BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 14 2006, pages 1784–1785
doi:10.1093/bioinformatics/btl180

Sequence analysis

TSSub: eukaryotic protein subcellular localization by extracting features from profiles

Jian Guo* and Yuanlie Lin

Laboratory of Statistical Computation & Bioinformatics, Department of Mathematical Sciences, Tsinghua University, Beijing 100084, China

Received on April 4, 2006; revised on April 4, 2006; accepted on May 4, 2006
Advance Access publication June 20, 2006
Associate Editor: Charlie Hodgman

ABSTRACT

Summary: This paper introduces a new subcellular localization system (TSSub) for eukaryotic proteins. This system extracts features from both profiles and amino acid sequences. Four different features are extracted from profiles by four probabilistic neural network (PNN) classifiers, respectively (the amino acid composition from whole profiles; the amino acid composition from the N-terminus of profiles; the dipeptide composition from whole profiles and the amino acid composition from fragments of profiles). In addition, a support vector machine (SVM) classifier is added to implement the residue-couple feature extracted from amino acid sequences. The results from the five classifiers are fused by an additional SVM classifier. The overall accuracies of this TSSub reach 93.0 and 77.4% on Reinhardt and Hubbard's eukaryotic protein dataset and Huang and Li's eukaryotic protein dataset, respectively. The comparison with existing methods results shows TSSub provides better prediction performance than existing methods.

Availability: The web server is available from http://166.111.24.5/webtools/TSSub/index.html

Contact: guojian99@tsinghua.org.cn

Supplementary Information: The Supplementary Data can be downloaded from http://166.111.24.5/webtools/TSSub/Supplementary.htm

INTRODUCTION

Subcellular location is a key functional characteristic of the proteins. As a complementary part of the time-consuming experimental techniques, a number of *in silico* subcellular localization methods have been introduced. These methods can be grouped into three categories. The first category is based on the existence of the sorting signals (Nakai, 2000), which include signal peptides, mitochondrial targeting peptides and chloroplast transit peptides (Emanuelsson *et al*., 2000; Nielsen *et al*., 1997, 1999). Nevertheless, the performances of these methods are highly dependent on the quality of the N-terminal sequence assignment (Hua and Sun, 2001). The second category study the whole sequence information such as the amino acid composition (Nakashima and Nishikawa, 1994; Cedano *et al*., 1997; Reinhardt and Hubbard, 1998; Chou and Elord, 1999; Hua and Sun, 2001; Guo *et al*., 2004), dipeptide (Yuan, 1999; Huang and Li, 2004), high rank 2-tuple composition (Guo *et al*., 2005; Park and Kanehisa, 2004) and so on. The third category fuses the results from different modules each of which extract a particular feature from the protein. They integrate signal peptide information or whole sequence information with other features such as physical and chemical properties (Chou, 2001; Feng and Zhang, 2001), protein domain information (Chou and Cai, 2002, 2003a, 2003b, 2004) and *n*-gram (Yu *et al*., 2004) etc. PSORT-B (Gardy *et al*., 2003) integrates the amino acid composition, the similarity to proteins of known localization, the presence of signal peptides and the transmembrane alpha-helices and motifs corresponding to specific localizations; Bhasin and Raghava (2004, 2005) and Garg *et al*. (2005) fuse amino acid composition, composition of physico-chemical properties, dipeptide composition, residue couples and EuPSI-BLAST. This paper introduced a new integrative method, TSSub (Tsinghua subcellular localization software), for eukaryotic protein subcellular localization. This system extracts features from both profiles and amino acid sequences in order to improve the prediction performance.

METHODS

The profiles are generated by PSI-BLAST program (Altschul *et al*., 1997). Each sequence is used as a seed to search the SWISSPROT 46.0 database to generate two profiles: position-specific scoring matrix (PSSM) and position-specific frequency matrix (PSFM). Both PSSM and PSFM are matrices with 20 rows and L columns, where L is the length of the query sequence. The elements of PSSM represents the log-likelihood of the residue substitutions at all positions in the template (query sequence) while PSFM contains both the sequence-weighted observed frequency as well as the pseudo-counts derived from the substitution matrix.

Four schemes are used to extract different features from the profiles. The first scheme extracts the amino acid composition from profiles. Specifically, the mean value along the 20 rows of PSSM is obtained as the feature vectors. The second scheme extracts amino acid composition from both the whole profile and the N-terminus of PSSM. The third scheme extracts dipeptide composition from PSFM, which can be regarded as an extension of the residue-couple model (Guo *et al*., 2005). The fourth scheme divides PSSM into several fragments along the rows and extracts the amino acid compositions of these profile fragments. To further improve the performance of TSSub, an additional scheme employs the residue-couple model (Guo *et al*., 2005) to extract the distribution of amino acid pairs from the amino acid sequence. The feature vectors of scheme 1 to scheme 4 are trained by probabilistic neural network (PNN) classifiers (Specht, 1990) and that of the last scheme is trained by support vector machine (SVM) classifier (Vapnik, 1995, 1998). The outputs from the five classifiers are fused by an additional SVM classifier. Readers can refer to the Supplementary Data for more details.

*To whom correspondence should be addressed.

Table 1. Comparison of TSSub (SVM fusion) with four existing methods on Reinhardt and Hubbard's eukaryotic dataset

Location		NNPSL	Subloc	FKNN	ESLpred	TSSub
Cyt	Acc%	55	76.9	86.7	85.2	89.7
	MCC	—	0.64	0.76	0.79	0.87
Ext	Acc%	75	80.0	83.7	88.9	94.2
	MCC	—	0.78	0.87	0.91	0.95
Mit	Acc%	61	56.7	60.4	68.2	85.1
	MCC	—	0.58	0.63	0.69	0.86
Nuc	Acc%	72	87.4	92.0	95.3	96.9
	MCC	—	0.75	0.83	0.87	0.92
OA%		66	79.4	85.2	88.0	93.0

RESULTS

The performance of TSSub was tested on Reinhardt and Hubbard's eukaryotic protein dataset (Reinhardt and Hubbard, 1998). The results of TSSub are compared with NNPSL (Reinhardt and Hubbard, 1998), Subloc (Hua and Sun, 2001), FKNN (Huang and Li, 2004) and ESLpred (Bhasin and Raghava, 2004) on the same dataset (Table 1). The overall accuracy of TSSub reaches 93.0%, which is 27.0, 13.6, 7.8 and 5.0% higher than NNPSL, Subloc, FKNN and ESLpred, respectively. If we focus on Mitochondria, the accuracy of TSSub achieved 85.1%, which is 24.1, 28.4, 24.7 and 16.9% higher than NNPSL, Subloc, FKNN and ESLpred, respectively.

The results of NNPSL and ESLpred were derived with 10- and 5-fold cross validation, respectively, and the results of other three methods were derived with leave-one-out cross validation. OA: overall accuracy; Acc: accuracy in a specific location; Cyt: cytoplasm; Ext: extracellular; Mit: mitochondria; Nuc: nuclear.

In addition, TSSub was also compared with the Huang and Li's (2004) FKNN method on their dataset with 11 subcellular locations. The overall accuracy of TSSub reaches 77.4%, which is 19.3% higher than that of FKNN.

In conclusion, a new method for eukaryotic protein subcellular localization has been introduced. This method integrates the prediction from five schemes, four of which extracts features from the profiles (PSSM and PSFM) and the rest of which extracts features from the amino acid sequences. The outputs from the five schemes are fused by an SVM classifier. By comparing with the prediction results from other methods, TSSub exhibits a better performance. We hope TSSub can become a complementary method with the existing experimental and computational methods for protein subcellular localization.

WEBSERVER

The webserver TSSub have been developed and can be accessed from: http://166.111.24.5/webtools/TSSub/index.htm.

ACKNOWLEDGEMENTS

The author would thank Prof. Xiangjun Liu who provided the high-performance workstation. We also thank Dr A. Reinhardt and Dr Y. Huang for sharing their eukaryotic protein datasets. This work was supported by Human Liver Proteome Project (2004BA711A21) and The National Nature Science Foundation of China (10371063).

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.

Bhasin,M. and Raghava,G.P.S. (2004) ESLpred: SVM based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Res*., **32**, W414–W419.

Bhasin,M. *et al*. (2005) PSLpred: prediction of subcellular localization of bacterial proteins. *Bioinformatics*, **21**, 2522–2524.

Chou,K.C. and Elord,D. (1999) Protein subcellular location prediction. *Protein Eng*., **12**, 107–118.

Chou,K.C. (2001) Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. *Biochem. Biophys. Res. Commun*., **278**, 477–483.

Chou,K.C. and Cai,Y.D. (2002) Using functional domain composition and support vector machines for prediction of protein subcellular location. *J. Biol. Chem*., **277**, 45765–45769.

Chou,K.C. and Cai,Y.D. (2003a) A new hybrid approach to predict subcellular localization of proteins by incorporating Gene ontology. *Biochem. Biophys. Res. Commun*., **311**, 743–747.

Chou,K.C. and Cai,Y.D. (2003b) Prediction and classification of protein subcellular location-sequence-order effect and pseudo amino acid composition. [Addendum (2004) *J. Cell. Biochem*., **91**, 1085] *J. Cell. Biochem*., **90**, 1250–1260.

Chou,K.C. and Cai,Y.D. (2004) Prediction of protein subcellular locations by GO-FunD-PseAA predicor. *Biochem. Biophys. Res. Commun*., **320**, 1236–1239.

Cedano,J. *et al*. (1997) Relation between amino acid composition and cellular location of proteins. *J. Mol. Biol*., **266**, 594–600.

Emanuelsson,O. *et al*. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol*., **300**, 1005–1016.

Feng,Z. and Zhang,C.T. (2001) Prediction of the subcellular location of prokaryotic proteins based on the hydrophobicity index of amino acids. *Int. J. biol. macromol*., **28**, 225–261.

Garg,A. *et al*. (2005) SVM-based method for subcellular localization of human proteins using amino acid compositions, their order and similarity search. *J. Biol. Chem*., **280**, 14427–14432.

Guo,J. *et al*. (2004) A novel method for protein subcellular localization based on boosting and probabilistic neural network. *Proc. APBC*, **2004**, 21–27.

Guo,J. *et al*. (2005) A novel method for protein subcellular localization: combining residue-couple model and SVM. *Proc. APBC*, **2005**, 117–129.

Hua,S.J. and Sun,Z.R. (2001) Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, **17**, 721–728.

Huang,Y. and Li,Y.D. (2004) Prediction of protein subcellular locations using fuzzy k-NN method. *Bioinformatics*, **20**, 21–28.

Gardy,J.L. *et al*. (2003) PSORT-B: improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acids Res*., **31**, 3613–3617.

Nakai,K. (2000) Protein sorting signals and prediction of subcellular localization. *Adv. Protein Chem*., **54**, 277–344.

Nakashima,H. and Nishikawa,K. (1994) Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. *J. Mol. Biol*., **238**, 54–61.

Nielsen,H. *et al*. (1997) A neural network method for identification of prokaryotic and eukaryotic signal perptides and prediction of their cleavage sites. *Int. J. Neural Sys*., **8**, 581–599.

Nielsen,H. *et al*. (1999) Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Eng*., **12**, 3–9.

Park,K.J. and Kanehisa,M. (2004) Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics*, **19**, 1656–1663.

Reinhardt,A. and Hubbard,T. (1998) Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Res*., **26**, 2230–2236.

Specht,D.F. (1990) Probabilistic Neural Networks. *Neural Networks*, **3**, 109–118.

Vapnik,V. (1995) *The Nature of Statistical Learning Theory*. Springer-Verlag, NY.

Vapnik,V. (1998) *Statistical Learning Theory*. John Wiley and Sons, Inc., NY.

Yu,C.S. *et al*. (2004) Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. *Protein Sci*., **13**, 1402–1406.

Yuan,Z. (1999) Prediction of protein subcellular locations using Markov chain models. *FEBS Lett*., **451**, 23–26.