Genomics Proteomics Bioinformatics 11 (2013) 385–390

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

METHOD

Bagging with CTD – A Novel Signature for the Hierarchical Prediction of Secreted Protein Trafficking in Eukaryotes

Geetha Govindan *, **Achuthsankar S. Nair**

Department of Computational Biology and Bioinformatics, University of Kerala, Thiruvananthapuram 695581, India

Received 2 February 2013; revised 1 July 2013; accepted 17 July 2013
Available online 6 December 2013

KEYWORDS

Sequence driven features;
Sequence mapped features;
Autocorrelation;
Ensemble classifier;
Pathways;
Protein sorting

Abstract Protein trafficking or protein sorting in eukaryotes is a complicated process and is carried out based on the information contained in the protein. Many methods reported prediction of the subcellular location of proteins from sequence information. However, most of these prediction methods use a flat structure or parallel architecture to perform prediction. In this work, we introduce ensemble classifiers with features that are extracted directly from full length protein sequences to predict locations in the protein-sorting pathway hierarchically. Sequence driven features, sequence mapped features and sequence autocorrelation features were tested with ensemble learners and their performances were compared. When evaluated by independent data testing, ensemble based-bagging algorithms with sequence feature composition, transition and distribution (CTD) successfully classified two datasets with accuracies greater than 90%. We compared our results with similar published methods, and our method equally performed with the others at two levels in the secreted pathway. This study shows that the feature CTD extracted from protein sequences is effective in capturing biological features among compartments in secreted pathways.

Introduction

Eukaryotic cells contain complex compartments called organelles enclosed within membranes. Protein trafficking or protein sorting is a biological process where newly formed proteins get sorted and delivered to various organelles in the intracellular and secretory pathways [1]. Prediction of these protein localization sites in the pathways from the full length amino acid sequence is a complex process, which has not been fully elucidated yet. In 1982, Nishikawa et al. [2] reported that amino acid composition correlates with localization sites and each localization site in a cell has a unique set of functions. Hence protein localization prediction has implications both for the function of the protein and its possibility of interacting with other proteins in the same compartment [3,4].

Major protein sorting pathways can be divided hierarchically into secretory and intracellular types [5,6]. In a secretory pathway, all non-secretory proteins are delivered to the endoplasmic reticulum (ER) and then transported to other related

* Corresponding author.
E-mail: geetha@sctimst.ac.in (Govindan G).

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

ELSEVIER Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.gpb.2013.07.005

locations, which is controlled by ER signal sequences located in the N-termini. On the other hand, in an intracellular pathway, proteins with organelle-specific signal sequences are imported into the nucleus or mitochondria, according to their signal sequence type. The remaining proteins lacking sorting signals are located in the cytosol [7,8].

The success of computational prediction relies on the extraction of biological features from the sequence and the computational technique used [9–13]. A wide variety of methods have been tried throughout the years in order to predict the subcellular localization of proteins from full length sequence features. Methods reported differ in terms of input data and the technique employed to make the prediction about subcellular location. According to studies reported by Nakashima and Nishawa [14], intracellular and secretory proteins differ significantly in their amino acid compositions and in residue pair frequencies. Therefore, in this study simpler and less expensive methods that can extract features from full length protein sequence were given priority. The main advantage of our feature extraction methods over existing techniques is that features are extracted from the full length protein sequence based on various coding schemes without referencing external databases. For computation, we used hierarchical ensemble learning [15–19] (**Figure 1**) by mimicking the protein trafficking phenomenon which is incorporated from the location descriptions provided by the Gene Ontology (GO) Consortium [20] with the sequence features as input.

Results and discussion

Two basic ensemble based classifiers, bagging and AdaBoost M1 were trained to classify the location compartment of proteins in the intracellular and secretory pathways using the Waikato environment for knowledge analysis (WEKA) [21]. Two tests were carried out with two datasets for performance evaluation. These include a 6-fold cross validation test, which means randomly partitioning the dataset into equally sized training and test sets, training on 5 sets and testing with the 6th set and averaging the results, and an independent data test, which means training on one set and testing with another set by dividing the dataset into two random groups. The performance evaluation parameters specificity (Sp), sensitivity (Sn), accuracy (Acc), Mathew's correlation coefficient (MCC), positive predictive value (PPV), negative predictive value (NPV) and receiver operating characteristic (ROC) were calculated at all levels for comparing our results with the published results.

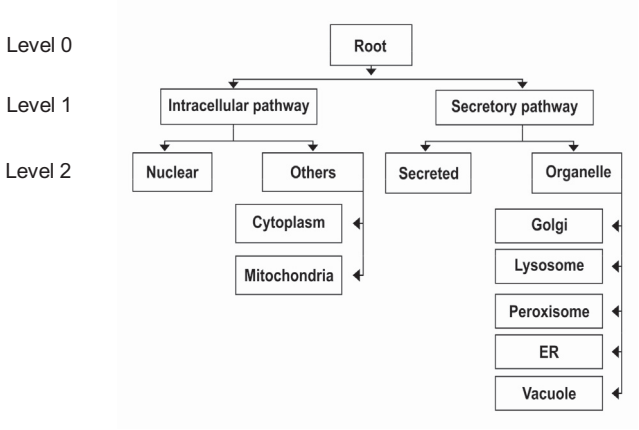

Figure 1 Hierarchical structures of compartments in protein trafficking

Adopted from [15–19]. Level 0, root of hierarchy; Level 1, first division; Level 2, second division.

Tables S1 and S2 show the average of the classifier performance parameters obtained from the two datasets at various levels of the pathway hierarchy in 6-fold cross validation and independent data test. These results were compared with the similar work of LOCtree [15] in Table S3. Table S4 shows the comparison of our classifier performance parameters with the LocTree2 [16] dataset for 5-fold cross validation.

Comparison with existing methods

Our method provides a hierarchical system for the prediction of protein subcellular localization with features generated exclusively from the full length sequence without using any server generated inputs. Similar classification work was reported by LOCtree [15] and LocTree2 [16]. LOCtree used the amino acid composition (20 units), composition of the 50 N-terminal residues (20 units) , amino acid composition from three secondary structure states and SignalP server [22] outputs as a feature vector on a support vector machine, whereas LocTree2 used the profiles created by BLAST-ing [23].

Although the results reported by LOCtree [15] are not directly comparable to ours in terms of features, selection of data, sizing of the data, and method of accuracy calculation, PPV, NPV and MCC reported by our method proved to be better at Level 0 and Level 1 of the hierarchy in the secreted pathway. The overall accuracy mentioned in LOCtree [15] is the PPV result based on the 6-fold cross validation experiments from a single dataset. At Level 0, our independent data testing results based on AdaBoost M1 and bagging reported average accuracies above 95% (Table S3) between the intracellular and secretory pathways with four of the sequence features. Bagging reported accuracy above 91% for classifying proteins between the secretory and organelle pathways with independent data testing. Because there is no result published for independent data tests by LOCtree [15], results obtained by this method cannot be compared.

For the 6-fold cross validation test (Table S3), our method reported accuracies above 92% at Level 0 for both bagging and AdaBoost M1 with an average MCC of 0.87, which was reported as 0.73 when using the LOCtree method. At Level 1, AdaBoost M1 and bagging reported PPVs above 90% with MCC above 0.70 while LOCtree reported an MCC of 0.55. Classifier bagging with sequence feature CTD performed better than LOCtree in differentiating the cytoplasm and mitochondrial pathways at Level 2.

LocTree2 is developed using a different hierarchical pathway and hence we could do the testing only for two levels using a LocTree2 dataset under 5-fold cross validation. Our method reported accuracies above 88% at Level 0 (Table S4) for all features under bagging while LocTree2 reported 90%. For level 1, bagging with feature vector CTD reported an accuracy of 82%, which is also comparable to that reported by LocTree2, 83%.

Conclusion

Previous protein localization prediction methods have been implemented using standard machine learning algorithms with

parallel architecture as a common practice in computer science [24–26]. Here novel systems of ensemble learners using hierarchical architecture from features extracted directly from full length protein sequences that can predict localization have been tested and the results have been compared.

Our testing results at the secretory pathway of hierarchy show that the prediction accuracy can be significantly improved by using the classifier bagging with feature vector CTD. The system achieved an overall accuracy above 90% with this sequence signature using bagging on independent data tests, suggesting that the native protein localization for each compartment is imprinted onto the features extracted from protein sequence. Feature generation methods described in this paper works independently and no server/external data reference is required for its extraction. Methods are based on the composition of amino acid. Additionally, this hierarchical structure has provided insights into the sorting process, such as the accurate distinction between the intracellular and secretory pathways. However, we observed that, as one descends the hierarchical path, the prediction accuracy progressively decreases as the classification task complexity increases. The best scoring decisions reported are at the top, and the worst are at the bottom. Thus, hierarchical model classification is unable to correct a prediction mistake made at the top node.

This study supports the hypothesis reported by Nakashima and Nishawa [14] that intracellular and secretory proteins differ significantly in their amino acid compositions. Both classifiers performed well using three sequence features at the top levels of hierarchy.

In the future, this classification method could be potentially extended to any level in the hierarchy using these sequence features and with the location descriptions provided by the Gene Ontology Consortium [20]. This method can predict the final localization of the protein as well as the mechanism underlying such localization. Our result may aid the development of more accurate predictors of protein function.

Materials and methods

Dataset construction

Two datasets (Table S5) were compiled for this study, which are denoted as ASN_G 1756 (Human) and ASN_G 1008 (Eukaryote). ASN_G (Human) is collected from a manually curated database for the subcellular localizations of proteins in human [27] and ASN_G (Eukaryote), which is from eSLDB [17], is a database for eukaryotic organisms. These are the only two manually curated public databases with experimental annotations reported in www.psort.org [28] for eukaryotes. ASN_G (Human) and ASN_G (Eukaryote) is maintained by the Rost lab of Columbia University Bioinformatics Centre and the Bologna Biocomputing Group, University of Bologna, respectively. These experimentally annotated proteins were finalized by verifying with UniProt (www.uniprot.org, release 2011-02 Sept–Oct) and by selecting the sequences that had a determined single subcellular location. Entries in the subcellular location that were annotated as ''putative'', ''potential'', ''possible'' and ''by similarity'' were eliminated to remove sequences with ambiguous and uncertain annotations.

We used the Cluster Database at High Identity with Tolerance (CD-HIT-2D) [29] web server to eliminate sequences in both datasets that displayed a similarity greater than or equal to 30%. The program (CD-HIT) takes a fasta format sequence database as input and produces a set of 'non-redundant' representative sequences as output by removing the highly similar sequences.

For comparing our results with the LocTree2, we downloaded 1682 sequences from the LocTree2 publication site [16] and generated a dataset with 1677 sequences (Table S7) after verifying the subcellular localizations with UniProt (March 2013).

Sequence feature formation

The features extracted from protein full length sequence can be classified into three groups. The first group consists of sequence driven features, which are generated directly from sequence through converting the protein sequence into a numeric sequence by replacing each amino acid with equivalent numeric values, counts, *etc*. The second group consists of sequence mapped features, which are generated by mapping amino acids into sub groups and the third group contains sequence autocorrelation features, which are obtained from calculations based on three types of spatial autocorrelation (Moreau-Broto, Moran and Geary).

Sequence driven features

There are two composition features considered, which include amino acid dipeptide composition (dipeptide descriptors) and composition of physico-chemical properties (amino acid index). Properties of dipeptides are determined by the amino acids forming the dipeptide. Dipeptide composition, which gives a fixed pattern length of 400 (20×20), encapsulates the global information about each protein sequence and the order it contains [30]. For example, in the sample protein sequence GCATGGTGCGAAACTTTGGCTG, 400 pairs of dipeptide occurrence frequency with no skips c0, are calculated by counting its presence in the sequence with no gaps. In **Figure 2**, the count of $c0_{GC}$ is 3, one skip $c1_{GC}$ is 1 and two skips $c2_{GC}$ is 1. The dipeptide count, 'cN_{xx}', counts pairs with N skips between them. The feature vector using the dipeptide occurrence frequency count for a protein sequence is represented as three separate numeric counts of its dipeptide c0, c1 and c2, each having 400 components. The final feature vector of 1200 com-

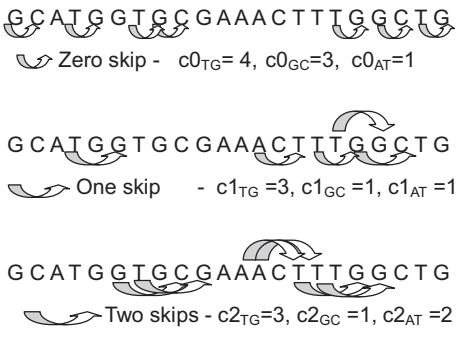

Figure 2 Amino acid di-peptide (GC, TG, AT) count with skips in a sample sequence

c0 indicates count of dipeptides with zero skip, c1 indicates count of dipeptides with one skip and c2 indicates count of dipeptides with two skips.

ponents is formed by concatenating the corresponding vectors c0, c1 and c2.

The Amino Acid Index (AAindex -1,2,3) is a database of numerical indices representing various physico-chemical and biochemical properties of amino acids and pairs of amino acids [31]. Physico-chemical properties derived from the AAindex1 database having 544 indices are used to compute the features. Feature vector having 544 components is represented as $\{f_1\ f_2\ f_3 \ldots f_{544}\}$ where f_1 is the physico-chemical property value for all residues of the sequence divided by the length of the sequence.

Sequence mapped features (CTD descriptors)

Structural variation in the R groups of amino acids is considered as the main factor for its difference in properties. From side chains we can classify amino acids into four groups (1) non-polar and neutral, (2) polar and neutral, (3) acidic and polar, and (4) basic and polar. The 20 amino acids forming the protein sequence can also be divided into several groups based on their other properties like (5) charge, (6) hydrophilicity or hydrophobicity, (7) size, and (8) functional groups. Twenty amino acids can be mapped into 1–3 groups by replacing each amino acid code with its group code. From the mapped sequence, features called composition, transition and distribution (CTD) can be calculated. Composition is determined as the number of amino acids of a particular property divided by total number of amino acids, whereas transition is determined as the number of transition from a particular property to different property divided by (total number of amino acids − 1). Distribution is the chain length within which the first, 25%, 50%, 75% and 100% of the amino acids of a particular property are located.

According to the property types, amino acids are divided into three groups and are marked as numeric indices 1, 2 and 3 (Table S6). Properties whose attributes can be grouped perfectly into three sets like charge, hydrophobicity, normalized van der Waals volume, polarity, polarizability, secondary structure and solvent accessibility are used for this mapping [32–35]. For example, according to secondary structure property grouping, the sample protein sequence HEAMRQLTIFVCYWNSPDDG is coded as "222222233 33333111111". In this example with the property of secondary structure, the total count of the coil is 6, the helix is 7 and the strand is 7. Hence the composition is calculated as 6/20, 7/20 and 7/20, where 20 is the total length of the sequence. Three numbers of composition descriptors are formed from three groups.

The transition from class 1 to 2 is the percentage frequency with which class 1 is followed by class 2 or class 2 is followed by class 1 in the encoded sequence, likewise the transition from class 3 to class 1 or class 1 to class 3, *etc*. For the sample sequence, the sum of transition from 2 to 3 and 3 to 2 is 1. Hence transition = 1/19.

The distribution descriptor describes the distribution of each property in the sequence. Five distribution descriptors are formed for each group, including the position percentages in the sequence for the first residue, 25% of the residues, 50% of the residues, 75% of the residues and 100% of the residues. Fifteen distribution descriptors are formed from three groups. In total 21 CTD descriptors are formed from a sequence.

For this study, CTD calculation is performed for 7 properties for each protein sequence after dividing each sequence into three equal segments. In total, 21×3 attributes for a sequence and 441 attributes for 7 properties compose the final feature vector.

Sequence autocorrelation features (autocorrelation descriptors)

Sequence autocorrelation-based features are based on the Tobler's first law of geography – "everything is related to everything else but nearby things are more related than distant things" [36]. Sequence autocorrelation-based features also assume that "the disturbances in each area are systematically related to those in adjacent areas" [37]. Spatial autocorrelation is the correlation of the variable with itself through space. Spatial autocorrelation measures the degree to which near and distant things are related, which is positive when nearby things are similar and negative when they are dissimilar. This concept helps to analyze the dependency among the features of sequences in each location.

Autocorrelation features are calculated based on the distribution of amino acid properties along the sequence. Thirty nine amino acid indices related to hydrophobicity are used for calculation after replacing each amino acid with its equivalent normalized index as P_i. Three autocorrelation descriptors are used as features, including normalized Moreau-Broto autocorrelation descriptors [38], Moran auto-correlation descriptors [39] and Geary autocorrelation descriptors [40].

The Moreau-Broto autocorrelation descriptor is defined as

$$\mathrm{MB}(d) = \sum_{i=1}^{N-d} P_i P_{i+d} \quad \text{where } d = 1,\ 2,\ 3 \text{ upto Max.lag}$$

where d is the lag of the autocorrelation, N is the length of the sequence, and P_i and P_{i+d} are the amino acid index value of the selected property at position i and $i + d$, respectively. Max.lag is the maximum value of the lags.

The normalized Moreau-Broto autocorrelation descriptors are defined as $\mathrm{MB}(d)/(N - d)$.

The Moran autocorrelation descriptor is defined as

$$\mathrm{Moran}\ (d) = \frac{\frac{1}{N-d}\sum_{i=1}^{N-d}(P_i - \bar{P})(P_{i+d} - \bar{P})}{\frac{1}{N}\sum_{i=1}^{N}(P_i - \bar{P})^2} \quad d = 1,\ 2,\ 3 \ldots, 30$$

$$\bar{P} = \frac{\sum_{i=1}^{N} P_i}{N}$$

where P_i and P_{i+d} have the same meaning as above.

The Geary autocorrelation descriptor is defined as

$$\mathrm{Geary}\,(d) = \frac{\frac{1}{2(N-d)}\sum_{i=1}^{N-d}(P_i - P_{i+d})^2}{\frac{1}{N-1}\sum_{i=1}^{N}(P_i - \bar{P})^2} \quad d = 1,\ 2,\ 3 \ldots, 30.$$

where $\bar{P}$, P_i and P_{i+d} have the same meaning as above. 3510 attributes from 39 amino acid properties with 30 lags compose the sequence feature vector for autocorrelation.

Computational techniques used

Among prediction algorithms, ensemble learning is a process by which multiple models such as classifiers are generated and combined to improve overall prediction accuracy [41]. Multiple learners (base learners) are trained to solve the same

problem by averaging over multiple classification models with different input feature vectors. These ensemble techniques reduce the small sample size problem which is critical in biological applications. This method reduces the over fitting of data. The three most popular classifiers based on the ensemble method, are bagging [42], AdaBoost M1 [43] and Random Forest [44]. In this study, two methods bagging and AdaBoost were used to predict protein trafficking at all levels of protein sorting pathway.

Bagging is the name derived from "bootstrap aggregation". This method uses multiple versions of a training set on different models by using the bootstrap (sampling with replacement). The outputs of the models are combined (average or vote) to create a single output. AdaBoost M1 adopts an adaptive sampling by using all instances of each iteration. In bagging, each classifier has the vote of the same strength, whereas AdaBoost M1 assigns different voting strengths to classifiers based on their accuracy.

Performance evaluation parameters

The classifier performance evaluation parameters specificity, sensitivity, accuracy, MCC [45], PPV [46], NPV [46] and ROC [47] were calculated at all levels as per the below equations. Specificity (Sp) is determined as (TN)/(TN + FP), where TN indicates true negative and FP means false positive. Sensitivity is defined as (TP)/(TP + FN), where TP means true positive and FN means false negative. Accuracy is defined as (TP + TN)/(TP + TN + FP + FN). PPV and NPV is calculated as (TP)/(TP + FP) and (TN)/(TN + FN), respectively. MCC is calculated as $\frac{TP\times TN - FP\times FN}{\mathrm{sqrt}((TP+FN)\times(TP\times FP)\times(TN\times FN)\times(TN\times FP))}$.

Authors' contributions

GG collected the dataset, conducted the data analysis, did machine learning experiments and wrote the manuscript. ASN conceived the original idea of using ensemble classifiers for the prediction of protein localization hierarchically. Both authors read and approved the final manuscript.

Competing interests

The authors declared that no competing interests exist.

Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.gpb.2013.07.005.

References

[1] Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. Molecular biology of the cell. 4th ed. New York: Garland Science; 2002.

[2] Nishikawa K, Kubota Y, Ooi T. Classification of proteins into groups based on amino acid composition and other characters. J Biochem 1983;94:997–1007.

[3] Bork P, Eisenhaber F. Wanted: subcellular localization of proteins based on sequence. Trends Cell Biol 1998;8:169–70.

[4] Drawid A, Gerstein M. A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. J Mol Biol 2000;301:1059–75.

[5] Rusch SL, Kendall DA. Protein transport via amino-terminal targeting sequences common themes in diverse systems. Mol Membr Biol 1995;12:295–307.

[6] Horton P, Nakai K. A probabilistic classification system for predicting the cellular localization sites of proteins. Proc Int Conf Intell Syst Mol Biol 1996;4:109–15.

[7] Lodish H, Berk A, Zipursky SL, Matsudaira P, Baltimore D, Darnell J. Molecular cell biology. 4th ed. New York: W.H. Freeman; 2000.

[8] Cooper GM, Hausman RE. The cell: a molecular approach. 5th ed. Washington: ASM Press; and Sunderland: Sinauer Associates; 2009.

[9] Reinhardt A, Hubbard T. Using neural networks for prediction of the subcellular location of proteins. Nucleic Acids Res 1998;26:2230–6.

[10] Emanuelsson O, Nielsen H, Brunak S, von Heijne G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 2000;300:1005–16.

[11] Nakai K. Prediction of in vivo fates of proteins in the era of genomics and proteomics. J Struct Biol 2001;134:103–16.

[12] Chou KC, Cai YD. Prediction and classification of protein subcellular location sequence order effect and pseudo amino acid composition. J Cell Biochem 2003;90:1250–60.

[13] Tantoso E, Li KB. AAindexLoc: predicting subcellular localization of proteins based on a new representation of sequences using amino acid indices. Amino Acids 2008;13:345–53.

[14] Nakashima H, Nishikawa K. Discrimination of intracellular and extracellular proteins using amino acid composition and residue pair frequencies. J Mol Biol 1994;238:54–61.

[15] Nair R, Rost B. Mimicking cellular sorting improves prediction of subcellular localization. J Mol Biol 2005;348:85–100.

[16] Goldberg T, Hamp T, Rost B. LocTree2 predicts localization for all domains of life. Bioinformatics 2012;28:458–65.

[17] Pierleoni A, Martelli PL, Fariselli P, Casadio R. ESLDB: eukaryotic subcellular localization database. Nucleic Acids Res 2006;35:208–12.

[18] Pierleoni A, Martelli PL, Fariselli P, Casadio R. BacelLo: a balanced subcellular localization prediction. Bioinformatics 2006;22:408–16.

[19] Lin T, Murphy RF, Bar-Joseph Z. Discriminative motif finding for predicting protein subcellular localization. IEEE/ACM Trans Comput Biol Bioinform 2011;8:441–51.

[20] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000;25:25–9.

[21] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. SIGKDD Explor 2009;11:10–8.

[22] Bendtsen JD, Nielsen H, von Heijne G, Brunak S. Improved prediction of signal peptides. SignalP 3.0. J Mol Biol 2004;340:783–95.

[23] Altschul SF. Basic local alignment search tool. J Mol Biol 1990;215:403–10.

[24] Cherian BS, Nair AS. Protein location prediction using atomic composition and global features of the amino acid sequence. Biochem Biophys Res Commun 2010;391:1670–4.

[25] Su EC, Chiu H, Lo A, Hwang J, Sung T, Hsu W. Protein subcellular localization prediction based on compartment-specific features and structure conservation. BMC Bioinformatics 2007;8:330.

[26] Blum T, Briesemeister S, Kohlbacher O. MultiLoc2: integrating phylogeny and gene ontology terms improves subcellular protein localization prediction. BMC Bioinformatics 2009;10:274.
[27] Rastogi S, Rost B. LocDB, experimental annotation of localization for *Homo sapiens* and *Arabidopsis thaliana*. Nucleic Acids Res 2011;39:D230–4.
[28] Nakai K, Horton P. PSORT: a program for detecting the sorting signals of proteins and predicting their subcellular localization. Trends Biochem Sci 1999;24:34–6.
[29] Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT suite: a web server for clustering and comparing biological sequences. Bioinformatics 2010;26:680–2.
[30] Ding Y, Cai Y, Zhang G, Xu W. The influence of dipeptide composition on protein thermostability. FEBS Lett 2004;569:284–8.
[31] Tomii K, Kanehisa M. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. Protein Eng 1996;9:27–36.
[32] Dubchak I, Muchnik I, Holbrook SR, Kim SH. Prediction of protein folding class using global description of amino acid sequence. Proc Natl Acad Sci U S A 1995;92:8700–4.
[33] Cai CZ, Han LY, Ji ZL, Chen X, Chen YZ. SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. Nucleic Acids Res 2003;31:3692–7.
[34] Lin HH, Han LY, Zhang HL, Zheng CJ, Xie B, Chen YZ. Prediction of the functional class of metal-binding proteins from sequence derived physicochemical properties by support vector machine approach. BMC Bioinformatics 2006;7:S13.
[35] Rao HB, Zhu F, Yang GB, Li ZR, Chen YZ. Update of PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. Nucleic Acids Res 2011;39:W385–90.
[36] Tobler W. A computer movie simulating urban growth in the Detroit region. Econ Geogr 1970;46:234–40.
[37] Loftin C, Ward SK. Spatial autocorrelation models for Galton's problem. Behav Sci Res 1981;16:105–41.
[38] Feng ZP, Zhang CT. Prediction of membrane protein types based on the hydrophobic index of amino acids. J Protein Chem 2000;19:269–75.
[39] Horne DS. Prediction of protein helix content from an autocorrelation analysis of sequence hydrophobicities. Biopolymers 1988;27:451–77.
[40] Sokal RR, Thomson BA. Population structure inferred by local spatial autocorrelation: an example from an Amerindian tribal population. Am J Phys Anthropol 2006;129:121–31.
[41] Yang P, Yang YH, Zhou BB, Zomaya AY. A review of ensemble methods in bioinformatics. Curr Bioinform 2010;5:296–308.
[42] Breiman L. Bagging predictors. Mach Learn 1996;26:123–40.
[43] Freund Y, Schapire R. Experiments with a new boosting algorithm. In: Proceedings of the thirteenth national conference on machine learning; 1996; p. 148–56.
[44] Breiman L. Random forests. Mach Learn 2001;45:5–32.
[45] Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 1975;405:442–51.
[46] Altman DG, Bland JM. Diagnostic tests 2: predictive values. Br Med J 1994;309:102.
[47] Spackman Kent A. Signal detection theory: valuable tools for evaluating inductive learning. In: Proceedings of the sixth international workshop on machine learning; 1989; p. 160–3.