J Ind Microbiol Biotechnol (2012) 39:579–584
DOI 10.1007/s10295-011-1047-z

BIOTECHNOLOGY METHODS

Prediction of ketoacyl synthase family using reduced amino acid alphabets

Wei Chen · Pengmian Feng · Hao Lin

Received: 10 August 2011 / Accepted: 4 October 2011 / Published online: 26 October 2011

Abstract Ketoacyl synthases are enzymes involved in fatty acid synthesis and can be classified into five families based on primary sequence similarity. Different families have different catalytic mechanisms. Developing cost-effective computational models to identify the family of ketoacyl synthases will be helpful for enzyme engineering and in knowing individual enzymes' catalytic mechanisms. In this work, a support vector machine-based method was developed to predict ketoacyl synthase family using the *n*-peptide composition of reduced amino acid alphabets. In jackknife cross-validation, the model based on the 2-peptide composition of a reduced amino acid alphabet of size 13 yielded the best overall accuracy of 96.44% with average accuracy of 93.36%, which is superior to other state-of-the-art methods. This result suggests that the information provided by *n*-peptide compositions of reduced amino acid alphabets provides efficient means for enzyme family classification and that the proposed model can be efficiently used for ketoacyl synthase family annotation.

W. Chen (✉)
Department of Physics, College of Sciences,
Center for Genomics and Computational Biology,
Hebei United University, Tangshan 063000, China
e-mail: chenwei_imu@yahoo.com.cn

P. Feng
Department of Preventive Medicine, College of Public Health,
Hebei United University, Tangshan 063000, China

H. Lin (✉)
Key Laboratory for NeuroInformation of Ministry of Education,
Center of Bioinformatics, School of Life Science and
Technology, University of Electronic Science and Technology
of China, Chengdu 610054, China
e-mail: hlin@uestc.edu.cn

Keywords Ketoacyl synthase family · Reduced amino acid alphabet · Support vector machine · *n*-Peptide

Introduction

Enzymes are proteins that catalyze (i.e., increase or decrease the rates of) chemical reactions. Almost all processes in a biological cell need enzymes to occur at significant rates. Enzymes are usually very specific as to which reactions they catalyze and the substrates that are involved in the reactions. Ketoacyl synthases (KSs) are key members of the fatty acid synthesis cycle used by organisms to form lipids [4]. In fatty acid synthesis, KSs condense the acyl-X chain with a carboxylated acyl-X chain [where X is either coenzyme A or acyl carrier protein (ACP)] and add two carbon atoms to the growing fatty acid chain by releasing a carbon dioxide and making ketoacyl-ACPs. Based on primary sequence similarity, KSs are divided into five families, KS1 to KS5 [4]. Knowing to which family an enzyme belongs casts new light on its catalytic specificity and gives clues to the relevant biological function. With the rapid increase in newly found protein sequences, the need for an automated and accurate tool to recognize enzyme families becomes increasingly important.

Over the past decade, many computational methods have been proposed for enzyme family and function prediction. By using the covariant discriminate function algorithm, Chou and Elord [10] predicted the subclasses of oxidoreductases. In follow-up work, Chou [8] improved the predictive accuracy to 70.6% by using amphiphilic pseudo amino acid composition. Afterwards, several other works were developed to predict enzyme subclasses. The GO-PseAA predictor was employed by Chou and Cai [9] to

predict enzyme subclasses. Later, by using functional domain composition and pseudo amino acid composition, Cai and Chou [2] obtained higher predictive accuracy. Recently, Shen and Chou [43] proposed a top-down approach for predicting enzyme functional classes, obtaining overall success rates ranging from 86.7% to 98.3%. The functions of proteins correlate with their three-dimensional (3D) structures. Based on the information of the 3D structure of proteins, González-Díaz and colleagues developed some models and web servers to discriminate between enzymes and nonenzymes [14, 15, 39], predict enzyme classes [13], and recognize protein kinases [26, 27]. They also developed some quantitative structure–activity relationship (QSAR)-based methods [16, 24, 25] to classify polygalacturonases and nonpolygalacturonases [1], discriminate dyneins from nondyneins [17], and predict RNase scores [23], achieving encouraging results.

However, to the best of our knowledge, there exists no theoretical method for KS family classification. In this article, we propose a support vector machine (SVM)-based method to identify KS families using reduced amino acid alphabets (RAAA) obtained by the protein blocks method [18, 19, 31, 47]. Compared with amino acid composition, RAAA can extract more useful information for protein sequences, eliminate some noise, reduce the dimension of the feature space, and improve the prediction accuracy. The performance of the proposed method was compared with that of other methods. Results demonstrate that this model could be a potentially useful tool for ketoacyl synthase family identification.

Materials and methods

Dataset

A total of 1,304 experimentally confirmed (evidence at transcript level or protein level) KSs were obtained from the ThYme database [4]. Highly similar data will surely lead to overestimation of the performance of the proposed method. To prepare a high-quality dataset, sequences with ≥60% identity were removed by using the CD-HIT program [34]. The final dataset contained 225 proteins. According to the database annotation, these proteins are classified into five families: 16 KS1, 30 KS2, 80 KS3, 29 KS4, and 70 KS5. If the sequence identity cutoff is set to a lower percentage (such as 25%), the results will be more objective and reliable. However, in this study we did not use such a stringent criterion because the currently available data do not allow this. Otherwise, the number of proteins for some subsets would be too few to have statistical significance.

Table 1 Scheme for reduced amino acid alphabets based on various protein blocks methods

Size	Protein blocks method
20	G-I–V-F-Y-W-A-L-M-E-Q-R-K–P-N-D-H–S-T-C
13	G-**IV**-**FYW**-A-L-M-E-**QRK**-P-**ND**-**HS**-T-C
11	G-**IV**-**FYW**-A-**LM**-**EQRK**-P-**ND**-**HS**-T-C
9	G-**IV**-**FYW**-**ALM**-**EQRK**-P-**ND**-**HS**-**TC**
8	G-**IV**-**FYW**-**ALM**-**EQRK**-P-**ND**-**HSTC**
5	G-**IVFYW**-**ALMEQRK**-P-**NDHSTC**

Clustered amino acids are shown in bold

Frequency of reduced amino acid alphabet

Reduced amino acid alphabets (RAAA) [21] clustered based on protein blocks [18, 19, 31] have been successfully employed in the area of protein annotation [7, 33, 40, 41, 47]. The RAAA scheme is shown in Table 1. Compared with the traditional amino acid composition, RAAA not only simplifies the complexity of the protein system but also improves the ability to find structurally conserved regions and structural similarity of entire proteins.

In this study, protein sequences are encoded by the frequency of n-peptide composition of RAAA of different sizes. For each value of n, the corresponding feature vector contains the fraction of each possible n-length substring in the sequence. The feature vector dimension (D) of the n-peptide composition obtained from RAAAs of different size (S) is listed in Table 2. The case $n = 1$ can be considered as the first-order approximation to the complete protein sequence. $n = 2$ gives the dipeptide composition, depicting the correlation of proximate residues. As n increases, n-peptides provide progressively more detailed sequential information. However, for $n \geq 3$, the amount of information parameters increases dramatically, and computation becomes not only impractical but also susceptible to the danger of overfitting. So, in the current study, we chose $n = 1$ and 2.

Classification protocol

SVM is a very powerful and popular method for supervised pattern recognition and has been widely used in the realm of bioinformatics [3, 6, 29, 37, 38, 42, 44–46]. To handle a multiclass problem, the one-versus-one (OVO) and one-versus-rest (OVR) approaches are generally applied to extend traditional SVM. In this work, the OVO strategy was employed to make predictions using radial basis functions (RBF). The SVM implementation was based on LibSVM written by Chang and Lin [5]. The grid search method was applied to tune the regularization parameter C and the kernel width parameter γ.

Table 2 Feature vector dimensions of *n*-peptide compositions with different RAAA sizes

n-Peptide	Dimension of different amino acid alphabet sizes (*S*)					
	$S = 20$	$S = 13$	$S = 11$	$S = 9$	$S = 8$	$S = 5$
$n = 1$	20	13	11	9	8	5
$n = 2$	400	169	121	81	64	25

Performance assessment

The capability of the method was evaluated using the sensitivity (Sn), specificity (Sp), overall accuracy (OA), average accuracy (AA), and Matthew's correlation coefficient (MCC). These measurements are expressed as follows:

$$\mathrm{Sn}(i) = \frac{\mathrm{TP}(i)}{\mathrm{TP}(i) + \mathrm{FN}(i)}, \quad (1)$$

$$\mathrm{Sp}(i) = \frac{\mathrm{TN}(i)}{\mathrm{TN}(i) + \mathrm{FP}(i)}, \quad (2)$$

$$\mathrm{OA} = \frac{1}{N}\sum_{i=1}^{k} \mathrm{TP}(i), \quad (3)$$

$$\mathrm{AA} = \frac{1}{N}\sum_{i=1}^{k} \mathrm{Sn}(i), \quad (4)$$

$$\mathrm{MCC}(i) = \frac{\mathrm{TP}(i) \times \mathrm{TN}(i) - \mathrm{FP}(i) \times \mathrm{FN}(i)}{\sqrt{(\mathrm{TP}(i) + \mathrm{FN}(i)) \times (\mathrm{TN}(i) + \mathrm{FP}(i)) \times (\mathrm{TP}(i) + \mathrm{FP}(i)) \times (\mathrm{TN}(i) + \mathrm{FN}(i))}}, \quad (5)$$

where k ($k = 5$) is the number of families and N is the total number of sequences in the final dataset. TP(i), TN(i), FP(i), and FN(i) represent the true positives, true negatives, false positives, and false negatives for family i.

Results and discussion

Three cross-validation methods, namely the subsampling test, independent dataset test, and jackknife test, are often employed to evaluate the predictive capability of a predictor. Among these three methods, the jackknife test is deemed as the most objective and rigorous [12] and can always yield a unique outcome, as demonstrated by a penetrating analysis in a recent comprehensive review [11]; it has been widely and increasingly adopted [20, 28, 30, 32, 35, 36, 43]. Accordingly, the jackknife test was used to examine the performance of the model proposed in this study. In the jackknife test, each sequence in the training dataset is in turn singled out as an independent test sample and all the rule parameters are calculated without using this one.

Ketoacyl synthase family classification

Each sequence in the dataset was translated into discrete feature vectors described by the frequencies of the *n*-peptide composition of RAAA. We firstly encoded protein sequences using the frequency of 20 amino acid and 400 dipeptide compositions. The jackknife test results are shown in the first two columns of Table 3. The overall accuracy of dipeptide composition reached 96.44% with average accuracy of 92.80%, which are higher than those of amino acid composition. Especially, all proteins in the KS2, KS3, and KS5 families were correctly recognized by using 400 dipeptides (i.e., sensitivity 100%).

To investigate whether a special class or property of amino acid affects the predictive accuracy and to determine the optimal amount of information, we compared the predictive capability of models trained by using the *n*-peptide ($n = 1, 2$) composition of RAAA of different sizes. The predictive sensitivity, specificity, average accuracy, and overall accuracy are listed in Table 3. As shown in Table 3, the best predictive results were obtained based on the 2-peptide composition of the reduced amino acid alphabet of size 13 ($n = 2$, $S = 13$). Although the overall accuracies were equal for the 2-peptide composition of RAAA with sizes $S = 11$, 13, and 20, the best MCC values for each family were obtained for the case $S = 13$ (Fig. 1). When $n = 2$ and $S = 13$, the MCC values for classification of KS1, KS2, KS3, KS4, and KS5 were 0.93, 1, 1, 0.88, and 1, respectively. Besides, the average accuracy of RAAA with $n = 2$ and $S = 13$ is the highest among all parameters. These results indicate that the *n*-peptide composition of reduced amino acid alphabets could extract more prominent structural and functional

Table 3 Result of SVM model based on different features

Family (%)	*n*-Peptide composition of RAAA with size *S* (*n*, *S*)											
	(1,20)	(2,20)	(1,13)	**(2,13)**	(1,11)	(2,11)	(1,9)	(2,9)	(1,8)	(2,8)	(1,5)	(2,5)
KS1												
Sn	50.00	81.25	50.00	**87.50**	56.25	81.25	62.50	62.50	56.25	68.75	43.75	37.50
Sp	99.47	100.00	98.97	**100.00**	97.87	100.00	98.92	100.00	99.47	100.00	98.79	99.47
KS2												
Sn	96.67	100.00	93.33	**100.00**	96.55	100.00	83.33	96.67	86.67	90.00	53.33	70.00
Sp	99.41	99.47	99.43	**100.00**	97.06	99.47	97.67	98.92	97.69	94.95	95.65	97.75
KS3												
Sn	90.00	100.00	95.00	**100.00**	88.75	100.00	95.00	100.00	92.50	96.25	88.75	98.75
Sp	87.41	95.14	90.58	**100.00**	87.14	95.14	83.57	92.31	87.05	92.31	73.88	84.67
KS4												
Sn	65.52	82.76	65.52	**79.31**	51.72	82.76	41.38	79.31	55.17	82.76	31.03	72.41
Sp	95.70	100.00	95.79	**100.00**	97.27	100.00	98.37	100.00	96.24	97.88	94.15	97.75
KS5												
Sn	98.57	100.00	100.00	**100.00**	100.00	100.00	100.00	100.00	100.00	100.00	95.71	97.14
Sp	100.00	100.00	100.00	**100.00**	100.00	100.00	100.00	100.00	100.00	100.00	99.04	100.00
OA	87.56	96.44	89.33	**96.44**	85.78	96.44	85.78	94.22	86.67	92.89	75.56	86.67
AA	80.15	92.80	80.77	**93.36**	78.65	92.80	76.44	87.70	78.12	87.55	62.51	75.16

The best results are shown in bold

information than the original amino acid or dipeptide composition.

Comparison with other methods

It is natural to ask whether the SVM model proposed here is superior to other state-of-the-art methods. For further investigation of the results, we compared the performance of SVM with that of random forest and naïve Bayes classifiers using the same features (2-peptide composition of the reduced amino acid alphabet of size 13). Random forest and naïve Bayes models were tested on the final dataset containing 225 sequences and implemented in WEKA [22]. The results are presented in Table 4. It is shown that the average accuracy of SVM is approximately 9% and 20% higher than the random forest and naïve Bayes classifiers, respectively. This result demonstrates that SVM can be used to identify ketoacyl synthase family with higher accuracy.

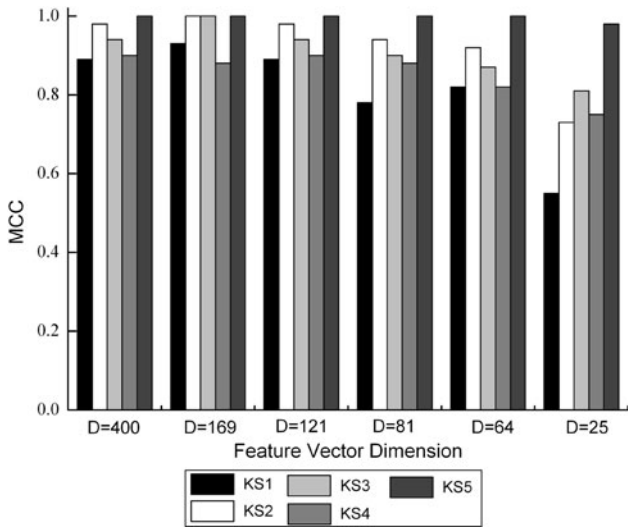

Fig. 1 MCC for five ketoacyl synthase families by using 2-peptide compositions with different RAAA sizes

Conclusions

We report a support vector machine-based approach for ketoacyl synthase family classification using reduced amino acid alphabets. The use of reduced amino acid alphabets not only provides an efficient and accurate way of protein vectorization for sequence-based protein classification systems but also achieves a remarkable improvement in terms of computational efficiency. High predictive accuracies show that the reduced *n*-peptide composition clustered based on protein blocks can extract more useful information than original 20 amino acid or 400 dipeptide compositions, and also demonstrate that our proposed method is a potentially useful tool for classification of ketoacyl synthase family. Moreover, as the dimension of the feature space was reduced by using RAAA (from 400 to 169 dimensions in this work), the reduced amino acid alphabet scheme could provide novel insights into proteomic classification tasks.

Table 4 Comparison of SVM with other methods for ketoacyl synthase family classification

Family	SVM			Random forest			Naïve Bayes		
	Sn (%)	Sp (%)	MCC	Sn (%)	Sp (%)	MCC	Sn (%)	Sp (%)	MCC
KS1	87.50	100.00	0.93	43.75	97.33	0.47	62.50	99.44	0.74
KS2	100.00	100.00	1.00	63.33	98.27	0.70	93.33	98.77	0.92
KS3	100.00	100.00	1.00	92.50	85.19	0.76	92.50	89.84	0.81
KS4	79.31	100.00	0.88	65.52	97.14	0.68	72.41	98.25	0.77
KS5	100.00	100.00	1.00	100.00	97.54	0.97	100.00	97.54	0.97
OA (%)		96.44			84.00			90.22	
AA (%)		93.36			73.02			84.15	

Acknowledgments The authors would like to thank three anonymous referees for constructive comments on the manuscript. This work was supported by the National Natural Science Foundation of China (no. 61100092), the Doctoral Scientific Research Start-up Foundation of Hebei United University (no. 10101115) to W.C., the Fundamental Research Funds for the Central Universities (ZYGX2009J081), and the Scientific Research Foundation of Sichuan Province (2009JY0013) to H.L.

References

1. Agüero-Chapin G, Varona-Santos J, de la Riva GA, Antunes A, González-Vlla T, Uriarte E, González-Díaz H (2009) Alignment-free prediction of polygalacturonases with pseudofolding topological indices: experimental isolation from coffea arabica and prediction of a new sequence. J Proteome Res 8:2122–2128
2. Cai YD, Chou KC (2005) Predicting enzyme subclass by functional domain composition and pseudo amino acid composition. J Proteome Res 4:967–971
3. Cai YD, Ricardo PW, Jen CH, Chou KC (2004) Application of SVM to predict membrane protein types. J Theor Biol 226:373–376
4. Cantu CD, Chen YF, Lemons ML, Reilly PJ (2011) ThYme: a database for thioester-active enzymes. Nucleic Acids Res 39:D342–D346
5. Chang CC, Lin CJ (2001) LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/_cjlin/libsvm
6. Chen W, Lin H (2010) Prediction of midbody, centrosome and kinetochore proteins using gene ontology. Biochem Biophys Res Commun 401:382–384
7. Chen YL, Li QZ, Zhang LQ (2010) Using increment of diversity to predict mitochondrial proteins of malaria parasite: integrating pseudo-amino acid composition and structural alphabet. Amino Acids [Epub ahead of print]. doi: 10.1007/s00726-010-0825-7
8. Chou KC (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21:10–19
9. Chou KC, Cai YD (2004) Using GO-PseAA predictor to predict enzyme sub-class. Biochem Biophys Res Commun 325:506–507
10. Chou KC, David WE (2003) Prediction of enzyme family classes. J Proteome Res 2:183–190
11. Chou KC, Shen HB (2007) Review: recent progress in protein subcellular location prediction. Anal Biochem 370:1–16
12. Chou KC, Zhang CT (1995) Prediction of protein structural classes. Crit Rev Biochem Mol Biol 30:275–349
13. Concu R, Dea-Ayuela MA, Perez-Montoto LG, Bolas-Fernández F, Prado–Prado FJ, Podda G, Uriarte E, Ubeira FM, González-Díaz H (2009) Prediction of enzyme classes from 3D structure: a general model and examples of experimental-theoretic scoring of peptide mass fingerprints of Leishmania proteins. J Proteome Res 8:4372–4382
14. Concu R, Dea-Ayuela MA, Perez-Montoto LG, Prado–Prado FJ, Uriarte E, Bolás-Fernández F, Podda G, Pazos A, Munteanu CR, Ubeira FM, González-Díaz H (2009) 3D entropy and moments prediction of enzyme classes and experimental-theoretic study of peptide fingerprints in Leishmania parasites. Biochim Biophys Acta 1794:1784–1794
15. Concu R, Podda G, Uriarte E, González-Díaz H (2009) Computational chemistry study of 3D-structure-function relationships for enzymes based on Markov models for protein electrostatic, HINT, and van der Waals potentials. J Comput Chem 30:1510–1520
16. Concu R, Podda G, Ubeira FM, González-Díaz H (2010) Review of QSAR models for enzyme classes of drug targets: theoretical background and applications in parasites, hosts and other organisms. Curr Pharm Des 16:2710–2723
17. Dea-Ayuela MA, Pérez-Castillo Y, Meneses-Marcel A, Ubeira FM, Bolas-Fernández F, Chou KC, González-Díaz H (2008) HP-Lattice QSAR for dynein proteins: experimental proteomics (2D-electrophoresis, mass spectrometry) and theoretic study of a Leishmania infantum sequence. Bioorg Med Chem 16:7770–7776
18. de Brevern AG, Etchebest C, Hazout S (2000) Bayesian probabilistic approach for prediction backbone structures in terms of protein blocks. Protein 41:271–287
19. de Brevern AG (2005) New assessment of a structural alphabe. In Silico Biol 5:283–289
20. Ding H, Liu L, Guo FB, Huang J, Lin H (2011) Identify golgi protein types with modified mahalanobis discriminant algorithm and pseudo amino acid composition. Protein Pept Lett 18:58–63
21. Etchebest C, Benros C, Bornot A, Camproux AC, de Brevern AG (2007) A reduced amino acid alphabet for understanding and designing protein adaptation to mutation. Eur Biophys J 36:1059–1069
22. Frank E, Hall M, Trigg L, Holmes G, Witten IH (2004) Data mining in bioinformatics using Weka. Bioinformatics 20:2479–2481
23. González-Díaz H, Dea-Ayuela MA, Pérez-Montoto LG, Prado–Prado FJ, Agüero-Chapín G, Bolas-Fernández F, Vazquez-Padrón RI, Ubeira FM (2010) QSAR for RNases and theoretic-experimental study of molecular diversity on peptide mass fingerprints of a new Leishmania infantum protein. Mol Divers 14:349–369
24. González-Díaz H, Duardo-Sanchez A, Ubeira FM, Prado–Prado F, Pérez-Montoto LG, Concu R, Podda G, Shen B (2010) Review of MARCH-INSIDE & complex networks prediction of drugs: ADMET, anti-parasite activity, metabolizing enzymes and cardiotoxicity proteome biomarkers. Curr Drug Metab 11:379–406

25. González-Díaz H, González-Díaz Y, Santana L, Ubeira FM, Uriarte E (2008) Proteomics, networks and connectivity indices. Proteomics 8:750–778
26. González-Díaz H, Saíz-Urra L, Molina R, González-Díaz Y, Sánchez-González A (2007) Computational chemistry approach to protein kinase recognition using 3D stochastic van der Waals spectral moments. J Comput Chem 28:1042–1048
27. Gonzalez-Díaz H, Saiz-Urra L, Molina R, Santana L, Uriarte E (2007) A model for the recognition of protein kinases based on the entropy of 3D van der Waals interactions. J Proteome Res 6(2):904–908
28. Gu Q, Ding YS, Zhang TL (2010) Prediction of G-protein-coupled receptor classes in low homology using Chou's pseudo amino acid composition with approximate entropy and hydrophobicity patterns. Protein Pept Lett 17:559–567
29. Hayat M, Khan A (2011) Predicting membrane protein types by fusing composite protein sequence features into pseudo amino acid composition. J Theor Biol 271:10–17
30. Hu LL, Niu S, Huang T, Wang K, Shi XH, Cai YD (2010) Prediction and analysis of protein hydroxyproline and hydroxylysine. PLoS One 5:e15917
31. Joseph AP, Agarwal G, Mahajan S, Gelly JC, Swapna LS, Offmann B, Cadet F, Bornot A, Tyagi M, Valadié H, Schneider B, Etchebest C, Srinivasan N, de Brevern AG (2010) A short survey on protein blocks. Biophys Rev 2:137–145
32. Kandaswamy KK, Chou KC, Martinetz T, Möller S, Suganthan PN, Sridharan S, Pugalenthi G (2011) AFP-Pred: a random forest approach for predicting antifreeze proteins from sequence-derived properties. J Theor Biol 270:56–62
33. Li J, Wang W (2007) Grouping of amino acids and recognition of protein structurally conserved regions by reduced alphabets of amino acids. Sci China C Life Sci 50:392–402
34. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659
35. Lin H (2008) The modified mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. J Theor Biol 252:350–356
36. Lin H, Chen W (2011) Prediction of thermophilic proteins using feature selection technique. J Microbiol Methods 84:67–70
37. Lin H, Ding H (2011) Predicting ion channels and their types by the dipeptide mode of pseudo amino acid composition. J Theor Biol 269:64–69
38. Mizianty MJ, Kurgan L (2011) Improved identification of outer membrane beta barrel proteins using primary sequence, predicted secondary structure, and evolutionary information. Proteins 79:294–303
39. Munteanu CR, González-Díaz H, Magalhães AL (2008) Enzymes/non-enzymes classification model complexity based on composition, sequence, 3D and topological indices. J Theor Biol 254:476–482
40. Nanni L, Lumini A (2008) A genetic approach for building different alphabets for peptide and protein classification. BMC Bioinformatics 9:45
41. Ogul H, Mumcuogu EU (2007) Subcellular localization prediction with new protein encoding schemes. IEEE/ACM Trans Comput Biol Bioinform 24:227–232
42. Park KJ, Gromiha MM, Horton P, Suwa M (2005) Discrimination of outer membrane proteins using support vector machines. Bioinformatics 21:4223–4229
43. Shen HB, Chou KC (2007) EzyPred: a top-down approach for predicting enzyme functional classes and subclasses. Biochem Biophys Res Comm 364:53–59
44. Xiong Y, Liu J, Wei DQ (2011) An accurate feature-based method for identifying DNA-binding residues on protein surfaces. Proteins 79:509–517
45. Zhu L, Yang J, Shen HB (2009) Multi label learning for prediction of human protein subcellular localizations. Protein J 28:384–390
46. Zou D, He Z, He J, Xia Y (2011) Supersecondary structure prediction using Chou's pseudo amino acid composition. J Comput Chem 32:271–278
47. Zuo YC, Li QZ (2010) Using K-minimum increment of diversity to predict secretory proteins of malaria parasite based on groupings of amino acids. Amino Acids 38:859–867