使用组合分类器预测蛋白质相互作用

周正荣,宋晓峰,王明浩
(南京航空航天大学自动化学院生物医学工程系,江苏南京 210016)

摘 要: 蛋白质必须通过与其他蛋白质之间的相互作用才能行使其功能,因此,对于蛋白质相互作用的研究显得尤为重要.针对蛋白质相互作用预测问题,本文提出了一种基于不同特征编码的组合分类器投票的预测方法.该方法综合考虑了蛋白质序列中氨基酸频率、氨基酸位置、氨基酸理化性质和氨基酸生物相似性等特征.在真实的蛋白质相互作用human数据集上,使用支持向量机根据不同特征编码建立的预测模型,分别作为组合分类器中的子分类器进行投票.结果表明,该方法能有效提高蛋白质相互作用预测的性能,且预测结果与其他方法相比也具有一定优势.

关键词: 蛋白质相互作用; 编码; 支持向量机; 组合分类器投票
中图分类号: TP274 **文献标识码:** A **文章编号:** 0372-2112 (2010) 06-1464-04

Predicting Protein-Protein Interactions Based on Ensemble Classifiers

ZHOU Zheng-rong, SONG Xiao-feng, WANG Ming-hao
(*Department of Biomedical Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing* 210016 *China*)

Abstract: Most proteins perform their function by interacting with other proteins. Thus, the research of Protein-Protein Interactions (PPI) is becoming more and more important. A new multi-encoding of sequence method, which represents the sequence of protein more completely, is presented to predict the protein-protein interactions. We train three different support vector systems based on multi-features and then combine three outputs of three different classifiers to vote the optimal result. Result shows that our method can improve predictive ability, moreover, our method show better performance than some of the previously developed methods in present research.

Key words: protein-protein interactions (PPI); encoding; support vector machine (SVM); vote

1 引言

蛋白质的功能必须通过其相互作用表现出来,这使得对蛋白质相互作用的研究显得尤为重要.随着高通量生物实验技术的应用,如酵母双杂交法[1]、质谱法[2]、蛋白芯片技术[3],人们获得了大量的蛋白质相互作用数据.然而,实验方法获得的结果中通常包含大量的假阳性和假阴性数据[4,5],并且这类方法获得的数据远远不够全面.因此,需要有效的计算方法来预测蛋白质相互作用.

现有的预测方法主要依据相互作用的蛋白质在其序列、结构、domain、功能、进化信息、亚细胞定位等方面有着相近或者相同的特征,通过研究这些特征来预测蛋白质相互作用.例如,Shawn M. et al[6]假定蛋白质相互作用网络中的每对蛋白质都有一定概率存在相互作用,通过表征网络拓扑结构的变量 global 和网络中局部蛋白质相互作用的变量 local 之间建立的概率模型来预测蛋白质的相互作用.文献[7~9]提出的方法主要是通过研究蛋白质序列和 domain 序列,构建相应的机器学习模型,来寻找蛋白质相互作用所遵循的规律.Shawn Martin et.al[7]根据蛋白质序列信息提出"signature product"的序列编码的方式,并用蛋白质之间编码的乘积作为 SVM 的核函数.Loris Nanni et.al[8]采用"2-Gram"编码方式对蛋白质序列进行编码,1-Gram 是氨基酸字母关键字,1-Gram 是该关键字在序列中出现的频率.然后,对线性分类器和"云"分类器加权融合.文献[9]把 SVM 核函数细分为蛋白质对核函数、序列核函数、非序列核函数.通过融合这些核函数来预测蛋白质相互作用.目前,研究人员认为蛋白质最小功能单元是其一段子序列 domain,domain 之间相互作用一定程度上反映对应蛋白质之间相互作用[10~12].文献[10,11]通过建立蛋白质相互作用与其 domain 相互作用的概率模型预测蛋白相互作用.WangLang et.al[12]依据氨基酸理化性质对 domain 序列编码,并假定蛋白质之间相互作用的概率是就是其 domain 发生相互作用的最大概率.然而,对于上述文献中提到的序列编码方式,文献[7,8]都只是考虑了蛋白质序列中氨基酸频率和位置特征,而没有包含氨基酸的生物学特征,如理化性质、生物相似性.文献[12]对 domain 的编码没有

收稿日期:2009-02-07;修回日期:2009-04-07

考虑氨基酸在序列中出现的位置、频率、生物相似性等特征.使用序列信息预测蛋白质相互作用时,对蛋白质序列的编码必须尽量全面、准确的反映蛋白质序列信息[7].上述方法种所提到的几种编码方式中没有全面考虑蛋白质序列信息."没有免费午餐定理"说明在缺少对样本足够的先验知识的情况下,我们不能偏爱某种算法或者模型[13].因此,对于这些没有全面反映序列特征的编码算法,同样也不能确定其对未知样本分类性能的优劣.为了提高预测算法的泛化能力,本文首先使用能较为全面反映序列信息的氨基酸位置、频率、氨基酸理化性质、氨基酸生物相似性等分别特征进行编码.然后,对不同编码结果训练 SVM 预测模型,通过组合分类器投票的方式预测蛋白质相互作用.最后使用与文献[7,14]相同的 human 数据集来验证了我们的算法,最终的分类结果不仅高于分别对序列不同特征编码的分类结果,而且优于文献[7,14].

2 序列编码算法

由于不同蛋白质序列长度可能不一致,所以在使用 SVM 预测蛋白质相互作用时,需把不同长度的蛋白质序列编码转变为长度一致的 SVM 训练数据.

2.1 K-Mers Count(KMC)

本算法主要对氨基酸的理化性质特征进行编码[12].20 种氨基酸按照其 m 种不同的物化性质可以被分为 n 组,分别用 $P_i(0<i<n+1)$表示.用表示一条由长度为 L 氨基酸构成的蛋白质序列.基于 m 种理化性质可以把表示成由字母 P_i 构成的字母序列($0<k<m+1$).对于 $\sum_0^k(0<k<m+1)$ 中的每条序列采用长度为 K 的滑动窗口 W 沿着序列每个字母滑动,就可以得到$(L-K+1)$个长度为 K 的字母关键字.例如,对于序列 $abcd$,长度为 2 的窗口滑动 W 可得到的字母关键字为 ab,bc,cd.统计这些关键字出现的频率,这样可以把一条蛋白质序列化为长度为 $m\times n^K$ 维 SVM 可训练的数字向量.本文采用五种物化性质[12,15],每种理化性质把氨基酸分为 3 组,即 $m=5$,$n=3$.窗口长度 $K=3$,那么不同长度的蛋白质序列就可以转化为维数相同的 $5\times3^3=135$ 维的特征向量.

2.2 K-Nearest-Neighbor(KNN)

本编码算法主要对蛋白质序列中氨基酸位置、频率特征编码.对长度为 L 的氨基酸序列,可以将其划分为 $L-2k$ 个由氨基酸及其 k 近邻构成的氨基酸字母关键字[7].用 $\sum_0$ 表示由氨基酸组成的蛋白质序列,A_i 为 $\sum_0$ 第 i 个氨基酸字母.若 $i=1$,A_i 的 1 近邻 A_{i-1}和 A_{i+1}可以组成字母关键字 A_iA_{i-1},A_{i+1}.为了简化维数大小,假定 $A_iA_{i-1}\ A_{i+1}=A_iA_{i+1}\ A_{i-1}$,统计关键字出现的频率.本文中,令 $k=1$,不同长度的氨基酸序列可以转化为 $C_{20}^1\times\sum_{i-1}^{20}i=4200$ 维的特征向量.

2.3 Biosimilarity(BIO)

此算法主要对氨基酸生物相似性特征进行编码.根据氨基酸之间的生物相似性可以把 20 种氨基酸分为 6 类,分别为 IVLM、FYW、HKR、DE、QNDP、ACGS[6,16],用 $A_i(0<<i<7)$表示这六类.由氨基酸构成蛋白质序列用 $\sum_0$ 表示.依照氨基酸生物相似性把 $\sum_0$ 表示成由 Ai 表示的字母序列 $\sum_0^b$.采用窗口长度为 k 的滑动窗口 W 对 $\sum_0^b$ 的每个字母依次滑动,得到 $L-k+1$ 个长度为 k 的字母关键字.例如,$k=3$,由 A_i 组成的字母序列 $A_1A_2A_3A_4A_5A_6$,其关键字分别为:$A_1A_2A_3$,$A_2A_3A_4$,$A_3A_4A_5$,$A_4A_5A_6$,统计这些关键字出现的频率.本文中,令 $k=3$,对于不同长度的蛋白质序列,可以转化为 $6^3=216$ 维数字向量.

表 1 氨基酸相似性的分类表

Bio-similarity					
IVLM	FYW	HKR	DE	QNDP	ACGS
A_1	A_2	A_3	A_4	A_5	A_6

3 SVM(Support Vector Machine)

SVM 是两类样本的分类法则,其分类方法思想是首先通过非线性变换将其输入空间变换为一个高维空间,然后在这个新空间中求取最优线性分类面,而这种非线性变换是通过定义适当的内积函数实现的.判别函数为:

$$f(x)=\mathrm{sgn}\left(\sum_{i=1}^{n}a_i^*y_ik(x_i,x)+b^*\right) \tag{1}$$

SVM 已经被成功用于解决很多生物信息学问题,如基因表达数据的分类[16]、同源性预测[17]、蛋白质相互作用预测[18].本文采用 libsvm-2.86 版本[19].在选取最优参数时,一种是采用交叉验证的方法,另外也可以采用其自动寻优工具寻找最优参数,gnu-plot 可动态显示其寻优过程.

4 技术与方法

预测蛋白质之间未知相互作用时,若其序列编码方式不能较为全面反映其特征,则可能会造成序列信息的缺失,从而导致更多的分类误差.本文基于蛋白质序列预测蛋白质之间相互作用,使用能较为全面反映蛋白质序列信息的氨基酸频率、位置、理化性质、生物相似性等特征分别编码,利用组合分类器投票的方法

预测蛋白质之间未知的相互作用.

4.1 组合分类器的基本结构

图 1 为组合分类器的基本结构,假设样本的每个分类模式都是取自组合分类器模型,训练样本 x 输出结果概率分布为 :

$$P(y\mid x,\Theta_0^0)=\sum_{r=1}^{k}P(r\mid x,\theta_0^0)P(y\mid x,\theta^0) \quad (2)$$

其中 $\Theta^0=(\Theta_1^0,\Theta_2^0\cdots\Theta_k^0)^t$ 表示全部有关的参数向量.

样本 x 被提供给 r 个分量分类器,每一个都输出 c 个标量的判别函数值.这样对分量分类器的 c 个判别值组织在一起,记作 $g(x,\theta_r)$,且有

$$\sum_{j=1}^{c}g_{rj}=1 \quad (3)$$

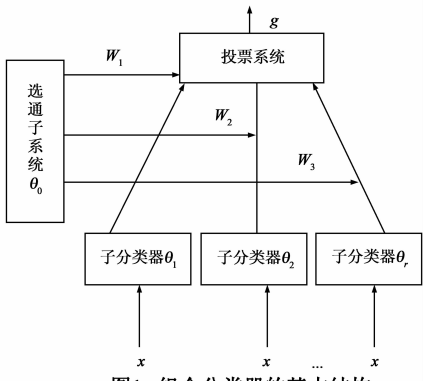

图1 组合分类器的基本结构

分量分类器的输出全部判别值都乘上权值 Wr,Wr 由"选通子系统"给定.在图 1 组合分类器结构图中,选通子系统表达了式(2)中参数 $P(r\mid x,\theta_0^0)$的模型.

4.2 组合分类器投票

本文采用组合分类器投票的方式来预测蛋白质的相互作用.在组合分类器基本结构中,三个子分类器分别是基于三种编码方式得到的 SVM 预测模型.

4.2.1 基于概率的组合分类器投票(Voting Based On Probability,VBP)

组合分类器投票系统中,对未知样本 x,设其有 $Cj,(j=1,2\cdots c)$个分类类别,假设子分类器 θ_r 把 x 划分为类别 $Cj(0<j<c+1)$的概率为,那么 $\sum_{j=1}^{c}g_{rj}=1$.所以,样本 x 组合分类器投票结果 g 可以表示为:

$$g=\sum_{r=1}^{n}W_r\sum_{j=1}^{c}g_{rj}C_j \quad (4)$$

因为 PPI 预测是一个二分类问题,所以,在使用 VBP 组合分类器投票时,可令 $C1=1$ 表示 x 为正样本,$C2=-1$ 表示 x 为负样本.当 $g>0$,表示样本 x 为正样本,当 $g<0$,表示 x 为负样本.

4.2.2 基于成员的组合分类器投票(Predicting Bbased On Classifiers' Ballot,PBB)

对于样本 x,令 $V(r,Cj)$表示把 x 判别为类别 Cj 的子分类器的数目.则在 PBB 投票中,若 $V(r,Cj)=\max\{V(r,C1),\cdots,V(r,Cc)\}$,且不存在任意 $V(r,Cj)=V(r,Ci)$,那么,就把样本 x 判定为类别 Cj .

对于样本 x,P_1,P_2,P_3 分别表示 KMC、KNN、BIO 三种编码的子分类器的分类准确率.P_0 表示 PBB 投票后的分类准确率,则理论上:

$$P_0=\sum_{i=1}^{3}P_i\sum_{j=i+1}^{3}P_j-2\times P_1P_2P_3 \quad (5)$$

理论上,要使得组合分类器分类准确率大于每个子分类器的准确率,则子分类器分类准确率应该满足以下约束条件:

$$P_0>P_{\max},(P_{\max}=\max\{P_1,P_2,P_3\}) \quad (6)$$

从式(6)可知,当 P_1、P_2、P_3 相差不大(大约 0 – 10%)且各子分类器先验分类结果都优于随机分类结果时,可以采用 PBB 组合分类器系统.

在使用 PBB 组合分类器投票时,若子分类器数目为偶数,则可能会出现最不确定性分类结果.即,存在任意 $V(r,Cj)=V(r,Ci)=\max\{V(r,C1),\cdots,V(r,Cc)\}$时,使用投票系统将无法对 x 的类别做出判断.使用 VBP 投票虽然不会出现最不确定性分类结果,但其权值不易确定.本文使用三个子分类器,每个子分类器分类准确率满足式(6),所以本文使用基于成员的组合分类器投票方式预测未知样本.

5 结果与分析

为了评估算法性能,本文实验数据采用了与文献[7]、[14]相一致的 human 数据集包含 800 个蛋白质之间的 1790 个蛋白质对.假样本数据由构成 1790 个真样本蛋白质对 800 个蛋白质随机组合而成,共 800 × 800 个,剔除重叠的正样本,为了平衡正负样本数目,随机选取其中 1790 个假样本.

采用广泛使用的性能评价指标对其进行测评,用 TP 表示预测正确的正样本数,TN 表示预测正确的负样本数,Acc 表示分类的准确率.三个常用的预测性能指标敏感度 Sn、专一性 SP 和准确率 Acc 分别定义为:

$$Sn=TP/(TP+FN) \quad (7)$$

$$SP=TP/(TP+FP) \quad (8)$$

$$Acc=\frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

由于我们训练和测试数据中相互作用和无相互作用的蛋白质数量一致,所以 Acc% 可以反映算法的综合性能.对于本文所选数据集,每次选取正负样本数目各为 150 的蛋白质对作为预测集,余下的为训练集.对预测集中的每组蛋白质,采用文中的三种编码方式,得到各自的分类结果.对照式(6),确定其满足组合分类器投票的约束条件,最终组合分类器投票的分类结果见表 2.从表 2 中可知,组合分类器投票结果优于三个子分类器各自的分类结果,并且我们样本的平均分类结果 73.1% 高于文献[7]同样数据集的 70.3%、文献[14]同样数据集的 72.1%.

表 2 不同预测样本分类结果

	Sample 1			Sample 2			Sample 3			Sample 4			Sample 5		
Encode	Sn	SP	Acc	Sn	SP	Acc	Sn	SP	Acc	Sn	SP	Acc	Sn	SP	Acc
KMC	0.671	0.671	0.671	0.784	0.717	0.750	0.560	0.843	0.710	0.590	0.810	0.718	0.573	0.803	0.705
KNN	0.504	0.853	0.681	0.627	0.849	0.743	0.660	0.720	0.700	0.606	0.750	0.711	0.607	0.860	0.731
BIO	0.559	0.812	0.688	0.706	0.792	0.750	0.664	0.719	0.697	691	0.709	0.703	0.592	0.853	0.716
Pbb	0.573	0.832	0.709	0.690	0.830	0.760	0.642	0.734	0.719	5	0.782	0.735	0.599	0.851	0.741

6 结论

本文使用组合分类器投票的方法预测蛋白质相互作用.本文中使用 KMC、KNN、BIO 这三种编码方式分别得到 SVM 预测模型,并作为组合分类器的三个子分类器成员.实验结果证明,组合分类器投票方法能够有效的提高最终的分类准确率,并且我们的结果要优于文献[7]、文献[14]的分类结果.

虽然取得了相对较好实验效果,但是预测结果仍然偏低.如果能找到更能反映序列特征的编码方式,并且在编码中结合更多的蛋白质生物学特征如结构、遗传信息、亚细胞定位等将可能会取得较好的预测效果.这将是我们下一步工作的重点.

参考文献:

[1] Fields S and Song O-K. A novel genetic system to detect protein-protein interactions[J]. Nature, 1989, 340: 245 – 246.

[2] Ho Y, Gruhler A, Heilbut A, Bader G D, Moore L, Adams S L, Millar A, Taylor P, Bennet K, Boutilier K et al. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry[J]. Nature, 2002, 415: 180 – 183.

[3] Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, Bertone P, Lan N, Jansen R, Bidlingmaier S, Houfek T et al. Global analysis of protein activities using proteome chips[J]. Science, 2001, 293: 2101 – 2105.

[4] Mrowka R, Patzak A, Herzel H. Is there a bias in proteome research[J]. Genome Res 2001, 11(12): 1971 – 1973.

[5] Edwards A M, Kus B, Jansen R, Greenbaum D, Greenblat J and Gerstein M. Bridging structural biology and genomics: assessing protein interaction data with known complexes [J]. Trends Genet, 2002, 18: 529 – 536.

[6] Shawn M Gomez, William Stafford Noble and Andrey Rzhetsky. Learning to predict protein-protein interactions from protein sequences[J]. Bioinformatics 2003, 19(15): 1875 – 1881.

[7] Shawn Martin, Diana Roe and Jean-Loup Faulon. Predicting PPI using signature products[J]. Bioinformatics, 2005, 21(2): 218 – 226.

[8] Loris Nann. Fusion of classifiers for predicting protein – protein interactions[J]. Neurocomputing, 2005, 68(7): 289 – 296.

[9] Asa Ben-Hur1, and William Stafford Noble. kernel methods for protein-protein interaction prediction[J]. Bioinformatics, 2005, 21(Suppl): i38 – i46.

[10] Gomez S M, Lo S H, Rzhetsky A. Probabilistic prediction of unknown metabolic and signal-transduction networks[J]. Genetics, Nov 2001, 159(3): 1291 – 1298.

[11] Gomez S M, Rzhetsky A. Towards of prediction of complete protein-protein interaction networks[J]. Pac Symp Biocomput, 2002, 7: 413 – 414.

[12] 王兰,刘融,周艳红.基于结构域组合信息预测蛋白质相互作用[J].生物信息学,2008,6(1):28 – 30.
WANGLan, LIU Rong, ZHOU Yan hong, Predict protein-protein interactions based on domain combination information [J]. China Journal of Bioinformatics, 2008, 6(1): 28 – 30. (in Chinese)

[13] Richard O Duda, Peter E Hart, and David G Stork. Pattern Classification[M]. Second Edition. VS: Wiley-Interscienle, 2003. 401 – 402.

[14] NI Qingshan, WANG Zhengzhi, WANG Guangyun, QIANG Bo. Prediction of protein-protein interactions based on local support vector machine[J]. Journal of Biomedical Engineering Research 2008, 27 (2): 69 – 73.

[15] Dubchak I, Muchnik I, Holbrook S R, et al. Prediction of protein folding class using global description of amino acid sequence[J]. Proc Natl Acad Sci, 1995, 92(19): 8700 – 8704.

[16] Taylor W R and Jones D T. Deriving an amino acid distance matrix [J]. J Theor Biol, 1993, 164(1): 65 – 83.

[17] Furey T, Cristianini N, Duffy N, Bednarski DW, Schummer M and Hau-ssler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data[J]. Bioinformatics, 2000, 16(10): 906 – 914.

[18] Leslie C, Eskin E, Weston J and Noble W. Mismatch string kernels for SVM protein classification[J]. Advances in Neural Information Processing Systems, 2003, 15: 1441 – 1448.

[19] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. A Practical Guide to Support Vector Classi cation[OL]. http://www.csie.ntu.edu.tw/~cjlin.

作者简介:

周正荣 男,1984 年生,硕士研究生,研究方向为智能生物信息处理.

宋晓峰(通信作者) 男,1968 年生,副教授,硕士生导师,研究方向为智能生物信息处理. E-mail: xfsong@nuaa.edu.cn

王明浩 男,1985 年生,研究方向为智能生物信息处理.