Available online at www.scholarsresearchlibrary.com

Scholars research library

Archives of Applied Science Research, 2011, 3 (6):392-401
(http://scholarsresearchlibrary.com/archive.html)

ISSN 0975-508X
CODEN (USA) AASRC9

Subcellular Localization of Proteins

Anubha Dubey* and Usha Chouhan**

**Department of Bioinformatics, MANIT, Bhopal*
***Department of Mathematics, MANIT, Bhopal*

ABSTRACT

Subcellular location prediction of proteins is an important and well-studied problem in bioinformatics. This is a problem of predicting which part in a cell a given protein is transported to, where an amino acid sequence of the protein is given as an input. This problem is becoming more important since information on subcellular location is helpful for annotation of proteins and genes and the number of complete genomes is rapidly increasing. Since existing predictors are based on various heuristics, it is important to develop a simple method with high prediction accuracies. Support vector machines play an important role in developing models to predict higher accuracies with different parameters of protein.

INTRODUCTION

The prediction of protein subcellular localization (PSL) focuses on determining localization sites of unknown proteins in a cell. The study of PSL is important for elucidating protein functions involved in various cellular processes. Despite recent technical advances, experimental determination of PSL remains time-consuming and labor-intensive. In addition, researches in the post-genomic era have yielded a tremendous amount of sequence data. Given the size and complexity of the data, many researchers would prefer to use prediction systems to identify and screen possible candidates for further analyses. Hence, computational approaches have become increasingly important.

Previous works
Extensive studies of PSL prediction have led to the development of several methods, which can be classified as follows.
1. *Amino acid composition-based methods* These methods utilize machine learning techniques, including neural networks [1] and support vector machines (SVM) [2-8]. This category includes methods like P-CLASSIFIER [6] and CELLO [7,8], which utilize *n*-peptide composition-based SVM approaches.

2. *Methods that integrate various protein characteristics* Several methods including expert systems [9,10], *k*-nearest neighbor [11-13], SVM [14-16], support vector data description [17], and Bayesian networks [18-21], integrate various biological features that influence localization. The features that characterize a protein can be extracted from biological literature, public databases, and related prediction systems. Both PSORTb [18,19] and PSLpred [14] integrate different analytical modules and demonstrate that the hybrid approaches perform better than each individual module.

3. *Sequence homology-based methods* It has been suggested that PSL is an evolutionary conserved trait [20,21]. Efforts to address the relationship between evolutionary information and localization identity have relied heavily on exploiting sequence similarity to infer PSL. Such methods include phylogenetic profiling [22], domain projection [23], and a sequence homology-based method [7]. Several other methods, such as PSORTb and PSLpred, also incorporate such sequence homology-based components in their analyses.

General biological features

1. *Amino acid (AA) composition*: Protein descriptors based on *n*-peptide compositions or their variations have proved effective in PSL prediction [8]. If *n* = 1, then the *n*-peptide composition reduces to amino acid composition, which generates a 21 dimensional feature vector (i.e., 20 amino acid types plus a symbol 'X', for others) that represents the occurrence frequency of amino acids in a protein sequence.
2. *Di-peptide (DP) composition*: Similar to amino acid composition, if *n* = 2, the di-peptide composition gives a fixed length of 21 × 21 di-peptides, which represent the occurrence frequency of amino acid pairs in a protein sequence.
3. Relative solvent accessibility (RSA): Proteins in different compartments have various buried and exposed residue compositions [24,25]. For example, CP proteins have a balance of acidic and basic surface residues, while EC proteins have a slight excess of acidic surface residues [26]. We use amino acid compositions of both buried and exposed residues, with a cutoff of 25% [27], to represent the results derived by SABLE II [28], a relative solvent accessibility prediction method.

4. *Secondary structure elements encoding scheme* 1 (SSE1): Transmembrane *a*-helices are frequently observed in IM proteins, while transmembrane *β*-barrels are primarily found in OM proteins [29]. Secondary structure elements (SSE) are crucial for detecting proteins localized in the IM and OM. We compute the amino acid compositions of three SSEs [15,38], *a*-helix, *β*-strand, and loop, based on the predictions of HYPROSP II [27], a knowledge-based SSE prediction approach.

5. *Secondary structure elements encoding scheme* 2 (SSE2): SSE1 alone cannot discriminate proteins that share similar SSE compositions and localize in different compartments. For example, the SSE compositions of OM proteins might be similar to proteins localized in other compartments, but OM proteins are characterized by *β*-strand repeats throughout the transmembrane domains. To further depict such properties in a protein, three descriptors, composition, transition, and distribution, are used to encode predictions of HYPROSP II. Composition describes the global composition of a given SSE type in a protein. Transition characterizes the percentage frequency that amino acids of a particular SSE type are followed by a different type. Distribution measures the chain length within which the first, 25, 50, 75 and 100% of the amino acids of a particular SSE type are location [28].

Compartment-specific biological features

1. *Signal peptides (SIG):* Signal peptides are N-terminal peptides, typically between 15 and 40 amino acids long, which target proteins for translocation through the general secretory pathway [1]. The presence of a signal peptide suggests that the protein does not reside in the CP and several prediction methods have been developed [27-29]. We employ SignalP 3.0 [27], a neural network- and hidden Markov model-based method, to predict the presence and location of signal peptide cleavage sites.

2. *Transmembrane* a-*helices (TMA):* Integral IM proteins are characterized by *a*-helices, typically 20–25 amino acids in length, which traverse the IM. The presence of one or more transmembrane *a*-helices implies that the protein is located in the IM. We apply TMHMM 2.0 [30], a hidden Markov model-based method, to identify potential transmembrane *a*-helices.

3. *Twin-arginine translocase (TAT) motifs:* The twin-arginine translocase system exports proteins from the CP to the PP. The proteins translocated by twin-arginine translocase bear a unique twin-arginine motif [31], the presence of which is a useful feature for distinguishing between PP and non-PP proteins. We use TatP 1.0 [32], a neural network-based method, to predict the presence of twin-arginine translocase motifs.

4. *Transmembrane β-barrels (TMB):* A large number of proteins residing in the OM are characterized by *β*-barrel structures; thus, they could be candidate features for detecting OM proteins. We adopt TMB-Hunt [33], a method that uses a *k*-nearest neighbor algorithm, to distinguish between transmembrane *β*-barrels and non-transmembrane *β*-barrels.

5. *Non-classical protein secretion (SEC):* For a long time, it was believed that an N-terminal signal peptide was absolutely necessary to export a protein to the extracellular space. However, recent studies have shown that several EC proteins can be secreted without a classical N-terminal signal peptide [34]. Identification of non-classical protein secretion could be a potential discriminator for CP and EC proteins. Predictions from SecretomeP 2.0 [35], a non-classical protein secretion prediction method, are incorporated in our method.

Sequence and structure conservation:

Because PSL tends to be evolutionary conserved, the known localization sites of homologous sequences could be useful indicators of the actual localization of an unknown protein. We apply both sequence and structural homology approaches to infer localization. For the sequence homology approach, we develop a prediction method, called PSLseq, which is based on pairwise sequence alignment of ClustalW. In the structural homology approach, we employ secondary structural similarity comparison, referred to as PSLsse. Based on secondary structure elements predicted by HYPROSP II, we use SSEA to perform pairwise secondary structure alignment. In the sequence and structural homology approaches, the known localization of the top-rank aligned protein is assigned to the query protein as its predicted localization.

Knowing the subcellular location of proteins is important for understanding their functions. Many methods have been described to predict subcellular location from sequence information. However, most of these methods either rely on global sequence properties or use a set of known protein targeting motifs to predict protein localization. Here we develop and test a novel method that identifies potential targeting motifs using a discriminative approach based on Hidden Markov models (discriminative HMMs). These models search for motifs that are present in a compartment but absent in other, nearby, compartments by utilizing an hierarchical structure that mimics the protein sorting mechanism. We show that both discriminative motif finding and the hierarchical structure improves localization prediction on a benchmark dataset of yeast proteins. The motifs identified can be mapped to known targeting motifs and they are more conserved

than the average protein sequence. Using our motif-based predictions we can identify what we believe are annotation errors in public databases for the location of some of the proteins.
The predictions methods are described below:

1. Target P
Predicts the subcellular location of eukaryotic proteins (use results of ChloroP and SignalP). Secretory signal peptides, mitochondrial targeting peptides and chloroplast transit peptides in eukaryotes. http://www.cbs.dtu.dk/services/TargetP

2. WolfPSort
Protein Subcellular Localization Prediction (plant, animal, fungi).http://wolfpsort.org/

3.PSORTb: Bacterial protein subcellular localization prediction.http://www.psort.org/psortb/

4. SecretomeP
The SecretomeP 2.0 server produces ab initio predictions of non-classical i.e. not signal peptide triggered protein secretion. The method queries a large number of other feature prediction servers to obtain information on various post-translational and localizational aspects of the protein, which are integrated into the final secretion prediction.
http://www.cbs.dtu.dk/services/SecretomeP/

5.LOCtree is a novel system of support vector machines (SVMs) that predict the subcellular localization of proteins, and DNA-binding propensity for nuclear proteins, by incorporating a hierarchical ontology of localization classes modeled onto biological processing pathways.
http://cubic.bioc.columbia.edu/services/loctree/

6.Twin-arginine translocation signal peptides in bacteria.http://www.cbs.dtu.dk/services/TatP/

7.BaCelLo is a predictor for the subcellular localization of proteins in eukaryotes. It is based on a decision tree of several support vector machines (SVMs), it classifies up to four localizations for Fungi and Metazoan proteins and five localizations for Plant ones
http://gpcr.biocomp.unibo.it/bacello/

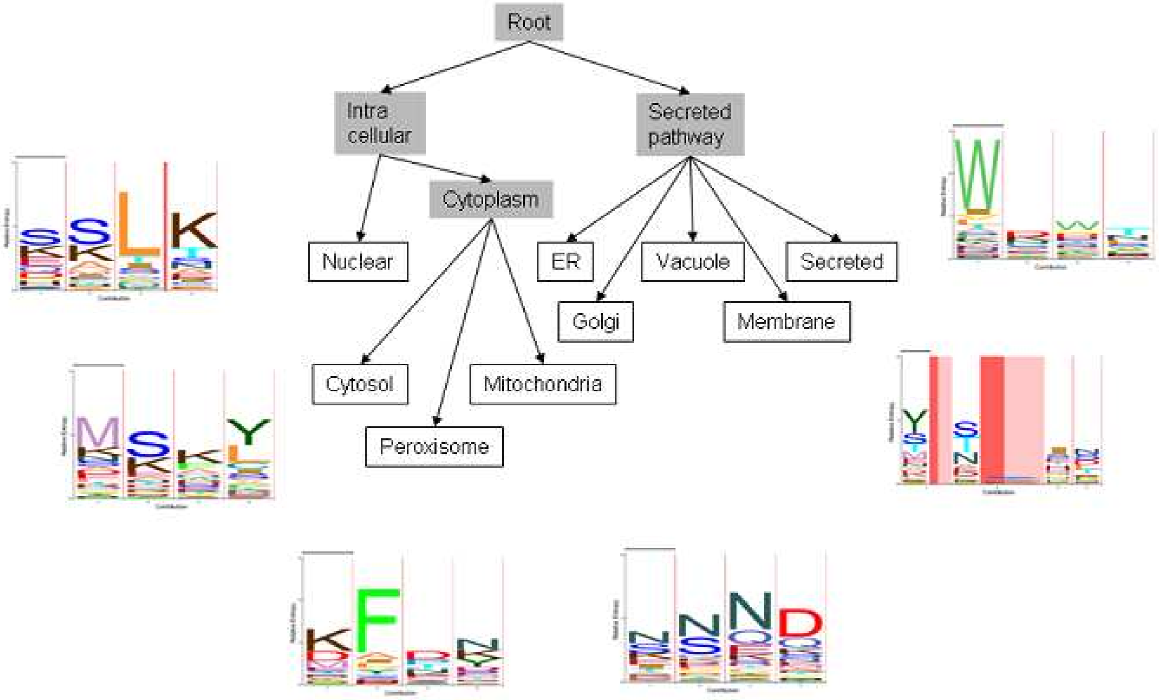
8. Protein Prowler
Subcellular Localisation Predictor (locations in plants and in other eukaryotes (TargetP data))..http://pprowler.imb.uq.edu.au/

9. CELLO is a multi-class SVM classification system. http://cello.life.nctu.edu.tw/

10.PA SUB: The Proteome Analyst Specialized Subcellular Localization Server (PA-SUB) is part of Proteome Analyst (PA). PA is a web server built to predict protein properties, such as general function, in a high-throughput fashion. PA-SUB is specialized to predict the subcellular localization of proteins using established machine learning techniques. http://www.cs.ualberta.ca/~bioinfo/PA/Sub/

11.Multiloc: Subcellular location in plants, other eukaryotes, fungi.
http://www-bs.informatik.uni-tuebingen.de/Services/MultiLoc/

12. PSLpred**:** PSLpred is a SVM based method to predict 5 major subcelullar localization (cytoplam, inner-membrane, outermembrane, extracellular, and periplasm) of Gram-negative bacteria. http://www.imtech.res.in/raghava/pslpred/



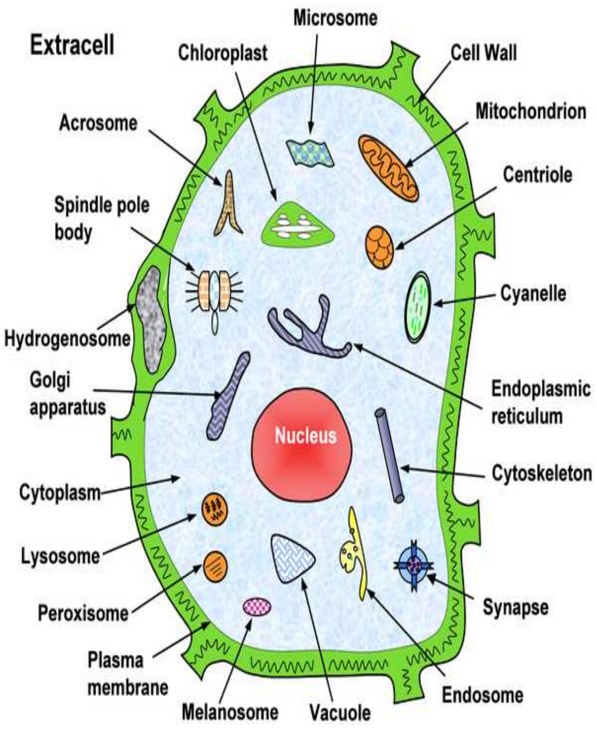


Figure 2: Illustration to show the 22 subcellular locations of eukaryotic proteins.

The 22 location sites are: (1) acrosome, (2) cell wall, (3) centriole, (4) chloroplast, (5) cyanelle, (6) cytoplasm, (7) cytoskeleton, (8) endoplasmic reticulum, (9) endosome, (10) extracell, (11) Golgi apparatus, (12) hydrogenosome, (13) lysosome, (14) melanosome, (15) microsome (16) mitochondria, (17) nucleus, (18) peroxisome, (19) plasma membrane, (20) plastid, (21) spindle pole body, and (22) vacuole [17].

13.pTARGET: It is a computational method to predict the subcellular localization of only eukaryotic proteins from animal species that include fungi and metazoans. Predictions are carried out based on the occurrence patterns of protein functional domains and the amino acid compositional differences in proteins from different subcellular locations. This method can predict proteins targeted to nine distinct subcellular locations that include cytoplasm, endoplasmic reticulum, extracellular/secreted, Golgi, lysosomes, mitochondria, nucleus, peroxysomes and plasma membrane. http://bioapps.rit.albany.edu/pTARGET//

14. Subloc**:** SubLoc is a prediction system for protein subcellular localization based on amino acid composition alone. http://www.bioinfo.tsinghua.edu.cn/SubLoc/
Figure 1 shows different subcellular localization positions of protein in Cell.

Information of subcellular locations of proteins is important for in-depth studies of cell biology. It is very useful for proteomics, system biology and drug development as well. However, most existing methods for predicting protein subcellular location can only cover 5 to 12 location sites. Also, they are limited to deal with single-location proteins and hence failed to work for multiplex proteins, which can simultaneously exist at, or move between, two or more location sites. Actually, multiplex proteins of this kind usually posses some important biological functions worthy of our special notice. A new predictor called "**Euk-mPLoc 2.0**" is developed by hybridizing the gene ontology information, functional domain information, and sequential evolutionary information through three different modes of pseudo amino acid composition. It can be used to identify eukaryotic proteins among the following 22 locations (as shown in Figure 2): (1) acrosome, (2) cell wall, (3) centriole, (4) chloroplast, (5) cyanelle, (6) cytoplasm, (7) cytoskeleton, (8) endoplasmic reticulum, (9) endosome, (10) extracell, (11) Golgi apparatus, (12) hydrogenosome, (13) lysosome, (14) melanosome, (15) microsome (16) mitochondria, (17) nucleus, (18) peroxisome, (19) plasma membrane, (20) plastid, (21) spindle pole body, and (22) vacuole. Compared with the existing methods for predicting eukaryotic protein subcellular localization, the new predictor is much more powerful and flexible, particularly in dealing with proteins with multiple locations and proteins without available accession numbers. For a newly-constructed stringent benchmark dataset which contains both single- and multiple-location proteins and in which none of proteins has pairwise sequence identity to any other in a same location, the overall jackknife success rate achieved by Euk-mPLoc 2.0 is more than 24% higher than those by any of the existing predictors. As a user-friendly web-server, Euk-mPLoc 2.0 is freely accessible at http://www.csbio.sjtu.edu.cn/bioinf/euk-multi-2/. For a query protein sequence of 400 amino acids, it will take about 15 seconds for the web-server to yield the predicted result; the longer the sequence is, the more time it may usually need. It is anticipated that the novel approach and the powerful predictor as presented in this paper will have a significant impact to Molecular Cell Biology, System Biology, Proteomics, Bioinformatics, and Drug Development.

Proteins are sorted into different cellular compartments such as cytoplasm, nuclear region, mitochondrion, etc. or may be secreted out of the cell, and their proper functioning relies on this precise process of subcellular localization. Hence, subcellular location information may imply the function. DBSubLoc (Guo et al, 2004) is a protein subcellular localization annotation database, which is available at http://www.bioinfo.tsinghua.edu.cn/dbsubloc.html. The database contains >60□000 protein sequences from virus, bacteria, fungi, plant and animal. However, its service is via WWW and the user interface is built on web browsers, which is designed to be accessed by humans, not by machines. Thus, it is troublesome for users to use DBSubLoc in an automated manner.

METHODOLOGY: Support vector machine (SVM) has been used to predict the subcellular location of eukaryotic proteins from their different features such as amino acid composition, dipeptide composition and physico-chemical properties. The SVM module based on dipeptide composition performed better than the SVM modules based on amino acid composition or physico-chemical properties. In addition, PSI-BLAST was also used to search the query sequence against the dataset of proteins (experimentally annotated proteins) to predict its subcellular location.

Evaluation of different prediction softwares:

The performance modules constructed in this study were evaluated using a 5-fold cross-validation technique. In the 5-fold cross-validation, the relevant dataset was partitioned randomly into five equally sized sets. The training and testing was carried out five times, each time using one distinct set for testing and the remaining four sets for training. For evaluating the performance of various modules, accuracy and Matthew's correlation coefficient (MCC) were calculated using the following equations: The fractions of all 20 natural amino acids were calculated by using Equation 1,

$$= \frac{Total\ Number\ of\ amino\ acid\ i}{Total\ number\ of\ amino\ acids\ in\ a\ protein} \qquad (1)$$

$$Accuracy(x) = \frac{tp + tn}{tp + tn + fp + fn} \qquad (2)$$

where x can be any subcellular location (nuclear, cytoplasm, extracellular and mitochondria), exp(x) is the number of sequences observed in location x, $p(x)$ is the number of correctly predicted sequences of location x, $n(x)$ is the number of correctly predicted sequences not of location x, $u(x)$ is the number of under-predicted sequences and $o(x)$ is the number of over-predicted sequences.

Support vector machine

SVMs are universal approximators based on statistical and optimization theory. The SVM is particularly attractive to biological sequence analysis due to its ability to handle noise, large dataset and large input spaces [38]. Further details about the SVM can be obtained from Vapnik's papers [39] or http://www.imtech.res.in/raghava/eslpred/algo.html. In the present study, we have used SVM_light to predict the subcellular localization of proteins. This software is freely downloadable from http://www.cs.cornell.edu/People/tj/svm_light/. The software enables the users to define a number of parameters and also allows a choice of inbuilt kernel function, including linear, RBF and Polynomial. The parameters except kernel functions and regulatory parameters C were kept constant during the training. The prediction of subcellular localization is a multi-class classification problem. We developed a series of binary classifiers to handle the multi-classification problem. We constructed *N* SVMs for *N*-class classification. Here, the class number was equal to four for eukaryotic sequences. The *i*th SVM was trained with all samples in the *i*th class with positive labels and all other samples with negative labels. In this way, four SVMs were constructed for subcellular localization of protein to nuclear, cytoplasm, extracellular and mitochondria. An unknown sample was classified into the class that corresponded to the SVM with highest output score.

Protein features

Amino acid composition

Amino acid composition is the fraction of each amino acid in a protein. The fraction of all 20 natural amino acids was calculated using the following equation:

$$= \frac{\text{Total Number of amino acid } i}{\text{Total number of amino acids in a protein}} \qquad (3)$$

where *i* can be any amino acid.

Composition of physico-chemical properties

The 33 physico-chemical properties were used to represent proteins as shown in Table S1 of the supplementary material [40]. The values of each physio-chemical property for all 20 amino acids were normalized between 0 and 1 using the standard conversion formula. The input vector has 33 scalar values, each representing the average value of a distinct physico-chemical property of protein.

Dipeptide composition

Dipeptide composition was used to encapsulate the global information about each protein sequence, which gives a fixed pattern length of 400 (20 x 20). This representation encompassed the information about amino acid composition along local order of amino acid. The fraction of each dipeptide was calculated using following equation:

$$\text{Fraction of dipeptide} = \frac{\text{total number of dipeptide (i)}}{\text{Total no. Of all possible dipeptides}} \qquad (4)$$

where dep(*i*) is one out of 400 dipeptides.

DISCUSSION

In general, artificial intelligence (AI) based techniques such as SVMs and neural networks are elegant approaches for the extraction of complex patterns from biological sequence data. These techniques are highly successful for residue state prediction where fixed window/pattern length is used [42]. The major limitation of the AI techniques is that they need patterns/input units of fixed length. This is the major reason for the failure of the AI techniques in the classification of proteins (e.g. subcellular localization prediction, fold recognition) because similar/homologous proteins often have variable length. In order to overcome this problem, a fixed-length pattern must be generated for proteins, for AI techniques to be implemented.

The percentage composition of amino acids, which gives a fixed pattern length of 20, is commonly used by AI techniques for the classification of proteins. This strategy has been used previously for developing the method for subcellular localization prediction of eukaryotic and prokaryotic proteins [47,48]. However, this approach provides information only about the amino acid frequency, but no information about the local order of amino acids [43]. To provide the information about frequency and local order of amino acids, dipeptide composition (instead of amino acid composition) can be used as the input unit to AI techniques. Dipeptide composition gives a fixed pattern length of 400. Dipeptide composition is widely used in the development of methods for fold prediction [44]. The prediction accuracy of the dipeptide composition-based method should be higher than that of amino acid composition based methods [45]. More information about the protein sequence can be encapsulated using tripeptide composition.

Tripeptide composition gives a fixed pattern length of 8000, which is commonly used in similarity searching in BLAST and FASTA (41,46). In the case of tripeptide composition, ANN and SVM are unable to handle the noise due to the large number of input units and number of missing tripeptides in a protein. The physico-chemical properties of a protein are yet another alternative way to provide the global information of a protein in the form of fixed pattern length. The prediction accuracy can further be improved, by devising methodologies to encapsulate more comprehensive information of a protein. A SVM-based module (hybrid) was constructed on the basis of comprehensive information about proteins including amino acid composition, physico-chemical properties, dipeptide composition and PSI-BLAST results. The hybrid module predicted the subcellular localization of a protein more accurately than the rest of the modules developed in this study. These results confirmed that the approach is capable of capturing more information about a protein that is crucial for detecting subcellular localization of proteins. Thus, providing more comprehensive information can be useful in enhancing the prediction accuracy of fold or tertiary structure prediction methods.

In conclusion, a new method for subcellular localization of a eukaryotic protein is presented. This method will nicely complement the existing subcellular localization prediction methods. It will assist in assigning the subcellular location or function of proteins more reliably. The authors believe that the prediction method presented here would be useful for the annotation of the piled-up genomic data.

REFERENCES

[1] Emanuelsson O, Nielsen H, and Brunak S, von Heijne G: *J Mol Biol* **2000**, 300(4):1005-1016. Pub Med Abstract | Publisher Full Text

[2] 2.Hoglund A, Donnes P, Blum T, Adolph HW, Kohlbacher O: *Bioinformatics* **2006**, 22(10):1158-1165.

[3] Wang J, Sung WK, Krishnan A, Li KB: *BMC Bioinformatics* **2005**, 6:174.

[4] Yu CS, Chen YC, Lu CH, Hwang JK: *Proteins* **2006**, 64(3):643-651.

[5] Yu CS, Lin CJ, Hwang JK: *Protein Sci* 2004, 13(5):1402-1406.

[6] Bannai H, Tamada Y, Maruyama O, Nakai K, Miyano S: *Bioinformatics* **2002**, 18(2):298-305.

[7] Nakai K, Kanehisa M: *Proteins* **1991**, 11(2):95-110.

[8] Chou KC, Cai YD: *Bioinformatics* **2005**, 21(7):944-950.

[9] Horton P, Park KJ, and Obayashi T, Nakai K: Protein subcellular localization prediction with WoLF PSORT.*Proceedings of the 4th Annual Asia Pacific Bioinformatics Conference (APBC'06): 13–16 February* **2006***; Taipei, Taiwan* 2006, 39-48.

[10] Chou KC, Shen HB: *Journal of Proteome Research* **2006**, 5:1888-1897.

[11] Bhasin M, Garg A, and Raghava GP: *Bioinformatics* **2005**, 21(10):2522-2524.

[12] Nair R, Rost B: *J Mol Biol* **2005**, 348(1):85-100.

[13] Su CY, Lo A, Chiu HS, Sung TY, Hsu WL: Protein subcellular localization prediction based on compartment-specific biological features. *Proceedings of IEEE Computational Systems Bioinformatics Conference (CSB'06): 14–18 August 2006; Stanford, California* **2006** , 325-330.

[14] Lee K, Kim DW, Na D, and Lee KH, Lee D: *Nucleic Acids Res* **2006**, 34(17):4655-4666.

[15] Gardy JL, Laird MR, Chen F, Rey S, Walsh CJ, Ester M, and Brinkman FS: *Bioinformatics* **2005**, 21(5):617-623.

[16] Gardy JL, Spencer C, Wang K, Ester M, Tusnady GE, Simon I, Hua S, deFays K, Lambert C, Nakai K, *et al*.: *Nucleic Acids Res* **2003**, 31(13):3613-3617.

[17] Gardy JL, Brinkman FS: *Nat Rev Microbial* **2006**, 4(10):741-751.

[18] Nair R, Rost B: *Protein Sci* **2002**, 11(12):2836-2847.
[19] Marcotte EM, Xenarios I, van Der Bliek AM, Eisenberg D: *Proc Natl Acad Sci USA 2000*, 97(22):12115-12120.
[20] Mott R, Schultz J, Bork P, Ponting CP: *Genome Res* **2002**, 12(8):1168-1174.
[21] Nair R, Rost B: *Nucleic Acids Res* **2003**, 31(13):3337-3340.
[22] Nair R, Rost B: *Proteins* **2003**, 53(4):917-930.
[23] Andrade MA, O'Donoghue SI, Rost B: *J Mol Biol* **1998**, 276(2):517-525.
[24] Cheng BY, Carbonell JG, Klein-Seetharaman J: *Proteins* **2005**, 58(4):955-970.
[25] Adamczak R, Porollo A, Meller J: *Proteins* **2005**, 59(3):467-475.
[26] Pautsch A, Schulz GE, *Nat Struct Biol* **1998** , 5(11):1013-1017.
[27] Lin HN, Chang JM, Wu KP, Sung TY, Hsu WL: *Bioinformatics* **2005**, 21(15):3227-3233.
[28] Bendtsen JD, Nielsen H, von Heijne G, Brunak S: *J Mol Biol* **2004**, 340(4):783-795.
[29] Chou KC, Shen HB: Signal-CF: *Biochem Biophys Res Comm* **2007**, 357:633-640.
[30] Zhang Z, Henzel WJ: *Protein Science* **2004**, 13:2819-2824.
[31] Krogh A, Larsson B, von Heijne G, Sonnhammer EL: *J Mol Biol* **2001**, 305(3):567-580.
[32] Berks BC: *Mol Microbiol* **1996**, 22(3):393-404.
[33] Bendtsen JD, Nielsen H, Widdick D, Palmer T, Brunak S: *BMC Bioinformatics* **2005**, 6:167.
[34] Garrow AG, Agnew A, Westhead DR: *Nucleic Acids Res* **2005**, (33 Web Servers):W188-192.
[35] Nickel W: *Eur J Biochem* **2003**, 270(10):2109-2119.
[36] Bendtsen JD, Kiemer L, Fausboll A, Brunak S: *BMC Microbiol* **2005**, 5:58.
[37] Nair R, Rost B: *Proteins* **2003**, 53(4):917-930.
[38] Zavaljevski, N., Stevens,F.J. and Reifman,J. ((**2002**)) *Bioinformatics*, , 18, , 689–696.
[39] Joachims, T. (**1999**) Making large-scale SVM learning practical. In Scholkopf, B., Burges,C. and Smola,A. (eds), *Advances in Kernel Methods—Support Vector Learning*. MIT Press, Cambridge, MA, London, England.
[40] Bhasin, M. and Raghava, G.P.S. ((**2004**)) *Protein Sci.*, , 13, , 596–607.
[41] Altschul, S.F., Gish, W., Miller,W., Myers,E.W. and Lipman,D.J., ((**1990**)). *J. Mol. Biol.*, , 215, , 403–410.
[42] Krogh, A. and Riis, S.K. ((**1996**)) Prediction of ß sheets in protein. In Touretzky, D.S., Mozer,M.C., Hasaselmo, M.E. (eds), *Advances in Neural Information Processing System 8*. MIT Press, Cambridge, MA, pp. 917–923.
[43] Shepherd, A.J., Gorse,D. and Thornton,J.M. ((**2003**)) *Proteins*, , 50, , 290–302.
[44] Reczko,M. and Bohr, H. ((**1995**)) *Nucleic Acid Res.*, , 22, , 3616–3619.
[45] Grassmann,J., Reczko,M., Suhai,S. and Edler,L. ((**1999**)) Protein fold class prediction: new methods of statistical classification. In Lengauer,T., Schneider,R., Bork,P., Brutlag,D.L., Glasgow,J.I., Mewes,H.-W. and Zimmer,R. (eds), *Proceedings of Seventh International Conference on Intelligent System for Molecular Biology* (ISMB'99), AAAI, Heidelberg, Germany, pp. 106–112.
[46] Pearson, W.R. and Lipman, D.J. ((**1988**)). *Proc. Natl Acad. Sci., USA*,, 85, , 2444–2448.
[47] 47.Hua,S. and Sun,Z. ((2001)) *Bioinformatics*, , 17, , 721–728.
[48] 48. Reinhardt,A. and Hubbard,T. ((**1998**)) *Nucleic Acids Res.*, , 26, , 2230–2236.