PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2018, 45(4): 453~459
www.pibb.ac.cn

基于间隔二肽组分和递归特征消除法的 DNA 结合蛋白的鉴定 *

汤亚东 [1)**] 刘 潇 [1)**] 刘太岗 [2)] 谢 鹭 [3)] 陈兰明 [1)***]

([1)] 农业部水产品贮藏保鲜质量安全风险评估重点实验室(上海)，上海海洋大学食品科学与技术学院，上海 201306；
[2)] 上海海洋大学信息学院，上海 201306；[3)] 上海生物信息技术研究中心，上海 201203)

摘要 DNA 结合蛋白(DNA-binding proteins，DBPs)的鉴定在原核和真核生物的基因和蛋白质功能注释研究中具有十分重要的意义．本研究首次运用间隔二肽组分(gapped-dipeptide composition，GapDPC)结合递归特征消除法(recursive feature elimination，RFE)鉴定 DBPs．首先获得待测蛋白质氨基酸序列的位置特异性得分矩阵(position specific scoring matrix，PSSM)，在此基础上提取蛋白质的 GapDPC 特征，通过 RFE 法选择最优特征，然后利用支持向量机(support vector machine，SVM)作为分类器，在蛋白质序列数据集 PDB396 和 LB1068 中进行夹克刀交叉验证(jackknife cross validation test)．研究结果显示，基于 PDB396 和 LB1068 数据集，DBPs 预测的准确率、Matthews 相关系数、敏感性和特异性分别达到 93.43%、0.86、89.04%和 96.00%，以及 86.33%、0.73、86.49%和 86.18%，明显优于文献报道中的相关方法，为 DBPs 的鉴定提供了新的模型．

关键词 DNA 结合蛋白，间隔二肽组分，位置特异性得分矩阵，递归特征消除法，支持向量机分类器
学科分类号 Q71，Q81 **DOI**: 10.16476/j.pibb.2017.0413

DNA 结合蛋白(DNA-binding proteins，DBPs)是原核和真核生物蛋白质组的重要组分[1-2].虽然 DBPs 在原核和真核生物细胞总蛋白质中的占比较低(分别为 2%～3%和 6%～7%)，但是它们发挥着不可替代的重要功能．例如，特定核苷酸序列的识别、DNA 复制、基因转录和表达调控等[3]．DBPs 可以通过实验方法被鉴定，例如过滤结合分析(filter binding assays)、微阵列染色质免疫沉淀(chromatin immune precipitation on microarrays)，以及 X 射线结晶学(X-ray crystallography)分析等．这些方法预测 DBPs 的准确率高，但是费用较高且耗时较长[4].随着基因组测序技术的突破[5-6]，蛋白质序列增长迅速，实验分析已经不能满足实际所需．因此，建立快速、有效的 DBPs 预测新方法非常必要．

迄今为止，国内外学者已报道一些预测 DBPs 的方法，其中构建蛋白质氨基酸序列的特征模型是关键，即把可变长度的蛋白质序列转换为固定长度的特征向量．例如，Kumar 等[1]建立了 DNAbiner 方法，该方法首次采用 PSSM 结合 SVM 鉴定 DBPs[7-8]．PSSM 矩阵反映了蛋白质序列在进化过程中氨基酸残基发生替换的得分情况，比蛋白质序列本身提供了更多的信息[9-10]．Kumar 等[11]还建立了基于随机森林法(random forest，RF)的 DNA-Prot 方法．Lin 等[12]结合 RF 法和灰色模型(grey model)建立了 DBPs 鉴定的 iDNA-Prot 方法．最近，Liu 等[13-14]先后建立了 DBPs 鉴定的 iDNA-Prot|dis 和 iDNAPro-PseAAC 方法．前者基于氨基酸距离对

* 国家自然科学基金(31671946, 11601324)和上海市科委基金(17050502200)资助项目.
** 并列第一作者.
*** 通讯联系人.
Tel: 021-61900504, E-mail: lmchen@shou.edu.cn
收稿日期：2017-11-07，接受日期：2018-03-09

(amino acid distance-pairs)和缩减的氨基酸字母表(reduced alphabet profile)构建蛋白质序列特征模型；后者基于 PSSM 矩阵的伪氨基酸组分(pseudo amino acid composition，PseAAC)特征，进一步改进了预测准确率(76.56%)．尽管上述方法不断提高了 DBPs 鉴定的准确率，但是其预测性能仍有较大的提升空间．本研究首次运用 GapDPC 特征[15]结合 RFE 法进一步挖掘蕴含在 PSSM 矩阵中的蛋白质序列进化信息和“序”信息，构建了 GapDPC-DBPs 模型，提高了 DBPs 的预测准确率．

1 材料与方法

1.1 数据集

本研究利用两个蛋白质氨基酸序列数据集来检验新模型的预测性能：数据集 PDB(protein data bank，PDB)396 由 Kumar 等[1]和 Stawiski 等[2]构建，含有 146 条 DBPs 序列，以及 250 条非 DBPs 序列．该数据集中的任意两条蛋白质序列的相似度≤25%；数据集 LB1068 由 Liu 等[14]构建，含有 525 条 DBPs 序列,以及 550 条非 DBPs 序列，该数据集中的蛋白质序列长度≥50，且任意两条蛋白质序列的相似度≤25%．通过序列分析，本研究删除了数据集 LB 1068 中 7 条可疑的序列(1AOII、3THWD、4FCYC、4GNXL、4GNXK、4JJNJ、4JJNI)，此 7 条序列分别由“AGCT”排序组成，有可能是核酸序列．因此，采用的基准数据集(LB1068)由 518条 DBPs 和 550 条非 DBPs 序列组成．

1.2 蛋白质序列特征向量的提取

参考 Liu 等[15]的间隔二肽提取方法．利用 PSI-BLAST 软件[16](ftp://www.ncbi.nlm.nih.gov/blast/executables/blast+/2.3.0/ncbi-blast-2.3.0+-x64-linux.tar.gz)获得待测蛋白质序列的 PSSM 矩阵，迭代次数设为 3，E 值设为 0.001，其余参数均为默认值．把基于蛋白质序列二肽组分(dipeptide composition，DPC)概念推广到基于 PSSM 矩阵的情形．推广后的 DPC 定义如下：

$$X=(x_{1,1}, \cdots, x_{1,20}, x_{2,1}, \cdots, x_{2,20}, \cdots, x_{20,20})$$

为了分析氨基酸“序”信息的影响，我们进一步提取了 GapDPC 特征以描述 2 个氨基酸残基之间的相关性．GapDPC 特征定义如下：

$$y_{i,j,g}=\sum_{k=1}^{L-g-1} p_{k,i} \times p_{k+g+1,j}(1\leqslant i, j\leqslant 20),$$

式中：g 代表氨基酸 i 和 j 之间的间隔；L 表示蛋白质序列的长度；k 表示蛋白质序列中氨基酸的下标(序列位置)，从第 1 个氨基酸开始(k=1)．

当 g 等于 0 时，GapDPC 退化为 DPC，即 2 个相邻氨基酸组成的二肽．提取自 PSSM 矩阵的三维矩阵[$y_{i,j,g}$]表示查询序列，当 g= 0, 1, 2,…, G (最大的 g)时，每条蛋白质序列对应一个 400×(G+1)维的特征向量．

1.3 支持向量机与递归特征消除法

采用 LIBSVM 软件包[17]执行 SVM 分类预测．LIBSVM 软件包提供了 4 种基本核函数，选择径向基函数(radial basis function，RBF)预测 DBPs．RBF 能准确地反映样本数据的分布情况，其中的惩罚系数 C 和 RBF 自带参数 γ 通过 grid 方法[20]获得，后者决定了数据映射到新的特征空间后的分布．本研究采用 Guyon 等[18]建立的 RFE 算法进行特征选择，该方法已被证明是用于“降维”的有效方法．

1.4 DBPs 预测性能的评价方法

采用夹克刀交叉验证方法[15]评价模型的预测性能．该方法每次选出数据集中一条蛋白质序列作为测试集，其余序列作为训练集，并依次轮流循环，直至所有序列测试完毕．本研究参考 Liu 等[19]的方法，选择敏感性(sensitivity，Sn)、特异性(specificity，Sp)、预测准确率(accuracy，Acc)以及 Matthews 相关系数(matthews correlation coefficient，MCC)作为预测模型的评价指标,它们分别定义如下：$Sn=TP/(TP+FN)$，$Sp=TN/(TN+FP)$，$Acc=(TP+TN)/(TP+TN+FP+FN)$，$MCC=[(TP\times TN)-(FP\times FN)]/\sqrt{[(TP+FP)(TP+FN)(TN+FP)(TN+FN)]}$，式中：$TP$、$TN$、$FP$、$FN$ 分别表示真阳性、真阴性、假阳性、假阴性的样本数．

此外，本研究采用受试者工作特征曲线(receiver operating characteristic curve，ROC)和 ROC 曲线下方的面积(area under the ROC curve，AUC)[20]进一步检测模型的预测性能．

2 结果与分析

2.1 GapDPC-DPBs 模型参数的选择

为了考察不同的氨基酸间隔对预测性能的影响，并且保证每条蛋白质序列的特征向量的维数不至过高，本研究设置了不同的 g 值(0, 1, 2, 3, 4)，每个 g 值对应 400 维特征向量．基于两个数据集 PDB396 和 LB 1068，DBPs 的预测结果显示，不同 g 值对应的二肽组分对 DBPs 均有一定的预测能力(预测准确率为 79.03%～81.82%)．因此，本文整

合不同 g 值对应的间隔二肽特征以进一步提高 GapDPC-DPBs 模型的预测性能．特征整合后，数据集中的每一条蛋白质序列均用一个 2 000(400×5)维的特征向量表征．利用 RFE 算法对这些特征向量进行排序，并采用夹克刀交叉验证法检验了不同维数(K=10，20，30，…，500)特征向量对预测准确率的影响．鉴于 K 值太大，导致维数太高，造成特征冗余，K 值太小，则出现所选特征不能完全反映与预测属性的相关度．因此，本研究选择 2 000 维特征向量中的前 500 维，即 K 值最大为 500．由图 1 可知，当 K 值为 370 时，数据集 PDB396 的预测准确率最高达到 92.93%．当 K 值为 490 时，数据集 LB1068 的预测准确率最高达到 87.54%．当 K 值为 370 时，数据集 LB1068 也取得了比较满意的结果(86.42%)．为了保证数据集的特征向量维数一致,且不会因为维数太高而导致计算量加大，本研究选取最优的 370 维特征向量表示数据集中的每一条蛋白质序列．

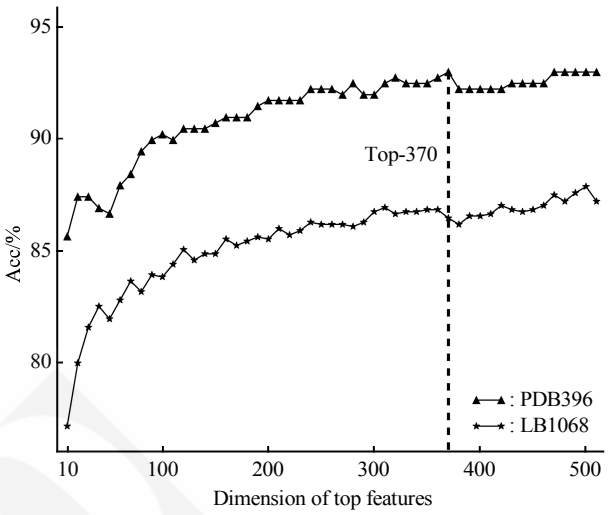

Fig. 1　The effects of feature dimensions on the prediction accuracy

Table 1　The effects of different GapDPC on the accuracy of DBPs prediction

Acc /%	*g*				
	0	1	2	3	4
PDB396	81.82	81.82	81.82	81.82	81.82
LB1068	79.12	79.03	79.31	79.04	79.21

2.2　GapDPC-DPBs 模型的预测结果

比较特征向量维数选择前后的模型，基于数据集 PDB396，研究结果显示，DBPs 预测的准确率从 83.33%提高到了 93.43%，MCC 从 0.64 提高到了 0.86，AUC 值从 0.903 提高到了 0.970(表 2)．基于数据集 LB1068，DBPs 的预测准确率从 79.96%提高到了 86.33%，MCC 从 0.60 提高到了 0.73，AUC 值从 0.865 提高到了 0.916(表 2)．

Table 2　The jackknife cross validation test on the datasets PDB396 and LB1068 before and after the selection of feature dimensions

Datasets	Dimensions	Acc/%	MCC	Sp/%	Sn/%	AUC
PDB396	Before selection	83.33	0.64	88.80	73.97	0.903
	After selection	93.43	0.86	96.00	89.04	0.970
LB1068	Before selection	79.96	0.60	79.09	80.89	0.865
	After selection	86.33	0.73	86.18	86.49	0.916

ROC 曲线以简单、直观的图形反映了 DBPs 预测的敏感性和特异性连续变化的情况(图 2)．AUC 值作为 ROC 曲线下面积，常用于评价模型整体性能，该值越大，表明分类器的预测性能越好．本研究基于数据集 PDB396 和 LB1068，AUC 值在特征选择后比特征选择前分别提高了约 7 和 5 个百分点．

2.3　GapDPC-DPBs 模型的预测性能评价

为了客观地评价本研究建立的 GapDPC-DPBs 模型的预测性能，我们利用相同数据集和夹克刀交叉验证方法，比较了该模型和文献报道中的预测方法．基于数据集 PDB396，尽管已有预测模型的评价指标较少，但是在准确率和 Matthews 相关系数这两个重要指标上，GapDPC-DPBs 比 Niu 等[21]建

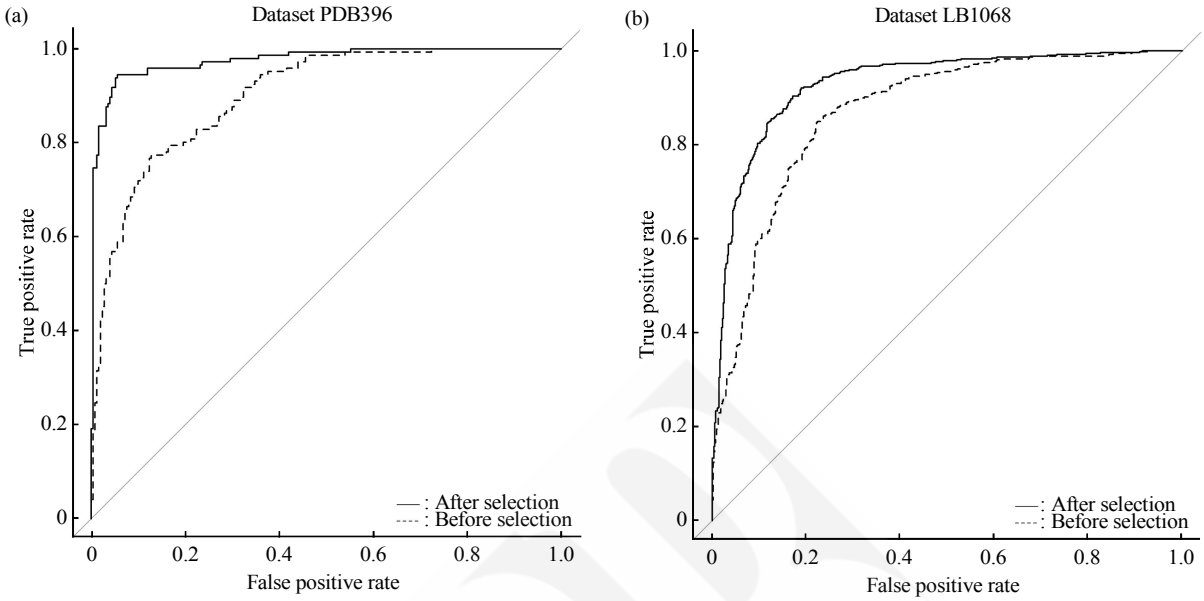

Fig. 2　The ROCs of the datasets PDB396 (a) and LB1068 (b) before and after the selection of feature dimensions

(a), (b) are datasets PDB396 and LB1068, respectively.

立的模型预测的准确率和 Matthews 相关系数分别提高了约 12 和 26 个百分点(表 3). 基于 LB1068 数据集，GapDPC-DPBs 的预测性能明显优于文献报道的 iDNAPro-PseAAC[14]、DNAbinder(dimension 21 和 dimension 400)[9]、DNA-Prot[11]以及 iDNA-Prot[12] 四种方法. 与预测性能最好的 iDNAPro-PseAAC[14] 相比，GapDPC-DPBs 在准确率、特异性、敏感性、Matthews 相关系数和 AUC 值等指标上，分别提升了约 10、9、10、20 和 7 个百分点(表 4).

Table 3　The comparison of different methods based on the dataset PDB396

Methods	Acc/%	MCC	Sp/%	Sn/%	AUC	Reference
DNA-Prot	80.31	–	–	–	–	[11]
Niu *et al*	81.82	0.60	–	–	–	[21]
GapDPC-DPBs	93.43	0.86	96.00	89.04	0.970	This study

– No data provided in the literature.

Table 4　The comparison of different methods based on the dataset LB1068

Methods	Acc/%	MCC	Sp/%	Sn/%	AUC	Reference
iDNAPro-PseAAC	76.56	0.53	77.45	75.62	0.839	[14]
DNAbinder(21)	73.95	0.48	79.09	68.57	0.814	[9]
DNAbiner(400)	73.58	0.47	80.36	66.47	0.815	[9]
DNA-Prot	72.55	0.44	59.76	82.67	0.789	[11]
iDNA-Prot	75.40	0.50	64.73	83.81	0.761	[12]
GapDPC-DPBs	86.33	0.73	86.18	86.49	0.916	This study

3 讨 论

迄今为止，已经有一些机器学习算法(machine learning algorithm，MLA)被用于预测蛋白质的结构和功能分类，例如 SVM[22]、随机森林(random forest)[23]、人工神经网络(artificial neural network，ANN)[24]、最近邻法(nearest neighbor method)[25]、朴素贝叶斯(naïve bayes)法[23]等．其中，由 Vapnik 等[22]建立的 SVM 应用最为广泛，并取得了较高的预测准确率[26]．MLA 的预测性能在一定程度上与特征向量的维数有关．维数越低，包含的信息量越少，不足以识别不同的样本；维数越高，不仅会增加噪声和信息冗余，而且还加大了计算复杂度．因此，在执行分类算法之前，选择合适的特征子集是非常有必要的．本研究利用 RFE 法对 2 000 维特征向量进行排序，并采用夹克刀交叉验证法对前 500 维特征向量对 DBPs 预测准确率分别进行了分析，发现 K 值为 370 时 2 个数据集中 DBPs 的预测准确率最优．

本研究采用 SVM 作为分类器在数据集 PDB396 和 LB1068 上进行夹克刀交叉验证，均取得了较好的预测效果．与其他交叉验证方法相比[例如，独立数据集测试(independent dataset test)、子集测试(sub-sampling test)等]，夹克刀交叉验证虽然耗时较长，但通常被认为最为客观和严格[15]．数据集 PDB396 的预测准确率为 93.43%，敏感性为 92.52%，特异性为 92.52%，Matthews 相关系数为 0.86，AUC 值为 0.970；数据集 LB1068 的预测准确率为 86.33%，敏感性为 86.34%，特异性为 86.34%，Matthews 相关系数为 0.73，AUC 值为 0.916，优于文献中报道的 4 种预测模型．GapDPC-DPBs 对 DBPs 的预测取得了较好的预测结果，主要原因可能在于：a．基于 PSSM 矩阵提取 GapDPC，不仅包含了待测蛋白质序列的组分特征、进化信息和“序”信息，而且在一定程度上反映了序列中 2 个相邻或间隔的氨基酸之间的相关性；b．RFE 法能够根据各个特征向量与预测属性的相关度选择最优的特征子集，既降低了计算复杂度又提高了预测准确率．综上表明，本研究结果发展并补充了 DBPs 的鉴定方法，为原核和真核生物的基因和蛋白质功能注释提供了技术支撑．实用模型的发展方向是提供公共的网络服务器和友好的用户界面，因此，在接下来的研究工作中，我们将致力于为本研究建立的新模型提供共享信息平台．

参 考 文 献

[1] Kumar M, Gromiha M M, Raghava G P, *et al*. Identification of DNA-binding proteins using support vector machines and evolutionary profiles. BMC Bioinformatics, 2007, **8**: 463−472

[2] Stawiski E W, Gregoret L M, Mandel G Y, *et al*. Annotating nucleic acid-binding function based on protein structure. Journal of Molecular Biology, 2003, **326**(4): 1065−1079

[3] Xu R, Zhou J, Liu B, *et al*. Identification of DNA-binding proteins by incorporating evolutionary information into pseudo amino acid composition *via* the top-n-gram approach. Journal of Biomolecular Structure & Dynamics, 2015, **33**(8): 1720−1730

[4] Langlois R E, Lu H. Boosting the prediction and understanding of DNA-binding domains from sequence. Nucleic Acids Res, 2010, **38**(10): 3149−3158

[5] Lin C, Chen W, Qiu C, *et al*. LibD3C: ensemble classifiers with a clustering and dynamic selection strategy. Neurocomputing, 2014, **123**: 424−435

[6] Li P, Guo M, Wang C, *et al*. An overview of SNP interactions in genome-wide association studies. Briefings in Functional Genomics, 2015, **14**(2): 143−155

[7] Nanni L, Lumini A. An ensemble of reduced alphabets with protein encoding based on grouped weight for predicting DNA-binding proteins. Amino Acids, 2009, **36**(2): 167−175

[8] Cai Y, Lin S L. Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequence. Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics, 2003, **1648**(1−2): 127−133

[9] Chen K, Kurgan L A, Ruan J, *et al*. Prediction of protein structural class using novel evolutionary collocation-based sequence representation. Journal of Computational Chemistry, 2008, **29**(10): 1596−1604

[10] Liu T, Geng X, Zheng X, *et al*. Accurate prediction of protein structural class using auto covariance transformation of PSI-BLAST profiles. Amino Acids, 2012, **42**(6): 2243−2249

[11] Kumar K K, Ganesan P, Suganthan P N, *et al*. DNA-Prot: identification of DNA binding proteins from protein sequence information using random forest. Journal of Biomolecular Structure & Dynamics, 2009, **26**(6): 679−686

[12] Lin W Z, Fang J, Xiao X, *et al*. iDNA-Prot: identification of DNA binding proteins using random forest with grey mode. Plos One, 2011, **6**(9): e24756

[13] Liu B, Xu J, Lan X, *et al*. iDNA-Prot|dis: identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition. Plos One, 2014, **9**(9): e106691

[14] Liu B, Wang S, Wang X, *et al*. DNA binding protein identification by combining pseudo amino acid composition and profile-based protein representation. Scientific Reports, 2015, **5**: 15479

[15] Liu T, Qin Y, Wang Y, *et al*. Prediction of protein structural class based on gapped-dipeptides and a recursive feature selection approach. International Journal of Molecular Sciences, 2015,

17(1): 15−23

[16] Altschul S F, Madden T L, Schaffer A A, *et al*. Gapped Blast and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research, 1997, **17**(25):3389−3402

[17] Chang C C, Lin C J. Libsvm: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2011, **2**(3): 1−27

[18] Guyon I, Jason W, Stephen B. Gene selection for cancer classification using support vector machines. Machine Learning, 2002, **46**(1−3): 389−422

[19] 刘光徽，胡 俊，於东军，等. 基于多视角特征组合与随机森林的 G 蛋白偶联受体与药物相互作用预测. 南京理工大学学报(自然科学版), 2016, **40**(1): 1−9
Liu G H, Hu J, Yu D J, *et al*. Journal of Nanjing University of Technology(Natural Science Edition), 2016, **40**(1): 1−9

[20] Wang L, You Z, Xia S, *et al*. Advancing the prediction accuracy of protein-protein interactions by utilizing evolutionary information from position-specific scoring matrix and ensemble classifier. Journal of Theoretical Biology, 2017, **418**(4): 105−110

[21] Niu X H, Hu X, Shi F, *et al*. Predicting DNA binding proteins using support vector machine with hybrid fractal features. Journal of Theoretical Biology, 2014, **343**(2): 186−192

[22] Vapnik V N, Vapnik V. Statistical learning theory. New York: Wiley, 1998

[23] Lou W, Wang X, Chen F, *et al*. Sequence based prediction of DNA-binding proteins based on hybrid feature selection using random forest and Gaussian naive Bayes. Plos One, 2014, **9** (1): e86703

[24] Xu, R, Zhou J, Liu B, *et al*. enDNA-Prot: identification of DNA-binding proteins by applying ensemble learning. Biomed Research International, 2014, **2014**(1): 294279

[25] Qian Z, Cai Y D, Li Y. A novel computational method to predict transcription factor DNA binding preference. Biochemical and Biophysical Research Communications, 2006, **348**(3): 1034−1037

[26] Zhang Y P, Wu Y, Wei Z, *et al*. gDNA-Prot: predict DNA-binding proteins by employing support vector machine and a novel numerical characterization of protein sequence. Journal of Theoretical Biology, 2016, **406**(1): 8−16

Identification of DNA-binding Proteins Using Gapped-dipeptide Composition and Recursive Feature Elimination Algorithm*

TANG Ya-Dong[1)**], LIU Xiao[1)**], LIU Tai-Gang[2)], XIE Lu[3)], CHEN Lan-Ming[1)***]

([1)] *Laboratory of Quality and Safety Risk Assessment for Aquatic Products on Storage and Preservation*(*Shanghai*), *Ministry of Agriculture*, *College of Food Science and Technology*, *Shanghai* 201306, China;

[2)] *College of Information Technology*, *Shanghai Ocean University*, *Shanghai* 201306, China;

[3)] *Shanghai Center for Bioinformation Technology*, *Shanghai* 201203, China)

Abstract The identification of DNA-binding proteins (DBPs) plays an important role in functional annotation of genes and proteins of prokaryote and eukaryote organisms. This study, for the first time, combined the gapped-dipeptide composition (GapDPC) and recursive feature elimination (RFE) to identify DBPs. The position specific scoring matrix (PSSM) of each tested amino acid sequence was obtained. Based on the PSSM, their GapDPC features of the amino acid sequences were extracted, and then the optimal features were selected using the RFE method. Subsequently, the support vector machine (SVM) was chosen as a classifier and the datasets PDB396 and LB1068 were tested using the jackknife cross validation test. The result showed that the values of accuracy, Matthews correlation coefficient, sensitivity, and specificity for the identification of DBPs were 93.43%, 0.86, 89.04% and 96%, and 86.33%, 0.73, 86.49% and 86.18% for the datasets PDB396 and LB1068, respectively, which were obviously superior to the methods reported previously in the literature. The new model established in this study improved the identification methods of DBPs.

Key words DNA-binding proteins, gapped-dipeptide composition, position specific score matrix, recursive feature elimination algorithm, support vector machine classifier

DOI: 10.16476/j.pibb.2017.0413

* This work was supported by grants from The National Natural Science Foundation of China (31671946, 11601324) and Shanghai Municipal Science and Technology Commission Foundation (17050502200).

**These authors contributed equally to this work.

***Corresponding author.

Tel: 86-21-61900504, E-mail: lmchen@shou.edu.cn

Received: November 7, 2017 Accepted: March 9, 2018