

# Eukaryotic Protein Subcellular Localization Based on Local Pairwise Profile Alignment SVM

Jian QUO and Man Wai MAK

Dept. of Electronic and Information  
Engineering, The Hong Kong  
Polytechnic University

Sun Yuan KUNG

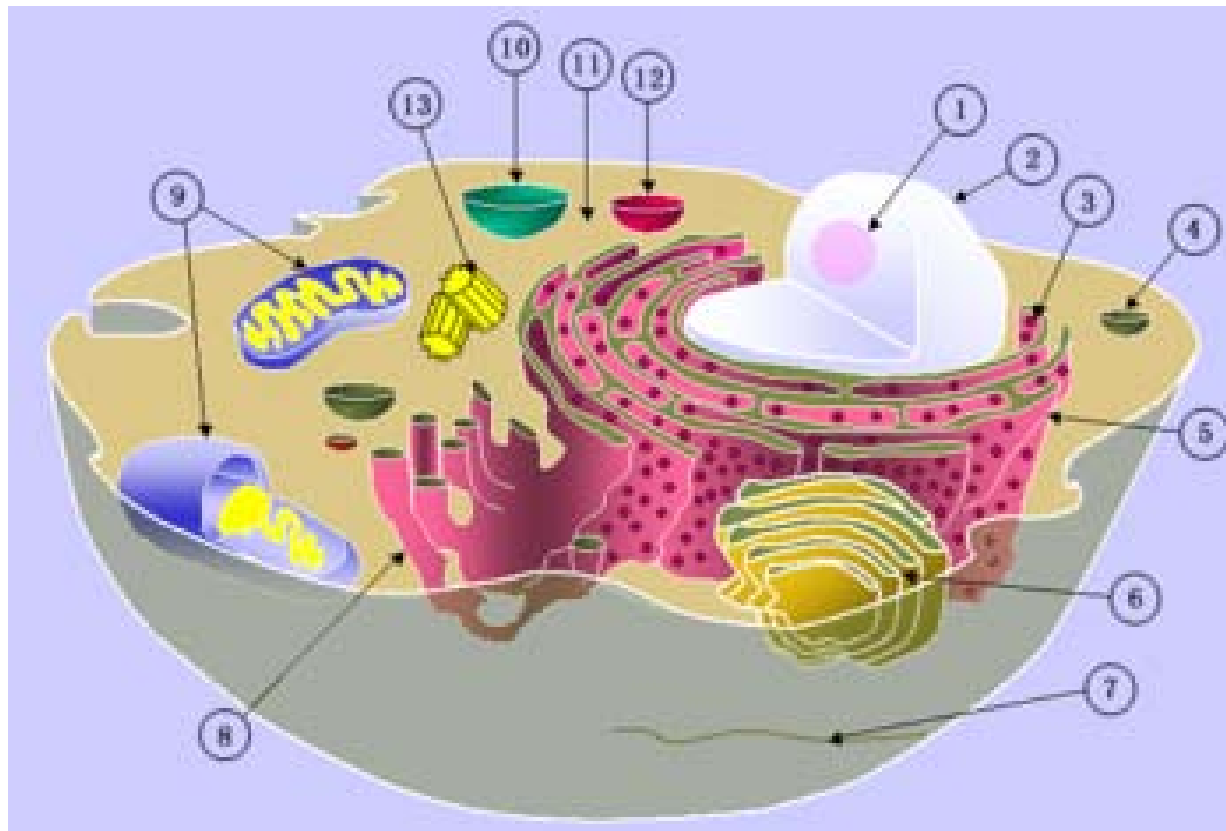
Dept. of Electrical Engineering,  
Princeton University

# Outline

- Why Subcellular Localization?
- Feature Extraction
  - By aligning protein sequences
  - By aligning the profiles of protein sequences
- 1-vs-rest SVM Classifiers
- Results and Conclusions

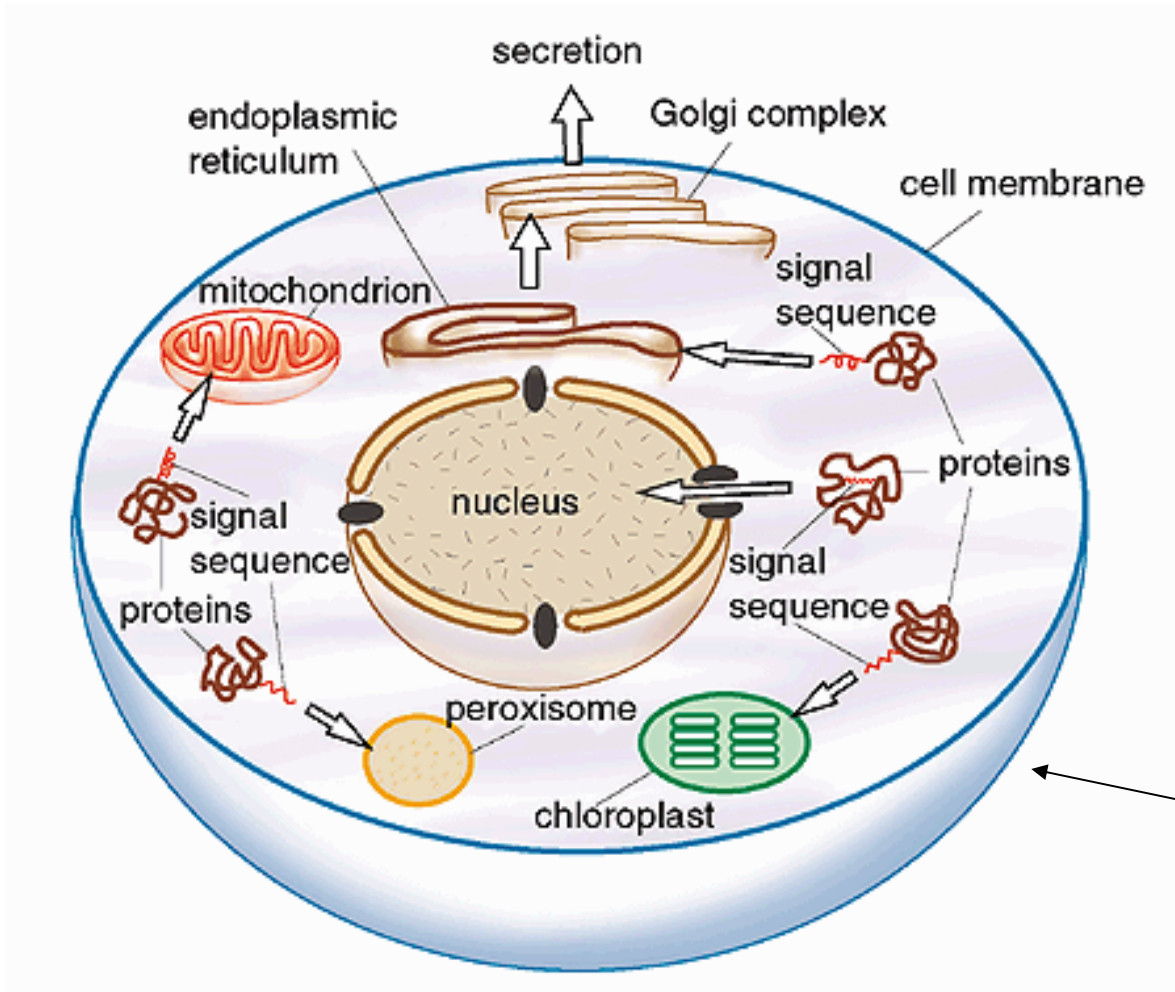
# Why Subcellular Localization?

- The human body contains many different organs with each organ performing a different function. **Cells** also have a set of "little organs," called **organelles**, that are adapted and/or specialized for carrying out one or more vital functions.



- (1) Nucleolus
- (2) Nucleus
- (3) Ribosome
- (4) Vesicle
- (5) Rough endoplasmic reticulum (ER)
- (6) Golgi apparatus
- (7) Cytoskeleton
- (8) Smooth ER
- (9) Mitochondria
- (10) Vacuole
- (11) Cytoplasm
- (12) Lysosome
- (13) Centrioles

# Why Subcellular Localization?



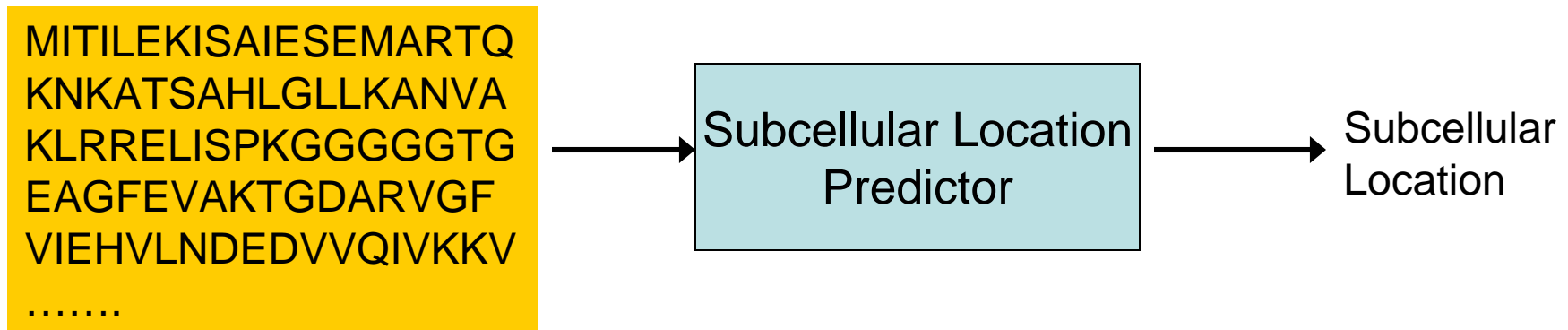
A protein consists of a sequence of amino acids

Amino acid sequence of a protein contains information about its subcellular location

Picture was extracted from [http://redpoll.pharmacy.ualberta.ca/lab\\_talks/ProteinSubcellularLocalization.ppt](http://redpoll.pharmacy.ualberta.ca/lab_talks/ProteinSubcellularLocalization.ppt)

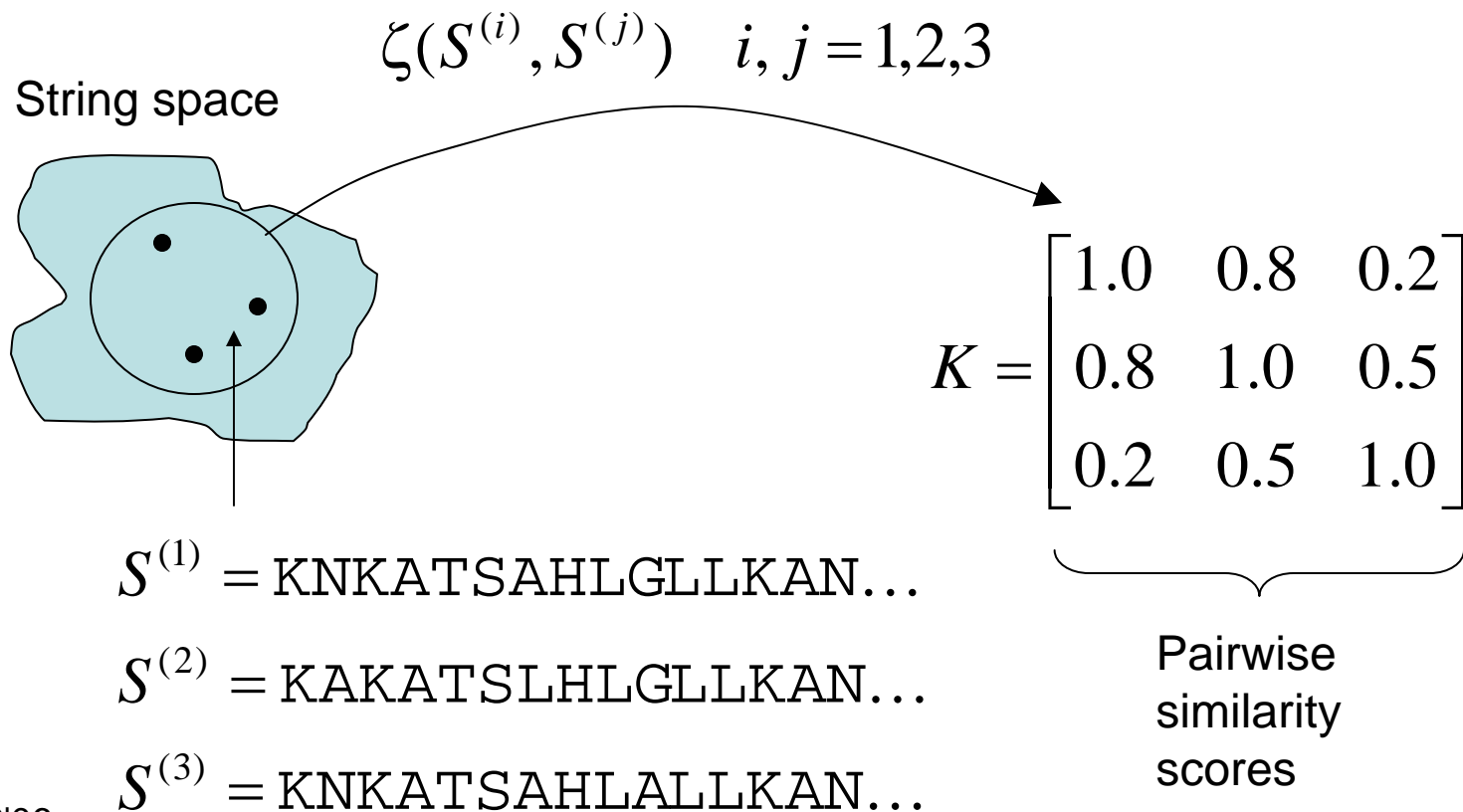
# Why Subcellular Localization?

- Knowledge of subcellular location of proteins has important implication to drug design and discovery of drug targets.
- However, determination of subcellular localization via experimental processes (e.g., microscopy via cell fractionation) is often time-consuming and laborious.
- This motivates the prediction of subcellular locations through amino acid sequences.



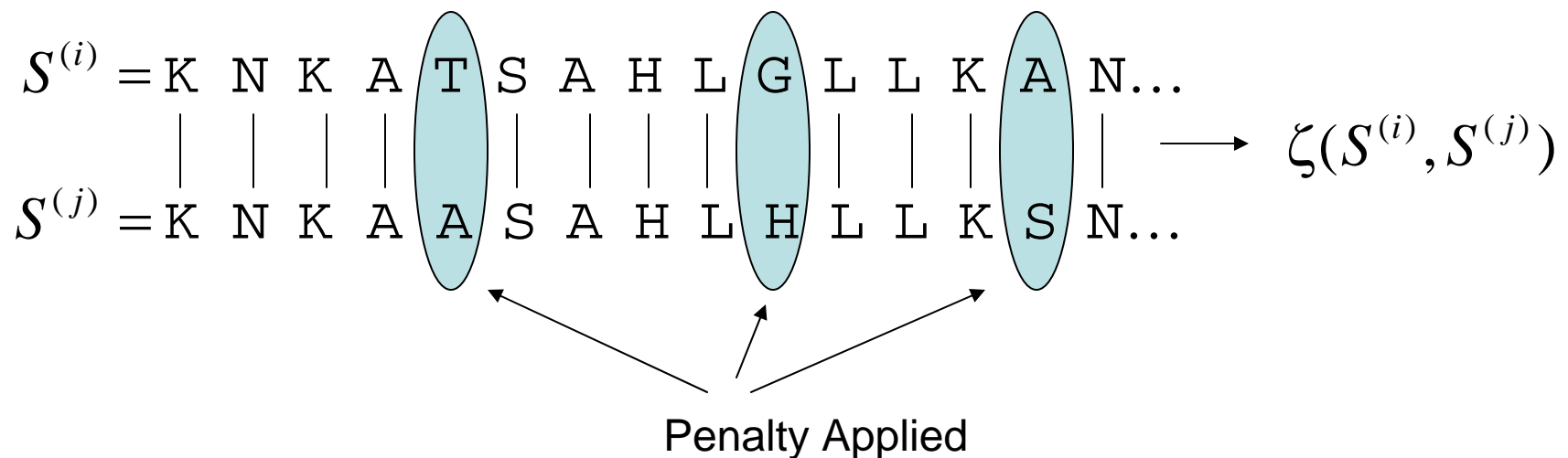
# Feature Extraction

- Because most classifiers work on numbers instead of strings, we need to convert sequences to numbers or vectors.
- This can be solved by kernel methods



# Feature Extraction by Sequence Alignment

- **Idea:** Given a query sequence, we align it against a set of sequences with known subcellular locations to infer its location.
- $\zeta(S^{(i)}, S^{(j)})$  gives the alignment score of sequences  $S^{(i)}$  and  $S^{(j)}$



# Feature Extraction by Sequence Alignment

- Five possible kernels:

$$K_1^{\text{seq}}(S^{(i)}, S^{(j)}) = \zeta(S^{(i)}, S^{(j)})$$

$$K_2^{\text{seq}}(S^{(i)}, S^{(j)}) = \max_{1 \leq l \leq T} \zeta(S^{(i)}, S^{(l)}) \zeta(S^{(j)}, S^{(l)})$$

$$K_3^{\text{seq}}(S^{(i)}, S^{(j)}) = (\zeta(S^{(i)}, S^{(j)}) + 1)^d$$

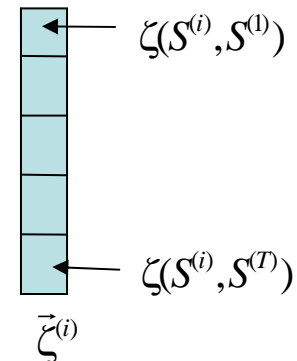
$$K_4^{\text{seq}}(S^{(i)}, S^{(j)}) = \left( \max_{1 \leq l \leq T} \zeta(S^{(i)}, S^{(l)}) \zeta(S^{(j)}, S^{(l)}) + 1 \right)^d$$

$$K_5^{\text{seq}}(S^{(i)}, S^{(j)}) = \sum_{t=1}^T \zeta(S^{(i)}, S^{(t)}) \zeta(S^{(j)}, S^{(t)})$$

Dot product:  $\langle \vec{\zeta}^{(i)}, \vec{\zeta}^{(j)} \rangle$

Special case  
of  $K_5^{\text{seq}}$

Emphasize/  
deemphasize off-  
diagonal  
elements



$T$  is the number of training sequences with known subcellular location

# Feature Extraction by Profile Alignment

- The sensitivity of detecting remote homolog can be improved by replacing **sequence** alignment (comparing amino-acid residues) with **profile** alignment.
- Given a query sequence  $S$  of length  $L$ , PSI-BLAST is used to obtain the **profile** of  $S$ , which is represented by two matrices: Position-Specific Scoring Matrix (PSSM) and Position-Specific Frequency Matrix (PSFM)
- PSSM is an  $L \times 20$  matrix in which the  $(i,j)$ -th entry represents the chance of amino acid in the  $j$ -th position of the query sequence being mutated to amino acid type  $i$  during the evolution process.
- PSFM is an  $L \times 20$  matrix in which the  $(i,j)$ -th entry represents the possibility of having amino acid type  $i$  in position  $j$  of the query sequence.

# Feature Extraction by Profile Alignment

Position-Specific Scoring Matrix (PSSM):

20 Amino Acid

Position  
L

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	\$
1 V	-1	-2	-3	-4	-1	-2	-3	-3	-3	2	1	-2	4	-1	-3	-2	-1	-3	-1	4	\$
2 L	-2	-3	-4	-4	-2	-3	-3	-4	-3	4	5	-3	2	0	-3	-3	-2	-2	-1	1	\$
3 I	-2	-3	-4	-4	-2	-3	-4	-4	-4	5	2	7	4	0	7	-3	-1	-3	-2	2	\$
4 K	-1	2	0	-1	-4	1	1	-2	-1	-3	-3	0	0	0	0	0	-1	-3	-2	-3	\$
5 E	-1	0	-1	1	-4	2	6	-3	0	-4	-3	0	0	0	0	0	-1	-3	-2	-3	\$
6 F	-2	-3	-3	-4	-3	-2	-3	-4	1	-1	-1	-3	-1	5	-4	-2	-2	2	6	-1	\$
7 R	-2	6	-1	-2	-4	1	0	-3	-1	-3	-3	2	-2	-3	-3	-1	-1	-3	-2	-3	\$
8 V	-1	-3	-3	-4	-1	-3	-3	-4	-4	4	1	-3	1	-1	-3	-2	-1	-3	-2	4	\$
9 V	-1	-3	-3	-3	-2	-2	-2	-3	-3	2	0	-2	0	-2	5	-2	-1	-4	-2	3	\$
10 L	-2	-3	-4	-4	-2	-3	-3	-4	-3	1	5	-3	2	0	-3	-3	-2	-2	-1	1	\$
11 P	-1	-3	-2	-2	-3	-2	-1	-3	-3	-3	-3	-1	-3	-4	8	-1	-1	-4	-3	-3	\$
12 C	-1	-3	-3	-4	7	-3	-4	-4	-3	1	2	-3	0	-1	-3	-2	-1	-3	-2	2	\$
13 S	1	-1	0	-1	-1	-1	-1	-1	-1	-2	-2	-1	-2	-3	-1	4	3	-3	-2	-1	\$
14 V	0	-3	-3	-4	-1	-3	-3	-4	-4	2	1	-3	0	-1	-3	-2	0	-3	-2	5	\$
15 Q	-1	0	0	3	-4	4	4	-2	0	-3	-3	1	-2	-4	-2	0	-1	-3	-2	-3	\$
16 E	-1	0	-1	1	-4	2	6	-3	0	-4	-3	1	-2	-4	-1	0	-1	-3	-2	-3	\$
17 Y	-2	-2	-2	-4	-3	-2	-2	-4	2	-2	-1	-2	-1	3	-3	-2	-2	2	8	-2	\$
18 Q	-1	4	0	-1	-4	5	1	-2	0	-3	-3	1	-1	-4	-2	-1	-1	-3	-2	-3	\$
19 V	-1	-3	-3	-4	-1	-3	-3	-4	-4	4	1	-3	1	-1	-3	-2	-1	-3	-2	4	\$
20 G	3	-2	-1	-2	-2	-2	-2	5	-2	-3	-3	-2	-2	-3	-2	0	-1	-3	-3	-2	\$
21 Q	-1	1	0	-1	-3	6	2	-2	0	-3	-3	1	-1	-4	-2	0	-1	-2	-2	-3	\$
22 L	-2	-3	-4	-4	-2	-3	-3	-4	-3	1	5	-3	2	0	-3	-3	-2	-2	-1	1	\$
23 Y	-2	-2	-2	-4	-3	-2	-2	-4	2	-2	-1	-2	-1	3	-3	-2	-2	2	8	-2	\$
24 S	0	-1	0	-1	-1	0	-1	-1	-1	-1	0	-1	4	-2	-2	4	1	-3	-2	-1	\$

# Profile Alignment Kernels

Denote the operation of PSI-BLAST search as

$$\phi^{(i)} \equiv \phi(S^{(i)}) \rightarrow \left\{ \begin{array}{l} \mathbf{P}^{(i)} \\ \text{PSSM} \end{array} , \begin{array}{l} \mathbf{Q}^{(i)} \\ \text{PSFM} \end{array} \right\}$$

The 5 profile alignment kernels are defined as

$$K_1^{\text{pro}}(\phi(S^{(i)}), \phi(S^{(j)})) = \zeta(\phi^{(i)}, \phi^{(j)})$$

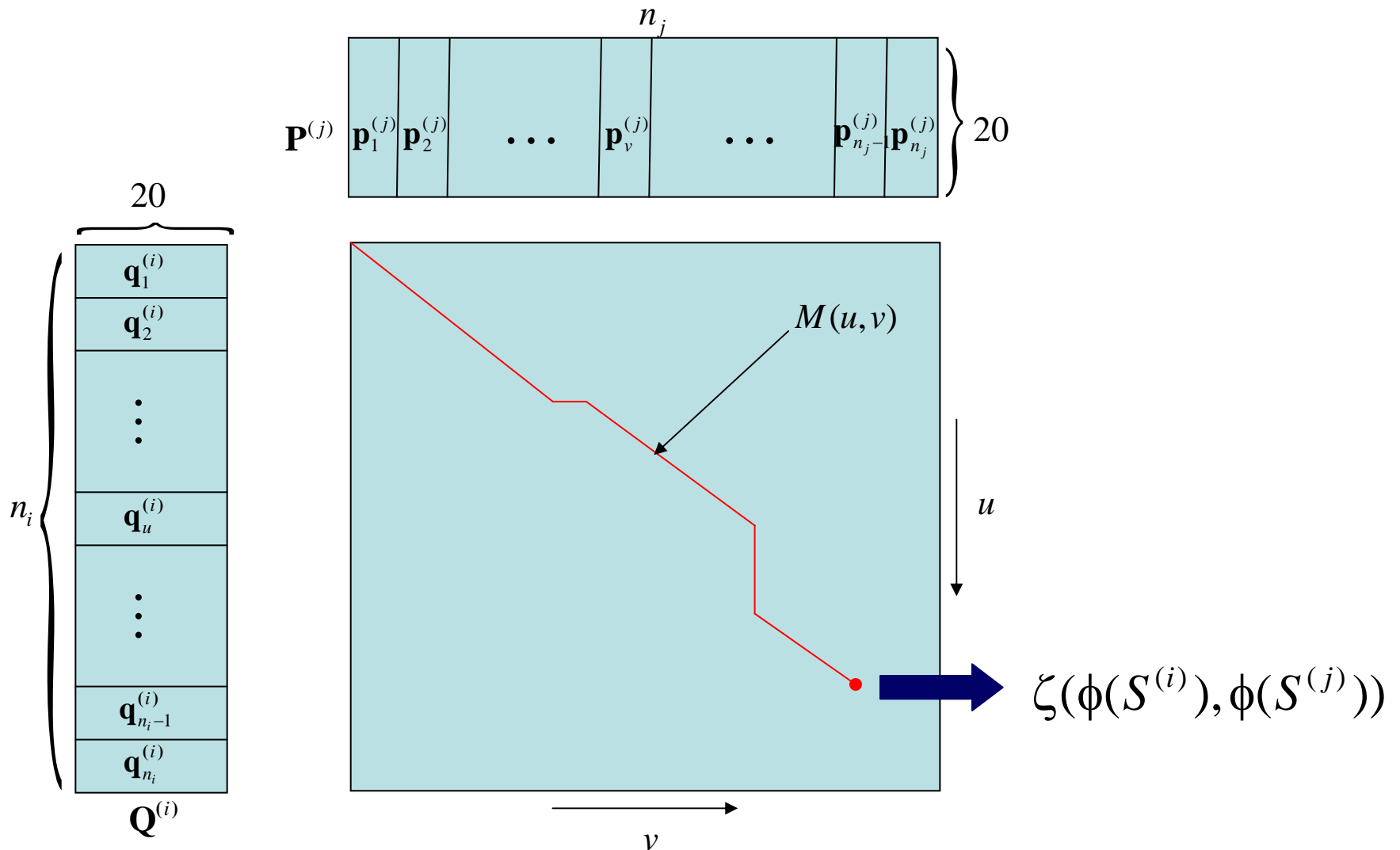
$$K_2^{\text{pro}}(\phi(S^{(i)}), \phi(S^{(j)})) = \max_{1 \leq l \leq T} \zeta(\phi^{(i)}, \phi^{(l)}) \zeta(\phi^{(j)}, \phi^{(l)})$$

$$K_3^{\text{pro}}(\phi(S^{(i)}), \phi(S^{(j)})) = \left( \zeta(\phi^{(i)}, \phi^{(j)}) + 1 \right)^d$$

$$K_4^{\text{pro}}(\phi(S^{(i)}), \phi(S^{(j)})) = \left( \max_{1 \leq l \leq T} \zeta(\phi^{(i)}, \phi^{(l)}) \zeta(\phi^{(j)}, \phi^{(l)}) + 1 \right)^d$$

$$K_5^{\text{pro}}(\phi(S^{(i)}), \phi(S^{(j)})) = \sum_{t=1}^T \zeta(\phi^{(i)}, \phi^{(t)}) \zeta(\phi^{(j)}, \phi^{(t)})$$

# Feature Extraction by Profile Alignment



# Sequence Kernels Vs. Profile Kernels

Sequence  
Kernels

$$K_1^{\text{seq}}(S^{(i)}, S^{(j)}) = \zeta(S^{(i)}, S^{(j)})$$

$$K_2^{\text{seq}}(S^{(i)}, S^{(j)}) = \max_{1 \leq l \leq T} \zeta(S^{(i)}, S^{(l)}) \zeta(S^{(j)}, S^{(l)})$$

$$K_3^{\text{seq}}(S^{(i)}, S^{(j)}) = \left( \zeta(S^{(i)}, S^{(j)}) + 1 \right)^d$$

$$K_4^{\text{seq}}(S^{(i)}, S^{(j)}) = \left( \max_{1 \leq l \leq T} \zeta(S^{(i)}, S^{(l)}) \zeta(S^{(j)}, S^{(l)}) + 1 \right)^d$$

$$K_5^{\text{seq}}(S^{(i)}, S^{(j)}) = \sum_{t=1}^T \zeta(S^{(i)}, S^{(t)}) \zeta(S^{(j)}, S^{(t)})$$

Profile  
Kernels

$$K_1^{\text{pro}}(\phi(S^{(i)}), \phi(S^{(j)})) = \zeta(\phi^{(i)}, \phi^{(j)})$$

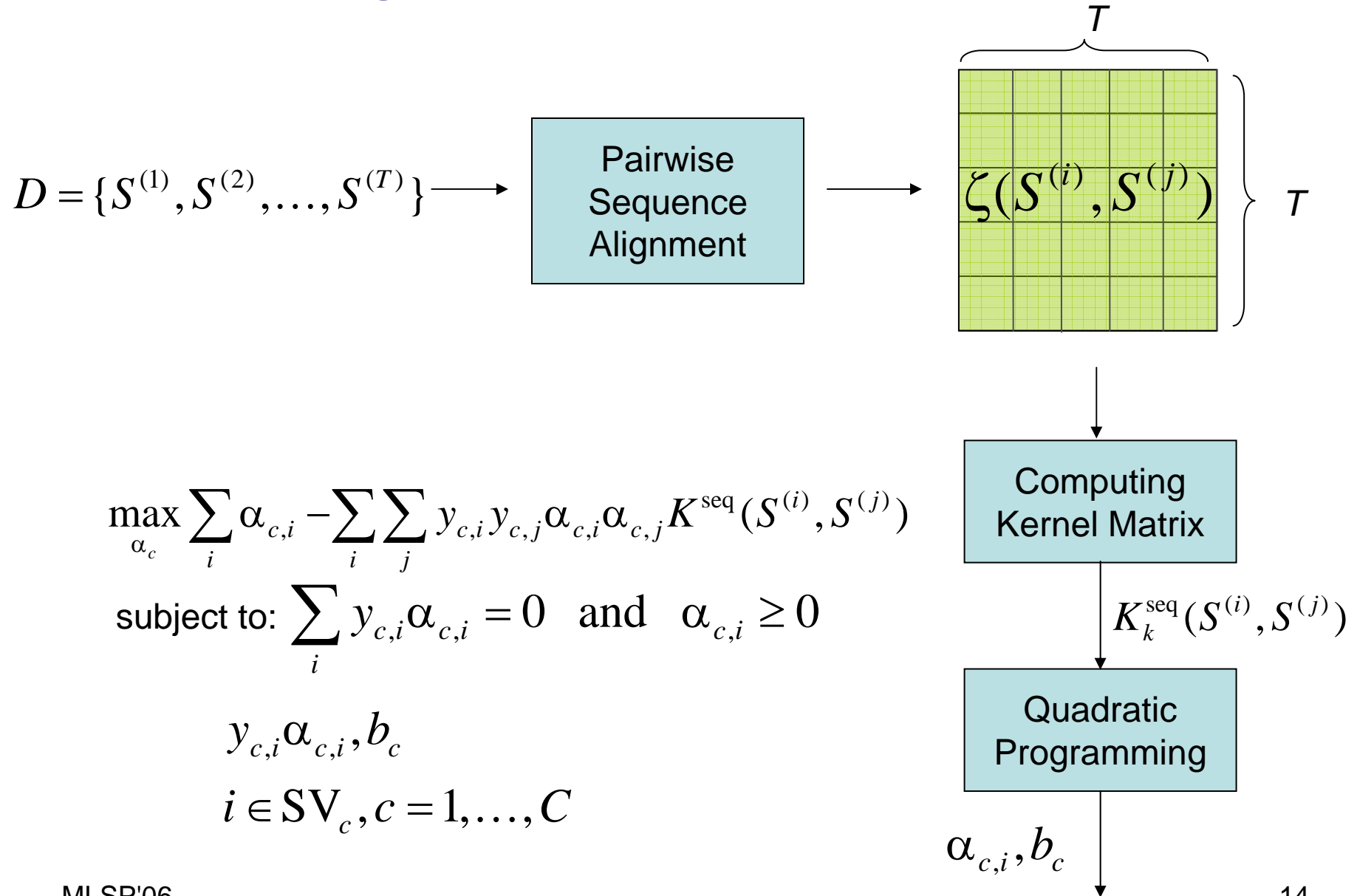
$$K_2^{\text{pro}}(\phi(S^{(i)}), \phi(S^{(j)})) = \max_{1 \leq l \leq T} \zeta(\phi^{(i)}, \phi^{(l)}) \zeta(\phi^{(j)}, \phi^{(l)})$$

$$K_3^{\text{pro}}(\phi(S^{(i)}), \phi(S^{(j)})) = \left( \zeta(\phi^{(i)}, \phi^{(j)}) + 1 \right)^d$$

$$K_4^{\text{pro}}(\phi(S^{(i)}), \phi(S^{(j)})) = \left( \max_{1 \leq l \leq T} \zeta(\phi^{(i)}, \phi^{(l)}) \zeta(\phi^{(j)}, \phi^{(l)}) + 1 \right)^d$$

$$K_5^{\text{pro}}(\phi(S^{(i)}), \phi(S^{(j)})) = \sum_{t=1}^T \zeta(\phi^{(i)}, \phi^{(t)}) \zeta(\phi^{(j)}, \phi^{(t)})$$

# Training 1-vs-Rest SVM Classifier



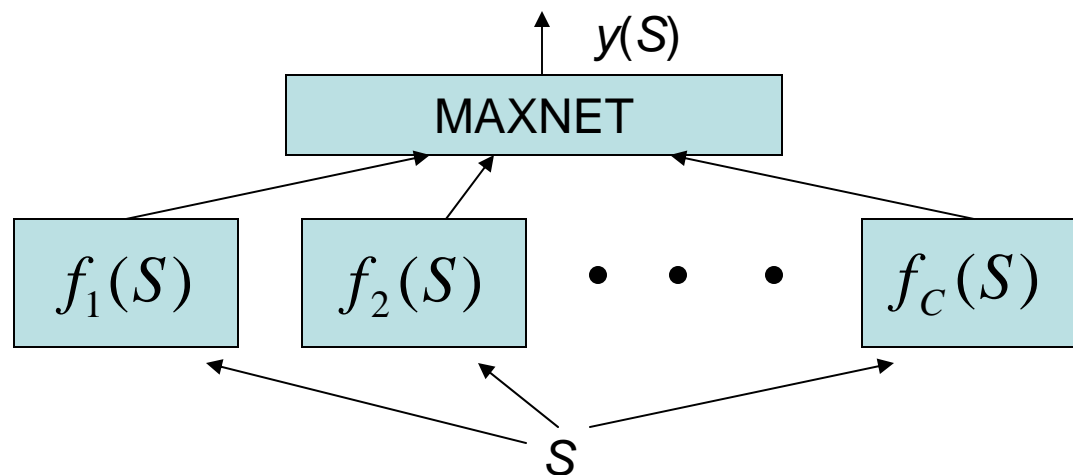
# Classification by 1-vs-Rest SVM

- Given an unknown sequence  $S$ , the score of the  $c$ -th SVM is given by

$$f_c(S) = \sum_{i \in SV_c} y_{c,i} \alpha_{c,i} K_k^{\text{seq}}(S^{(i)}, S) + b_c$$

- Prediction is based on

$$y(S) = \arg \max_{c=1}^C f_c(S)$$



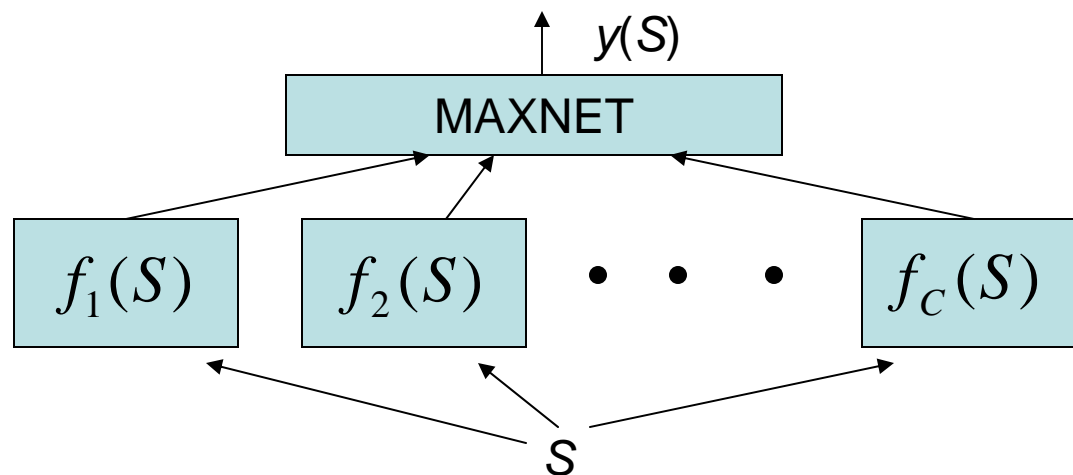
# Classification by 1-vs-Rest SVM

- Given an unknown sequence  $S$ , the score of the  $c$ -th SVM is given by

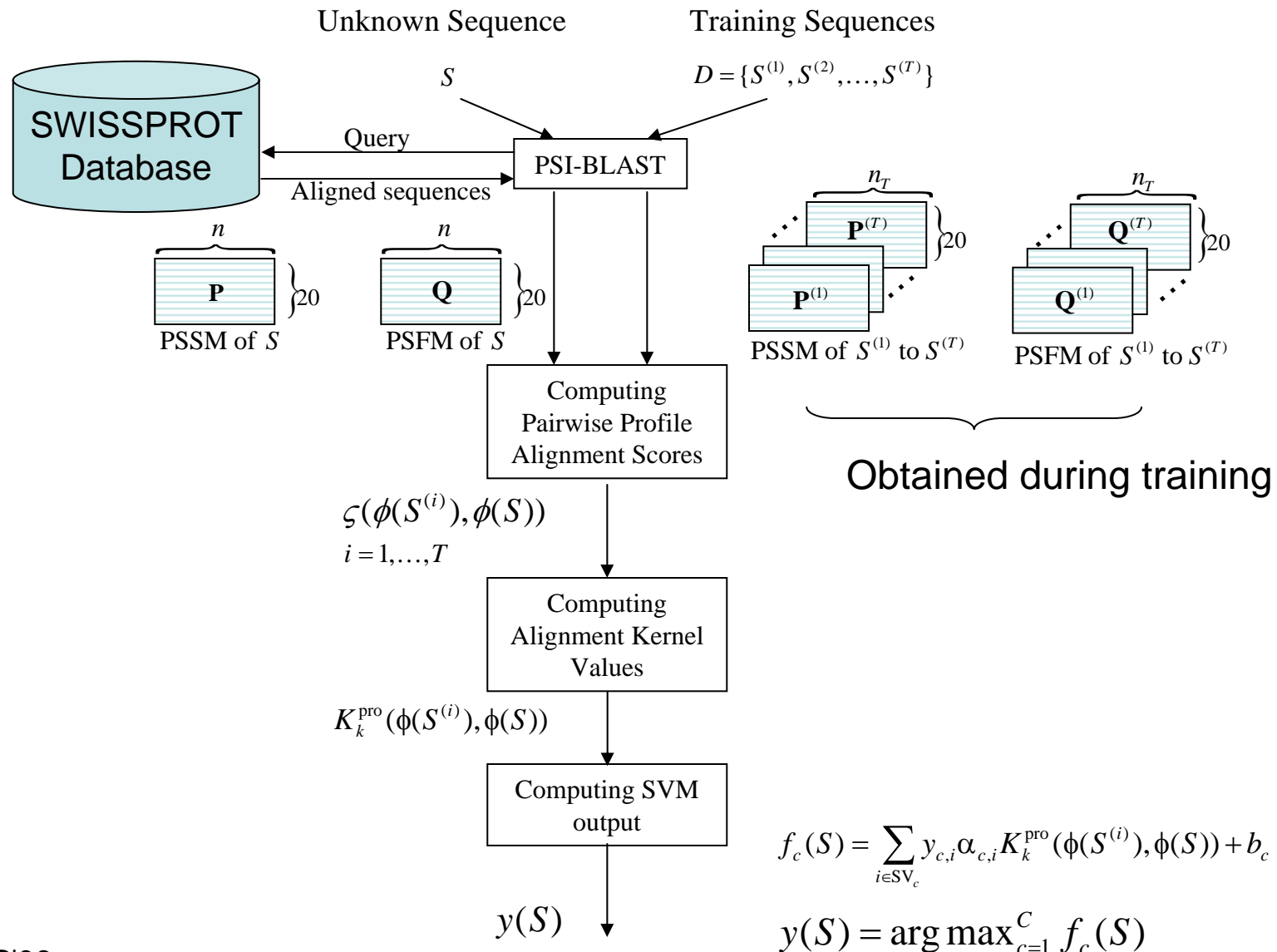
$$f_c(S) = \sum_{i \in \text{SV}_c} y_{c,i} \alpha_{c,i} K_k^{\text{pro}}(\phi(S^{(i)}), \phi(S)) + b_c$$

- Prediction is based on

$$y(S) = \arg \max_{c=1}^C f_c(S)$$



# Prediction by Profile Alignment SVM



# Experiments

- We applied the sequence alignment SVM and profile alignment SVM to a eukaryotic protein dataset (Reinhardt and Hubbard, 1998).
- The dataset comprises 2427 annotated sequences extracted from SWISSPORT 33.0, which amounts to 684 cytoplasm, 325 extracellular, 321 mitochondrial, and 1097 nuclear proteins.
- To mitigate homology bias, we constructed two redundancy-removed datasets by eliminating the most similar sequences.
- 5-Fold cross validation was used to obtain the accuracy.

# Results: Sequence Vs Profile

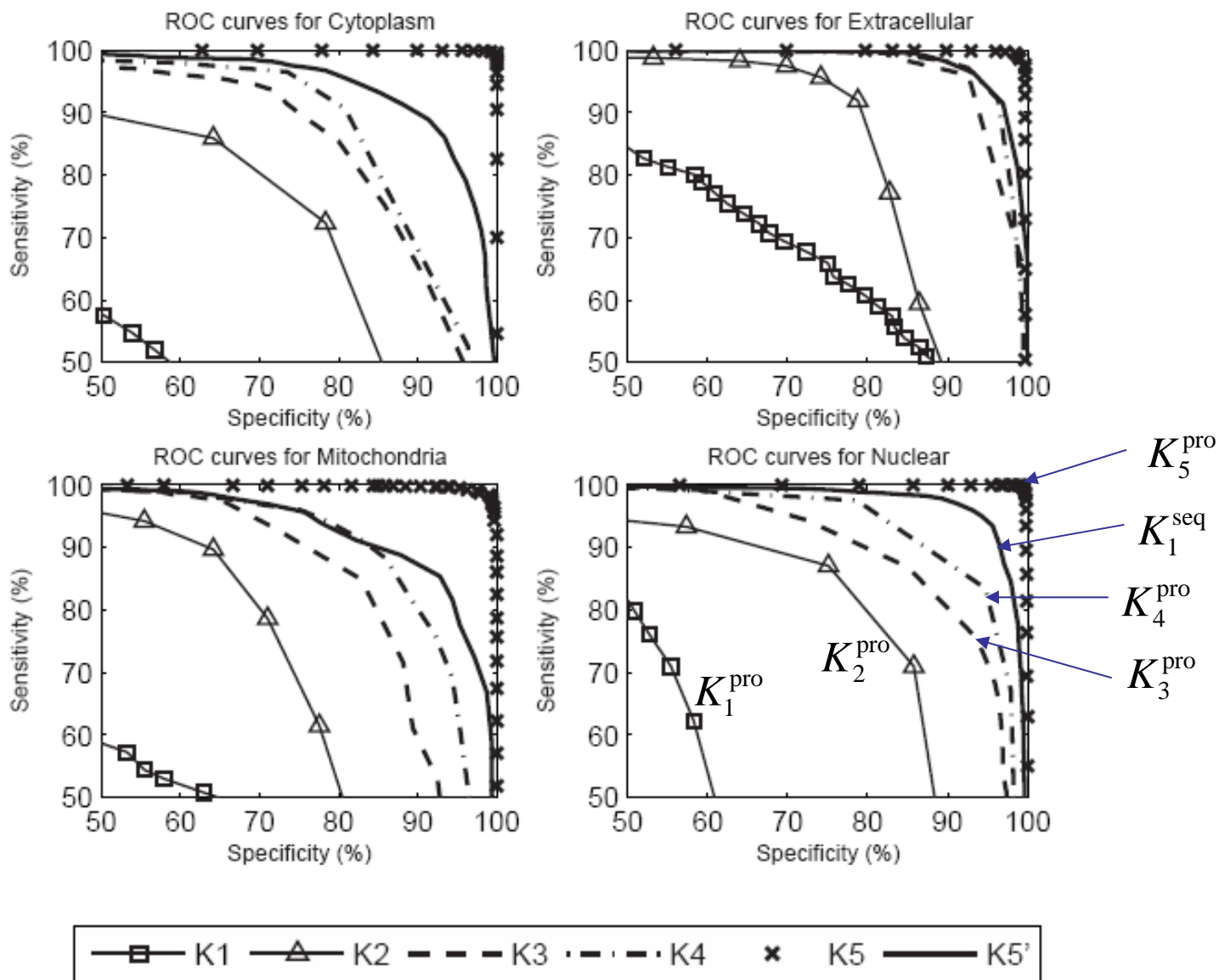
	Sequence Alignment Kernel				
	$K_1^{\text{seq}}$	$K_2^{\text{seq}}$	$K_3^{\text{seq}}$	$K_4^{\text{seq}}$	$K_5^{\text{seq}}$
Accuracy	87.0%	87.1%	87.0%	87.1%	87.1%
% of negative eigenvalues in $K^{\text{seq}}$	0	0	0	0	0
Meeting Mercer's condition	Yes	Yes	Yes	Yes	Yes

	Profile Alignment Kernel				
	$K_1^{\text{pro}}$	$K_2^{\text{pro}}$	$K_3^{\text{pro}}$	$K_4^{\text{pro}}$	$K_5^{\text{pro}}$
Accuracy	45.9%	73.6%	77.1%	82.9%	99.1%
% of negative eigenvalues in $K^{\text{pro}}$	8.5	6.2	0.3	0.2	0.0
Meeting Mercer's condition	No	No	No	No	Yes

## *Key Observations:*

1. The performance of profile alignment SVM is **sensitive** (and somewhat proportional) to the degree of its kernel matrix meeting the Mercer's condition.
2. When Mercer's condition is met, profile alignment SVM achieve higher prediction accuracy.

# Results:ROC



# Results: Comparison with Existing Methods

Subcellular Location	NNPSL	SubLoc		Fuzzy K-NN	
	Acc(%)	Acc(%)	MCC	Acc(%)	MCC
Cytoplasm	55	76.9	0.64	86.7	0.76
Extracellular	75	80.0	0.78	83.7	0.87
Mitochondria	61	56.7	0.58	60.4	0.63
Nuclear	72	87.4	0.75	92.0	0.83
Overall	66	79.4	–	85.2	–
Weighted Average	–	–	0.70	–	0.79

Subcellular Location	ESLpred		PairSeqSVM ( $K_5^{\text{seq}}$ )		PairProSVM ( $K_5^{\text{pro}}$ )	
	Acc(%)	MCC	Acc(%)	MCC	Acc(%)	MCC
Cytoplasm	85.2	0.79	83.2	0.78	100.0	1.00
Extracellular	88.9	0.91	84.3	0.89	98.2	0.97
Mitochondria	68.2	0.69	61.7	0.73	96.9	0.97
Nuclear	95.3	0.87	97.8	0.83	99.5	1.00
Overall	88.0	–	87.1	0.83	99.1	0.99
Weighted Average	–	0.83	–	0.81	–	0.99

# Conclusions

## **Sequence-Based Method (Direct)**

Pros: Simpler and Meet Mercer's condition

Cons: Could not capture remote homology information for subcellular localization

## **Profile-Based Method (Indirect)**

Pros: PSI-BLAST makes use of un-annotated sequences in database to capture richer subcellular localization information

Cons: Most kernels do not meet Mercer's condition. Only K5 does, but it is computationally expensive.

# Further Information

<http://www.eie.polyu.edu.hk/~mwmak/BSIG/PairProSVM.htm>

The screenshot shows a Microsoft Internet Explorer browser window with the title "PairProSVM Supplementary Materials - Microsoft Internet Explorer". The address bar contains the URL <http://www.eie.polyu.edu.hk/~mwmak/BSIG/PairProSVM.htm>. The main content area displays the following text:

## PairProSVM: A New Method for Eukaryotic Protein Subcellular Localization Based on Local Pairwise Profile Alignment and SVM

[Jian Guo](#)<sup>1</sup>, [Man-Wai Mak](#)<sup>1</sup>, [Sun-Yuan Kung](#)<sup>2</sup>

1. Dept. of Electronic and Information Engineering, The Hong Kong Polytechnic University  
2. Dept. of Electrical Engineering, Princeton University

This page provides some supplementary materials for the paper "PairProSVM: A New Method for Eukaryotic Protein Subcellular Localization Based on Local Pairwise Profile Alignment and SVM".

The alignment score matrices (in Matlab .mat and ASCII formats) used in this paper can be downloaded here:

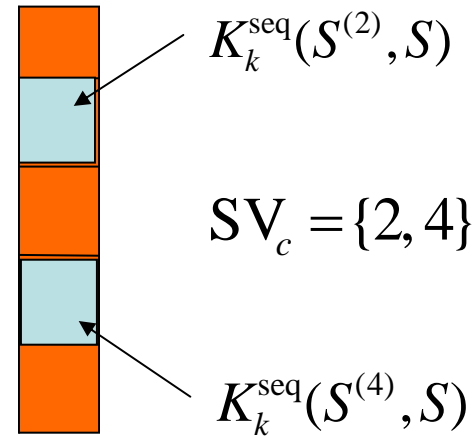
Reinhardt and Hubbard's dataset:	<a href="#">Sequence alignment score matrices for K1, K3, and K5</a>	<a href="#">Profile alignment score matrices for K1, K3, and K5</a>
	Sequence alignment score matrices for K2 and K4: <a href="#">CV1</a> , <a href="#">CV2</a> , <a href="#">CV3</a> , <a href="#">CV4</a> , <a href="#">CV5</a>	Profile alignment score matrices for K2 and K4: <a href="#">CV1</a> , <a href="#">CV2</a> , <a href="#">CV3</a> , <a href="#">CV4</a> , <a href="#">CV5</a>
Huang and Li's dataset:	<a href="#">Sequence alignment score matrices for K1, K3, and K5</a>	<a href="#">Profile alignment score matrices for K1, K3, and K5</a>

The Matlab programs used in this paper can be downloaded [downloaded](#) here

# Complexity Analysis

For  $K_1^{\text{seq}}$  to  $K_4^{\text{seq}}$ ,  $S$  only needs to be aligned with the training sequences that correspond to the support vectors

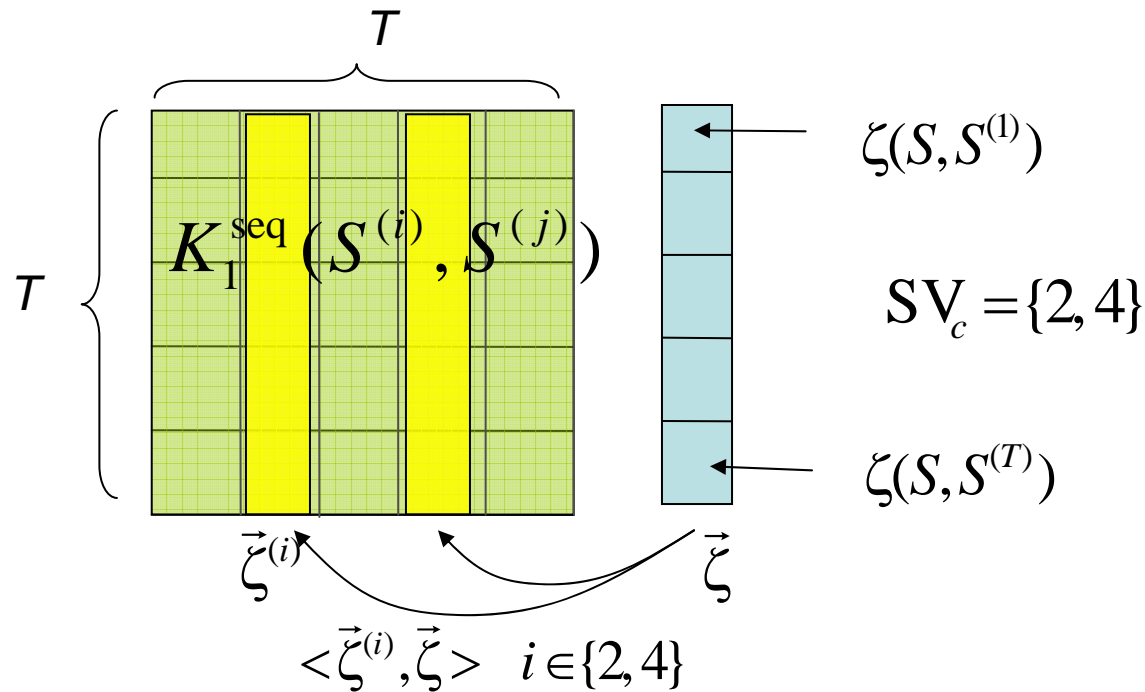
$$\begin{aligned}
 K_1^{\text{seq}}(S^{(i)}, S^{(j)}) &= \zeta(S^{(i)}, S^{(j)}) \\
 K_2^{\text{seq}}(S^{(i)}, S^{(j)}) &= \max_{1 \leq l \leq T} \zeta(S^{(i)}, S^{(l)}) \zeta(S^{(j)}, S^{(l)}) \\
 K_3^{\text{seq}}(S^{(i)}, S^{(j)}) &= (\zeta(S^{(i)}, S^{(j)}) + 1)^d \\
 K_4^{\text{seq}}(S^{(i)}, S^{(j)}) &= \left( \max_{1 \leq l \leq T} \zeta(S^{(i)}, S^{(l)}) \zeta(S^{(j)}, S^{(l)}) + 1 \right)^d
 \end{aligned}$$



$$f_c(S) = \sum_{i \in SV_c} y_{c,i} \alpha_{c,i} K_k^{\text{seq}}(S^{(i)}, S) + b_c, \quad k = 1, \dots, 4$$

# Complexity Analysis

For  $K_5^{\text{seq}}$ ,  $S$  needs to be aligned with **all** training sequences



$$\begin{aligned}
 f_c(S) &= \sum_{i \in SV_c} y_{c,i} \alpha_{c,i} K_5^{\text{seq}}(S^{(i)}, S) + b_c = \sum_{i \in SV_c} y_{c,i} \alpha_{c,i} \sum_{t=1}^T \zeta(S^{(i)}, S^{(t)}) \zeta(S, S^{(t)}) + b_c \\
 &= \sum_{i \in SV_c} y_{c,i} \alpha_{c,i} \langle \vec{\zeta}^{(i)}, \vec{\zeta} \rangle + b_c
 \end{aligned}$$

# Global Vs. Local Alignment

