保守权都可以通过 HSSP 文件获得。

Zhu Hanxi 等人[12]以多序列比对信息和保守权为网络输入,综合 MNN 技术,取得了66%左右的 Q_3。

4.2 插入删除信息(indel information)

前面的多序列比对是不允许对残基进行插入和删除的,如一级序列 LNNTEGDWW 和 LEEHGEWW,如下进行比对,有三个相同残基。

LNNTEGDWW　　LNNTEGDWW

LEEHGEWW　　LEEHGEWW

按允许插入和删除的方法进行比对,删除第一个序列中的 T,

则两序列有4个相同残基。

插入和删除更多地发生在规则二级结构的环状区域，所以，某个残基位置的插入删除比对信息携带着二级结构的信息：插入删除越多，越有可能是一个环状区域[19]。

某残基位置的插入删除信息：

$s_{ins}=N_{ins}/N_{all}$　$s_{del}=N_{del}/N_{all}$

其中，N_{ins}表示在该残基位置有插入的比对序列数目；N_{del}表示在该残基位置有删除的比对序列数目；N_{all}表示共有的比对序列数目。

Rost & Sander[19]在已经取得较高预测精度的方法中，加入插入删除信息，使Q_3又提高了0.5个百分点——主要来自对环状区域预测精度的提高。这也证实：插入删除信息的加入可以提高对环状区域的预测精度。

4.3 三级结构类信息

Kneller 等人[20]通过加入非局部信息——三级结构类信息(tertiary structural class information)，提高了网络预测精度。其中，蛋白质被分为：全α类(几乎所有二级结构都是α螺旋)，全β类(几乎所有二级结构都是β折叠)，和αβ混合类(没有哪种结构占绝对优势)。为每类分别构建网络，全α类只有H和C两个输出，同理，全β类只有E和C两个输出，αβ混合类三种输出均有。该方法由于对全α类和全β类只有两个输出，与三种输出均有(未加入三级结构类信息)的网络相比，提高了4个百分点的预测精度[21]。

4.4 残基疏水模式(hydrophobic pattern)信息

残基疏水性是蛋白质折叠的主要驱动力[22]。根据 Lim[23]的理论，疏水模式(I,I+2)和(I,I+2,I+4)倾向于在β折叠中发生；(I,I+3)，(I,I+3,I+4)，(I,I+1,I+4)更多发生在α螺旋中；(I,I+5)可帮助区别αβ混合类与其它。Brent & Peter[21]把残基疏水模式信息用窗口中符合模式的残基数与窗口大小的比值表示，综合了三级结构信息和多序列比对信息，获得了平均74%的预测精度。

4.5 密码子编码法

基因是DNA碱基(A,C,G,T)序列。基因中每三个连续的碱基叫做一个密码子(codon)。当DNA经过转录、翻译为蛋白质时，每个密码子翻译成一个氨基酸。64种密码子只能翻译为20种氨基酸，因此密码子携带着冗余的翻译信息。

如果只用翻译后的蛋白质一级序列预测二级结构，显然丢失密码子冗余信息。为了保留这部分信息，Owen[24]用密码子对氨基酸进行编码，如碱基表示：A1000000000，C0100000000，G 0010000000，T 0001000000，A或G 0000100000，A或T 0000010000，C或G 0000001000，C或T 0000000100，A或C或T 0000000010，A或C或G或T 0000000001。则氨基酸A可编码为0010000000 0100000000 0000000001，氨基酸C可编码为0001000000 0010000000 0000000100，等等。

Owen 用密码子编码法取得了与传统的正交编码法相近的预测精度，但在识别残基非局部相互作用信息上，密码子编码法有一定优势。

5 三级网络

三级网络由 Rost 和 Sander[14]在1993年提出，影响极为深远，现在已基本成为蛋白质二级结构预测普遍采用的方法，故我们在这里把它单独列出。三级网络是由多个二级网络组成会议机。

图5是 Rost 和 Sander 的实验结果，从中可以对几种蛋白质二级结构预测技术加以比较。

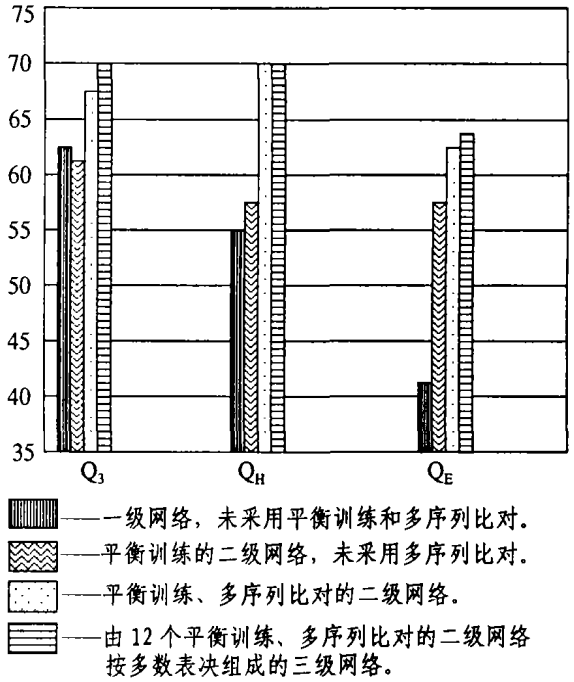


图5

从上图的实验结果中不难得出以下结论：

1. 采用平衡训练技术可以明显提高最难预测的E类的预测精度，使三类二级结构预测精度差别不是太大；

2. 多序列比对携带着更多的二级结构信息，可以明显提高预测精度。

3. 多数表决的三级网络可改善单个网络可能陷入局部最小的情况，在二级网络的基础上，预测精度再提高几个百分点。

小结 Rost 把蛋白质二级结构预测划分为三个阶段：第一阶段(1960—1970)，基于单个残基进行预测，精度较低；第二阶段(1970—90年代初)，基于3—51个残基的窗口大小进行预测，由于局部信息的有限性，对三种结构(H,E,C)的预测精度停滞在60%多一点；第三阶段(约1993—现在)，综合了大型数据库，加入了进化信息，预测精度突破了70%。State-of-the-art 方法获得的$Q_3=76\%$基本是目前最高的预测精度。

本文对神经网络用于蛋白质二级结构预测的基本方法和基本技术做了介绍，综合大型数据库的三级网络是现在最常用的网络拓扑结构(每级中最常用 BP 网络)；正交编码是最常使用的氨基酸编码方法；Q_3、C、和 Sov 是最常用的预测精度衡量指标；为了预测结果的更普遍性和相互之间的可比性，对二级结构分类方法、序列相似性、训练测试集和测试方法做了一定规范；网络结构的改进和进化信息的加入是提高预测精度的两个主要方面，近来这方面的许多工作引领着二级结构预测精度的一步步攀升，这方面的新成果也在不断涌现。

由于篇幅原因，还有一些好的工作和思想没有在本文中介绍，Ruggiero[25]提出了一种以α螺旋百分度组织训练集的方法；David[26]提出了一种基于从 PSI-BLAST 获取的 position specific scoring matrices 的预测方法；Rost 提出的预测置信度指标 RI(reliability index)；还有诸如 PHD、HMMSTR、PSI-BLAST、GOR、Sspro 等一些常用的预测方法。

氨基酸编码、网络结构和进化信息是用神经网络进行蛋白质结构预测的三个重要方面，但编码方法和网络结构从1988年到现在变化并不大。近来，Owen 提出的密码子编码方

法和 Sujun Hua 等人提出的支撑向量机方法都取得了很好的预测结果，可见，氨基酸编码方法和网络结构还有很大的探索空间；综合生物学方面的进展，加入更重要的进化信息，蛋白质二级结构的预测精度将会有新的突破。

参 考 文 献

1 Kabsch W, Sander C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen Bonded and Geometrical Features. Biopolymers, 1983, 22: 2577～2637
2 Bernstein F C, et al. The Protein data bank: a computer-based archival file for macromolecular structures. J. Mol. Biol., 1977, 112: 535～542
3 Anfisen C B, et al. Proc. Natl. Acad. Sci. U. S. A, 1961, 47: 1309～1314
4 Epstein C J, Goldberger R F, Anfinsen C B. Cold Spring Harb. Symb. Quant. Biol, 1963, 28: 439
5 Qian N, Sejnowski T J. Predicting the secondary structure of globular proteins using neural network models. J. Mol. Biol., 1988, 202: 865～884
6 Matthews B W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta, 1975, 405: 442～451
7 Rost B, Schneider R, Sander C. Redifining the goals of protein secondary structure prediction. J. Mol. Biol., 1993
8 Rost, B, Sander C. Prediction of Secondary structure at better than 70% accuracy. J. Mol. Biol, 1993, 232: 584～599
9 Cuff J A, Barton G J. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. Proteins: Struct. Funct. Genet, 1999, 34: 509～519
10 Salzberg S, Cost, S. J. Mol. Biol, 1992, 227: 371～374
11 Hayward S, Collins J F. Proteins, 1992, 14: 372～381
12 Zhu H, Yoshihara I, Yamamori K. Prediction of Protein Secondary structure by Multi-Modal Neural Networks. IEEE. Trans, 2002, 1: 280～285
13 Riis S K. Combining Neural Networks for Protein Secondary Structure Prediction. Neural Networks, Proceedings., 1995, 4: 1744～1748
14 Rost B, Sander C. Proc. Natl. Acad. Sci. USA, Biothysics, 1993, 90: 7558～7562
15 Hua S, Sun Z. A Novel Method of Protein Secondary Structure Prediction with High Segment Overlap Measure: Support Machine Approach. J. Mol. Biol., 2001, 308: 397～407
16 Rost B. Protein structure prediction in 1D, 2D, and 3D, in Encyclopedia of Computational Chemistry, 1998. 2242～2255
17 Sander B. R C. Third Generation Prediction of Secondary structure. in Protein structure prediction: methods and protocols, Humana Press, 2000. 71～95
18 Chothia C, Lesk A M. EMBO J, 1986, 5: 823～826
19 Rost B, Sander C, Schneider R. Evolution and Neural Networks: Protein Secondary Structure Prediction Above 71% Accuracy. In: Proc. of Twenty-Seventh Annual Hawaii Intl. Conf. of System Sciences, 1994, 5: 385～394
20 Kneller D G, Cohen F E, Langridge R. J. Mol. Biol, 1990, 214: 171～182
21 Metfessel B A, Saurugger P N. Pattern Recognition in the Prediction of Protein Structural Class. In: Proc. of the Twenty-sixth Hawaii Intl. conf. on, 1993, 1: 679～688
22 Dill K A. Biochemistry, 1990, 29: 7133～7155
23 Deleage G, Roux B. Protein Eng., 1987, 1: 289～294
24 Owen L, Liang H. H, Matthew B. 2001. Data Representation Influences Secondary Structure Prediction using Artificial Neural Networks. In: Seventh Australian and New Zealand Intelligent Information Systems Conf. 2001. 411～416
25 Ruggiero C. Peptides Secondary Structure Prediction with Neural Networks: A Criterion for Building Appropriate Learning Sets. Biomedical Engineering, IEEE Trans. On, 1993, 40(11): 1114～1121
26 Jones D T. Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices. J. Mol. Biol, 1999, 292: 195～202

（上接第31页）

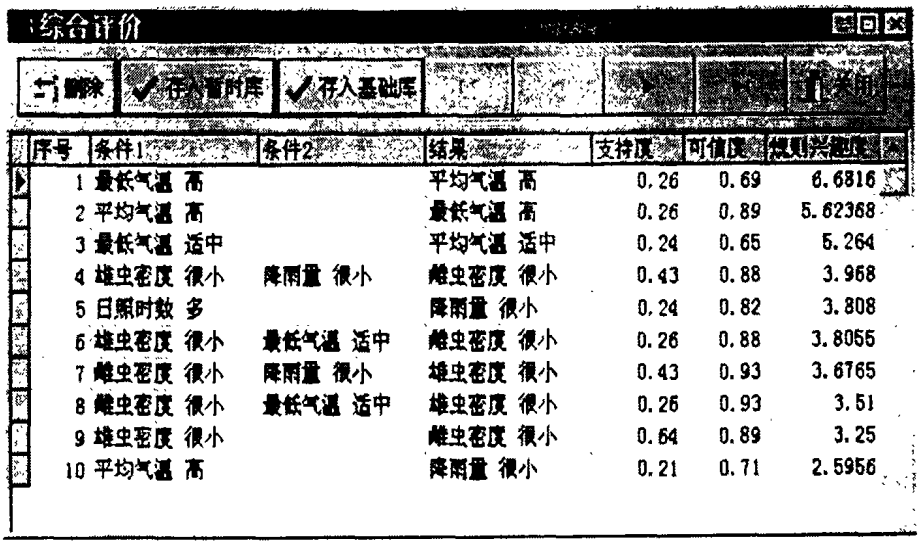
法置于有基础知识库参与的开放系统中（不局限于 KDD 自身的封闭系统中）；

（2）将主观的和客观的度量指标统一在一个三层次的评价体系统中，使它们不再是孤立的分析，而是将它们紧密地结合在一起；

（3）将评价过程自动化，所有的相关计算都由计算机编程实现，最终以一个统一的综合指标呈现给用户，便于用户比较选择；

（4）在主要是自动完成的前提下，也充分地发挥了用户的作用，并且用户所参与的部分（提供了各个语言变量和语言值的感兴趣度等）具有明确的意义，很容易实现。

（5）综合评价的指标是可扩展的，即可以根据需要添加新的指标，可简单地实现与原有的其它指标的综合。



序号	条件1	条件2	结果	支持度	可信度	规则兴趣度
1	最低气温 高		平均气温 高	0.26	0.69	6.6816
2	平均气温 高		最低气温 高	0.26	0.89	5.62368
3	最低气温 适中		平均气温 适中	0.24	0.65	5.264
4	雄虫密度 很小	降雨量 很小	雌虫密度 很小	0.43	0.88	3.968
5	日照时数 多		降雨量 很小	0.24	0.82	3.808
6	雄虫密度 很小	最低气温 适中	雌虫密度 很小	0.26	0.88	3.8055
7	雌虫密度 很小	降雨量 很小	雄虫密度 很小	0.43	0.93	3.6765
8	雌虫密度 很小	最低气温 适中	雄虫密度 很小	0.26	0.93	3.51
9	雄虫密度 很小		雌虫密度 很小	0.64	0.89	3.25
10	平均气温 高		降雨量 很小	0.21	0.71	2.5956

图4 综合评价结果

参 考 文 献

1 Padmanabhan B, Tuzhilin A. Unexpectedness as a measure of interestingness in knowledge discovery. Decision support systems, 1999, 27: 303～318
2 Freitas A A. On rule interestingness measures. Knowledge-Based Systems, 1999, 12: 309～315
3 Agrawal R, Imielinski T, Swami A. Database mining: a performance perspective. IEEE Trans. Knowledge and Data Eng., 1993, 5(6): 914～925
4 Quinllan J R. C4.5: program for machine learning. Morgan Kaufmann, 1992
5 Dhar V, Tuzhilin A. Abstract-driven pattern discovery in databases. IEEE Trans. Knowledge and Data Eng., 1993, 5(6): 926～938
6 Silberschatz A, Tuzhilin A. On subjective measures of interestingness in knowledge discovery. In: Proc of the 1st Int'l Conf on Knowledge Discovery and Data Mining. 275～286
7 Liu Bing, Hsu W, et al. Finding interesting patterns using user expectations. IEEE Trans. Knowledge and Data Eng., 1999, 11 (6): 817～832
8 Yang Bingru. A study on double bases cooperating mechanism in KDD(I). Engineering Science, 2002, 4(4): 41～45