PairProSVM: Protein Subcellular Localization Based on Local Pairwise Profile Alignment and SVM

Man-Wai Mak, Jian Guo, and Sun-Yuan Kung

Abstract—The subcellular locations of proteins are important functional annotations. An effective and reliable subcellular localization method is necessary for proteomics research. This paper introduces a new method—PairProSVM—to automatically predict the subcellular locations of proteins. The profiles of all protein sequences in the training set are constructed by PSI-BLAST, and the pairwise profile alignment scores are used to form feature vectors for training a support vector machine (SVM) classifier. It was found that PairProSVM outperforms the methods that are based on sequence alignment and amino acid compositions even if most of the homologous sequences have been removed. PairProSVM was evaluated on Huang and Li's and Gardy et al.'s protein data sets. The overall accuracies on these data sets reach 75.3 percent and 91.9 percent, respectively, which are higher than or comparable to those obtained by sequence alignment and composition-based methods.

Index Terms—Protein subcellular localization, sequence alignment, profile alignment, kernel methods, support vector machines.

✦

---

1 INTRODUCTION

THE subcellular locations of proteins have a significant influence on their functional characteristics, interaction partners, and potential roles in the cellular machinery. The determination of subcellular localization via experimental processes is often time consuming and laborious; therefore, a number of in silico subcellular localization methods have been proposed. These methods can be generally divided into the following categories:

1. *Sorting-signal-based methods.* This group of methods locates the proteins based on the existence of sorting signals [1]. PSORT, proposed by Nakai and Kanehisa in 1991 [2], [3], is the earliest predictor that uses sorting signals. Subsequent approaches use signal peptides, mitochondrial targeting peptides, and chloroplast transit peptides [4], [5], [6]. More recently, Horton et al. [7] have proposed a sorting-signal-composition-based method called WoLF PSORT for subcellular localization. WoLF is essentially a feature selector that selects features derived from PSORT, iPSORT, the amino acid content, and the sequence length. Horton et al. demonstrated that WoLF PSORT can be easily combined with BLAST [8] for subcellular localization.

2. *Composition-based methods.* This category studies the relationship between the subcellular locations and 1) amino acid compositions [9], [10], [11], [12], 2) amino acid pair compositions (dipeptide) [13], [14], [15], and 3) gapped amino acid pair compositions [14]. The concept of amino acid composition has been extended to pseudo amino acid compositions [16], from which information relevant to subcellular locations is extracted through a set of sequence-order-correlated factors and biochemical properties.

3. *Functional-domain-based methods.* This category looks at the correlation between the function of a protein and its subcellular location. Specifically, a protein is represented as a point in a high-dimensional space in which each basis is defined by one of the functional domains obtained from the functional domain database, the gene ontology database, or their combination [17].

4. *Homology-based methods.* This category is based on the notion that homologous sequences are also likely to have the same subcellular location. This property was first studied by Nair and Rost in 2002 [18], and subsequently, a number of methods based on this property have been proposed. For example, Proteome Analyst (PA) [19] uses the presence or absence of the tokens from certain fields of the homologous sequences in the SWISSPROT database as a means to compute features for classification. In the work of Kim et al. [20], an unknown protein sequence is aligned with every training sequence (with known subcellular locations) to create a feature vector for classification.

5. *Fusion-based methods.* This group integrates signal peptide information or whole sequence information with other features. For example, Gardy et al. [21]

- *M.-W. Mak is with the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hung Hom, Hong Kong. E-mail: enmwmak@polyu.edu.hk.*
- *J. Guo is with the Department of Statistics, University of Michigan at Ann Arbor, 436 West Hall, 1085 South University Ave., Ann Arbor, MI 48109. E-mail: guojian@umich.edu.*
- *S.-Y. Kung is with the Department of Electrical Engineering, Princeton University, Room B 230, Engineering Quadrangle, Princeton, NJ 08544. E-mail: kung@princeton.edu.*

developed PSORT-B that integrates amino acid compositions, similarity to proteins of known locations, signal peptides, transmembrane alpha-helices, and motifs corresponding to specific localizations. Bhasin and Raghava [22] and Garg et al. [23] predicted subcellular locations by fusing amino acid compositions, the composition of physicochemical properties, dipeptide compositions, residue couples, and PSI-BLAST search.

This paper applies pairwise profile alignment, which has been used successfully in remote homology detection [24], [25], [26], to predict protein subcellular locations. A profile is a matrix in which elements in a column specify the frequency of each amino acid that appears in that sequence position. In this work, the profile of a query sequence is generated by PSI-BLAST [27]; the resulting profile is then aligned with the profile of every training sequence to obtain a vector of alignment scores; finally, linear support vector machines (SVMs) are used to classify the vector. Our experimental results demonstrate the advantage of this vectorization scheme and the benefit of using profile alignment as compared to sequence alignment. This paper also serves to answer the following question: Can homology-based methods outperform nonhomology-based methods in subcellular localization, even on nonredundant data sets? The results in this paper suggest that the answer is yes.

The paper is organized as follows: Section 2 details the procedures of sequence alignment and profile alignment. It also outlines the vectorization process and the one-versus-rest SVM classifier based on the alignment scores. In Section 3, experimental evaluations based on two protein data sets are reported, and the performance between profile alignment, sequence alignment, and composition-based methods are compared. Finally, concluding remarks are drawn in Section 4.

2 KERNEL METHODS FOR SEQUENCE CLASSIFICATION

2.1 Sequence Alignment Kernels

Pairwise sequence alignment has been widely used for computing the similarity between two DNA or two protein sequences. It finds the best match between two sequences by inserting some gaps into proper positions of the two sequences. Denote

$$\mathcal{D} = \{S^{(1)}, \ldots, S^{(i)}, \ldots, S^{(j)}, \ldots, S^{(T)}\}$$

as a training set containing T sequences. Here, the ith protein sequence is denoted as

$$S^{(i)} = S_1^{(i)}, S_2^{(i)}, \ldots, S_{n_i}^{(i)}, \quad 1 \leq i \leq T,$$

where $S_k^{(i)} \in \mathcal{A}$, which is the set of 20 amino acid symbols, and n_i is the length of $S^{(i)}$. Using the BLOSUM62 substitution matrix [28], a set of similarity scores $\varepsilon'(S_u^{(i)}, S_v^{(j)})$ between position u of $S^{(i)}$ and position v of $S^{(j)}$ can be obtained.[1] Then, based on these scores and the Smith-Waterman alignment algorithm [29] with affine gap

extension [30], a sequence alignment score $\rho'(S^{(i)}, S^{(j)})$ can be obtained. Then, borrowing the idea from Shpaer et al. [31], we obtain the following normalized alignment score:

$$\zeta'(S^{(i)}, S^{(j)}) = \frac{\rho'(S^{(i)}, S^{(j)})}{\ln(n_i)\ln(n_j)}, \tag{1}$$

where n_i and n_j are the length of the ith and jth sequences. The normalization makes the alignment scores of unrelated sequences less dependent on the sequence length [31], thus allowing us to compare the alignment scores arising from sequences of different lengths. To facilitate SVM classification, we use a linear kernel of the form:

$$K'(S^{(i)}, S^{(j)}) = \sum_{t=1}^{T} \zeta'(S^{(i)}, S^{(t)})\zeta'(S^{(j)}, S^{(t)}). \tag{2}$$

Note that this kernel maps the variable-length sequence $S^{(i)}$ to a vector of alignment scores:

$$\zeta'^{(i)} = [\zeta'(S^{(i)}, S^{(1)}) \ \ldots \ \zeta'(S^{(i)}, S^{(T)})]^{\mathrm{T}}.$$

By aligning $S^{(i)}$ with each of the sequences in the training set. A kernel inner product between $S^{(i)}$ and $S^{(j)}$ can then be naturally obtained as $\langle \zeta'^{(i)}, \zeta'^{(j)} \rangle$. This leads to a class of algorithms referred to as SVM-pairwise, adopted in [20] and [32].

The sensitivity of detecting subtle homogenous segments can be improved by replacing pairwise sequence alignment with pairwise profile alignment. In Section 2.2, we will use the similarity scores of pairwise profile alignment to generate kernel matrices for SVM classification.

2.2 Profile Alignment Kernels

Following [33], here, we use a protein sequence (called query sequence) as a seed to search and align homogenous sequences from SWISSPROT 46.0 [34] using the PSI-BLAST program [27] with parameters h and j set to 0.001 and 3, respectively. The homolog information pertaining to the aligned sequences can be represented by two matrices (profiles): a position-specific scoring matrix (PSSM) and a position-specific frequency matrix (PSFM). Both PSSM and PSFM have 20 rows and L columns, where L is the number of amino acids in the query sequence. Each column of a PSSM represents the log likelihood of the residue substitutions at the corresponding positions in the query sequence. The (i, j)th entry of the matrix represents the chance of the amino acid in the jth position of the query sequence being mutated to amino acid type i during the evolution process. The PSFM contains the weighted observation frequencies of each position of the aligned sequences. Specifically, the (i, j)th entry of PSFM represents the possibility of having amino acid type i in position j of the query sequence.

Let us denote the operation of PSI-BLAST search given the query sequence $S^{(i)}$ of length n_i as

$$\phi^{(i)} \equiv \phi(S^{(i)}) : S^{(i)} \longrightarrow \{\mathbf{P}^{(i)}, \mathbf{Q}^{(i)}\}, \tag{3}$$

where $\mathbf{P}^{(i)}$ and $\mathbf{Q}^{(i)}$ are the PSSM and PSFM of $S^{(i)}$, respectively. Using the profile alignment algorithm specified in the Appendix, we obtain the profile alignment scores $\rho(\phi(S^{(i)}), \phi(S^{(j)}))$. Then, similar to sequence alignment, the following normalized alignment scores are obtained:

---

1. Hereafter, symbols with a prime mark (′) denote functions or variables associated with sequence alignment.

$$\zeta(\phi^{(i)}, \phi^{(j)}) = \frac{\rho(\phi(S^{(i)}), \phi(S^{(j)}))}{\ln(n_i)\ln(n_j)}. \quad (4)$$

A linear kernel based on the normalized scores (4) is then constructed for training SVM classifiers:

$$K(\phi(S^{(i)}), \phi(S^{(j)})) = \sum_{t=1}^{T} \zeta(\phi^{(i)}, \phi^{(t)})\zeta(\phi^{(j)}, \phi^{(t)}). \quad (5)$$

2.3 Multiclassification Using SVM

The multiclass problem can be solved by the one-versus-rest approach. Specifically, for a C-class problem (here, C is the number of subcellular locations), C independent SVM classifiers are constructed. During prediction, given an unknown protein sequence S, the output of the cth SVM is computed as[2]

$$f_c(S) = \sum_{i \in \mathcal{S}_c} y_{c,i}\alpha_{c,i}K(\phi(S^{(i)}), \phi(S)) + b_c, \quad (6)$$

where $\mathcal{S}_c$ is a set composed of the indices of the support vectors, $y_{c,i} \in \{-1, +1\}$ is the label of the ith training sequence, and $\alpha_{c,i}$ is the ith Lagrange multiplier of the cth SVM. The predicted class of S is given by

$$y(S) = \arg\max_c f_c(S), \quad c = 1, \dots, C.$$

In the following, we refer to the $y(S)$ with kernel $K(\phi(S^{(i)}), \phi(S))$ as the pairwise profile alignment SVM (or, simply, PairProSVM) and to the $y(S)$ with kernel $K'(S^{(i)}, S)$ as the pairwise sequence alignment SVM (PairSeqSVM).

The Spider Toolbox[3] was used to implement the SVM classifiers.

3 EXPERIMENTS AND RESULTS

3.1 Data Sets

Two data sets were used to evaluate the performance of the proposed method. The first one is introduced by Huang and Li [15]. This data set was created by selecting all eukaryotic proteins with annotated subcellular locations from SWIS-SPROT 41.0 and by setting the identity cutoff to 50 percent. The data set comprises 3,572 proteins (622 cytoplasm, 1,188 nuclear, 424 mitochondria, 915 extracellular, 26 golgi apparatus, 225 chloroplast, 45 endoplasmic reticulum, 7 cytoskeleton, 29 vacuole, 47 peroxisome, and 44 lysosome). The second data set was prepared by Gardy et al. [21] in 2003. It contains 1,443 Gram-negative bacterial sequences extracted from SWISS-PROT Release 40.29.

3.2 Performance Metric

Fivefold cross validation was used to evaluate the performance of PairProSVM and PairSeqSVM. The performance measures include the overall prediction accuracy (OA), the accuracy for each subcellular location (Acc), and Matthew's correlation coefficient (MCC) [35]. MCC [35] can overcome the shortcoming of accuracy on unbalanced data. For example, a classifier predicting all samples as positive cannot be regarded as a good classifier unless it can also

predict negative samples accurately. In this case, the accuracy and MCC of the positive class are 100 percent and 0 percent, respectively. Therefore, MCC is a better measure for unbalanced classification.

Denote $\mathbf{M} \in \Re^{C \times C}$ as the confusion matrix of the prediction result, where C is the number of classes. Then, $M_{i,j}$ $(1 \leq i, j \leq C)$ represents the number of proteins that actually belong to class i but are predicted as class j. We further denote

$$p_c = M_{c,c}, \quad q_c = \sum_{i=1,i\neq c}^{C}\sum_{j=1,j\neq c}^{C} M_{i,j},$$
$$r_c = \sum_{i=1,i\neq c}^{C} M_{i,c}, \quad s_c = \sum_{j=1,j\neq c}^{C} M_{c,j}, \quad (7)$$

where c $(1 \leq c \leq C)$ is the index of a particular class. For class c, p_c is the number of true positives, q_c is the number of true negatives, r_c is the number of false positives, and s_c is the number of false negatives. Based on the notations above, the OA, the accuracy of class c (Acc_c), the MCC of class c (MCC_c), the overall MCC (OMCC), and the weighted average MCC (WAMCC) are defined, respectively, as

$$\text{OA} = \frac{\sum_{c=1}^{C} M_{c,c}}{\sum_{i=1}^{C}\sum_{j=1}^{C} M_{i,j}}, \quad (8)$$

$$\text{Acc}_c = \frac{M_{c,c}}{\sum_{j=1}^{C} M_{c,j}}, \quad (9)$$

$$\text{MCC}_c = \frac{p_c q_c - r_c s_c}{\sqrt{(p_c + s_c)(p_c + r_c)(q_c + s_c)(q_c + r_c)}}, \quad (10)$$

$$\text{OMCC} = \frac{\hat{p}\,\hat{q} - \hat{r}\hat{s}}{\sqrt{(\hat{p} + \hat{s})(\hat{p} + \hat{r})(\hat{q} + \hat{s})(\hat{q} + \hat{r})}}, \quad (11)$$

$$\text{WAMCC} = \sum_{c=1}^{C} \frac{p_c + s_c}{N} \text{MCC}_c, \quad (12)$$

where $N = \sum_{c=1}^{C}(p_c + s_c)$, $\hat{p} = \sum_{c=1}^{C} p_c$, $\hat{q} = \sum_{c=1}^{C} q_c$, $\hat{r} = \sum_{c=1}^{C} r_c$, and $\hat{s} = \sum_{c=1}^{C} s_c$.

3.3 Results on Eukaryotic Proteins

The performance of Fuzzy K-NN [15], SubLoc [12], PairSeqSVM (K'), and PairProSVM (K) on Huang and Li's data set are shown in Table 1. The results of SubLoc were obtained by submitting the sequences of the first four classes in the data set to the SubLoc server (http://www.bioinfo.tsinghua.edu.cn/SubLoc/). Because SubLoc can only classify cytoplasm, nuclear, mitochondria, and extracellular, only the results corresponding to the first four classes are reported. The OA of PairProSVM and PairSeqSVM reaches 75.3 percent and 71.8 percent, which are significantly higher than that of Fuzzy K-NN and SubLoc. In addition, the OA of PairProSVM is 3.5 percent higher than that of PairSeqSVM. PairProSVM and PairSeqSVM outperform Fuzzy K-NN in 8 out of 11 subcellular locations. However, they perform poorly on golgi apparatus, cytoskeleton, and vacuole. This is mainly due to the lack of data

---

2. For sequence alignment, the kernel function becomes $K'(S^{(i)}, S)$ in (2).
3. http://www.kyb.tuebingen.mpg.de/bs/people/spider/.

TABLE 1
Comparison of Fuzzy K-NN, SubLoc, PairSeqSVM, and PairProSVM on Huang and Li's Data Set
at 50 Percent and 15 Percent Sequence Identities

Subcellular Location	50% Sequence Identity								15% Sequence Identity		
	Number of Sequences	Fuzzy K-NN		SubLoc	PairSeqSVM (K')		PairProSVM (K)		Number of Sequences	PairProSVM (K)	
		Acc	MCC	Acc	Acc	MCC	Acc	MCC		Acc	MCC
Cytoplasm	622	35.4%	0.31	60.8%	49.5%	0.45	51.1%	0.49	255	34.1%	0.28
Nuclear	1188	71.5%	0.58	78.7%	93.1%	0.71	90.3%	0.75	478	88.5%	0.65
Mitochondria	424	36.6%	0.30	64.4%	53.3%	0.56	66.5%	0.64	194	49.0%	0.47
Extracellular	915	81.6%	0.54	53.8%	87.4%	0.78	89.7%	0.84	365	87.1%	0.77
Golgi Apparatus	26	15.4%	0.27	–	3.8%	0.11	15.4%	0.29	15	13.3%	0.20
Chloroplast	225	32.4%	0.36	–	35.6%	0.50	59.6%	0.62	86	32.6%	0.40
End. Reticulum	45	11.1%	0.22	–	26.7%	0.43	44.4%	0.47	22	9.1%	0.15
Cytoskeleton	7	28.6%	0.44	–	0.0%	0.00	0.0%	0.00	2	0.0%	0.00
Vacuole	29	6.9 %	0.16	–	0.0%	0.00	0.0%	0.01	12	16.7%	0.25
Peroxisome	47	14.9%	0.27	–	42.6%	0.58	46.8%	0.51	19	31.6%	0.34
Lysosome	44	20.5%	0.31	–	22.7%	0.32	36.4%	0.37	22	36.4%	0.38
Overall	3572	58.1%	–	66.0%	71.8%	0.70	75.3%	0.73	1470	66.1%	0.63
Weighted Average	–	–	0.46	–	–	0.63	–	0.68	–	–	0.56

Acc: Accuracy; MCC: Matthew's correlation coefficient. For SubLoc, the OA is based on the first four classes.

TABLE 2
Accuracies of SubLoc, PSORT-B, PA, PSLpred, CELLO, PairSeqSVM (K'), and PairProSVM (K) in Gardy et al.'s Data Set

Subcellular Location	No. of Seq.	PSORT-B	SubLoc	CELLO	PA	PSLpred	PairSeqSVM (K')	PairProSVM (K)
Cytoplasm	248	69.4%	97.6%	90.7%	85.3%	90.7%	97.2%	93.2%
Inner Membrane	268	78.7%	–	88.4%	90.6%	90.3%	93.7%	92.9%
Periplasmic	244	75.6%	83.2%	86.9%	86.0%	90.6%	79.1%	87.7%
Outer Membrane	352	90.3%	–	94.6%	94.7%	95.2%	95.7%	96.0%
Extracellular	190	70.0%	67.5%	78.9%	88.0%	86.8%	83.7%	86.8%
Overall Accuracy	–	74.8%	84.1%	88.9%	89.5%	91.2%	90.7%	91.9%

in these three classes in the data set (26, 7, and 29 sequences only). PairProSVM also outperforms SubLoc in three out of four classes, leading to a significantly higher OA.

3.4 Results on Prokaryotic Proteins

To investigate whether profile alignment is also appropriate for predicting the subcellular localization of prokaryotic proteins, we performed experiments on a prokaryotic data set created by Gardy et al. [21]. The results of SubLoc [12], PSORT-B 1.0 [21], PA [19], PSLpred [36], CELLO [37], PairSeqSVM, and PairProSVM on this data set are shown in Table 2. Again, the results of SubLoc were obtained by presenting the sequences of the data set to SubLoc's Web server. The results suggest that the performance of PairProSVM is comparable to that of PA and PSLpred and is significantly better than that of PSORT-B and SubLoc. Also, note that the performance difference between Pair-SeqSVM and PairProSVM is only 1 percent, confirming that remote homology seems to be more helpful for classifying eukaryotic proteins, where over 3 percent difference was achieved.

3.5 Results on Redundancy-Removed Data Sets

Huang and Li's data set covers 11 location sites, allowing a sequence identity of up to 50 percent. To mitigate homology bias, we constructed a series of redundancy-removed data sets by eliminating the most similar sequences. Specifically, any pairs of sequences in a redundancy-removed data set should not have an identity higher than λ, where λ is a filtering threshold. The NCBI Blastclust program was used to implement the filtering process (blastclust -L 0 -S λ). Different values of λ, from 15 percent to 45 percent with intervals of 5 percent for Huang and Li's data set, were tested. Note that the number of proteins in each subcellular location becomes smaller when the threshold decreases. The numbers of proteins in each subcellular location at 50 percent and 15 percent sequence identities are shown in Table 1.

Fig. 1 shows the OA of PairProSVM (K) and Pair-SeqSVM (K') in the redundancy-removed data sets with different values of λ ($\lambda = 50$ percent means that the original data set was used). The result shows that the accuracy of both methods decreases when the sequences become less similar to each other. However, the rate of performance degradation is more dramatic in PairSeqSVM than in
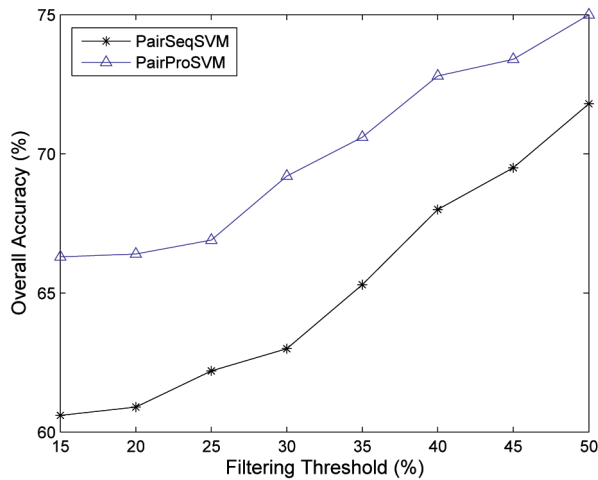
Fig. 1. The relationship between the filtering threshold λ and the OA of PairSeqSVM and PairProSVM on Huang and Li's data sets.

PairProSVM. In particular, when the filtering threshold drops from 50 percent to 15 percent, the accuracy of PairProSVM drops from 75.3 percent to 66.1 percent (a 12 percent reduction), whereas the accuracy of PairSeqSVM drops from 71.8 percent to 60.6 percent (a 16 percent reduction). This suggests that PairProSVM is less sensitive to the similarity among the training sequences than PairSeqSVM.

3.6 Computational Complexity

The computation in PairProSVM can be divided into three phases: 1) profile generation, 2) profile alignment, and 3) SVM training and classification. Among the three phases, the alignment process is the most computationally intensive, which amounts to 94 percent of the total computation time. We note that PairProSVM is more computationally intensive than PairSeqSVM. In particular, the profile alignment time is about three times that of the sequence alignment time. PairProSVM also requires extra computation time to produce the profiles using PSI-BLAST, which amounts to about 4 percent of the total computation time.

In the training phase, each entry of the kernel matrix (both K' and K) requires $O(n_i n_j)$ computation time, where n_i and n_j are the lengths of the ith and jth sequences in the training set, respectively. Therefore, the computational complexity of the whole kernel matrix is $O(\sum_i \sum_j n_i n_j) = O(L^2)$, where L is the sum of the length of all training sequences. In the prediction phase, an unknown sequence (profile) of n residues needs to be aligned with all training sequences (profiles) to produce a feature vector for SVM classification, which requires $O(\sum_j n n_j) = O(nL)$ operations. Assuming that there are V support vectors in the SVM, the overall complexity in the prediction phase is $O(nL + VT)$, where T is the number of training sequences.

While alignment-based methods are more computationally intensive than composition-based methods, we strongly believe that accuracy is by far much more important than speed in subcellular localization, because the latter can be easily solved by the rapid improvement in CPU performance. Moreover, recent advances in feature selection methods for pairwise scoring matrices [38] can also help alleviate this computation limitation.

4   CONCLUSIONS

This paper applies SVMs with profile alignment kernels to predict proteins' subcellular locations. Profiles are calculated by searching the SWISSPROT database using PSI-BLAST. Then, the scores of sequence and profile alignment are computed, which in turn are used to construct the kernels of an SVM classifier. Evaluations on eukaryotic and prokaryotic data sets show that profile-based methods are superior to sequence-alignment-based methods and composition-based methods. This paper also addresses a concern about homology-based methods raised in [39]: proteins with high sequence homology do not necessarily share the same localization. Results of this paper, however, show that if rich localization information can be extracted from homologous sequences (such as profiles), homology-based methods can outperform nonhomology-based methods significantly.

The kernel matrices used by PairProSVM and other supplementary materials can be found in http://www.eie. polyu.edu.hk/~mwmak/BSIG/PairProSVM.htm.

APPENDIX

Let us denote the operation of PSI-BLAST search given the query sequence $S^{(i)}$ of length n_i as

$$\phi^{(i)} \equiv \phi(S^{(i)}) : S^{(i)} \longrightarrow \{\mathbf{P}^{(i)}, \mathbf{Q}^{(i)}\}, \qquad (13)$$

where

$$\mathbf{P}^{(i)} = [\mathbf{p}_1^{(i)}, \mathbf{p}_2^{(i)}, \dots, \mathbf{p}_{n_i}^{(i)}],$$
$$\mathbf{Q}^{(i)} = [\mathbf{q}_1^{(i)}, \mathbf{q}_2^{(i)}, \dots, \mathbf{q}_{n_i}^{(i)}]$$

are the PSSM and PSFM of $S^{(i)}$, respectively. The elements in $\mathbf{p}_u^{(i)}$ and $\mathbf{q}_v^{(i)}$ can be expressed as

$$\mathbf{p}_u^{(i)} = [p_{u,1}^{(i)}, p_{u,2}^{(i)}, \dots, p_{u,20}^{(i)}]^{\mathrm{T}}, \quad 1 \le u \le n_i,$$
$$\mathbf{q}_v^{(i)} = [q_{v,1}^{(i)}, q_{v,2}^{(i)}, \dots, q_{v,20}^{(i)}]^{\mathrm{T}}, \quad 1 \le v \le n_i.$$

Let us assume that we need to align the profiles of two sequences $S^{(i)}$ and $S^{(j)}$. Define partial matrices $\hat{\mathbf{P}}_u^{(i)} = [\mathbf{p}_1^{(i)} \dots \mathbf{p}_u^{(i)}]$ and $\hat{\mathbf{Q}}_v^{(i)} = [\mathbf{q}_1^{(j)} \dots \mathbf{q}_v^{(j)}]$ corresponding to $S^{(i)}$ and $S^{(j)}$, respectively. Define an $(n_i + 1) \times (n_j + 1)$ matrix $\mathbf{M}$ whose (u, v)th element $M(u, v)$ for $u = 1, \dots, n_i$ and $v = 1, \dots, n_j$ represents the score of an optimal profile alignment between $\hat{\mathbf{P}}_u^{(i)}$ and $\hat{\mathbf{Q}}_v^{(j)}$ and between $\hat{\mathbf{P}}_u^{(j)}$ and $\hat{\mathbf{Q}}_v^{(i)}$, given that the alignment ends with $\mathbf{p}_u^{(i)}$ aligned to $\mathbf{q}_v^{(j)}$ and $\mathbf{p}_v^{(j)}$ aligned to $\mathbf{q}_u^{(i)}$. The scoring function introduced in [25] can be adopted to compute the similarity score between $\mathbf{p}_u^{(i)}$, $\mathbf{q}_v^{(j)}$, $\mathbf{p}_v^{(j)}$, and $\mathbf{q}_u^{(i)}$ as follows:

$$\varepsilon_{u,v}^{(i,j)} = \sum_{h=1}^{20} \left(p_{u,h}^{(i)} q_{v,h}^{(j)} + p_{v,h}^{(j)} q_{u,h}^{(i)} \right). \qquad (14)$$

Define an $n_i \times n_j$ matrix $\mathbf{I}$ whose (u, v)th element represents the score of an optimal alignment, given that the alignment ends with $\mathbf{p}_u^{(i)}$ or $\mathbf{q}_u^{(i)}$ aligned to a gap. Similarly, define an $n_i \times n_j$ matrix $\mathbf{J}$ whose (u, v)th element represents the score of an optimal alignment given that the alignment ends with $\mathbf{p}_v^{(j)}$ or $\mathbf{q}_v^{(j)}$ aligned to a gap.

With the above definitions, the profile alignment algorithm is specified as follows [29]:

1. Initialize the accumulative score matrix M:

$$M(0,0) = 0,$$
$$M(u,0) = -g_{\text{open}} - (u-1)g_{\text{ext}},$$
$$M(0,v) = -g_{\text{open}} - (v-1)g_{\text{ext}},$$

where $u = 1, \ldots, n_i$, $v = 1, \ldots, n_j$, and g_{open} and g_{ext} are two user-defined parameters representing the gap opening penalty and gap extension penalty, respectively.

2. Calculate $M(u,v)$ recursively as follows:

$$M(u,v) = \max \begin{cases} 0, \\ M(u-1,v-1) + \varepsilon_{u,v}^{(i,j)}, \\ I(u-1,v-1) + \varepsilon_{u,v}^{(i,j)}, \\ J(u-1,v-1) + \varepsilon_{u,v}^{(i,j)}, \end{cases}$$

where

$$I(u,v) = \max \begin{cases} 0, \\ M(u-1,v) - g_{\text{open}}, \\ I(u-1,v) - g_{\text{ext}}, \end{cases}$$

$$J(u,v) = \max \begin{cases} 0, \\ M(u,v-1) - g_{\text{open}}, \\ J(u,v-1) - g_{\text{ext}}. \end{cases}$$

3. Obtain the profile alignment score of $S^{(i)}$ and $S^{(j)}$ as follows:

$$\rho(\phi(S^{(i)}), \phi(S^{(j)})) = \max\{M(\hat{u}_i, \hat{v}_j), I(\hat{u}_i, \hat{v}_j), J(\hat{u}_i, \hat{v}_j)\},$$

where $(\hat{u}_i, \hat{v}_j)$ is the position in $\mathbf{M}$ corresponding to the maximum alignment score, i.e.,

$$(\hat{u}_i, \hat{v}_j) = \arg \max_{1 \leq u \leq n_i;\, 1 \leq v \leq n_j} \{M(u,v), I(u,v), J(u,v)\}.$$

In this work, the open gap (g_{open}) and extension gap penalties (g_{ext}) were set to 11 and 1, respectively.

ACKNOWLEDGMENTS

REFERENCES

[1] K. Nakai, "Protein Sorting Signals and Prediction of Subcellular Localization," *Advances in Protein Chemistry,* vol. 54, no. 1, pp. 277-344, 2000.

[2] K. Nakai and M. Kanehisa, "Expert System for Predicting Protein Localization Sites in Gram-Negative Bacteria," *Proteins: Structure, Function, and Genetics,* vol. 11, no. 2, pp. 95-110, 1991.

[3] K. Nakai and M. Kanehisa, "A Knowledge Base for Predicting Protein Localization Sites in Eukaryotic Cells," *Genomics,* vol. 14, pp. 897-911, 1992.

[4] O. Emanuelsson, H. Nielsen, S. Brunak, and G. von Heijne, "Predicting Subcellular Localization of Proteins Based on Their N-Terminal Amino Acid Sequence," *J. Molecular Biology,* vol. 300, pp. 1005-1016, 1997.

[5] H. Nielsen, J. Engelbrecht, S. Brunak, and G. von Heijne, "A Neural Network Method for Identification of Prokaryotic and Eukaryotic Signal Peptides and Prediction of Their Cleavage Sites," *Int'l J. Neural Systems,* vol. 8, pp. 581-599, 1997.

[6] H. Nielsen, S. Brunak, and G. von Heijne, "Machine Learning Approaches for the Prediction of Signal Peptides and Other Protein Sorting Signals," *Protein Eng.,* vol. 12, no. 1, pp. 3-9, 1999.

[7] P. Horton, K.J. Park, T. Obayashi, and K. Nakai, "Protein Subcellular Localization Prediction with WoLF PSORT," *Proc. Fourth Ann. Asia Pacific Bioinformatics Conf. (APBC '06),* pp. 39-48, 2006.

[8] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman, "Basic Local Alignment Search Tool," *J. Molecular Biology,* vol. 215, pp. 403-410, 1990.

[9] H. Nakashima and K. Nishikawa, "Discrimination of Intracellular and Extracellular Proteins Using Amino Acid Composition and Residue-Pair Frequencies," *J. Molecular Biology,* vol. 238, pp. 54-61, 1994.

[10] J. Cedano, P. Aloy, J.A. Perez-Pons, and E. Querol, "Relation between Amino Acid Composition and Cellular Location of Proteins," *J. Molecular Biology,* vol. 266, pp. 594-600, 1997.

[11] A. Reinhardt and T. Hubbard, "Using Neural Networks for Prediction of the Subcellular Location of Proteins," *Nucleic Acids Research,* vol. 26, pp. 2230-2236, 1998.

[12] S.J. Hua and Z.R. Sun, "Support Vector Machine Approach for Protein Subcellular Localization Prediction," *Bioinformatics,* vol. 17, pp. 721-728, 2001.

[13] Z. Yuan, "Prediction of Protein Subcellular Locations Using Markov Chain Models," *FEBS Letters,* vol. 451, no. 1, pp. 23-26, 1999.

[14] 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, no. 13, pp. 1656-1663, 2003.

[15] Y. Huang and Y.D. Li, "Prediction of Protein Subcellular Locations Using Fuzzy K-NN Method," *Bioinformatics,* vol. 20, no. 1, pp. 21-28, 2004.

[16] K.C. Chou, "Prediction of Protein Cellular Attributes Using Pseudo-Amino Acid Composition," *Proteins: Structure, Function, and Genetics,* vol. 43, pp. 246-255, 2001.

[17] Y.D. Cai and K.C. Chou, "Predicting Subcellular Localization of Proteins in a Hybridization Space," *Bioinformatics,* vol. 20, pp. 1151-1156, 2004.

[18] R. Nair and B. Rost, "Sequence Conserved for Subcellular Localization," *Protein Science,* vol. 11, pp. 2836-2847, 2002.

[19] Z. Lu, D. Szafron, R. Greiner, P. Lu, D.S. Wishart, B. Poulin, J. Anvik, C. Macdonell, and R. Eisner, "Predicting Subcellular Localization of Proteins Using Machine-Learned Classifiers," *Bioinformatics,* vol. 20, no. 4, pp. 547-556, 2004.

[20] J.K. Kim, G.P.S. Raghava, S.Y. Bang, and S. Choi, "Prediction of Subcellular Localization of Proteins Using Pairwise Sequence Alignment and Support Vector Machine," *Pattern Recognition Letters,* vol. 27, no. 9, pp. 996-1001, 2006.

[21] J.L. Gardy, C. Spencer, K. Wang, M. Ester, G.E. Tusnady, I. Simon, S.J. 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, no. 13, pp. 3613-3617, 2003.

[22] 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, Webserver Issue, pp. 414-419, 2004.

[23] A. Garg, M. Bhasin, and G.P.S. Raghava, "SVM-Based Method for Subcellular Localization of Human Proteins Using Amino Acid Compositions, Their Order and Similarity Search," *J. Biological Chemistry,* vol. 280, pp. 14427-14432, 2005.

[24] S. Busuttil, J. Abela, and G.J. Pace, "Support Vector Machines with Profile-Based Kernels for Remote Protein Homology Detection," *Genome Informatics,* vol. 15, no. 2, pp. 191-200, 2004.

[25] H. Rangwala and G. Karypis, "Profile-Based Direct Kernels for Remote Homology Detection and Fold Recognition," *Bioinformatics,* vol. 21, no. 23, pp. 4239-4247, 2005.

[26] R. Kuang, E. Ie, K. Wang, K. Wang, M. Siddiqi, Y. Freund, and C. Leslie, "Profile-Based String Kernels for Remote Homology Detection and Motif Extraction," *J. Bioinformatics and Computational Biology,* vol. 3, pp. 527-550, 2005.

[27] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman, "Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs," *Nucleic Acids Research,* vol. 25, pp. 3389-3402, 1997.

[28] S. Henikoff and J.G. Henikoff, "Amino Acid Substitution Matrices from Protein Blocks," *Proc. Nat'l Academy of Sciences,* pp. 10915-10919, 1992.

[29] T.F. Smith and M.S. Waterman, "Comparison of Biosequences," *Advances in Applied Math.,* vol. 2, pp. 482-489, 1981.

[30] O. Gotoh, "An Improved Algorithm for Matching Biological Sequences," *J. Molecular Biology,* vol. 162, pp. 705-708, 1982.

[31] E.G. Shpaer, M. Robinson, D. Yee, J.D. Candlin, R. Mines, and T. Hunkapiller, "Sensitivity and Selectivity in Protein Similarity Searches: A Comparison of Smith-Waterman in Hardware to BLAST and FASTA," *Genomics,* vol. 38, pp. 179-191, 1996.

[32] L. Liao and W.S. Noble, "Combining Pairwise Sequence Similarity and Support Vector Machines for Detecting Remote Protein Evolutionary and Structural Relationships," *J. Computational Biology,* vol. 10, no. 6, pp. 857-868, 2003.

[33] L. Rychlewski, B. Zhang, and A. Godzik, "Fold and Function Predictions for Mycoplasma Genitalium Proteins," *Folding and Design,* vol. 3, no. 4, pp. 229-238, 1998.

[34] B. Boeckmann, A. Bairoch, R. Apweiler, M.C. Blatter, A. Estreicher, E. Gasteiger, M.J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider, "The SWISS-PROT Protein Knowledgebase and Its Supplement TrEMBL in 2003," *Nucleic Acids Research,* vol. 31, pp. 365-370, 2003.

[35] B.W. Matthews, "Comparison of Predicted and Observed Secondary Structure of T4 Phage Lysozyme," *Biochimica et Biophysica Acta,* vol. 405, pp. 442-451, 1975.

[36] M. Bhasin, A. Garg, and G.P.S. Raghava, "PSLpred: Prediction of Subcellular Localization of Bacterial Proteins," *Bioinformatics,* vol. 21, no. 10, pp. 2522-2524, 2005.

[37] C.S. Yu, 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 Science,* vol. 13, pp. 1402-1406, 2004.

[38] S.Y. Kung and M.W. Mak, "Feature Selection for Pairwise Scoring Kernels with Applications to Protein Subcellular Localization," *Proc. IEEE Int'l Conf. Acoustic, Speech, and Signal Processing (ICASSP '07),* pp. 569-572, 2007.

[39] P. Donnes and A. Hoglund, "Predicting Protein Subcellular Localization: Past, Present, and Future," *Genomics, Proteomics, and Bioinformatics,* vol. 2, no. 4, pp. 209-215, 2004.

Man-Wai Mak received the BEng (Hons) degree in electronic engineering from Newcastle Upon Tyne Polytechnic and the PhD degree in electronic engineering from the University of Northumbria at Newcastle. He joined the Department of Electronic Engineering, Hong Kong Polytechnic University, as a lecturer in 1993 and as an assistant professor in 1995. His research interests include speaker recognition, machine learning, and bioinformatics. He has authored more than 100 technical papers in speaker recognition, machine learning, and bioinformatics. He is also a coauthor of the postgraduate textbook *Biometric Authentication: A Machine Learning Approach* (Prentice Hall, 2005). He received the Faculty of Engineering Research Grant Achievement Award in 2003. Since 1995, he has been an executive committee member of the IEEE Hong Kong Section Computer Chapter. He was the chairman of the IEEE Hong Kong Section Computer Chapter from 2003 to 2005 and is currently a member of the IEEE Machine Learning for Signal Processing Technical Committee. He is a member of the IEEE.

Jian Guo received the bachelor's degree in applied mathematics. He is a research assistant in the Department of Electronic and Information Engineering, Hong Kong Polytechnic University. His research focuses on methodology of machine learning and its application to bioinformatics.

Sun-Yuan Kung is a professor in Department of Electrical Engineering, Princeton University. His research areas include VLSI array processors, system modeling and identification, neural networks, wireless communication, sensor array processing, multimedia signal processing, bioinformatic data mining, and biometric authentication. He has authored and coauthored more than 400 technical publications and numerous textbooks, including *VLSI and Modern Signal Processing* (Prentice Hall, 1985), *VLSI Array Processors* (Prentice Hall, 1988), *Digital Neural Networks* (Prentice Hall, 1993), *Principal Component Neural Networks* (John Wiley & Sons, 1996), and *Biometric Authentication: A Machine Learning Approach* (Prentice Hall, 2004). Since 1990, he has been the editor in chief of the *Journal of VLSI Signal Processing Systems*. He was appointed as the first associate editor in VLSI area (1984) and, later, the first associate editor in neural network (1991) for the *IEEE Transactions on Signal Processing*. He was a recipient of the IEEE Signal Processing Society's Technical Achievement Award for the contributions on "parallel processing and neural network algorithms for signal processing" (1992), a Distinguished Lecturer of IEEE Signal Processing Society (1994), a recipient of the IEEE Signal Processing Society's Best Paper Award for his publication on principal component neural networks (1996), and a recipient of the IEEE Third Millennium Medal (2000). He was a founding member of several technical committees of the IEEE Signal Processing Society. He has been a fellow of the IEEE since 1988. He served as a member of the board of governors of the IEEE Signal Processing Society from 1989 to 1991.