

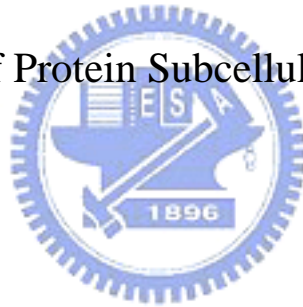
國立交通大學

生物科技系

博士論文

蛋白質於細胞位置之預測

Prediction of Protein Subcellular Localization



研究生：游景盛

指導教授：黃鎮剛 教授

中華民國九十六年七月

蛋白質於細胞位置之預測  
Prediction of Protein Subcellular Localization


研究生：游景盛

Student : Chin-Sheng Yu

指導教授：黃鎮剛

Advisor : Jenn-Kang Hwang

國立交通大學  
生物科技系  
博士論文



A Dissertation  
Submitted to Department of Biological Science and Technology  
College of Biological Science and Technology  
National Chiao Tung University  
in partial Fulfillment of the Requirements  
for the Degree of  
Ph.D.  
in

Biological Science and Technology

July 2007

Hsinchu, Taiwan, Republic of China

中華民國九十六年七月

# 蛋白質於細胞位置之預測

學生：游景盛

指導教授：黃鎮剛博士

國立交通大學 生物科技學系 博士班

## 摘 要

蛋白質在細胞體內的位置(subcellular localization)與其生理功能有著密不可分的關係，一般相信，在此有利的線索下，能幫助研究者更快速有效的分析該蛋白質的功能；在此需求下，近年來不斷的有相關的預測工具被發展出來，胺基酸與核甘酸序列數目快速增長的當前，透過計算工具直接針對序列作預測和分析尤其重要，而這些根據不同演算法(algorithms)所發展的各種方法差異極大，應用的序列來源物種和各細胞位置的預測結果也有很大的變異。在這篇論文中，首先我利用支持向量機(Support Vector Machines)，根據各種不同多樣特性的 n-peptide 組成份，並針對已知的幾個標準資料群作測試，都能比原有的方法得到更好的預測結果。接下來，為了持續改進原有的方法同時也希望深入探討這個方法的優缺點和限制，便更進一步利用序列比對的方式檢測這些標準資料群，發現目前各廣泛使用的資料群中均存在非常高比例的同源性序列(highly homologous sequences)，以至於造成高估的預測結果；而前人的研究中，如 Rost and Nair (Protein Sci, 11:2836-47 (2002))曾探討蛋白質序列的相似度與細胞位置的關係，即在相同細胞位置的蛋白質帶有著較為保留的胺基酸序列(conserved sequence)，對於序列和細胞位置的關係，明確劃定了一個序列比對可辨識的相似度界限(threshold)；同時我發展了一個雙層支持向量機(two-level support vector machine)系統，於第一層由胺基酸序列所轉換的不同特徵向量(feature vectors)製造數種有效的支持向量機分類器(SVM classifiers)，第二層將上述的分類器的預測結果由支持向量機結合，藉此得到一個預測某蛋白質可能細胞位置的機率分布值；再將序列比對與此方法分別應用於目前兩個非常常用的標準資料群，前者並全面性的兩兩配對比較其序列相似度(sequence identity)，與前人研究相符，在胺基酸序列相似度小於30%時，序列比對所能判別該蛋白質所在細胞位置的能力急劇降低，而雙層的支持向量機預測力則不受影響並遠優於序列相似度判別的效力，利用這樣的特性進一步將兩者有效結合，此合成方法得到了極佳的預測結果。我們將此方法所發展的工具建置成網頁伺服器，命名為 CELLO (subCELLular Localization predictive system)，供研究者使用，對於大量 high-throughput 的蛋白質體和基因體資料分析應有相當的幫助。

# Prediction of Protein Subcellular Localization

student : Chin-Sheng Yu

Advisors : Dr. Jenn-Kang Hwang

Department of Biological Science and Technology  
National Chiao Tung University

## ABSTRACT

Since the protein's function is usually related to its subcellular localization, the ability to predict subcellular localization directly from protein sequences will be useful to biologists to infer protein function. Recent years we have seen a surging interest in the development of novel computational tools to predict subcellular localization. With the rapid increase of sequenced genomic data, the need for an automated and accurate tool to predict subcellular localization becomes increasingly important. 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. In this thesis, I used support vector machine (SVM) method based on  $n$ -peptide composition in predicting the subcellular locations of proteins. For an unbiased assessment of the results, we apply our approach to several independent data sets in the beginning. In those data sets, our approach gives superior performance compared with other approaches. A number of authors have noticed that sequence similarity is useful in predicting subcellular localization. For example, Rost and Nair (Protein Sci, 11:2836-47 (2002)) have carried out extensive analysis of the relation between sequence similarity and identity in subcellular localization and 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 developed an approach based on a two-level 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

often-used 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 check the relationship between sequence homology and localization. Our results, which are consistent with previous studies, indicate that the homology search approach performs surprisingly well for sequences sharing homology as low as 30%, but its performance deteriorates considerably for sequences sharing lower sequence identity. A data set of high homology levels will obviously lead to biased assessment of the performances of the predictive approaches - especially those relying on homology search or sequence annotations. Since our two-level classification system based on SVM does not rely on homology search, its performance remains relatively unaffected by sequence homology. When compared with other approaches, our approach outperformed other existing approaches, even though some of which use homology search as part of their algorithms. Furthermore, for the practical purpose, we also develop a practical hybrid method that pipelines the two-level SVM classifier and the homology search method in sequential order as a general tool for the sequence annotation of subcellular localization. Our approaches should be valuable in the high throughput analysis of genomics and proteomics.



## Acknowledgement

終於還是到了離別的時候，有點感傷也有些期待，不能免俗的，要感謝所有在我身邊的人。

首先最感謝的，還是黃鎮剛教授，自考上清華開始就蒙他照顧，七年的時間不算短，特別感謝他的寬容與耐心，這樣不厭其煩的指導我這個笨學生這麼久，在他的指導下，引領我進入學術的殿堂，並提供了一個讓人安心研究、快樂工作的環境，還有豐富的資源讓學生充分發揮，充分的信任並鼓勵啟發我們發揮想像力，也時時提醒我們從事科學研究應保有的態度和熱情，這一切除了感謝，還是感謝。

許多實驗室夥伴早已畢業離開，在竹銘館共度的兩年時光令人難忘，鎮熊學長教導我這個門外漢許多該有的知識，學會使用及管理工作站和廣泛的生物資訊訊息，春吟和文忠的耐心解說，讓我好不容易學會程式語言，勇欣學長讓我聽聞了另一個領域的學術層次和嚴謹，讓我警惕自己的渺小，淵仁學長教了我許許多多事物並常帶我爬山，晚上在實驗室努力的時候還有朝鈞學長陪伴，玉菁、航申、理漢、星男還有大家一起去三棧溪，一同去賞鳥，雖然已是陳年往事，想到仍十分窩心。

玉菁、志豪、涵堃、志鵬、志杰、力彰、蔚倫、建華、存操、少偉、啟德、思樸、書璋，和各位的研究討論令我獲益良多，和大家一同出國開會，一起規劃行程，都是令人難忘的回憶；啟文和士中為了我的口試幫我張羅一切，讓我的口試順利進行；還要特別感謝禎祥學長常告訴我許多新知，並盡力為我們維護機器，讓我們有穩定快速的工作站可用。

研究的過程中接受了許許多多的師長照顧，如台大資工林智仁教授和王榮英學長解答我許多 SVM 的問題，楊進木教授、林彩雲教授、林立元教授和盧錦隆教授提供我許多建議和寶貴意見，楊勻良教授、彭慧玲教授、林志生教授和黃憲達教授在之前的口試也給予許多幫助，多次麻煩各位，非常的感謝您們。

詩欣，感謝你陪在我身邊度過這段求學過程，除了要忍受我任性的濫好人性格，還要包容與我相處的枯燥乏味，這段日子有苦有樂，辛苦你了；最後要感謝我的父母家人，沒有你們的支持，我無法順利完成學位，有你們的陪伴和鼓勵，我感到非常的幸福。

僅以此論文作為這段求學旅程的總結，並獻給所有曾幫助過我的人。

07年7月18日

## Contents

中文摘要	.....	i
Abstract	.....	ii
Acknowledgement	.....	iv
Contents	.....	v
Abbreviations	.....	vi
Chapter 1	GENERAL INTRODUCTION.....	1
Chapter 2	Prediction subcellular localization of proteins for different datasets by support vector machines based on n-peptide compositions INTRODUCTION.....	6
	MATERIALS AND METHODS.....	8
	RESULTS AND DISCUSSION.....	12
	CONCLUSION.....	17
Chapter 3	Improvement and analysis for prediction subcellular localization INTRODUCTION.....	20
	MATERIALS AND METHODS.....	21
	RESULTS.....	25
	DISCUSSION.....	31
References	.....	32
Tables	.....	36
Figures	.....	51

## Abbreviations

ER	: Endoplasmic Reticulum
HMM	: hidden Markov model
HSSP	: Homology derived Secondary Structure of Proteins
MCC	: Matthew's correlation coefficient
nm	: nanometer
RBF	: Radial Basis Function
SI	: Sequence Identity
SVM	: Support Vector Machine
SW41	: SWISSPROT protein sequence database release 41.0





# Chapter 1

## GENERAL INTRODUCTION

The cell is the basic unit of life. Based on the structure differences of cell, organisms can be divided into two broad groups, the prokaryotes (bacteria) and the eukaryotes (all other forms of life, like plants and animals), which the former live itself as a single cell and the later live with others as an organization with several degrees of differentiating complex. Both of the single one are so small that invisible to the human eye. Roughly, bacterial cells are in the size of a few micrometers ( $10^{-6}$  m), and eukaryotic cells are 10- to 20-fold larger for any single one.

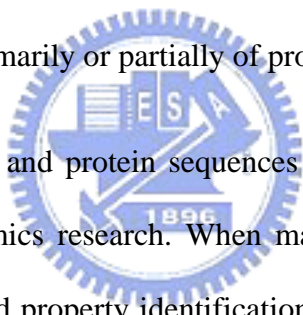
There are some appearance differences in structure between the two groups. One of the most fundamental distinctions is the real, membrane-enveloped nucleus to box the inheritance materials only for eukaryotic cells. In addition, most eukaryotic cells have many similar membrane-bound internal compartments term as organelles, and they are highly specialized for particular functions, whereas the prokaryotic cells do not have. Biologists can recognize them with the assistance of electron microscopy. By the way, essential materials for corresponding cellular processes are bounded and therefore high concentrations for those are kept inside. For example, chemical reactions in mitochondria involved in energy production, photosynthesis in chloroplasts involved in converting solar energy into sugars, proteins and lipids synthesized in endoplasmic reticulum (ER), secretory proteins and polysaccharides synthesized in Golgi complex, hydrolases storing and molecules digesting in lysosomes, something temporary storage or transport in vacuoles, and an internal framework called cytoskeleton, helps to establish cell shape and maintain cell moment and cell division.

However, most cellular functions act in the cytoplasm or on the plasma membrane of prokaryotes. And the inheritance materials gather near the center of the cell nakedly, instead of a special internal membrane-enclosed nucleus. There exist so many significant structural, biochemical, and genetic differences in cellular form that people distinguish them easily. And almost both researches are usually discussed separately.

Even relative few cause damages healthily or economically, most of the many thousand bacterial species known as harmless. Microbiologists study for medical diagnosis and treatment or other academic or industrial purposes through a series of experiments. Differential staining techniques were developed in order to isolate, enumerate, and identify targets from the samples. Some chemical materials force dye inside or outside the bacterial cells makes us distinguish them by color from others. For instance, the well known acid-fast stain is for the bacteria that cause tuberculosis. Besides, Gram staining is one of the most important and widely used staining techniques for bacteria. After a process of more than one dye solution smearing and washing out, Gram-positive bacteria retain the crystal violet dyed deep violet purple color; and Gram-negative bacteria lost the dye and appear red with safranin. Both cases just differ in thickness and substances composition of the cell wall. Simply, the cell use membrane as boundary to form a closed room against environment. The boundary of cytoplasmic area contains essential material for a whole life. In bacteria, there is usually a layer of cell wall outside the membrane. Thus the great difference in the staining appearance between the two bacterial groups makes us detect the structure of the cells definitely. The cell walls of Gram-negative ones are generally thinner (10 to 15 nm) than those of Gram-positive bacteria (20 to 25 nm). The former is obviously more complex seen by electron microscopy. An outer membrane inside the wall contains a thin layer of peptidoglycan. And there is also an additional periplasmic space between the cytoplasmic membrane and the outer

membrane. The later do not have this space, that is, there is no outer membrane inside their cell wall.

Cell structure and function diverge for different organisms of various size, shapes, and forms, even for different individual in the same body. Cells share common chemical molecules as building blocks and physical universe as interacting behaves. The major bases – proteins, nucleic acids, and polysaccharides are synthesized by series of chemical reactions for maintaining normal cellular organization and function. Most of them are too small or too thin to be seen under the light microscope and the sizes are about in the range of one to ten nanometers. Proteins are considered to be necessary everywhere in the cell. In addition to enzymes, proteins form the basis of most cellular structures. Connective tissue, muscle fibrils, cilia, flagella – all are made primarily or partially of proteins.



Tremendous amounts of DNA and protein sequences data have come out from experiment upon recent progress in genomics research. When many molecular sequences are in long-winded prospecting for role and property identification, much more mines are waiting in the process simultaneously. Hence, to develop useful computational tools to extract relevant biological information from sequences in a short time becomes even more important nowadays. Since 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 and interactions. Besides providing the clues of cell physiological properties, it will be helpful for the design of protein isolation and data analysis in experiment, furthermore, in medical researches.

In recent years, many efforts[1-21] have been made to predict protein subcellular locations based on the cell structure definition as described above. These approaches cover

various types of algorithms such as the knowledge-based expert system[15], the artificial neural networks[13, 16, 18], the support vector machines (SVM)[11, 17, 20, 21], the covariant discriminant algorithm[2, 5], or the Bayesian networks[8, 9, 19]. The most used features are the short N-terminal amino acid sequences[6, 7, 14-16] (i.e., the sorting or signal peptides), or the amino acid compositions[2, 5, 10, 11, 13, 17, 18] (or the general  $n$ -peptide compositions[20, 21]) derived from the whole amino acid sequences. Other approaches make use of additional information like sequence profiles derived from PSI-BLAST[1, 8-10], the ontology labels or the text annotations of the sequence databases[12, 19, 22]. In this thesis, we improved present approach through characteristics extraction from sequences and feedbacks to ensure some information correlating to protein localization.



## Chapter 2

Prediction for subcellular localization of proteins from different datasets by support vector machines based on  $n$ -peptide compositions



## INTRODUCTION

Some studies[11, 18] have shown that methods based on the amino acid composition appear to be more robust to errors in 5' gene annotation than those based on targeting sequences. Recently, Andrade et al[23] found that the total amino acid composition of the surface residues carries a signal that could help to identify the subcellular location, and they postulated that proteins in each location adapt their structures to their environmental variations throughout evolution. A number of studies[2, 5, 11, 18, 23] have shown that amino acid composition is a useful feature vector in the prediction of protein subcellular location as well as other protein global properties, such as protein folds[24, 25], disulfide bridges[26] and protein thermophilicity[27].

Reinhardt and Hubbard[18] applied neural networks to the prediction of subcellular location of proteins and obtained a prediction accuracy of 81% for three subcellular locations in prokaryotic organisms and 66% for four locations in eukaryotic proteins. Using the same data set, Hua and Sun obtained a prediction accuracy of 91.4% for prokaryotic organisms and 79.4% for eukaryotic organisms using SVM based on amino acid composition. Cedano et al[2] carried out a correlation analysis of the amino acid composition and the cellular location of five protein classes and have developed a program ProLock to predict the cellular locations of proteins. However, there are concerns that the methods based on the amino acid composition could have an intrinsic limitation on their predictive performance, because the amino acid composition does not have sequence-order information. Chou and coworkers developed approaches based on the pseudo amino acid composition[28], which is designed to include sequence correlation effects. For a data set[5, 28] comprising 12 location categories, the prediction accuracy reached 73.0% based on the pseudo amino acid composition, which is

significantly higher than those results based on general amino acid composition.

Gram-negative bacteria have five major subcellular localization sites that include the cytoplasm, the inner membrane, the outer membrane, the periplasm, and the extracellular space. PSORT I[29] has been the most widely used predictive tool for Gram-negative bacteria. However, it does not predict extracellular sequences, and its predictive performance reaches only 61% in overall prediction accuracy for a standard data set [9]. Recently Gardy et al.[9], combining different algorithms and input information, developed a multimodular method PSORT-B. This approach comprises six modules examining the query sequence specifically for different characteristics such as amino acid composition, similarity to proteins of known localization, presence of a signal peptide, transmembrane  $\alpha$ -helices, and motifs corresponding to specific localizations. This program then constructs a Bayesian network to generate a final probability value for each localization site. This approach yields an overall prediction accuracy of 75% for all location sites, significantly improving the previous results of PSORT I by 14%. However, despite the great improvement, PSORT-B gives modest prediction for some subcellular locations. For example, it gives a poor predictive accuracy of 58% for periplasmic sequences and 69% for cytoplasmic sequences.

Recently, we have developed an SVM method based on the  $n$ -peptide composition encoding scheme[25]. This coding scheme has the advantage of incorporating global sequence in a systematic way, which has been successfully applied to the prediction of protein folds[25]. In this work, we extend the approach to the prediction of protein subcellular locations. In order to get unbiased assessment of the results, we applied our approach to three independent data sets: the first set consisting of 997 prokaryotic proteins in three localization categories and 2427 eukaryotic proteins in four location categories[18]; the second set comprising 2191

proteins in 12 subcellular locations[30]; and the third set including 1443 protein sequences in five localization sites[9]. In those data sets, our approach gives superior performance (accuracy) compared with other approaches.

## MATERIALS AND METHODS

### Support Vector Machine (SVM)

Given training vectors  $x_i$ ,  $i=1,\dots,l$  and a vector  $y$  defined as:  $y_i=1$  if  $x_i$  is in one class, and  $y_i=-1$  if  $x_i$  is in the other class. The support vector technique[31] tries to find the separating hyperplane  $w^T x_i + b = 0$  with the largest distance between two classes, measured along a line perpendicular to this hyperplane. This requirement is equivalent to minimizing  $\frac{1}{2} w^T w$  with respect to  $w$  and  $b$  under the constraint that  $y_i(w^T x_i + b) \geq 1$ . However, in practice, these data to be classified may not be linearly separable. To overcome this difficulty, SVM non-linearly transforms the original input space into a higher dimensional feature space by  $\phi(x) = (\phi_1(x), \phi_2(x), \dots)$  and tries to minimize the object function  $\frac{1}{2} w^T w + C \sum_{i=1}^l \xi_i$  with respect to  $w$ ,  $b$  and  $\xi$ , under the constraint that  $y_i[w^T \phi(x_i) + b] \geq 1 - \xi_i$ , where  $\xi_i \geq 0$ . The function  $K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j)$  is usually called the kernel function. Note that training data  $x$  is mapped into a (possibly infinite) vector in a higher dimensional space; since in the higher dimensional space, it is more possible that data can be linearly separated. This procedure has the advantage of allowing training errors, since we do not require that training data should be always on the correct side of the separating hyperplane  $w^T x + b = 0$ , and we also try to minimize the training error  $\sum_{i=1}^l \xi_i$  in the objective function. In the end, the decision



function is written as  $f(x) = \text{sign}(w^T \phi(x) + b)$ . In other words, for a test vector  $w^T x + b > 0$ , we classified it to be in class 1. Otherwise, we think it is in the second class. Those  $x_i$ 's that are used to construct  $w$  and  $b$  are called support vectors.

In the training process, only part of the training data are used to construct the hyperplane, hence avoiding the overfitting problem usually plaguing other machine learning methods. These data constructing the classifier are called support vectors. The preliminary tests showed that the radial basis function (RBF) kernel gave results better than other kernels. Therefore, in this work we used the RBF kernel for all the experiments.

An important issue of optimizing SVMs is the selection of parameters. For SVM training, a few parameters such as the penalty parameter and the kernel parameter of the RBF function must be determined in advance. Choosing optimal parameters for SVM is an important step in SVM design. We used the cross-validation on different parameters for the model selection[32]. In this work, all SVM calculations were performed by using LIBSVM[33], a general library for support vector classification and regression.

### **Coding schemes**

In the past study[25], we have shown that protein descriptors based on the  $n$ -peptide composition are effective in predicting protein folds. For the simplest case  $n=1$ , the  $n$ -peptide composition were reduced to the amino acid composition, which can also be considered as a first-order approximation to the global protein sequence. When  $n=2$ , the  $n$ -peptide composition gives  $20 \times 20 = 400$  dipeptide composition. When  $n$  gets larger, the  $n$ -peptide compositions will cover more global sequence information, but at the same time, such a coding scheme becomes not only impractical from a computational viewpoint but also undoable from a learning viewpoint. However, the size problem can be overcome if we

regroup the amino acids into a smaller group of classes, according to their physico-chemical properties or the structural properties. In this work, we use the notation  $A_n$  to denote the  $n$ -peptide composition of amino acids;  $F_n$  to denote the reduced amino acid composition in which 20 amino acids are classified into four groups: polar, nonpolar, acid and base; and  $X_k$  to denote the partitioned amino acid composition in which the sequence is partitioned into  $k$  regions of equal length, and each partitioned sequence described by its amino acid composition are concatenated together. For example, the notation  $X_5$  denotes that the sequence is divided into 5 subsequences, each of which is encoded by  $A_1$  (note that  $X_1$  is equivalent to  $A_1$ ). Similar sequence coding schemes such as the  $n$ -gram hashing function has also been successfully applied to the protein classification[34, 35]. And further, amino acid composition can be substituted for reduced amino acid composition in this partitioned feature. Like  $F_n$  and  $X_k$ , the combined new one denotes as  $F_n X_k$ . These input vectors can be concatenated into one long input vector and fed into SVM. In this work, we used  $A + B + CD$  to denote three SVM classifiers, which are trained with input vector  $A$ ,  $B$  and the combined input vector  $CD$ , and the final prediction is decided by the jury votes from the classifiers.

### **Training and testing the SVM classifiers**

For multi-class SVM classification, we use the one-against-one (OAO) method[25]. Given  $J$  classes of subcellular locations, we construct  $J(J-1)/2$  SVM classifiers for a given type of input vector and train with proteins from two different subcellular locations. For each penalty parameter and kernel parameter, cross validation combining with the OAO method is used for estimating the performance of the model. Therefore, for each model,  $J(J-1)/2$  decision functions share the same parameter. Each protein in the test set always get a vote from each binary classifier. In the end, we used the jury voting to determine the final assignment of

locations to each sequence in the test set. In the case of identical votes, we gave more weight to the vote from  $X_4$  because of the better performance. The general architecture of our predictive system is shown in Figure 1. We use the standard  $Q_i$  percentage accuracy[36-38] and the Matthew's correlation coefficient[39]  $MCC$  for assessing the accuracy of subcellular location identification:  $Q_i = c_i/n_i$ , where  $n_i$  is the number of test data in the  $i^{th}$  subcellular location and  $c_i$  the number correctly predicted. The overall  $Q_i$  is given by  $Q_i = \sum_i^F w_i Q_i$ , where  $w = n_i/N$ . The Matthew's correlation coefficient  $MCC$  is given by[39]

$$MCC_i = \frac{tp_i tn_i - fp_i fn_i}{\sqrt{(tp_i + fn_i)(tp_i + fp_i)(tn_i + fp_i)(tn_i + fn_i)}}$$

where  $tp_i$  is the true positives in location  $i$ ,  $tn_i$  is the true negatives in location  $i$ ,  $fp_i$  is the false positives and  $fn_i$  is the false negatives.  $MCC_i$  is one for a perfect prediction, and zero for a completely random prediction.

We also assess the performance of the classifiers by the Jackknife test, which measure the prediction accuracy systematically by singling out one sequence as a test case from the data set during the training process and then testing the classifiers against this single protein. The Jackknife test is considered as the most objective and effective method in assessing the statistical prediction[40], and all our results reported here are done with the jackknife tests. All computations are done on a 48 double-CPU PC cluster running in the Linux system. And for convenience, we denoted our Subcellular Localization Predictive System as CELLO.

## Data sets

We use three independent data sets for the assessment of our approach. The first data set is

that of Reinhardt and Hubbard[18], called the RH dataset. The RH data set consists of two parts: the prokaryotic set and the eukaryotic set. The prokaryotic set includes 997 protein sequences in three categories: 688 cytoplasmic proteins, 202 periplasmic proteins and 107 extracellular proteins; the eukaryotic set includes 2427 sequences in four location categories: 1097 nuclear proteins, 684 cytoplasmic proteins, 321 mitochondrial proteins and 325 extracellular proteins. The pair sequence identities are less than 90% among the data set to avoid a bias towards large sequence families with high similarity. The second data set is from Chou and Elrod[5, 28], referred to as the CE data set. This data set contains 2191 protein sequences in 12 categories, which consists of 145 chloroplast proteins, 571 cytoplasm, 34 cytoskeleton, 49 endoplasmic reticulum (ER), 224 extracellular, 25 Golgi apparatus, 37 lysosome, 84 mitochondria, 272 nuclear proteins, 27 peroxisome, 699 plasma membrane, and 24 vacuole. In this data set, there are sequence pairs with sequence identity  $> 90\%$ , though the average sequence identity in each category is less than 12%. The third data set we used is the same with Gardy et al.[9], termed the PS 1.0 dataset (for the version 1.0), extracted from SWISS-PROT release 40.29[41]. This data set consists of 1443 protein sequences: 1302 proteins localized in a single subcellular site, which are 248 cytoplasmic, 268 inner membrane, 244 periplasmic, 352 outer membrane, and 190 extracellular. This data set also includes a further 141 proteins resident at multiple localization sites: 14 cytoplasmic/inner membrane, 50 inner membrane/periplasmic, and 77 outer membrane/extracellular.

## RESULTS AND DISCUSSION

### *The RH data set*

In order to have an unbiased assessment, we tested our approach on two data sets. The first one is the RH data set, a benchmark data set studied by a number of investigators[5, 11, 18, 42, 43]. In Table 1 and 2, we summarize the prediction accuracies using different input vectors by Jackknife tests on the eukaryotic and prokaryotic set, respectively. Only the best results of the given coding schemes are reported. For eukaryotic sequences (Table 1), the overall prediction accuracy with the portioned amino acid composition  $X_4$  is the best among the single parameter sets. The SVM based on the dipeptide composition  $A_2$  performs slightly better than based on the amino acid composition  $A_1$ . The SVM based on the reduced amino acid representation  $F_3X_5$  gives relatively poor prediction accuracy. The multiple input vectors  $A_1 + A_2 + X_4 + F_3X_3$  give prediction accuracies higher than those of the single parameter sets – the best overall prediction accuracy for eukaryotic sequences is 87.0%. Among the 4 subcellular locations, the prediction accuracy for the nuclear sequences reaches 96.0%. The prediction accuracies for mitochondrial location are relatively lower – the best prediction accuracy is 69.5%.

For prokaryotic proteins (Table 2), the single input vector  $A_1$ ,  $A_2$  and  $X_4$  already give excellent prediction accuracies around 91-92%, while both multiple input vectors gives the slightly better overall prediction accuracies. The prediction accuracy for cytoplasmic sequences can reach 99.7%. In general, the prediction accuracies for prokaryotic sequences are higher than those for eukaryotic sequences.

#### *Comparison with other approaches*

The RH data set has been studied by a number of investigators. There are the neural

network method described by Reinhardt and Hubbard[18], the SVM by Hua and Sun[11], the covariant discriminant method by Chou and Elrod[5], the Markov chains model by Yuan[43] and the nearest neighbor algorithm by Chou and Cai[42] using a hybrid method of the function domain composition and pseudo amino acid composition. Tables 3 and 4 summarize the predictive performance of these approaches for the RH data set. All results are obtained by the Jackknife tests except those of the Reinhardt and Hubbard, which are computed with six-fold validation. It should be also noted that Chou and Cai did not report any results for individual location category.

For the eukaryotic sequences (Table 3), our overall prediction accuracy is favored as compared to other approaches. Our prediction accuracy  $Q_i = 88.1\%$  is 21% higher than that of Reinhardt & Hubbard, 14% higher than that of Yuan, and 7.9% higher than that of Hua and Sun. For the subcellular locations, our prediction accuracy reaches 96.0% for the nuclear location – almost 10% higher than that of Hua & Sun. Our *MCC*'s for subcellular locations are also significantly higher than those of other approaches. For example, our value of *MCC* for cytoplasmic sequences is 0.80, which is higher than both Yuan (0.60), and Hua and Sun (0.64). For mitochondria locations, our prediction accuracy (69.5%) is similar to that of Yuan, but our correlation coefficient (0.77) for this location is significantly higher than that of Yuan (0.53). Though Chou & Cai obtained an overall prediction accuracy better than our approach (about 2%), they reported neither  $Q$  nor *MCC* for each subcellular location, so it is not possible to make any further comparison between these two approaches.

For the prokaryotic sequences (Table 4), our method gives the superior predictive performance in both  $Q$ 's and *MCC*'s. Our Matthews correlation coefficients for subcellular locations are higher than those of all other approaches. For example, our value of *MCC* for the

cytoplasmic location is 0.90, higher than both Yuan (0.83), and Hua and Sun (0.86). Though our prediction accuracy for periplasmic proteins is lower than that of Reinhardt and Hubbard, we like to mention that their results are obtained with six-fold cross validation instead of the Jackknife tests.

### *The CE dataset*

To include sequence-order effects, Chou has developed the coding scheme based on the pseudo amino acid composition[28], which consists of  $20 + \lambda$  discrete numbers, where the first 20 numbers are identical with those in the amino acid composition and the remaining numbers represent  $\lambda$  different ranks of sequence correlation factors. Using the pseudo amino acid composition, Chou[28] have extensively studied the CE data set by different approaches, such as the least Euclidean distance method developed by Nakashima et al[44], the ProtLock by Cedano et al[2] and the covariant discriminant approach[5] by Chou. Table 5 compares these results by Jackknife tests. This table lists complete only  $Q$ 's and  $MCC$ 's for subcellular locations by our approach, this is because that Chou[28] reported only the total prediction accuracy  $Q$ 's. Our approach gives the best overall prediction accuracy  $Q_t = 83.2\%$ , significantly higher than those of other approaches. Both  $Q$ 's and  $MCC$ 's vary considerably with regards to the locations. This is due to the highly uneven distributions of sequences in each location category (varying from 699 sequences in membrane to 24 in vacuole) – for example, the first five locations (as indicated by italics in the table) contain 87% of all sequences in 12 locations. If considered only the five most populated categories, our approach gives excellent results – the overall prediction accuracy is 90.3%, which is higher than 80.9% by Chou[28]. The CE data set, unlike the RH data set, includes many location categories that contain very small number of sequences (7 out of 12 locations contain only 13% of total

sequences), and this is the main reason that our prediction accuracies for these locations are generally poor. But we expect the prediction will greatly improve when more data are included. Despite the deficiency of the CE data set, our prediction accuracies for subcellular locations are still significantly higher than other approaches.

#### *The PS 1.0 dataset*

In Table 6, we compared the predictive performances of CELLO, PSORT I, PSORT-B, and SubLoc for five subcellular localization sites. Because the original SubLoc for prokaryotes predicts only three subcellular localization sites (cytoplasmic, periplasmic, and extracellular), we used the  $A_1$  SVM classifier for the current data set. The results are obtained with fivefold cross-validation. The overall prediction accuracy of CELLO reached 89%, which is 14% higher than that of PSORT-B, 28% higher than that of PSORT I, and 10% higher than that of SubLoc. In general, CELLO achieves better prediction accuracy for all subcellular localization sites than do the other approaches. Noticeably, our prediction accuracy for cytoplasmic location ( $Q_i = 91\%$ ) is 22% higher than that of PSORT-B, and for periplasmic location ( $Q_i = 87\%$ ) is 30% higher. These are very significant improvements on the previous results. In CELLO, the only prediction  $<80\%$  is for extracellular location ( $Q_i = 79\%$ ), but it is still 9% higher than that of PSORT-B. Although the prediction accuracy  $Q_i$  offers a convenient measure for predictive performances, one should be careful in drawing hasty conclusion from  $Q_i$ , because it overlooks overpredictions.  $MCC$ , taking into account of both under- and overpredictions, offers a complementary measurement for the predictive performances. For example, PSORT I gives a remarkable prediction accuracy,  $Q_i = 95\%$ , for inner membrane, but, due to overpredictions, it gives a less impressive  $MCC = 0.64$ , which is much lower than that of CELLO ( $MCC = 0.92$ ) and other approaches. CELLO also performs



better than other approaches in terms of *MCCs*. The *MCCs* of CELLO ranges consistently between 0.80 and 0.92, but the *MCCs* of PSORT-B deviate greatly among location sites (the difference between *MCCs* reached 0.24). PSORT-B gives a particularly poor prediction for periplasmic location (*MCC* = 0.69), compared with that of CELLO (*MCC* = 0.80). The inconsistent prediction accuracies of PSORT-B for different localization sites may reflect the uneven predictive performances of different modules in PSORT-B. It is also worth noting that even though PSORT-B uses different modules and input information tuned up for specific localization sites, CELLO, a single module approach, achieved better predictive performances. For example, PSORT-B uses HMMTOP[45] to predict inner membrane sequences, HMMTOP being a well-known hidden Markov model approach specifically designed to identify transmembrane proteins, but CELLO still gives better results,  $Q_i = 88\%$  and *MCC* = 0.92, compared with  $Q_i = 79\%$  and *MCC* = 0.85 obtained by PSORT-B. It is interesting to note that SubLoc shows a better overall performance than the more complicated multimodular PSORT-B. SubLoc can be seen as a special case of CELLO, because SubLoc uses amino acid compositions as the only input vectors. This surprisingly good predictive performances support previous observations that amino acid composition is indeed a good discriminator for subcellular localization.

## CONCLUSION

In this chapter, we apply the SVM approach based on  $n$ -peptide composition to the prediction of subcellular locations. For an unbiased assessment of the results, we test our approach by Jackknife tests on three independent data sets. Our approach yields significantly better prediction performance for all data sets than existing approaches in both overall prediction

accuracy and the correlation coefficients for associated subcellular locations. It is worth noting that our approach based on  $n$ -peptide composition also outperforms those approaches based on Markov chains model and pseudo amino acid composition, which include the order information. In addition, CELLO is a simple, straightforward implementation of a single module (SVM) based on multiple  $n$ -peptide composition to predict subcellular localization. It does not need specialized algorithms or particular input vectors for each subcellular localization site. Compared with CELLO, PSORT-B comprises six modules, with different modules examining specific localization sites, the results of which are then used to construct a Bayesian network to generate a final probability for localization sites. However, it is remarkable that CELLO gives significantly better predictive performances. Because CELLO is a simple straightforward implementation of SVM classifiers, one can easily extend CELLO to other organisms. An interesting question is whether CELLO, trained specifically for Gram-negative bacteria, can also predict heterologous expression of proteins in prokaryotic hosts. The availability of such predictive system would surely be helpful to researchers working on recombinant protein expression. Unfortunately, such study is presently hindered by the relatively scant amount of relevant testing data. However, it is expected that with more data accumulated in the future, such study will become more feasible. We have implemented a CELLO Web server, which is available at <http://cello.life.nctu.edu.tw>.

## Chapter 3

Improvement and analysis for prediction subcellular localization



## INTRODUCTION

Many efforts attempted to improve the prediction of protein subcellular localization through methods combination recently. For instance, some approaches make use of additional information like sequence profiles derived from PSI-BLAST[1, 8-10] or the ontology labels or the text annotations of the sequence databases[12, 19, 22]. In general, these approaches perform well for specific organisms and for certain localization categories. However, it is noticed that the benchmark data sets used for the assessment of the predictive performances of most methods usually contain highly homologous sequences. For example, the data set of Reinhardt and Hubbard[18] as well as that of Garg *et al*[10] include sequences up to 90% sequence identity, and the data set of Park and Kanehisa[17] comprises sequences up to 80% sequence identity. Several groups[46, 47] 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[46] 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[46], a measure for sequence similarity accounting for pairwise sequence identity and alignment length. It is well known in the study of secondary structure prediction[38, 48, 49] that the homologous sequences are meticulously removed from the testing-training data sets. For example, the popular benchmark RS126 set[38] comprises sequences that 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 over-prediction, i.e., the positive predictions may 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. 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 previous 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 denoted this 2-level SVM predictor of subcellular localization as CELLO II. Furthermore, using the relationship between sequence similarity and identity in subcellular localization[46, 47], we propose a practical pipeline approach combining CELLO II and the sequence alignment method to predict subcellular localization.

## MATERIALS AND METHODS



### Support Vector Machines

(Please see in MATERIALS AND METHODS in Chapter 2.)

### 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 discriminate protein properties in a number of applications[20, 21, 25, 50]. In the case of  $n = 1$ , the  $A_1$  coding is reduced 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 over-fitting. We can overcome the size problem by regrouping the amino acids into smaller number of classes according to their physico-chemical properties. In this work, we use the following classification schemes of the amino acids based on their physico-chemical properties - we use  $H_n$  for polar (RKEDQN), neutral (GASTPHY) and hydrophobic (CVLIMFW)[51];  $V_n$  for small (GASCTPD), medium (NVEQIL) and large van der Waals force (MHKFRYW)[51];  $Z_n$  for of low polarizability (GASDT), medium (CPNVEQIL) and high (KMHFRYW)[51];  $P_n$  for low polarity (LIFWCMVY), neutral (PATGS) and high polarity (HQRKNED)[51];  $F_n$  for acidic (DE), basic (HKR), polar (CGNQSTY) and nonpolar (AFILMPVW);  $S_n$  for acidic (DE), basic (HKR), aromatic (FWY), small hydroxyl (ST), sulfur-containing (CM) and aliphatic (AGPILV);  $E_n$  for acidic (DE), basic (HKR), aromatic (FWY), small hydroxyl (ST), sulfur-containing (CM), aliphatic 1 (AGP) and aliphatic 2 (ILV). For clarity, these coding schemes are summarized in Table 7.

### *The partitioned amino acid composition*

We use  $X_k^Y$  to denote the partitioned amino acid composition in which the sequence is partitioned into  $k$  subsequences of equal length, and each fragment encoded by the particular amino acid composition  $Y$ . For example, the notation  $X_5^{A_1}$  denotes that the sequence is divided into 5 subsequences, each of which is encoded by  $A_1$  (note that  $X_1^{A_1}$  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 of the flanking sequences of a given amino acid type. Note that when  $l$  is the length  $L$  of the whole sequence,  $W_L$  reduced to  $A_1$ .

### **The two level SVM classifier system**

The first level SVM classifiers comprise a number of separate SVM classifiers, each based on a specific sequence coding as described in the previous section. For the sake of notation simplicity, we use the coding symbol to represent the SVM classifier based on that coding. For example, we denote the SVM system comprising 3 classifiers, say,  $A$ ,  $B$  and  $C$  by the shorthand symbol  $A + B + C$ . In this work, the first level classifiers consist of the following

$$\text{SVMs: } \sum_{k=1}^9 X_k^{A_1} + \sum_{k=0}^6 D_k + \sum_{x \in S} X_5^x + \sum_{l \in S'} W_l, \text{ where } S = \{H_3, P_3, F_3, S_2, E_2\} \text{ and } S' = \{7, \dots, 15\} .$$

Each SVM generates a probability distribution [20, 21, 25] of the subcellular localization based on its particular sequence coding. A second SVM (i.e. the jury SVM) is used to process these probability distributions to generate the final probability distribution and the location

with the largest probability is used as the prediction. The two-level SVM system is shown schematically in Fig. 2.

### Performance assessment

Following the previous works[21, 25], 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  $i^{th}$  subcellular location and  $n_i$  is the number of sequences in that location. The overall prediction accuracy is given by

$$P = \sum_i f_i Q_i \quad (1)$$

where  $f_i = n_i/N$  and  $N$  is the total number of sequences. Though the percentage accuracy ( $Q_i$  or  $P$ ) provides a convenient measure for predictive performance, the Matthew's Correlation Coefficient[39] ( $MCC$ ) gives a more precise measurement for predictive performance:

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

where  $TP_i$  is the true positives in location  $i$ ,  $TN_i$  is the true negatives in location  $i$ ,  $FP_i$  is the false positives and  $FN_i$  is the false negatives. The value of  $MCC_i$  is 1 for a perfect prediction, 0 for a completely random prediction and  $-1$  for a perfect reverse correlation.

### The sequence-localization relationship

The query sequence is aligned against the data set of 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 as a negative hit. We performed all-against-all sequence alignment using the global alignment program ALIGN developed by Myers and Miller[52].

## **Data sets**

Two data sets were used in the experiment. The first data set, referred to as PS 2.0 data set (the version 2.0), is composed of Gram-negative sequences[8]. We selected from the data set only those sequences with a single localization (there are 4 groups of sequences with double localization, the average of which accounts for about 1% of the original data set). This resultant data set comprises 1444 protein sequences for five subcellular locations: extracellular (190), cytoplasmic (278), cytoplasmic membrane (309), periplasmic (276) and outer membrane (391). The second data set is from Park and Kanehisa, referred to as the PK dataset[17]. The sequences are selected from SWISSPROT[53] release 39.0 in such a way that the pair wise 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 space (862), Golgi apparatus (48), lysosome (93), mitochondria (727), nucleus (1932), peroxisome (125), plasma membrane (1677) and vacuoles proteins (54). We followed the same validation procedures for predictive performances as those of the previous works[8, 13, 17].

## **RESULTS**

### *The sequence-localization relationship*

The sequence homology of a data set can be easily inspected using the pair distribution of sequence identities, which shows the relative numbers of sequence pairs that share a given percentage sequence identity. Fig. 3 shows the pair distributions of the sequence identities of the PS (Fig. 3A) and the PK (Fig. 3B) 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  $\geq 30\%$  in both data sets. Using all-against-all sequence alignment through ALIGN, we compared the identity in sequence against the identity in localization for the PS and PK data sets (Figs. 4A and 4B). In general, when sequence identity  $\geq 25\%$ , the sequences usually share identical localization (however, in the PK data set, the abnormal behaviors of points at sequence identities  $\geq 80\%$  are due to the relatively smaller example sizes at those regions). We observe that the relationships between sequence identity and identity in localization are quite similar for these data sets. We also 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[13]. The resultant data set (referred to as SW41) comprises 9851 eukaryotic proteins sequences distributing in 5 locations: extracellular, cytoplasmic, mitochondria, nuclear and others. We also observe in the SW41 data set a similar relationship (Fig. 4C) between sequence identity and localization identity, i.e. when sequence identity  $\geq 25\%$ , it is very likely that the sequences also share identical localization.

#### *Comparison of different coding schemes*

Tables 8 and 9 compare the performances of different coding schemes for the PK and the PS data sets. The SVM based on the multiple coding schemes uses a second level SVM (i.e., CELLO II) to make the jury decision of the final prediction, while the SVM based on the

single parameter set use the output with the largest probability as the prediction. The general trends of the overall performances of the single parameter sets are quite similar for both data sets (Table 8 and 9). For certain rows of subcellular compartments, the performances of the single parameter sets fluctuate considerably. For example, in Table 8, the prediction accuracy for chloroplast ranges from 57% to 72 %. On the other hand, the single parameter sets perform similarly for some subcellular compartments like, for example, the plasma membrane (Table 8), for which the overall prediction accuracy ranges from 86% to 92%. The overall prediction accuracy with the partitioned amino acid composition  $X_4^{A_1}$  is the best among the single parameter sets for both data sets. In general, the results based on the multiple feature vector coding schemes are consistently better than those based on the single feature vector. This is probably due to the complementarity of information encoded in the single parameter sets. Our results are consistent with previous studies [3, 21, 25, 50, 54] that SVM based on the multiple feature vectors usually performs better than that based on the single feature vector.

#### *Comparison of CELLO II and ALIGN*

In Fig. 5, we compare the predictive performances of CELLO II and ALIGN for the PS and PK data sets, respectively. The predictive performances of ALIGN are estimated with simple procedures as follows. We take the top 1 hit from the all-against-all alignments from ALIGN and, if the localization of the hit sequence is identical to that of query sequence, it is counted as a positive hit, otherwise 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 can obtain for any given sequence  $N-1$  sequence identities  $si_i$ , where  $i = 1 \dots N-1$ . 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 are plotted against their associated *SIs*. For both data sets, we observe that, when the sequence identity is  $\geq 30\%$ , ALIGN generally performs slightly better than CELLO II does. However, the predictive performances of ALIGN drop considerably when sequence identity is below 20%. On the other hand, the predictive performances of CELLO II remain more consistent throughout the sequence identity range.

In Tables 10 and 11, we list the prediction accuracies of CELLO II and ALIGN for each individual subcellular location with sequence identity  $\geq 30\%$  and  $< 30\%$ , respectively. For sequence identity  $\geq 30\%$ , ALIGN performs slightly better than CELLO II does, though both CELLO II and ALIGN perform well. However, when sequence identity  $< 30\%$ , CELLO II performs significantly better than ALIGN. For example, the MCCs of CELLO II for cytoplasmic and cytoplasmic membrane localizations are both 0.85, while those of ALIGN for these two localizations are 0.41 and 0.62 in PS data set, respectively. The MCCs of CELLO II are in general 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 use ALIGN to search against the data set composed of sequences of known subcellular localization. As shown Fig. 4C, the accuracy shows a transition in the SW41 data set, which comprises 9851 sequences, at around 30% sequences identity from accurate distinction to inaccurate distinction. The localization annotation of the top hit sequence sharing a 30% or greater sequence identity with the query sequence was used as the prediction of its localization. However, Rost[55] have previously shown that the

results from sequence alignment really depends on the choice of sequence database and its corresponding annotation set[55]. In practice, we can use the more sophisticated similarity measure like HSSP distance developed by Nair and Rost[46]. Also, we can always construct an updated sequence database comprising large amount of sequences– for example, like the SW41 data set with the sequences of dubious annotations removed. If ALIGN fails, the *ab initio* CELLO II will be used to predict subcellular localization of the query sequence. We simply refer the approach as HYBRID. Table 12 compares the results of CELLO II, ALIGN and PSORTb 2 and the hybrid method for the PS data set. All results are averaged over the 5-fold cross validation. As expected, the hybrid method gives the best overall prediction accuracy (92%), followed by CELLO II (90%), 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 that of PSORTb 2. However, the good performances of ALIGN are due to the relatively high sequence homology bias inherent in the PS data set (see Fig. 3A). However, it is noted that PSORTb 2 also contains a sequence comparison module SCL-BLAST, which performs a BLASTP search against the expanded PSORTdb database of known localization. Among these methods, CELLO II is the only method that does not rely on homology search; however, its performance is still among the best next only to HYBRID. CELLO II performs especially well for the cytoplasmic localization, yielding a prediction accuracy 95% and MCC 0.89 (in comparison, PSORTb 2 gives 70% and 0.77, respectively, for the same localization).

In Table 13 we compare the results of HYBRID, CELLO II, ALIGN and the PK method for the eukaryotic PK data set. The PK method used SVM based on compositions of amino acids and amino acid pairs to predict protein subcellular localization. HYBRID as expected 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 obviously due to the even higher homology levels of the PK data set (Fig. 3B). In fact, when the homologous sequences (sequence identity  $\geq 30\%$ ) are removed in the PK data set, the overall prediction accuracy of ALIGN drops to 57%. Both CELLO II and the PK method belong to the class of the *ab initio* methods and do not rely on homology search; CELLO II performs significantly better.

It is interesting to note that different approaches produce quite similar trends of prediction accuracies for a number of subcellular compartments (Tables 12 and 13). For example, all approaches perform well for subcellular compartments associated with membranes (cytoplasmic membrane or outer membrane in Table 12, and plasma membrane in Table 13). 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[45, 56, 57]. We also found that the nuclear, extracellular and chloroplast localization are among the best predicted in the eukaryotes (Table 13). On the other hand, Golgi and vacuole are among the worst predicted in the eukaryotes. The poor performances are probably due to the relatively small number of sequences in the data set and the possible multiple localizations of these sequences. It is expected that the prediction will improve when more sequence data with reliable localization annotations are coming in.

At present, our program does not deal with proteins with multiple subcellular localizations[8, 54]. It is quite straightforward to extend our approach to the cases of multiple localizations. Since our output is in fact a probability distribution of a given set of localizations, it is possible, instead of taking the single one with highest probability, to set a probability

threshold value to determine the possible multiple localizations.

## DISCUSSION

In this work, we found that sequence identity is quite useful in identifying subcellular localization of homologous sequences down to 25% sequence identity. Furthermore, we found that the homology search method gives surprisingly good results for the two popular benchmark data sets. However, on closer inspection, these good performances are in fact due to the relatively high homology levels in the data sets and the performances go down drastically when the homologous sequences are removed from the data sets. We have developed an *ab initio* approach CELLO II based on a 2-level SVM system to predict protein subcellular localization. Its performance is comparable to the homology search method in the high homology regions and much superior to the homology search method for sequences in the low homology regions. We also showed that CELLO II performs better than other current methods for these data sets. For practical purpose, we also develop a hybrid approach combining CELLO II and the sequence alignment method, which may be applied to a wide range of sequence identity and provide a useful tool for biologists.

## References

1. Bhasin, M., A. Garg, and G.P. Raghava, *PSLPred: prediction of subcellular localization of bacterial proteins*. *Bioinformatics*, 2005. **21**(10): p. 2522-4.
2. Cedano, J., et al., *Relation between amino acid composition and cellular location of proteins*. *J Mol Biol*, 1997. **266**(3): p. 594-600.
3. Chen, Y.C., et al., *Prediction of the bonding states of cysteines using the support vector machines based on multiple feature vectors and cysteine state sequences*. *Proteins*, 2004. **55**(4): p. 1036-42.
4. Chou, K.C., *Prediction of protein subcellular locations by incorporating quasi-sequence-order effect*. *Biochem Biophys Res Commun*, 2000. **278**(2): p. 477-83.
5. Chou, K.C. and D.W. Elrod, *Protein subcellular location prediction*. *Protein Eng*, 1999. **12**(2): p. 107-18.
6. Emanuelsson, O., et al., *Predicting subcellular localization of proteins based on their N-terminal amino acid sequence*. *J Mol Biol*, 2000. **300**(4): p. 1005-16.
7. Emanuelsson, O., H. Nielsen, and G. von Heijne, *ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites*. *Protein Sci*, 1999. **8**(5): p. 978-84.
8. Gardy, J.L., et al., *PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis*. *Bioinformatics*, 2005. **21**(5): p. 617-23.
9. Gardy, J.L., et al., *PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria*. *Nucleic Acids Res*, 2003. **31**(13): p. 3613-7.
10. Garg, A., M. Bhasin, and G.P. Raghava, *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**(15): p. 14427-32.
11. Hua, S. and Z. Sun, *Support vector machine approach for protein subcellular localization prediction*. *Bioinformatics*, 2001. **17**(8): p. 721-8.
12. Lu, Z., et al., *Predicting subcellular localization of proteins using machine-learned classifiers*. *Bioinformatics*, 2004. **20**(4): p. 547-56.
13. Nair, R. and B. Rost, *Better prediction of sub-cellular localization by combining evolutionary and structural information*. *Proteins*, 2003. **53**(4): p. 917-30.
14. Nakai, K., *Protein sorting signals and prediction of subcellular localization*. *Adv Protein Chem*, 2000. **54**: p. 277-344.
15. Nakai, K. and M. Kanehisa, *A knowledge base for predicting protein localization sites in eukaryotic cells*. *Genomics*, 1992. **14**(4): p. 897-911.



16. Nielsen, H., et al., *Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites*. Protein Eng, 1997. **10**(1): p. 1-6.
17. Park, K.J. and M. Kanehisa, *Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs*. Bioinformatics, 2003. **19**(13): p. 1656-63.
18. Reinhardt, A. and T. Hubbard, *Using neural networks for prediction of the subcellular location of proteins*. Nucleic Acids Res, 1998. **26**(9): p. 2230-6.
19. Scott, M.S., D.Y. Thomas, and M.T. Hallett, *Predicting subcellular localization via protein motif co-occurrence*. Genome Res, 2004. **14**(10A): p. 1957-66.
20. Yu, C.S., et al., *Prediction of protein subcellular localization*. Proteins, 2006. **64**(3): p. 643-51.
21. Yu, C.S., C.J. Lin, and J.K. Hwang, *Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions*. Protein Sci, 2004. **13**(5): p. 1402-6.
22. Chou, K.C. and Y.D. Cai, *Using GO-PseAA predictor to predict enzyme sub-class*. Biochem Biophys Res Commun, 2004. **325**(2): p. 506-9.
23. Andrade, M.A., S.I. O'Donoghue, and B. Rost, *Adaptation of protein surfaces to subcellular location*. J Mol Biol, 1998. **276**(2): p. 517-25.
24. Dubchak, I., S.R. Holbrook, and S.H. Kim, *Prediction of protein folding class from amino acid composition*. Proteins, 1993. **16**(1): p. 79-91.
25. Yu, C.S., et al., *Fine-grained protein fold assignment by support vector machines using generalized npeptide coding schemes and jury voting from multiple-parameter sets*. Proteins, 2003. **50**(4): p. 531-6.
26. Mucchielli-Giorgi, M.H., S. Hazout, and P. Tuffery, *Predicting the disulfide bonding state of cysteines using protein descriptors*. Proteins, 2002. **46**(3): p. 243-9.
27. Kreil, D.P. and C.A. Ouzounis, *Identification of thermophilic species by the amino acid compositions deduced from their genomes*. Nucleic Acids Res, 2001. **29**(7): p. 1608-15.
28. Chou, K.C., *Prediction of protein cellular attributes using pseudo-amino acid composition*. Proteins, 2001. **43**(3): p. 246-55.
29. Nakai, K. and M. Kanehisa, *Expert system for predicting protein localization sites in gram-negative bacteria*. Proteins, 1991. **11**(2): p. 95-110.
30. Chou, K.C. and Y.D. Cai, *Using functional domain composition and support vector machines for prediction of protein subcellular location*. J Biol Chem, 2002. **277**(48): p. 45765-9.

31. Vapnik, V., *The nature of statistical learning theory*. 1995, New York: Springer.
32. Duan, K., Keerthi, S. S. and Poo, A. N., *Evaluation of simple performance measures for tuning SVM hyperparameters*. Neurocomputing, 2003. **51**: p. 41-59.
33. Chang, C.C.a.L., C. J., *LIBSVM: a library for support vector machines. Software available from <http://www.csie.ntu.edu.tw/~cjlin/libsvm>*. 2001.
34. Wu, C., et al., *Protein classification artificial neural system*. Protein Sci, 1992. **1**(5): p. 667-77.
35. Wu, C.H., et al., *Motif identification neural design for rapid and sensitive protein family search*. Comput Appl Biosci, 1996. **12**(2): p. 109-18.
36. Baldi, P., et al., *Assessing the accuracy of prediction algorithms for classification: an overview*. Bioinformatics, 2000. **16**(5): p. 412-24.
37. Ding, C.H. and I. Dubchak, *Multi-class protein fold recognition using support vector machines and neural networks*. Bioinformatics, 2001. **17**(4): p. 349-58.
38. Rost, B. and C. Sander, *Prediction of protein secondary structure at better than 70% accuracy*. J Mol Biol, 1993. **232**(2): p. 584-99.
39. Matthews, B.W., *Comparison of the predicted and observed secondary structure of T4 phage lysozyme*. Biochim Biophys Acta, 1975. **405**(2): p. 442-51.
40. Mardia, K.V., Kent, J. T. and Bibby, J. M., *Multivariate analysis*. London; Academic Press, 1979. **322**: p. 381.
41. Bairoch, A. and R. Apweiler, *The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000*. Nucleic Acids Res, 2000. **28**(1): p. 45-8.
42. Cai, Y.D. and K.C. Chou, *Nearest neighbour algorithm for predicting protein subcellular location by combining functional domain composition and pseudo-amino acid composition*. Biochem Biophys Res Commun, 2003. **305**(2): p. 407-11.
43. Yuan, Z., *Prediction of protein subcellular locations using Markov chain models*. FEBS Lett, 1999. **451**(1): p. 23-6.
44. Nakashima, H., K. Nishikawa, and T. Ooi, *The folding type of a protein is relevant to the amino acid composition*. J Biochem (Tokyo), 1986. **99**(1): p. 153-62.
45. Tusnady, G.E. and I. Simon, *The HMMTOP transmembrane topology prediction server*. Bioinformatics, 2001. **17**(9): p. 849-50.
46. Nair, R. and B. Rost, *Sequence conserved for subcellular localization*. Protein Sci, 2002. **11**(12): p. 2836-47.
47. Nielsen, H., et al., *Defining a similarity threshold for a functional protein sequence pattern: the signal peptide cleavage site*. Proteins, 1996. **24**(2): p. 165-77.

48. Hua, S. and Z. Sun, *A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach*. J Mol Biol, 2001. **308**(2): p. 397-407.
49. Jones, D.T., *Protein secondary structure prediction based on position-specific scoring matrices*. J Mol Biol, 1999. **292**(2): p. 195-202.
50. Chen, Y.C. and J.K. Hwang, *Prediction of disulfide connectivity from protein sequences*. Proteins, 2005. **61**(3): p. 507-12.
51. Dubchak, I., et al., *Recognition of a protein fold in the context of the Structural Classification of Proteins (SCOP) classification*. Proteins, 1999. **35**(4): p. 401-7.
52. Myers, E.W. and W. Miller, *Optimal alignments in linear space*. Comput Appl Biosci, 1988. **4**(1): p. 11-7.
53. Boeckmann, B., et al., *The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003*. Nucleic Acids Res, 2003. **31**(1): p. 365-70.
54. Lei, Z. and Y. Dai, *An SVM-based system for predicting protein subnuclear localizations*. BMC Bioinformatics, 2005. **6**: p. 291.
55. Rost, B., *Enzyme function less conserved than anticipated*. J Mol Biol, 2002. **318**(2): p. 595-608.
56. Adamian, L. and J. Liang, *Prediction of buried helices in multispan alpha helical membrane proteins*. Proteins, 2006. **63**(1): p. 1-5.
57. Krogh, A., et al., *Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes*. J Mol Biol, 2001. **305**(3): p. 567-80.

Table 1. Comparison of predictive performances using different n-peptide coding schemes by jackknife test for the RH eukaryotic data set.

Localizations	$A_1$	$A_2$	$S_1$	$F_3X_5$	$X_4$	$A_1+A_2+X_4$	$A_1+A_2+X_4+F_3X_5$
Cytoplasm	79.2	79.8	79.2	76.2	82.5	85.1	86.3
Extracellular	80.6	79.7	78.8	73.8	84.0	84.3	84.0
Mitochondria	59.2	62.6	60.1	53.9	66.7	63.2	69.5
Nucleus	90.5	93.9	90.3	86.8	94.3	96.0	96.0
Overall accuracy	81.9	83.9	81.7	77.7	86.0	87.0	<b>88.1</b>



Table 2. Comparison of predictive performances using different  $n$ -peptide coding schemes by jackknife tests for the RH prokaryotic data set.

Localizations	$A_1$	$A_2$	$S_1$	$F_5$	$X_4$	$A_1+A_2+X_4$	$A_1+A_2+X_4+F_3X_3$
Cytoplasm	98.7	98.1	98.8	94.2	98.0	99.3	99.7
Periplasm	76.7	73.8	76.2	60.9	81.2	80.7	80.2
Extracellular	75.7	77.6	74.8	72.9	72.9	76.6	75.7
Overall	91.8	91.0	91.7	85.2	91.9	93.1	<b>93.2</b>



Table 3. Comparison of different approaches in the prediction of subcellular localizations for the RH eukaryotic sequences

Localizations <sup>†</sup>	CELLO		Reinhardt & Hubbard*[18]		Yuan[43]		Hua & Sun[11]		Chou & Cai[42]	
	<i>Q</i> (%)	MCC	<i>Q</i> (%)	MCC	<i>Q</i> (%)	MCC	<i>Q</i> (%)	MCC	<i>Q</i> (%)	MCC
Cytoplasm (1097)	86.3	0.80	55	–	78.1	0.60	76.9	0.64	–	–
Extracellular (325)	84.0	0.89	75	–	62.2	0.63	80.0	0.78	–	–
Mitochondria (321)	69.5	0.77	61	–	69.2	0.53	56.7	0.58	–	–
Nucleus (1097)	96.0	0.83	72	–	74.1	0.68	87.4	0.75	–	–
Overall	<b>88.1</b>		66		73.0		79.4		90.4	

<sup>†</sup>The number of sequences is indicated in the parenthesis.

\*The results are obtained with 6-fold cross validation.

Table 4. Comparison of different approaches in the prediction of subcellular localizations for the RH prokaryotic sequences

Localizations†	CELLO		Reinhardt & Hubbard*[18]		Yuan[43]		Hua & Sun[11]		Chou & Cai[42]	
	<i>Q</i> (%)	<i>MCC</i>	<i>Q</i> (%)	<i>MCC</i>	<i>Q</i> (%)	<i>MCC</i>	<i>Q</i> (%)	<i>MCC</i>	<i>Q</i> (%)	<i>MCC</i>
Cytoplasm(688)	99.7	0.90	80	–	93.6	0.83	97.5	0.86	–	–
Periplasm(202)	80.2	0.81	85	–	79.7	0.69	78.7	0.78	–	–
Extracellular(107)	75.7	0.81	77	–	77.6	0.77	75.7	0.77	–	–
Overall	<b>93.1</b>	–	80.9	–	89.1	–	91.4	–	89.3	–

†The number of sequences is indicated in the parenthesis

\*The results are obtained with 6-fold cross validation, while all other are obtained with the Jackknife tests.

Table 5. Comparison of different approaches in the prediction of subcellular localizations by jackknife tests on CE data set

Localizations <sup>†</sup>	CELLO <sup>§</sup>		ProtLock[2, 28]	Least Euclidean distance[28, 44]	Augmented covariant-discriminant[4, 28]
	$Q(\%)$	MCC	$Q(\%)/MCC$	$Q(\%)/MCC$	$Q(\%)/MCC$
<i>Plasma membrane (699)</i>	95.6	0.93	–	–	–
<i>Cytoplasm (571)</i>	95.1	0.77	–	–	–
<i>Nucleus (272)</i>	89.8	0.80	–	–	–
<i>Extracellular (224)</i>	75.1	0.75	–	–	–
<i>Chloroplast (145)</i>	70.7	0.81	–	–	–
<i>Mitochondria (84)</i>	38.1	0.59	–	–	–
ER (49)	37.7	0.60	–	–	–
Lysosome (37)	34.2	0.54	–	–	–
Cytoskeleton (34)	36.1	0.60	–	–	–
Golgi (25)	19.2	0.40	–	–	–
Peroxisome (27)	33.3	0.58	–	–	–
Vacuole (24)	24.0	0.45	–	–	–
$Q_i(\%)$	<b>83.2</b>	–	48.7	49.1	73.0

<sup>†</sup>The number of sequences is indicated in the parenthesis. <sup>§</sup>Using  $A_1 + A_2 + X_4 + F_3X_5$ .



Table 6. The comparison of predictive performances of different approaches in the prediction of subcellular localization for Gram-negative bacteria (PS 1.0 data set)

Localizations	CELLO		PSORT-B		PSORT I		SubLoc <sup>a</sup>	
	<i>Q</i> (%)	<i>MCC</i>	<i>Q</i> (%)	<i>MCC</i> <sup>b</sup>	<i>Q</i> (%)	<i>MCC</i> <sup>b</sup>	<i>Q</i> (%)	<i>MCC</i>
Cytoplasm	90.7	0.85	69.4	0.79	75.4	0.58	75.0	0.74
Inner membrane	88.4	0.92	78.7	0.85	95.1	0.64	82.8	0.89
Periplasm	86.9	0.80	57.6	0.69	66.4	0.55	68.9	0.71
Outer membrane	94.6	0.90	90.3	0.93	54.5	0.47	89.1	0.86
Extracellular	78.9	0.82	70.0	0.79	–	–	69.5	0.78
Overall	<b>88.9</b>	–	74.8	–	60.9	–	78.5	–

<sup>a</sup> The original SubLoc for prokaryotes predicts only three subcellular localization sites, therefore, we retrained the  $A_1$  SVM for this data set using the one-against-one method, which is different from the original one-against-all method.

<sup>b</sup> *MCCs* are calculated using the precision and recall values reported in Gardy et al.[9].

Accuracy is in %.

Table 7. The coding schemes of the amino acids compositions based on different classification definitions

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

Table 8. Comparison of predictive performances (accuracies) using 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
Cytoplasm	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
Lysosome	61.3	65.6	64.5	44.1	51.6	55.9	47.3	49.5	69.9	68.8
Mitochondria	44.8	53.1	59.4	49.7	51.0	60.5	34.7	40.0	51.6	72.9
Nucleus	86.7	87.0	89.8	78.9	84.2	86.4	84.9	85.7	89.9	91.0
Peroxisome	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 9. Comparison of predictive performances (*accuracies*) using different coding schemes for the PS 2.0 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 Membrane	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



Table 10-1. Comparison of CELLO II and ALIGN for the sequences with sequence identity  $\geq 30\%$  in the PS 2.0 data set

Localization	Amount	CELLO II		ALIGN*	
		<i>Accuracy</i>	<i>MCC</i>	<i>Accuracy</i>	<i>MCC</i>
Cytoplasm	74	94.6	0.92	93.2	0.93
Cytoplasmic Membrane	156	98.1	0.97	99.4	0.99
Periplasm	180	92.8	0.90	94.4	0.94
Outer Membrane	316	96.5	0.96	99.4	0.99
Extracellular	139	90.6	0.90	98.6	0.98
Overall	865	94.9	-	97.7	-

\*The localization annotation of the top hit of the alignment list is used as the predicted localization.



Table 10-2. Comparison of CELLO II and ALIGN for the sequences with sequence identity  $\geq 30\%$  in the PK data set.

Localization	Amount	CELLO II		ALIGN*	
		<i>Accuracy</i>	<i>MCC</i>	<i>Accuracy</i>	<i>MCC</i>
Chloroplast	602	83.2	0.85	94.5	0.94
Cytoplasm	991	84.6	0.79	93.5	0.91
Cytoskeleton	31	80.7	0.90	96.8	0.97
ER	98	77.6	0.86	92.9	0.93
Extracellular	730	93.8	0.92	97.7	0.98
Golgi	35	65.7	0.79	94.3	0.93
Lysosome	77	71.4	0.81	93.5	0.93
Mitochondria	539	78.7	0.79	88.9	0.89
Nucleus	1358	94.0	0.89	99.0	0.99
Peroxisome	103	57.3	0.70	90.3	0.89
Plasma membrane	984	99.2	0.97	99.6	0.99
Vacuole	39	71.8	0.78	89.7	0.89
Overall	5587	88.9	-	96.0	-

\*The localization annotation of the top hit of the alignment list is used as the predicted localization.

Table 11-1. Comparison of CELLO II and ALIGN for the sequences with sequence identity < 30% in the PS 2.0 data set.

Localization	Amount	CELLO II		ALIGN*	
		<i>Accuracy</i>	<i>MCC</i>	<i>Accuracy</i>	<i>MCC</i>
Cytoplasm	204	95.6	0.85	42.2	0.41
Cytoplasmic Membrane	153	81.7	0.85	68.6	0.62
Periplasm	96	78.1	0.68	54.2	0.38
Outer Membrane	75	77.3	0.72	81.3	0.46
Extracellular	51	49.0	0.56	43.1	0.40
Overall	579	82.6	-	56.3	-

\*The localization annotation of the top hit of the alignment list is used as the predicted localization.



Table 11-2. Comparison of CELLO II and ALIGN for the sequences with sequence identity < 30% in the PK data set

Localization	Amount	CELLO II		ALIGN*	
		<i>Accuracy</i>	<i>MCC</i>	<i>Accuracy</i>	<i>MCC</i>
Chloroplast	69	50.7	0.47	40.6	0.25
Cytoplasm	250	48.0	0.42	34.4	0.24
Cytoskeleton	9	22.2	0.38	33.3	0.20
ER	16	6.25	0.12	37.5	0.30
Extracellular	131	70.2	0.66	55.7	0.41
Golgi	12	16.7	0.29	41.7	0.37
Lysosome	16	56.3	0.65	37.5	0.32
Mitochondria	188	56.4	0.53	34.6	0.27
Nucleus	574	83.8	0.68	63.1	0.54
Peroxisome	22	0	0	31.8	0.26
Plasma membrane	691	91.2	0.89	71.8	0.70
Vacuole	15	0	0	0	0
Overall	1993	74.2	-	57.1	-

\*The localization annotation of the top hit of the alignment list is used as the predicted localization.



Table 12. Comparison of the prediction accuracies of different approaches in the prediction of subcellular locations for the PS 2.0 data set

Localization	HYBRID		CELLO II		ALIGN*		PSORTb 2[8]	
	<i>Accuracy</i>	<i>MCC</i>	<i>Accuracy</i>	<i>MCC</i>	<i>Accuracy</i>	<i>MCC</i>	<i>Accuracy</i>	<i>MCC</i>
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	-

\*The localization annotation of the top hit of the alignment list is used as the predicted localization.

Table 13. Comparison of prediction accuracies of different approaches in the prediction of subcellular localizations for the PK dataset

Localization	HYBRID		CELLO II		ALIGN*		The PK method[17]	
	<i>Accuracy</i>	<i>MCC</i>	<i>Accuracy</i>	<i>MCC</i>	<i>Accuracy</i>	<i>MCC</i>	<i>Accuracy</i>	<i>MCC</i>
Chloroplast	90.0	0.88	79.9	0.81	89.0	0.83	72.3	-
Cytoplasm	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	-
Lysosome	87.1	0.89	68.8	0.78	83.9	0.81	61.8	-
Mitochondria	80.5	0.80	72.9	0.72	74.8	0.73	57.4	-
Nucleus	94.5	0.90	91.0	0.83	88.3	0.86	89.6	-
Peroxisome	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	-

\*The localization annotation of the top hit of the alignment list is used as the predicted localization.

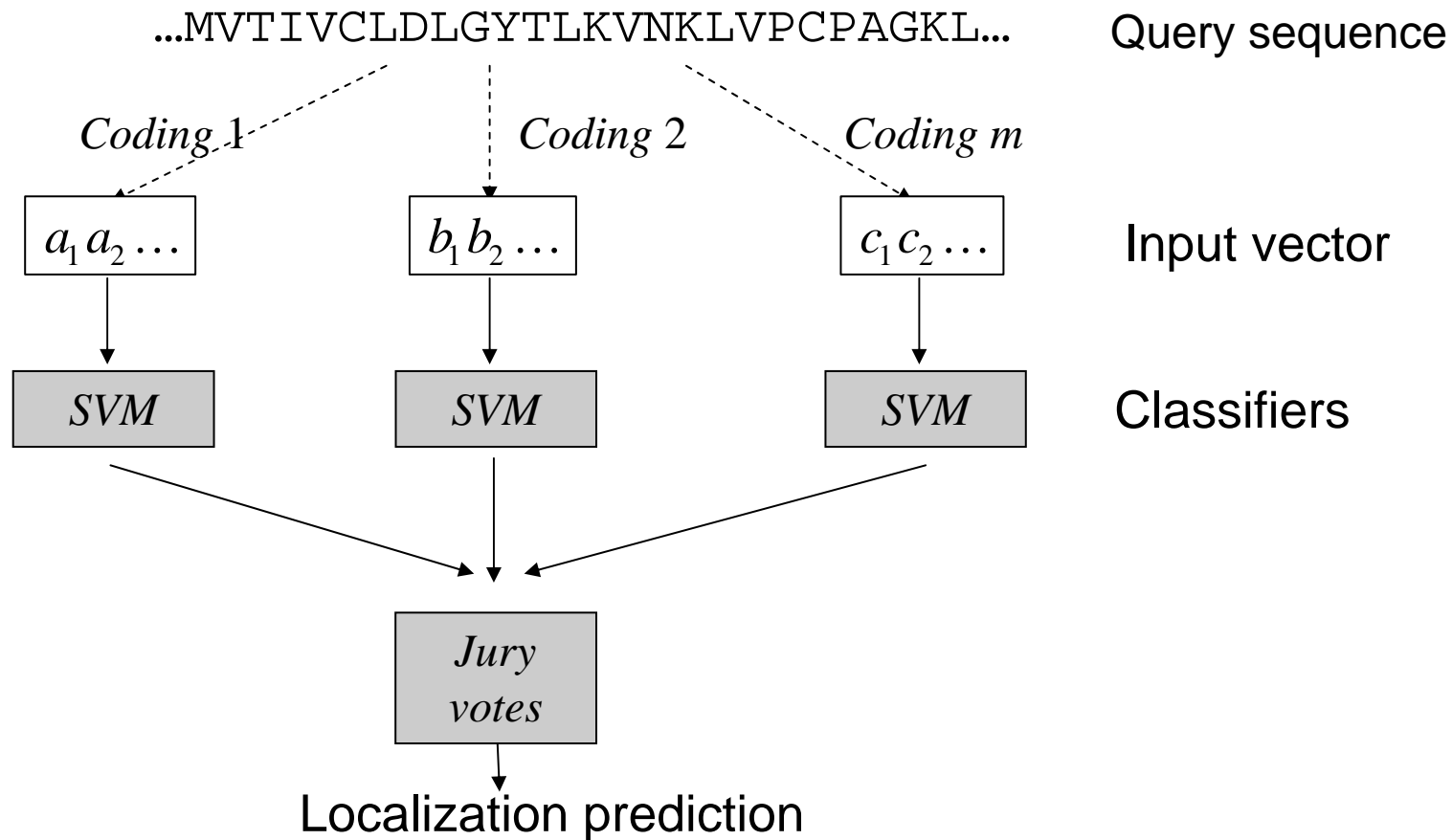


Figure 1. The query sequence is encoded by different coding schemes to obtain  $(a_1 a_2 \dots)$ ,  $(b_1 b_2 \dots)$ , and  $(c_1 c_2 \dots)$ , which are used to train the SVM classifiers. We combine votes from these classifiers and use the jury votes to determine the final assignment. We use four coding schemes in this work, which are  $A_1$ ,  $A_2$ ,  $X_4$ , and  $F_3 X_5$ . Because we use the one-against-one methods, we construct SVM classifiers for the prediction of  $J(J-1)/2$  subcellular localization sites.

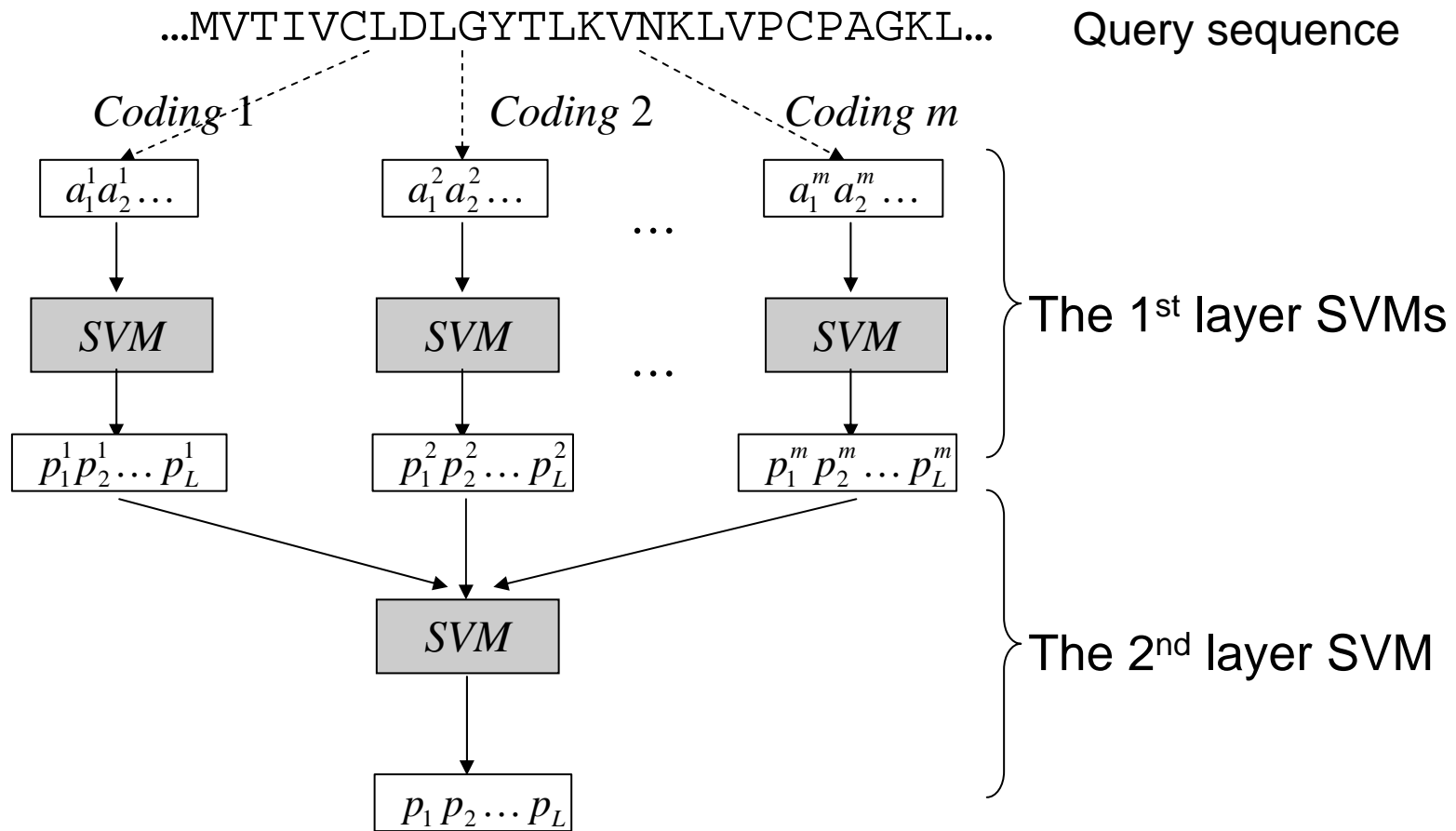


Figure 2. The first level classification system comprises SVMs based on different feature vectors:  $(a_1^1 a_2^1 \dots)$ ,  $(a_1^2 a_2^2 \dots)$ , ... and  $(a_1^m a_2^m \dots)$ . These SVMs generate probability distributions  $(a_1^1 a_2^1 \dots)$ ,  $(a_1^2 a_2^2 \dots)$ , ... and  $(a_1^m a_2^m \dots)$  of subcellular localizations. A second layer SVM (as a jury SVM) is used to process these probability distributions to generate the final probability distribution .

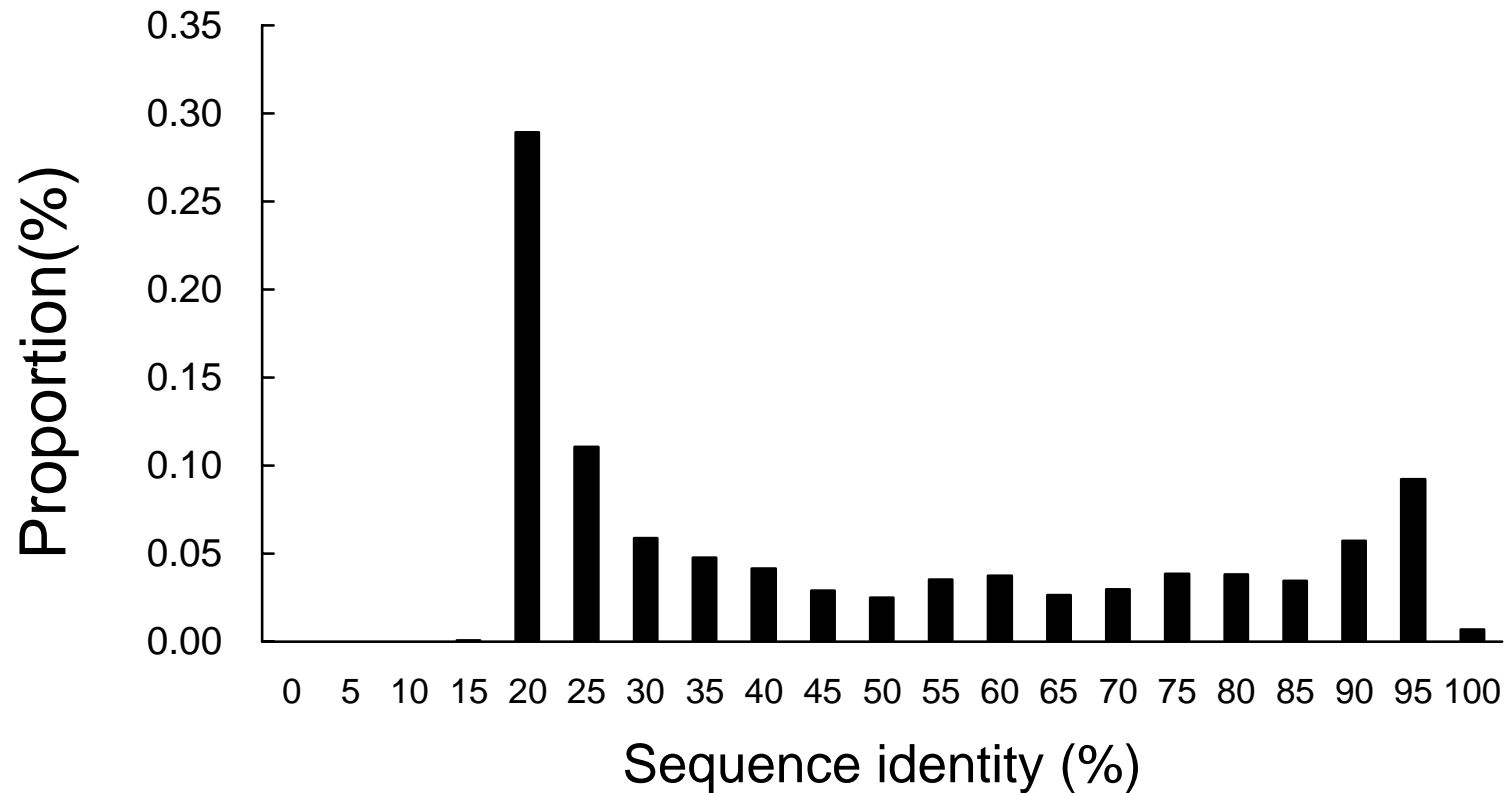


Figure 3. (A) The pair distribution of the sequence identities of the PS data set. Each bin (the width set to 5% sequence identity) represents the relative amount of the sequence pairs that share a given percentage sequence identity. For example, all sequences in each bin (say 20%) will share a pair sequence identity between 17.5% and 22.5% against each other. The value of the pair distribution is normalized by averaged over the total area under the distribution curve. Note that there are a few examples in the 15% and 100% sequence identity bins.

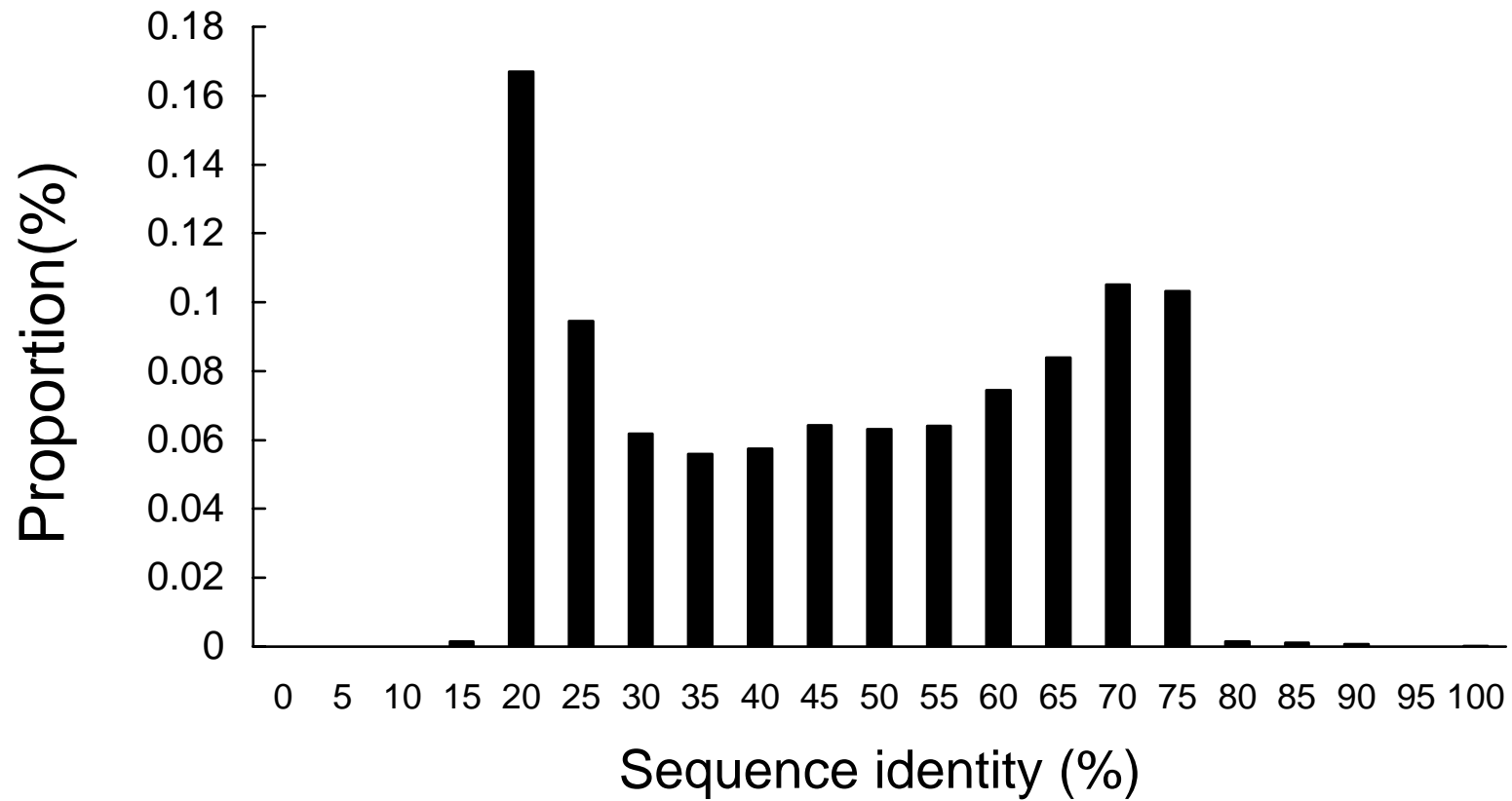


Figure 3. (B) The pair distribution of the sequence identities of the PK data set. There are a few examples in the 15% and 80-100% sequence identity bins.

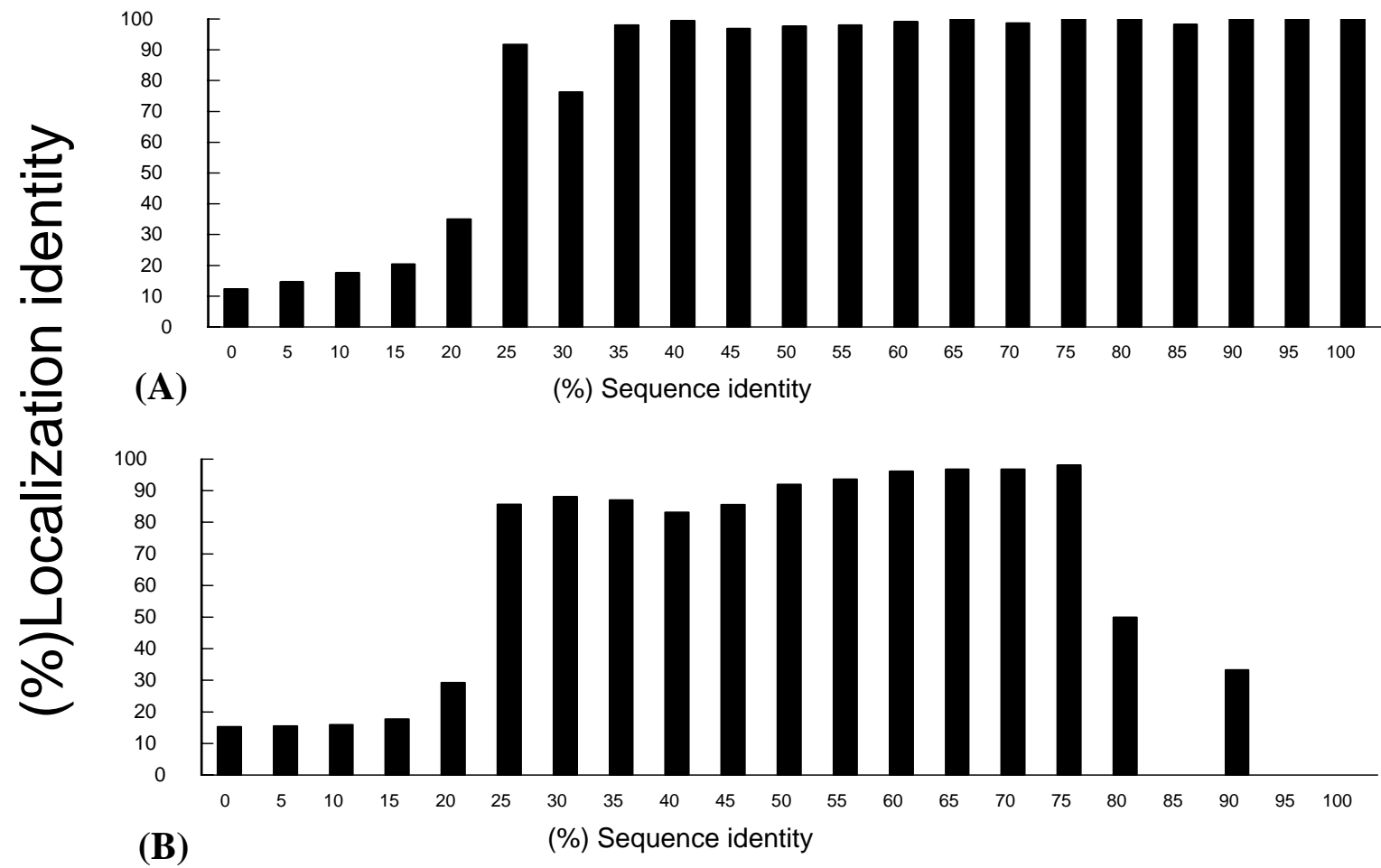


Figure 4. (A) The bar charts of localization identity vs. sequence identity for the PS data set.  
 (B) The bar charts of localization identity vs. sequence identity for the PK data set.

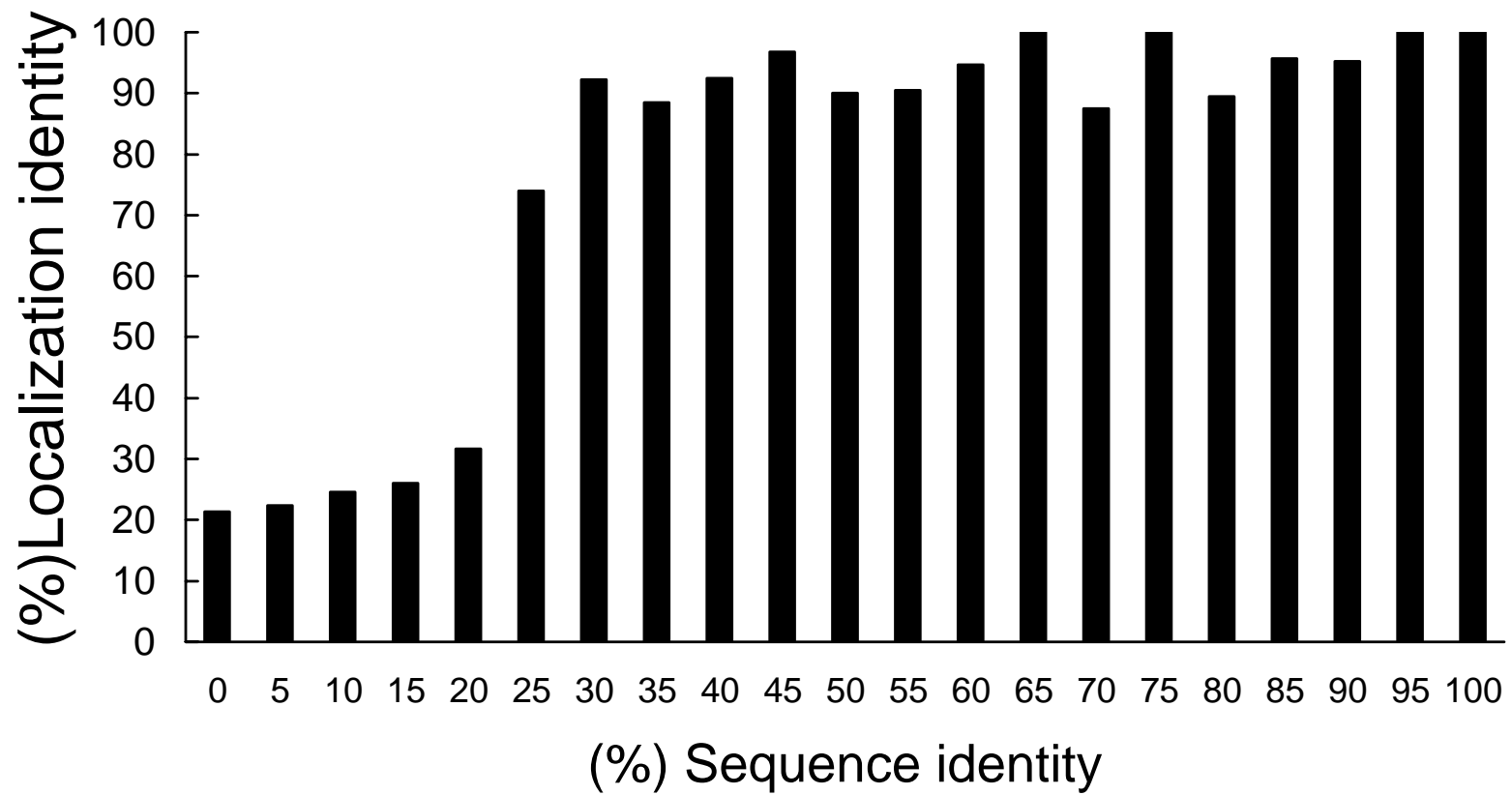


Figure 4. (C) The bar charts of localization identity vs. sequence identity for the SW41 data set.



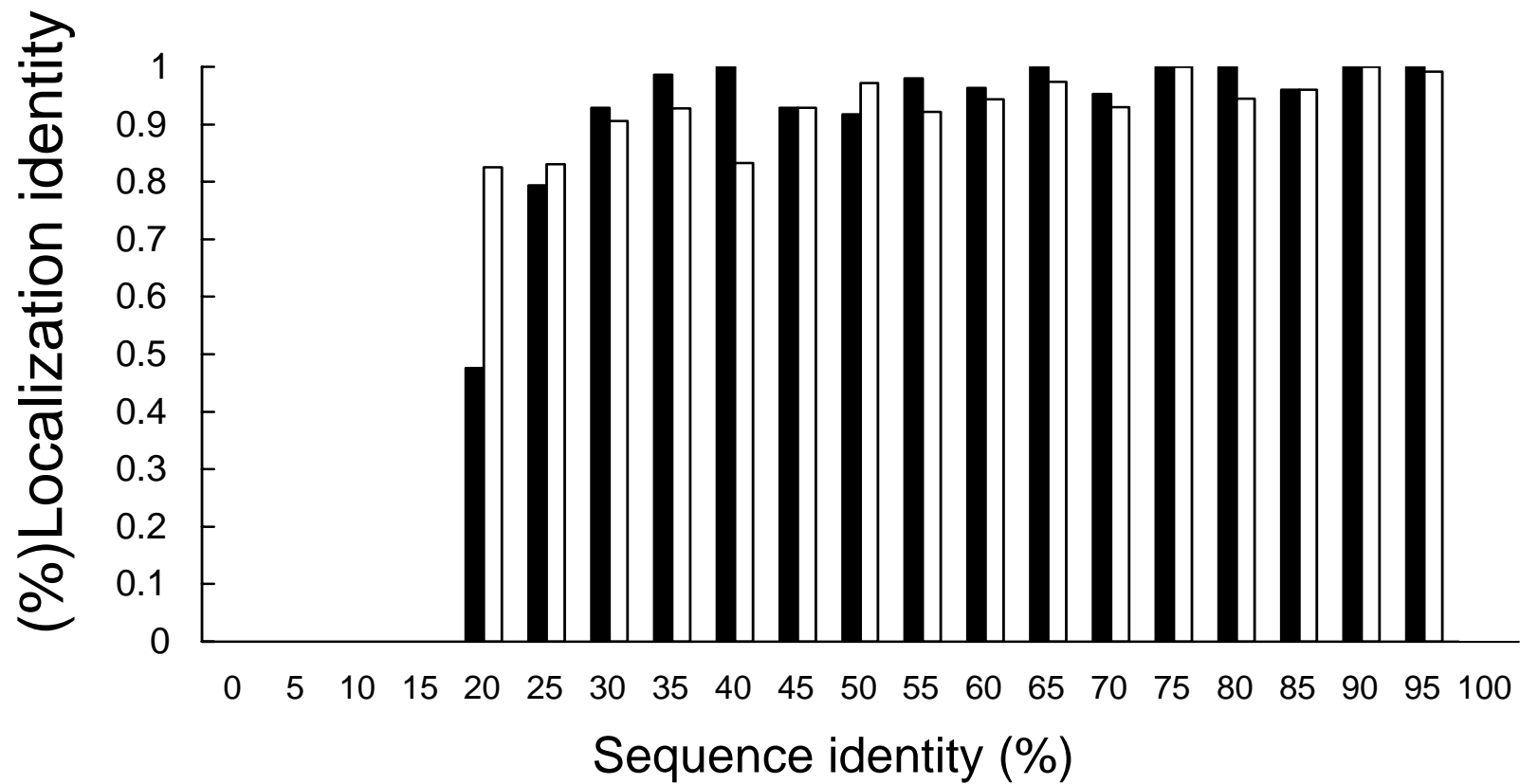


Figure 5. (A) The distributions of prediction accuracies as a function of sequence identity of both CELLO II (white bar) and ALIGN (black bar) for the PS data set. Note that we did not plot the prediction accuracies for those sequence identity bins that have relatively small example sizes as mentioned in the figure caption of Fig. 2.

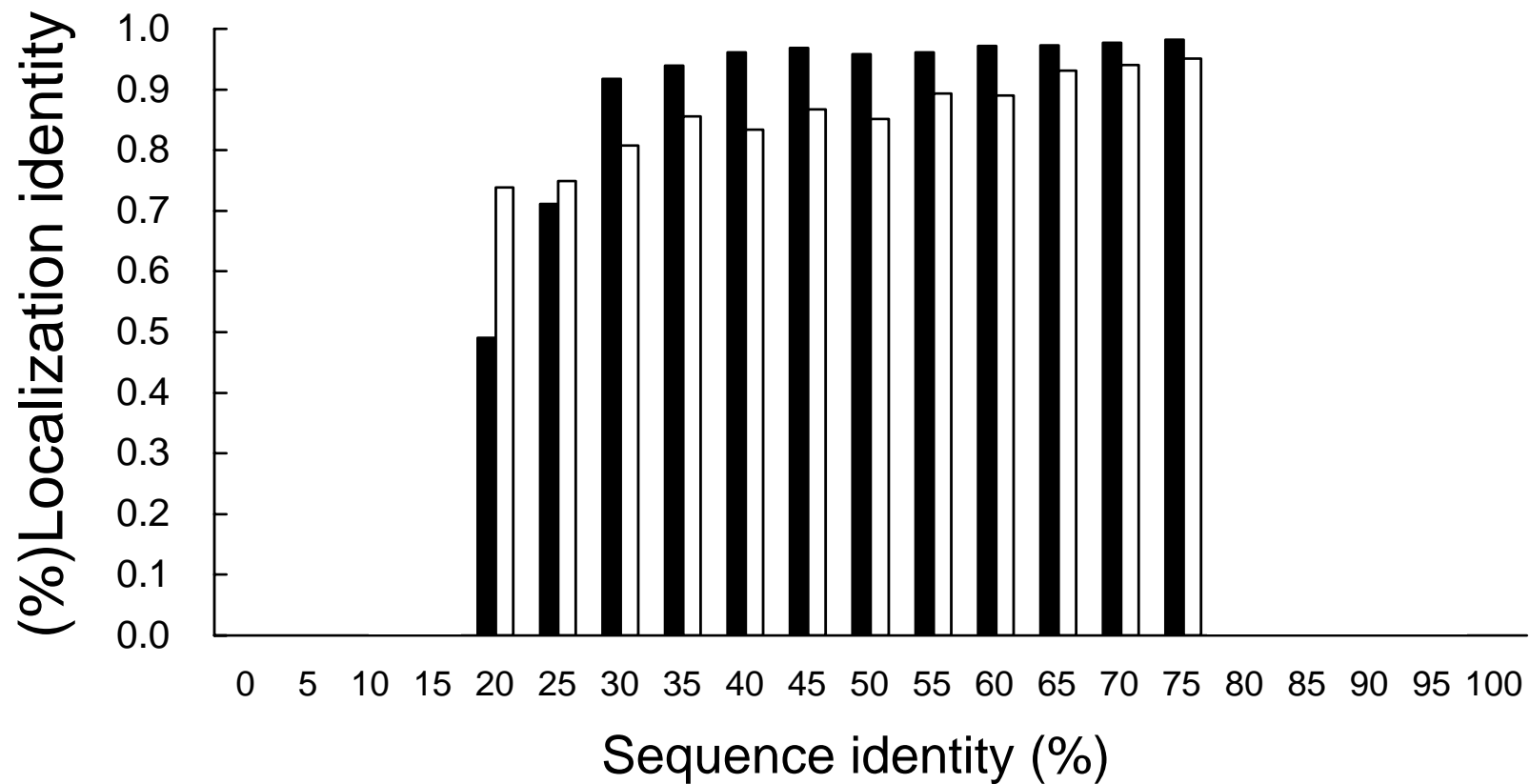


Figure 5. (B) The distributions of prediction accuracies as a function of sequence identity of both CELLO II (white bar) and ALIGN (black bar) for the PK data set. Note that we did not plot the prediction accuracies for those sequence identity bins that have relatively small example sizes as mentioned in the figure caption of Fig. 2.