一种序列相关支持向量机的β桶状跨膜蛋白预测方法

吴宏杰[1,3] 吕 强[1,2] 唐剑锋[1] 钱培德[1,2]
(苏州大学计算机科学与技术学院 苏州 215006)[1]
(江苏省计算机信息处理技术重点实验室 苏州 215006)[2]
(苏州科技学院电子与信息工程学院 苏州 215011)[3]

摘 要 膜蛋白是一种具有重要生物功能的蛋白质,根据蛋白质的序列信息预测其是否属于β桶状跨膜蛋白是结构预测与功能分析的重要先导步骤,也是蛋白质预测领域中的一个挑战性问题。针对这两类问题,提取了208条β桶状跨膜蛋白序列的氨基酸位置与理化特征。利用支持向量机(SVM)进行了预测,结果表明二分类精度与相关系数分别达到了88.36%与0.7723。

关键词 跨膜蛋白,支持向量机,β桶状
中图法分类号 TP305 **文献标识码** A

Prediction of β-barrel Transmembrane Protein from Sequence Based on SVM

WU Hong-jie[1,3] LV Qiang[1,2] TANG Jian-feng[1] QIAN Pei-de[1,2]
(School of Computer Science and Technology, Soochow University, Suzhou 215006, China)[1]
(Jiangsu Provincial Key Lab for Information Processing Technologies, Suzhou 215006, China)[2]
(College of Electronic and Information Engineering, Suzhou University of Science and Technology, Suzhou 215011, China)[3]

Abstract Membrane protein is a sort of protein with important biological functions. To predict if the protein belong to β-barrel transmembrane protein according to sequence is the important precursor step for predicting 3D structure of β-barrel transmembrane protein, also is a challenging job in computational protein field. This paper introduced the feature extraction of 208β-barrel transmembrane protein sequences and prediction by SVM. The features consist of the position information in sequence, the physical and chemical properties of amino acid residues. And the results show that the accuracy and MCC of the method are 88.36% and 0.7723 respectively.

Keywords Transmembrane protein, Support vector machine, β-barrels

1 引言

膜蛋白(IMP)是嵌入在生物膜中的多肽链,在细胞中具有重要的生物功能,它们构成了各种神经信号分子、激素和受体,是各种离子跨膜的通道,也是许多药物分子的靶点。大多数膜蛋白可以分为周边膜蛋白和内在膜蛋白两类。内在膜蛋白的跨膜区结构主要由α螺旋和β折叠构成。β桶状结构只能在2个相邻的β折叠之间形成链间氢键,因此每个β折叠必须有2个相邻的β折叠才能形成一个β桶状结构,桶轴垂直于膜平面。β桶状结构的膜蛋白见于Gram-negative等细菌中,未发现同时具有α螺旋和β桶状的膜蛋白。目前只有在革兰阴性菌和线粒体以及叶绿体外膜中发现β桶状跨膜蛋白,数据量比较少。因此,如何使用少数已知的β桶状跨膜蛋白,根据蛋白质的序列信息来预测其是否属于β桶状跨膜蛋白,该问题是β桶状跨膜蛋白结构预测与功能分析的重要先导步骤,也是蛋白质预测领域中的一个挑战性问题。

现在对β桶状跨膜蛋白预测的研究有了很多进展。Diederichs等[1]提出用神经网络预测外膜β桶状蛋白拓扑结构,预测了外膜蛋白中α碳原子的z坐标,并且指出中等的z值代表跨膜蛋白β折叠。Jacoboni等[2]将神经网络法与动态规划法结合预测β桶状跨膜蛋白。Natt等[3]在人工神经网络法上加入支持向量机方法,预测β桶状跨膜蛋白。Daisuke等[4,5]提出了模糊支持向量机的方法,将模糊隶属度引入到支持向量机中预测。

由于跨膜蛋白所属类别与序列中各个氨基酸的物理、化学性质密切相关,而且这种相关性并不是简单的位置、属性值之间的一一对应关系,因此本文提出一种以氨基酸物理化学指数相关性为特征的支持向量机预测方法。方法中使用氨基酸组分融合氨基酸物理化学特性方法对208条β桶状跨膜蛋

到稿日期:2011-09-23 返修日期:2011-12-23 本文受国家自然科学基金项目(60970055,61170125)资助。
吴宏杰(1977-),男,博士生,主要研究方向为生物信息学、模式识别;吕 强(1965-),男,博士,教授,博士生导师,主要研究方向为生物信息计算、元启发搜索、并行计算;唐剑锋(1982-),男,硕士生,主要研究方向为生物信息学,E-mail:jerry_tang1209@126.com;钱培德(1947-),男,教授,博士生导师,主要研究方向为中文信息处理。

白序列进行特征提取，然后利用高斯径向基为核函数的支持向量机对β桶状跨膜蛋白进行预测，其二分类精度与相关系数分别达到了88.36%与0.7723，高于现有的主流方法。

2 数据和方法

2.1 数据集

本文用于训练集和测试集使用了由Gromiha和Suwa收集[6]，并且由Park等人[7]通过CD-HIT程序(http://bioinformatics.org/cd-hit/)去除序列相似度大于40%的冗余序列后得到的一个数据集[7]。该数据集合组成：TMB(208条)，GP673条，其中155条全α、156条全β、183条α+β和179条α/β，此球状蛋白序列用于数据集的负样本)。该非冗余数据集中蛋白质详细的3D结构信息可以从PDB数据库中获得。

氨基酸残基指数是反映氨基酸不同物理化学性质的数值，我们从Kawashima等[8]的氨基酸残基指数数据库中，选取5种最具代表性的指数，即疏水性(Hydrophobicity)、亲水性(Hydrophilicity)、柔性(Flexibility)、转移动力自由能(Free energy transfer)和极性(Polarity)，具体数值如表1所列。

表1 5种氨基酸残基指数值

指数名称	指数值									
	A	R	N	D	C	Q	E	G	H	I
	L	K	M	F	P	S	T	W	Y	V
疏水性[9]	0.25	−1.76	−0.64	−0.72	0.04	−0.69	−0.62	0.16	−0.4	0.73
	0.53	−1.1	0.26	0.61	−0.07	−0.26	−0.18	0.37	0.02	0.54
亲水性[10]	−0.5	3	0.2	3	−1	0.2	3	0	−0.5	−1.8
	−1.8	3	−1.3	−2.5	0	0.3	−0.4	−3.4	−2.3	−1.5
柔性[11]	0.357	0.529	0.463	0.511	0.346	0.493	0.497	0.544	0.323	0.462
	0.365	0.466	0.295	0.314	0.509	0.507	0.444	0.305	0.42	0.386
转移动力自由能[12]	−0.2	−0.12	0.08	−0.2	−0.45	0.16	−0.3	0	−0.12	−2.26
	−2.46	−0.35	−1.47	−2.33	−0.98	−0.39	−0.52	−2.01	−2.24	−1.56
极性[13]	8.1	10.5	11.6	13	5.5	10.5	12.3	9	10.4	5.2
	4.9	11.3	5.7	5.2	8	9.2	8.6	5.4	6.2	5.9

2.2 支持向量机的分类方法

支持向量机(SVM)是一种有效的机器学习分类方法[14-16]，它将原训练数据映射到较高的维，在新的维上搜索线性最佳边缘超平面。支持向量机常用的核函数有：线性核：$K(x,y)=x\cdot y$；多项式核：$K(x,y)=[(x\cdot y)+1]^d$，$d=1,2,\cdots$；高斯径向基核函数(RBF)：$K(x,y)=\exp(-\gamma|x-y|^2)$。本文采用高斯径向基核函数，参数选择时尝试了几组不同的参数值[17]，当C为30，Gamma g为0.08，J为0.1时，取得最佳预测结果，详见第3节结果与讨论部分。使用了SVM_Light[18]工具。

假设有多条蛋白质序列$x_i\in R^d(i=1,2,\cdots,l)$，对应的目标值$y_i\in\{\langle targetvalue\rangle\}$，SVM对氨基酸序列$X$学习。这里，目标值设为+1和−1。+1代表是跨膜β桶状蛋白(正样本)，−1代表非跨膜β桶状蛋白(负样本)。

2.3 基于氨基酸组成位置和物理化学特性的特征

首先，基于氨基酸组分(AAC)的特征提取算法最早是由Nakashima和Nishikawa[19]提出来的，将蛋白质序列表示为$X=[f_1,f_2,\cdots,f_i,\cdots,f_{20}]^{\mathrm{T}}$，其中$f_i$是第$i(i=1,2,\cdots,20)$种氨基酸在蛋白质序列中出现的频率。$f_i=\frac{n_i}{L}$，其中$n_i$为第$i$种氨基酸在序列中出现的次数，$L$为蛋白质序列中氨基酸残基总数。特征向量$X$中元素的顺序按照20种天然氨基酸字母顺序排列。

其次，氨基酸的物理化学特性是影响跨膜蛋白类型的重要因素。由于氨基酸特性的差异，各种不同组合的氨基酸序列具有不同的结构和各自特定的生理功能。计算氨基酸残基指数相关系数如下：

第一步，将β桶状跨膜蛋白氨基酸序列映射为相应的数值序列：$h_1,h_2,\cdots,h_i,\cdots,h_L$，其中$h_i$表示序列中第$i$个残基对应的残基指数，$L$为跨膜蛋白序列的长度。

第二步，利用自相关函数计算β桶状跨膜蛋白序列中氨基酸残基之间的顺序相关性：$r_n=\frac{1}{L-n}\sum_{i=1}^{L-n}h_ih_{i+n}(n=1,2,\cdots,m)$，其中$m$为相关系数的阶数，$m<L$。当$m=1$时为第一阶相关系数，反映了序列中所有相邻的两个氨基酸残基之间的相关性；当$m=2$时为第二阶相关系数，反映了序列中所有相邻的3个残基之间的相关性，其他阶数依次类推。

第三步，利用所选的5种氨基酸残基指数计算相应序列相关系数，将跨膜蛋白序列S_k特征提取为一个$5\times m$维的特征向量，表征为：

$$AR_k=[r_{11}^k,r_{12}^k,\cdots,r_{1m}^k,\cdots r_{w1}^k,r_{w2}^k,\cdots,r_{wm}^k]^{\mathrm{T}}$$

其中，$w=5$。

最后，综合上述的氨基酸组分与氨基酸物理化学性质的相关系数，一条β桶状跨膜蛋白序列表征为一个$20+5\times m$维的特征向量，其中前面20维是氨基酸组分，后面$5\times m$维是5种氨基酸残基指数计算所得的相关系数，即特征向量表示为：

$$X=[c_1^k,c_2^k,\cdots,c_{20}^k,r_{11}^k,r_{12}^k,\cdots,r_{1m}^k,\cdots,r_{w1}^k,r_{w2}^k,\cdots,r_{wm}^k]^{\mathrm{T}}$$

其中，$w=5$。考虑到所选β桶状跨膜蛋白序列的长度值，选定m值为30。

3 结果与讨论

3.1 检验标准

通常检验分类模型性能的方法有自检验和交叉检验。交叉检验验证模型的泛化能力，其中Jackknife检验是被认为最合理的交叉检验方法[20,21]，消除了自检验中“记忆影响”，然而Jackknife检验的不足是时间开销太大。在分类研究中常采用该检验方法的变形k-叠交叉检验(k-fold cross-validation)。首先将样本数据随机分为大概相等的k个子集，依次取出一个子集作为测试集，其余$k-1$个子集作为训练集，交

替反复 k 次，将各次的精确度作平均。实际上，当 k 等于样本总数时，k-叠交叉检验就是 Jackknife 检验。我们采用了[22] TP(真阳性)、TN(真阴性)、FP(假阳性)、FN(假阴性)、分类精确度(ACC)、相关系数(MCC)、正样本灵敏度 Sensitivity(+)、正样本特异度 Specificity(+)、负样本灵敏度 Sensitivity(−)、负样本特异度 Specificity(−)对实验结果进行比较。

3.2 结果

在参数选择时，本文尝试了两种不同的参数设置，第一组为 RBF 核函数($C=30, G=0.08, J=0.1$)；第二组为 RBF 核函数(默认参数)。使用 10 叠交叉检验的方式来评价分类预测结构的优劣，实验结果如表 2 所列。

表 2 支持向量机对 β 桶状跨膜蛋白的预测结果(10 叠交叉检验)

评价指标	RBF 核函数 (C=30,G=0.08,J=0.1)	RBF 核函数 (默认参数)
TP	184	12
TN	182	208
FP	26	0
FN	24	196
Specificity(+)	88.24%	100%
Specificity(−)	89.53%	51.49%
Sensitivity(+)	88.26%	5.71%
Sensitivity(−)	87.53%	100%
Accuracy	88.36%	52.86%
MCC	0.772325	—

从表 2 的数据可以看出，同样使用 RBF 核函数，默认参数的预测结果中，正确预测正样本数很少，导致正样本灵敏度只有 5.71%，几乎无法识别出 β 桶状跨膜蛋白，分类精度只有 52.86%，远远低于调整参数后($C=30, G=0.08, J=0.1$)的预测结果；MCC 相关系数也无法计算出，无法体现预测结果和原始数据的相关性。默认参数的预测结果不满意，所以本文采用 RBF 核函数，参数设为 $C=30, G=0.08, J=0.1$。

从灵敏度和特异性来看，正、负样本都比较高，分别达到了 88.26%和 88.24%，87.53%和 89.53%。灵敏度反映了模型不弃真的程度，这说明该模型不管从正样本中识别 β 桶状蛋白还是从负样本中识别非 β 桶状蛋白，其能力都比较强。特异性反映了模型有效性的程度。从预测结构数据来看，正、负样本的特异性都比较理想。

表 3 不同特征输入的 FSVM 和 SVM 在 TMB 数据集上的交叉检验测试结果比较[23]

分类算法	ACC	MCC
SVM(one-vs-one)[23]	83.2	0.664
FSVM(cf)[23]	86.1	0.673
FSVM(aa)[23]	80.3	0.626
FSVM(aa+dispep)[23]	84.6	0.648
SVM(AAC+Hydi+Hydo+Flex+EnTr+Pol)	88.36	0.7723

表 3 结果显示，采用支持向量机的分类算法，基于本文的组合特征的分类方法，其分类精度和 MCC 相关系数均好于只利用氨基酸组分成分特征的分类方法，也优于使用氨基酸和二肽组成特征相结合的分类方法。这是因为其不但考虑了氨基酸在序列中的频数信息，还考虑了氨基酸的物理化学性质的作用，利用了更多的序列信息，而且氨基酸物理化学性质的选择也影响着预测结果。结果说明，使用氨基酸组分融合氨基酸物理化学特性的方法是有效果的。

结束语 采用多特征融合的特征提取方法(氨基酸组分和氨基酸残基指数相关参数)，对跨膜蛋白序列进行特征提取，结合支持向量机分类算法构建了一种新颖的分类模型，该方法能够更加有效地刻画跨膜蛋白序列中所蕴含的特征信息。与已有跨膜蛋白分类预测方法相比较，该方法具有更强的自适应、泛化和推广能力，获得了较好的分类预测精度。另外由于基于 SVM 的方法与其它机器学习方法都不同程度地受训练集大小的影响[24]，而且当今蛋白质序列数据库中某些类别跨膜蛋白序列的数目相对较少，因此所构建的模型的预测成功率有所欠缺，但是随着新发现的跨膜蛋白序列数目的增加，数据集将逐渐地得到完善。因此，β 桶状跨膜蛋白预测的成功率在以后将会有很大的提升空间。

参 考 文 献

[1] Diederichs K, Freigang J, Umhau S, et al. Prediction by a neural network of outer membrane beta-strand protein topology [J]. Protein Sci, 1998, 7(11): 2413-2420

[2] Jacoboni I, Martelli P L, Fariselli P, et al. Prediction of the transmembrane regions of beta-barrel membrane proteins with a neural network-based predictor [J]. Protein Sci, 2001, 10(10): 779-787

[3] Natt N K, Kaur H, Raghava G P S. Prediction of transmembrane regions of beta-barrel proteins using ANN- and SVM-based methods [J]. Proteins, 2004, 56(1): 11-18

[4] Lin C F, Wang S D. Fuzzy support vector machines [J]. IEEE Trans on Neural Networks, 2002, 13(2): 464-471

[5] Daisuke T, Shigeo A. Fuzzy lest squares support vector machines for multiclass problems [J]. Neural Networks, 2003, 16(5): 785-792

[6] Gromiha M M, Suwa M. A simple statistical method for discriminating outer membrane proteins with better accuracy [J]. Bioinformatics, 2005, 21: 961-968

[7] Park K J, Gromiha M M, Horton P, et al. Discrimination of outer membrane proteins using support vector machines [J]. Bioinformatics, 2005, 21(23): 4223-4229

[8] Kawashima S, Ogata H, Kanehisa M. AAindex: amino acid index database [J]. Nucleic Acids Research, 1999, 27(1): 368-369

[9] Eisenberg D, Weiss R M, Terwilliger T C. The hydrophobic moment detects periodicity in protein hydrophobicity [J]. Proc. Natl. Acad. Sci. USA, 1984, 81(1): 140-144

[10] Hopp T P, Woods K R. Prediction of protein antigenic determinants from amino acid sequecces [J]. Proc. Natl. Acad. Sci. USA, 1981, 78(6): 3824-3828

[11] Bhaskaran R, Ponnuswamy P K. Positional flexibilities of amino acid residues in globular proteins [J]. Int. J. Peptide Protein Research, 1988, 32(4): 241-255

[12] Bull H B, Breese K. Surface tension of amino acid solutions: A hydrophobicity scale of the amino acid residues [J]. Archives of biochemistry and biophysics, 1974, 161(2): 665-670

[13] Grantham R. Amino acid difference formula to help explain protein evolution [J]. Science, 1974, 185(4154): 862-864

[14] Vapnik V. The nature of statistical learning theory [M]. New York: Springer-Verlag, 1995

[15] Vapnik V. Statistical learning theory [M]. New York: Wiley, 1998

[16] 顾亚祥,丁世飞. 支持向量机研究进展[J]. 计算机科学,2011,38(2):14-17

[17] Ren Y,Liu H,Xue C,et al. Classification study of skin sensitizers based on support vector machine and linear discriminant analysis[J]. Anal Chim Acta,2006,572(2):272-282

[18] Joachims T. Text categorization with support vector machines: learning with many relevant features[C]// Proceedings of the European conference on machine learning. Berlin:Springer,1998

[19] Nakashima H,Nishikawa K. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies [J]. Journal of Molecular Biology,1994,238(1):54-61

[20] Chou K C, Maggiora G M. Domain structural class prediction [J]. Protein Engineering,1998,11(7):523-538

[21] Chou K C,Liu W M,Maggiora G M,et al. Prediction and classification of domain structural classes [J]. Proteins: Structure, Function and Bioinformatics,1998,31(1):97-103

[22] Baldi P,Brunak S,Chauvin Y,et al. Assessing the accuracy of prediction algorithms for classification: an overview [J]. Bioinformatics,2000,16(5):412-424

[23] 邹凌云,王正志,黄教民. 基于模糊支持向量机的膜蛋白折叠类型预测[J]. 生命科学研究,2007,11(4):306-310

[24] 李亚飞,吕强,苏伟峰,等. 一种小规模数据集下的贝叶斯网络学习方法及其应用[J]. 计算机科学,2011,38(7):181-184,234

(上接第 227 页)

系数比较大,因此其抗攻击性比 Regular 网络强。这与 3.5 节论述的平均度与紧致系数越大,则网络越鲁棒;同一平均度下网络紧致系数越大,则网络越鲁棒相符。

4.4 网络紧致系数、平均度与网络抗攻击性的关系

根据 3.1 节—3.5 节的理论性讨论和 4.2 节、4.3 节的仿真实验验证可知,针对不同类型网络,若平均度、网络紧致系数均比较大,则采用 ID、IB、IC 3 种攻击策略攻击网络均会花费比较高的攻击代价,且攻击效果较差;反之,则会有攻击策略以较低的攻击代价取得较好的攻击效果。其现实意义为:针对"安全系数低"的网络进行攻击,会达到事半功倍的效果。

本文在 3.2 节定性分析了网络紧致系数、平均度对网络抗攻击性的影响。图 1—图 4 表明,代价下网络面对选择性攻击时,网络性能迅速下降,且其攻击代价会迅速上升,这与现实网络遭受选择性攻击所表现的现象更为相符。图 5—图 8 和表 3 表明,代价下平均度、网络紧致系数越大,则网络越鲁棒;同一平均度下网络紧致系数越大,则网络越鲁棒。

综上所述,不同类型的网络,其平均度、网络紧致系数越大,则网络越鲁棒;同一平均度下,网络紧致系数越大,则网络越鲁棒。

结束语 本文在考虑攻击代价下,研究了主流的 3 类复杂网络模型——无标度网络、Regular 网络、ER 网络的抗攻击性。网络在面对多种攻击策略时,其抗攻击性可能在一种攻击策略下比较强,但是在其他攻击策略下却很弱。为了正确衡量网络整体抗攻击性,提出了度量网络鲁棒性的重要指标——网络紧致系数、平均度。进一步,为了实证网络紧致系数、平均度与网络抗攻击性间的关系,定性分析了前者与后者之间的相关关系,并通过实验实证了前者与后者间的紧密关系,进而得出结论:不同类型网络,平均度与紧致系数越大,则网络越鲁棒;同一平均度下网络紧致系数越大,则网络越鲁棒。该结论对如何提高网络的抗攻击性和防止黑客攻击有重要指导意义。

本文分别就 ID、IB、IC 3 种选择性攻击下的复杂网络抗攻击性进行了研究。然而,复杂网络面对的攻击策略众多,如 ID、IB、IC 任意两种攻击策略交替攻击复杂网络,或者 3 种攻击策略交替攻击复杂网络等,对以上攻击策略下复杂网络的抗攻击性仍需要进一步研究。我们将进一步从网络紧致系数、平均度出发,研究边攻击下网络的抗攻击性、边攻击下攻击策略的攻击强度等系列课题。

参 考 文 献

[1] Criado R,Garcia del Amo A. New results on computable efficiency and it's stability for complex networks[J]. Journal of Computational and Applied Mathematics,2006,192(1):59-74

[2] Wasserman S, Faust K. Social network analysis: methods and applications [D]. Cambridge University Press,1994

[3] Vazquez A, Pastor-Satorras R, Vespignani A. Large-scale topological and dynamical properties of the internet[J]. Phys. Rev. E,2002,65(6):066130

[4] Adamic L A, Huberman B A. Power-law distribution of the world wide web [J]. Science,2000,287(5461):2115

[5] Sporns O. Network analysis,complexity,and brain function[J]. Complexity,2002,8(1):56-60

[6] Lew I,Sexton,Thomas R. Network DEA: Efficiency analysis of organizations with complex internal structure[J]. Computersand Operations Research,2004,31(9):1365-1380

[7] Xia Yong-xiang. Attack Vulnerability of Complex Communication Networks[J]. IEEE Circuits and Systems,2008,55(1):65-69

[8] Schneider C M. The Robustness of Complex Networks[D]. 2011

[9] Matthew J F,Shweta B,Lauren A M. Network frailty and the geometry of herd immunity [J]. Proc Biol Sci,2006,273(1602):2743-2748

[10] Zheng Bo-jin,Huang Dan. Some scale-free networks could be robust under selective node attacks[J]. EPL,2011,94:28010-28015

[11] Wang Xiao-fan, Chen Guan-rong. Complex networks: small-world, scale-free and beyond[J]. IEEE Circuits and Systems Magazine,2003,3(1):6-20

[12] Wang Li,Yan Pei-zhou,Li Ying-hong,et al. Signal sub-control-area division of traffic complex network based on nodes importance assessment[C]// 2011 30th Control Conference. China,2011:5606-5609

[13] Holme P, Kim B J. Attack vulnerability of complex networks [J]. Phys. Rev. E,2002,65(5):1-14

[14] Repperger, Daniel W. New Results in Understanding Performance and Vulnerability in Complex Networks[C]// CIRA International Symposium. Jacksonville,Florida,U. S. A,2007