BIOINFORMATICS

Vol. 19 no. 13 2003, pages 1656–1663
DOI: 10.1093/bioinformatics/btg222

Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs

*Keun-Joon Park and Minoru Kanehisa**

Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan

Received on November 6, 2002; revised on January 24, 2003; accepted on March 17, 2003

ABSTRACT

Motivation: The subcellular location of a protein is closely correlated to its function. Thus, computational prediction of subcellular locations from the amino acid sequence information would help annotation and functional prediction of protein coding genes in complete genomes. We have developed a method based on support vector machines (SVMs).

Results: We considered 12 subcellular locations in eukaryotic cells: chloroplast, cytoplasm, cytoskeleton, endoplasmic reticulum, extracellular medium, Golgi apparatus, lysosome, mitochondrion, nucleus, peroxisome, plasma membrane, and vacuole. We constructed a data set of proteins with known locations from the SWISS-PROT database. A set of SVMs was trained to predict the subcellular location of a given protein based on its amino acid, amino acid pair, and gapped amino acid pair compositions. The predictors based on these different compositions were then combined using a voting scheme. Results obtained through 5-fold cross-validation tests showed an improvement in prediction accuracy over the algorithm based on the amino acid composition only. This prediction method is available via the Internet.

Availability: http://www.genome.ad.jp/SIT/ploc.html

Contact: kanehisa@kuicr.kyoto-u.ac.jp

Supplementary information: http://web.kuicr.kyoto-u.ac.jp/~park/Seqdata/

INTRODUCTION

Most of the proteins in a eukaryotic cell are synthesized in the cytoplasm. Newly synthesized proteins are targeted to the correct subcellular compartments and play their biological roles. Thus, computational methods for predicting protein subcellular locations are valuable tools for obtaining functional clues from the amino acid sequence information. Nakai and Kanehisa were the first to propose a computational method, named PSORT, based on sequence motifs and amino acid compositions reflecting sorting signals and other information (Nakai and Kanehisa, 1992; Nakai, 2000). They constructed a knowledge base by organizing experimental and computational observations as a collection of if–then rules. When sorting signals were not well characterized experimentally, various sequence features were computationally derived from a training data set.

Reinhardt and Hubbard (1998) used neural networks and showed that amino acid compositions alone contained information to distinguish proteins of different subcellular locations, although the method was not reliable enough for eukaryotic proteins. They stated that a method based on the amino acid composition would be more useful in practical applications, because automatically assigned protein sequences from genome projects are often unreliable for the 5′ regions. Chou and Elrod (1999a) constructed a data set of 12 subcellular locations, which accounted for most organelles and subcellular compartments in an animal or plant cell, and proposed a covariant discriminant algorithm to predict the subcellular location of a query protein from its amino acid composition. They also tried prediction of membrane protein types and subcellular locations with another membrane protein data set (Chou and Elrod, 1999b). Furthermore, Chou (2000, 2001) observed improvements of prediction accuracy when correlations of residue pairs were considered in addition to the amino acid composition. Yuan (1999) introduced a new method using Markov chain models. As for proteins from eukaryotic cells, the prediction rates were 73.0 and 78.7% corresponding to, respectively, four and three location categories. The four location categories were cytoplasmic, extracellular, nuclear and mitochondrial proteins, and the three categories contained the mixture of cytoplasmic and mitochondrial proteins. These accuracies were measured by the jack-knife (leave-one-out) test on the data set used by Reinhardt and Hubbard. Emanuelsson *et al*. (2000) proposed an integrated prediction method, TargetP, using neural networks based on individual sorting signal predictions. This method discriminated between proteins destined for mitochondrial, chloroplast, secretory pathway, and other localizations.

* To whom correspondence should be addressed.

Hua and Sun (2001a) constructed a subcellular localization prediction system using support vector machines (SVMs) based on amino acid compositions. They tested for four locations in eukaryotic cells: cytoplasmic, extracellular, mitochondrial, and nuclear. The prediction accuracy by the jack-knife test was 79.4% using the radial basis function (RBF) kernel. More recently, Cai *et al*. (2002) developed a new prediction method also using support vector machines. They used the data set of Chou and Elrod, and the total of 2191 proteins were classified into 12 groups. The prediction accuracy by the jack-knife test was 75%.

SVMs are a new generation of machine learning algorithms, which is gaining popularity in the analysis of biological problems such as gene and tissue classifications from microarray expression data (Brown *et al*., 2000; Furey *et al*., 2000), protein fold recognition (Ding and Dubchak, 2001), and protein secondary structure prediction (Hua and Sun, 2001b), as well as protein localization prediction mentioned above. Here we also use the SVM learning algorithm to extract sequence features from the training data set of proteins, whose subcellular locations are classified into 12 groups as in the work of Chou and Elrod (1999a). Specifically, we examine the validity of using different SVM kernel functions and parameters, and also using different sequence features represented by the compositions of amino acids, amino acid pairs, and gapped amino acid pairs.

DATA SET

Initial data collected from SWISS-PROT

All protein sequences were collected from the SWISS-PROT database (Bairoch and Apweiler, 2000) release 39.0. We identified eukaryotic proteins with specific subcellular locations according to the annotation information in the CC (comments or notes) and OC (organism classification) fields of SWISS-PROT. Table 1 summarizes the keywords that we used to search against the categorization of subcellular locations (-!- SUBCELLULAR LOCATION) in the CC field in order to collect proteins in 12 subcellular locations: chloroplast, cytoplasmic, cytoskeleton, endoplasmic reticulum, extracellular, Golgi apparatus, lysosomal, mitochondrial, nuclear, peroxisomal, plasma membrane and vacuolar proteins. When multiple keywords are shown in separate lines in Table 1, proteins that match any of the keywords were selected. We also checked the OC field to remove prokaryotic proteins. Proteins annotated with two or more subcellular locations were not included in the current data set. For example, a protein entry annotated with 'SUBCELLULAR LOCATION: NUCLEAR AND CYTOPLASMIC' in the CC field was not included. All protein entries computationally selected were then manually examined.

Table 1. Selection of proteins with known subcellular locations from SWISS-PROT

Subcellular location	Keywords
Chloroplast	Chloroplast
Cytoplasmic	Cytoplasmic
Cytoskeleton	Cytoskeleton
	Filament
	Microtubule
Endoplasmic reticulum	Endoplasmic reticulum
Extracellular	Extracellular
	Secreted
Golgi apparatus	Golgi
Lysosomal	Lysosomal
Mitochondrial	Mitochondrial
Nuclear	Nuclear
Peroxisomal	Peroxisomal
	Microsomes
	Glyoxysomal
	Glycosomal
Plasma membrane	Integral membrane
Vacuolar	Vacuolar
	Vacuole

Keywords were used to search against the CC field of the SWISS-PROT database.

Removal of highly similar sequences

Sequences with a high degree of similarity to other sequences were removed by all-to-all sequence similarity search using the program ALIGN, which produces an optimal global alignment between two protein or DNA sequences, using a modification of the algorithm described by Myers and Miller (1988). First, we grouped all proteins by ALIGN with full length matches of 80% similarity, and considered the proteins in the same group as too similar to each other for use in the SVM training. Then, we selected only one protein entry randomly from each group. We did not consider protein entries containing X, Z or B in the amino acid sequence.

After the sequence similarity check operation, we shuffled the order of all sequence entries in the data set, because similar SWISS-PROT entry names tend to represent similar proteins. This operation was desirable for the cross-validation test after construction of SVMs.

The total number of proteins in the final data set was 7589 for the 12 subcellular locations as summarized in Table 2. We constructed the data set without restriction of organisms. The number of different organisms in the data set was 709. The top ranking five were 1027 yeast (*Saccharomyces cerevisiae*) proteins, 1006 human (*Homo sapiens*) proteins, 592 mouse (*Mus musculus*) proteins, 570 rat (*Rattus norvegicus*) proteins, and 309 worm (*Caenorhabditis elegans*) proteins.

SUPPORT VECTOR MACHINE

Kernel functions

SVM is a learning algorithm (Cristianini and Shawe-Taylor, 2000), which from a set of positively and negatively labeled training vectors learns a classifier that can be used to classify

Table 2. The number of proteins used in the data set

Subcellular location	No. of entries
Chloroplast	671
Cytoplasmic	1245
Cytoskeleton	41
Endoplasmic reticulum	114
Extracellular	862
Golgi apparatus	48
Lysosomal	93
Mitochondrial	727
Nuclear	1932
Peroxisomal	125
Plasma membrane	1677
Vacuolar	54
Total	7589

Chloroplast proteins exist only in a plant cell, and vacuolar proteins exist in a plant or fungal cell. The animal cells contain lysosomes that are corresponding to vacuoles in fungal or plant cells.

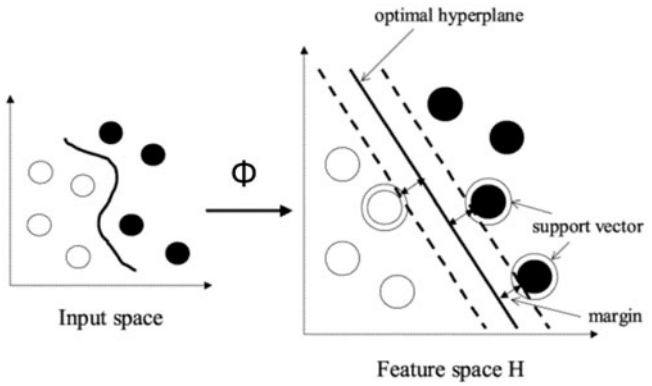

Fig. 1. An illustration of the SVM training method. Given a nonlinear classification problem in the input space, the SVM method defines a mapping Φ, and constructs the optimal separating hyperplane in the high-dimensional feature space H. Black and white circles indicate positive and negative samples to be classified, each of which is characterized by a vector. Support vectors are indicated by extra circles.

new unlabeled test samples. SVM learns the classifier by mapping the input training samples $\{x_1, \ldots, x_n\}$ into a possibly high-dimensional feature space, and seeking a hyperplane in this space which separates the positive examples from the negative ones with the largest possible margin, i.e. distance to the nearest point (Fig. 1). If the training set is not linearly separable, SVM finds a hyperplane, which optimizes a trade-off between good classification and large margin. In Figure 1 the SVM method defines a mapping Φ, and builds a linear SVM in the high-dimensional feature space. Black circles are positive examples, and white circles are negative examples. Support vectors are indicated by extra circles.

Instead of explicitly mapping the objects to the possibly high-dimensional feature space H, SVM usually works implicitly in the feature space by only computing the corresponding kernel $K(\vec{x}, \vec{y})$ between any two objects x and y, defined by:

$$K(\vec{x}, \vec{y}) = \Phi(\vec{x}) \cdot \Phi(\vec{y}),$$

where Φ is the mapping to the feature space H.

Popular kernels used in most SVM packages include the linear kernel, the polynomial kernel, and the RBF kernel. The linear kernel is defined by

$$K(\vec{x}, \vec{y}) = \vec{x} \cdot \vec{y},$$

which is not really mapping the objects to a high-dimensional feature space because $\Phi(\vec{x}) = \vec{x}$. The polynomial kernel is defined by

$$K(\vec{x}, \vec{y}) = (\vec{x} \cdot \vec{y} + 1)^d.$$

We tested this kernel with various values of the degree d. The shape of the decision boundary in the input space becomes more complex as the degree increases. We also tested the RBF kernel, which is defined by

$$K(\vec{x}, \vec{y}) = \exp(-\gamma \|\vec{x} - \vec{y}\|^2),$$

where $\gamma = 1/\sigma^2$ is a parameter and σ is called the width of the kernel. A smaller γ or a larger σ makes the decision boundary smoother.

For actual implementation we used the SVM-light package version 5.00 by Joachims (1999), which could be downloaded from http://svmlight.joachims.org/ for scientific use. SVM-light consists of a learning module (svm_learn) and a classification module (svm_classify). The classification module can be used to apply the learned model to new examples. We tested linear, polynomial and RBF kernels with various parameters. The parameter C, which controls the trade-off between training error and margin, was always set to its default value, namely

$$C = \frac{N}{\sum_{i=1}^{N} K(x_i, x_i)},$$

where N is the size of the training set.

Compositions of amino acids and amino acid pairs

Each protein in the training data set of N proteins is characterized by a vector $\vec{x}_i (i = 1, \ldots, N)$ representing certain sequence features, together with the positive label '+1' or the negative label '−1' for discriminating two different subcellular locations. In addition to the amino acid composition, which has been often utilized for protein localization predictions (see Introduction), we consider the amino acid pair composition and the gapped amino acid pair compositions

corresponding, respectively, to two adjacent amino acids (dipeptides) and two amino acids separated by one or more intervening residue positions. We expect that these additional compositions will detect different sequence features. In particular, the gapped amino acid pair composition will detect periodic appearance of certain amino acids in the sequence. In the present study, we examined the gapped amino acid pair composition with 1–3 intervening residues. The vector $\vec{x}_i$ has 20 coordinates for the amino acid composition and 400 coordinates for the four kinds of amino acid pair compositions.

Voting scheme

In the training of SVMs, we use the method of one versus the others, or one versus the rest. For example, an SVM for the chloroplast protein group is trained with the chloroplast protein sequences used as positive samples and proteins in the other 11 subcellular location groups used as negative samples, because SVMs basically train classifiers between only two different samples. Thus, we build 60 SVM classifiers corresponding to 12 subcellular locations and five different types of compositions: amino acids, amino acid pairs, one gapped amino acid pairs, two gapped amino acid pairs, and three gapped amino acid pairs. For each of the five different compositions, a query protein is tested against 12 SVM classifiers and assigned to the subcellular location that corresponds to the highest output value. After repeating this step for the five different compositions, the results are combined by a voting scheme.

The voting scheme utilized here is the following. When a single location receives five, four or three votes out of five votes, or it receives only two votes but the rest of the votes are split, the query protein is predicted to belong to this location. When two locations receive two votes each, the prediction is either of these locations. When all votes are split in five ways, the subcellular location is unknown. We expect that such a voting scheme would capture different sequence features for different locations, and that it is more stable to the change of data sets than the method using, for example, the amino acid composition alone.

5-fold cross-validation test

The prediction performance was examined by the 5-fold cross-validation test, in which the data set of 7589 proteins for the 12 subcellular locations was divided into five subsets of approximately equal size. This means that the data was partitioned into training and test data in five different ways. After training the SVMs with a collection of four subsets, the performance of the SVMs was tested against the fifth subset. This process is repeated five times so that every subset is once used as the test data.

In order to assess the accuracy of prediction methods we use two measures, the total accuracy defined by

$$\mathrm{TA} = \frac{\sum_{i=1}^{k} T_i}{N},$$

and the location accuracy defined by

$$\mathrm{LA} = \frac{\sum_{i=1}^{k} P_i}{k},$$

where

$$P_i = \frac{T_i}{n_i}.$$

Here N is the total number of proteins in the data set ($N = 7589$), k is the number of subcellular locations ($k = 12$), n_i is the number of proteins in each location i (Table 2), and T_i is the number of correctly predicted proteins (true positives) in each location i. When the correct location is one of the two alternative locations predicted in the voting scheme mentioned above, the score of 0.5 is given when counting T_i. In previous studies the total accuracy TA has usually been mentioned as the performance of the predictors, but TA depends on the location groups with large numbers of entries. For example, if a prediction method is optimized for the plasma membrane group (1677 entries in our data set), the total accuracy TA will rise dramatically. In contrast, the location accuracy LA has opposite aspects reflecting the performances of small groups equally important to those of large groups. We try to consider both measures to find the best condition of the prediction method.

RESULTS

Kernel selection

We begin by selection of a kernel from the three possibilities: the simple linear kernel, the polynomial kernel, and the RBF kernel. A typical result of prediction accuracies for 7589 protein sequences with different types of kernel functions is summarized in Table 3. The performance of each classifier was measured by examining how well the classifier identified positive examples in the test sets, or by P_i at each subcellular location, according to the 5-fold cross-validation test. To judge the overall performance, both TA (total accuracy) and LA (locations accuracy) were computed. The result of Table 3 was obtained based on the amino acid composition information only. We also examined the performance of different kernel functions based on the information about each of the four types of amino acid pair compositions, and observed a similar tendency. In Table 3 we chose the parameter $d = 5$ for the polynomial kernel, which gave the best result when d was changed from 1 to 6. Various values of the parameter γ were also tested for the RBF kernel ranging from 0.01 to 0.1, and our choice was $\gamma = 0.02$ or 0.03. In general, RBF kernel SVM classifiers performed better than linear and polynomial kernel SVM classifiers, as indicated by the TA and LA values in Table 3.

However, when individual locations were examined the prediction ability became worse with the RBF kernel for some small groups, such as peroxisomal, Golgi apparatus, and vacuolar protein groups. It appeared that the balance

Table 3. Prediction accuracies (%) for the 12 subcellular locations with different types of SVM kernels based on amino acid composition information

Location (No. of entries)	Linear	Polynomial $d = 5$	RBF		
			$\gamma = 0.02$	$\gamma = 0.03$	Mixture
Chloroplast (671)	35.0	51.0	62.3	62.7	64.5
Cytoplasmic (1245)	34.1	49.2	65.9	67.8	63.9
Cytoskeleton (41)	29.3	58.5	51.2	53.7	53.7
ER (114)	11.4	36.8	46.5	46.5	51.8
Extracellular (862)	34.7	60.6	75.8	73.9	72.6
Golgi apparatus (48)	37.5	41.7	25.0	22.9	27.1
Lysosomal (93)	54.8	60.2	57.0	61.3	63.4
Mitochondrial (727)	24.9	35.2	48.8	43.9	51.7
Nuclear (1932)	54.2	71.3	82.5	85.2	83.4
Peroxisomal (125)	36.8	45.6	23.2	24.0	24.8
Plasma membrane (1677)	81.3	83.1	87.7	87.4	86.6
Vacuolar (54)	38.9	37.0	31.5	25.9	37.0
Total accuracy, TA	48.9	62.2	72.4	72.7	72.4
Location accuracy, LA	39.4	52.5	54.8	54.6	56.7

between large groups and small groups was not optimal with a single parameter value of γ. The differences between $\gamma = 0.02$ and 0.03 tended to be larger for smaller groups in Table 3 while the prediction ability for large groups, such as cytoplasmic, nuclear, plasma membrane, and extracellular protein groups, was relatively stable. Thus, we tried the parameter mixture in the RBF kernel where a smaller γ value was used for smaller groups to avoid over-fitting and a larger γ value was used for larger groups to better capture a more complex decision boundary. In particular we used $\gamma = 0.03$ for the four large groups (cytoplasmic, extracellular, nuclear, and plasma membrane protein groups), and $\gamma = 0.02$ for the rest. As the result we could obtain a modest improvement of the LA while the TA was not much affected (see Table 3).

Use of amino acid pair compositions

Table 4 shows the result of the 5-fold cross-validation tests for the RBF kernel SVM classifiers with the parameter mixture ($\gamma = 0.02$ and 0.03), using five different types of compositions: amino acids, amino acid pairs, one gapped amino acid pairs, two gapped amino acid pairs, and three gapped amino acid pairs. Although there was not a large difference, TA became somewhat better for the amino acid pair compositions than for the amino acid composition, and it was opposite for LA. We had expected that known signals of amphipathic helices, for example, in mitochondrial proteins would be reflected in two or three gapped amino acid pairs, but there was not such a clear pattern probably because we computed compositions for the entire amino acid sequence. In any case, the information about amino acid pair compositions was potentially as useful as the information about amino acid compositions in distinguishing subcellular locations of proteins.

Table 4. Comparison of different composition information

Location	Amino acid	Amino acid pair	One gapped amino acid pair	Two gapped amino acid pair	Three gapped amino acid pair
Chloroplast	64.5	68.7	67.1	66.9	67.1
Cytoplasmic	63.9	69.6	66.4	67.3	67.8
Cytoskeleton	53.7	61.0	58.5	58.5	63.4
ER	51.8	49.1	43.9	45.6	42.1
Extracellular	72.6	73.3	77.0	74.6	73.3
Golgi apparatus	27.1	14.6	20.8	10.4	12.5
Lysosomal	63.4	55.9	58.1	49.5	58.1
Mitochondrial	51.7	51.2	54.1	51.3	48.4
Nuclear	83.4	88.7	88.8	88.2	87.5
Peroxisomal	24.8	24.8	20.0	24.8	20.8
Plasma membrane	86.6	91.1	91.3	90.8	91.2
Vacuolar	37.0	33.3	24.1	13.0	16.7
TA (%)	72.4	75.9	75.8	75.0	74.7
LA (%)	56.7	56.8	55.8	53.4	54.1

Table 5. Results of voting for predicting different numbers of subcellular locations

Accuracy	12 locations (7589 entries)	11 locations for plant cells (7496 entries)	10 locations for fungal cells (6825 entries)	10 locations for animal cells (6864 entries)
TA (%)	78.2 ± 0.9	78.5 ± 0.9	79.5 ± 0.9	79.6 ± 0.9
LA (%)	57.9 ± 2.1	57.9 ± 1.3	56.8 ± 1.9	59.9 ± 3.3

The average and the SD were computed by the 5-fold cross-validation test.

Improvement by voting

In order to best utilize the potentials of the five different compositions in Table 4, we devised a voting scheme as described above. As shown in Table 5 there was a definite trend of the improved TA, from 72.4% without voting (amino acid composition only shown in Table 4) to 78.2% with voting (five types of compositions), while the LA remained at a similar level, 56.7% without voting and 57.9% with voting. The average and the SD of the TA or the LA were computed from five trials.

Our voting scheme involves 60 SVM classifiers, 12 classifiers for each of the five compositions. In the first step a vote is cast by the best score among a set of 12 classifiers and in the second step the final prediction is made by the most votes. Apparently, this scheme was able to better capture sequence features that might be different in different subcellular locations. Voting was also effective in alleviating the dependency of SVM training on the data set. Changes or updates of entries

within our training data set could cause large changes of the prediction accuracy in some types of composition information used separately, especially for small groups including cytoskeleton and Golgi apparatus (data not shown). The voting scheme apparently averaged out such large changes.

We also tested for more realistic repertoires of subcellular locations in different cell types, 11 subcellular locations excluding lysosome for a plant cell, 10 locations excluding chloroplast and lysosome for a fungal cell, and 10 locations excluding chloroplast and vacuole for an animal cell. Note that vacuoles in fungi or plants are thought to correspond to lysosomes in animals. We performed identical procedures for training and testing with smaller data sets, 7496, 6825, and 6864 entries for the 11 (plant), 10 (fungal), and 10 (animal) locations, respectively. Here we constructed the data sets without considering actual organism groups; for example, the data set for the animal cell type of 10 locations actually contained entries from plants or fungi. The result is shown in Table 5. Overall, the accuracy of prediction was comparable to the case of 12 locations.

Table 6. Prediction accuracy with nine subcellular locations

Location	12 locations	9 locations
Chloroplast	72.3	70.3
Cytoplasmic[a]	72.2	73.9
Cytoskeleton	58.5	59.8
ER	46.5	39.0
Extracellular	78.0	77.1
Golgi apparatus	14.6	—
Lysosomal	61.8	62.4
Mitochondrial	57.4	53.5
Nuclear	89.6	89.0
Peroxisomal	25.2	—
Plasma membrane	92.2	91.9
Vacuolar	25.0	—
TA (%)	78.2	79.1
LA (%)	57.9	68.5

[a]Golgi apparatus, peroxisomal and vacuolar proteins are included in cytoplasmic proteins (total of 1472 entries) in the prediction of nine locations.

Implementation

The prediction method presented in this paper is implemented as a computer program named PLOC and the web service is made available at http://www.genome.ad.jp/SIT/ploc.html. Given an amino acid sequence and a cell type (plant, fungal, or animal cell) the program reports the most probable subcellular location according to the voting of five predictions based on the compositions of amino acids and four types of amino acid pairs, using the SVM classifiers with the RBF kernel and the parameter mixture of $\gamma = 0.02$ and 0.03. The predicted location is marked on a schematic drawing of the cell, and the result is associated with the details of five votes and 60 SVM classifiers. The data sets used in this paper are also available at http://web.kuicr.kyoto-u.ac.jp/~park/Seqdata/

DISCUSSION

Measure of prediction accuracy

Because there was a large difference, up to 45 times, in the data size of each subcellular location group (Table 2), we tried to balance the TA and the LA when selecting an optimal condition for our prediction method. If the number of proteins in our data set corresponds to the actual frequency of proteins in the 12 subcellular locations in living cells, TA would represent an expected degree of prediction accuracy in practical applications. In fact TA has usually been utilized in previous methods as a measure of prediction accuracy. However, TA could easily be optimized too much for large groups at the expense of small groups. Thus, we used LA to measure the balance of prediction accuracies between large and small groups.

We found two ways to improve trade-off between TA and LA. One was the parameter mixture for the RBF kernel. By assigning a larger γ value for the four largest groups and a smaller γ value for the rest, LA was improved without much affecting TA (Table 3). The other was the voting scheme. By combining five predictions based on different types of compositions, TA was improved without much affecting LA (Tables 4 and 5). Both seemed to be related to dependency of SVM training with small data sets. A smaller γ, that is, a larger $\sigma(\gamma = 1/\sigma^2)$ for the RBF kernel was effective to avoid overfitting, and the voting represented a type of averaging to lessen the dependency on training data.

Prediction of nine locations

Despite the improvements obtained by the parameter mixture and the voting scheme, prediction rates were not satisfactory for some groups, especially Golgi apparatus, peroxisomal, and vacuolar protein groups. When details of the prediction results were examined, these proteins were often assigned to cytoplasmic proteins and vice versa (data not shown). Thus, we constructed a data set of nine subcellular locations where cytoplasmic, Golgi apparatus, peroxisomal, and vacuolar proteins were combined into a single group. Table 6 shows the result of the 5-fold cross-validation test with the same voting scheme and the RBF kernel mixture SVMs. As expected, the LA exhibited a large increase, from 58 to 69%. We did not, however, include this prediction scheme in the Web service of our PLOC program because our intention is to increase, rather than decrease, the number of subcellular locations in future updates (see later).

Comparison with other methods

In order to check the performance of our method, we made comparisons with other methods, especially the method by Cai *et al*. (2002) who had also used SVMs and a data set

Table 7. Comparison of our method with a previous method

Location	Cai *et al*. (2002)		Our method	
	No. of entries (total 2191)	Jack-knife (%)	No. of entries (total 7589)	5-Fold cross (%)
Chloroplast	145	57	671	72
Cytoplasmic	571	88	1245	72
Cytoskeleton	34	44	41	59
ER	49	31	114	47
Extracellular	224	57	862	78
Golgi apparatus	25	12	48	15
Lysosomal	37	54	93	62
Mitochondrial	84	42	727	57
Nuclear	272	73	1932	90
Peroxisomal	27	4	125	25
Plasma membrane	699	91	1677	92
Vacuolar	24	25	54	25
TA		75		78
LA		48		58

of 12 locations by Chou and Elrod (1999a). The comparison is summarized in Table 7. Although the size of the data set was different (2191 and 7589 entries) and the test method was different (jack-knife test and 5-fold cross-validation test), our method achieved a significant improvement in the LA, from 48 to 58%, together with a small increase in the TA, from 75 to 78%. Their method is apparently optimized for large groups, especially plasma membrane (91%) and cytoplasm (88%), at the expense of small groups, such as peroxisome (4%), Golgi apparatus (12%), and vacuole (25%). In contrast, our method is more balanced, although the poor performance for the three problematic groups remains the same.

Because the repertoire of subcellular locations was different in different prediction methods, the reported values of prediction accuracy (see Introduction) cannot be readily compared. In addition, different levels of sequence similarities in different data sets also complicate the comparison. We removed highly similar sequences and the level of sequence similarity was, at most, 80% in the data set reported here. In contrast, the data set used by Cai *et al*. (2002) apparently contained a large number of similar sequence pairs above the 80% level, although it was not possible to entirely reconstruct their data set (Chou and Elrod, 1999a) because some of the SWISS-PROT identifiers no longer existed. In order to estimate the effect of similarity level on the prediction accuracy we prepared our data sets with different similarity thresholds, 70 or 60%, for the four largest groups, cytoplasmic, extracellular, nuclear, and plasma membrane protein groups, while keeping the 80% level for the rest. The TA decreased from 78.2 (Table 5) to 76.0 and 73.5%, respectively, with the 70 and 60% levels, and the LA varied from 57.9 to 57.9% and 56.5%. Perhaps we should not emphasize too much about the absolute value of the prediction accuracy, which is difficult to compare among different studies, but we can conclude our approach to optimize both TA and LA was successful to cover one of the largest repertoires of subcellular locations with prediction accuracy as good as the previous best methods.

Further improvements

In the training of SVM classifiers, we adopted the one-versus-others method. For each of the 12 subcellular locations, an SVM was trained with positive data in that location and negative data in the other 11 locations. Therefore, a protein in the test set is checked against 12 SVM classifiers. We also tested the all-versus-all method as a different way of SVM training. In this method each pair of subcellular locations is used as positive and negative data. A protein in the test set is now checked against $66 (12 \times 11/2)$ SVM classifiers. However, this method showed lower prediction accuracy as a whole. We have examined, in a fairly comprehensive way, various SVM kernels and parameters, and it is unlikely that a significant improvement will be obtained by changing training methods alone. There is still room for improvements in defining sequence features, but our voting scheme with the compositions of amino acids and four types of amino acid pairs is already an improvement over other methods mostly based on the amino acid composition.

Perhaps further improvements can best be obtained by preparing data sets of higher quality. It should be possible to increase the number of data entries from updated databases, especially for the small groups with low prediction accuracies, including Golgi apparatus, peroxisomal, and vacuolar protein groups. Adding new subcellular locations or defining finer classifications, such as subdividing mitochondrial proteins into inner/outer membrane and matrix proteins, will also be important in practical applications of gene annotations and functional predictions. At the same time it would become necessary to consider protein groups that inherently belong to multiple locations, such as those that move between cytoplasm and nucleus under different conditions. We think that our method is general enough to cope with these different data sets and also simple enough to automate retraining of SVMs so that we can update our Web service accordingly.

ACKNOWLEDGEMENTS

We thank Jean-Philippe Vert, Masahiro Hattori, and Shuichi Kawashima for helpful discussions and comments on the manuscript. This work was supported by grants from the Ministry of Education, Culture, Sports, Science, and Technology of Japan, the Japan Society for the Promotion of Science, and the Japan Science and Technology Corporation.

REFERENCES

Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.

Brown,M.P.S., Grundy,W.N., Lin,D., Cristianini,N., Sugnet,C.W., Furey,T.S., Ares,M. and Haussler,D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci. USA*, **97**, 262–267.

Cai,Y.D., Liu,X.J., Xu,X.B. and Chou,K.C. (2002) Support vector machines for prediction of protein subcellular location by incorporating quasi-sequence-order effect. *J. Cell Biochem.*, **84**, 343–348.

Chou,K.C. (2000) Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. *Biochem. Biophys. Res. Commun.*, **278**, 477–483.

Chou,K.C. (2001) Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins*, **43**, 246–255.

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

Chou,K.C. and Elrod,D.W. (1999b) Prediction of membrane protein types and subcellular locations, *Proteins*, **34**, 137–153.

Cristianini,N. and Shawe-Taylor,J. (2000) *An Introduction to Support Vector Machines and other Kernel-Based Learning Methods.* Cambridge University Press, Cambridge, MA.

Ding,C. and Dubchak,I. (2001) Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics*, **17**, 349–358.

Emanuelsson,O., Nielsen,H., Brunak,S. and von Heijne,G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.*, **300**, 1005–1016.

Furey,T.S., Cristianini,N., Duffy,N., Bednarski,D.W., Schummer,M. and Haussler,D. (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **16**, 906–914.

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

Hua,S. and Sun,Z. (2001b) A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. *J. Mol. Biol.*, **308**, 397–407.

Joachims,T. (1999). In Schölkopf,B., Burges,C. and Smola,A. (ed.), *Making Large-Scale SVM Learning Practical. Advances in Kernel Methods–Support Vector Learning*. MIT Press, Cambridge MA.

Myers,E.W. and Miller,W. (1988) Optimal alignments in linear space. *CABIOS*, **4**, 11–17.

Nakai,K. and Kanehisa,M. (1992) A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics*, **14**, 897–911.

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

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

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