基于扩展的 Markov 逻辑网的蛋白质 beta 结构关联预测方法研究*

熊忠阳，万文为，张玉芳，刘　君
（重庆大学 计算机学院，重庆 400044）

摘　要：传统机器学习方法在蛋白质关联图预测中要求满足独立一致性的条件，为了克服传统机器学习独立一致性假设，并且利用关联残基之间的规则约束，提出一种基于扩展的 Markov 逻辑网的蛋白质 beta 关联预测方法，该方法能够利用 Markov 统计关系学习框架来实现蛋白质 beta 结构关联预测。实验结果表明，利用该方法能够获得较好的 beta 关联预测效果，实验预测精度能够达到 45.91%，较 BetaPro 能提高 8%，是随机预测精度的 14 倍。

关键词：Markov 逻辑网；蛋白质关联预测；beta-残基；神经网络
中图分类号：TP18；TP319　　**文献标志码**：A　　**文章编号**：1001-3695(2012)04-1216-04
doi:10.3969/j.issn.1001-3695.2012.04.004

Research on method of protein contact prediction of beta structure by extend Markov logic network

XIONG Zhong-yang，WAN Wen-wei，ZHANG Yu-fang，LIU Jun
（*College of Computer Science*，*Chongqing University*，*Chongqing* 400044，*China*）

Abstract：In order to overcome the independent and identically distributed assumption and take advantage of rules and constrains in the protein，this paper presented a Markov logic network method to predict the beta-residue contact. This method could take advantage of Markov statistical relational learning framework to achieve the prediction of protein beta-residue contact. Experimental results show that this method can get better performance than other methods previously reported in the protein beta-residue contact prediction. It can be increased by 8% compared with BetaPro and 14 times of random prediction accuracy.

Key words：Markov logic network；protein contact prediction；beta-residue；neural network

0　引言

蛋白质关联预测是蛋白质三维结构预测的一个重要的中间步骤，该问题的解决对蛋白质空间结构预测和蛋白质功能预测都有重要意义。alpha-螺旋和 beta-折叠是两种主要的蛋白质二级结构。对于 alpha-螺旋结构，残基关联预测精度相对比较高；但是对于 beta 结构，关联的残基因为存在远程相互作用，从而令预测精度不高，成为蛋白质关联预测的瓶颈。

传统的机器学习方法如贝叶斯方法、HMM、SVM、神经网络方法等在关联预测中要求氨基酸残基满足相互独立和一致分布的假设，但是关联的残基之间往往不满足相互独立的假设。如在 alpha-螺旋中，位置上相距近的残基存在的关联概率大；而在 beta-折叠中，如平行和反平行 beta 结构，关联呈现一定的规律，残基之间明显不满足相互独立的假设。Markov 逻辑网是将 Markov 逻辑网和一阶谓词逻辑结合的新的统计关系学习方法[1,2]，它的优势在于能够利用谓词逻辑很好地处理规则约束的问题，将规则通过谓词公式进行表示，通过谓词逻辑进行学习推导，从而克服传统机器学习方法的不足。

1　蛋白质 beta 结构关联

1.1　beta 结构关联定义

蛋白质 beta 结构关联指的是蛋白质链上的任意两个 beta 残基，如果它们之间的空间距离小于某个给定的阈值，则认为这两个残基关联[3,4]。当然这个距离可能是两个残基的 α 碳原子之间的距离，也可能是 β 碳原子之间的距离，或者是两个残基空间结构中心的距离。

在本文的研究中，beta 结构关联的定义基于氨基酸残基的氢键，具体是按照 DSSP[4] 程序的结果给出：beta 残基 a 和 b 关联，当且仅当 b 被列为 a 的两个 β-桥（DSSP 程序输出中的 BP1 和 BP2 字段）当中的一个。本文的工作即预测蛋白质 beta 结构中任何两个 beta 残基，看它们之间是否存在关联。

1.2　研究方法与现状

最早对 beta 结构关联问题进行研究的是 Baldi[5]，他从氨

收稿日期：2011-06-23；**修回日期**：2011-09-10　　**基金项目**：中央高校研究生科技创新基金资助项目（CDJXS11181163）；重庆市自然科学基金资助项目（CSTC 2008BB2191）

作者简介：熊忠阳（1962-），男，重庆人，教授，博导，主要研究方向为网格与并行处理、数据挖掘、互联网；万文为（1986-），男，湖南岳阳人，硕士研究生，主要研究方向为数据挖掘、生物信息学（wanwenwei@cqu.edu.cn）；张玉芳（1965-），女，上海人，教授，主要研究方向为数据挖掘、网络入侵检测；刘君（1977-），女，讲师，博士研究生，主要研究方向为生物信息学、模式识别.

基酸序列着手,采用前馈神经网络作为二元分类器来判断任意两个氨基酸残基 i 和 j 是否存在关联。输入包括以目标 i 和 j 为中心残基的两个窗口。但是该方法存在其局限性:a)方法采用的数据集是高度不平衡的,真正关联的残基对只有 37 000,而不关联的残基对数目达到 44 000 000;b)残基之间的这种关联并不是相互独立的,如两个残基若它们在序列上相隔近就更有可能存在关联关系,忽视这种联系可能在预测时导致高偏差。Cheng 等人[6] 在 2005 年进行了另外一个研究,他们假设蛋白质氨基酸序列和氨基酸的二级结构是已知的,采用二维递归神经网络(2D-RNN)来进行预测,该神经网络结构采用二维网格来表示蛋白质序列,输入包括氨基酸序列位置 i 和 j 的残基特征,还包括周围一个小矩阵的氨基酸特征。它与之前的前馈神经网络的最大差别是残基对(i,j)的预测不但依赖于较近的那些位置的残基,而且还受矩阵的邻接单元的潜在影响。2D-RNN 被看做是命题和统计关系学习的一种折中方法,但是该方法也存在其缺陷,其中一个就是常见的梯度消失(vanishing gradients)问题,再一个就是该方法不能很好地处理残基关联中存在的明显的约束条件,如任何一个残基都不可能有超过两个以上的关联的残基、同一个 beta 残基片段上不可能存在相互关联的残基、残基关联存在对称性等。

2　Markov 逻辑网

2.1　Markov 逻辑网简介

Markov 逻辑网[7] 是公式附加权值的一阶逻辑知识库。在一阶逻辑中,通常一个世界只要违反了一个公式,该世界发生的概率就为 0。Markov 逻辑网的基本思想是软化这个约束:当一个世界违反了知识库中的一个公式,该世界发生的概率降低但不是不可能;而违反的公式越少,则发生的概率越大。

标准的 Markov 逻辑网由以下部分组成:a)一个基于知识的规则库 $F=\{F_1,F_2,\cdots,F_n\}$;b)一个有限常量集合 C;c)一个用来表达规则 F 强度的权值向量 $\boldsymbol{w}=\{w_1,w_2,\cdots,w_n\}$。Markov 逻辑网也可以看做是一个 Markov 网,这个 Markov 网的结构来自于规则 F 和常量 C,并以权值向量 $\boldsymbol{w}$ 进行参数化。特别地,Markov 网上所有的节点集合构成了所有可能的闭谓词。如果两个节点之间存在边相连,则表示对应的两个闭谓词必然存在于 F 中的某个闭谓词公式中。

Markov 逻辑网是一种条件概率模型,这个模型对于给定的证据谓词 X,求查询谓词 Y 的条件概率。条件概率可以通过式(1)求出。

$$P(Y=y|X=x;w)=(1/Z_x)\times e^{(\sum_{F_i\in F_y}w_i n_i(x,y))} \tag{1}$$

其中:w_i 是谓词公式 F_i 的权值,F_y 是所有的包含查询谓词的谓词公式,$n_i(x,y)$是谓词公式 F_i 满足世界(x,y)的所有的闭谓词公式个数。该公式可以看做给定证据谓词 x 以及对应的权值向量函数 w,查询谓词的条件似然。

2.2　Markov 逻辑网参数学习

对于 Markov 逻辑网来说,参数就是知识库中各公式的权值,因此参数学习任务就是估计出知识库中所有公式的权值。一般情况下,参数的学习通过最大似然方法求得,但是由于在单一数据库中计算公式的真闭公式个数是 NP 问题,所以可以利用最大伪似然估计方法和判别式训练方法来进行 Markov 逻辑网参数学习。

2.3　Markov 逻辑网推理算法

Markov 逻辑网可以回答任意“给定公式 F_1 的情况下公式 F_2 成立的概率是多少?”这样的查询问题。如果 F_1 和 F_2 是两个一阶逻辑公式,C 表示出现在 F_1 和 F_2 中个体常项的有限集合,L 表示 Markov 逻辑网,则有如下公式:

$$\begin{aligned}P(F_1|F_2,L,C)&=P(F_1|F_2,M_{L,C})=\\&P(F_1\wedge F_2|M_{L,C})/P(F_2|M_{L,C})=\\&\sum_{x\in\chi_{F_1}\cap\chi_{F_2}}P(X=x|M_{L,C})/\sum_{x\in\chi_{F_2}}P(X=x|M_{L,C})\end{aligned}\tag{2}$$

其中:$y=x$ 表示使公式 F_i 成立的世界集合,$P(x|M_{L,C})$由式(1)给出。

Markov 逻辑网的推理是 NP 完全问题[1],计算 u 可以近似地使用 Markov 链蒙特卡洛方法[8](Markov chain Monte Carlo, MCMC),不同的转移核导致不同的 MCMC 方法,如有 MC-SAT、吉布斯抽样、模拟退火等不同 MCMC 推理算法。

3　基于扩展的 Markov 逻辑网的 beta 结构关联预测

3.1　基于闭谓词公式的 Markov 逻辑网

对于 beta 结构关联预测问题,可以简单地将问题通过以下谓词公式表达出来:

$$\begin{aligned}&\text{window}(a,a_{-k},\cdots,a_{+k})\wedge\text{window}(b,b_{-k},\cdots,b_{+k})\Rightarrow\\&\qquad\text{partner}(a,b)\end{aligned}\tag{3}$$

其中:a_l 和 $b_l(l=-k,-k+1,\cdots,k)$表示氨基酸类型的变量(这里的氨基酸指的是常见的 20 种氨基酸);变量 a、b 表示氨基酸残基类型变量,本文取 a、b 分别为对应的两段长度为 $2k+1$ 的氨基酸残基片段的中心残基;window$(a,+a_{-k},\cdots,+a_{+k})$为真当且仅当残基 a 是 $2k+1$ 个氨基酸残基的中心残基;partner(a,b)为真当且仅当残基 a 和 b 存在关联。

如果简单地将以上谓词公式赋予一个权值的话,对应的 Markov 逻辑网将产生过于简单的模型,这个模型只能对数据集中出现频率最多的氨基酸对进行关联预测,模型退化为简单统计模型。在 Markov 逻辑网的许多应用中,谓词公式中的变量被常量取代,从而给不同的闭谓词公式以不同的权值,这个在 Alchemy 系统中可以通过在变量前添加“+”符号来实现。含有“+”符号的谓词公式的变量将会被该变量同类型的常量取代,从而产生不同权值的闭谓词公式。按照这个方法,谓词公式将转换为以下形式:

$$\begin{aligned}&\text{window}(a,+a_{-k},\cdots,+a_{+k})\wedge\text{window}(b,+b_{-k},\cdots,+b_{+k})\Rightarrow\\&\qquad\text{partner}(a,b)\end{aligned}\tag{4}$$

但是将问题简单地通过闭谓词公式处理并不能解决问题,因为对于本文的 beta 残基关联预测,在窗口大小为 $2k+1$ 的情况下,将扩展出 20^{4k+2} 个不同权值的闭谓词公式,闭谓词公式的指数增长,使得该方法不能应用到本文的问题当中。出现上面的闭谓词公式指数增长的主要原因是一个谓词公式当中出现了高元谓词,解决这类问题常用的方法是将高元谓词分解为多个二元谓词,每个二元谓词出现在一个谓词公式中。假设定义谓词 $AA(a,a_l)$,这个谓词为真,当且仅当氨基酸残基 a 是 a_l,则可以将式(4)分解得到以下谓词公式:

$$\begin{aligned}&AA(a,+a_k)\Rightarrow\text{partner}(a,b)\\&\qquad\vdots\\&AA(a,+b_k)\Rightarrow\text{partner}(a,b)\end{aligned}\tag{5}$$

经过分解以后,对应的 Markov 逻辑网只会产生 20 ×

$(2k+1)$个闭谓词公式。但是这种方法也有其缺点:a)会导致线性回归问题;b)在本文的 beta 残基关联预测问题中,一些特征的非线性组合能够极大地提高预测精度,但是分解后的线性模型显然不能满足这个要求;c)标准的 Markov 逻辑网不能处理连续型变量,而本文的问题中,如果想提高预测的精度,需要引入连续实数型的多序列比对结果。于是本文采用扩展的 Markov 逻辑网,通过与神经网络结合以克服上述问题。

3.2 扩展的 Markov 逻辑网

标准 Markov 逻辑网不能处理常量值为连续实数型的问题,而对于 beta 残基关联预测问题,引进蛋白质氨基酸的物理化学性质以及生物进化信息能够明显地提高预测精度。这些信息一般是通过实数向量形式给出的。为了解决这个问题,本文使用扩展的 Markov 逻辑网结构。扩展后的 Markov 逻辑网的条件概率为

$$P(Y=y|X=x;w)=(1/Z_x)\times e^{\sum_{F_i\in F_y}\sum_j w_i(g_{ij},\theta_i)n_{ij}(x,y)} \tag{6}$$

其中:g_{ij}表示谓词公式 F_i 对应的某个闭谓词公式的常量集合;w_i 是一个实值型函数,返回某个闭谓词公式下对应的权值;θ_i 是参数向量。当 θ_i 是标量时,$w_i=\theta_i$,扩展的 Markov 逻辑网退化为式(1)表示的标准 Markov 逻辑网,所有标准 Markov 逻辑网是扩展 Markov 逻辑网的特殊形式。θ_i 可以通过核心逻辑回归 KLR 和神经网络方法获得。

扩展 Markov 逻辑网在标准 Markov 逻辑网的基础上,增加了“ * ”和“ $ ”符号,“ * ”用来进行谓词声明,“ $ ”表示对应的闭谓词公式的权值需要通过神经网络学习来获得[9],这里的神经网络已经嵌入到 Alchemy 系统中。神经网络神经元的输入可以是蛋白质对应常量的某种形式的编码,以及相关物理化学性质的特征向量等。扩展后的 Markov 逻辑网谓词公式为

$$\text{window}(a,\$pa)\wedge\text{window}(b,\$pb)\Rightarrow\text{partner}(a,b) \tag{7}$$

3.3 Markov 逻辑网知识库的构建

蛋白质关联关系是相互影响的,蛋白质关联残基之间存在很强的规则约束,传统机器学习方法在蛋白质关联预测中要求满足独立一致性的条件,不能有效利用问题中存在的规则约束条件。Markov 逻辑网相对于传统机器方法的优势在于能够有效地处理规则约束问题,将规则约束通过谓词公式表示,通过相关的学习推理获得问题的解决方法。

针对 beta 残基关联预测问题,可以构建相应的 Markov 逻辑网知识库。本文主要定义 Residue 和 Strand 两种对象类型,Residue 表示氨基酸变量,Strand 表示蛋白质链中的 beta 残基片段变量,DSSP 中定义 Strand 为蛋白质氨基酸序列中连续的 E 和 B 子串。本文通过对 beta 残基关联相关知识进行研究,总结出以下知识库关联规则:

a)beta 关联存在对称性和反自反性,所以本文得出以下两个知识库规则:

$$\text{partner}(a,b)\Rightarrow\text{partner}(b,a) \tag{8}$$

$$\text{partner}(a,a) \tag{9}$$

b)同一个 beta 残基片段上不可能存在两个关联对,同时,同一个残基片段上也不可能存在两个残基与另一个残基片段上的同一个残基都关联,可以得出以下两条规则:

$$\text{belongToStrand}(b_1,g)\wedge\text{belongToStrand}(b_2,g)\Rightarrow !\,\text{partner}(b_1,b_2) \tag{10}$$

$$\text{partner}(b_1,b_2)\wedge\text{belongToStrand}(b_2,g)\wedge\text{belongToStrand}(b_3,g)\wedge b_2 != b_3\Rightarrow !\,\text{partner}(b_3,b_1) \tag{11}$$

c)因为 beta 残基不允许出现交叉关联的情况,所以本文可以得出以下一条重要的规则:

$$\text{partner}(b_1,q_1)\wedge\text{partner}(b_2,q_2)\wedge\text{succ}(b_1,b_2)\wedge\text{succ}(q_1,q_2)\wedge\text{succ}(b_2,b_3)\wedge\text{succ}(q_3,q_1)\Rightarrow !\,\text{partners}(b_3,q_3) \tag{12}$$

d)两个 beta 残基片段,如果相隔一个残基的两个残基对都关联,则中间的残基对也一定关联,可以得出以下规则:

$$\text{partner}(b_1,q_1)\wedge\text{succ}(b_1,b_2)\wedge\text{succ}(b_2,b_3)\wedge\text{succ}(q_1,q_2)\wedge\text{succ}(q_2,q_3)\wedge\text{partners}(b_3,q_3)\Rightarrow\text{partner}(b_2,q_2) \tag{13}$$

e)beta 发夹(β-hairpins)是一种特殊的 beta 结构,它的特殊性在于相邻的两个残基片段以反平行的方式一一关联。研究表明,当两个残基片段之间的距离小于 6 时,70% 以上的 beta 发夹结构符合上面的规律,如果相邻的两个残基片段的长度都大于 4 的话,符合的比例达到 77%;进一步,如果在这两段残基片段中存在甘氨酸的话,这种比例将达到 83%。所以可以得出以下规则:

$$\text{belongToStrand}(b_1,g_1)\wedge\text{belongToStrand}(b_2,g_2)\wedge\text{partner}(b_1,b_2)\wedge\text{glycineWithin}(g_1,g_2)\wedge\text{interStrandDistance}(g_1,g_2,d)\wedge\text{lessThanSix}(d)\wedge\text{succ}(b_1,q_1)\wedge\text{succ}(q_2,b_2)\Rightarrow\text{partner}(q_1,q_2) \tag{14}$$

f)β-α-β 也是一种特殊的 beta 结构,其特殊性在于相邻的两个残基片段会以平行的方式进行关联。研究表明,当两个残基片段之间的长度大于 4 且中间被 alpha 螺旋隔开时,对应的残基片段和中间 alpha 螺旋一起构成 β-α-β 结构。平行的残基片段之间存在着明显的关联关系,可以得到下面一个明显的关系:

$$\text{helixWithin}(s_1,s_2)\wedge\text{strandLength}(s_1,n)\wedge\text{strandLength}(s_2,n)\wedge\text{moreThanFour}(n)\wedge\text{firstOfStrand}(f_1,s_1)\wedge\text{firstOfStrand}(f_2,s_2)\wedge\Rightarrow\text{partner}(f_1,f_2) \tag{15}$$

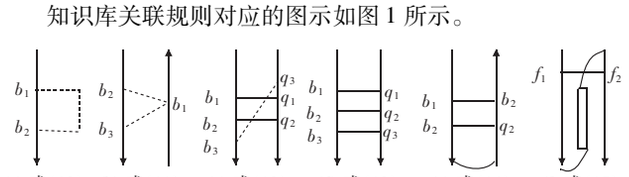
g)每一个 beta 残基只可能存在 0、1、2 个与它关联的残基。本文得出以下规则:

$$\text{partner}(a,b)\wedge\text{partner}(a,c)\wedge b != d\wedge c != d\Rightarrow !\,\text{partner}(a,d) \tag{16}$$

h)整个预测模型中最核心的部分是本文前面介绍的通过扩展 Markov 逻辑网模型来实现 beta 关联预测的规则。

$$\text{window}(b_1,\$p_1)\wedge\text{window}(b_2,\$p_2)\wedge\text{belongToSequence}(b_1,s)\wedge\text{sequenceFeatures}(s,\$f)\Rightarrow\text{partner}(b_1,b_2) \tag{17}$$

知识库关联规则对应的图示如图 1 所示。



(a) 式(10) (b) 式(11) (c) 式(12) (d) 式(13) (e) 式(14) (f) 式(15)

图 1 部分关联规则图示

4 实验及分析

4.1 实验数据集

本文实验中采用的蛋白质数据集是 Cheng 等人在 2005 年做 beta 预测所用的数据集,该数据集是从 Protein Data Bank 中抽取出来的。蛋白质通过 X 射线解析,只有解析度大于 2.5Å 的才保留,蛋白质序列如果包含有未知的以及非标准的氨基酸,或者序列长度小于 50 都被过滤掉。氨基酸的二级结构和疏水信息是通过 DSSP 程序获得的,DSSP 中二级结构为 E 和 B 的被认为是 beta 残基。根据 DSSP,每一个 beta 残基可能存在 0、1、2 个与它关联的残基。通过过滤程序可以选择至少包含

10 ~ 100 个 beta 残基的蛋白质链,这些链中 90% 以上的至少包含一对关联的残基。通过 UniqueProt 工具,本文去掉了相似度高的冗余数据。为了获得蛋白质的生物进化信息,利用 PSI-BLAST 工具在非冗余 NCBI 蛋白质数据库上进行多序列比对从而获得蛋白质的序列谱。最后将获得的数据集随机地切分成 10 部分,从而进行 10 倍交叉验证实验。该数据集相关信息如表 1 所示。

4.2　实验特征提取

为了能够提高实验的预测精度,本文实验所采用的输入特征包括以下几个部分:

a)氨基酸残基特征。一个氨基酸的结构与它所处的上下文环境有关,本文采用常用的滑动窗口机制,窗口大小为 9。氨基酸残基特征包括:(a)多序列比对信息,它包含了丰富的生物进化信息,通过多序列比对,可以增加生物同源信息,而同源的蛋白质具有相同的结构特征,所以多序列比对信息可以用来提高预测的精度;(b)氨基酸残基二级机构,本文的实验中将二级结构分为 alpha 螺旋、beta 折叠和 coil 卷曲三类;(c)氨基酸残基疏水性,疏水性是一种侧链非极性的度量,是蛋白质折叠的必要驱动力,常作为蛋白质结构预测的特有信息。因为残基的疏水性影响环境间的非共价结合,因此对于一个残基与其他残基关联的确定,疏水性是重要的因素。

b)氨基酸残基对特征是指残基 i 和 j 表示的残基对信息。残基对特征包括:(a)i 和 j 之间的序列间距,为了便于处理,将间距除以 200;(b)i 和 j 之间的 beta 残基片段数,并将这个值除以 10 作为输入;(c)残基对类型特征[10],根据物理化学性质将氨基酸分为极性、非极性、酸性和其他,这四种性质无序组合出 10 种类别;(d)以 $(i+j)/2$ 为中心残基、直径为 5 的窗口片段信息,包括每个氨基酸的编码、二级结构和疏水信息;(e)20 种氨基酸构成比例;(f)i 到 j 之间二级结构构成比例;(g)i 到 j 之间疏水氨基酸构成比例。

c)氨基酸序列编码。其中包括:(a)序列中 20 种氨基酸构成比例;(b)序列中二级结构构成比例编码;(c)序列中疏水氨基酸构成比例;(d)氨基酸序列中 beta 残基个数编码,按照个数将氨基酸序列分为四类($L \leqslant 50$, $50 < L \leqslant 100$, $100 < L \leqslant 150$, $L > 150$)。

4.3　实验方法

本实验平台采用华盛顿大学 Domingos 和 Kok 等人开发的基于 Markov 逻辑网的 Alchemy 软件包[10,11],该软件包提供了一系列统计关系学习和概率逻辑推理方面的算法。实验具体过程如下:

a)蛋白质 beta 关联预测的知识库的构建。即构建. mln 文件,该文件由谓词声明和谓词公式构成,具体参考本文 Markov 逻辑网知识库的构建介绍。

b)进行实验数据预处理。主要是从数据集中将蛋白质序列解析出来,并通过 PSI-BLAST 进行多序列比对,获得氨基酸序列的序列谱(profile)信息。

c)生成 Alchemy 训练数据和测试数据知识库 DB 文件。在生成这两种数据文件时,还需要将对应的 Markov 逻辑网特征向量文件保存起来。

d)建立扩展 Markov 逻辑网模型内嵌神经网络权值学习参数文件。该文件包括四行,第一行为选择的神经网络类型,第二行为隐单元个数,第三行为学习率,第四行为特征向量文件存放的具体路径。

e)权值学习。本文采用的权值学习算法为判别式训练方法,对训练知识库进行权值学习。

f)谓词推理。根据上一步学习到的权值和测试知识库,分别采用 MC-SAT、吉布斯抽样和模拟退火三种推理算法进行推理。

4.4　实验结果分析

本文实验采用 10 倍交叉验证方法,实验数据采用的是与 Cheng 等人在 2005 年作 beta 预测相同的训练测试数据集。本文采用与蛋白质 beta 关联预测中有良好预测精度的 BetaPro 进行实验比对,实验的结果通过精确度 $P = TP/(TP + FP)$、召回率 $R = TP/(TP + FN)$、$F1 = 2PR/(P + R)$ 值表示,实验结果如表 2 所示。

表 1　蛋白质数据集

数据集信息	数量
蛋白质链数	916
氨基酸残基数	187 516
beta 残基数	48 996
beta 残基关联对	31 638
beta 残基片段	8 172

表 2　BetaPro 和 Markov 逻辑网实验对比

参数	BetaPro	Markov 逻辑网		
		MC-SAT	吉布斯抽样	模拟退火
P	38.03	43.49	45.85	45.91
R	44.18	40.37	38.96	38.83
$F1$	40.87	41.87	42.13	42.07

通过实验结果可以知道,对于蛋白质 beta 关联预测问题,通过 Markov 逻辑网能够取得较 BetaPro 更好的预测精度,预测精度提高比较显著,较 BetaPro 方法最大能提高约 8% 。

分析这两种不同的蛋白质 beta 关联预测方法之所以会产生不同的预测效果,主要是因为 Markov 逻辑网通过引进规则约束移除了错误的预测实例。如式(16)所示,与每个氨基酸残基关联的残基数不可能大于等于 3,在 BetaPro 中测出平均每个蛋白质上存在 7 个违反这个规则的氨基酸残基,而 Markov逻辑网平均只有 0.1 个;BetaPro 测出 61% 的蛋白质存在这样的违反规则的氨基酸残基,而本文中提出的方法只有 2.8% 。由表 2 可以看出,P 值在提高时,R 值相应在降低,这表明本文的方法取得了更好的预测精度值,但也牺牲了一定的召回率。

通过对比 Markov 逻辑网的三种推理算法,可以得出模拟退火和吉布斯抽样推理算法较 MC-SAT 有更高的精确度值,这主要是因为 MC-SAT 为了加速推理的过程[9],损失了一定的精度。

5　结束语

本文通过对蛋白质 beta 结构关联预测的研究,提出了一种基于一阶逻辑和概率统计学习模型的扩展的 Markov 逻辑网方法,该方法达到了对 beta 结构关联进行预测的目的。实验结果表明,基于扩展的 Markov 逻辑网的蛋白质 beta 结构关联预测能够取得较好的实验效果。实际上,本文提出的方法具有通用性,可以将该方法应用到其他的研究领域,如数据挖掘、信息检索、自然语言处理、社会网络分析以及生物信息学等。本文的进一步研究工作是应用扩展的 Markov 逻辑网对整个蛋白质关联进行预测,以期获得好的预测效果。

参考文献:

[1] RICHARDSON M, DOMINGOS P. Markov logic networks[J]. Machine Learning, 2006, 62(1): 107-136.

[2] GETOOR L, TASKAR B. Introduction to statistical relational learning [M]. Cambridge: MIT Press, 2007: 339-371.　(下转第 1255 页)

子一个初始速度,即一个交换序;

b)利用式(1)来评价粒子,对不符合约束条件的粒子采取惩罚操作,求出初始的 *p*best 和 *g*best;

c)对每个微粒做初始速度操作,即经过交换序之后得到新的微粒;

d)对新的微粒和个体最优解 *p*best 按概率 c_1 做类 OX 交叉操作,对不做交叉操作的微粒进行变异,得到新的微粒再参照全局最优解 *g*best 按概率 c_2 进行类 OX 交叉,同样对不做这个操作的微粒进行变异;

e)按式(1) 评价每个粒子的适应值,对不符合约束条件的粒子的适应值加惩罚值,并与 *p*best 和 *g*best 进行比较,如果较好,则更新 *p*best 和 *g*best;

f)若满足停止条件,搜索停止输出结果,否则返回 c)继续搜索。

4　仿真实例

用 MATLAB 7.0 编程,在 CPU 为 E5800 3.20 GHz,内存为 2.96 GB,操作系统为 Windows XP 的微机上验证了算法的性能,求解实例采用了 Solomon 提供的 25 个客户点规模的标准测试问题,具体客户与车场坐标、客户需求、时间窗等相关数据可以参看文献[9]所给出的网站。算法参数设置如下:粒子数目取 40,迭代次数取 400,w 取值 0.75,c_1 和 c_2 均取 0.8。对测试实例 C101(25 客户)进行 20 次运算,其中有 13 次达到了本算法的最优解:使用 3 辆车,总行 197.4,对应的车辆路线图如图 2 所示。

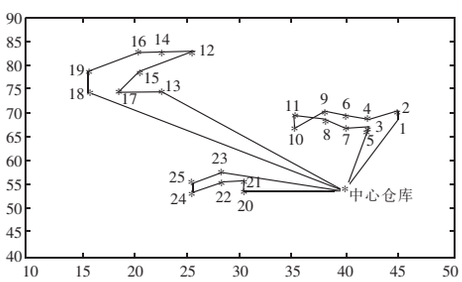


图 2　测试实例 C101(25)车辆路径

同时采用了遗传算法和粒子群算法对 25 个客户点随机几个算例进行了计算,设定两种算法的迭代次数均为 400 代。遗传算法的种群规模为 40,交叉概率为 0.78,变异概率为 0.2,采用轮盘赌为选择策略;粒子群算法对更新得到的粒子速度和位置进行取整和按边界取值操作,粒子数目为 40,惯性权重取 0.75,学习因子 $c_1 = c_2 = 2$,分别进行 20 次运算,结果如表 1 所示。

表 1　各算法最优结果比较

实例名称	已知最优解	遗传算法	粒子群算法	本文算法
C101	191.3/3	210.7/3	248.0/3	197.4/3
C201	214.7/2	242.4/2	257.2/2	216.4/1
C104	186.9/3	202.6/3	228.5/3	188.2/3
R104	416.9/4	428.6/4	446.0/4	415.8/4
R208	328.2/1	335.4/1	344.6/1	330.2/1
R201	463.3/4	470.4/4	482.7/4	469.3/4
RC107	298.3/3	320.7/3	337.1/3	302.8/3
RC204	299.7/3	318.6/3	323.5/3	305.2/3

从表 1 情况来看,本文算法得到的解的质量要高于其他算法,并且很接近已知最优解。

5　结束语

本文提出了基于粒子碰撞的离散粒子群算法,对不做碰撞的粒子进行变异操作,避免陷入局部解,并成功地求解了 Solomon's VRPTW 问题集中的 25 客户规模的问题,结果证明了该算法应用于带时间窗车辆调度问题的有效性。

参考文献:

[1] 李宁,邹彤,孙德宝. 带时间窗车辆路径问题的粒子群算法[J]. 系统工程理论与实践,2004,24(4):130-135.

[2] 刘志雄. 基于粒子群算法的物流配送车辆优化调度研究[J]. 武汉科技大学学报,2009,32(6):615-618.

[3] 马炫,彭芃,刘庆. 求解带时间窗车辆路径问题的改进粒子群算法[J]. 计算机工程与应用,2009,45(27):200-202.

[4] CHERC M. Discrete particle swarm optimization illustrated by traveling salesman problem[M]. Berlin: Springer-Verlag,2004.

[5] 黄岚,王康平,周春光,等. 粒子群优化算法求解旅行商问题[J]. 吉林大学学报:理学版,2003,41(4):477-480.

[6] 肖健梅,黄有方,李军军,等. 基于离散微粒群优化的物流配送车辆路径问题[J]. 系统工程,2005,23(4):97-100.

[7] 郎茂祥. 配送车辆优化调度模型与算法[M]. 北京:电子工业出版社,2009.

[8] 余伶俐,蔡自兴,高平安,等. 当代学习自适应混合离散粒子群算法研究[J]. 小型微型计算机系统,2009,30(9):1800-1804.

[9] 张有华,张翠军,崔忠强. 有时间窗车辆路径问题的混合智能算法[J]. 计算机工程与应用,2010,26(4):221-223.

(上接第 1219 页)

[3] FAURE G,BORNOT A,De BREVERN A G. Protein contacts, inter-residue interactions and side-chain modeling[J]. Biochimie,2008,90(8):626-639.

[4] 刘桂霞,于哲舟,周春光. 基于带偏差递归神经网络蛋白质关联图的预测[J]. 吉林大学学报:理学版,2008,46(2):265-271.

[5] BALDI P,POLLASTRI G,ANDERSEN C A,*et al*. Matching protein betasheet partners by feedforward and recurrent neural networks [C]//Proc of the 8th International Conference on Intelligent for Molecular Biology. 2000:25-36.

[6] CHENG Jian-lin, BALDI P. Three-stage prediction of protein beta-sheets by neural networks, alignments and graph algorithms [J]. Bioinformatics,2005,21(1):75-84.

[7] RICHARDSON M, DOMINGOS P. Markov logic networks [D]. Washington DC:University of Washington,2004.

[8] ANDRIEU C,De FREITAS N,DOUCET A,*et al*. An introduction to MCMC for machine learning[J]. Machine Learning,2003,50(1):5-43.

[9] 张玉芳,黄涛,艾东梅,等. Markov 逻辑网及其在文本分类中的应用[J]. 计算机应用,2009,29(10):2729-2736.

[10] DOMINGOS P,KOK S,LOWD D,*et al*. Markov logic[M]//Probabilistic Inductive Logic Programming. Berlin:Springer,2008:92-117.

[11] KOK S,SUMNER M,RICHARDSON M,*et al*. The Alchemy system for statistical relational AI[R]. Washington DC:University of Washington,2007.

[12] KABSCH W,SANDER C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features[J]. Biopolymers,1983,22(12):2577-2637.

[13] CHENG Jian-lin,BALDI P. Improved residue contact prediction using support vector machines and a large feature set[J]. BMC Bioinformatics,2007,8(1):113-122.

[14] FARISELLI P,OLMEA O,VALENCIA A,*et al*. Prediction of contact maps with neural networks and correlated mutations[J]. Protein Engineering,2001,14(11):835-843.