doi:10.3969/j.issn.1672-5565.2013.04.10

利用伪氨基酸组分和支持向量机预测抗冻蛋白

许 嘉
(内蒙古科技大学 分析测试中心,内蒙古 包头 014010)

摘 要:抗冻蛋白是一类具有提高生物抗冻能力的蛋白质。抗冻蛋白能够特异性的与冰晶相结合,进而阻止体液内冰核的形成与生长。因此,对抗冻蛋白的生物信息学研究对生物工程发展,提高作物抗冻性有重要的推动作用。本文采用由400条抗冻蛋白序列和400条非抗冻蛋白序列构成数据集,以伪氨基酸组分为特征,利用支持向量机分类算法预测抗冻蛋白,对训练集预测精度达到91.3%,对测试集预测精度达到78.8%。该结果证明伪氨基酸组分能够很好的反映抗冻蛋白特性,并能够用于预测抗冻蛋白。

关键词:抗冻蛋白;伪氨基酸组分;支持向量机
中图分类号:Q753 **文献标识码**:A **文章编号**:1672-5565(2013)-04-297-03

Predicting antifreeze proteins by using pseudo amino acid composition and support vector machine

XU Jia
(*Analysis and Testing Center*, *Inner Mongolia University of Science and Technology*, *Baotou* 014010, *China*)

Abstract: Antifreeze protein (AFP) is a kind of protein that can improve the antifreeze capability of organisms. They specifically bind to ice crystals to inhibit growth and recrystallization of ice. It is very important for bioengineering and for improving antifreeze capability of crop to accurately identify AFPs. The present study constructed a benchmark dataset including 400 AFPs and 400 non-AFPs. By using pseudo amino acid composition as parameters, support vector machine was applied to perform prediction. We finally achieved overall accuracies of 91.3% and 78.8%, respectively for training set and test set. These results suggest that pseudo amino acid composition can describe the characteristics of AFPs and can be used for AFPs prediction.

Keywords: Antifreeze protein; Pseudo amino acid composition; Support vector machine

抗冻蛋白(Antifreeze protein, AFP)是一类能够特异性结合冰晶、提高生物抗冻能力的蛋白质[1]。这类蛋白最初是在南北极的海洋鱼类血清中发现,近年来,在昆虫、真菌、细菌和某些植物体内也均发现存在抗冻蛋白。这类蛋白通过与冰晶的特异性相互作用,阻止生物体内冰核的形成与生长,维持生物体内的溶液状态。因此,对抗冻蛋白的理论研究有助于揭示抗冻蛋白的活性和抗冻机理。

正确判断一条新测序的蛋白质是否为抗冻蛋白对于生物工程发展、作物的改造十分重要。然而,利用实验手段来判断是否是抗冻蛋白不但费时,而且会消耗很多资源。随着大量生物基因组测序的完成,海量基因组、蛋白质组、转录组数据的产生,利用机器学习算法来预测蛋白质的类型和功能不仅节约了实验成本,而且能够大大提高实验效率。后基因组时代为我们提供了大量蛋白质序列和注释信息,同时为理论预测抗冻蛋白提供了可能性[2]。

目前,已有一些判别方法用于抗冻蛋白的预测[3-4],且取得了一定的结果。然而,仍缺乏对抗冻蛋白有效的描述。本文利用伪氨基酸组分来描述抗冻蛋白序列,并利用支持向量机来对抗冻蛋白进行预测。

收稿日期:2012-11-15;**修回日期**:2012-12-18.
基金项目:内蒙古科技大学青年创新基金(2011NCL048)资助。
作者简介:许 嘉,女,硕士,讲师,研究方向:界面化学;E-mail:xujia2006@imust.cn.

1 数据库

抗冻蛋白原始数据从 http://www3.ntu.edu.sg/home/EPNSugan/index_files/AFP-Pred.htm[3] 下载。该数据集包含了 481 条抗冻蛋白序列和 9 193 条非抗冻蛋白序列,这些数据的序列一致性低于 40%。如果正负数据集的数目偏差过大,会导致错误的评估预测模型。因此,为了平衡正负集数据,分别选取 400 条抗冻蛋白和 400 条非抗冻蛋白作为基准数据集,并进一步将正负数据集随机分为训练集和测试。这两集合分别包含 200 条抗冻蛋白和 200 条非抗冻蛋白。

2 预测算法

2.1 特征提取

伪氨基酸组分(PseAAC)[5] 是 Chou 教授提出的一种能够很好地表征蛋白质序列的信息参数。它不但能够描述蛋白质序列的氨基酸组成,而且能够描述蛋白质氨基酸序列的物理化学性质的关联。下面对伪氨基酸组分进行描述。

如果将一个氨基酸残基数为 L 的蛋白质 X 表示成,$R_1\ R_2\ R_3\cdots R_L$那么,这条蛋白质序列就可以表示成由 $20+\lambda$ 个离散数值定义的一个 $20+\lambda$ 维向量,定义形式如下:

$$X=[x_1\cdots x_20\ x_20+1\cdots x_{20+\lambda}]^T \tag{1}$$

这里

$$x_u=\begin{cases}\dfrac{f_u}{\sum_{i=1}^{20}f_i+\omega\sum_{j=1}^{\lambda}\theta_j}, & (1\leqslant u\leqslant 20)\\ \dfrac{\omega\theta_{u-20}}{\sum_{i=1}^{20}f_i+\omega\sum_{j=1}^{\lambda}\theta_j}, & (21\leqslant u\leqslant 20+\lambda)\end{cases} \tag{2}$$

其中,f_i表示 20 种不同氨基酸残基在蛋白质 X 中出现的频率。ω 是蛋白质序列关联的权重因子。通常,权重因子的选择范围定在 $\omega=0.05$ 到 0.7 之间,这里我们选取 $\omega=0.05$。θ_j是 j 阶序列相关系数:

$$\theta_j=\frac{1}{L-j}\sum_{i=1}^{L-j}\Theta(R_i,R_{i+j}),\quad (j<L) \tag{3}$$

公式(3)中相关性函数 $\Theta(R_i,R_{i+j})$是可以由以下公式得出:

$$\Theta(R_i,R_{i+j})=\frac{1}{k}\sum_{l=1}^{K}[H_l(R_{i+j})-H_l(R_i)]^2 \tag{4}$$

其中,k 是因子个数,$H_l(R_i)$是第 i 个氨基酸残基所具有的任一种物理化学特征。这些物理化学特征主要包括亲水性,疏水性,侧链聚集度,a-COOH 基的 PK 值,α-NH_3^+基的 PK 值,温度为 25 ℃时的 pI 值。这些物化性质的值需经过标准化处理,公式如下:

$$H_l(R_i)=\frac{H_l^0(i)-\sum_{i=1}^{20}(H_l^0(i)/20)}{\sqrt{\dfrac{\sum_{i=1}^{20}[H_l^0(i)-\sum_{i=1}^{20}(H_l^0(i)/20)]^2}{20}}} \tag{5}$$

这里 $H_l^0(i)$是第 i 个氨基酸残基物理化学特征值的原始值,可从网站 http://chou.med.harvard.edu/bioinf/PseAAC/获得。

2.2 支持向量机

支持向量机是一种优秀的机器学习方法,并已广泛运用于生物信息学的领域,比如:转录起始点和蛋白质亚细胞定位等多个方面。其优点在于能够同时最小化经验误差与最大化几何边缘区,因此支持向量机也被称为最大边缘区分类器。其基本思想是将向量映射到一个更高维的空间里,使得不同类型的向量在高维空间中线性可分。对于待分类样本,其判别函数具有如下形式:

$$f(x)=\mathrm{sgn}(\sum_{i=1}^{k}\alpha_i\ y_i\ k(x,x_i)+b) \tag{6}$$

其中,$k(x,x_i)$称为核函数,通过选取不同的核函数可以得到不同的支持向量机,常用的核函数有以下几种形式:

多项式核函数:$K(x_i,x_j)=(x_i\cdot x_j+1)^d$ (7)

径向基核函数:$K(x_i,x_j)=\exp(-\gamma\parallel x_i-x_j\parallel^2)$ (8)

Sigmoid 核函数:$K(x_i,x_j)=\tanh(b(x_i\cdot x_j)+c)$ (9)

(6)~(8)式中,d、γ、b 和 c 分别为三种核函数的可调参数。本文采用由 Chang 和 Lin 开发的 LIBSVM 软件包[6],选取径向基函数(RBF)作为支持向量机的核函数,调整误差惩罚参数 C 及核函数参数 γ,可得到最佳预测模型。这里使用 LIBSVM 中的 grid-search 程序来优化参数 C 和 γ。

2.3 精度估计

利用敏感性(Sensitivity,Sn)、特异性(Specificity,Sp)和总体准确率(Overall accuracy,OA)为评价指标测试模型的预测性能,其定义如下:

敏感性(Sn):$Sn=TP/(TP+FN)$ (10)

特异性(Sp):$Sp=TN/(TN+FP)$ (11)

总体准确率(OA):$OA=(TP+TN)/(TN+TP+FP+FN)$ (12)

其中,TP、TN、FP 和 FN 分别为正确预测抗冻蛋白数目,正确预测的非抗冻蛋白,非抗冻蛋白预测成为抗冻蛋白的数目和抗冻蛋白预测成非抗冻蛋白的数目。

3 结果与讨论

以伪氨基酸组分为特征,利用支持向量机进行分

类。利用 grid 方法对训练集进行参数寻优,建立最优模型。发现当 $C=32\ 768$ 且 $\gamma=0.001\ 953\ 125$时,模型的预测精度最高,对训练集预测精度达到 91.3%。为检验模型的推广能力,我们利用构建好的模型对 400 条测试序列进行预测,结果表明有 78.8%的蛋白质被预测成功,其中 75.1%的抗冻蛋白和 83.6%的非抗冻蛋白能够被正确预测。该结果证明伪氨基酸组分可用于抗冻蛋白的预测。

AFP-Pred 是第一款用于抗冻蛋白预测的软件[3],其构建基于 300 条抗冻蛋白和 300 条抗冻蛋白。通过使用随机森林算法对抗冻蛋白进行预测,对训练集的预测精度达到 81.3%,对测试集的预测精度达到 83.4%。最近,Zhao Xiaowei 等开发了 AFP_PSSM 来预测抗冻蛋白[4],对训练集的预测精度为 82.7%,对测试集的预测精度达到 93.0%。

尽管已有对测试集的预测精度高于本研究结果,但对于训练集,本研究结果仍具备优势。此外,这些方法大多使用了蛋白质序列的进化信息和预测的二级结构信息,这些信息的获得和提取比本研究使用的伪氨基酸组分要更加复杂。特别是当查询的数据库中没有待查询序列的同源序列时,进化信息将不可用;当二级结构预测软件错误的预测了蛋白质结构时,那么提取的二级结构信息也不可信。因此,只从蛋白质一级序列出发来预测抗冻蛋白,能够避免以上问题的出现。

尽管目前的研究结果还不十分令人满意,但随着蛋白质序列数据库的不断充实,将考虑更多的信息,如寡肽频率、氨基酸约化等信息,以期提高分类模型的预测准确率。

参考文献(References)

[1] Carvajal-Rondanelli PA, Marshall SH, Guzman F. Antifreeze glycoprotein agents: structural requirements for activity[J]. Journal Science Food Agricuture, 2011, 91(14):2507-2510.

[2] Garner J, Harding MM. Design and synthesis of antifreeze glycoproteins and mimics[J]. Chembiochem, 2010, 11(18):2489-2498.

[3] Kandaswamy KK, Chou KC, Martinetz T, Möller S, Suganthan PN, Sridharan S, Pugalenthi G. AFP-Pred: A random forest approach for predicting antifreeze proteins from sequence-derived properties[J]. Journal of Theoretical Biology, 2011, 270(1):56-62.

[4] Zhao Xiaowei, Ma Zhiqiang, Yin Minghao. Using support vector machine and evolutionary profiles to predict antifreeze protein sequences[J]. Interntional Journal of Molecular Science, 2012, 13(2):2196-2207.

[5] Chou KC. Prediction of protein cellular attributes using pseudo-amino acid composition[J]. Proteins, 2001, 43(3): 246-255.

[6] Fan RE, Chen PH, Lin CJ. Working set selection using the second order information for training SVM[J]. Journal of Multivariate Analysis, 2005, 6: 1889-1918.