JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 24, 1361-1375 (2008)

Predicting Subcellular Locations of Eukaryotic Proteins Using Bayesian and *k*-Nearest Neighbor Classifiers

HAN C. W. HSIAO, SHIH-HAO CHEN, JUDSON PEI-CHUN CHANG AND JEFFREY J. P. TSAI[+]

Department of Bioinformatics
Asia University
Wufeng, 413 Taiwan
[+]*Department of Computer Science*
University of Illinois at Chicago
Chicago, IL 60607, U.S.A.

Biologically, the function of a protein is highly related to its subcellular location. It is of necessity to develop a reliable method for protein subcellular location prediction, especially when a large amount of proteins are to be analyzed. Various methods have been proposed to perform the task. The results, however, are not satisfactory in terms of effectiveness and efficiency. A hybrid approach combining naïve Bayesian classifier and *k*-nearest neighbor classifier is proposed to classify eukaryotic proteins represented as a combination of amino acid composition, dipeptide composition, and functional domain composition. Experimental results show that the total accuracy of a set of 17,655 proteins can reach up to 91.5%.

Keywords: subcellular location prediction, naïve Bayesian classifier, *k*-nearest neighbor classifier, functional domain, feature reduction

1. INTRODUCTION

According to the central dogma of molecular biology, a eukaryotic DNA sequence is transcribed to an RNA sequence, and then translated to a protein, which can be addressed for specific target organelle or compartment, depending on its function [1]. Conventionally, three techniques are necessary to identify the subcellular location of a protein [2]: cell fractionation, electron and fluorescence microscopy images. Kumar *et al*. [3], for example, determined the subcellular location of 2,744 yeast proteins by immunolocalization of tagged gene products. However, wet lab experiments are usually time-consuming and not cost-effective. It is possible to alleviate these problems and reduce the time span of research by using *in silico* approaches to perform prediction tasks prior to conducting laboratory experiments. Nakai and Kanehisa [4, 5] were the first to predict protein location sites systematically, and the results were helpful to the biologists. Later, numerous works regarding subcellular location prediction for protein function analysis have been reported in the literature [6-15].

In general, these works can be reviewed from two perspectives: feature representation and classification approach. It is known that the subcellular location of a protein is determined by sorting signal [16] of around 15 to 60 amino acid residues at the *N*-terminal. It was found out that positively charged amino acid residues tend to congregate at

Received November 7, 2006; revised March 14, 2007; accepted June 27, 2007.
Communicated by Tsan-sheng Hsu.

the front part of the signal peptide, hydrophobic at the middle, and polar at the rear [17]. Another discovery is that Ser and Thr are frequent in chloroplast targeting peptides, while the negatively charged amino acid residues, like Asp and Glu, are less [18]. Consequently, recognizing sorting signal as a feature of protein is meaningful [19, 20]. A serious problem of using sorting signal is that protein sequences from draft genomes are usually incomplete and lack of signal peptides [21]. Amino acid composition is the simplest feature by calculating the frequencies of 20 kinds of amino acid residues [22-24]. It has a very high coverage but is not characteristic [24]. Dipeptide composition [25] is a variation that the frequencies of 202 combinations of two adjacent amino acid residues form a vector in a feature space of 400 dimensions. Different features may be extracted by a generalized autocorrelation function of the physicochemical properties, such as hydrophobicity and polarity, along an amino acid sequence [9, 26, 27]. Recently, functional domain composition [28, 29] has been utilized popularly due to its representative characteristics that proteins having particular domains often localize to specific organelles [30].

Various algorithms have been applied to the problem of protein subcellular location prediction. Reinhardt and Hubbard [22] used amino acid composition in combination with neural network approach to predict subcellular location of prokaryotic and eukaryotic proteins. The total accuracies achieved 81% for three subcellular locations in prokaryotic organisms and 66% for four locations in eukaryotic organisms. Yuan [31] adopted Markov chain models to achieve better accuracies of 89% and 73% for prokaryotic and eukaryotic sequences on the same data set. Lately, support vector machines (SVM) introduced by Vapnik [32] have been gaining popularity in the academic areas due to many attractive features for supervised classification and regression. Hua and Sun [24] proposed a method using amino acid composition and SVM to work on the same data set. The total prediction accuracies reached 91.4% for three subcellular locations in prokaryotic organisms and 79.4% for four locations in eukaryotic organisms. Different from the previous cases, Chou and Cai [28] used functional domain composition and SVM for subcellular location prediction. Later, Cai *et al*. [29] applied the same approach to predict membrane protein types by supervised classification performed in a hyperspace with dimensions up to 2005, and reached an overall accuracy up to 87.3%. Instead of using functional domain composition, Mott *et al*. [30] turned their attention to the co-occurrence of pairs of domains extracted from the SMART domain databank [33]. The co-occurrence matrix was then projected to a two-dimensional Euclidean space by multidimensional scaling, and thus the domains scattered as clusters of dots over this space. A number of Gaussian kernels as probability functions were applied to fit to the clusters. Finally, the subcellular location of a protein was determined by its domains locating on which Gaussian hills of the landscape and each hill represented a specific subcellular location. Their approach achieved an accuracy of 92% but was inefficient. Huang and Li [34] utilized dipeptide composition and fuzzy *k*-nearest neighbor method to achieve an overall accuracy about 80% in a jackknife test. Cai and Chou [35] used both functional domain composition and pseudo-amino acid composition as features to perform prediction separately by using the nearest neighborhood method. The result of using pseudo-amino acid composition as feature, however, was a supplement only when proteins without (known) functional domain.

For the sake of performing subcellular location prediction of a large volume of eukaryotic proteins, a considerably straightforward yet greatly effective approach combin-

ing naïve Bayesian classifier and *k*-nearest neighborhood method is proposed. Three commonly used features are considered as protein sequence descriptors, *i.e.* amino acid composition, dipeptide composition, and functional domain composition. The paper is organized as follows. The following section describes the proposed methods, including the evaluation method utilized in this study. The experimental results for three datasets as well as discussion are provided subsequently. The final section concludes the study.

2. METHODOLOGY

As described previously, different combinations of amino acid composition (AAC), dipeptide composition (DC), and functional domain composition (FDC) were considered as features for protein classification to predict the associated subcellular location. To retrieve the functional domains of each protein from the Conserved Domain Database (cdd v1.65), the Reverse Position Specific BLAST (RPS-BLAST 2.2.6) was utilized with E value set as 10^{-2}. The integrated feature vector, however, could be up to thousands of dimensions. Due to the curse of dimensionality, it is difficult to perform a classification task in such extremely high-dimensional feature space. The naïve Bayesian classifier was adopted in the first phase of classification task because of its demonstrated performance in many areas [36]. Suppose that a protein is expressed as a set of attributes $\langle a_1, a_2, \ldots, a_n \rangle$ with an annotated subcellular location $s \in S$, where S is a finite set of n_s subcellular locations. The Bayesian approach to classifying the subcellular location of a new protein, giving its attributes, is to maximize *a posterior* hypothesis of assigning the most probable s_{MAP} to the protein.

$$\begin{aligned} s_{MAP} &= \arg\max_{s_j \in S} P(s_j \mid a_1, a_2, \ldots, a_n) \\ &= \arg\max_{s_j \in S} \frac{P(a_1, a_2, \ldots, a_n \mid s_j) P(s_j)}{P(a_1, a_2, \ldots, a_n)} \\ &= \arg\max_{s_j \in S} P(a_1, a_2, \ldots, a_n \mid s_j) P(s_j) \end{aligned} \tag{1}$$

where $j = 1, \ldots, n_s$ and $P(a_1, a_2, \ldots, a_n)$ is a constant that can be disregarded. The two terms in the above equation can be estimated from the training dataset. Estimating $P(s_j)$ is straightforward, while the $P(a_1, a_2, \ldots, a_n \mid s_j)$ term is not that simple. An assumption to simplify the computation is that the attributes are conditionally independent, which can yield a result of replacing the first term by the product of the conditional probability for the individual attributes, *i.e.*

$$s_{NB} = \arg\max_{s_j \in S} P(s_j) \prod_{i=1}^{n} P(a_i \mid s_j) \tag{2}$$

where s_{NB} means the output of naïve Bayesian classifier. In the case of binary attributes a_i representing FDC, $P(a_i \mid s_j)$ is basically the portion of proteins belonging to subcellular location s_j. For numeric attributes a_i describing AAC and DC, $P(a_i \mid s_j)$ is calculated from a Gaussian density function of a_i

$$P(a_i \mid s_j) = g(a_i, \mu_{ij}, \sigma_{ij}) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} \exp\left[-\frac{(a_i - \mu_{ij})^2}{2\sigma_{ij}^2} \right] \tag{3}$$

where μ_{ij} and σ_{ij} are the mean and standard deviation of attribute a_i of training samples belonging to subcellular location s_j. Preferably, the values for comparison are normalized within a finite range. In reality, one protein could be represented by thousands of attributes, and thus the product of n probabilities would result in a numerical problem. In order to remedy this problem, Eq. (2) is modified as follows:

$$\begin{aligned} s_{NB} &= \arg\max_{s_j \in S}(p_j) \\ p_j &= \exp\left\{ \frac{1}{\alpha} \left[\log P(s_j) + \sum_{i=1}^{n} \log P(a_i \mid s_j) \right] \right\} \end{aligned} \tag{4}$$

where $0 < p_j < 1$ and α is a regulatory parameter. Moreover, one rational assumption is that some attributes a_i may be strongly associated with a specific subcellular location s_j. Instead of treating each attribute as equally important, a weighted version of Eq. (4) is presented as

$$\begin{aligned} s_{NB} &= \arg\max_{s_j \in S}(p_j) \\ p_j &= \exp\left\{ \frac{1}{\alpha} \left[\log P(s_j) + \sum_{i=1}^{n} w_i \log P(a_i \mid s_j) \right] \right\} \end{aligned} \tag{5}$$

where the weights w_i should be self-decided by the dataset. Numerous attribute selection techniques have been reviewed in the literature [37]. The information gain ratio is considered for its simplicity. First, the information (or entropy) of n_p proteins relative to classification of n_s subcellular locations is defined as

$$E = -\sum_{j=1}^{n_s} P(s_j) \log P(s_j). \tag{6}$$

After partitioning the set of proteins in accordance with n_o outcomes of an attribute a_i, the expected information can be acquired as the weighted sum of entropies over the subsets

$$E(a_i) = -\sum_{k=1}^{n_o} P(o_k) \sum_{j=1}^{n_s} P(s_j \mid o_k) \log P(s_j \mid o_k) \tag{7}$$

where o_k is one of n_o outcomes of attribute a_i and $P(o_k)$ is the portion of proteins having outcome o_k in attribute a_i. For binary attributes a_i, it is rather simple to get $P(s_j \mid o_k)$ by counting the number of proteins at subcellular location s_j under the condition of outcome o_k. Datasets with numeric attributes should be converted to discrete values prior to information calculation. A simple conversion of a_i to o_i is defined as

$$o_i = \begin{cases} 1 & \text{if } (a_i - \mu_{ij}) \le \sigma_{ij} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

where μ_{ij} and σ_{ij} are mentioned previously. The information gain with respect to attribute a_i is obtained as

$$Gain(a_i) = E - E(a_i). \tag{9}$$

The information gain is further normalized by the split information to obtain the gain ratio. The larger the gain ratio value is, the more important the attribute is.

$$\begin{aligned} GainRatio(a_i) &= \frac{Gain(a_i)}{SplitInfo(a_i)}, \\ SplitInfo(a_i) &= -\sum_{k=1}^{n_o} P(o_k) \log P(o_k). \end{aligned} \tag{10}$$

The result of naïve Bayesian classifier with information gain ratio is considered as the input for the second phase of classification. It is believed that the output generated in the first phase may contain rich information that can be reused as another feature to improve the accuracy. In other words, the output for a protein can be treated as a vector in a 12-dimensional hypercube ($[0, 1]^{12}$) in this study. Well classified proteins will definitely tend to gather at the 1 side; whereas others will distribute within the hypercube. Moreover, similar samples should distribute together even though they have represented in different way. Combining the features of amino acid composition and dipeptide composition, the dimensionality of input vector can still be up to hundreds of dimensions. Unfortunately, such representation may result in an unpredictably nonlinear distribution. Using *k*-nearest neighbor classifier is thus a compromising consideration of effectiveness,

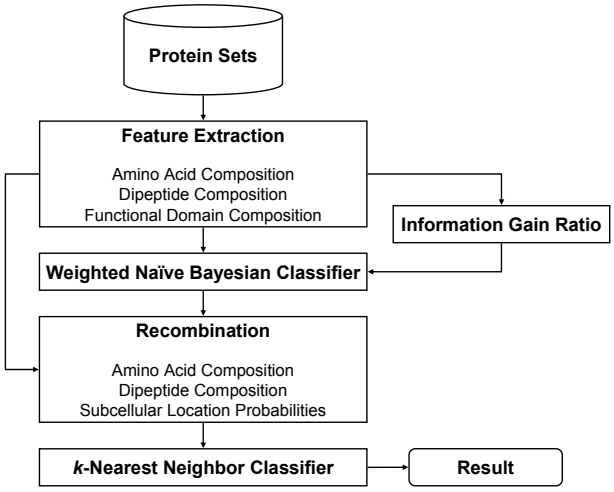


Fig. 1. Flow chart for predicting subcellular location.

efficiency, and simplicity. Note that this part of proteins within the hypercube cannot be correctly classified by weighted naïve Bayesian classifier. The nearest neighbors of an instance in such a space $\Re^n$ are defined in terms of the Euclidean distance defined in Eq. (11). The overall procedure for subcellular location prediction is illustrated in Fig. 1.

$$d(x_i, x_j) \equiv \sqrt{\sum_{k=1}^{n} [a_k(x_i) - a_k(x_j)]^2} \tag{11}$$

To compare with other works [6, 7, 25, 28], the experimental results were evaluated by the same testing procedures [24, 25]. Let N be the total number of protein sequences in a dataset and S be the category number of subcellular locations, *i.e.* $S = 12$ in this study. Moreover, TP_i, TN_i, FP_i, and FN_i are defined as the numbers of true positive, true negative, false positive, and false negative sequences of location i, respectively. Accordingly, the individual accuracy (IA_i), location accuracy (LA), total accuracy (TA), and Matthew's Correlation Coefficient (MCC_i) [38] are defined as follows:

$$\begin{aligned} IA_i &= \frac{TP_i}{TP_i + FP_i} \\ LA &= \frac{\sum_{i=1}^{S} IA_i}{S} \\ TA &= \frac{\sum_{i=1}^{S} TP_i}{N} \\ MCC_i &= \frac{TP_i \times TN_i + FP_i \times FN_i}{\sqrt{(TP_i + FN_i)(TP_i + FP_i)(TN_i + FN_i)(TN_i + FP_i)}} \end{aligned} \tag{12}$$

where $i = 1, \ldots, S$. It is clear from the definition of MCC that a completely correct classification yields a result of one.

3. EXPERIMENTAL RESULTS AND DISCUSSION

As listed in Table 1, three different datasets of eukaryotic proteins were extracted from the SWISS-PROT databank [39]. The first dataset D_1 from release 39.0 was prepared by Park and Kanehisa [25]; while the second dataset D_2 from release 35.0 was generated by Chou and Elrod [6] and later utilized by Chou and Cai [28]. To evaluate the performance of the proposed approach, a third dataset D_3 was generated from release 43.0 for testing. All eukaryotic proteins having single subcellular location were extracted out first, excluding those annotated as PROBABLE, POTENTIAL, POSSIBLE, or BY SIMILARITY. In addition, proteins having amino acids marked as X, Z, or B were removed as well. To include more sequences, some proteins having different keywords for location description were regarded as the same class [25]. Accordingly, all proteins in each dataset were then divided into twelve categories: chloroplast, cytoplasm, cytoskeleton

Table 1. The numbers of protein sequences in three datasets generated from the SWISS-PROT databank.

Subcellular location	Number of proteins		
	D_1	D_2	D_3
Chloroplast	671	138	1609
Cytoplasm	1241	535	2632
Cytoskeleton	40	29	81
Endoplasmic reticulum	114	42	432
Extracellular	861	206	3926
Golgi apparatus	47	20	186
Lysosome	93	36	151
Mitochondria	727	78	1414
Nucleus	1932	246	3331
Peroxisome	125	25	321
Plasma membrane	1674	627	3493
Vacuole	54	23	79
Total	7579	2005	17655

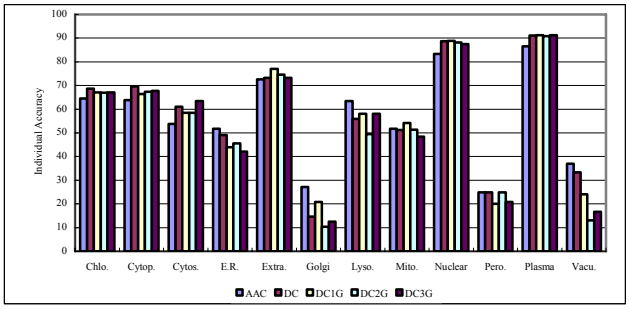
(filament, microtubule), endoplasmic reticulum, extracellular (secreted), Golgi apparatus, lysosome, mitochondria, nucleus, peroxisome (microsomes, glyoxysomal, glycosomal), plasma membrane (integral membrane), and vacuole.

Note that proteins in dataset D_2 were elaborately selected by the following five criteria [6, 28]; however, only the first three criteria were applied to datasets D_1 and D_3.

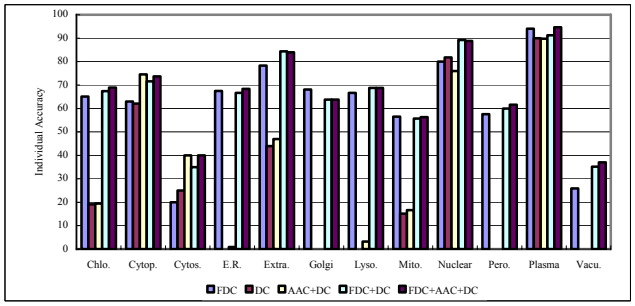
1. Sequences annotated as fragments were not selected;
2. Sequences contained ambiguous amino acids were not considered;
3. Sequences annotated multiple subcellular locations were filtered out;
4. Sequences without clear annotation were removed; and
5. Only one sequence was selected from a number of proteins with identical name but from different species.

After extracting functional domains from each protein by RPS-BLAST 2.2.6, there were 3824, 2968, and 8418 different conserved domains found from the protein datasets D_1, D_2, and D_3, respectively. The AAC and DC of each protein also formed a feature vector of 20 + 400 dimensions. For example, the vector dimensionality describing each protein in dataset D_1 was 3824 + 20 + 400. Each dataset was classified by weighted naïve Bayesian classifier with information gain ratio. The result for each protein was in fact a vector of probabilities with respect to twelve subcellular locations. This vector was then recombined with the 420-dimensional vector obtained previously to form a vector of 432 dimensions. Finally, all proteins represented by a 432-dimensional vector were classified by the k-nearest neighborhood method, where $k = 3$ was the optimal number to achieve the best result.

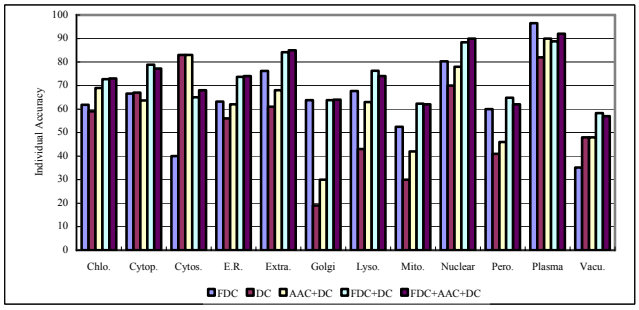
The first experiment utilizing the dataset D_1 was to evaluate the overall performance of the proposed approach in terms of feature extraction and protein classification. Fig. 2 demonstrates the individual accuracy of predicting each subcellular location with different combinations of features and classifiers. Fig. 2 (a) replicates the prediction results by



(a) Replicated from Park and Kanehisa [25] by SVM.



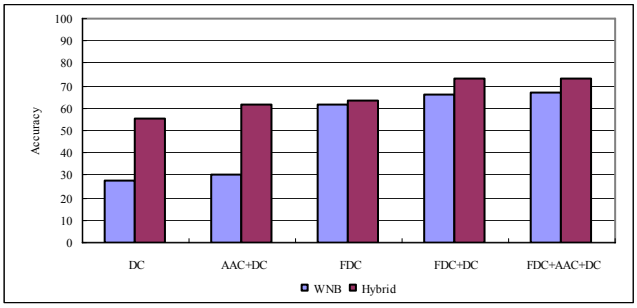
(b) By weighted naïve Bayesian classifier only.



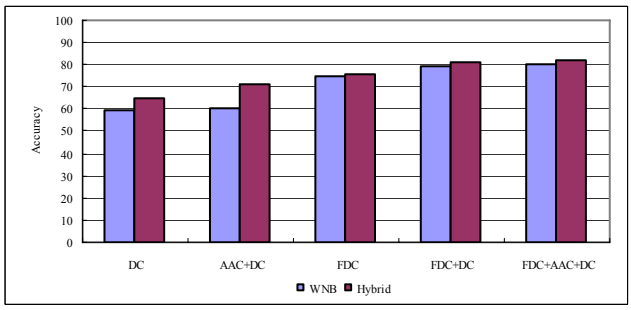
(c) By the proposed hybrid approach.

Fig. 2. Comparison of three approaches with different combinations of AAC, DC, FDC.

Park and Kanehisa [25] for performance comparison, where their task was done by employing SVM using RBF-kernel. The extracted features were AAC, DC with 0, 1, 2, and 3 gaps (*i.e.* DC1G, DC2G, DC3G). Figs. 2 (b) and (c) illustrate the results of applying weighted naïve Bayesian classifier without and with *k*-nearest neighbor method (*i.e.* hybrid classifier), respectively. Although the technique of SVM has demonstrated its outstanding performance to solve classification problems [40, 41], Fig. 2 reveals that more representative features can result in an improved accuracy. In comparison with Figs. 2 (a) and (b), the usage of weighted naïve Bayesian classifier with AAC + DC + FDC as the feature has achieved a better result, especially in the cases of endoplasmic reticulum, extracellular, Golgi apparatus, and peroxisome. The overall accuracy of the hybrid classifier using AAC + DC + FDC is even higher than that of weighted naïve Bayesian classifier, as shown in Fig. 2 (c).



(b) Location accuracy.



(b) Total accuracy.

Fig. 3. Accuracy comparison between weighted naïve Bayesian classifier and hybrid classifier with respect to different features.

The location accuracy and total accuracy by using weighted naïve Bayesian classifier without and with *k*-nearest neighbor method and different feature combinations are shown in Figs. 3 (a) and (b), respectively. The result indicates that FDC is a representative feature, whereas AAC and DC can contribute minor improvement. Moreover, the proposed hybrid classifier has substantial contribution to improve the accuracy, especially when the utilized feature is not so effective. Table 2 shows the confusion matrix [42] for location prediction of the first dataset using the hybrid classifier with AAC + DC + FDC as the input feature. It is interesting to note that the classes of plasma membrane, extracellular, and nucleus have a very high accuracy. On the other hand, proteins belonging to the categories of chloroplast, mitochondria, and nucleus tend to be incorrectly predicted as that of cytoplasm, and vice versa. Furthermore, the classes of vacuole, cytoskeleton, peroxisome, endoplasmic reticulum, lysosomal, and Golgi apparatus contain very few proteins and have an unreliable and inaccurate result. The phenomena may be possibly attributed to those proteins that are driven by protein interactions or involved in a secretory pathway.

Table 3 shows another comparison results for both datasets D_1 and D_2, where the prediction accuracies were evaluated by using a 5-fold cross validation test. Because of the changes of code naming, very few proteins in the original datasets could not be found. In the second dataset, Chou and Cai [28] performed the classification task by employing SVM with FDC as the feature for each protein and reached a total accuracy of 75%. The proposed hybrid approach combining three compositions could achieve better total accuracies of 82% and 79% for D_1 and D_2, respectively, especially for those location classes having fewer proteins. The same feature and approach were then applied to the dataset

Table 2. Confusion matrix for location prediction of dataset D_1 using hybrid classifier with three compositions as the feature. Twelve subcellular locations are chloroplast (Ch), cytoplasm (Cp), cytoskeleton (Cs), endoplasmic reticulum (ER), extracellular (Ex), Golgi apparatus (GA), lysosome (Ly), mitochondria (Mi), nucleus (Nu), peroxisome (Pe), plasma membrane (PM), and vacuole (Va).

		Predicted												
		Ch	Cp	Cs	ER	Ex	GA	Ly	Mi	Nu	Pe	PM	Va	Sum
Actual	Ch	**489**	33	0	1	3	0	0	42	9	4	2	1	584
	Cp	70	**958**	7	18	37	9	1	139	108	18	53	9	1427
	Cs	3	10	**27**	0	5	0	0	4	10	0	3	1	63
	ER	3	9	0	**84**	4	0	2	0	1	0	0	0	103
	Ex	9	26	1	3	**731**	3	12	6	23	3	19	5	841
	GA	1	1	0	0	1	**30**	0	0	0	0	0	0	33
	Ly	0	1	0	2	5	0	**69**	0	0	0	1	1	79
	Mi	47	67	1	1	4	0	0	**451**	14	9	12	0	606
	Nu	24	90	4	1	34	2	0	48	**1737**	5	40	4	1989
	Pe	2	12	0	0	5	0	0	11	0	**77**	5	0	112
	PM	22	25	0	3	28	2	6	25	26	9	**1534**	2	1682
	Va	1	9	0	1	4	1	3	1	4	0	5	**31**	30
	Sum	671	1241	40	114	861	47	93	727	1932	125	1674	54	**7579**

Table 3. Comparison of prediction accuracy for the datasets D_1 [25] and D_2 [28]. The difference in number of proteins N_p of each location class reflects the unavailability of data access due to the changes of code naming. The results marked in gray are cited from Table 7 in [25].

Location	Dataset D_1				Dataset D_2			
	Park & Kanehisa		Proposed		Chou & Cai		Proposed	
	N_p	%[†]	N_p	%[‡]	N_p	%[‡]	N_p	%[‡]
Chloroplast	671	72	671	73	145	57	138	49
Cytoplasmic	1245	72	1241	77	571	88	535	71
Cytoskeletal	41	59	40	68	34	44	29	68
ER[*]	114	47	114	74	49	31	42	74
Extracellular	862	78	861	85	224	57	206	81
Golgi apparatus	48	15	47	64	25	12	20	73
Lysosomal	93	62	93	74	37	54	36	67
Mitochondrial	727	57	727	62	84	42	78	83
Nuclear	1932	90	1932	90	272	73	246	93
Peroxisomal	125	25	125	62	27	4	25	53
Plasma membrane	1677	92	1674	92	699	91	627	90
Vacuole	54	25	54	57	24	25	23	57
Total	7589	78	7579	82	2191	75	2005	79

[*] Endoplasmic reticulum is abbreviated as ER
[+] Jack-knife
[‡] Five-fold cross validation

Table 4. The prediction accuracy of subcellular locations for dataset D_3 by the proposed hybrid classifier with three compositions as the feature.

Subcellular location	Number of proteins	Accuracy	MCC
Chloroplast	1609	92.3%	0.85
Cytoplasmic	2632	85.7%	0.84
Cytoskeletal	81	70.4%	0.71
Endoplasmic Reticulum	432	86.3%	0.86
Extracellular	3926	96.3%	0.96
Golgi Apparatus	186	69.9%	0.74
Lysosomal	151	92.1%	0.92
Mitochondrial	1414	84.5%	0.86
Nuclear	3331	90.2%	0.89
Peroxisomal	321	80.4%	0.84
Plasma Membrane	3493	97.7%	0.97
Vacuole	79	60.8%	0.70
Total accuracy		91.5%	

D_3, where valid sequences were arbitrarily selected as many as possible from the SWISS-PROT databank. The predicted results shown in Table 4 were evaluated by a 5-fold cross validation test as well, and the total accuracy of the set of 17,655 proteins could achieve up to 91.5%. In the worst classes of cytoskeleton, Golgi apparatus, and vacuole, insufficient number of proteins yields a low accuracy around 70% and MCC value above 0.7. This result might be improved by increasing the number of samples.

4. CONCLUSION

In this paper, a hybrid approach of two-stage classification for predicting subcellular location of eukaryotic proteins is presented. Two datasets reported in the literature were utilized for comparison, whereas a third dataset containing a large amount of proteins was generated for performance evaluation. Different combinations of amino acid composition, dipeptide composition, and functional domain composition have been tested. The results indicate that the combined usage of three compositions as protein feature for subcellular location prediction can reach a satisfactory accuracy. More precisely, the functional domain composition is of significant importance, while the other two compositions can contribute minor improvement. Without sophisticated mathematics, the proposed approach seamlessly combines weighted naïve Bayesian classifier and *k*-nearest neighbor classifier to perform classification task in a 432-dimensional feature space. Because of its simplicity, the proposed hybrid classifier is computationally efficient and easy to be implemented. Nor is it required to set up parameters for classification. The total accuracy for the third dataset of 17,655 proteins has achieved 91.5%. The future work will be directed towards refinements to the prediction of multiple subcellular locations at different status.

REFERENCES

1. F. Eisenhaber and P. Bork, "Wanted: subcellular localization of proteins based on

sequence," *Trends in Cell Biology*, Vol. 8, 1998, pp. 169-170.
2. R. F. Murphy, M. V. Boland, and M. Velliste, "Towards a systematics for protein subcelluar location: quantitative description of protein localization patterns and automated analysis of fluorescence microscope images," in *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology*, 2000, pp. 251-259.
3. A. Kumar, S. Agarwal, J. A. Heyman, S. Matson, M. Heidtman, S. Piccirillo, L. Umansky, A. Drawid, R. Jansen, Y. Liu, K. H. Cheung, P. Miller, M. Gerstein, G. S. Roeder, and M. Snyder, "Subcellular localization of the yeast proteome," *Genes and Development*, Vol. 16, 2002, pp. 707-719.
4. K. Nakai and M. Kanehisa, "Expert system for predicting protein localization sites in gram-negative bacteria," *Proteins: Structure*, *Function*, *and Genetics*, Vol. 11, 1991, pp. 95-110.
5. K. Nakai and M. Kanehisa, "A knowledge base for predicting protein localization sites in eukaryotic cells," *Genomics*, Vol. 14, 1992, pp. 897-911.
6. K. C. Chou and D. W. Elrod, "Prediction of membrane protein types and subcellular locations," *Proteins: Structure*, *Function*, *and Genetics*, Vol. 34, 1999, pp. 137-153.
7. K. C. Chou and D. W. Elrod, "Protein subcellular location prediction," *Protein Engineering*, Vol. 12, 1999, pp. 107-118.
8. K. C. Chou, "Review: prediction of protein structural classes and subcellular locations," *Current Protein and Peptide Science*, Vol. 1, 2000, pp. 171-208.
9. K. C. Chou, "Prediction of protein subcellular locations by incorporating quasi sequence order effect," *Biochemical and Biophysical Research Communications*, Vol. 278, 2000, pp. 477-483.
10. Z. P. Feng, "An overview on predicting the subcellular location of a protein," *Silico Biology*, Vol. 2, 2002, pp. 0027.
11. J. L. Gardy, C. Spencer, K. Wang, M. Ester, G. E. Tusnády, I. Simon, S. Hua, K. deFays, C. Lambert, K. Nakai, and F. S. L. Brinkman, "PSORT-B: improving protein subcellular localization prediction for Gram-negative bacteria," *Nucleic Acids Research*, Vol. 31, 2003, pp. 3613-3617.
12. M. Bhasin and G. P. S. Raghava, "ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST," *Nucleic Acids Research*, Vol. 32, 2004, pp. W414-W419.
13. K. C. Chou and Y. D. Cai, "Predicting protein localization in budding yeast," *Bioinformatics*, Vol. 21, 2005, pp. 944-950.
14. C. Guda and S. Subramaniam, "TARGET: a new method for predicting protein subcellular localization in eukaryotes," *Bioinformatics*, Vol. 21, 2005, pp. 3963-3969.
15. A. Höglund, P. Dönnes, T. Blum, H. W. Adolph, and O. Kohlbacher, "MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition," *Bioinformatics*, Vol. 22, 2006, pp. 1158-1165.
16. B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell*, Garlan Science, New York, 2002.
17. G. von Heijne, "The signal peptide," *Journal of Membrane Biology*, Vol. 115, 1990, pp. 195-201.
18. G. von Heijne, J. Steppuhn, and S. G. Hermann, "Domain structure of mitochondrial

and chloroplast targeting peptides," *European Journal of Biochemistry*, Vol. 180, 1989, pp. 535-545.
19. O. Emanuelsson, H. Nielsen, S. Brunak, and G. von Heijne, "Predicting subcellular localization of proteins based on their N-terminal amino acid sequence," *Journal of Molecular Biology*, Vol. 300, 2000, pp. 1005-1016.
20. K. Nakai, "Protein sorting signals and prediction of subcellular localization," *Advances in Protein Chemistry*, Vol. 54, 2000, pp. 277-344.
21. E. S. Lander, *et al*., "Initial sequencing and analysis of the human genome," *Nature*, Vol. 409, 2001, pp. 860-921.
22. A. Reinhardt and T. Hubbard, "Using neural networks for prediction of the subcellular location of proteins," *Nucleic Acids Research*, Vol. 26, 1998, pp. 2230-2236.
23. G. Schneider, "How many potentially secreted proteins are contained in a bacterial genome?" *Gene*, Vol. 237, 1999, pp. 113-121.
24. S. Hua and Z. Sun, "Support vector machine approach for protein subcellular localization prediction," *Bioinformatics*, Vol. 17, 2001, pp. 721-728.
25. K. J. Park and M. Kanehisa, "Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs," *Bioinformatics*, Vol. 19, 2003, pp. 1656-1663.
26. K. C. Chou and C. T. Zhang, "Prediction of protein structural classes," *Critical Reviews in Biochemistry and Molecular Biology*, Vol. 30, 1995, pp. 275-349.
27. K. C. Chou, "Prediction of protein cellular attributes using pseudo-amino acid composition," *Proteins*, Vol. 43, 2001, pp. 246-255.
28. K. C. Chou and Y. D. Cai, "Using functional domain composition and support vector machines for prediction of protein subcellular location," *Journal of Biological Chemistry*, Vol. 277, 2002, pp. 45765-45769.
29. Y. D. Cai, G. P. Zhou, and K. C. Chou, "Support vector machines for predicting membrane protein types by using functional domain composition," *Biophysical Journal*, Vol. 84, 2003, pp. 3257-3263.
30. R. Mott, J. Schultz, P. Bork, and C. P. Ponting, "Predicting protein cellular localization using a domain projection method," *Genome Research*, Vol. 12, 2002, pp. 1168-1174.
31. Z. Yuan, "Prediction of protein subcellular locations using Markov chain models," *FEBS Letter*, Vol. 451, 1999, pp. 23-26.
32. V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, 1995.
33. J. Schultz, R. R. Copley, T. Doerks, C. P. Ponting, and P. Bork, "SMART: a web-based tool for the study of genetically mobile domains," *Nucleic Acids Research*, Vol. 28, 2000, pp. 231-234.
34. Y. Huang and Y. Li, "Prediction of protein subcellular locations using fuzzy *k*-NN method," *Bioinformatics*, Vol. 20, 2004, pp. 21-28.
35. Y. D. Cai and K. C. Chou, "Nearest neighbour algorithm for predicting protein subcellular location by combining functional domain composition and pseudo-amino acid composition," *Biochemical and Biophysical Research Communications*, Vol. 305, 2003, pp. 407-411.
36. N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine Learning*, Vol. 29, 1997, pp. 131-163.
37. M. A. Hall and G. Holmes, "Benchmarking attribute selection techniques for discrete

class data mining," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 15, 2003, pp. 1437-1447.
38. B. W. Matthews, "Comparison of predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta*, Vol. 405, 1975, pp. 442-451.
39. A. Bairoch and R. Apweiler, "The SWISS-PROT protein sequence data bank and its supplement TrEMBL," *Nucleic Acids Research*, Vol. 25, 1997, pp. 31-36.
40. N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press, Cambridge, 2000.
41. S. T. Li, W. Shiue, and M. H. Huang, "The evaluation of consumer loans using support vector machines," *Expert Systems with Applications*, Vol. 30, 2006, pp. 772-782.
42. R. Kohavi and F. Provost, "Glossary of terms," *Machine Learning*, Vol. 30, 1998, pp. 271-274.

Han C. W. Hsiao (蕭震緯) received a Ph.D. degree in Civil Engineering with a Specialization in Geoinformatics from the University of Illinois at Urbana-Champaign (UIUC) in 1999. During his graduate study in the US, he worked respectively for the Robotics Center and the Business Process Division of Construction Engineering Research Laboratories (CERL) of Corps of Engineers, US Army. He also participated in the UIUC Digital Library Initiative (DLI) Project for more than two years. After graduation, he joined as a postdoctoral researcher in the information system research team of the Office of the National Science and Technology Program for Hazards Mitigation. He has been with Asia University (formerly Taichung Healthcare and Management University) since fall 2001, where he is currently an associate professor in the Department of Bioinformatics, and also holds a joint appointment with the Department of Computer Science and Information Engineering. He served as the registration chairs of IEEE BIBE 2004, IEEE SUTC 2006, IEEE ISM 2007, the panel co-chair of IEEE ICSC 2007, the publication and registration chair of IEEE SUTC 2008, the program committee members of IEEE APSCC 2008, IEEE ISM 2008. His research interests are in the areas of geoinformatics, bioinformatics, pattern recognition, data mining, and image processing. Recently, his research direction focuses on semantic computing for biomedical and multimedia data. Dr. Hsiao is a member of the Chinese Image Processing and Pattern Recognition Society.

Shih-Hao Chen (陳世豪) received a B.S. degree in Information Management from Chien Kuo Institute of Technology in 2002 and a M.S. degree in Bioinformatics from Asia University (formerly Taichung Healthcare and Management University) in 2004. He is currently a research assistant with the Institute of Harbor and Marine Technology, Institute of Transportation, MOTC. His research interests include data mining and bioinformatics.

Pei-Chun Chang (張培均) received the Ph.D. degree in Physical Chemistry from National Taiwan University in 1998. After that, he obtained the postdoctoral fellow from the Institute of Biomedical Sciences at Academic Sinica in Taiwan for four years. Dr. Chang is currently an associate professor in the Department of Bioinformatics at Asia University. His research interests include analysis of DNA and protein sequences, analysis of microarray data, cancer genomics, dynamics of nonlinear systems, and systems biology.

Jeffrey J. P. Tsai (蔡進發) received a Ph.D. degree in Computer Science from Northwestern University, Evanston, Illinois. He is a professor in the Department of Electrical Engineering and Computer Science at the University of Illinois at Chicago, where he is also the director of the Distributed Real-Time Intelligent Systems Laboratory. He coauthored Knowledge-Based Software Development for Real-Time Distributed Systems (World Scientific, 1993), Distributed Real-Time Systems: Monitoring, Visualization, Debugging, and Analysis (John Wiley & Sons, Inc., 1996), Compositional Verification of Concurrent and Real-Time Systems (Kluwer, 2001), coedited Monitoring and Debugging Distributed Real-Time Systems (IEEE/CS Press, 1995), and has published extensively in the areas of knowledge-based software engineering, software architecture, requirements engineering, formal methods, agent-based systems, and distributed real-time systems. Dr. Tsai was the recipient of a University Scholar Award from the University of Illinois in 1994 and was presented a Technical Achievement Award from the IEEE Computer Society in 1997. He is currently the co-editor-in-chief of the International Journal of Artificial Intelligence Tools. He is also an editor of the Annals of Software Engineering, the International Journal of Software Engineering and Knowledge Engineering, and the International Journal of Systems Integration. He is a fellow of the IEEE, the AAAS, and the SDPS.