Journal of Computational Biology and Bioinformatics Research Vol. 3(2), pp. 15-24, February 2011
Available online http://www.academicjournals.org/jcbbr
ISSN-2141-2227 ©2011 Academic Journals

Full Length Research Paper

Prediction of eukaryotic protein subcellular multi-localisation with a combined KNN-SVM ensemble classifier

Liqi Li[1], Hong Kuang[2], Yuan Zhang[1]*, Yue Zhou[1], Kaifa Wang[3] and Ying Wan[4]

[1]Department of Orthopaedics, Xinqiao Hospital, Third Military Medical University, Chongqing, China.
[2]Central Laboratory, 452nd Hospital of Chinese PLA, Chengdu, Sichuan, China.
[3]Department of Mathematics, Third Military Medical University, Chongqing, China.
[4]Department of Immunology, Third Military Medical University, Chongqing, China.

Accepted 15 December, 2010

Proteins may exist in or shift among two or more different subcellular locations, and this phenomenon is closely related to biological function. It is challenging to deal with multiple locations during eukaryotic protein subcellular localisation prediction with routine methods; therefore, a reliable and automatic ensemble classifier for protein subcellular localisation is needed. We propose a new ensemble classifier combined with the KNN (K-nearest neighbour) and SVM (support vector machine) algorithms to predict the subcellular localisation of eukaryotic proteins from the GO (gene ontology) annotations. This method was developed by fusing basic individual classifiers through a voting system. The overall prediction accuracies thus obtained via the jackknife test and resubstitution test were 70.5 and 77.6% for eukaryotic proteins respectively, which are significantly higher than other methods presented in the previous studies and reveal that our strategy better predicts eukaryotic protein subcellular localisation.

Key words: Gene ontology, multiple subcellular localisation, K-nearest neighbour, support vector machine, ensemble classifier.

INTRODUCTION

The amount of protein sequence data is increasing rapidly with the progression of genome projects. However, traditional experimental methods, including cell fractionation, electron microscopy and fluorescence microscopy, are time-consuming and expensive and cannot meet the research demands of the enormous amount of raw protein sequences (Lin et al., 2009; Xu et al., 2009). Thus, it is essential to find computational techniques to effectively analyse these data.

Because a protein needs to be transported to the correct cellular location to properly perform its functions, the prediction of protein subcellular localisation is an important aspect of protein bioinformatics and biofunctionality. Compared to experimental methods, computational prediction techniques can predict protein subcellular localisation more quickly and accurately. Moreover, it can effectively analyse proteome sequences on a large scale. Currently, many computational techniques, such as the neural network (Ma and Gu, 2010), support vector machine (SVM) (Shen and Burger, 2010) and hidden Markov models (HMM) (Rashid et al., 2007), have been developed for the prediction of protein subcellular localisation.

However, the neural network can suffer from multiple local minima (Marinov and Weeks, 2001), the solution to an SVM is unique and global. While the number of parameters that need to be evaluated in an HMM is large (Mount, 2009). In contrast to HMM, only the kernel function and the regularization parameter C are selected to specify one SVM (Hua and Sun, 2001). The reason

*Corresponding author: E-mail: qiangx163@yahoo.cn. Fax: +86-23-68755608.

Abbreviation: GO, Gene ontology; **KNN,** K-nearest neighbour; **SSL,** subset subcellular location; **SVM,** support vector machine.

that SVMs often outperform other computational techniques in practice is that SVMs are less prone to over fitting (Huang and Kecman, 2005). While the classification decision of KNN is based on a small neighbourhood of similar objects. The advantages of KNN are that the training is fast and it is well-suited for multi-modal classes (Wang and Yang, 2009). Therefore, SVM and KNN were introduced to predict eukaryotic protein sub cellular localisation. We developed a novel predictor named the KNN-SVM ensemble classifier, which is a combination of the KNN and SVM algorithms (Qiu et al., 2010; Zheng et al., 2009).

This predictor uses the gene ontology (Kim et al., 2010; Torto-Alalibo et al., 2010) database, which is based on the three related ontologies of molecular function, biological process, and cellular component, to improve prediction performance. Although there are three different types of gene ontology GO IDs for each sequence, we have not chosen only one type of GO ID. In this strategy, the GO numbers of the eukaryotic protein dataset covering 22 subcellular locations were extracted from the free GO database at ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/UNIPROT/ (released on March 15, 2010) and then were transformed into a 7166-dimension input vector (Chou and Shen, 2007). For the 22 subcellular locations in this paper, the one-versus-one method (Su et al., 2007) was employed to construct the classifiers. Thus, a total of 231 binary classifiers were prepared. In addition, either KNN or SVM, depending on which gave the higher accuracy rate, was used to predict the classifier of the 231 binary classifiers. As a result, a significant improvement in prediction quality was achieved, which makes the KNN-SVM ensemble classifier another powerful method for subcellular localisation prediction.

MATERIALS AND METHODS

Dataset

The free dataset we used was generated by Chou and Shen (Chou and Shen, 2007; Chou and Shen, 2008), which can be downloaded from http://www.csbio.sjtu.edu.cn/bioinf/euk-multi/Download.htm. The dataset contained 5618 eukaryotic protein sequences belonging to 22 location categories: 17 acrosome, 53 cell wall, 64 centriole, 501 chloroplast, 85 cyanelle, 1060 cytoplasm, 74 cytoskeleton, 364 endoplasmic reticulum, 89 endosome, 640 extracellular, 254 Golgi apparatus, 13 hydrogenosome, 80 lysosome, 13 melanosome, 31 microsome, 535 mitochondria, 1333 nucleus, 97 peroxisome, 725 plasma membrane, 15 synapse, 36 spindle pole body, and 102 vacuole (Figure 1). Although the dataset generated by Chou and Shen is 2 years old, it is a classical dataset cited by many related articles (Blum et al., 2009; Briesemeister et al., 2010; He and De Buck, 2010; He et al., 2010; Huang et al., 2008; Sharpe et al., 2010).

These sequences were extracted from SWISSPROT (version 50.7) and included only sequences that appeared to be complete and were annotated with reliable experimental observations. In this dataset, none of proteins had 25% or more sequence identity to any other protein in a same subcellular location, with the exception of three locations: acrosome, melanosome, and synapse. Otherwise, the numbers of proteins in the three locations would be insufficient to reach the statistical requirement. Because some proteins in this dataset could exist in or shift among two or more different subcellular locations, it was necessary to introduce the concept of a “locative protein”. If a protein exists in or shifts between two different subcellular locations, it will be counted as two locative proteins; likewise, if it exists in or shifts among three locations, three locative proteins will be counted. In this study, 5091 proteins belonged to 1 subcellular location, 495 to 2 locations, 28 to 3 locations, and 4 to 4 locations.

Gene ontology

Gene ontology, which is a controlled vocabulary, is used to describe the biology of a gene product in any organism. The gene ontology annotation is an effective protein descriptor and has been applied to a wide variety of biological sequence analyses (Rastogi and Rost, 2010). Moreover, most eukaryotic protein sequences in the UniProtKB/Swiss-Prot database have annotated GO terms. For example, the percentage of the 2423 proteins that were not annotated by GO terms was only 3.96% (Huang et al., 2009). Although, gene ontology annotation has been used for prediction of subnuclear localisation in many papers (Huang et al., 2009; Lei and Dai, 2006).

Gene ontology, which can effectively grasp the core features of proteins closely related to the subcellular localisation, is an effective and useful descriptor of eukaryotic proteins. Therefore, in this paper, the classifier is applied to proteins that have corresponding GO terms. While a small number of proteins without annotated GO terms can be predicted based on the existing sequenced-based prediction methods (Li and Li, 2008), we used the gene ontology annotations to improve the prediction quality for protein subcellular localisation. By mapping the 5618 eukaryotic protein entries to the GO database at ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/UNIPROT/ (released on March 15, 2010), one can obtain a list in which each eukaryotic protein entry corresponds to one or more GO numbers. For example, the eukaryotic protein entry “Q80UF4” corresponds to four GO numbers, that is, GO: 0005737, GO: 0005813, GO: 0005815 and GO: 0005856, while the eukaryotic protein entry “P40152” corresponds to three GO numbers, that is, GO: 0000324, GO: 0005773 and GO: 0016787. Another eukaryotic protein entry “Q5VT06” corresponds to five GO numbers, that is, GO: 0005634, GO: 0005737, GO: 0005815, GO: 0005819 and GO: 0005856. It is obvious that the total number of GO terms for the three eukaryotic protein entries described above is nine, that is, GO: 0005737, GO: 0005813, GO: 0005815, GO: 0005856, GO: 0000324, GO: 0005773, GO: 0016787, GO: 0005634 and GO: 0005819. Thus, we obtained the GO terms that are annotated for each eukaryotic protein from the GO database.

The total number of GO terms that appeared for the 5618 eukaryotic proteins was 7166. The simplest approach was to use a binary feature component for a protein, in which a value of 1 is used if the corresponding GO number appears or 0 if it does not appear. For example, the eukaryotic protein entry “Q80UF4” corresponds to four GO numbers, that is, GO: 0005737, GO: 0005813, GO: 0005815 and GO: 0005856, so the four corresponding components for the protein were assigned a value of 1 and the other 7162 with a value of 0. Thus, the GO terms annotated for each protein were transformed into a 7166-dimension input vector, in which the value of each element is 0 or 1. In other words, the input vectors for the 5618 eukaryotic proteins had equal lengths, and each protein entry corresponded to a 7166-dimension input vector. However, most of the features in the 7166-dimension input vectors remained null. Therefore, the key was to find an effective approach to incorporate the GO information into the prediction algorithm.

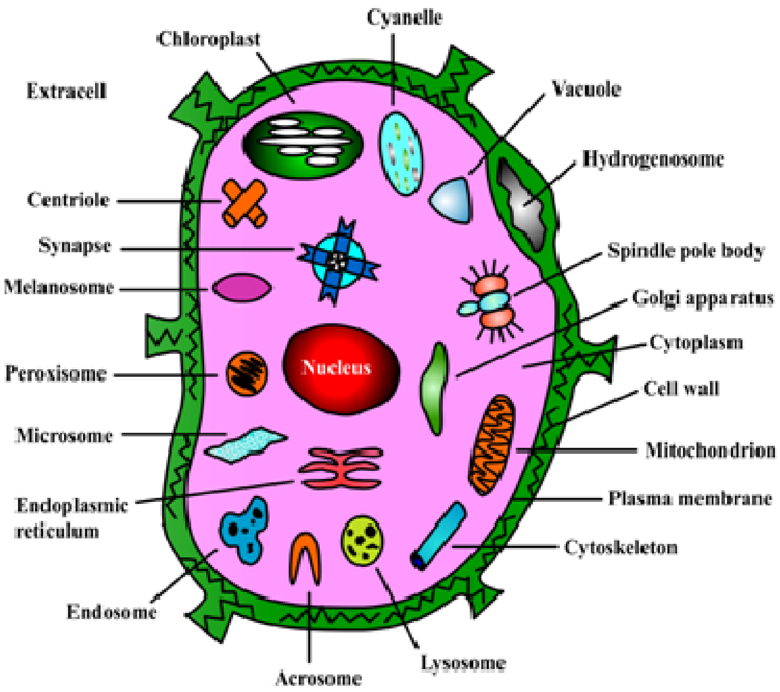

Figure 1. Schematic illustration of a eukaryotic cell containing the 22 subcellular locations. (1) acrosome, (2) cell wall, (3) centriole, (4) chloroplast, (5) cyanelle, (6) cytoplasm, (7) cytoskeleton, (8) endoplasmic reticulum, (9) endosome, (10) extracellular, (11) Golgi apparatus, (12) hydrogenosome, (13) lysosome, (14) melanosome, (15) microsome, (16) mitochondria, (17) nucleus, (18) peroxisome, (19) plasma membrane, (20) synapse, (21) spindle pole body, and (22) vacuole. Note that this model cell is not a specific eukaryotic cell. For example, "Chloroplast" proteins would not be included in the repertoire for animal proteins.

This was realised using the KNN-SVM ensemble classifier, in which the length of final input feature vector could be optimised. In this work, the multi-classification problem was solved by utilising a series of binary classifiers from KNN or SVM. Thus, the length of the final input feature vectors in every binary classifier was determined by the total number of GO terms that appeared for all of the proteins in the two categories. Let us assume that the numbers of proteins in two subsets (subcellular locations) are two and three respectively, and that the initial length of each input vector is nine. The two input vectors in subset 1 are and, while the three in subset 2 are $\{0,0,0,1,1,1,0,0,0\}$, $\{0,0,0,1,1,0,0,0,0\}$ and $\{0,0,0,1,0,1,0,0,0\}$. All five input vectors in the two subsets contain zero elements in the last three units, which remain null and lead to noise in the classifying protein entities into the two subsets. Therefore, the length of the final input vectors in the corresponding binary classifier is 6 = 9 – 3.

The KNN-SVM ensemble classifier

The key to the formulation of a powerful algorithm for predicting eukaryotic protein subcellular localisation is the selection of algorithms. A few parameters, such as the parameter of the kernel function in the SVM algorithm (Hua and Sun, 2001) and the parameter in the KNN algorithm (Wang and Yang, 2010), should be further optimised. Here, five-fold cross validation has been used to select algorithms and parameters for the limited computational power. In this technique, the dataset was divided randomly into five sets that consisted of nearly equally sized subsets. Subset 1, 2, 3, 4, and 5 contained 1124, 1124, 1124, 1124, and 1122 different proteins respectively. Subset 1 was firstly selected for testing, and the remaining four sets were used for training; then Subset 2 for testing, and the remaining four sets for training; and so Subset 3, 4, and 5.

This means that the data were further portioned into training and test datasets in five different ways. The training and testing was performed five times at a particular value of in the kernel function in the SVM algorithm and in the KNN algorithm. Each set was in turn selected for testing, and all rule parameters were calculated based on the remaining four sets. The overall performance was then obtained by averaging the performances of the five test sets. It is instructive to note that during the five-fold cross validation, each of the 5618 different proteins was selected only once for testing, although it may exist in or shift among two or more different sub cellular locations and correspond to several locative proteins. This means that the proteins used for training and those used for final evaluation could not overlap.

The prediction of eukaryotic protein subcellular localisation is a multi-classification problem. This problem can be solved by utilising a series of binary classifiers of KNN or SVM. In this work, we adopted the 'one-versus-one' method to transform this multi-classification problem into a two-class problem because it avoids the so-called 'false positive' problem in the 'one-versus-rest' method (Park and Kanehisa, 2003). For a k-class problem, classifiers were constructed by the 'one-versus-one' method:

$$n=\begin{cases} j-1, i=1 \\ \sum_{h=1}^{i-1}(k-h)+(j-i), i>1 \end{cases} \qquad (1)$$

The nth (Equation 1) classifier was trained by considering all proteins in the ith class as positive samples and all proteins in the jth class as negative samples. Here, $n=1,2,...,k\times(k-1)/2; i=1,...,k-1; j=i+1,...,k$. For the 22 subcellular locations in this paper, there were 231 (22×21/2=231) binary classifiers that needed to be constructed. To obtain a high overall prediction accuracy, the following scheme was utilised. For each binary classifier, the dataset was tested with the KNN and SVM methods. Different values for parameter k in the KNN method and the three common kernel (linear, polynomial and RBF) functions of SVM were chosen to test the dataset. It is instructive to note that the length of the final input feature vectors in every binary classifier is not related to the method that was finally chosen for the binary classifier. As described above, the length was determined only by the total number of GO terms that appeared for all proteins in the two categories. For example, the number of GO terms that appeared for all proteins in the two subsets 'acrosome' and 'cell wall' was 215, so the length of the input vectors in the corresponding binary classifier was 215. Then, the binary classifier was trained with the KNN and SVM methods. For the same binary classifier, we compared the results predicted from different methods, which were trained on the same length of input vectors. The best method, from which the highest prediction accuracy was obtained by five-fold cross validation, was chosen for each binary classifier.

The accuracy of binary classifier n can be represented by Equation (5). The process of the KNN-SVM ensemble classifier is described as follows. Suppose that the predicted classification results for the query protein **P** for the 231 binary classifiers are $R_1, R_2, ..., R_{231}$, respectively; that is:

$$\{R_1, R_2, ..., R_{231}\} \in \{S_1, S_2, ..., S_{22}\} \qquad (2)$$

where $S_1, S_2, ..., S_{22}$ represent the 22 subcellular locations. The voting score for the protein "P" belonging to class a is defined as

$$G_a=\sum_{p=1}^{231}\delta(R_p, S_a) \qquad (a=1,2,...,22) \qquad (3)$$

where the δ function in Equation (3) is given by

$$\delta(R_p, S_a)=\begin{cases} 1, R_p \in S_a \\ 0, R_p \notin S_a \end{cases} \qquad (4)$$

Subsequently, the query protein "P" will be assigned to the class that gives the highest score for Equation (3) of the 231 binary classifiers. Let us assume that there are four subsets and $6=4\times(4-1)/2$ classifiers are constructed. The predicted classification results for a query protein P with the six binary classifiers are R_1={1,0,0,0,},R_2={0,0,1,0},R_3={0,0,0,1},R_4={0,0,1,0},R_5={0,1,0,0},R_6 ={0,0,0,1}, respectively; that is, classifiers 1, 2, 3, 4, 5 and 6 assign protein P to subsets 1, 3, 4, 3, 2 and 4, respectively.

Accordingly, the voting scores for protein P are $G_1=1, G_2=1, G_3=2, G_4=2$. Therefore, protein P will be assigned to classes 3 and 4, which both give the highest score of $G_3=G_4=2$. In other words, according to the KNN-SVM ensemble classifier, it is a multiple-site protein and is predicted as belonging to both subsets 3 and 4. The KNN-SVM ensemble classifier was developed by fusing 231 binary classifiers through a voting system as described above. The 'engine' of each binary classifier was operated by the SVM or KNN rule, depending on which produced a higher accuracy rate. The software used to implement KNN and SVM was MATLAB R2009a, which can be downloaded from http://www.mathworks.com/ for academic purposes. Figure 2 shows the flow chart for application of KNN and SVM algorithms in MATLAB R2009a software.

Assessment of prediction performances

To measure the quality of the eukaryotic protein subcellular localisation prediction, it is convenient to introduce accuracy, overall accuracy and Matthew's Correlation Coefficient (MCC) (Restrepo-Montoya et al., 2009), which can be represented as

$$accuracy \quad (n)=\frac{p_n(i)+p_n(j)}{m(i)+m(j)} \qquad (5)$$

$$overall \quad accuracy = \frac{\sum_{a=1}^{k} TP_a}{N} \qquad (6)$$

$$MCC(a)=\frac{TP_a\times TN_a-FP_a\times FN_a}{\sqrt{(TP_a+FP_a)(TP_a+FN_a)(TN_a+FP_a)(TN_a+FN_a)}} \qquad (7)$$

where N is the total number of locative proteins, k is the class number, $m(i)$ and $m(j)$ are the numbers of the locative proteins in class i and class j, $p_n(i)$ and $p_n(j)$ are the numbers of the correctly predicted locative proteins of class i and class j by binary classifier n. TP_a, FP_a, TN_a, and FN_a are the number of true positives, false positives, true negatives, and false negatives in class a by the KNN-SVM ensemble classifier respectively.

RESULTS

Selection of kernel functions and parameters

We chose the three common kernel functions (RBF, linear and polynomial) of SVM to test the dataset. For the dataset used here, SVM with the linear and polynomial kernels was unable to unravel the overall protein localisation problem because the maximum numbers of iterations were exceeded. Thus, the RBF kernel function was selected to test the dataset. Figure 3 shows the prediction accuracies for 5618 eukaryotic proteins with different values of γ in the RBF kernel function by five-fold cross validation. Because preliminary tests with this

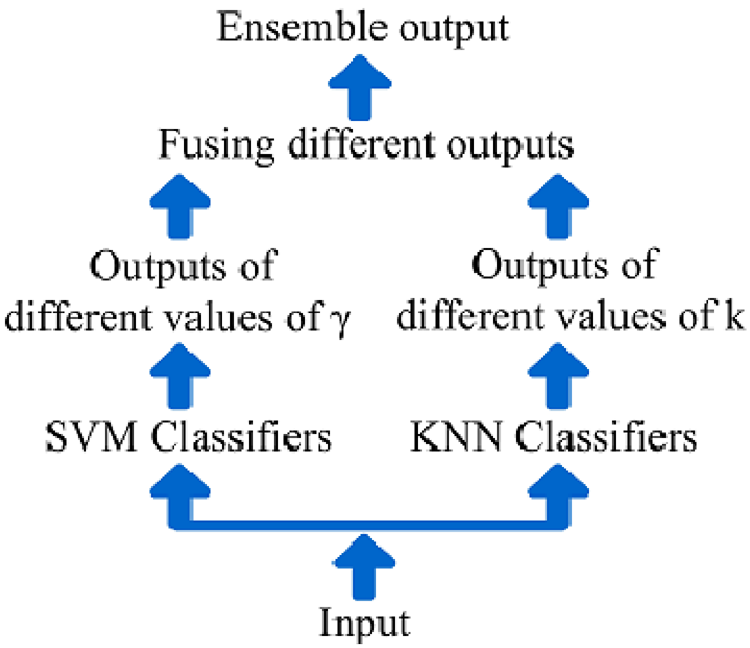

Figure 2. Flow chart for application of KNN and SVM algorithms.

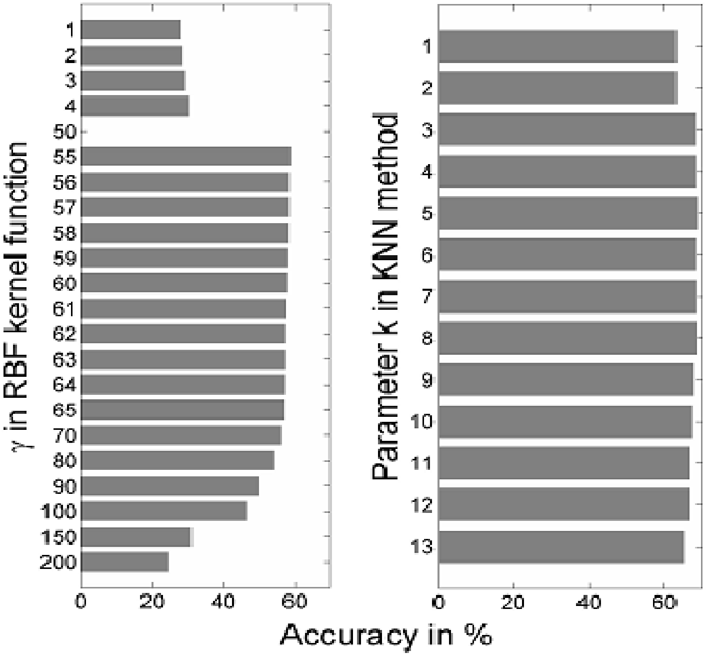

Figure 3. Prediction accuracies for 5618 eukaryotic proteins with different values for γ in the RBF kernel function and k in the KNN method by five-fold cross validation. For the dataset, this method was unable to solve the optimisation problem with the RBF kernel ($\gamma = 50$) and KNN method ($k = 14$).

dataset indicated that higher accuracies were obtained when γ in the SVM method (with the RBF kernel) and k in the KNN method were certain values, we optimised the sets as $\gamma \in \{55,56,57,58,59,60,61,62,63,64,65\}$ and

$k \in \{3,4,5,6,7,8\}$ (Figure 3).

Selection of prediction methods

Various approaches have been proposed for predicting protein subcellular localisation (Ma and Gu, 2010; Qiu et al., 2010; Shen and Chou, 2010; Wang and Yang, 2010). However, the existing predictors were established mostly based on a single theory, such as the neural network, KNN, SVM and Markov chain models. Obviously, the prediction performances of these predictors would be not desirable. Thus, we constructed the KNN-SVM ensemble classifier, which is based on two theories: the KNN and SVM algorithms. For the 22 subcellular locations in this work, the one-versus-one method was used. Thus, a total of 231 binary classifiers were constructed.

Although 5618 eukaryotic proteins in the dataset S corresponded to 7166 GO terms, the length of final input feature vector could be optimised. Only the GO terms that appeared for any of the proteins in the two categories were selected by the corresponding binary classifier. Either KNN or SVM, depending on which produced a higher accuracy rate, was then used to predict a classifier of the 231 binary classifiers. For example, the 82nd binary classifier was constructed using the SVM method with the RBF kernel ($\gamma = 57$) because the best accuracy of Equation (5) was obtained with the current method with the five-fold cross validation test, while the 99th binary classifier was constructed with the KNN method ($k = 4$) for the same reason.

Comparison with other methods

The prediction performance of the KNN-SVM ensemble classifier presented in this study was compared with that of other prediction methods. The dataset utilised here was also tested by Euk-mPloc (Chou and Shen, 2007). The current method was compared with the Euk-mPloc, KNN binary classifiers and SVM binary classifiers, and the results are listed in Table 1.

In statistical prediction, there are four methods that are often used for validation: the independent dataset test, the subsampling test, the jackknife test, and the resubstitution test (Chou and Shen, 2006). Of these, the jackknife test has been deemed the most rigorous and objective (Kandaswamy et al., 2010) and has been used increasingly by investigators for assessing the prediction performances of various methods (Cai et al., 2010; Kandaswamy et al., 2010; Qiu et al., 2010). In the jackknife test, each protein in the dataset was omitted as a query protein, and all of the rule parameters were obtained based on the remaining proteins. In this work, each of the 5618 different eukaryotic proteins was omitted only once for jackknife testing, although that protein may correspond to more than one locative protein. The resubstitution test is another important method which reflects the self-consistency of a classification method. In the resubstitution test, same proteins were used to construct the model and to test themselves. Although this test could give the higher accuracy, it represents the self-consistency of the identification method. For the ensemble classifier that we proposed, several results for the same dataset were compared, and the accuracies and the MCCs are also given in Table 1.

As shown in Table 1, the overall accuracy obtained by the current method with the jackknife test was 70.5%, which was nearly 10% higher than that of SVM (RBF kernel with) and 3.1% higher than that of Euk-mPloc. As compared with the SVM method, the accuracy of each subcellular location was improved significantly (except the cytoskeleton and nucleus). For Class 5, the predictive accuracy was even improved to 91.8%. In addition, the resubstitution test was performed with the KNN-SVM ensemble classifier on the dataset. The overall accuracy was improved to 77.6%. The accuracies and the MCCs were also improved. We could not compare our results with Euk-mPloc in detail because the previous study (Chou and Shen, 2007) did not show the accuracy of each subcellular location; however, the overall accuracy that we obtained was still higher than that of Euk-mPloc.

Table 2 shows a comparison of KNN-SVM ensemble classifier with other methods. Although the overall accuracies achieved by the KNN-SVM ensemble classifier were lower than those by OWFKNN (Nasibov and Kandemir-Cavas, 2008), PSP-WNN (Zou et al., 2007), and AdaBoost (Niu et al., 2008). It should be pointed out that the datasets used in OWFKNN, PSP-WNN, and AdaBoost all came from the Reinhardt and Hubbard database (Reinhardt and Hubbard, 1998), which indeed was 12 years old. Furthermore it contained homologous proteins with up to 90% sequence identity and covered only 4 subcellular location sites. The other dataset used in PSP-WNN was 7 years old and contained homologous proteins with up to 80% sequence identity. It covered only 12 subcellular location sites. These methods only covered a limited number of location sites and will fail to work if a query protein is outside their coverage. In contrast to these, none of proteins in the dataset has ≥ 25% sequence identity to any other in a same subcellular location. Although the dataset is 2 years old, it covers 22 location sites and is a classical dataset cited by many related articles (Blum et al., 2009; Briesemeister et al., 2010; He and De Buck, 2010; He et al., 2010; Huang et al., 2008; Sharpe et al., 2010; Zhou et al., 2008). We have also compared our method with Euk-mPLoc and Euk-mPLoc 2.0 (Chou and Shen, 2010) recently proposed by Chou. Although the KNN-SVM ensemble classifier and Euk-mPLoc were tested by the same dataset and the dataset used in Euk-mPLoc 2.0 contained more protein sequences, the overall accuracy

Table 1. Comparison of prediction performance for different methods on the dataset S .

Order	Subcellular location	Number of samples	Euk-mPloc		SVM (RBF kernel with $\gamma = 55$)		KNN-SVM ensemble classifier			
			Jackknife		Jackknife		Jackknife		Resubstitution	
			Accuracy (%)	MCC	Accuracy (%)	MCC	Accuracy (%)	MCC	Accuracy (%)	MCC
1	Acrosome	17	-	-	0	-	41.2	0.641	76.5	0.874
2	Cell wall	53	-	-	11.3	0.335	67.9	0.711	88.7	0.903
3	Centriole	64	-	-	15.6	0.393	62.5	0.690	81.3	0.786
4	Chloroplast	501	-	-	89.2	0.857	97.4	0.879	99.0	0.918
5	Cyanelle	85	-	-	0	-	91.8	0.957	91.8	0.957
6	Cytoplasm	1060	-	-	77.3	0.517	88.2	0.640	91.8	0.729
7	Cytoskeleton	74	-	-	27.0	0.517	24.3	0.491	41.9	0.645
8	Endoplasmic reticulum	364	-	-	41.8	0.564	79.7	0.776	86.8	0.839
9	Endosome	89	-	-	31.5	0.558	62.9	0.770	67.4	0.812
10	Golgi apparatus	254	-	-	32.7	0.529	74.0	0.802	79.5	0.828
11	Hydrogenosome	13	-	-	0	-	38.5	0.620	69.2	0.692
12	Lysosome	80	-	-	8.8	0.294	65.0	0.662	72.5	0.772
13	Melanosome	13	-	-	0	-	53.9	0.733	84.6	0.880
14	Microsome	31	-	-	9.7	0.310	19.4	0.380	41.9	0.647
15	Mitochondria	535	-	-	68.8	0.777	85.1	0.872	87.5	0.910
16	Nucleus	1333	-	-	93.0	0.759	84.6	0.824	85.7	0.862
17	Peroxisome	97	-	-	5.2	0.225	37.1	0.589	74.2	0.860
18	Plasma membrane	725	-	-	77.8	0.637	81.4	0.766	84.4	0.817
19	Extracell	640	-	-	80.8	0.789	83.3	0.864	85.9	0.894
20	Spindle pole body	36	-	-	44.4	0.666	50.0	0.669	75.0	0.850
21	Synapse	15	-	-	13.3	0.365	66.7	0.816	66.7	0.816
22	Vacuole	102	-	-	6.9	0.260	42.2	0.610	82.4	0.865
Overall	accuracy	-	67.4	-	61.9	-	70.5	-	77.6	-

that we obtained was still higher than that of Euk-mPloc and Euk-mPLoc 2.0.

All of the results indicated that the KNN-SVM ensemble classifier might be better than SVM and Euk-mPloc for the prediction of eukaryotic protein subcellular localisation. Table 3 shows the results predicted by the KNN-SVM ensemble classifier for the proteins that we investigated in previous studies (Zhang et al., 2010; Zhang et al., 2009) and the annotations for the corresponding GO numbers in the GO database.

Dealing with multiple locations

Most of the existing methods for predicting protein subcellular localisation are limited to a single location. It is instructive to note that the KNN-SVM ensemble classifier proposed here can be used effectively to deal with multiple locations as well. For multiple locations, the predicted result for query protein P may belong to one or more subcellular locations. In this work, it would be assigned to the class set A formed by:

Table 2. Comparisons with other methods.

Method	Input form	Number of subcellular locations	Test method	Overall accuracy (%)
OWFKNN	amino acidcomposition	4	Jackknife	86.2
PSP-WNN	Position-specific profiles	4	Jackknife	88.4
PSP-WNN	Position-specific profiles	12	five-fold cross validation	83.3
AdaBoost	amino acid composition	4	Jackknife	80.8
AdaBoost	amino acid composition	4	resubstitution test	100
Euk-mPLoc	gene ontology information	22	Jackknife	67.4
Euk-mPLoc 2.0	pseudo amino acid composition	22	Jackknife	64.2
KNN-SVM ensemble classifier	gene ontology information	22	Jackknife	70.5
KNN-SVM ensemble classifier	gene ontology information	22	Resubstitution	77.6

Table 3. Examples to show the predicted results by KNN-SVM ensemble classifier.

Accession number	Entry name	Swiss-Prot annotation	GO number	GO annotation	Identified location by KNN-SVM ensemble classifier
P55288	Cadherin11_mouse	Plasma membrane	0016021 0005509 0005515 0007156	integral to membrane calcium ion binding protein binding homophilic cell adhesion	Plasma membrane
P02751	Fibronectin_human	Extracell	0005793 0005577 0031093 0005578 0005518 0005201 0008201 0006953 0016477 0018149 0008360 0034446	ER-Golgi intermediate compartment fibrinogen complex platelet alpha granule lumen proteinaceous extracellularmatrix collagen binding extracellular matrix structural constituent heparin binding acute-phase response cell migration peptide cross-linking regulation of cell shape substrate adhesion-dependent cell spreading	Extracell

$$A = \{S_a\} \tag{8}$$

in which every S_a gives the highest score of Equaton (3) for the 231 binary classifiers. The number of elements in the set A may be one or more, meaning that the query protein P will be assigned to one or more subcellular locations. For example, the real subcellular locations to which the protein entry "P13395" belongs are $\{S_7, S_{10}, S_{18}\}$, and the predicted subcellular locations for "P13395" by the KNN-SVM ensemble classifier are $\{S_6, S_7, S_{10}\}$, in which S_6, S_7, S_{10} give the highest score ($G_6 = G_7 = G_{10} = 20$) for Equation (3). While the real subcellular location to which the protein entry "P31412" belongs is S_{22}, the predicted subcellular location is also S_{22} because only S_{22} gives the highest score ($G_{22} = 21$) for Equation (3).

DISCUSSION AND CONCLUSION

KNN and SVM are powerful statistical learning methods. Gene ontology, which effectively grasps the core features closely related to the subcellular localisation, is an effective and useful descriptor of eukaryotic proteins. In this work, an ensemble classifier was proposed for the prediction of eukaryotic protein subcellular localisation by coupling two powerful algorithms with the information derived from gene ontology. We also compared our method with other methods. Euk-mPloc and Euk-mPLoc 2.0 were formed by fusing KNN classifiers only. If newly found proteins cannot be classified accurately using the KNN classifier, both Euk-mPloc and Euk-mPLoc 2.0 will fail and we could use the SVM classifier, which is another powerful one.

In addition, the length of the final input feature vectors in every binary classifier was determined by the total number of GO terms that appeared for all of the proteins in the two categories. In other words, although 5618 eukaryotic proteins in the dataset S corresponded to 7166 GO terms, the GO terms that did not appear for all of the proteins in the two categories remained null and were not selected by the corresponding binary classifier. Although the length of final input feature vector was optimised in the KNN-SVM ensemble classifier, filtering approach like information gain could be tried to select top features in the future. These features can be biologically correlated with subcellular localisation and the KNN-SVM ensemble classifier will become even more powerful.

In comparison with previous predictors, significant improvement in prediction quality was achieved. In addition, according to biological experiments, more proteins will be found in multiple subcellular locations. The prediction of eukaryotic protein subcellular localisation by considering multiple location sites has been discussed recently (Chou and Shen, 2008). In this work, 231 binary classifiers were constructed. However, it should be pointed out that the numbers of proteins in 'acrosome', 'melanosome', and 'synapse' locations were not sufficiently large to train the classifiers in a more effective way. The corresponding classifiers might be biased. It is expected that the situation will be improved with more protein entries available for the three locations in the future. In this work, the KNN-SVM ensemble classifier and assessment of predictive performances for multiple-site proteins have been introduced, and the results indicated that the KNN-SVM ensemble classifier is a powerful tool for the prediction of eukaryotic protein subcellular localisation, especially for proteins with multiple locations.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (grant no. 30901512) and Research Project of Sichuan Provincial Health Department (grant no. 090178).

REFERENCES

Blum T, Briesemeister S, Kohlbacher O (2009). MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction. BMC Bioinformatics., 10: 274.

Briesemeister S, Rahnenfuhrer J, Kohlbacher O (2010). Going from where to why--interpretable prediction of protein subcellular localization. Bioinformatics, 26: 1232-1238.

Briesemeister S, Rahnenfuhrer J, Kohlbacher O (2010). YLoc--an interpretable web server for predicting subcellular localization. Nucleic Acids Res., 38: W497-502.

Cai Y, He J, Li X, Feng K, Lu L, Kong X, Lu W (2010). Predicting protein subcellular locations with feature selection and analysis. Protein Pept Lett., 17: 464-472.

Chou KC, Shen HB (2006). Predicting protein subcellular location by fusing multiple classifiers. J. Cell Biochem., 99: 517-527.

Chou KC, Shen HB (2007). Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. J. Proteome. Res., 6: 1728-1734.

Chou KC, Shen HB (2008). Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. Nat. Protoc., 3: 153-162.

Chou KC, Shen HB (2010). A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mPLoc 2.0. PLoS One 5: e9931.

He Z, De Buck J (2010). Localization of proteins in the cell wall of Mycobacterium avium subsp. paratuberculosis K10 by proteomic analysis. Proteome Sci., 8: 21.

He Z, Zhang J, Shi XH, Hu LL, Kong X, Cai YD, Chou KC (2010). Predicting drug-target interaction networks based on functional groups and biological features. PLoS One 5: e9603.

Hua S, Sun Z (2001). Support vector machine approach for protein subcellular localization prediction. Bioinformatics, 17: 721-728.

Huang TM, Kecman V (2005). Gene extraction for cancer diagnosis by support vector machines--an improvement. Artif. Intell. Med., 35: 185-194.

Huang WL, Tung CW, Ho SW, Hwang SF, Ho SY (2008). ProLoc-GO: utilizing informative Gene Ontology terms for sequence-based prediction of protein subcellular localization. BMC Bioinformatics, 9: 80.

Huang WL, Tung CW, Huang HL, Ho SY (2009). Predicting protein subnuclear localization using GO-amino-acid composition features. Biosystems, 98: 73-79.

Kandaswamy KK, Pugalenthi G, Moller S, Hartmann E, Kalies KU, Suganthan PN, Martinetz T (2010). Prediction of apoptosis protein locations with genetic algorithms and support vector machines through a new mode of pseudo amino Acid composition. Protein Pept. Lett., 17: 1473-1479.

Kim S, Min WK, Chun S, Lee W, Chung HJ, Choi SJ, Yang SE, Yang YS, Yoo JI (2010). Protein expression profiles during osteogenic differentiation of mesenchymal stem cells derived from human umbilical cord blood. Tohoku J. Exp. Med., 221: 141-150.

Lei Z, Dai Y (2006). Assessing protein similarity with Gene Ontology and its use in subnuclear localization prediction. BMC Bioinformatics, 7: 491.

Li FM, Li QZ (2008). Predicting protein subcellular location using Chou's pseudo amino acid composition and improved hybrid approach. Protein Pept. Lett., 15: 612-616.

Lin HN, Chen CT, Sung TY, Ho SY, Hsu WL (2009). Protein subcellular localization prediction of eukaryotes using a knowledge-based approach. BMC Bioinformatics, 10(15): S8.

Ma J, Gu H (2010). A novel method for predicting protein subcellular localization based on pseudo amino acid composition. BMB Rep., 43: 670-676.

Marinov M, Weeks DE (2001). The complexity of linkage analysis with neural networks. Hum. Hered., 51: 169-176.

Mount DW (2009). Using hidden Markov models to align multiple sequences. Cold Spring Harb Protoc 2009: pdb top41.

Nasibov E, Kandemir-Cavas C (2008). Protein subcellular location prediction using optimally weighted fuzzy k-NN algorithm. Comput. Biol., Chem., 32: 448-451.

Niu B, Jin YH, Feng KY, Lu WC, Cai YD, Li GZ (2008). Using AdaBoost for the prediction of subcellular location of prokaryotic and eukaryotic proteins. Mol. Divers., 12: 41-45.

Park KJ, Kanehisa M (2003). Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. Bioinformatics, 19: 1656-1663.

Qiu JD, Luo SH, Huang JH, Sun XY, Liang RP (2010). Predicting subcellular location of apoptosis proteins based on wavelet transform and support vector machine. Amino Acids, 38: 1201-1208.

Rashid M, Saha S, Raghava GP (2007). Support Vector Machine-based method for predicting subcellular localization of mycobacterial proteins using evolutionary information and motifs. BMC Bioinformatics, 8: 337.

Rastogi S, Rost B (2011). LocDB: experimental annotations of localization for Homo sapiens and Arabidopsis thaliana. Nucleic Acids Res., 35(1):230-234

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

Restrepo-Montoya D, Vizcaino C, Nino LF, Ocampo M, Patarroyo ME, Patarroyo MA (2009). Validating subcellular localization prediction tools with mycobacterial proteins. BMC Bioinformatics, 10: 134.

Sharpe HJ, Stevens TJ, Munro S (2010). A comprehensive comparison of transmembrane domains reveals organelle-specific properties. Cell, 142: 158-169.

Shen HB, Chou KC (2010). Virus-mPLoc: a fusion classifier for viral protein subcellular location prediction by incorporating multiple sites. J. Biomol. Struct. Dyn., 28: 175-186.

Shen YQ, Burger G (2010). TESTLoc: protein subcellular localization prediction from EST data. BMC Bioinformatics, 11: 563.

Su EC, Chiu HS, Lo A, Hwang JK, Sung TY, Hsu WL (2007). Protein subcellular localization prediction based on compartment-specific features and structure conservation. BMC Bioinformatics, 8: 330.

Torto-Alalibo T, Collmer CW, Gwinn-Giglio M, Lindeberg M, Meng S, Chibucos MC, Tseng TT, Lomax J, Biehl B, Ireland A, Bird D, Dean RA, Glasner JD, Perna N, Setubal JC, Collmer A, Tyler BM (2010). Unifying themes in microbial associations with animal and plant hosts described using the gene ontology. Microbiol. Mol. Biol. Rev., 74: 479-503.

Wang T, Yang J (2009). Using the nonlinear dimensionality reduction method for the prediction of subcellular localization of Gram-negative bacterial proteins. Mol. Divers., 13: 475-481.

Wang T, Yang J (2010). Predicting subcellular localization of gram-negative bacterial proteins by linear dimensionality reduction method. Protein Pept. Lett., 17: 32-37.

Xu Q, Hu DH, Xue H, Yu W, Yang Q (2009). Semi-supervised protein subcellular localization. BMC Bioinformatics, 10(1): S47.

Zhang Y, Xiang Q, Dong S, Li C, Zhou Y (2010). Fabrication and characterization of a recombinant fibronectin/cadherin bio-inspired ceramic surface and its influence on adhesion and ossification *in vitro*. Acta Biomater., 6: 776-785.

Zhang Y, Zhou Y, Zhu J, Dong S, Li C, Xiang Q (2009). Effect of a novel recombinant protein of fibronectinIII7-10/cadherin 11 EC1-2 on osteoblastic adhesion and differentiation. Biosci. Biotechnol. Biochem., 73: 1999-2006.

Zheng X, Liu T, Wang J (2009). A complexity-based method for predicting protein subcellular location. Amino Acids, 37: 427-433.

Zhou M, Boekhorst J, Francke C, Siezen RJ (2008). LocateP: genome-scale subcellular-location predictor for bacterial proteins. BMC Bioinformatics, 9: 173.

Zou L, Wang Z, Huang J (2007). Prediction of subcellular localization of eukaryotic proteins using position-specific profiles and neural network with weighted inputs. J. Genet. Genomics, 34: 1080-1087.