基于条件随机场进行蛋白质二级结构预测 *

罗 亮, 邵泽辉
(华中科技大学 控制科学与工程系, 武汉 430074)

摘 要: 使用一种新的概率图模型——条件随机场对蛋白质二级结构进行预测,并给出了模型的构建、训练以及解码的算法。应用这一模型对一个典型的蛋白质数据集 CB513 的二级结构进行了预测,并将预测结果与其他方法进行比较,预测准确度有明显的提高。

关键词: 二级结构预测; 条件随机场; 概率图模型
中图分类号: O224;TP301 **文献标志码**: A **文章编号**: 1001-3695(2009)03-0832-02

Protein secondary structure predict based on conditional random fields

LUO Liang, SHAO Ze-hui
(*Dept. of Control Science & Engineering, Huazhong University of Science & Technology, Wuhan* 430074, *China*)

Abstract: Conditional random fields(CRFs) is a new probabilistic graph model and this work introduce conditional random fields in protein secondary structure prediction. This paper gave the method of constructing the model and the algorithms to train and decode the model and used the model to predict the second structure of a famous protein dataset(CB513). Finally compared the results with some other methods.

Key words: prediction of protein secondary structure; conditional random fields(CRFs); probabilistic graph model

蛋白质是生命科学研究的重要对象,研究蛋白质的功能需要了解它们的结构,特别是空间结构,因为结构决定功能。X 射线晶体衍射分析法和多维核磁共振技术是测定蛋白质空间结构的两种主要实验方法,然而实验方法不仅耗资耗时,还受实验条件的限制,而且用实验方法测定结构的速度和人类的测序速度之间还存在很大的差距。而蛋白质的空间折叠结构取决于构成该蛋白质的氨基酸序列[1], 即蛋白质的空间折叠结构的全部信息都隐藏在氨基酸序列中。因此蛋白质结构预测问题是当前国际上最具有理论研究和实际应用价值的问题之一。目前,解决蛋白质二级结构预测方法非常多,也取得了比较好的预测结果,如神经网络[2,3]、支持向量机[4~8]、隐马尔可夫模型(HMM)[9~12]、半马尔可夫模型(SSMM)[13]、最大熵模型(MEM)[14]等。

隐马尔可夫模型一个最大的缺点就是由于其输出独立性假设,导致其不能考虑上下文的特征,限制了特征的选择。而最大熵模型则解决了这一问题,可以任意地选择特征,但由于其在每一节点都要进行归一化,只能找到局部的最优值,同时也带来了标记偏见的问题(label bias),即凡是训练集中未出现的情况全都忽略掉。条件随机场(CRFs)则能很好地解决了这一问题。条件随机场是一种新的概率图模型[15],它并不在每一个节点进行归一化,而是所有特征进行全局归一化,具有表达元素长距离依赖性和交叠性特征的能力, 能方便地在模型中包含领域知识,因此可以求得全局的最优值。Liu 等人[14]应用条件随机场对蛋白质折叠进行预测,取得了比较好的预测结果。本文将条件随机场用于三状态二级结构预测,也提高了预测的精度。

1 材料和方法

1.1 线性条件随机场

条件随机场是一种以给定的输入节点值为条件来预测输出节点值概率的无向图模型。条件随机场是一个典型的判别式模型,其联合概率可以写成若干势函数联乘的形式,其中每个最大团都是两个相邻变量以一条边连接构成的子图。图 1 是一种最简单且最重要的 CRFs,称为线链 CRFs。线链 CRFs 假设在各个输出节点之间存在一阶马尔可夫独立性,其输出节点被无向边连接成一条线性链。

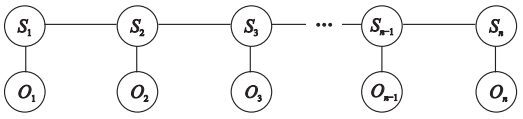

图1 线性CRFs的图形结构

在图 1 中,$S_1, S_2, \cdots, S_n$ 为 CRF 的 N 个状态,$O_1, O_2, \cdots, O_n$ 为 CRFs 的 N 个观测字符。观测字符从字符集 $\{V_1, \cdots, V_M\}$ 中取出,对于蛋白质结构预测问题,$M=20$,即字符集为包含了 20 个氨基酸残基符号的字符集。

CRFs 模型由 N 个状态组成,这样,在一个输入序列给定的情况下,对于参数为 $\lambda = \{1, 2, \cdots, K\}$ 的线性链 CRFs, 其状态序列的条件概率由式(1)得到:

$$P(S|Q) = \exp\left[\sum_{t=1}^{T}\sum_{k=1}^{K}\lambda_k f_k(s_{t-1}, s_t, o, t)\right]/Z_0 \quad (1)$$

其中 Z_0 为归一化因子,由式(2)得到:

$$Z_0 = \sum_{s}\exp\left[\sum_{t=1}^{T}\sum_{k=1}^{K}\lambda_k f_k(s_{t-1}, s_t, o, t)\right] \quad (2)$$

收稿日期: 2008-03-20; **修回日期**: 2008-05-26 **基金项目**: 国家自然科学基金资助项目(60373089, 60674106, 60533010, 60703047)

作者简介: 罗亮(1980-),男,湖北武汉人,博士研究生,主要研究方向为生物信息学、生物计算(luoliang@mail.hust.edu.cn);邵泽辉(1978-),男,博士研究生,主要研究方向为智能计算、图论等.

$\lambda_i f_i(S,O)$在线性链条件随机场模型中包括定义在点上的特征和定义在边上的特征,一般都取 0、1 特征。

1.2　基于序列模体的条件随机场模型

使用 CRFs 模型前,先将待预测以及用于训练的蛋白质序列转换为表示进化信息的序列模体。本文使用序列对比程序 PSI-BLAST 产生序列模体,程序的下载地址为 ftp://ftp.ncbi.nih.gov/blast。给定一个长度为 L 的蛋白质序列,使用 PSI-BLAST 程序与 NCBI 数据库中的蛋白质序列进行对比,可以得到一个 $20\times L$ 的位置得分矩阵 $\mathcal{S}$(PSSM):

$$\mathcal{S}=\begin{bmatrix}(S_1(1),S_1(2),\cdots,S_1(M))\\(S_2(1),S_2(2),\cdots,S_2(M))\\ \text{ú}\qquad\text{ú}\qquad\text{ú}\\(S_L(1),S_L(2),\cdots,S_L(M))\end{bmatrix}\tag{3}$$

每个向量 $\mathcal{s}_t$的分量为正数,并且其和为常数 S,其值独立于向量的位置 t。

按照 DSSP[16] 的方法,蛋白质的二级结构分为八种(H,G,I,E,B,T,S,_),本文按照 Rost 等人[17] 的分类方法把八种二阶结构简化为三种(H,E,C)。其中,H、G 归入 H;E、B 归入 E;剩下的归入 C。按照这种方法,长度为 L 的蛋白质序列的二级结构可以用列向量 $\mathcal{O}=(o_1,o_2,\cdots,o_L)^{\mathrm{T}}$表示。

1.3　RCFs 的参数训练与解码

与 HMM、MEMM 类似,在进行二级结构预测之前需要首先对 RCFs 的参数进行训练。参数训练是从训练数据集 D 学习每个特征的权重参数,求解向量 $\Lambda=\lambda_1,\lambda_2,\cdots,\lambda_K$ 的过程,这个过程通过最大对数似然估计实现。

将训练数据集表示为 $D=\{O^{(i)},S^{(i)}\}_{i=1}^{N}$。其中:$O^{(i)}=\{o_1^{(i)},o_2^{(i)},\cdots,o_L^{i}\}$是第 i 个输入数据序列,$S^{(i)}=\{s_1^{(i)},s_2^{(i)},\cdots,s_L^{i}\}$是相应的输出数据序列。在训练数据集 D 下对数似然为

$$L=\sum_{i=1}^{N}\log P(S^{(i)}\mid O^{(i)})\tag{4}$$

将式(1)(2)带入并进行高斯先验调整,式(4)变为

$$L=\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{k=1}^{K}\lambda_k f_k[s_{t-1}^{(i)},s_t^{i},o^{(i)},t]-\sum_{i=1}^{N}\log Z_{0(i)}-\sum_k(k^2/2\sigma^2)\tag{5}$$

其中最后一项是用于进行调整的特征参数的高斯先验值,σ^2表示先验方差。

Lafferty 建立 CRFs 时给出了一种训练 CRFs 的算法,然而这种算法对于参数较多的 RCFs 来说非常慢。2003 年 Wallach[18] 给出了一种 L-BFGS 算法对 CRFs 的参数进行训练。L-BFGS 算法通过充分利用以前的梯度和修正值来近似曲率值的二阶方法,因而使用 L-BFGS 算法进行 CRFs 训练只要求提供似然函数的一阶导数,很大程度上提高了参数训练的速度。

对于训练好的 CRFs 模型,给定待预测的序列,应用 Viterbi 算法求出产生此序列的最佳状态路径,即解码问题,其算法复杂度为 $O(\mathrm{NFL}^C)$。对于本文使用的线性链 RCFs,$C=2$。根据路径中各状态的类型即可确定序列中每个残基的预测结构。

$$S=\arg\max(S/O)=$$
$$\arg\max\sum_{j=1}^{n}\log\{\exp[\sum_{i=1}^{k}\lambda_i f_i(s_{c1},s_{c2},O)]\}/Z(o)\tag{6}$$

1.4　CB513 数据集

由于膜蛋白从组成和结构上都与球蛋白有很大区别,本文研究的是球形蛋白的二级结构,本文使用的数据从不包含膜蛋白的数据集 CB513 中进行选取。这个数据集可以从 http://www.compbio.dundee.ac.uk 下载。在使用 PSI-BLAST 进行序列对比时,删除了残基数小于 40 的序列。

2　结果与讨论

为了评价预测的准确性,使用 Q_3、C_H、C_C、C_E 及 SOV 五个指标。Q_3是测量每一个残基预测的执行情况(不是 1 就是 0),当 $x=y$,$\delta(x,y)=1$;当 $x\neq y$,$\delta(x,y)=0$。总体上蛋白质结构预测的执行情况可以用 Q_3的平均值$\overline{Q_3}$表示。其中 x 表示正确的结构;y 表示预测的结构。

$$Q_3=\delta[x,y]\tag{7}$$

$$\overline{Q_3}=(\sum Q_3)/N\tag{8}$$

其中:C_{H} 表示 H 结构中残基预测的准确度,其定义为

$$C_{\mathrm{H}}=(P_{\mathrm{H}}n_{\mathrm{H}}-u_{\mathrm{H}}o_{\mathrm{H}})/([n_{\mathrm{H}}+u_{\mathrm{H}}][n_{\mathrm{H}}+o_{\mathrm{H}}][p_{\mathrm{H}}+u_{\mathrm{H}}][p_{\mathrm{H}}+o_{\mathrm{H}}])\tag{9}$$

其中:p_{H} 为正确预测的数目,n_{H} 为正确负预测的数目,o_{H} 是过度预测(假阳性)数目,而 u_{H} 是不定预测残基(已丢失)的数目。该系数值越接近 1,说明该方法越能成功预测螺旋残基。类似地,C_{C}、C_{E} 显示对于 C 和 E 两种结构预测的准确度。

基于片断重叠的测度 SOV 考虑了 β-折叠和 α-螺旋的连续片断特性,更符合蛋白质二级结构的真实情况,其定义为式(10),更加详细的说明可以参考文献[19]。

$$\mathrm{SOV}=[(1/N)\sum_{1\in\{\mathrm{H,E,C}\}}\sum_{S(i)}[\mathrm{minov}(s_1+s_2)+\delta(s_1,s_2)]/\mathrm{minov}(s_1,s_2)\times\mathrm{len}(s_1)]\times100\tag{10}$$

五个指标从不同方面反映预测的准确度。本文将筛选过的 CB513 分为八组,每组约为 50 条链,每组分别作为测试集,剩下的七组作为训练样本集进行预测。预测结果如表 1 所示。

表1　预测结果

	Q_3	C_{H}	C_{E}	C_{C}	SOV		Q_3	C_{H}	C_{E}	C_{C}	SOV
1	76.8	80.4	66.3	80.3	73.8	6	78.2	80.2	63.7	82.4	75.1
2	78.3	78.6	63.4	81.6	72.1	7	77.4	78.9	68.5	80.5	74.3
3	77.4	79.3	67.9	79.7	72.9	8	78.1	80.7	66.5	80.3	74.7
4	76.7	81.2	65.8	79.8	73.7	最小值	76.7	78.6	63.4	79.7	72.1
5	77.5	79.9	66.2	80.4	71.9	平均值	77.5	79.9	66.0	80.6	73.6

本文进一步将 Q_3 和 SOV 的平均值与其他文献中使用支持向量机[6] 和最大熵模型[8] 得到的结果进行比较,比较结果如表 2 所示。从表中可以看到,两个指标均得到了提高,特别是 Q_3提高了约 1%。

表 2　与其他预测方法的比较

方法	Q_3	SOV
支持向量机	76.6	73.5
最大熵模型	76.9	76.1
条件随机场	77.7	73.6

3　结束语

本文成功地将条件随机场应用于蛋白质三状态二级结构预测,并取得了比较好的预测结果,类似的方法可以扩展到对蛋白质八状态的二级结构预测。本文发现,对于 E 的预测准确度低于对于 C 和 H 的预测,并且 SOV 的提高并不明显,需要进一步研究改进蛋白质结构图模型的建立,提高预测准确度。对于 RCFs 的训练的算法速度有待进一步的提高。(下转第 839 页)

的方法扩展到在最小化聚类数的同时最优化适应度函数的多目标聚类问题上。

参考文献：

[1] PRIBYL O, GOULIAS K G. Simulation of daily activity patterns incorporating interactions within households: algorithm overview and performance[J]. Journal of Trans on Research Board, 2005, 1926(1):135-141.

[2] EVERITT B S, LANDAU S, LEESE M. Cluster analysis[M]. London: Hodder Arnold, 2001.

[3] KAUFMAN L, ROUSSEEUW P J. Finding groups in data: an introduction to cluster analysis[M]. New York: Wiley, 1990.

[4] 陈群，姚加林，晏克非. 基于遗传算法的通道上停车换乘量确定方法[J]. 计算机工程，2008，34(4):201-206.

[5] 王　，段常贵. 改进遗传算法在燃气管网布局优化中的应用[J]. 哈尔滨工业大学学报，2006，38(1):46-48.

[6] WANG H, XUE D. An intelligent zone-based delivery scheduling approach[J]. Computers in Industry, 2002, 48(2):109-125.

[7] 张伟，廖晓峰，吴中福. 一种基于遗传算法的聚类新方法[J]. 计算机科学，2002，29(6):114-116.

[8] 张强，李淼. 基于遗传算法和遗传模糊聚类的混合聚类算法[J]. 计算机工程与应用，2007，43(3):164-166.

[9] 刘婷，郭海湘，诸克军，等. 一种改进的遗传 K-means 聚类算法[J]. 数学的实践与认识，2007，37(8):106-111.

[10] SCHWANEN T, DIJST M. Time window in workers' activity patterns: empirical evidence from the Netherlands[J]. Transportation, 2003, 30(3):261-283.

[11] 邵昀泓. 基于活动的出行需求分析及信息影响研究[D]. 南京：东南大学，2006.

[12] HUANG Zhe-xue. Clustering large data sets with mixed numeric and categorical values[C]//Proc of the 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining. 1997:283-304.

[13] ZHOU En-wang, KHOTANZAD A. Fuzzy classifier design using genetic algorithms[J]. Pattern Recognition, 2007, 40(12):3401-3414.

[14] MAULIK U, BANDYOPADHYAY S. Genetic algorithm-based clustering technique[J]. Pattern Recognition, 2000, 33(9):1883-1893.

[15] LUCASIUS C B, DANE A D, KATEMAN G. On K-medoid clustering of large data sets with the aid of genetic algorithm: background, feasibility and comparison[J]. Analytica Chimica Acta, 1993, 282(3):647-669.

[16] 傅景广，许刚，王裕国. 基于遗传算法的聚类分析[J]. 计算机工程，2004，30(4):122-124.

[17] ESTVILL-CASTRO V, MURRAY A T. Spatial clustering for data mining with genetic algorithms[C]//Proc of the International ICSC Symposium on Engineering of Intelligent Systems. 1998:3-9.

[18] 王小平，曹立明. 遗传算法：理论、应用及软件实现[M]. 西安：西安交通大学出版社，2002.

[19] 雷英杰，张善文，李续武，等. MATLAB 遗传算法工具箱及应用[M]. 西安：西安电子科技大学出版社，2006.

[20] PRINZIE A, POEL A V D. Incorporating sequential information into traditional classification models by using an element/position-sensitive SAM[J]. Decision Support Systems, 2006, 42(2):508-526.

(上接第 833 页)

参考文献：

[1] UNGER R, MOULT J. Genetic algorithms for protein folding simulations[J]. Journal of Moleculor Biology, 1993, 231(1):75-81.

[2] JONES D T. Protein secondary structure prediction based on position-specific scoring matrices[J]. Journal of Moleculor Biology, 1999, 292(2):195-202.

[3] LIANG Fa-ming, WONG W H. Evolutionary Monte Carlo for protein folding simulations[J]. Journal of Chemical Physics, 2001, 115(7):3374-3380.

[4] HUA Su-jun, SUN Zhi-rong. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach[J]. Journal of Moleculor Biology, 2001, 308(2):397-407.

[5] WARD J J, McGUFFIN L J, BUXTON B F, *et al*. Secondary structure prediction with support vector machines[J]. Bioinformatics, 2003, 19(13):1650-1655.

[6] KIM H, PARK H. Protein secondary structure prediction based on an improved support vector machines approach[J]. Protein Engineering, 2003, 16(8):553-560.

[7] GUO Jian, CHEN Hu, SUN Zhi-rong, *et al*. A novel method for protein secondary structure prediction using dual-layer SVM and profiles[J]. Proteins, 2004, 54(4):738-743.

[8] QIN San-bo, HE Yun, PAN Xian-ming. Prediction protein secondary structure and solvent accessibility with an improved multiple linear regression method[J]. Proteins, 2005, 61(3):473-480.

[9] KARPLUS K, BARRETT C, HUGHEY R. Hidden Markov models for detecting remote protein homologies[J]. Bioinformatics, 1998, 14(10):846-856.

[10] DURBIN R, EDDY S, KROGH A, *et al*. Biological sequence analysis: probabilistic models of proteins and nucleic acids[M]. Cambridge: Cambridge University Press, 1998.

[11] BYSTROS C, THORSSON V, BAKER D. HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins[J]. Journal of Moleculor Biology, 2000, 301(1):173-190.

[12] CHRISTOS L, COSTAS P, THEMIS P, *et al*. Sequence-based protein structure prediction using a reduced state-space hidden Markov model[J]. Computers in Biology and Medicine, 2007, 37(9):1211-1224.

[13] CHU Wei, GHAHRAMANI Z, DAVID L, *et al*. A graphical model for protein secondary structure prediction[C]//Proc of the 21st International Conference on Machine Learning. 2004:161-168.

[14] LIU Yan, CARBONELL J, KLEIN-SEETHARAMAN J, *et al*. Comparison of probabilistic combination methods for protein secondary structure prediction[J]. Bioinformatics, 2004, 20(17):3099-3107.

[15] LAFFERTY J, McCALLUM A, PEREIRA F. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]//Proc of the 18th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 2001:282-289.

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

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

[18] WALLACH H M. Efficient training of conditional random fields[D]. Edinburgh: University of Edinburgh, 2003.

[19] ZEMLA A, VENCLOVAS C, FIDELIS K, *et al*. A modified definition of SOV, a segment-based measure for protein secondary structure prediction assessment[J]. Proteins: Structure, Function, and Bioinformatics, 1999, 34(2):220-223.