

Profile 覆盖算法在蛋白质二级结构预测中的应用

郑婷婷^{1,2}, 毛军军^{1,2}, 吴涛^{1,2}, 程家兴²

(1. 安徽大学 数学与计算科学学院, 安徽 合肥 230039;

2. 安徽大学 计算智能与信号处理教育部重点实验室, 安徽 合肥 230039)

摘要:介绍了构造性机器学习方法——覆盖算法在蛋白质二级结构预测中的应用。相比普通的神经网络, 这种方法直观且运算简单, 对训练样本可100%识别。同时, 考虑到同源家族的结构应该比单条序列结构预测更准确, 采用了基于概率的Profile编码方式, 相比以往的预测方法, 具有更好的稳定性和精确性。

关键词:氨基酸序列; 蛋白质二级结构; 覆盖算法; Profile编码

中图分类号: TP18

文献标识码: A

文章编号: 1673-629X(2007)09-0171-03

Application of Profile Covering Method in Protein Secondary Structure Prediction

ZHENG Ting-ting^{1,2}, MAO Jun-jun^{1,2}, WU Tao^{1,2}, CHENG Jia-xing²

(1. School of Mathematics and Computational Science, Anhui Univ., Hefei 230039, China;

2. Ministry of Education Key Lab. of Intelligent Computing & Signal Processing, Anhui Univ., Hefei 230039, China)

Abstract: Mainly introduces protein secondary structure prediction based on structural machine learning—covering algorithm. Compared with common neural network, this approach is more easily understood and simpler, and its identification can get to 100%. At the same time, considering accuracy rate of homologous family structure prediction should be higher than of single sequence structure prediction, use Profile encoding, compared with obvious ways, which has better stability and higher accuracy rate.

Key words: amino acid sequences; protein secondary structure; covering algorithm; Profile encoding

0 引言

在分子生物学中, 通过有效地预测蛋白质二级结构可以比较准确地预测蛋白质分子的三维空间模型, 对蛋白质序列的分析、结构序列的缠绕及确定蛋白质分子功能具有重要意义^[1]。蛋白质是在分子级执行最基本生命功能的多肽链, 多肽链被认为是20个氨基酸字符(每个氨基酸用一个字符表示)的线性序列, 它折叠成为与其功能相应的复杂的三维结构。预测蛋白质如何折叠的关键一步是预测它的二级结构, 二级结构由局部折叠规则构成, 常常由氢键维持, 并且传统上被分为 α 螺旋, β 片段, γ 卷。蛋白质的氨基酸字符与它

的二级结构具有相关性。因此, 二级结构预测是计算分子生物学中的分类问题或者说是数学中的多维空间非线性映射问题。

近来, 许多机器学习的智能模型(如: 人工神经网络、隐马尔可夫模型、支持向量机、遗传算法、概率图模型等)都被用于二级结构预测中^[2]。总结起来, 有三类方法比较常用:

(1) 基于统计的预测方法^[3,4], 如 Chou-Fasman 法, GOR 法, 最小邻近法, HMM 法, 机器学习法。

(2) 基于知识的预测方法^[5,6], 如 Lim 法, Cohen 法。

(3) 混合方法^[7], 即选择性合并以上的若干种方法。

文中利用覆盖算法结合 Profile 编码对蛋白质二级结构进行预测, 取得了较好的效果。

1 覆盖算法

传统神经网络结构复杂、学习速度慢、运行效率

收稿日期: 2006-11-24

基金项目: 国家自然科学基金(60475017); 安徽省自然科学基金(050420208); 安徽省高校省级自然科学基金(2006KJ244B); 安徽大学人才队伍建设基金

作者简介: 郑婷婷(1978-), 女, 安徽合肥人, 讲师, 博士研究生, 研究方向为智能计算、神经网络理论与应用、生物信息分析; 程家兴, 教授, 研究方向为智能计算、算法分析与设计、最优化方法。

低、难以解决海量数据的处理。张铃教授提出的基于覆盖的构造性机器学习方法^[6],针对样本本身的特点构造神经网络,几何意义明确,在学习过程中只需内积运算,不需优化等过程,方法直观且运算简单,对训练样本可 100% 识别,这是一般的分类器难以做到的。而且这种分类方法可以适用于多类别情形,从而有效地解决了神经网络结构难以确定和运行速度慢等问题^[9]。

设学习样本共有 N 类,记为: $X = \{X_1, X_2, \dots, X_N\}$,交叉覆盖算法是用“球形领域作神经元,构造三层网络,将各类样本分开。其基本思路是:依次轮流构造各类别的球形领域,直至所构造的领域系能够盖住所有的训练样本点。学习过程中构造第 k 类学习样本 X_k 的“球形领域”的方法是:任取 X_k 中尚未被覆盖的点 a_i ,按公式:

$$d^1(\omega) = \max_{x \in X_k} |< a_i, x >|$$

$$d^2(\omega) = \min_{x \in X_k} |< a_i, x >| < a_i, x > > d^1(\omega) \}$$

$$d(\omega) = \frac{1}{2} (d^1(\omega) + d^2(\omega))$$

计算 $d(\omega)$,作以 a_i 为中心、阈值 $\theta = d(\omega)$ 的覆盖 $C(a_i): < \omega, x > - \theta > 0$ 。并通过求球形领域重心和平移,使之可以覆盖更多的样本点,并按此方法求出样本的全部覆盖。取功能函数为:

$$F: y = \sigma(< \omega, x > - \theta), \sigma(x) = \begin{cases} x, & x > 0 \\ 0, & \text{其它} \end{cases}$$

识别的方法是:给定一个样本,若它属于某类覆盖的一个“球形领域”,即可确定其类别属性。

2 数据集和编码

2.1 数据集

对于检验模型的预测精确,数据集的选择是一个比较难的问题,需要机器学习领域的知识和生物学领域的知识。如果选择的数据集不能反映实际的数据分布情况或包含矛盾的信息,那么就会导致训练的分类模式偏向于某一特定类别。另外数据中不能包含人造相关性,否则将导致过高估计分类器的有效性。我们所使用的蛋白质数据是生物信息领域内常用的一个数据集——RS126。这是 Rost 和 Sander^[10]提出的,而后被广泛用于蛋白质二级结构预测研究的一个典型数据集,共包含 126 条序列,其中长度大于 80 个氨基酸的序列的同源性低于 25%。

有关二级结构的分类有多种方法^[11],这里利用 DSSP 的分类方法将 PDB 库中的结构分为 Helices(H) 和 Sheets(E),既非 Helices 又非 Sheets 的全部归于 Coils 类(C)。

2.2 编码方式

用构造性学习算法进行预测时,编码技术也是相当关键。常用的编码方式有:正交编码法、极性正交编码法、5 位编码法、体积极性正交编码法等。这里采用的编码方式是 Profile 编码。所谓 Profile 编码是指在氨基酸序列的每个位置上一个氨基酸类型出现的相对概率。蛋白质的基本信息来自序列,但从单个序列中很难获取,只有通过多个序列的比较,才能获取有用信息,所以建立多序列比对是识别结构域的重要手段。这种编码方式正是源自多重序列比对的位置相关的频率向量,首先从 PDB 数据库中寻找同源序列进行比对。对于某一个位置上的氨基酸,求出所有序列中 20 种氨基酸各自所占的比重。将这个比重值(总共有 20 个)就作为原序列中这一位置的氨基酸表示方式,显然,每个氨基酸应该是一个 20 维的数组。

3 结果与讨论

采用覆盖算法作为分类器,选取 RS126 中的长度大于 80 的 90 个序列,共 16585 个氨基酸作为数据集。这里对分类器的评价采用了 7 次交叉验证的方法。

文中采用的是被广泛采用的评估公式来预测精度:

$$Q_3 = \frac{P_e + P_\beta + P_C}{N}$$

其中: Q_3 表示总的预测精度, P_e, P_β, P_C 分别表示判断正确的 α 类(H)、 β 类(E) 和 coils 类(C) 的氨基酸的个数, N 是所有氨基酸的总和。

为了说明覆盖算法在蛋白质二级结构预测中的优越性,与前人所做出的结果进行了比较。结果如下:

正确率 (%)	Multi-model ^[12]	BP 神经网络 ^[13]	Psipred	覆盖算法
Q_3	66.2	70.3	76.5	77.223

可以看出,覆盖算法在提高预测精度上,明显好于其他的方法。

另外,在这 90 个序列中挑选有代表性的几个蛋白质家族的蛋白作为测试集。按照 SCOP 分类法,将不同结构的蛋白质分别做测试,主要包括 α , 全 α , α/β 以及 $\alpha + \beta$ 四种结构类型。试验结果如下:

	覆盖数	训练时间(s)	正确率 (%)
全 α	1182	24.826	77.955
全 β	2436	112.98	70.853
α/β	1674	57.857	78.053
$\alpha + \beta$	1392	38.158	76.013

由此,我们认为覆盖算法在四种类型的识别上,预测精度差别应该不是很大。全 β 的识别率稍低,而覆盖数却相对大,可能与其结构本身的复杂度有关,因为相对其他几种结构来说, β 折叠更具有延展性。

相信对覆盖算法的进一步改进以及对数据集进行基于物化性质的预处理后,会对精度提高有更大的帮助。

参考文献:

- [1] Scott C S. Bayesian segmentation of protein secondary structure[J]. Journal of Computational Biology, 2000, 7(1/2): 233-248.
- [2] Geoffrey B J. Protein Secondary Prediction[J]. Curr Opin in Struct Biol, 1995, 5(3): 372-376.
- [3] Bohr H, Bohr J, Brunak S, et al. Protein secondary structures and homology by neural networks: the α -helices in rhodopsin[J]. FEBS Letters, 1988, 241(2): 223-228.
- [4] Baldi P, Chauvin Y, Hunkapilla T, et al. Hidden Markov models of G-protein-coupled receptor family[J]. Comput Biol 1994(1): 311-335.
- [5] Baldi P, Brunak S, Frasconi P, et al. Bidirectional, dynamics for protein secondary structure prediction[C]// In Sun R, Giles C L. Sequence Learning: Paradigms, Algorithms, and Application. New York: Springer Verlag, 2000: 99-120.
- [6] Baldi P, Brunak S, Frasconi P, et al. Exploiting the past and the future in protein secondary structure prediction[J]. Bioinformatics, 1999, 15: 937-946.
- [7] Levin J, Pascarella S, Argos P, et al. Quantification of secondary structure prediction improvement using multiple alignments[J]. Prot Eng, 1993, 6: 849-854.
- [8] Zhang L, Zhang B. A geometrical representation of McCulloch-Pitts neural model and its application[J]. IEEE Trans on Neural Networks, 1999, 10(4): 925-929.
- [9] 张铃, 张钹. M-P神经网络模型的几何意义及其应用[J]. 软件学报, 1998, 9(5): 334-338.
- [10] Rost B, Sander C. Improved prediction of protein secondary structure by use of sequence profiles and neural network[J]. Proc Nat Acad Sci USA, 1993, 90(16): 7558-7562.
- [11] Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features[J]. Biopolymers, 1983, 22(12): 2577-2637.
- [12] Zhu H X, Yoshihara I. Prediction of protein secondary structure by multi-model neural networks[C]// International Joint Conference on Neural Networks. Montréal: [s. n.], 2002: 280-285.
- [13] Rost B, Sander C. Prediction of Protein Secondary Structure at Better than 70% Accuracy[J]. J Mol Biol, 1993, 232: 584-599.

(上接第148页)

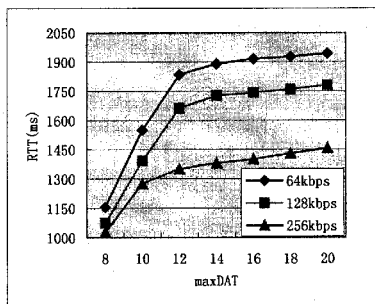


图5 BLER=10%时不同带宽下的RTT值

4 结束语

分别从理论推导和仿真建模两方面分析了max-DAT的不同取值对系统性能的影响。重点考察了无线链路的BLER、吞吐量以及RTT与maxDAT之间的关系,还考察了同一BLER条件下不同带宽下max-DAT与吞吐量和RTT之间的关系。由于无线链路的

复杂性,同时对于不同业务其性能要求也会有所不同,因此,在实际应用中要充分考虑不同的业务特性,从而选择合适的maxDAT值,使得系统性能达到最优化。

参考文献:

- [1] 3GPP TS 25.301 v4.12.0. Radio Interface Protocol Architecture[S]. 3GPP Organizational Partners (ARIB, CCSA, ETSI, T1, TTA, TTC). 2004.
- [2] 3GPP TS 25.401 v6.0. UTRAN Overall Description[S]. 3GPP Organizational Partners (ARIB, CCSA, ETSI, T1, TTA, TTC). 2003.
- [3] 3GPP TS 25.322 v4.6.0. Radio Link Control (RLC) Protocol Specification[S]. 3GPP Organizational Partners (ARIB, CCSA, ETSI, T1, TTA, TTC). 2002.
- [4] Xylomenous G. TCP Performance Issues over Wireless Links[J]. IEEE Communications Magazine, 2001, 8(2): 133-145.
- [5] LEFEVRE F, Guillaume. Optimizing UMTS Link Layer Parameters for a TCP Connection[C]//Proc. IEEE Conf on Vehicular Technology (VTC 2001 Spring). [s. l.]: [s. n.], 2001: 161-164.