Amino Acids (2010) 38:347–350
DOI 10.1007/s00726-009-0238-7

SHORT COMMUNICATION

Predicting protein subcellular location: exploiting amino acid based sequence of feature spaces and fusion of diverse classifiers

Asifullah Khan · Abdul Majid · Tae-Sun Choi

Received: 27 October 2008 / Accepted: 12 January 2009 / Published online: 7 February 2009

Abstract A novel approach *CE-Ploc* is proposed for predicting protein subcellular locations by exploiting diversity both in feature and decision spaces. The diversity in a sequence of feature spaces is exploited using hydrophobicity and hydrophilicity of amphiphilic pseudo amino acid composition and a specific learning mechanism. Diversity in learning mechanisms is exploited by fusion of classifiers that are based on different learning mechanisms. Significant improvement in prediction performance is observed using jackknife and independent dataset tests.

Keywords Amphiphilic pseudo amino acid composition · Protein subcellular location · Ensemble classifier

Introduction

Considering the fact that the number of newly found proteins is consistently increasing; with an increase of more than 50 times in the last two decades, the importance of automatically annotating the subcellular attributes of uncharacterized proteins and their timely utilization in drug discovery is self evident. Various attempts are made for predicting protein subcellular localization consisting of individual (Shi et al. 2007) and ensemble classifiers (Chou and Shen 2006; Nanni and Lumini 2007; Shen and Chou 2007a, b, c; Shen et al. 2007; Shen and Burger 2007; Yu et al. 2004). Other attractive illustrations of ensembles are proposed in (Nanni and Lumini 2006; Nanni and Lumini 2008b; Nanni and Lumini 2008c). Support vector machines (SVM) based ensembles are that of *CELLO* (Yu et al. 2004) and the combination that exploits different physicochemical properties (Nanni and Lumini 2008a).

However, most of the above mentioned ensemble systems do not combine classifiers that are individually trained on different feature spaces and possess different learning mechanisms. In contrast, our fusion of classifier scheme possesses both these attributes and thus is more effective compared to individual classifiers and those ensemble classification systems that either exploit diversity in feature or decision space. Mathematical description for developing *CE-Ploc* is provided. Our results show that the performance of loc-predictors can be enhanced by the exploitation of both the available feature and decision spaces.

Materials and methods

We test the performance of the proposed *CE* ensemble using the same training and testing data as investigated in (Chou and Shen 2006). They reported that most sequences have quite low sequence identity for both the training (3,799) and testing (4,498) datasets indicating exclusion of

Electronic supplementary material The online version of this article (doi:10.1007/s00726-009-0238-7) contains supplementary material, which is available to authorized users.

A. Khan · A. Majid · T.-S. Choi (✉)
Department of Mechatronics,
Gwangju Institute of Science and Technology,
1 Oryong-Dong, Buk-Gu, Gwangju 500-712,
Republic of Korea
e-mail: tschoi@gist.ac.kr

A. Khan
e-mail: asifullah@gist.ac.kr

A. Khan · A. Majid
Department of Information and Computer Sciences,
Pakistan Institute of Engineering and Applied Sciences,
Nilore, Islamabad, Pakistan

redundant and homologous sequences (Supplementary Table 1).

The basic architecture of the proposed *CE-PLoc* scheme is shown in Supplementary Fig. 1. We have used four different learning mechanisms; Nearest Neighbor (NN), Probabilistic Neural Network (PNN), SVM and Covariant Discriminant Classifier (CDC) to develop the base classifiers (Supplementary material). Individual ensemble (IE) classifier is produced by using a fixed learning mechanism which exploits diversity in feature spaces. This step is repeated by changing the learning mechanism. In the second stage, combined ensemble (CE) is developed by fusing either all individual classifiers (entire-pool voting) or the IE ensembles (sub-pool voting), thus exploiting diversity in decision spaces. These decision spaces are created by using amphiphilic pseudo amino acid composition (PseAA) with dimension $\Phi = 20 + 2(i - 1)$, where $i = 1, 2, \ldots, \xi$. Here $\xi = 22$ represents the number of individual classifiers used for developing IE.

Let us denote a feature vector for an input protein as **P**, the protein dataset by Q, the predicted results by classifier as R. A query protein out of the N proteins can then be assigned to any of the subcellular locations

$$Q = Q_1 \cup Q_2 \cup Q_3 \cup Q_4 \cup \ldots \cup Q_V \tag{1}$$

where V is the number of classes. A kth subcellular protein feature vector from class category v can be expressed as

$$\mathbf{P}_v^k = \left[p_{v,1}^k p_{v,2}^k \ldots p_{v,20}^k \ldots p_{v,\Phi}^k\right]^{\mathbf{T}} \tag{2}$$

where $p_{v,1}$, $p_{v,2}$, …, $p_{v,20}$ are the frequencies of occurrence of 20 amino acids, while the elements $p_{v,21}$, $p_{v,22}$, …, $p_{v,(\xi-1)}$ are the first-tier to Φ-tier correlation factors of an amino acid sequence in the protein chain based on two indices of hydrophobicity and hydrophilicity. First-tier correlation factors represent the sequence order correlation between all the first most contiguous residues along a protein chain, while $(\xi - 1)$-tier represent the same between all the $(\xi - 1)$ most contiguous residues (Chou and Shen 2006).

Development of IE

The ensemble IE has ξ classifiers and thus their individual predictions can be expressed as

$$\{R_1, R_2, R_3, \ldots, R_\xi\} \in \{Q_1, Q_2, Q_3, \ldots, Q_V\} \tag{3}$$

Now the IE-based voting mechanism for the kth protein feature vector can be formulated as

$$Z_j^{\mathrm{IE}} = \sum_{i=1}^{\xi} w_i \Delta(R_i, Q_j), \quad j = 1, 2, \ldots V \tag{4}$$

where w_i represents weight factor and $\Delta(R_i, Q_j) = \begin{cases} 1, & \text{if } R_i \in Q_j \\ 0, & \text{otherwise} \end{cases}$.

Finally, the class of the query protein is assigned by IE to the class γ that obtains maximum votes:

$$Z_\gamma^{\mathrm{IE}} = \mathrm{Max}\left\{Z_1^{\mathrm{IE}}, Z_2^{\mathrm{IE}}, \ldots, Z_V^{\mathrm{IE}}\right\} \tag{5}$$

Development of CE

We first describe the entire-pool voting scheme for developing CE. Let $l = 1, 2, 3, \ldots, L$ represents the number of different base learners. We compute the votes for each class as

$$Z_j^{\mathrm{CE}} = \sum_{i=1}^{L*\xi} w_i \Delta(R_i, Q_j), \quad j = 1, 2, \ldots, V \tag{6}$$

The predicted class τ is decided by using the Max function

$$Z_\tau^{\mathrm{CE}} = \mathrm{Max}\left\{Z_1^{\mathrm{CE}}, Z_2^{\mathrm{CE}}, \ldots, Z_V^{\mathrm{CE}}\right\} \tag{7}$$

Similarly, we also perform the fusion of the IE classifiers based on sub-pool voting mechanism with the aim of exploiting the variation in the decision spaces of the IE classifiers. Using Eqs. (4) and (8), respectively, we thus generate and combine IE ensembles to develop CE

$$Z_j^{\mathrm{CE}} = \sum_{l=1}^{L} w_l Z_j^{\mathrm{IE}_l}, \qquad j = 1, 2, \ldots, V \tag{8}$$

In this work, all weight factors are set to unity. However, the performance of the proposed CE could be further improved, if an intelligent weight optimization strategy is employed. Besides accuracy, MCC, Q-statistics (Kuncheva and Whitaker 2003), sensitivity, and specificity measures are also used to analyze the prediction performance of the loc-predictors (Supplementary material).

Results and discussion

First we discuss the performance of classifiers based on a specific learning mechanism. The performance of the proposed CE ensembles is then analyzed. It is found that the performance of CE is better using entire-pool voting as against sub-pool voting. Perhaps in case of sub-pool voting, the selection of optimal weights is more important than that of entire-pool voting. Therefore, results of CE are reported using entire-pool voting strategy.

Table 1 shows that in case of jackknife test, $\mathrm{IE}^{\mathrm{NN}}$ achieves an overall accuracy of 78.55% for the 14 subcellular locations. Thus, an improvement of 8.35 and 7.4%, respectively, is obtained compared with those of the

Table 1 Classification performance for the 14 subcellular locations of proteins

Type of classifier	Data sampling method					
	Jackknife test			Independent dataset test		
	Correct predictions	Accuracy %	Avg. Q statistics	Correct predictions	Accuracy %	Avg. Q statistics
CDC ensemble (Chou and Shen 2006)	2,666	70.20	–	3,331	74.10	–
Weighted CDC ensemble (Khan et al. 2007)	2,704	71.15	–	3,377	75.07	–
IE^{SVM}	3,072	80.86	0.928	1,950	43.35	0.936
IE^{CDC}	2,694	70.91	0.926	3,340	74.25	0.984
IE^{NN}	2,984	78.55	0.935	3,774	83.90	0.988
IE^{PNN}	2,969	78.15	0.936	3,747	83.30	0.988
Proposed CE (entire-pool)[a]	3,132	82.44	0.804	3,918	87.11	0.959
Proposed CE*(entire-pool)	3,191	83.99	0.818	3,930	87.33	0.881

[a] In addition to NN, PNN, and CDC, SVM is also used as a base classifier. However, in case of independent dataset, only the first three SVMs are included in CE*. The parameters of SVM are optimized using grid search (Supplementary material)

IE^{CDC} (Chou and Shen 2006; Khan et al. 2007). Similarly, the improvement in case IE^{PNN} is 7.95 and 7.0%. The performance of IE^{SVM} using jackknife test is the highest among all the IEs.

In case of independent dataset, the results in Table 1 show that IE^{NN} outperforms IE^{CDC} by 9.8% and weighted IE^{CDC} by 8.83% in overall accuracy. The improvement in case of IE^{PNN} is also considerable. However, in case of independent dataset, the performance of the one-versus-all strategy based SVM is not productive. This is because by increasing margin of separation, keeping in view the one-versus-all case for 14 classes, SVM may not always provide good results (Khan et al. 2008). Supplementary Tables 2–3 present classification results for the individual classifiers.

As regards jackknife test, CE attains an overall accuracy of 83.99% (Table 1). Performance of CE classifier is higher than the highest-performing IE classifier, i.e., IE^{SVM} (80.86%). This shows that by combining different classification approaches, we are able to exploit the diversity in learning mechanisms/decision spaces. In case of independent dataset test, the CE classifier achieves an overall accuracy of 87.11%. Q-statistics in Table 1 shows a partial diversity in the individual learners. Likewise, Supplementary Table 1 shows the MCC, sensitivity, and specificity based prediction performance.

Conclusion

This study validates that exploiting diversities both in feature and decision spaces improves not only the accuracy but also the generalization capability of a protein prediction system. Although we have used four types of base classifiers and one feature extraction strategy, by including various types of base classifiers and multiple feature extraction strategies, it is likely to improve the performance of the classification system.

Acknowledgments This work was supported by the Bio Imaging Research Center at Gwangju Institute of Science and Technology (GIST), South Korea.

References

Chou KC, Shen HB (2006) Predicting protein subcellular location by fusing multiple classifiers. J Cell Biochem 99:517–527. doi:10.1002/jcb.20879

Khan MF, Mujahid A, Khan A, Bangash A (2007) Prediction of protein sub-cellular localization through weighted combination of classifiers. International Conference on Electrical Engineering, ICEE

Khan A, Khan MF, Choi TS (2008) Proximity based GPCRs prediction in transform domain. Biochem Biophys Res Commun 371:411–415. doi:10.1016/j.bbrc.2008.04.074

Kuncheva LI, Whitaker CJ (2003) Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Mach Learn 51:181–207. doi:10.1023/A:1022859003006

Nanni L, Lumini A (2006) An ensemble of K-local hyperplane for predicting protein–protein interactions. Bioinformatics 22:1207–1210. doi:10.1093/bioinformatics/btl055

Nanni L, Lumini A (2007) Ensemblator: an ensemble of classifiers for reliable classification of biological data. Pattern Recognit Lett 28:622–630. doi:10.1016/j.patrec.2006.10.012

Nanni L, Lumini A (2008a) An ensemble of support vector machines for predicting the membrane protein type directly from the amino acid sequence. Amino Acids 35:573–580. doi:10.1007/s00726-008-0083-0

Nanni L, Lumini A (2008b) A genetic approach for building different alphabets for peptide and protein classification. BMC Bioinformatics 9:45. doi:10.1186/1471-2105-9-45

Nanni L, Lumini A (2008c) Using ensemble of classifiers in Bioinformatics. In: Vogel HPAM (ed) Machine learning research progress. Nova Science Publishers Inc, New York

Shen Y, Burger G (2007) 'Unite and conquer': enhanced prediction of protein subcellular localization by integrating multiple specialized tools. BMC Bioinformatics 8:420. doi:10.1186/1471-2105-8-420

Shen HB, Chou KC (2007a) Gpos-PLoc: an ensemble classifier for predicting subcellular localization of Gram-positive bacterial proteins. Protein Eng Des Sel 20:39–46. doi:10.1093/protein/gzl053

Shen HB, Chou KC (2007b) Hum-mPLoc: an ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. Biochem Biophys Res Commun 355:1006–1011. doi:10.1016/j.bbrc.2007.02.071

Shen HB, Chou KC (2007c) Virus-PLoc: a fusion classifier for predicting the subcellular localization of viral proteins within host and virus-infected cells. Biopolymers 85:233–240. doi:10.1002/bip.20640

Shen HB, Yang J, Chou KC (2007) Euk-PLoc: an ensemble classifier for large-scale eukaryotic protein subcellular location prediction. Amino Acids 33:57–67. doi:10.1007/s00726-006-0478-8

Shi JY, Zhang SW, Pan Q, Cheng YM, Xie J (2007) Prediction of protein subcellular localization by support vector machines using multi-scale energy and pseudo amino acid composition. Amino Acids 33:69–74. doi:10.1007/s00726-006-0475-y

Yu C-S, Lin C-J, Hwang J-K (2004) Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. Protein Sci 13:1402–1406. doi:10.1110/ps.03479604