PROTEINS: Structure, Function, and Bioinformatics 64:643–651 (2006)

Prediction of Protein Subcellular Localization

Chin-Sheng Yu,[1] Yu-Ching Chen,[2] Chih-Hao Lu,[2] Jenn-Kang Hwang[1,2,3]*
[1]*Department of Biological Science and Technology, National Chiao Tung University, Hsinchu, Taiwan, Republic of China*
[2]*Institute of Bioinformatics, National Chiao Tung University, Hsinchu, Taiwan, Republic of China*
[3]*Core Facility for Structural Bioinformatics, National Chiao Tung University, Hsinchu, Taiwan, Republic of China*

ABSTRACT **Because the protein's function is usually related to its subcellular localization, the ability to predict subcellular localization directly from protein sequences will be useful for inferring protein functions. Recent years have seen a surging interest in the development of novel computational tools to predict subcellular localization. At present, these approaches, based on a wide range of algorithms, have achieved varying degrees of success for specific organisms and for certain localization categories. A number of authors have noticed that sequence similarity is useful in predicting subcellular localization. For example, Nair and Rost (Protein Sci 2002;11:2836–2847) have carried out extensive analysis of the relation between sequence similarity and identity in subcellular localization, and have found a close relationship between them above a certain similarity threshold. However, many existing benchmark data sets used for the prediction accuracy assessment contain highly homologous sequences—some data sets comprising sequences up to 80–90% sequence identity. Using these benchmark test data will surely lead to overestimation of the performance of the methods considered. Here, we develop an approach based on a two-level support vector machine (SVM) system: the first level comprises a number of SVM classifiers, each based on a specific type of feature vectors derived from sequences; the second level SVM classifier functions as the jury machine to generate the probability distribution of decisions for possible localizations. We compare our approach with a global sequence alignment approach and other existing approaches for two benchmark data sets—one comprising prokaryotic sequences and the other eukaryotic sequences. Furthermore, we carried out all-against-all sequence alignment for several data sets to investigate the relationship between sequence homology and subcellular localization. Our results, which are consistent with previous studies, indicate that the homology search approach performs well down to 30% sequence identity, although its performance deteriorates considerably for sequences sharing lower sequence identity. A data set of high homology levels will undoubtedly lead to biased assessment of the performances of the predictive approaches—especially those relying on homology search or sequence annotations. Our two-level classification system based on SVM does not rely on homology search; therefore, its performance remains relatively unaffected by sequence homology. When compared with other approaches, our approach performed significantly better. Furthermore, we also develop a practical hybrid method, which combines the two-level SVM classifier and the homology search method, as a general tool for the sequence annotation of subcellular localization. Proteins 2006;64:643–651.** © 2006 Wiley-Liss, Inc.

Key words: support vector machines; subcellular localization; sequence alignment

INTRODUCTION

Due to the rapid advances in genomic and proteomic research in recent years, tremendous amounts of DNA and protein sequences have accumulated in databases. It becomes increasingly important for computational biologists to develop practical tools to efficiently extract relevant biological information from sequences for functional annotation. Because the protein's function is closely associated with its subcellular localization, the ability to predict protein subcellular localization will be useful in the characterization of the expressed sequences of unknown functions. In recent years, many efforts[1–19] have been made to develop novel approaches to predict protein subcellular localization. These approaches cover various types of algorithms such as the knowledge-based expert system,[1] the artificial neural networks,[3,4,11] the support vector machines (SVM),[9,12,16] the covariant discriminant algorithm,[2,5] and the Bayesian networks.[15,18] Some of the approaches use the short N-terminal amino acid sequences[1,3,6–8] (i.e., the sorting or signal peptides), the amino acid compositions,[2,4,5,9,11,12,19] or the general *n*-peptide compositions[16] derived from the whole amino acid sequences. Other approaches make use of additional information like sequence profiles derived from PSI-BLAST,[10,17,19] the ontology labels, or the text annotations of the sequence databases.[13–15] In general, these approaches perform well for specific organisms and for certain localization categories. However, it is noticed that

*Correspondence to: Jenn-Kang Hwang, Institute of Bioinformatics, National Chiao Tung University, Hsinchu, Taiwan, ROC. E-mail: jkhwang@cc.nctu.edu.tw

Received 22 December 2005; Revised 8 March 2006; Accepted 13 March 2006

Published online 2 June 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21018

the benchmark data sets used for the assessment of the predictive performances of many methods usually contain highly homologous sequences. For example, the data set of Reinhardt and Hubbard[4] as well as that of Garg et al.[19] include sequences with up to 90% sequence identity, and the data set of Park and Kanehisa[12] comprises sequences with up to 80% sequence identity. Several groups[20,21] have already pointed out that there is a close relationship between sequence similarity and identity in both subcellular localization and the signal peptide cleavage sites. For example, Nair and Rost[21] have performed large-scale analysis of the relation between sequence similarity and identity in subcellular localization. Their results show that one can accurately infer the subcellular compartment of a protein if one can find close homologs of experimentally verified localization using the HSSP distance,[21] a measure for sequence similarity accounting for pairwise sequence identity and alignment length. It is well known in the study of secondary structure prediction[22–24] that the homologous sequences are meticulously removed from the testing-training data sets. For example, the popular benchmark RS126 set[22] comprises sequences in which no sequence pairs share more than 25% sequence identity (over a length of more than 80 residues). The training-testing data sets of high homology will obviously lead to overprediction, that is, the positive predictions may be due to the presence of highly similar sequences in both training and testing sets instead of the effectiveness of the approaches in extracting key features associated with the investigated properties. Unfortunately, such is not the case of the study of subcellular localization. In this work, we developed a two-level SVM system to predict subcellular localization: the first level comprises a number of SVM classifiers, each based on a distinctive set of feature vectors derived from sequences. The second level consists of a jury SVM that processes the outputs from the first level SVM classifiers to generate the probability distribution of subcellular localization. We showed that this two-level approach performs better than other approaches for data sets comprising sequences of low homology. We will refer to this *two*-level SVM predictor of sub*cel*lular *lo*calization as CELLO II. Furthermore, using the relationship between sequence similarity and identity in subcellular localization,[20,21] we propose a practical pipeline approach combining CELLO II and the sequence alignment method to predict subcellular localization.

METHODS

Support Vector Machines

We will briefly explain SVM here. The idea of the SVM goes as follows: given training vectors $\mathbf{x}_i$ where $i = 1, \ldots, l$, and a vector $\mathbf{y} = (y_1, \ldots, y_i)$ defined as: $y_i = 1$ if $\mathbf{x}_i$ is in one class, and $y_i = -1$ if $\mathbf{x}_i$ is in the other class. The support vector technique tries to find the optimal separating-hyperplane $\mathbf{w}^T\mathbf{x}_i + b = 0$ with the largest distance between two classes, measured along a line perpendicular to this hyperplane. This requirement is equivalent to solving the equation: $\min_{w,b} 1/2\ \mathbf{w}^T\mathbf{w}$ under the constraints $y_i[(\mathbf{w}^T\mathbf{x}_i) + b] \geq 1$, $i = 1, \ldots, l$. However, in practice, these data to be classified may not be linearly separable. To overcome this difficulty, SVM transforms the original input space into a higher dimensional feature space using the function $\Phi(\mathbf{x}) = (\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \ldots)$. The optimization equation is now written as:

$$\min_{w,b,\xi} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i$$

under the constraints $y_i[(\mathbf{w}^T\mathbf{\Phi}(\mathbf{x}_i)) + b] \geq 1 - \xi_i$, $i = 1, \ldots, l$. The constraints are much more relaxed and allow some training data to be on the incorrect side of the separating hyperplane $\mathbf{w}^T\mathbf{x} + b = 0$. If the penalty parameter C is large enough and the data is linearly separable, all ξ_is will be zero.[25] In practice, the explicit form of $\mathbf{\Phi}(\mathbf{x})$ is not required, and we only need to calculate the function $K(\mathbf{x}_i,\mathbf{x}_j) = \mathbf{\Phi}(\mathbf{x}_i)^T\ \mathbf{\Phi}(\mathbf{x}_j)$ called the kernel function. Note that the training data are mapped into a vector in a higher dimensional space, since in a higher dimensional space, the data may be linearly separated. This procedure has the advantage of allowing training errors, because we do not require the training data to be always on the correct side of the separating hyperplane. Thus, we minimize the training error $\Sigma_{i=1}^{l}\xi_i$ in the objective function. In the end, the decision function is written as $f(\mathbf{x}) = \mathrm{sgn}(\mathbf{w}^T\mathbf{\Phi}(\mathbf{x}) + b)$. Those $\mathbf{x}_i$s that are used to construct $\mathbf{w}$ and b are called support vectors. All the SVM calculations are performed using LIBSVM,[26] a general library for support vector classification and regression from Lin's lab. Here, we use the radial basis function kernel given by $\exp(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2)$ for all our calculations. An important issue is the selection of parameters. For SVM training, the penalty parameter C and the kernel parameter γ of the RBF function must be determined in advance. We use the cross validation on different parameters for the model selection.[27]

Coding Schemes

The n*-Peptide Composition*

We have previously developed a general global sequence descriptor based on the n-peptide composition codings (denoted by A_n) to predict protein properties.[16,28,29] In the case of $n = 1$, the A_1 coding reduces to the usual amino acid composition, which can be considered as the first-order approximation to the complete protein sequence. The A_2 coding gives the dipeptide composition. As n increases, the A_n coding provides progressively more detailed sequential information. In the limit that n is the whole length of the sequence, the A_n becomes the sequence itself. The A_n coding scheme has the advantage of systematically extracting more information from sequences when n increases. In the case of $n \geq 3$, the computation of A_n becomes not only impractical from a learning viewpoint but also susceptible to the danger of overfitting. We can overcome the size problem by regrouping the amino acids into smaller number of classes according to their physicochemical properties. In this work, we use the following classification schemes based on the physicochemical properties of amino acid—we use H_n for polar (RKEDQN), neutral (GASTPHY), and hydrophobic (CVLIMFW);[30] V_n for small (GASCTPD), medium (NVEQIL), and large (MHK-

TABLE I. The Coding Schemes of the Amino Acids Compositions Based on Different Classification Types

Coding Schemes	Classification types	Amino acid types
	Polar	RKEDQN
H	Neutral	GASTPHY
	Hydrophobic	CVLIMFW
	Small	GASCTPD
V	Medium	NVEQIL
	Large	MHKFRYW
	Low polarizability	GASDT
Z	Medium polarizability	CPNVEQIL
	High polarizability	KMHFRYW
	Low polarity	LIFWCMVY
P	Neutral polarity	PATGS
	High polarity	HQRKNED
	Acidic	DE
F	Basic	HKR
	Polar	CGNQSTY
	Nonpolar	AFILMPVW
	Acidic	DE
	Basic	HKR
	Aromatic	FWY
S	Amide	NQ
	Small hydroxyl	ST
	Sulfur-containing	CM
	Aliphatic	AGPILV
	Acidic	DE
	Basic	HKR
	Aromatic	FWY
E	Amide	NQ
	Small hydroxyl	ST
	Sulfur-containing	CM
	Aliphatic 1	AGP
	Aliphatic 2	ILV

FRYW);[30] Z_n for low polarizability (GASDT),[30] medium (CPNVEQIL), and high (KMHFRYW);[30] P_n for low polarity (LIFWCMVY), neutral (PATGS), and high polarity (HQRKNED);[30] F_n for acidic (DE), basic (HKR), polar (CGNQSTY), and nonpolar (AFILMPVW); S_n for acidic (DE), basic (HKR), aromatic (FWY), amide (NQ), small hydroxyl (ST), sulfur-containing (CM), and aliphatic (AGPILV); E_n for acidic (DE), basic (HKR), aromatic (FWY), amide (NQ), small hydroxyl (ST), sulfur-containing (CM), aliphatic 1 (AGP), and aliphatic 2 (ILV). For clarity, these coding schemes are summarized in Table I.

The partitioned amino acid composition

We use X_k^Y to denote the partitioned amino acid composition—the sequence is partitioned into k subsequences of equal length, and each fragment is encoded by the particular amino acid composition Y. For example, the notation X_5^{A1} denotes that the sequence is divided into five subsequences, each of which is encoded by A_1 (note that X_1^{A1} is equivalent to A_1). The coding X_k^Y provides information about the local properties of sequences.

The g-gap dipeptide composition

Another generalized sequence composition is the g-gap dipeptide compositions, denoted by D_g, in which we compute the composition of the sequence of the form $a(x)_g b$, where a and b denote two specific amino acid types, and $(x)_g$ denotes g intervening amino acids of arbitrary type x. Note that in the special case of g = 0, D_0 is equivalent to A_2.

The local amino acid composition

We use W_l to denote the amino acid composition of a sliding window of length l centered on a given amino acid type. The W_l provides information on the flanking sequences of a given amino acid type. Note that when l is the length L of the whole sequence, W_L reduces to A_1.

The two-level SVM classifier system

The SVM classifiers in the first level comprise a number of SVM classifiers, each based on a specific sequence coding as described in the previous section. For the sake of notation simplicity, we will use the coding symbol to represent the SVM classifier based on that coding. For example, we will denote the SVM system comprising three classifiers, say, A, B, and C by the shorthand symbol $A + B + C$. In this work, the first level classifiers consist of the following SVMs:

$$\sum_{k=1}^{9} X_k^{a_1} + \sum_{k=0}^{6} D_k + \sum_{x \in S} X_S^x + \sum_{l \in S'} W_l,$$

where $S = \{H_3, P_3, F_3, S_2, E_2\}$ and $S' = \{7, \ldots, 15\}$. Each SVM generates a probability distribution[16,28] of the subcellular localization based on its particular sequence coding. A second SVM (i.e., the jury SVM) is used to process these probabilities to generate the final probability distribution of subcellular compartment. The location with the largest probability is used as the prediction. The two-level SVM system is shown schematically in Figure 1.

Performance assessment

Following the previous works,[16,28] we use the percentage accuracy to assess the accuracy of the subcellular localization identification: $Q_i = c_i/n_i$, where c_i is the number correctly predicted in the ith subcellular location, and n_i is the number of sequences in that location. The overall prediction accuracy is given by

$$P = \sum_i f_i Q_i \tag{1}$$

where $f_i = n_i/N$ and N is the total number of sequences. Although the percentage accuracy (Q_i or P) provides a convenient measure for predictive performance, the Matthews Correlation Coefficient[31] (MCC) gives a more informative measure for predictive performance:

$$MCC_i = \frac{TP_i TN_i - FP_i FN_i}{\sqrt{(TP_i + FN_i)(TP_i + FP_i)(TN_i + FP_i)(TN_i + FN_i)}} \tag{2}$$

where TP_i is the true positives in location i, TN_i is the true negatives in location i, FP_i is the false positive, and FN_i is the false negative. The value of MCC_i is 1 for a perfect

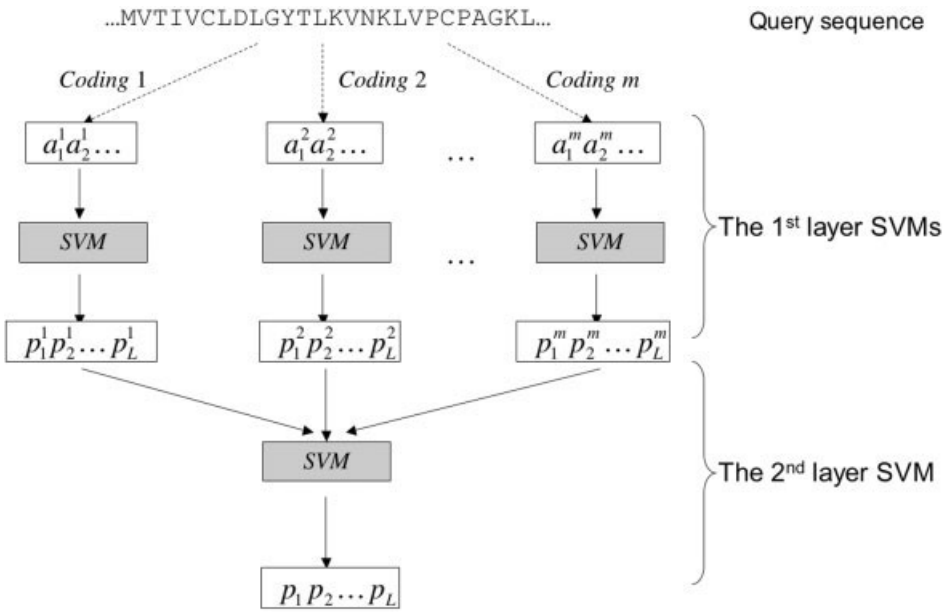

Fig. 1. The first level classification system comprises m SVMs based on different feature vectors: ($a_1^1 a_2^1$ . . .), ($a_1^2 a_2^2$. . .), and ($a^m{}_1 a^m{}_2$. . .). These SVMs generate m probability distributions ($p_1^1 p_2^1$. . .$p^1{}_L$), ($p_1^2 p_2^2$. . .$p^2{}_L$), and ($p^m{}_1 p^m{}_2$. . . p_L^m) of L subcellular localizations. A second layer SVM (as a jury SVM) is used to process these probability distributions to generate the final probability distribution ($p_1 p_2$ . . . p_L).

prediction, 0 for a completely random prediction and −1 for a perfectly reverse correlation.

The sequence–localization relationship

The query sequence is aligned against sequences of known localization. If the top-ranking aligned sequence has an identical localization with the query sequence, the sequence pair will be counted as a positive hit, or else a negative hit. We performed all-against-all sequence alignment using the global alignment program ALIGN developed by Myers and Miller.[32]

Data sets

Two data sets were used in the experiment. The first data set, referred to as the PS data set, is composed of Gram-negative sequences.[18] We selected from the data set only those sequences with a single localization (there are four groups of sequences with double localization, the average of which accounts for about 1% of the original data set). The resultant data set comprises 1444 protein sequences for five subcellular compartments: extracellular (190), cytoplasmic (278), cytoplasmic membrane (309), periplasmic (276), and outer membrane (391). The second data set is from Park and Kanehisa,[12] referred to as the PK dataset. The sequences are selected from SWISSPROT[33] release 39.0 in such as way that the pairwise sequence identities are below 80%. The PK dataset contains 7589 eukaryotic protein sequences for 12 subcellular locations—chloroplast (671), cytoplasmic (1245), cytoskeleton (41), endoplasmic reticulum (ER) (114), extracellular (862), Golgi apparatus (48), lysosomal (93), mitochondrial (727), nuclear (1932), peroxisomal (125), plasma membrane (1677), and vacuolar proteins (54). We followed the same validation procedures for predictive performances as those of the previous works.[12,18]

RESULTS

The Sequence–Localization Relationship

The sequence homology of a data set can easily be inspected using the pair distribution of sequence identities, which shows the relative numbers of sequence pairs that share a given range of sequence identity. Figure 2 shows the pair distributions of the sequence identities of the PS [Fig. 2(A)] and the PK [Fig. 2(B)] data sets. Both data sets peak at 20% sequence identity. However, it is easy to see that significant amount of sequences have a sequence identity ≥30% in both data sets. Performing all-against-all sequence alignments using ALIGN, we plot sequence similarity against identity in localization for the PS and PK data sets [Fig. 3(A) and (B)]. In general, when sequence identity ≥25%, the sequences usually share identical subcellular compartments (however, in the PK data set, the abnormal behaviors of data at sequence identities ≥80% are due to the relatively smaller example sizes at those regions). We also observe that the relationships between sequence identity and identity in localization are quite similar for these data sets.

We built a much larger data set from SWISSPROT release 41.0 by excluding any sequences annotated as MEMBRANE, POSSIBLE, PROBABLE, SPECIFIC PERIODS, or BY SIMILARITY.[11] The resultant data set (referred to as SW41) comprises 9851 eukaryotic proteins sequences distributed in five subcellular compartments: extracelluar, cytoplasmic, mitochondria, nuclear, and others. The larger SW41 data set also shows a similar

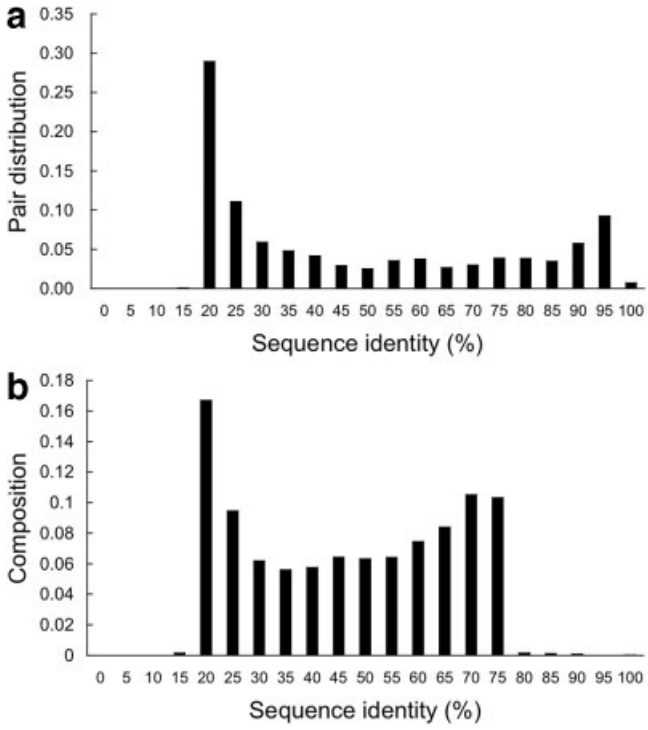

Fig. 2. (**A**) The pair distribution of the sequence identities of the PS data set. Each bin (the width set to 5% sequence identity) represents the relative amount of the sequence pairs that share a given range of sequence identity. For example, all sequences in each bin (say 20%) will share a pair sequence identity between 20 and 25% against each other. The value of the pair distribution is normalized over the total area under the distribution curve. (**B**) The pair distribution of the sequence identities of the PK data set.

relationship [Fig. 3(C)] between identity in sequence and subcellular localization as the PS and PK data sets do.

Comparison of Different Coding Schemes

Tables II and III compare the performances of different coding schemes for the PK and the PS data sets. CELLO II uses a second-level SVM to decide the final prediction, while the SVM based on a single parameter set uses the output with the largest probability as the prediction. We observe similar trends in the overall performance of all single parameter sets for two data sets (Tables II and III). For example, all single parameter sets perform similarly for the plasma membrane compartment for which the overall prediction accuracy ranges from 86 to 92%. The coding X_4^{A1} has the best overall performance among the single parameter sets for both data sets. On the other hand, for certain rows of subcellular compartments, the performances of the single parameter sets fluctuate considerably. For example, in Table II, the prediction accuracy for chloroplast ranges from 57 to 72 %. CELLO II, based on the multiple feature vector coding schemes, consistently outperforms those based on the single parameter set. This is obviously due to the complementarity of information encoded in the single parameter sets. Our results are consistent with previous studies[16,28,29,34,35] that SVMs based on multiple parameters usually perform better than those based on a single parameter.

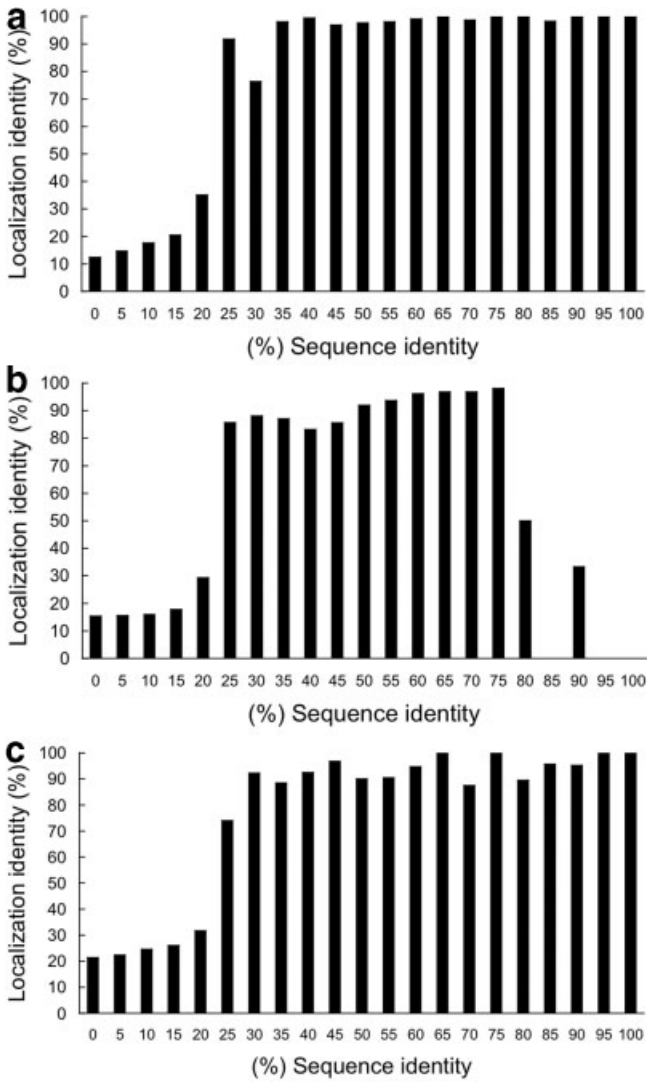

Fig. 3. The bar charts of sequence identity versus identity in localization for (**A**) the PS data set, (**B**) the PK data set, and (**C**) the SW41 data set.

Comparison of CELLO II and ALIGN

In Figure 4, we compare the predictive performances of CELLO II and ALIGN for the PS and PK data sets. The predictive performances of ALIGN are estimated as follows: we take the top hit from all-against-all alignment from ALIGN, and if the localization of the hit sequence is identical to that of the query sequence, it will be counted as a positive hit, or a negative hit. For the sake of comparison, we plot the prediction accuracies of both methods as a function of sequence identity. The procedures go as follows: assume that there are N sequences in the data set. By performing all-against-all sequence alignments, we obtain for any given sequence $N-1$ sequence identities si_i, where $i = 1 \ldots N-1$. The value $SI = \max(si_i)$ sets the upper limit of the sequence identity for the specific sequence sharing with the other sequences. The prediction accuracies of CELLO II and ALIGN for the sequences will be plotted against their associated SI. For both data sets,

TABLE II. Comparison of Predictive Performance of SVMs Based on Different Coding Schemes for the PK Data Set

	A_1	A_2	$X_4^{A_1}$	$X_5^{F_3}$	$X_5^{S_2}$	$X_5^{E_2}$	$X_5^{H_3}$	$X_5^{P_3}$	W_{13}	CELLO II
Chloroplast	62.0	67.4	72.0	56.8	66.5	69.3	57.5	59.6	69.6	79.9
Cytoplasmic	67.5	69.8	70.1	66.3	67.7	70.5	62.2	63.3	69.9	77.2
Cytoskeleton	60.0	47.5	65.0	45.0	45.0	47.5	40.0	35.0	67.5	67.5
ER	48.2	65.8	60.5	56.1	55.3	60.5	52.6	55.3	55.3	67.5
Extracellular	75.1	76.8	82.1	75.3	76.3	82.8	78.4	78.7	80.7	90.2
Golgi	17.0	21.3	38.3	29.8	29.8	36.2	23.4	27.7	27.7	53.2
Lysosomal	61.3	65.6	64.5	44.1	51.6	55.9	47.3	49.5	69.9	68.8
Mitochondrial	44.8	53.1	59.4	49.7	51.0	60.5	34.7	40.0	51.6	72.9
Nuclear	86.7	87.0	89.8	78.9	84.2	86.4	84.9	85.7	89.9	91.0
Peroxisomal	16.0	30.4	35.2	30.4	41.6	40.0	28.0	31.2	32.0	47.2
Plasma membrane	88.4	89.3	90.3	85.6	87.3	89.6	89.0	90.0	92.2	95.9
Vacuole	31.5	50.0	35.2	25.9	33.3	33.3	20.4	18.5	44.4	51.9
Overall	73.4	76.1	78.8	70.7	74.1	77.7	71.1	72.6	78.1	85.0

TABLE III. Comparison of Predictive Performance of SVMs Based on Different Coding Schemes for the PS Data Set

	A_1	A_2	$X_4^{A_1}$	$X_5^{F_3}$	$X_5^{S_2}$	$X_5^{E_2}$	$X_5^{H_3}$	$X_5^{P_3}$	W_{13}	CELLO II
Cytoplasm	86.7	82.7	90.3	80.9	81.3	80.6	82.0	79.1	84.5	95.3
Cytoplasmic	90.0	89.3	87.4	87.7	89.3	90.6	88.3	88.7	90.0	90.0
Periplasm	79.3	79.3	84.1	71.4	72.8	79.0	68.8	72.1	81.2	87.7
Outer membrane	90.5	92.8	91.3	86.2	89.0	91.6	83.6	85.9	88.5	92.8
Extracellular	76.8	74.7	78.9	66.8	71.1	74.7	61.6	67.4	76.3	79.5
Overall	85.7	85.2	87.3	80.1	82.1	84.6	78.6	80.1	85.0	90.0

we observe that, when the sequence identity is ≥30%, ALIGN performs slightly better than CELLO II does. However, the predictive performances of ALIGN drop considerably when sequence identity is below 20%, while the predictive performances CELLO II are consistent throughout the whole range of sequence similarity.

We compare the performances of CELLO II and ALIGN in each subcellular compartment for sequence identity ≥30% (Table IV) and <30% (Table V), respectively. For sequence identity ≥30%, ALIGN performs slightly better than CELLO II does. However, when sequence identity is <30%, CELLO II performs significantly better than ALIGN. For example, the MCCs of CELLO II for cytoplasmic and cytoplasmic membrane localizations are both 0.85, but those of ALIGN for these two localizations reach only 0.41 and 0.62, respectively. The MCCs of CELLO II are generally higher than those of ALIGN by 16–44% in the low homology region (i.e., sequence identity <30%).

Comparison with Other Approaches

The previous results suggest a simple hybrid procedure to predict subcellular localization: for a query sequence, we first use ALIGN to search against the data set composed of sequences of known subcellular localization. The localization of the top hit sequence that shares a 30% or greater sequence identity with the query sequence will be used as the prediction of its localization. These procedures suffer from the drawback that the sequence alignment results usually depend on the choice of sequence database and its corresponding annotation set.[36] In practice, we can use a

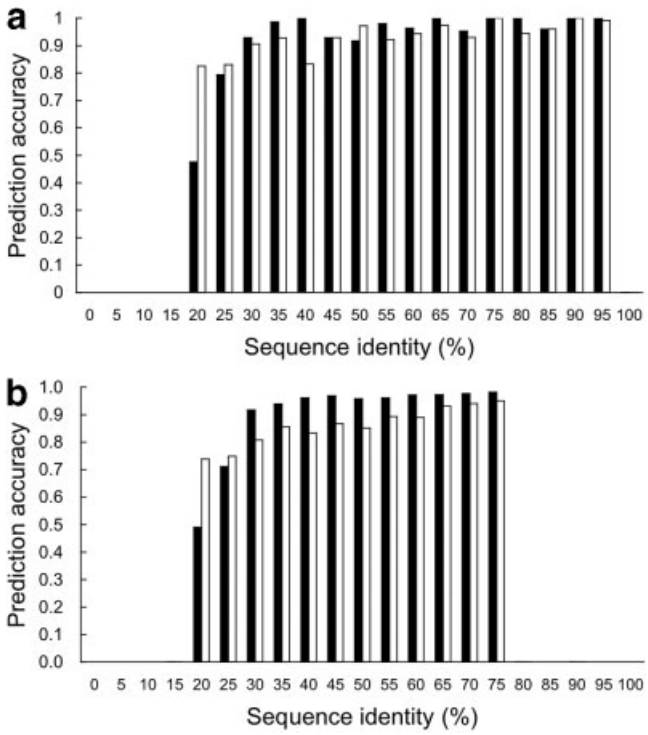

Fig. 4. The distribution of prediction accuracy as a function of sequence identity of both CELLO II (white bar) and ALIGN (black bar) for the (**A**) PS data set and (**B**) the PK data set. Note that we did not plot the prediction accuracies for those sequence identity bins that have relatively small example sizes as mentioned in the figure caption of Figure 2.

TABLE IV. Comparison of CELLO II and ALIGN for the Sequences with Sequence Identity ≥30% in the PS Data Set

	CELLO II		Align	
Localization	Accuracy	MCC	Accuracy	MCC
Cytoplasm	94.6	0.92	93.2	0.93
Cytoplasmic membrane	98.1	0.97	99.4	0.99
Periplasm	92.8	0.90	94.4	0.94
Outer membrane	96.5	0.96	99.4	0.99
Extracellular	90.6	0.90	98.6	0.98
Overall	94.9	—	97.7	—

TABLE V. Comparison of CELLO II and Align for the Sequences with Sequence Identity <30% in the PS Data Set

	CELLO II		ALIGN	
Localization	Accuracy	MCC	Accuracy	MCC
Cytoplasm	95.6	0.85	42.2	0.41
Cytoplasmic membrane	81.7	0.85	68.6	0.62
Periplasm	78.1	0.68	54.2	0.38
Outer membrane	77.3	0.72	81.3	0.46
Extracellular	49.0	0.56	43.1	0.40
Overall	82.6	—	56.3	—

more sophisticated similarity measure like HSSP distance developed by Nair and Rost.[21] On the other hand, we can always construct an updated data set comprising a large amount of sequences—like the SW41 data set. If ALIGN cannot identify any homologous sequences, we will use CELLO II to predict the subcellular localization of the query sequence. We will refer to this approach as HYBRID, because it combines CELLO II and ALIGN. Table VI compares the results of CELLO II, ALIGN, PSORTb 2, and HYBRID for the PS data set. All results are averaged over the fivefold crossvalidation. As expected, HYBRID gives the best overall performance (92%), then CELLO II (90%), followed by PSORTb 2 (83%) and ALIGN (81%). It is interesting to note that ALIGN appears to perform surprisingly well for the PS2 data set in comparison with the more sophisticated PSORTb 2. However, the good performances of ALIGN are due to the high homology bias inherent in the PS data set [see Fig. 2(A)]. On the other hand, it is noted that PSORTb 2 also contains a sequence comparison module SCL-BLAST, which performs a BLASTP search against the expanded PSORTdb database. CELLO II is the only method that does not rely on homology search. CELLO II performs especially well for the cytoplasmic localization, yielding a prediction accuracy 95% and MCC = 0.89, in comparison, PSORTb 2 yields 70% and 0.77 respectively, for the same localization.

In Table VII we compare the performances of HYBRID, CELLO II, ALIGN, and the PK method[12] for the eukaryotic PK data set. The PK method used SVM based on the compositions of both amino acids and amino acid pairs to predict protein subcellular localization. As expected, HYBRID gives the best overall performance (91.6%). ALIGN (85.8%) performs slightly better than CELLO II (85.0%). The PK method gives a 78.2% overall prediction accuracy. However, the good performance of ALIGN is due to the even higher homology levels of the PK data set [Fig. 2(B)]. In fact, when the homologous sequences (sequence identity ≥30%) are removed in the PK data set, the overall prediction accuracy of ALIGN drops to 57%.

It is interesting to note that all approaches perform well for some subcellular compartments and poorly for some other (Tables VI and VII). For example, all approaches perform well for subcellular compartments associated with membranes (cytoplasmic membrane or outer membrane in Table VI, and plasma membrane in Table VII). The good prediction accuracies are probably due to the distinct sequence features of the membrane proteins. Indeed, even the topology of the transmembrane proteins can be predicted with relatively good accuracy from protein sequences.[37–39] We also found that the nuclear, extracellular and chloroplast localization are among the best predicted in the eukaryotes (Table VII). On the other hand, Golgi and vacuole localizations are among the worst predicted in the eukaryotes. These poor performances are presumably due to the relatively small number of sequences in the data set and possibly multiple localizations of these sequences.

At present, our program does not deal with multiple subcellular localizations.[18,35] However, it is straightforward to extend our approach to the case of multiple localization; because our SVM output is, in fact, a probability distribution of subcellular localization, we can set a proper probability threshold to determine the possible subcellular compartment candidates.

DISCUSSION

Sequence similarity is useful in predicting subcellular localization for sequences sharing ≥30% sequence identity. We showed that the homology search method perform surprisingly well for two popular benchmark data sets. However, on closer inspection, these seemingly good performances are, in fact, due to the high homology levels inherent in the data sets. The performances deteriorate rapidly with the homologous sequences removed from the data sets. We have developed a two-level SVM system CELLO II to predict protein subcellular localization. Its performance is comparable to the homology search method in the high homology regions and better than the homology search method in the low homology regions. We showed that CELLO II performs better than other current methods. We have also developed a hybrid approach combining CELLO II and ALIGN, which may be applied to a wide range of sequence identity and thus provide a practical tool for biologists.

ACKNOWLEDGMENTS

This work is supported by National Science Council, National Research Program of Genomic Medicine, and the University System at Taiwan–Veteran General Hospital Grant. We are grateful to both the hardware and software supports of the Structural Bioinformatics Core Facility at National Chiao Tung University.

TABLE VI. Comparison of the Predictive Performance of Different Approaches in the Prediction of Subcellular Locations for the PS Data Set

	HYBRID		CELLO II		ALIGN		PSORTb 2[18]	
Localization	Accuracy	MCC	Accuracy	MCC	Accuracy	MCC	Accuracy	MCC
Cytoplasm	95.0	0.89	95.3	0.89	55.8	0.62	70.1	0.77
Cytoplasmic membrane	90.6	0.92	90.0	0.91	84.1	0.82	92.6	0.92
Periplasm	88.8	0.84	87.7	0.82	80.4	0.73	69.2	0.78
Outer membrane	95.1	0.93	92.8	0.90	95.9	0.81	94.9	0.95
Extracellular	85.3	0.87	79.5	0.82	83.7	0.82	78.9	0.86
Overall	91.6	—	90.0	—	81.1	—	82.6	—

TABLE VII. Comparison of Predictive Performance of Different Approaches in the Prediction of Subcellular Locations for the PK Dataset

	HYBRID		CELLO II		ALIGN		The PK method[12]	
Localization	Accuracy	MCC	Accuracy	MCC	Accuracy	MCC	Accuracy	MCC
Chloroplast	90.0	0.88	79.9	0.81	89.0	0.83	72.3	—
Cytoplasmic	84.4	0.81	77.2	0.71	81.6	0.77	72.2	—
Cytoskeleton	80.0	0.87	67.5	0.81	82.5	0.71	58.5	—
ER	80.7	0.85	67.5	0.78	85.1	0.82	46.5	—
Extracellular	93.5	0.93	90.2	0.88	91.3	0.87	78.0	—
Golgi	74.5	0.81	53.2	0.69	80.9	0.77	14.6	—
Lysosomal	87.1	0.89	68.8	0.78	83.9	0.81	61.8	—
Mitochondrial	80.5	0.80	72.9	0.72	74.8	0.73	57.4	—
Nuclear	94.5	0.90	91.0	0.83	88.3	0.86	89.6	—
Peroxisomal	74.4	0.80	47.2	0.63	80.0	0.76	25.2	—
Plasma membrane	96.1	0.96	95.9	0.94	88.1	0.89	92.2	—
Vacuole	64.8	0.75	51.9	0.66	64.8	0.72	25.0	—
Overall	90.3	—	85.0	—	85.8	—	78.2	—

REFERENCES

1. Nakai K, Kanehisa M. A knowledge base for predicting protein localization sites in eukaryotic cells. Genomics 1992;14:897–911.
2. Cedano J, Aloy P, Perez-Pons JA, Querol E. Relation between amino acid composition and cellular location of proteins. J Mol Biol 1997;266:594–600.
3. Nielsen H, Engelbrecht J, Brunak S, von Heijne G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng 1997;10:1–6.
4. Reinhardt A, Hubbard T. Using neural networks for prediction of the subcellular location of proteins. Nucleic Acids Res 1998;26: 2230–2236.
5. Chou KC, Elrod DW. Protein subcellular location prediction. Protein Eng 1999;12:107–118.
6. Emanuelsson O, Nielsen H, von Heijne G. ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. Protein Sci 1999;8:978–984.
7. Emanuelsson O, Nielsen H, Brunak S, von Heijne G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 2000;300:1005–1016.
8. Nakai K. Protein sorting signals and prediction of subcellular localization. Adv Protein Chem 2000;2000:277–344.
9. Hua S, Sun Z. Support vector machine approach for protein subcellular localization prediction. Bioinformatics 2001;17:721–728.
10. Gardy JL, Spencer C, Wang K, Ester M, Tusnady GE, Simon I, Hua S, deFays K, Lambert C, Nakai K, Brinkman FS. PSORT-B: improving protein subcellular localization prediction for Gram-negative bacteria. Nucleic Acid Res 2003;31:3613–3617.
11. Nair R, Rost B. Better prediction of sub-cellular localization by combining evolutionary and structural information. Proteins 2003; 53:917–930.
12. Park KJ, Kanehisa M. Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. Bioinformatics 2003;19:1656–1663.
13. Chou KC, Cai YD. Using GO-PseAA predictor to predict enzyme sub-class. Biochem Biophys Res Commun 2004;325:506–509.
14. Lu Z, Szafron D, Greiner R, Lu P, Wishart DS, Poulin B, Anvik J, Macdonell C, Eisner R. Predicting subcellular localization of proteins using machine-learned classifiers. Bioinformatics 2004;20: 547–556.
15. Scott MS, Thomas DY, Hallett MT. Predicting subcellular localization via protein motif co-occurrence. Genome Res 2004;14:1957–1966.
16. Yu CS, Lin CJ, Hwang JK. Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. Protein Sci 2004;13:1402–1406.
17. Bhasin M, Garg A, Raghava GP. PSLpred: prediction of subcellular localization of bacterial proteins. Bioinformatics 2005;21:2522–2524.
18. Gardy JL, Laird MR, Chen F, Rey S, Walsh CJ, Ester M, Brinkman FS. PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. Bioinformatics 2005;21:617–623.
19. Garg A, Bhasin M, Raghava GP. Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. J Biol Chem 2005;280:14427–14432.
20. Nielsen H, Engelbrecht J, von Heijne G, Brunak S. Defining a similarity threshold for a functional protein sequence pattern: the signal peptide cleavage site. Proteins 1996;24:165–177.
21. Nair R, Rost B. Sequence conserved for subcellular localization. Protein Sci 2002;11:2836–2847.
22. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 1993;232:584–599.
23. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
24. Hua S, Sun Z. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. J Mol Biol 2001;308:397–407.

25. Lin C-J. Formulations of support vector machines: a note from an optimization point of view. Neural Comput 2001;13:307–317.
26. Chang C-C, Lin C-J. LIBSVM v. 2.81: a library for support vector machines. (http://www.csie.ntu.edu.tw/~cjlin/libsvm) 2005.
27. Duan K, Keerthi SS, Poo AN. Evaluation of simple performance measures for tuning SVM hyperparameters. Neurocomputing 2003;51:41–59.
28. Yu CS, Wang JY, Yang JM, Lyu PC, Lin CJ, Hwang JK. Fine-grained protein fold assignment by support vector machines using generalized npeptide coding schemes and jury voting from multiple-parameter sets. Proteins 2003;50:531–536.
29. Chen YC, Lin YS, Lin CJ, Hwang JK. Prediction of the bonding states of cysteines using the support vector machines based on multiple feature vectors and cysteine state sequences. Proteins 2004;55:1036–1042.
30. Dubchak I, Muchnik I, Mayor C, Dralyuk I, Kim S-H. Recognition of a protein fold in the context of the structural classification of proteins (SCOP). Proteins 1999;35:401–407.
31. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 1975;405:442–451.
32. Myers EW, Miller W. Optimal alignments in linear space. Comput Appl Biosci 1988;4:11–17.
33. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, M. S. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acid Res 2003;31:365–370.
34. Chen YC, Hwang JK. Prediction of disulfide connectivity from protein sequences. Proteins 2005;61:507–512.
35. Lei Z, Dai Y. An SVM-based system for predicting protein subnuclear localizations. BMC Bioinformatics 2005;6:291.
36. Rost B. Enzyme function less conserved than anticipated. J Mol Biol 2002;318:595–608.
37. Krogh A, Larsson B, von Heijne G, Sonnhammer E. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 2001;305.
38. Tusnady GE, Simon I. The HMMTOP transmembrane topology prediction server. Bioinformatics 2001;17:849–850.
39. Adamian L, Liang J. Prediction of buried helices in multispan alpha helical membrane proteins. Proteins 2006;63:1–5.