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,

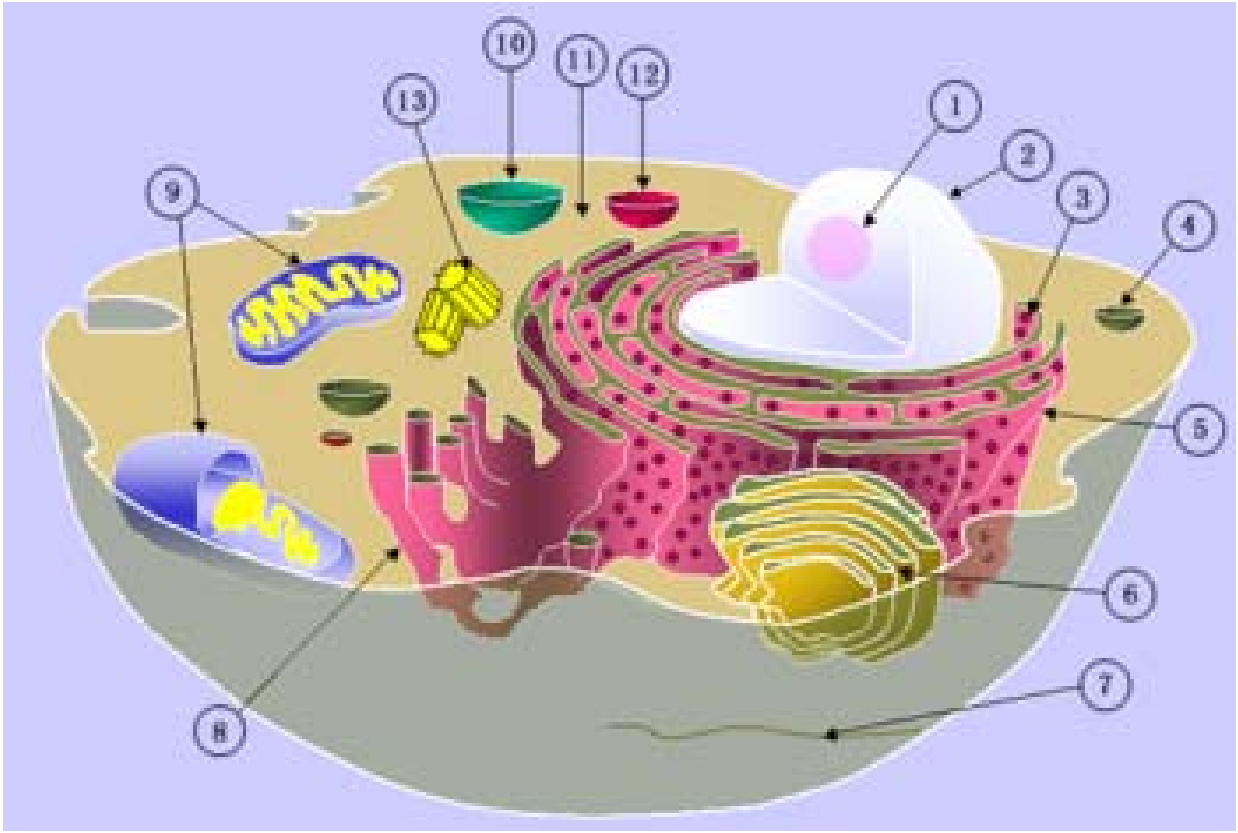
Princeton University

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?

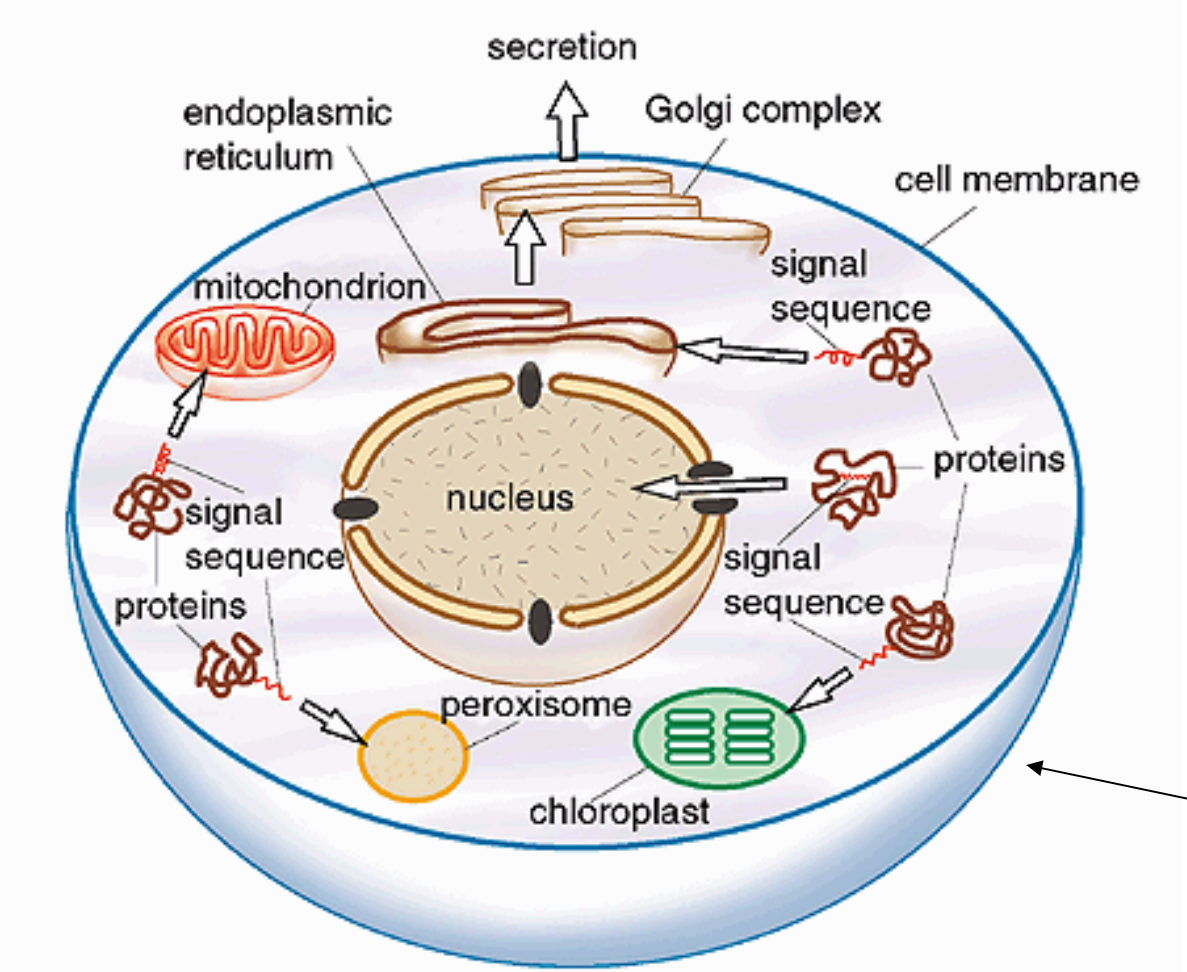
- The human body contains many different organs with each organ performing a different function. **Cells** also have a set of "little organs," called **organelles**, that are adapted and/or specialized for carrying out one or more vital functions.



(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

Picture from http://en.wikipedia.org/wiki/Cell_(biology)

Why Subcellular Localization?



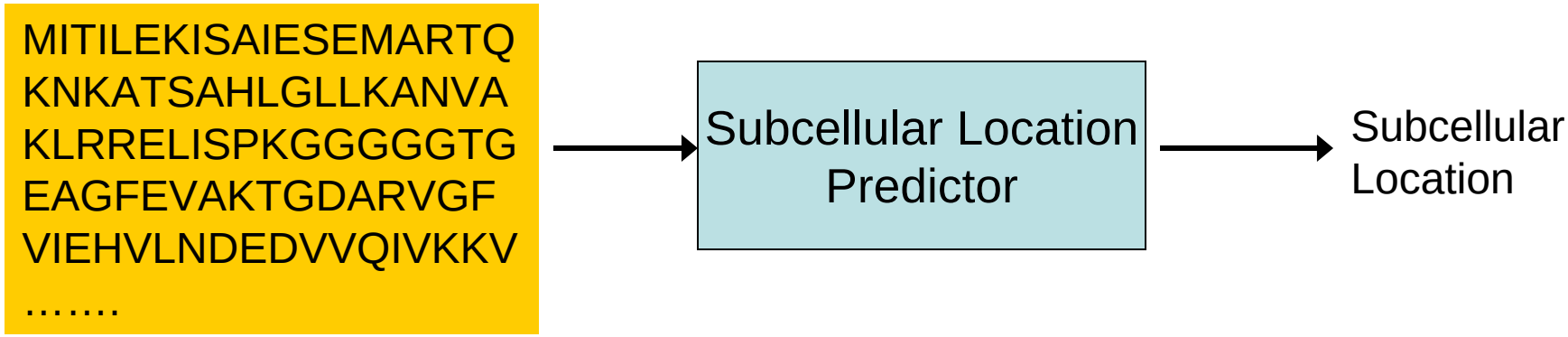
A protein consists of a sequence of amino acids

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

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.

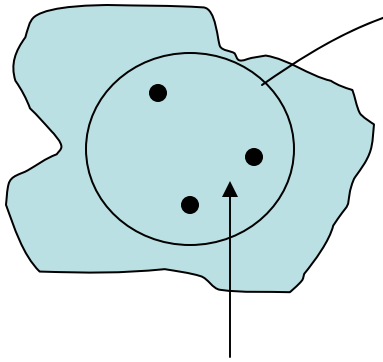


Feature Extraction

- 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

$$\zeta(S^{(i)}, S^{(j)}) \quad i, j = 1,2,3$$

String space



$$K = \begin{bmatrix} 1.0 & 0.8 & 0.2 \\ 0.8 & 1.0 & 0.5 \\ 0.2 & 0.5 & 1.0 \end{bmatrix}$$

Pairwise similarity scores

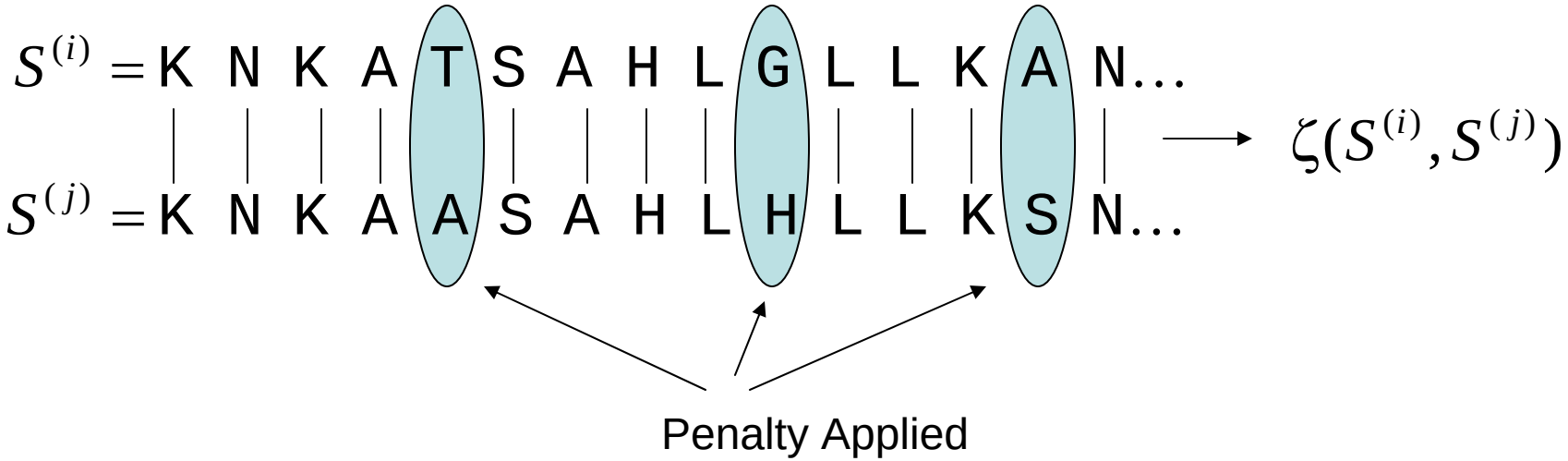
$$S^{(1)} = \text{KNKATSAHLGLLKAN}\ldots$$

$$S^{(2)} = \text{KAKATSLHLGLLKAN}\ldots$$

$$S^{(3)} = \text{KNKATSAHLALLKAN}\ldots$$

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:

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

Special case of K_5^{seq}

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

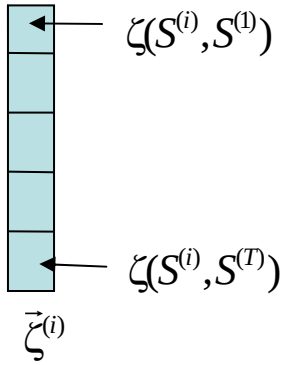
$$K_3^{\text{seq}}(S^{(i)}, S^{(j)}) = \left(\zeta(S^{(i)}, S^{(j)}) + 1\right)^d$$

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

Emphasize/ deemphasize off-diagonal elements

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

Dot product: $< \vec{\zeta}^{(i)}, \vec{\zeta}^{(j)} >$



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 *Lx*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 *Lx*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.

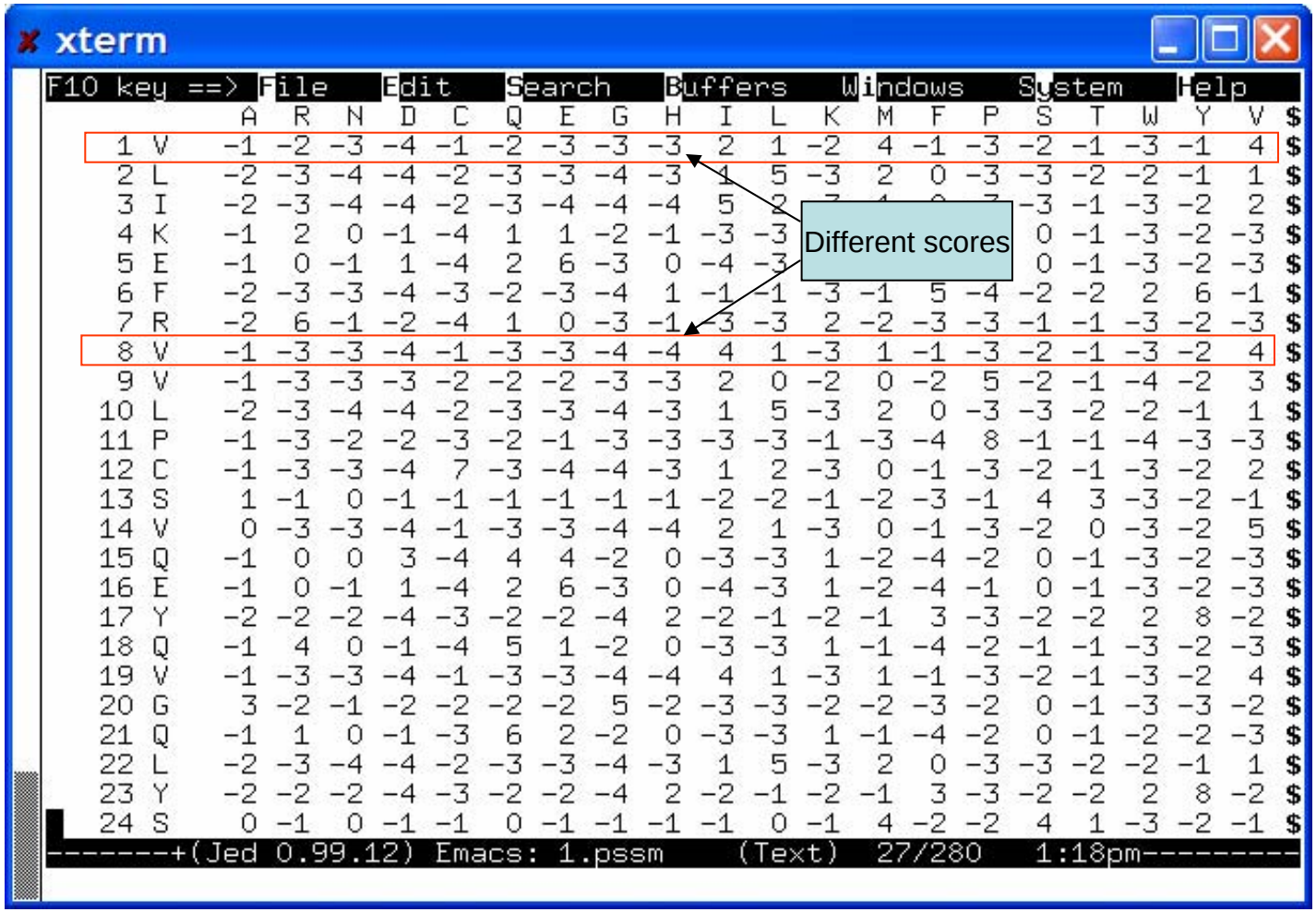
Feature Extraction by Profile Alignment

Position-Specific Scoring Matrix (PSSM):

20 Amino Acid

Position

L



Profile Alignment Kernels

Denote the operation of PSI-BLAST search as

$$\phi^{(i)} \equiv \phi(S^{(i)}) \rightarrow \{\underbrace{\mathbf{P}^{(i)}}_{\text{PSSM}}, \underbrace{\mathbf{Q}^{(i)}}_{\text{PSFM}}\}$$

The 5 profile alignment kernels are defined as

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

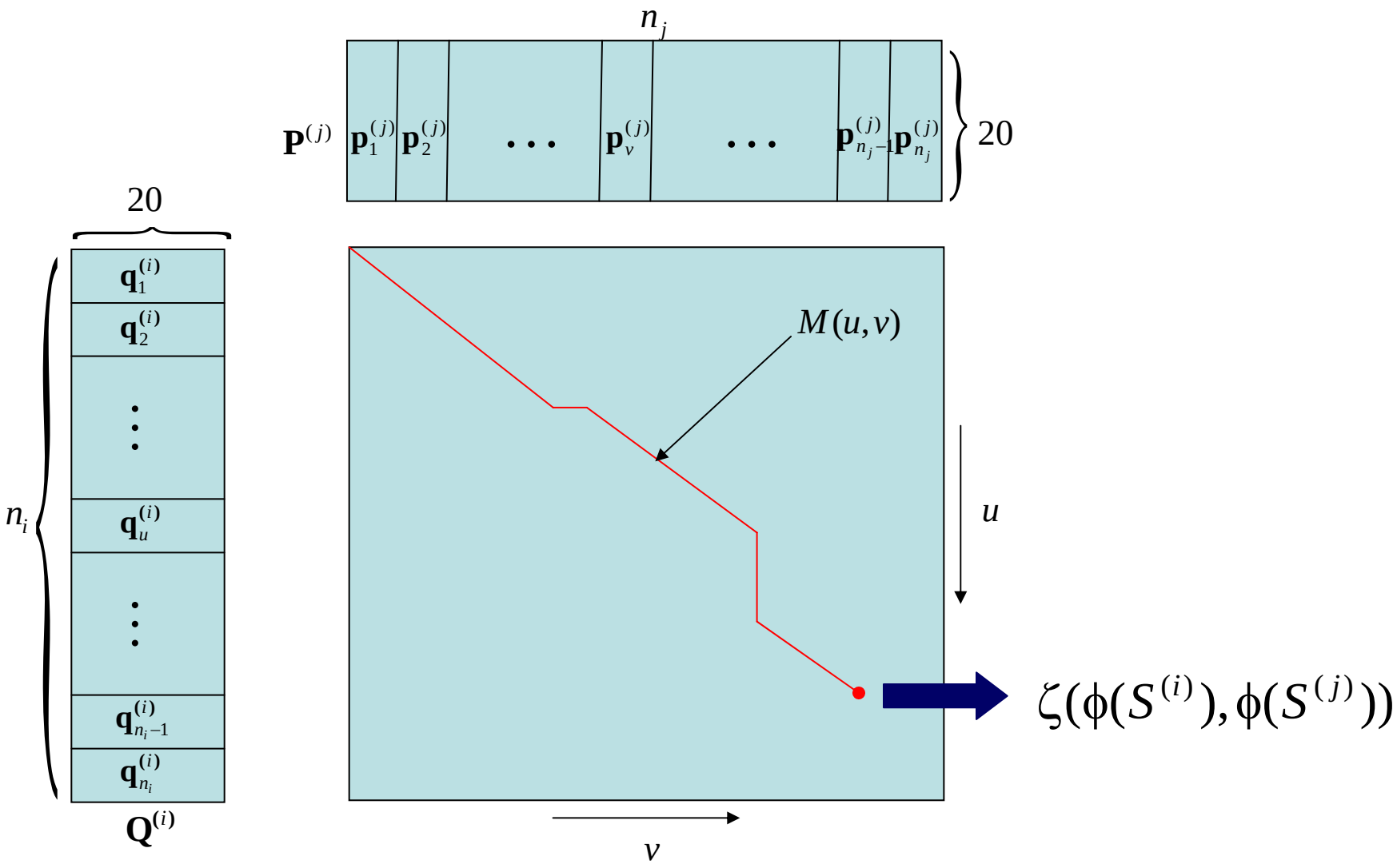
$$K_2^{\text{pro}}(\phi(S^{(i)}), \phi(S^{(j)})) = \max_{1 \le l \le T} \zeta(\phi^{(i)}, \phi^{(l)}) \zeta(\phi^{(j)}, \phi^{(l)})$$

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

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

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

Feature Extraction by Profile Alignment



Sequence Kernels Vs. Profile Kernels

Sequence Kernels

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

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

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

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

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

Profile Kernels

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

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

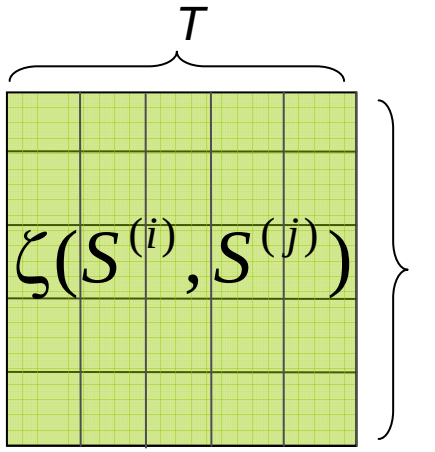
$$K_3^{\text{pro}}(\phi(S^{(i)}), \phi(S^{(j)})) = \left(\zeta(\phi^{(i)}, \phi^{(j)}) + 1\right)^d$$

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

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

Training 1-vs-Rest SVM Classifier

$D = \{S^{(1)}, S^{(2)}, \ldots, S^{(T)}\}$ → Pairwise Sequence Alignment →



Computing Kernel Matrix

$K_k^{\text{seq}}(S^{(i)}, S^{(j)})$

Quadratic Programming

$\alpha_{c,i}, b_c$

$$\max_{\alpha_c} \sum_i \alpha_{c,i} - \sum_i \sum_j y_{c,i} y_{c,j} \alpha_{c,i} \alpha_{c,j} K^{\text{seq}}(S^{(i)}, S^{(j)})$$

subject to: $\sum_i y_{c,i} \alpha_{c,i} = 0 \quad \text{and} \quad \alpha_{c,i} \geq 0$

$$y_{c,i} \alpha_{c,i}, b_c$$

$$i \in \text{SV}_c, c = 1, \ldots, C$$

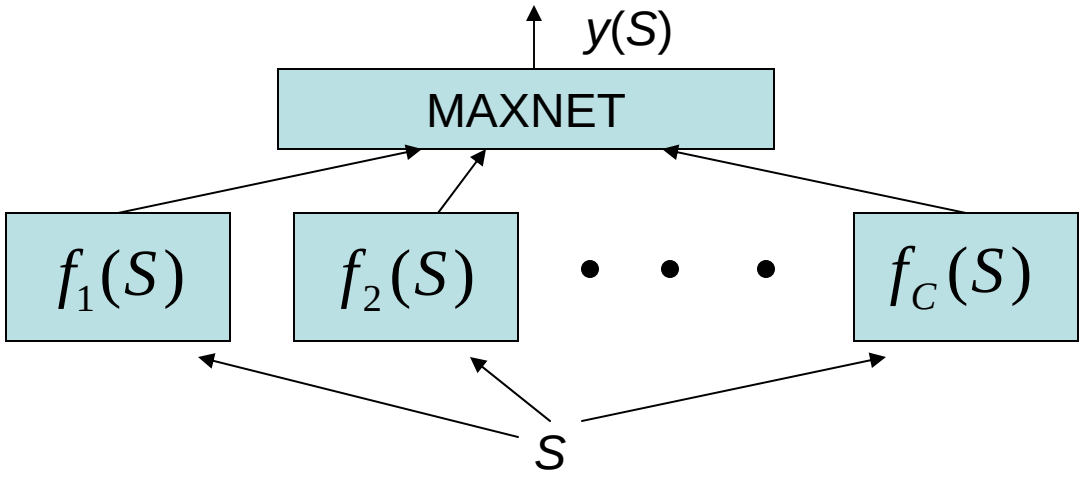
Classification by 1-vs-Rest SVM

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

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

- Prediction is based on

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



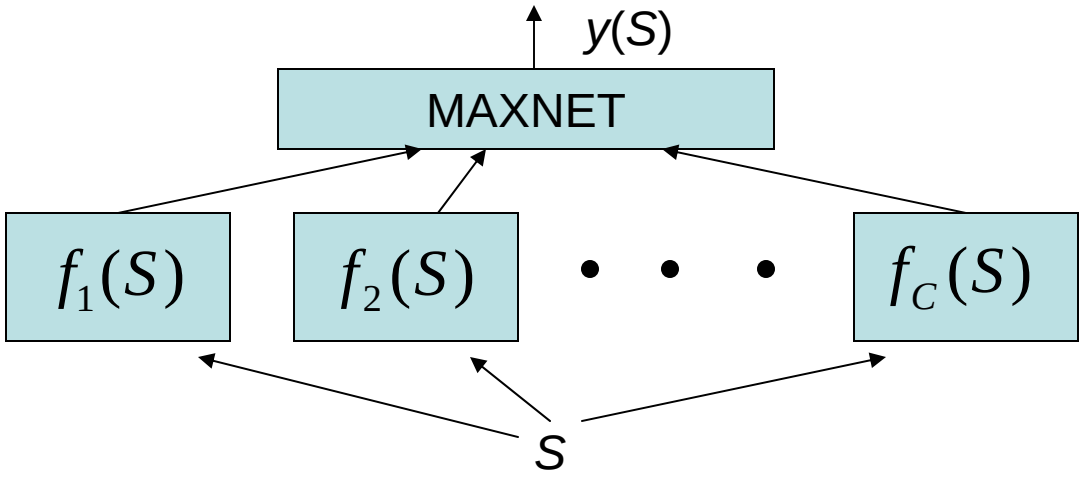
Classification by 1-vs-Rest SVM

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

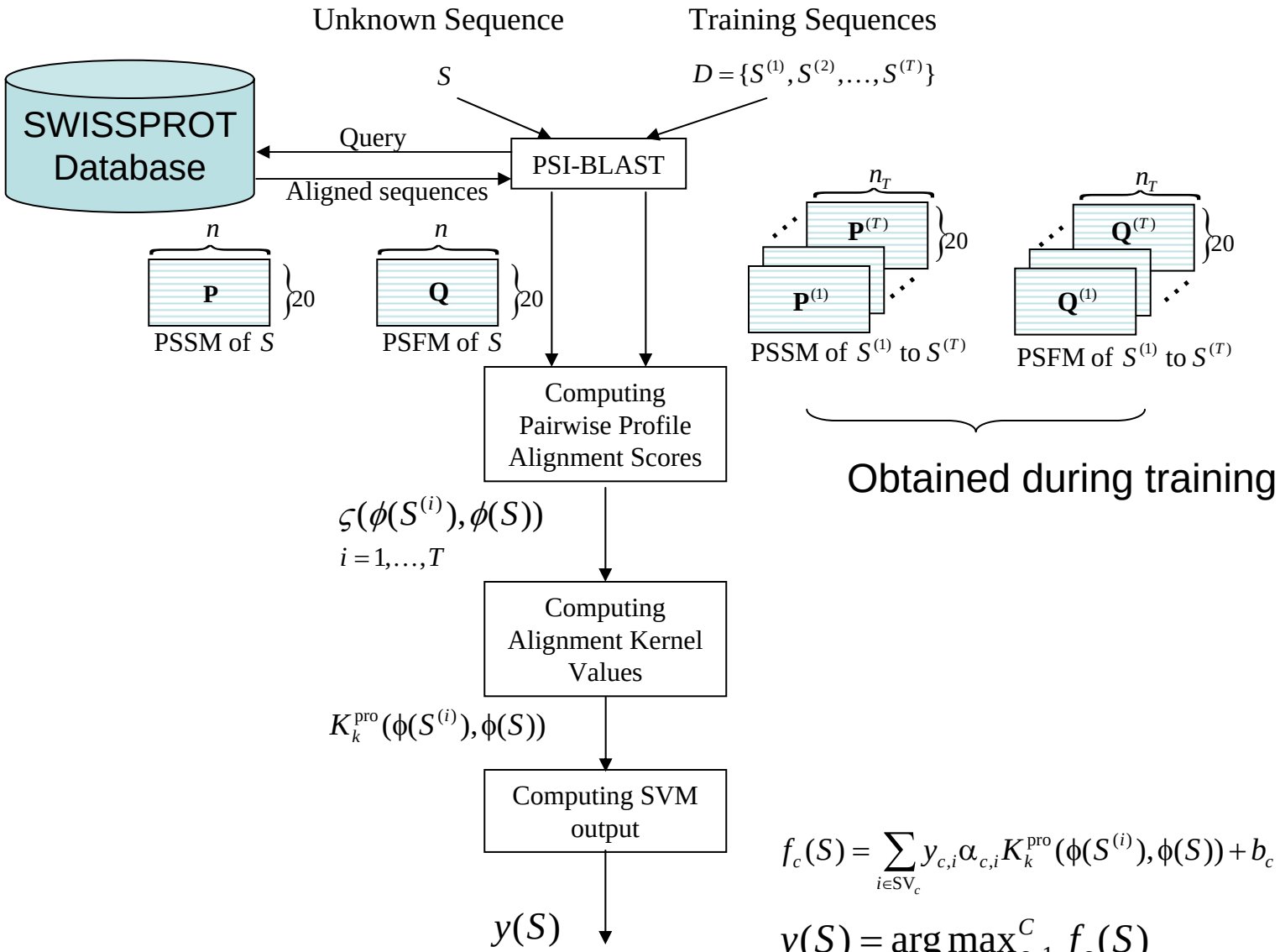
$$f_c(S) = \sum_{i \in \mathrm{SV}_c} y_{c,i} \alpha_{c,i} K_k^{\mathrm{pro}}(\phi(S^{(i)}), \phi(S)) + b_c$$

- 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

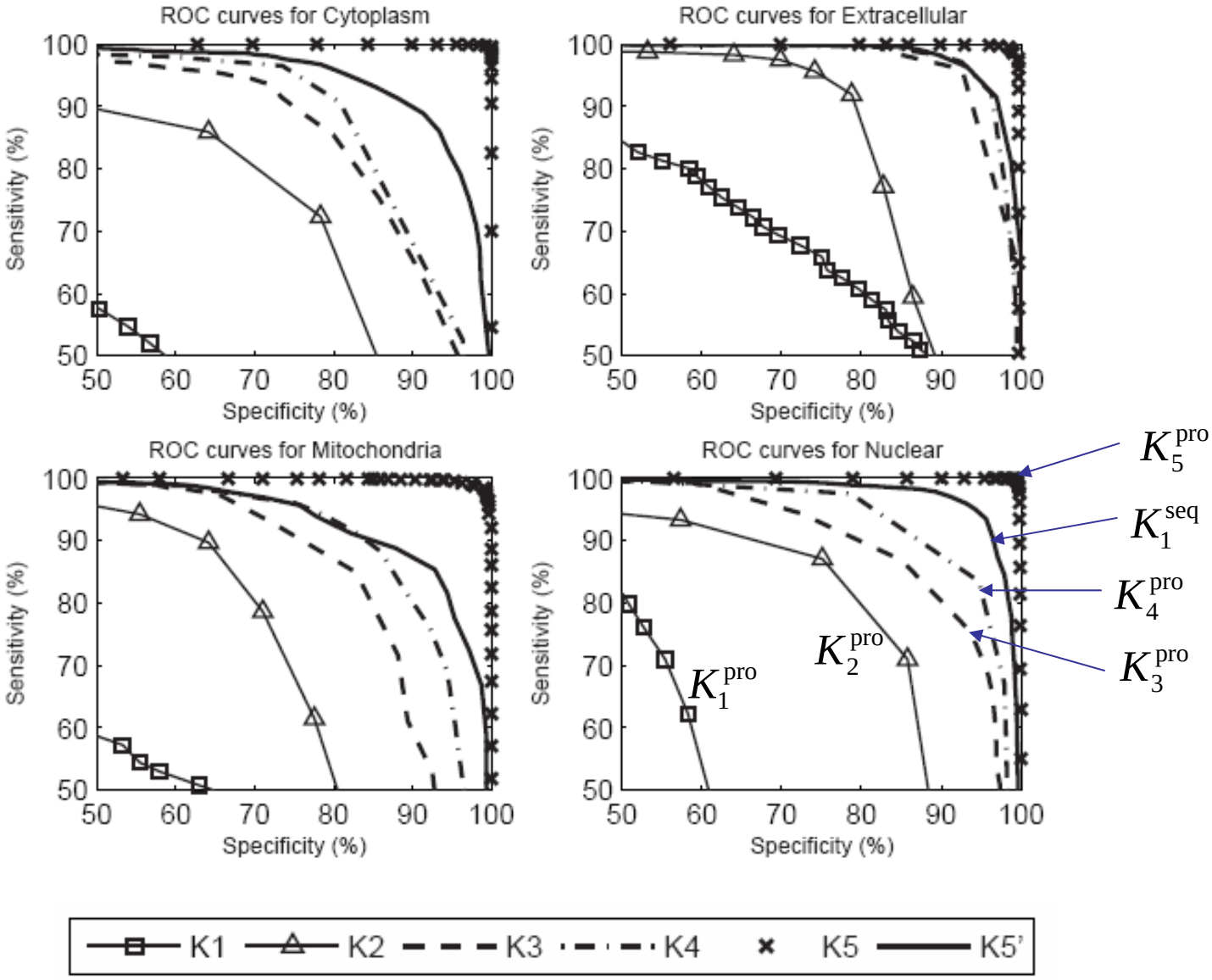
	Sequence Alignment Kernel				
	K_1^{seq}	K_2^{seq}	K_3^{seq}	K_4^{seq}	K_5^{seq}
Accuracy	87.0%	87.1%	87.0%	87.1%	87.1%
% of negative eigenvalues in K^{seq}	0	0	0	0	0
Meeting Mercer's condition	Yes	Yes	Yes	Yes	Yes

	Profile Alignment Kernel				
	K_1^{pro}	K_2^{pro}	K_3^{pro}	K_4^{pro}	K_5^{pro}
Accuracy	45.9%	73.6%	77.1%	82.9%	99.1%
% of negative eigenvalues in K^{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


ROC curves for Cytoplasm
ROC curves for Extracellular
ROC curves for Mitochondria
ROC curves for Nuclear
Sensitivity (%)
Specificity (%)
100
90
80
70
60
50
50
60
70
80
90
100
K_5^{pro}
K_1^{seq}
K_4^{pro}
K_3^{pro}
K_1^{pro}
K_2^{pro}
K1
K2
K3
K4
K5
K5'

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
Mitochrondria	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^{seq})		PairProSVM (K_5^{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
Mitochrondria	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

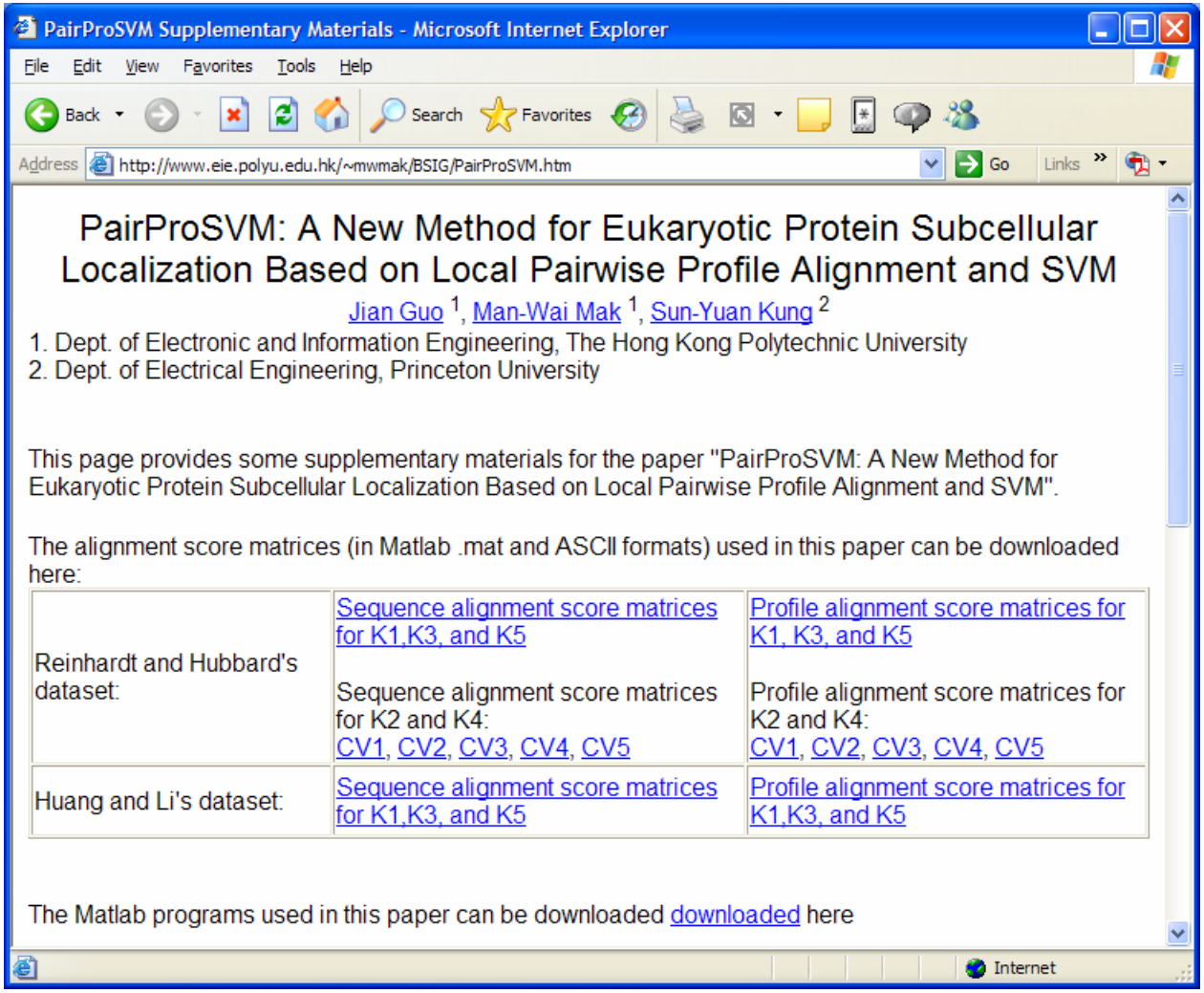
Profile-Based Method (Indirect)

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.

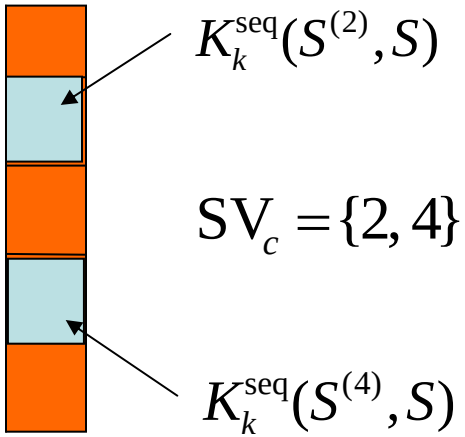
Further Information

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



Complexity Analysis

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

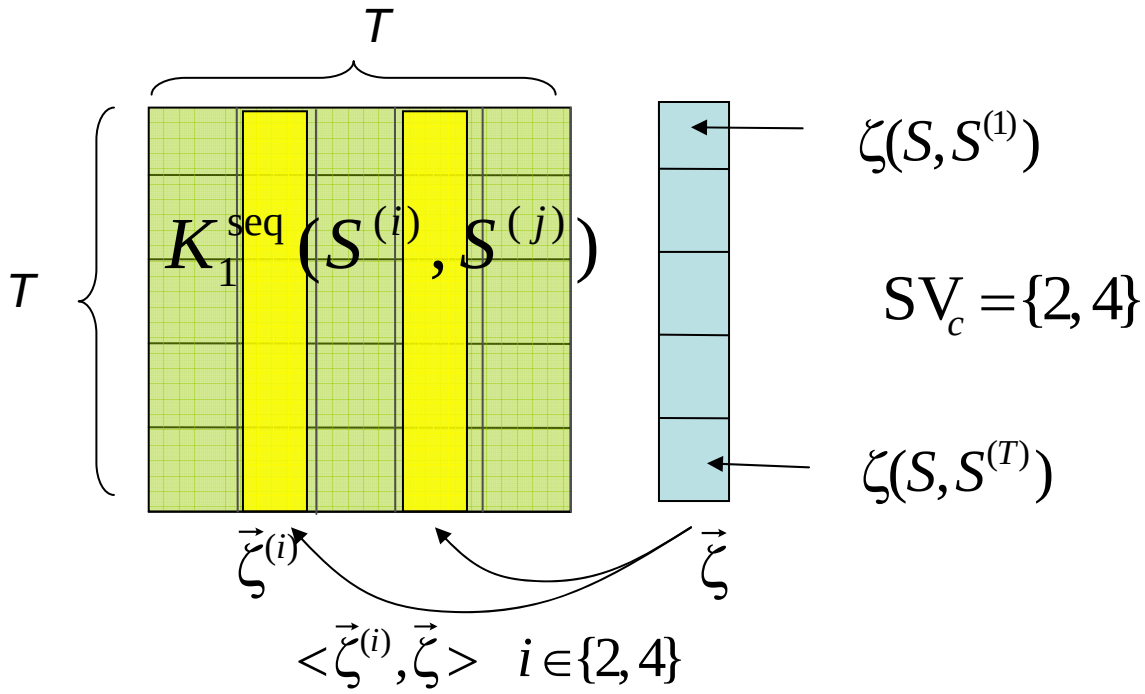


$$K_1^{\text{seq}}(S^{(i)}, S^{(j)}) = \zeta(S^{(i)}, S^{(j)})$$
$$K_2^{\text{seq}}(S^{(i)}, S^{(j)}) = \max_{1 \le l \le T} \zeta(S^{(i)}, S^{(l)})\zeta(S^{(j)}, S^{(l)})$$
$$K_3^{\text{seq}}(S^{(i)}, S^{(j)}) = \left(\zeta(S^{(i)}, S^{(j)}) + 1\right)^d$$
$$K_4^{\text{seq}}(S^{(i)}, S^{(j)}) = \left(\max_{1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

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

Complexity Analysis

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



$$f_c(S) = \sum_{i \in SV_c} y_{c,i} \alpha_{c,i} K_5^{\text{seq}}(S^{(i)}, S) + b_c = \sum_{i \in SV_c} y_{c,i} \alpha_{c,i} \sum_{t=1}^{T} \zeta(S^{(i)}, S^{(t)}) \zeta(S, S^{(t)}) + b_c$$

$$= \sum_{i \in SV_c} y_{c,i} \alpha_{c,i} <\vec{\zeta}^{(i)}, \vec{\zeta}> + b_c$$

Global Vs. Local Alignment