文章编号: 1000-5013(2009)02-0161-05

采用 BP 算法的多层感知机模型的蛋白识别

张光亚, 葛慧华, 方柏山

(华侨大学 工业生物技术研究所, 福建 泉州 362021)

摘要: 采用误差反传(BP)算法的多层感知机模型,对嗜热蛋白和常温蛋白进行模式识别. 通过增加训练数据及多种检验方法检验模型稳定性及泛化能力,探讨蛋白分子大小对识别效果影响. 结果表明,当动态参数为0.2,学习速率为0.5,隐含层节点数为11时,该模型在自一致性检验、交叉验证和独立样本测试3种检验方法中的识别精度分别为91.5%,88.2%和92.1%,其表现优于一些常见的模式识别算法,且具有良好的稳定性及泛化能力. 此外,对于较大的或者中等大小蛋白质分子,其识别的精度都较高;而对于较小的蛋白分子,其识别效果较差.

关键词: BP算法; 多层感知机; 模式识别; 蛋白质; 热稳定性

中图分类号: Q 811.212 文献标识码: A

大部分蛋白在高温下易失活,这成为拓展其极端条件下工业领域应用的瓶颈. 如何提高酶蛋白的热稳定性,一直是分子生物学、生物工程学和化学工业等领域所关注的课题之一[1]. 有研究表明,蛋白质氨基酸组成与其热稳定性密切相关[2-3]. Kumar 等[2]认为,随着最适温度的升高,蛋白质中 Ile, Tyr, Lys 和 Glu 含量增高,而 Gln 和 Cys 含量降低. Thompson 等[3]发现,嗜热蛋白中含有更多的 Glu, Val, Arg 和 Gly,而含有更少的 Gln, Ser, Asp 和 Lys. 然而,从蛋白质一级结构出发,利用这些研究结果对嗜热和常温蛋白进行识别的研究较少[4]. 如果蛋白质的热稳定性能通过其一级结构进行识别,就可以设计一种基于计算机的筛选方法. 借助计算机的高速运算能力,这种方法能筛选的序列数量可达 10^{80},能显著提高寻找热稳定性很好的蛋白的几率,并预测未知来源蛋白序列(如来源于宏基因组的蛋白)的热稳定性. 由此可见,通过一级结构的信息预测蛋白质的热稳定性具有重要意义. 本文在前期研究基础上[5],探讨基于误差反传(BP)算法的多层感知机模型在嗜热蛋白和常温蛋白识别中的应用.

1 材料与方法

1.1 数据来源

训练数据来源于9种常温微生物和15种嗜热微生物,如表1所示. 表1中,n_S 为蛋白质序列数,n_{CI} 为正确识别数,η_{ACC} 为精确率,序列来源于一个非冗余专家库(Swiss-Prot). 为了进一步减少信息冗余,剔除了所有注释为推测的(Putative)、可能的(Probable)、假设的(Hypothetical)、部分的(Partial)和片断的(Fragment)蛋白质序列,最后得到3 521条嗜热蛋白序列和4 895条常温蛋白序列,共计8 416组.

1.2 结合 BP 算法的多层感知机

多层感知机(MLP)模型结合误差反传(BP)算法训练神经网络,是目前神经网络最成功的应用之一[6]. MLP 是前向神经网络,通常有一个输入层、一个(或多个)隐含层和一个输出层. 网络中隐含层和输出层神经元的传递函数采用 Sigmoid 型函数 $f(x)=[1+e^{x}]^{-1}$. 神经网络为3层结构,采用误差方向传播,其输出层权的调整值 $\Delta W_{j,k}^{0}=\eta\delta_{ky}^{0}y_j$,隐含层权的调整值 $\Delta W_{i,j}=\eta\delta_j x_i$. 其中,$\delta$ 为误差信号,η 为学习速率,$\boldsymbol{x}$ 为输入向量,$\boldsymbol{y}$ 为隐层向量.

收稿日期: 2008-04-17

通信作者: 张光亚(1975-),男,副教授,主要从事酶工程与生物信息学的研究. E-mail: zhgyghh@hqu.edu.cn.

基金项目: 国务院侨办科研基金资助项目(05Q0018)

表1 训练数据的来源

Tab.1 Sources of the training datasets

种类	蛋白名称	n_S	n_{CI}	η_{ACC}/%
嗜热蛋白	*Aeropyrum pernix*	235	224	95.32
	Archaeoglobus fulgidus	245	229	93.47
	M. thermoautotrophicum	46	35	76.09
	Methanococcus jannaschii	354	335	94.63
	Methanopyrus kandleri	331	317	95.77
	Pyrobaculum aerophilum	201	191	95.02
	Pyrococcus abyssi	298	288	96.64
	Pyrococcus furiosus	266	259	97.37
	Sulfolobus acidocaldarius	90	80	88.89
	Sulfolobus solfataricus	287	255	88.85
	Sulfolobus tokodaii	233	213	91.42
	Thermoplasma acidophilum	216	169	78.24
	Thermoplasma volcanium	182	149	81.87
	Thermotoga maritima	367	330	89.92
	Thermus thermophilus	170	139	81.76
常温蛋白	*Bacillus halodurans*	511	433	84.74
	Chlamydia trachomatis	343	314	91.55
	Deinococcus radiodurans	367	334	91.01
	Lactococcus lactis	585	527	90.09
	Mycoplasma genitalium	241	227	94.19
	Rickettsia prowazekii	346	290	83.82
	Shigella flexneri	1 157	1 086	93.86
	Synechocystis sp.	640	599	93.59
	Yersinia pestis	705	679	96.31
	总数	8 416	7 702	91.52

1.3 有效性检验

模型的稳定性及泛化能力,采用以下3种方法进行检验.

(1) 自一致性检验. 用于训练的3 521条嗜热蛋白和4 895条常温蛋白的数据,同时也被用来进行预测并判断是否为嗜热蛋白或常温蛋白.

(2) 交叉验证. 采用5倍交叉验证[7],将训练的3 521条嗜热蛋白和4 895条常温蛋白随机分为5组(每组约包含704个嗜热蛋白和979常温蛋白). 然后,采用"留一法"进行验证,每次留出1组作为测试数据,另外4组作为训练数据,轮流进行5次,使得每组数据都能作为测试数据进行预测.

(3) 独立测试. 为了进一步验证模型的稳定性,另外采用859组数据进行预测,而这些数据在上述8 416组数据中从未出现.

这些训练的数据来源于两部分. 其中,一部分来源于嗜热微生物 *Aquifex aeolicus*(最适生长温度为95 ℃)[1]和常温微生物 *Xylella fastidiosa*(最适生长温度为26 ℃)[8]. 同样剔除了带有上述注释的蛋白序列,最终分别得到382条嗜热蛋白和325条常温蛋白. 另一部分包含76对嗜热蛋白和常温蛋白,这152组数据来源于文[9].

1.4 识别效果评估

各模型的最终表现通过敏感性(K_{SE})、特异性(K_{SP})、准确率(η_{ACC})和Matthew相关系数(x_{MCC})表达,则有

$$K_{SE} = \frac{TP}{TP + FN}, \tag{1}$$

$$K_{SP} = \frac{TN}{TN + FP}, \tag{2}$$

$$\eta_{ACC} = \frac{TP + TN}{TP + FP + TN + FN}, \tag{3}$$

$$x_{MCC} = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FN) \cdot (TN + FP) \cdot (TP + FP) \cdot (TN + FN)}}. \tag{4}$$

上式中, TP 为真阳性, 指嗜热蛋白被预测为嗜热蛋白; FN 为假阴性, 指嗜热蛋白被预测为常温蛋白; TN 为真阴性, 指常温蛋白被预测为常温蛋白; FP 为假阳性, 指常温蛋白被预测为嗜热蛋白.

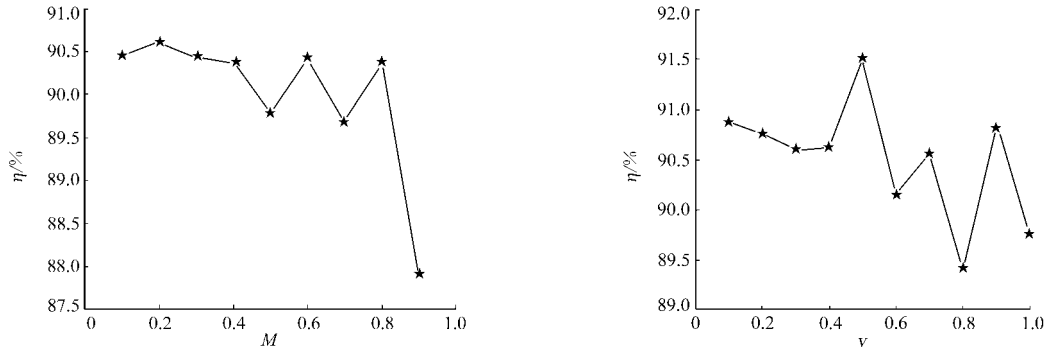
文中实现所有算法的软件均来自于 Weka(http://www.cs.waikato.ac.nz/ml/weka/), 该程序包是基于 JAVA 虚拟机开发的[10]; 使用的计算机为 Pentium V 2.7 GHz, 512 MB RAM.

2　结果与分析

2.1　运行参数的优化

按照软件的默认值, 隐含层的数量为 11, 故 MLP 神经网络模型的拓扑结构为"20-11-1"; 在层过程中, 动态参数和学习速率的选择对网络的识别效果有较大影响. 因此, 对这两个参数进行优化.

动态参数(M)对识别效果(η)的影响, 如图 1(a)所示, 此时, 学习速率(v)的值固定为 0.3. 由图 1(a)可知, 选择不同的动态参数值对识别效果有较大影响, 且当动态参数为 0.2 时, 其识别精度达到最佳(90.6%). 学习速率对识别效果的影响, 如图 1(b)所示. 不同的学习速率值对识别精度也存在较大影响, 当学习速率为 0.5 时, 其精度最佳(91.5%). 因此, 最终选择动态参数为 0.2, 学习速率为 0.5.



(a) 动态参数　　(b) 学习速率

图 1　学习速率和动态参数的变化

Fig. 1　Optimization of learning rate and momentum

2.2　基于 BP 算法的多层感知机模型的识别

基于 BP 算法的多层感知机模型识别效果, 如图 2 所示. 图 2 中, iv为自一致性检验, ㈤为 5 倍交叉验证, ㈣为独立测试. 在自一致性检验中, 所有训练数据均用作测试. 结果表明, 该模型整体识别精度达到了 91.5%. 说明经过足够的训练, 该分类器已掌握氨基酸组成和蛋白质热稳定性之间的复杂关系, 并取得较好的识别效果. 同时说明, 尽管采取了一些措施减少信息冗余, 但仍有部分噪音存在于样本中, 并影响了识别效果. 该模型对各微生物蛋白质组识别的效果, 如表 1 所示. 由表 1 可知, 不同蛋白质组的识别效果从 76.1% 到 96.4% 不等, 存在一定的差异.

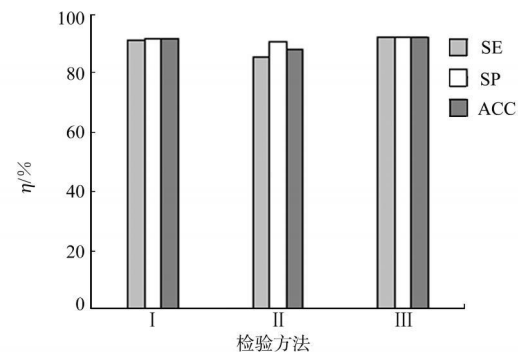


图 2　MLP 模型的识别效果

Fig. 2　Performance of the MLP model

交叉验证的结果表明, 该集成分类器分别正确识别出 3 006 个嗜热蛋白和 4 418 个常温蛋白, 整体识别精度为 88.2%. 交叉验证精度的下降, 可能是由于每次参与训练的数据减少(约 1 683 个), 导致模型训练不充分. 敏感性为 85.4%, 说明该 MLP 模型能识别约 85.4% 的嗜热蛋白; 特异性为 90.3%, 说明它能识别约 90.3% 的常温蛋白. 这意味着不借助蛋白质的结构信息而仅依赖其序列信息, 就可以达到一个较高的识

别精度.

对另外一组独立数据进行测试,进一步验证该集成分类器的实用性及稳定性.在此过程中,8 416组训练数据均参与训练,然后对859组测试数据进行预测.结果表明,该分类器从458个嗜热蛋白中成功识别出422个,从401个常温蛋白中成功识别出369个,正确率分别为92.1%和92.0%,总体识别精度为92.1%.相比交叉验证,其精度提高了3.9%,这可能是由于训练数据增加的缘故.

2.3 与其他模式识别算法的对比

基于BP算法的MLP模型与*K*-近邻(*K*-NN)、径向基神经网络(RBF-NN)及简单贝叶斯(Naive Bayes)3种模型识别效果,如表2所示.在自一致性检验中,*K*-NN模型的识别精度为100%,比MLP模型的精度提高了8.5%,说明*K*-NN算法对训练样本中噪音的抗干扰能力优于MLP模型.但是,*K*-NN模型在交叉验证及独立样本测试中效果,略逊于MLP模型.

表2 MLP模型与其他算法的比较

Tab.2 Comparison between the MLP model and other algorithms

检验方法	算法	K_{SE}	K_{SP}	η_{ACC}	x_{MCC}
自一致性检验	MLP	91.3	91.7	91.5	0.83
	K-NN	100.0	100.0	100.0	1.00
	RBF-NN	84.0	86.7	85.6	0.70
	Naive Bayes	85.6	84.7	85.1	0.70
交叉检验	MLP	85.4	90.3	88.2	0.76
	K-NN	86.5	88.6	87.7	0.75
	RBF-NN	84.0	86.4	85.4	0.70
	Naive Bayes	85.3	84.8	85.0	0.70
独立样本测试	MLP	92.1	92.0	92.1	0.84
	K-NN	89.1	90.3	89.6	0.79
	RBF NN	88.4	92.3	90.2	0.81
	Naive Bayes	89.5	90.5	90.0	0.80

RBF模型作为另一种常见的神经网络模型,在3种检验方法中,其精度分别为85.6%,85.4%和90.2%,表现均劣于MLP模型;尤其是在自一致性检验中,其效果明显较差,识别精度分别比MLP模型低5.9%.简单贝叶斯模型在3种检验方法中表现也逊于MLP模型.但是,MLP模型也存在一定的缺陷.在运算过程中,MLP神经网络对计算机资源的消耗较大,在几种算法中运算时间最久,其运算过程约13 min,而简单贝叶斯模型仅需1 min左右的时间.因此,提高计算机硬件可以在一定程度上缓解该矛盾.

2.4 蛋白分子大小对识别效果的影响

由于此方法是完全基于蛋白质的序列信息,因此,识别效果与样本中蛋白质分子的大小可能存在某种联系.将用于训练的8 416个样本及用于预测的859个样本按照分子大小分为4类,探讨蛋白质分子大小对识别效果的影响,结果如表3所示.表3中,*N*为蛋白分子总数,蛋白分子大小用氨基酸数量表征.

表3 蛋白分子大小对识别效果的影响

Tab.3 Influence of protein size on prediction accuracy

检验方法	分子尺寸	*N*	n_{CI}	η_{ACC}/%
自一致性检验	$L \geqslant 800$	344	336	97.7
	$500 \leqslant L < 800$	922	884	95.9
	$200 \leqslant L < 500$	4 653	4 299	92.4
	$L < 200$	2 497	2 183	87.4
独立样本测试	$L \geqslant 800$	29	29	100.0
	$500 \leqslant L < 800$	112	107	95.5
	$200 \leqslant L < 500$	551	516	93.6
	$L < 200$	167	140	83.8

对于较大的蛋白质分子($\geqslant$800个氨基酸),其识别的精度都很高,分别为97.7%和100%;对于氨基酸数量在500到800之间的蛋白质分子,该方法从922,112个蛋白分子中分别正确识别出了884,107个,正确率分别为95.9%和95.5%;对于中等大小(氨基酸数量在200到500之间)的蛋白分子而言,其识别精度也较令人满意,分别达到了92.4%和93.6%;而对较小(少于200个氨基酸)的蛋白分子,其识别效果较差,分别为87.4%和83.8%,比各自平均正确率分别低4.1%和8.3%.

对于MLP模型,在自一致性检验中也出现错误识别.如对于*M. thermoautotrophicum*蛋白质组产生错误识别的11个蛋白中,有4个蛋白(约占36.4%)的氨基酸数量少于200个,属于较小蛋白分子;而对于*T. thermophilus*蛋白质组产生错误识别的31个蛋白中,有12个蛋白(约占38.7%)属于较小蛋白分子;对于*B. halodurans*蛋白质组产生错误识别的78个蛋白中,有38个蛋白(约占48.7%)属于较小蛋白分子;而对于*R. prowazekii*蛋白质组产生错误识别的56个蛋白中,竟有29个蛋白(约占51.8%)也属于较小蛋白质分子.从信息学的角度来理解,可认为较小蛋白分子所包含的信息量较少,从

中提取的 20 个特征向量(指 20 种氨基酸组成)不足以反映其特性.

3　结束语

通过自一致性检验、交叉验证和独立样本测试 3 种验证方法, MLP 模型的准确率分别为 91.5%, 88.2%和 92.1%. 除在自一致性检验中其效果逊于 *K*-NN 模型外, MLP 模型均优于 3 种算法, 这提供了一个较好识别嗜热蛋白和常温蛋白的系统. 但是, 该系统对较小的蛋白分子识别效果一般, 如何提高对较小蛋白的识别精度是今后研究的重点.

参考文献:

[1] ATOMI H. Recent progress towards the application of hyperthermophiles and their enzymes[J]. Curr Opin in Chem Biol, 2005, 9: 1-8.

[2] KUMAR S, NUSSINOV R. How do thermophilic proteins deal with heat? [J]. Cell Mol Life Sci, 2001, 58: 1216-1233.

[3] THOMPSON M J, EISENBERG D. Transproteomic evidence of a loop-deletion mechanism for enhancing protein thermostability[J]. J Mol Biol, 1999, 290: 595-604.

[4] 丁彦蕊, 蔡宇杰, 须文波. 基于氨基酸组成预测蛋白质热稳定性的 ν-支持向量机方法(英文)[J]. 计算机与应用化学, 2005, 22(6): 51-57.

[5] 张光亚, 刘桂兰, 方柏山. 基于支持向量机识别嗜热和常温蛋白的研究[J]. 计算机与应用化学, 2006, 23(8): 707-710.

[6] 方　宁, 李景治, 贺贵明. 简化的广义多层感知机模型及其学习算法[J]. 计算机工程, 2004, 30: 50-52.

[7] PARK K J, KANEHISA M. Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs[J]. Bioinformatics, 2003, 19: 1656-1663.

[8] DAVID J L, GREGORY A S, DONAL A H. Synonymous codon usage is subject to selection in thermophilic bacteria[J]. Nucleic Acids Res, 2002, 30(19): 4272-4277.

[9] ZHANG Guang-ya, FANG Bai-shan. Discrimination of thermophilic and mesophilic proteins via pattern recognition methods[J]. Process Biochem, 2006, 41: 552-556.

[10] INAMDAR N M, EHRLICH K C, EHRLICH M, et al. Data mining in bioinformatics using Weka[J]. Bioinformatics, 2004, 20: 2479-2481.

Application of a BP Algorithm Based Multi-Layer Perceptron Model to Discriminate Thermophilic and Mesophilic Proteins

ZHANG Guang-ya, GE Hui-hua, FANG Bai-shan

(Institute of Industrial Biotechnology, Huaqiao University, Quanzhou 362021, China)

Abstract: In this paper, a back-propagation (BP) algorithm based multi-layer perceptron model was proposed to discriminate thermophilic and mesophilic proteins. When the momentum parameter, learning rate and the number of the hidden layer nodes were 0.2, 0.5 and 11, respectively, the model had the best performance. The success rate for self-consistency check, cross-validation and independent test with other dataset was 91.5%, 88.2% and 92.1%, respectively. It outperformed other pattern recognition methods such as *K*-nearest neighbors, Naive Bayes and RBF neural network. The model was robust and has good generalization. The influence of protein size on prediction accuracy was also addressed. For big and moderate protein, the prediction accuracy was hjgh, whereas for small protein, it was low.

Keywords: back-propagation algorithm; multi-layer perceptron; pattern recognition; thermostability

(责任编辑: 钱　筠　　英文审校: 陈国华)