Life Science Informatics Publications

Research Journal of Life Sciences, Bioinformatics, Pharmaceutical and Chemical Sciences

Journal Home page http://www.rjlbpcs.com/

RJLBPCS

ISSN 2454-6348

Original Research Article **DOI - 10.26479/2016.0105.01**

DEVELOPMENT OF A COMPUTATIONAL METHOD FOR LIPID-BINDING PROTEIN PREDICTION

K. Ueki[1], K. Sato[1], S. Nakamura[1], T. Terada[1], K. Sumikoshi[1] and K. Shimizu[1*]

1. Department of Biotechnology, Graduate School of Agricultural and Life Sciences,

The University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-8657, Japan.

ABSTRACT: Lipid-binding proteins (LBPs) perform various essential functions in organisms not only in cellular lipid uptake, lipid transport, and lipid metabolism but also in gene expression regulation, cell signaling, and innate immune response to bacterial infection. Because conducting experiments for identifying LBP functions is time-consuming and costly, computational methods may be useful for predicting lipid-associated functions. Here, we propose a method for predicting whether a given protein is an LBP from its amino acid sequence. Our method is based on the support vector machine (SVM), a machine learning algorithm, which is widely used in several prediction tools of bioinformatics. For SVM, feature selection is important for the accuracy of prediction. Our method uses the distribution of position-specific scoring matrix (PSSM) scores called as the position-specific score distribution (PSSD) as the input feature of SVM. PSSD is calculated from a PSSM which is generated by a multiple sequence alignment and summarizes the contents of PSSM. PSSD takes into account the homolog information while reducing the dimensions of the feature vector. Using the PSSD, our method achieved a value of the area under the receiver operating characteristic curve of 0.98 in a five-fold cross-validation test. In addition, our method achieved better performance in LBP function class prediction than that previously reported. We also examined one- and two-spectrum kernels, which have been widely used in protein function prediction, and showed that our method using the PSSD outperforms the existing methods.

LBPredictor: Available at http://www.bi.a.u-tokyo.ac.jp/software/

KEYWORDS: lipid-binding protein, function prediction, support vector machine, machine learning.

***Corresponding Author: Prof. Dr. Kentaro Shimizu** Ph. D.

Department of Biotechnology, Graduate School of Agricultural and Life Sciences,

The University of Tokyo, Japan. Email Address: shimizu@bi.a.u-tokyo.ac.jp

1. INTRODUCTION

Lipid-binding proteins (LBP) perform vital functions in organisms not only with regard to cellular lipid uptake, lipid transport, and lipid metabolism but also with regard to gene expression regulation, cell signaling, and innate immune response to bacterial infection [1-3]. LBPs have also been studied for the identification of therapeutic targets [4, 5]. However, because conducting experiments for identifying LBP functions is time-consuming and costly, methods of bioinformatics may be useful for predicting whether a protein is an LBP and what type of LBP functions does the protein performs have. A basic approach for predicting protein function is to use sequence similarity between the target protein and those already present in a database. However, in cases of low homology between the protein and those in the database, identifying function using this approach is difficult. LBPs have great variety in sequence and structure, and the detection of new LBPs in the database is not easy [6-9]. Therefore, we developed a prediction system using machine learning that can automatically extract information from sequence data.Machine learning methods have been used to predict protein function, functional binding sites, localization, and structural classification of target proteins as well as to produce highly accurate results in each kind of prediction [10-14]. When machine learning methods are applied to a problem, a numerical descriptor called the feature vector must be computed. Various descriptors, such as amino acid composition, physicochemical properties, Gene Ontology terms, and position-specific scoring matrix (PSSM) have been proposed [13, 15, 16]. The evolutionary information from PSSM has been considered as essential for discriminating a functional binding site from a nonbinding site, particularly in binding site prediction [17-20]. PSSM has also been used to predict protein function and localization by encoding PSSM to the fixed size input feature vector [21-23]. Here, we propose a new, simple method of converting PSSM to a feature vector effective for predicting LBP functions [1]. To our knowledge, prediction methods for LBP function classification have been proposed in two previous studies. Lin et al. [24] classified LBPs into the following nine classes: lipid degradation (LD), lipid metabolism (LM), lipid synthesis (LS), lipid transport (LT), lipid binding (LB), lipopolysaccharide biosynthesis (LPB), lipoprotein (LP), lipoyl, and all LBPs. They proposed a predictor using support vector machine (SVM) that consider physicochemical properties including hydrophobicity and polarity as inputs; SVM is a machine learning method applied to classification problems that searches for an optimized hyper plane in the feature space using a kernel function that defines the similarity between training samples. Lin et al. defined three groups of amino acids based on these properties and computed the composition, transition, and distribution in the sequence as descriptors of the proteins. The sensitivities and specificities of their predictor for each LBP class were in the ranges 76.6–90.6% and 97–99.9%, respectively. Bakhtiarizadeh et al. [25] used various properties extended from those used in the study by Lin et al., including amino acid composition, dipeptide composition, normalized Moreau–Broto autocorrelation, Moran autocorrelation, and other features obtained from a sequence. Bakhtiarizadeh et al. compared

the performance of two machine learning methods, neural network and SVM, and concluded that their predictor gave better performance using SVM than those using neural networks and reported by Lin et al. [26]. Here, we report the development of a predictor for LBP classification that uses new sequence features computed from a PSSM and compare the performance of our predictor with that of the previous study by Bakhtiarizadeh [25], hereafter referred to as "the previous study".

2. MATERIALS AND METHODS

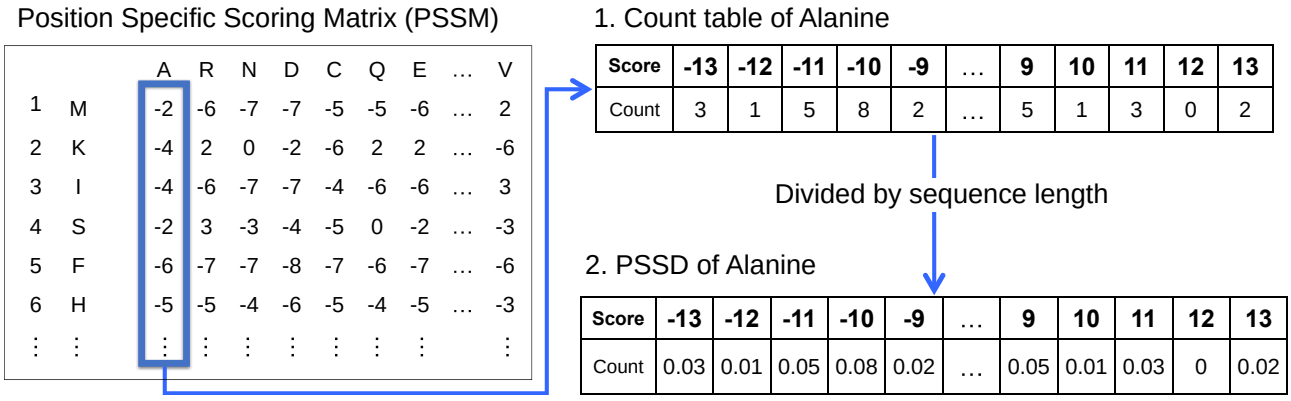
Dataset

We collected 10,603 LBPs as a positive dataset and 53,015 non-LBPs as a negative dataset. The positive dataset comprised the following eight LBP classes: (1) LB, (2) LD, (3) LM, (4) LS, (5) LT, (6) LP, (7) LPB, and (8) lipoyl. This classification is based on the annotation (description of the function of a protein) of the Swiss-Prot database; the datasets were constructed using a keyword search of the database. Redundancy in each LBP class was removed by clustering sequences with a similarity of >90% using CD-HIT [27]. The negative dataset of non-LBP proteins was constructed and further refined by the removal of proteins having common domains with proteins in the positive dataset [28]. The negative dataset also comprised eight groups, each corresponding to an LBP class. Each negative dataset comprised five subsets of the five-fold cross-validation (CV) test described in section 2.3, later collected independently to prevent probable bias from selecting the negative data. These datasets were based on the previous study [25], but 17 proteins (13 instances of negative data and 4 proteins in LP) lacking any homolog in the NCBI NR database were not used.

Feature Extraction

The PSSM of each protein was calculated from the multiple alignment of the sequences obtained by two iterations of Delta-Blast [29] against the NCBI NR database with an E-value cutoff of 0.05 for alignments with conserved domains and an E-value cutoff of 0.002 for pairwise alignments. PSSM is widely used to predict the function of a protein, its functional binding site, and its subcellular localization [12, 30]. The rows and columns of a PSSM describe alignment position and type of amino acid, respectively. PSSMs are often used as an input to machine learning. In the prediction of a functional binding site, a PSSM is generally converted to a feature vector with a fixed window size to extract a regionally conserved pattern. In contrast, with regard to the prediction of binding of a given protein to other molecules, the feature vector should be taken from wider regions of proteins. Accordingly, we extracted all conserved information called as position-specific score distribution (PSSD) from PSSM rather than the regional conservation information of fixed-size windows.

Fig. 1 Flow chart demonstrating the calculation of alanine-PSSD



		A	R	N	D	C	Q	E	…	V
1	M	-2	-6	-7	-7	-5	-5	-6	…	2
2	K	-4	2	0	-2	-6	2	2	…	-6
3	I	-4	-6	-7	-7	-4	-6	-6	…	3
4	S	-2	3	-3	-4	-5	0	-2	…	-3
5	F	-6	-7	-7	-8	-7	-6	-7	…	-6
6	H	-5	-5	-4	-6	-5	-4	-5	…	-3
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮

Score	-13	-12	-11	-10	-9	…	9	10	11	12	13
Count	3	1	5	8	2	…	5	1	3	0	2

Score	-13	-12	-11	-10	-9	…	9	10	11	12	13
Count	0.03	0.01	0.05	0.08	0.02	…	0.05	0.01	0.03	0	0.02

PSSD is a frequency distribution of all the scores for each amino acid type in PSSM. Each element of a PSSD is a relative frequency of all the PSSM scores of an amino acid and corresponds to an element of the feature vector. Fig. 1 illustrates the method of calculation of a PSSD for alanine. In this method, we first calculate the number of occurrences of each PSSM score (the count table) and divide it by the sequence length (the total number of occurrences) to calculate the relative frequency (PSSD).

Fig. 2 Relative frequencies of PSSM scores of all LBPs

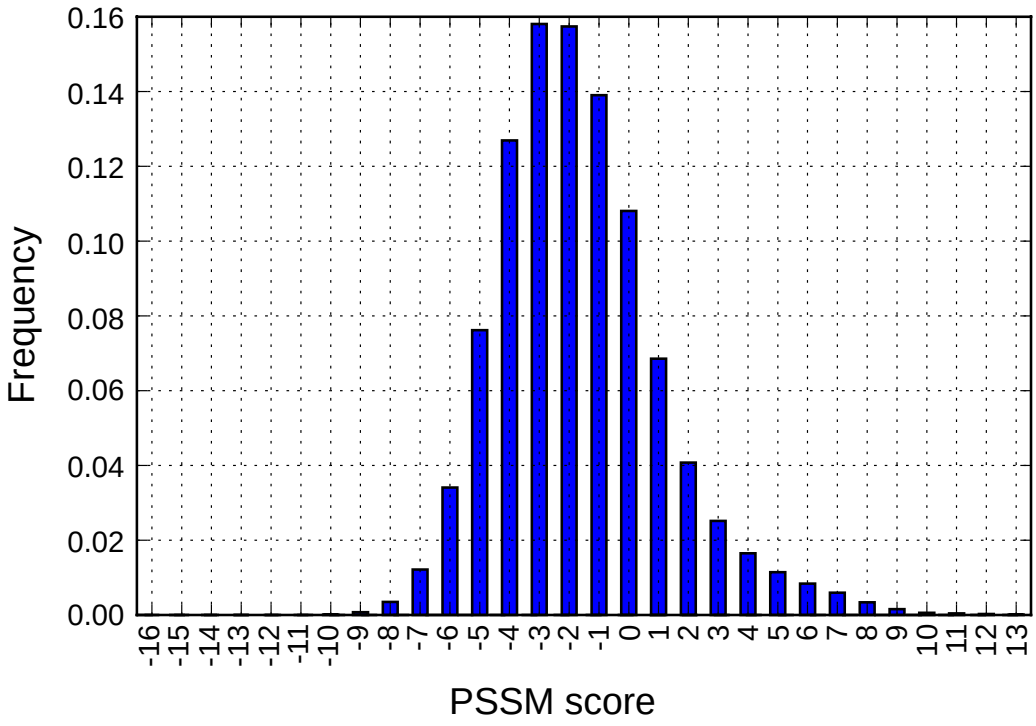


Fig 2 Relative frequencies of PSSM scores of all LBPs. We found that over 99.99% of scores were between −13 and 13 and decided to converting scores, the which greater than 13 and less than −13 to 13 and −13, respectively.

Support Vector Machine

We employed the radial basis function (RBF) that is capable of separating the training samples nonlinearly, as the kernel function. We used a library for support vector machines in the scikit-learn module [31, 32] to construct the prediction system with SVM. The cost parameter that determines the misclassification penalty and the gamma parameter used in the RBF kernel function were optimized on the basis of a five-fold CV test. We implemented a five-fold CV test for each LBP class and chose the optimal parameters based on the area under the receiver operating characteristic curve (AUC). Thus, these two parameters differed among the LBP classes.

Evaluation

The performance of our prediction system was measured using the five-fold CV test in which the complete datasets were randomly divided into five parts. One of the five parts was then used as a test set and the rest as training sets. This procedure was repeated five times until all parts were used as the test set. AUC is a measurement independent of the threshold of the decision value. AUC represents the separation ability of the prediction system, and an AUC value closer to 1 indicates a better prediction. We calculated the mean of AUC for five subsets of each LBP class. An SVM prediction system requires the threshold of its decision value to be fixed. The default value is usually set to 0, but we chose a decision value threshold based on the Matthews correlation coefficient (MCC)[13]. MCC is a balanced measurement used to assess the effectiveness of the prediction system. Three other measurements including accuracy (ACC), sensitivity (SE), and specificity (SP) were also computed with this threshold. These values were defined as follows:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}$$

$$ACC = \frac{TP + TN}{TP + FN + TN + FP}$$

$$SE = \frac{TP}{TP + FN}$$

$$SP = \frac{TN}{TN + FP}$$

where *TP*, *FN*, *TN*, and *FP* refer to the numbers of true positives, false negatives, true negatives, and false positives, respectively period. These values were computed five times using each test set in the five-fold CV, and the means of those measurements were computed. Then, we calculated the average

performance of the five subsets of each LBP class and compared the results with those of the previous study [25].

3. RESULTS AND DISCUSION

Performance of PSSD-based prediction

Table 1 The performance of the position-specific score distribution-based and previous method

	AUC		ACC		SE		SP		MCC
Lipid class	Previous	PSSD	Previous	PSSD	Previous	PSSD	Previous	PSSD	PSSD
All LBPs	0.95	**0.980**	89.28	**93.30**	89.2	**91.25**	89.22	**95.35**	**0.867**
Lipid binding (LB)	0.96	**0.984**	90.61	**95.13**	90.55	**94.46**	90.68	**95.81**	**0.904**
Lipid degradation (LD)	0.970	**0.992**	92.25	**96.54**	92.85	**96.74**	91.65	**96.34**	**0.931**
Lipid metabolism (LM)	0.96	**0.980**	89.15	**94.51**	90.55	**93.25**	88.75	**95.78**	**0.892**
Lipid synthesis (LS)	0.98	**0.996**	94.74	**98.09**	94.08	**97.41**	95.41	**98.77**	**0.962**
Lipid transport (LT)	0.96	**0.971**	88.84	**94.08**	87.55	**91.74**	90.11	**96.42**	**0.884**
Lipoprotein (LP)	0.96	**0.969**	89.06	**90.78**	**88.85**	88.77	89.06	**92.79**	**0.817**
Lipopolysaccharide biosynthesis (LPB)	0.97	**0.984**	93.26	**95.21**	92.58	**93.23**	93.94	**97.18**	**0.906**
lipoyl	**1**	0.999	98.58	**99.40**	97.76	**99.34**	99.4	**99.46**	**0.988**

The performance results of our prediction method and the method of Bakhtiarizadeh et al. are shown in Table 1. The PSSD-based method achieved ACC, SE, SP, and AUC of 93.30%, 91.25%, 95.35%, and 0.980, respectively, for the “all LBPs” class. These performances were higher than those reported in the previous study [25]; ACC and AUC were improved by 4.02 and 0.03 percentage points, respectively. ACC values of the individual LBP classes were improved by 0.86 (lipoyl class) to 5.49

(LM class) percentage points, and AUC of individual classes, except for the lipoyl class improved from 0.009 (LP class) to 0.024 (LB class). This result shows that the PSSD-based method succeeded in learning features shared among all LBPs but not among non-LBPs. Similarly, SE and SP of each LBP class showed higher performances by 0.64–4.10 and 0.12–6.77 percentage points, respectively, except for SE of LP. With regard to AUC, all LBP classes, except lipoyl, were improved by 0.008–0.025. Notably, AUCs achieved in this study were the mean of five subsets of each LBP class based on the five-fold CV test, although the previous study [25] used the AUC of a subset with the highest overall accuracy in an independent evaluation test. MCCs are also shown in Table 1, although not reported in the previous study [25]. These results showed that PSSD is more effective for predicting LBPs than the combined features, including compositions and various physicochemical properties. Moreover, the abovementioned results suggest that PSSD possesses some specific patterns or characteristics of LBPs.

Comparison with spectrum kernels

Table 2 The performance of spectrum kernels with k = 1 and k = 2 and position-specific score distribution (PSSD)

Lipid class	Method	AUC	ACC	SE	SP	MCC
All LBPs	k-spectrum kernel (k = 1)	0.655	60.78	62.96	58.61	0.227
	k-spectrum kernel (k = 2)	0.753	69.73	70.49	68.98	0.397
	PSSD	**0.980**	**93.30**	**91.25**	**95.35**	**0.867**

There are several methods appropriate for representing sequence features. One basic method is the k-spectrum kernel, which is widely used in sequence-based protein classification [34]. The kernel function of the k-spectrum is computed using k-mers that correspond to the substrings of k contiguous symbols occurring in a sequence. The dimension of the feature vector is 20^k, and over fitting is severe for large k. In this study, we implemented the prediction method using k-spectrum kernels with k = 1 and k = 2. The evaluation method was the same as that for PSSD. Table 2 compares the performance results of PSSD and the spectrum kernels with k = 1 and k = 2. As shown in the table, the spectrum kernel with k = 2 outperforms that with k = 1. This is because the spectrum kernel with k = 2 can extract more sequence features than k = 1. However, performance is further improved using the PSSD.

AUC values of PSSD are 0.325 and 0.227 higher than those of the spectrum kernels with k = 1 and k = 2, respectively.

DISCUSSION

The AUC value of 0.98 achieved by our method for predicting LBPs indicates that our method successfully discriminated LBPs from other proteins. Our method is based on the support vector machine (SVM), which is a machine learning algorithm widely used in several prediction tools of bioinformatics. In SVM, feature selection is important for prediction accuracy. We propose a new method to use the distribution of the PSSM scores called as PSSD as the feature input for SVM. Several previous studies have shown that PSSM can extract the sequence features effectively and have concluded that PSSM leads to better performance. PSSM is defined as a matrix that contains probability information of amino acids at each position in a multiple sequence alignment. PSSM describes the propensities of the residue substitutions at each position using information of homologs and is often used with windowing, particularly for functional site predictions including protein–protein and protein–ligand binding-site predictions. Window size is an important parameter. Functional sites can be affected by several residues, but large window sizes lead to overfitting owing to the large dimension of the feature vector. With regard to the problems of binding prediction in which the function of an entire protein is the target of the prediction, it is not practical to use windowing because the window size becomes large (up to the length of the amino acid sequence of the whole protein). Because our PSSD is based on PSSM, it inherits the merits of PSSM. PSSD is calculated from the PSSM of the whole sequence and summarizes the contents of PSSM. PSSD considers homolog information and also reduces the dimensions of the feature vector. These are the probable reasons for the greater effectiveness of PSSDs than the protein features previously reported [25] and k-spectrum kernels.

Table 3 Numbers of proteins and sequence diversity of lipid-binding protein function groups

Lipid class	No. of proteins	Average sequence similarity
All LBPs	10,603	-
Lipid binding (LB)	777	47.6
Lipid degradation (LD)	706	40.4
Lipid metabolism (LM)	616	37.5
Lipid synthesis (LS)	3,355	42.6
Lipid transport (LT)	235	42.5
Lipoprotein (LP)	4,026	41.0
Lipopolysaccharide biosynthesis (LPB)	553	43.6
lipoyl	335	56.9

As for LBP function class prediction, 0.969–0.999 of AUC values were achieved for all classes of LBPs. These values were obtained even when it was unknown whether a given protein was an LBP. Table 3 shows the number of sequences and their average similarity in each function class. The average similarity was calculated by averaging the sequence similarities obtained by the Smith–Waterman algorithm [35] of all-against-all pairs in a class. The finding that among the eight function classes, lipoyl can be predicted with a high AUC may be because of the higher similarity among lipoyl sequences than that found in other function classes. The high AUC of LS may be because of a large number of positive data. Lipid transfer (LT) had a small number of positive data, and its sequence similarity is small; AUC is lower than that of other function classes. LP, which had the largest number of positive data, had the lowest AUC, although its value of 0.969 was not lower than that of other function prediction tools. LPs include several enzymes, transporters, antigens, adhesins, and structural proteins; because of their variety, prediction tends to become difficult.

4. CONCLUSION

We developed an SVM-based system for predicting LBPs from their sequence. This system is intended to provide information to support laboratory experiments. As more data from high-throughput lipid proteomics becomes available and more knowledge is acquired, the reliability of predictions from our systems should improve because SVM performance depends on features extracted and training dataset quality. Here, we have described a new feature called PSSD obtained from PSSM. PSSD-based prediction for LBPs achieved better performance than the performance reported in the previous study. The overall AUC and ACC of prediction for all LBPs were 0.98 and 93.3%, respectively, which were higher by 0.03 and 4.02 percentage points, respectively, than those reported in the previous study [25]. PSSD can be applied to other protein functional predictions and is easily applicable to genome-wide predictions because it requires low computation cost. We have developed a web server, LBPredictor, to predict LBPs based on this study, available at http://www.bi.a.u-tokyo.ac.jp/software/.

ACKNOWLEDGEMENTS

We thank Masaki Banno and Wayne Dawson, past members of the Bioinformation Engineering Laboratory, for their support and valuable discussion. This study was supported by the Platform for Drug Discovery, Informatics, and Structural Life Science from the Ministry of Education, Culture, Sports, Science, and Technology, Japan.

CONFLICT OF INTEREST

The authors declare that no competing financial interests exist.

REFERENCES

1. Bingle CD, and Craven CJ (2004) Meet the relatives: a family of BPI- and LBP-related proteins. Trends in Immunology 25 : 53-55

2. Glatz JFC, Luiken J, van Bilsen M, and van der Vusse GJ (2002) Cellular lipid binding proteins as facilitators and regulators of lipid metabolism. Molecular and Cellular Biochemistry 239 : 3-7

3. Haunerland NH, and Spener F (2004) Fatty acid-binding proteins - insights from genetic manipulations. Progress in Lipid Research 43 : 328-349

4. Furuhashi M, and Hotamisligil GS (2008) Fatty acid-binding proteins: role in metabolic diseases and potential as drug targets. Nature Reviews Drug Discovery 7 : 489-503

5. Wolfrum C, Borrmann CM, Borchers T, and Spener F (2001) Fatty acids and hypolipidemic drugs regulate peroxisome proliferator-activated receptors alpha- and gamma-mediated gene expression via liver fatty acid binding protein: A signaling path to the nucleus. Proceedings of the National Academy of Sciences of the United States of America 98 : 2323-2328

6. Bingle CD, Bingle L, and Craven CJ (2011) Distant cousins: genomic and sequence diversity within the BPI fold-containing (BPIF)/PLUNC protein family. Biochemical Society Transactions 39 : 961-965

7. Niggli V (2001) Structural properties of lipid-binding sites in cytoskeletal proteins. Trends in Biochemical Sciences 26 : 604-611

8. Palsdottir H, and Hunte C (2004) Lipids in membrane protein structures. Biochimica Et Biophysica Acta-Biomembranes 1666 : 2-18

9. Reese AJ, and Banaszak LJ (2004) Specificity determinants for lipids bound to beta-barrel proteins. Journal of Lipid Research 45 : 232-243

10. Chen K, Mizianty MJ, and Kurgan L (2011) ATPsite: sequence-based prediction of ATP-binding residues. Proteome Science 9

11. Cui J, Han L, Lin H, Tang Z, Ji Z, Cao Z, Li Y, and Chen Y (2007) Advances in exploration of

machine learning methods for predicting functional class and interaction profiles of proteins and peptides irrespective of sequence homology. Current Bioinformatics 2 : 95-112

12. Huang C, and Yuan J (2013) Using radial basis function on the general form of Chou's pseudo amino acid composition and PSSM to predict subcellular locations of proteins with both single and multiple sites. Biosystems 113 : 50-57

13. Li ZR, Lin HH, Han LY, Jiang L, Chen X, and Chen YZ (2006) PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. Nucleic Acids Research 34 : W32-W37

14. Xiong W, Guo Y, and Li M (2010) Prediction of Lipid-Binding Sites Based on Support Vector Machine and Position Specific Scoring Matrix. Protein Journal 29 : 427-431

15. Li L, Cui X, Yu S, Zhang Y, Luo Z, Yang H, Zhou Y, and Zheng X (2014) PSSP-RFE: Accurate Prediction of Protein Structural Class by Recursive Feature Extraction from PSI-BLAST Profile, Physical-Chemical Property and Functional Annotations. Plos One 9

16. Rao HB, Zhu F, Yang GB, Li ZR, and Chen YZ (2011) Update of PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. Nucleic Acids Research 39 : W385-W390

17. Ahmad S, and Sarai A (2005) PSSM-based prediction of DNA binding sites in proteins. Bmc Bioinformatics 6:33

18. Fang C, Noguchi T, and Yamana H (2014) Simplified sequence-based method for ATP-binding prediction using contextual local evolutionary conservation. Algorithms for Molecular Biology 9

19. Kumar M, Gromiha AM, and Raghava GPS (2008) Prediction of RNA binding sites in a protein using SVM and PSSM profile. Proteins-Structure Function and Bioinformatics 71 : 189-194

20. Sodhi JS, Bryson K, McGuffin LJ, Ward JJ, Wernisch L, and Jones DT (2004) Predicting metal-binding site residues in low-resolution structural models. Journal of Molecular Biology 342 : 307-320

21. Chen S-A, Ou Y-Y, Lee T-Y, and Gromiha MM (2011) Prediction of transporter targets using

efficient RBF networks with PSSM profiles and biochemical properties. Bioinformatics 27 : 2062-2067

22. Kumar M, Gromiha MM, and Raghava GPS (2011) SVM based prediction of RNA-binding proteins using binding residues and evolutionary information. Journal of Molecular Recognition 24 : 303-313

23. Rashid M, Saha S, and Raghava GPS (2007) Support Vector Machine-based method for predicting subcellular localization of mycobacterial proteins using evolutionary information and motifs. Bmc Bioinformatics 8

24. Lin HH, Han LY, Zhang HL, Zheng CJ, Xie B, Cao ZW, and Chen YZ (2006) Prediction of the functional class of metal-binding proteins from sequence derived physicochemical properties by support vector machine approach. Bmc Bioinformatics 7

25. Bakhtiarizadeh MR, Moradi-Shahrbabak M, Ebrahimi M, and Ebrahimie E (2014) Neural network and SVM classifiers accurately predict lipid binding proteins, irrespective of sequence homology. Journal of Theoretical Biology 356 : 213-222

26. Lin HH, Han LY, Zhang HL, Zheng CJ, Xie B, and Chen YZ (2006) Prediction of the functional class of lipid binding proteins from sequence-derived properties irrespective of sequence similarity. Journal of Lipid Research 47 : 824-831

27. Huang Y, Niu B, Gao Y, Fu L, and Li W (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics 26 : 680-682

28. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer ELL, Tate J, and Punta M (2014) Pfam: the protein families database. Nucleic Acids Research 42 : D222-D230

29. Boratyn GM, Schaeffer AA, Agarwala R, Altschul SF, Lipman DJ, and Madden TL (2012) Domain enhanced lookup time accelerated BLAST. Biology Direct 7

30. Wang CC, Fang YP, Xiao JM, and Li ML (2011) Identification of RNA-binding sites in proteins by integrating various sequence information. Amino Acids 40 : 239-248

31. Chang C-C, and Lin C-J (2011) LIBSVM: A Library for Support Vector Machines. Acm Transactions on Intelligent Systems and Technology 2

32. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, and Duchesnay E (2011) Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12 : 2825-2830

33. Matthews BW (1975) Comparison of predicted and observed secondary structure of T4 phage lysozyme. Biochimica Et Biophysica Acta 405 : 442-451

34. Leslie C, Eskin E, and Noble WS (2002) The spectrum kernel: a string kernel for SVM protein classification. Pac Symp Biocomput : 564-575

35. Smith TF, and Waterman MS (1981) Identification of common molecular subsequences. Journal of Molecular Biology 147 : 195-197