A solution to the Curse of Dimensionality Problem in Pairwise Scoring Techniques

Man Wai MAK

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

Sun Yuan KUNG

Dept. of Electrical Engineering,

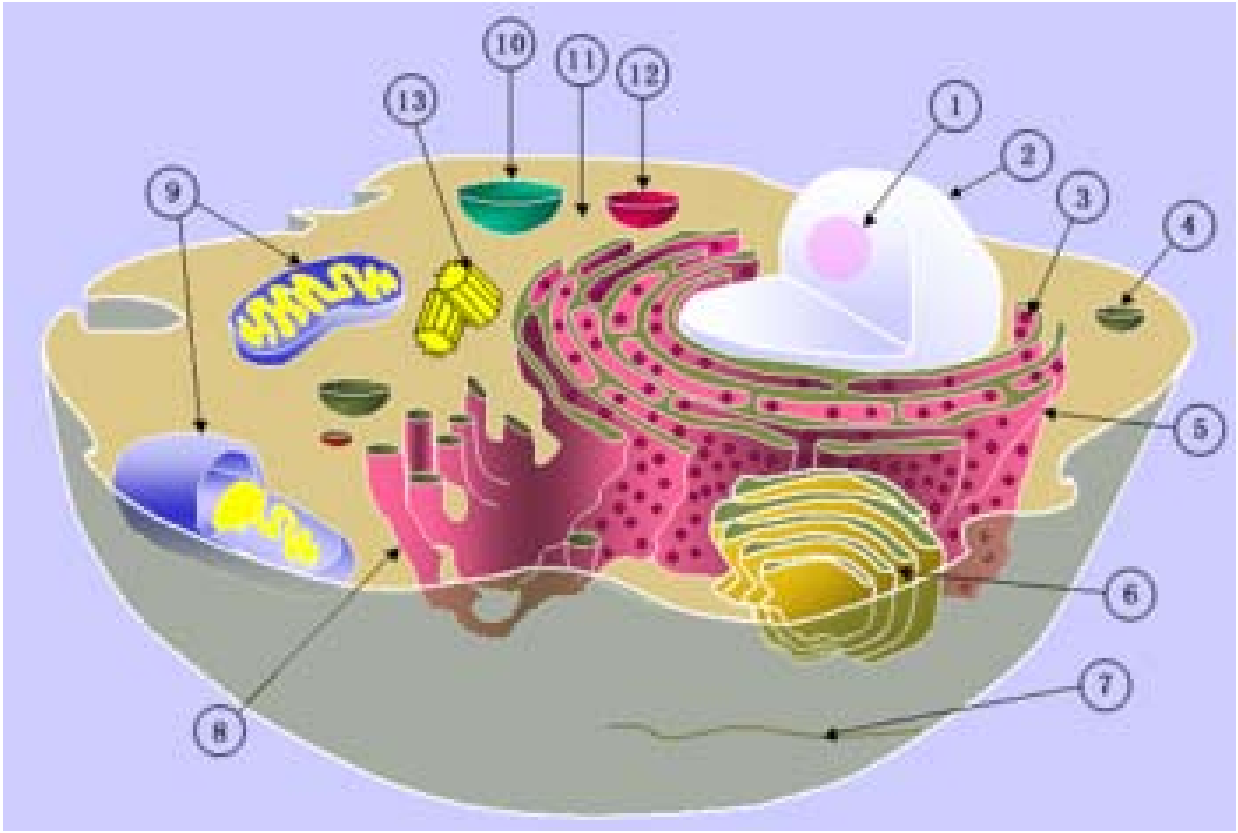
Princeton University

Outline

- Protein Sequences and Subcellular Localization
- Pairwise Scoring Kernels
- Feature Selection
- Results and Conclusions

Cells

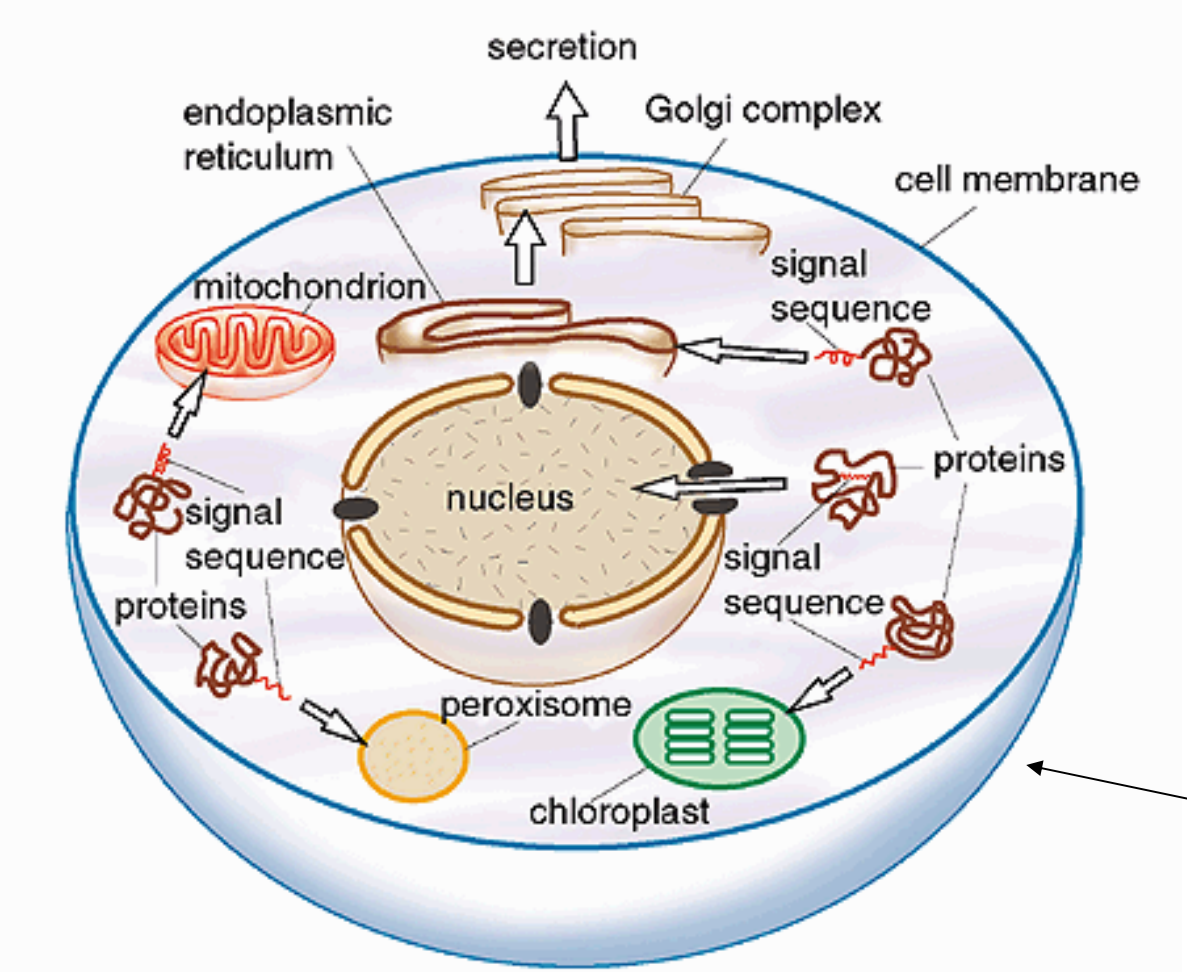
- The human body contains many different organs with each organ performing a different function. **Cells** also have a set of "little organs," called **organelles**, that are adapted and/or specialized for carrying out one or more vital functions.



(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)

Cell and Protein Sequences



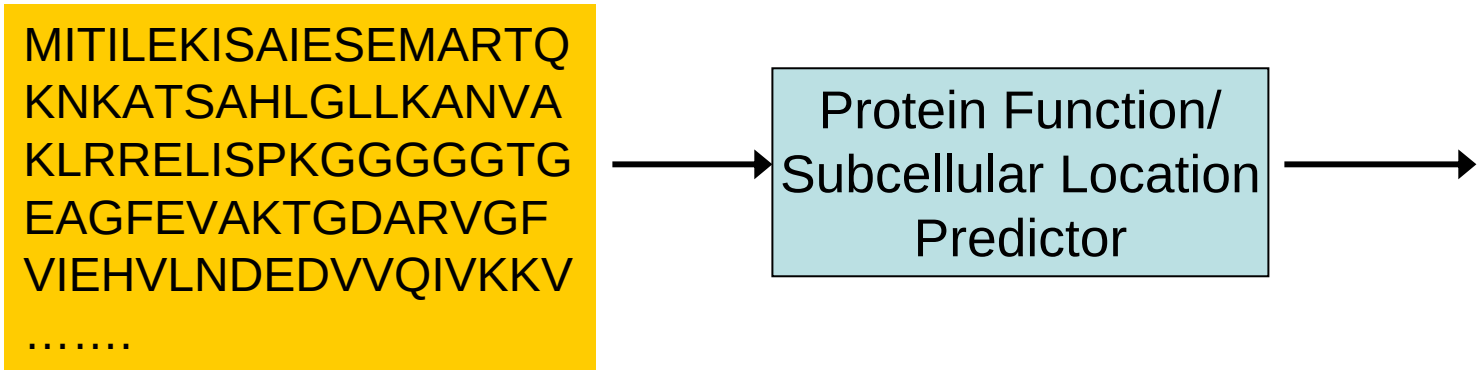
A protein consists of a sequence of amino acids

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

Protein Sequences

- Proteins are represented by sequences of 20 alphabets (amino acid).
- The function and subcellular locations of proteins can be predicted by looking at their corresponding sequences.

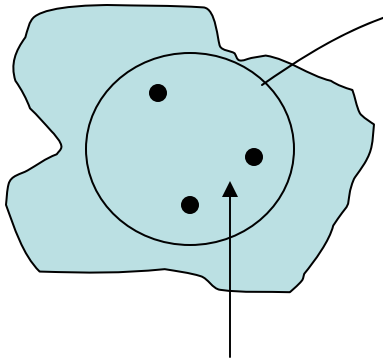


Feature Extraction

- 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

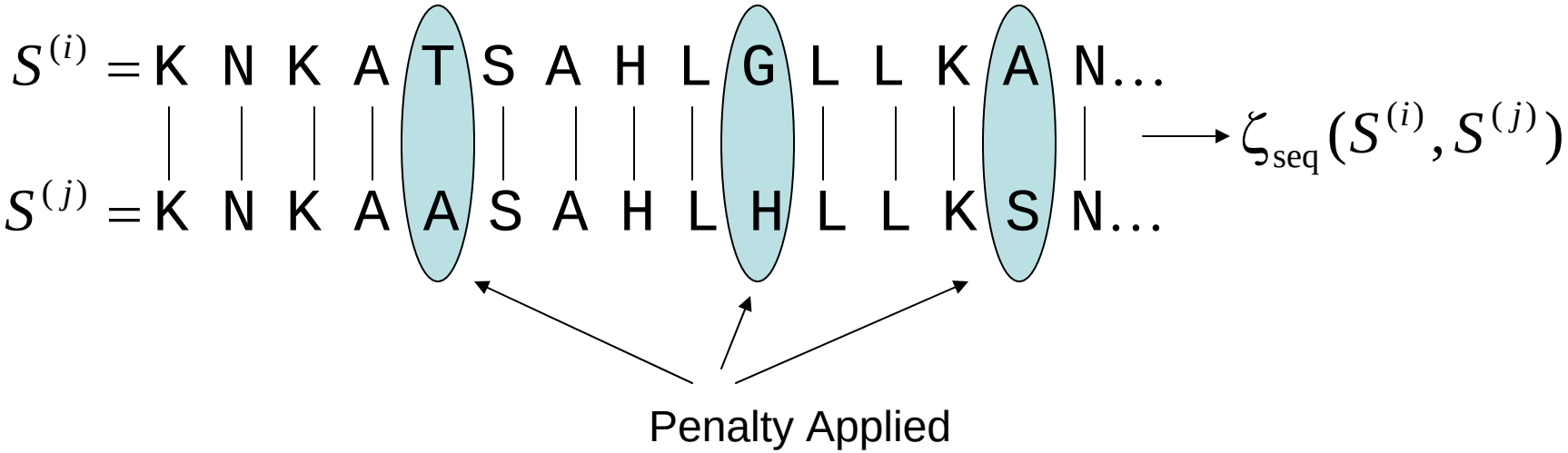
$S^{(1)} = \text{KNKATSAHLGLLKAN}\ldots$

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

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

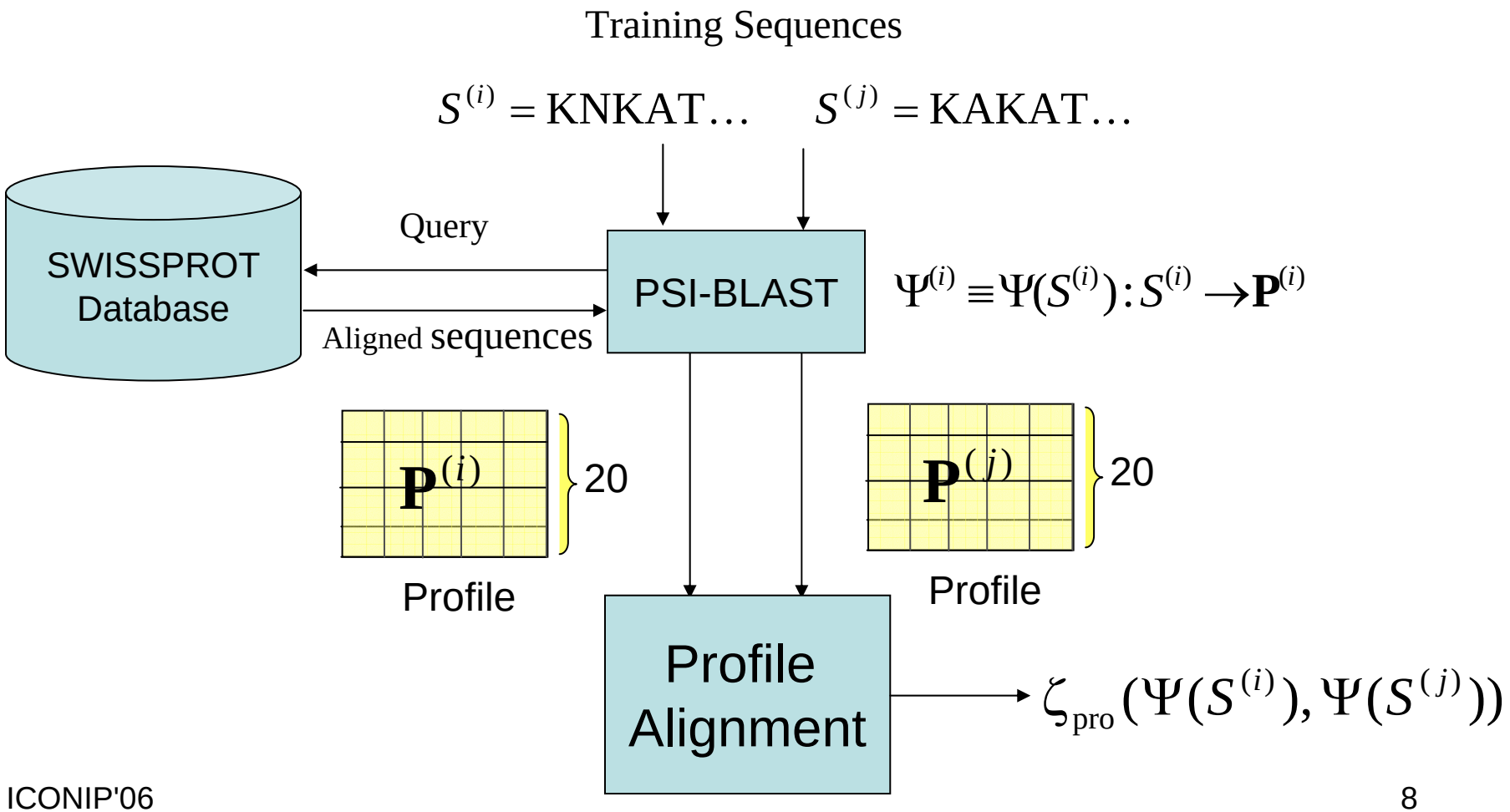
Feature Extraction by Sequence Alignment

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



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.



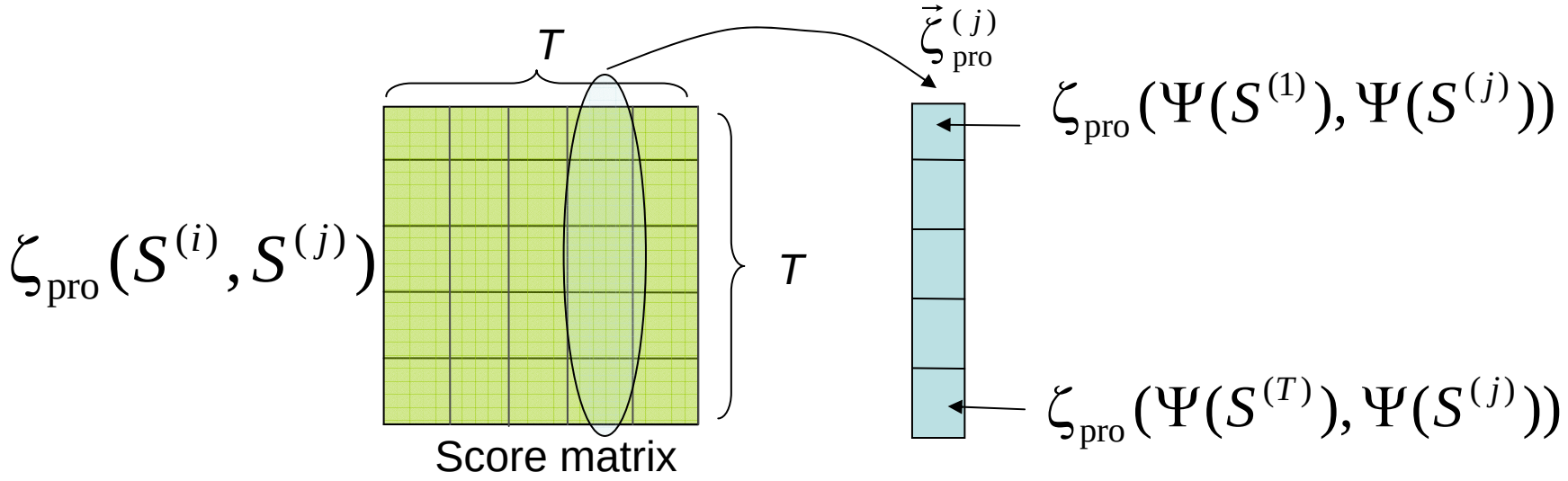
Feature Extraction by Profile Alignment

- Profile Alignment kernel:

$$K_{\text{pro}}(\Psi(S^{(i)}),\Psi(S^{(j)})) = <\vec{\zeta}_{\text{pro}}^{(i)},\vec{\zeta}_{\text{pro}}^{(j)}>$$

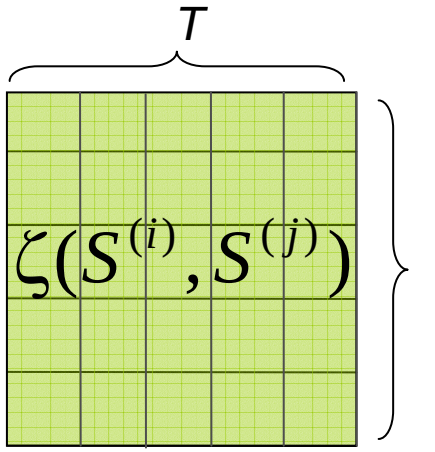
$$= \sum_{t=1}^{T} \zeta_{\text{pro}}(\Psi(S^{(i)}),\Psi(S^{(t)}))\zeta_{\text{pro}}(\Psi(S^{(t)}),\Psi(S^{(j)}))$$

T is the number of training sequences with known subcellular location



Training 1-vs-Rest SVM Classifier

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



Computing Kernel Matrix

$K(S^{(i)}, S^{(j)})$

Quadratic Programming

$\alpha_{c,i}, b_c$

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

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

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

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

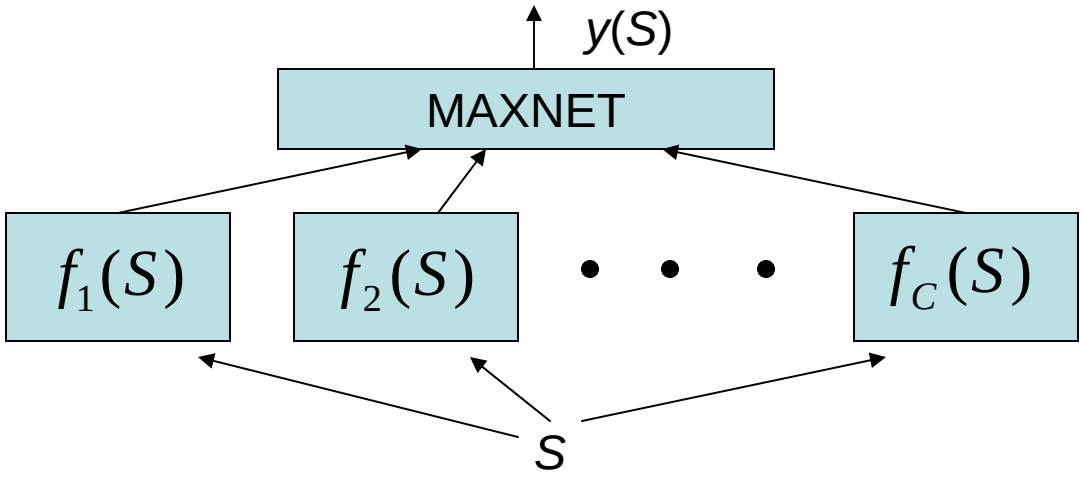
Classification by 1-vs-Rest SVM

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

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

- Prediction is based on

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



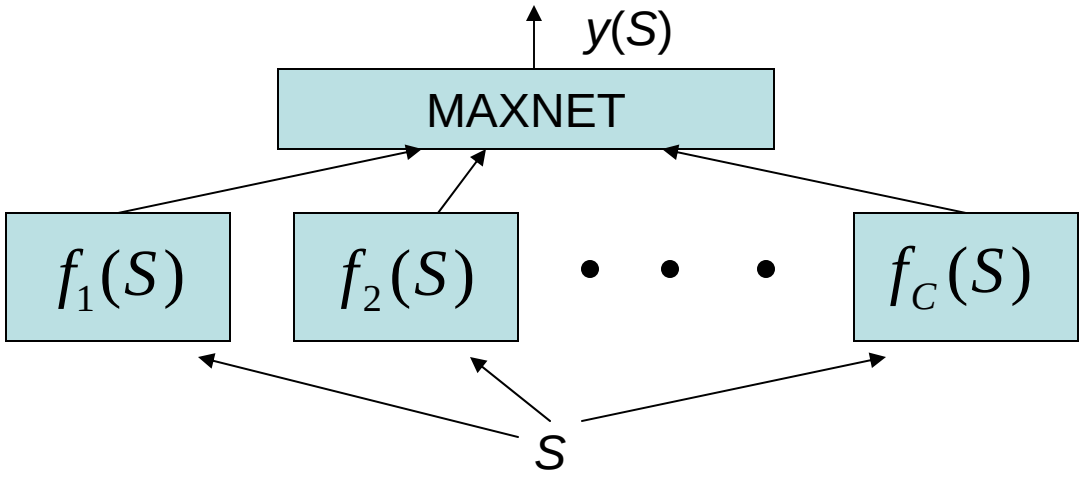
Classification by 1-vs-Rest SVM

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

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

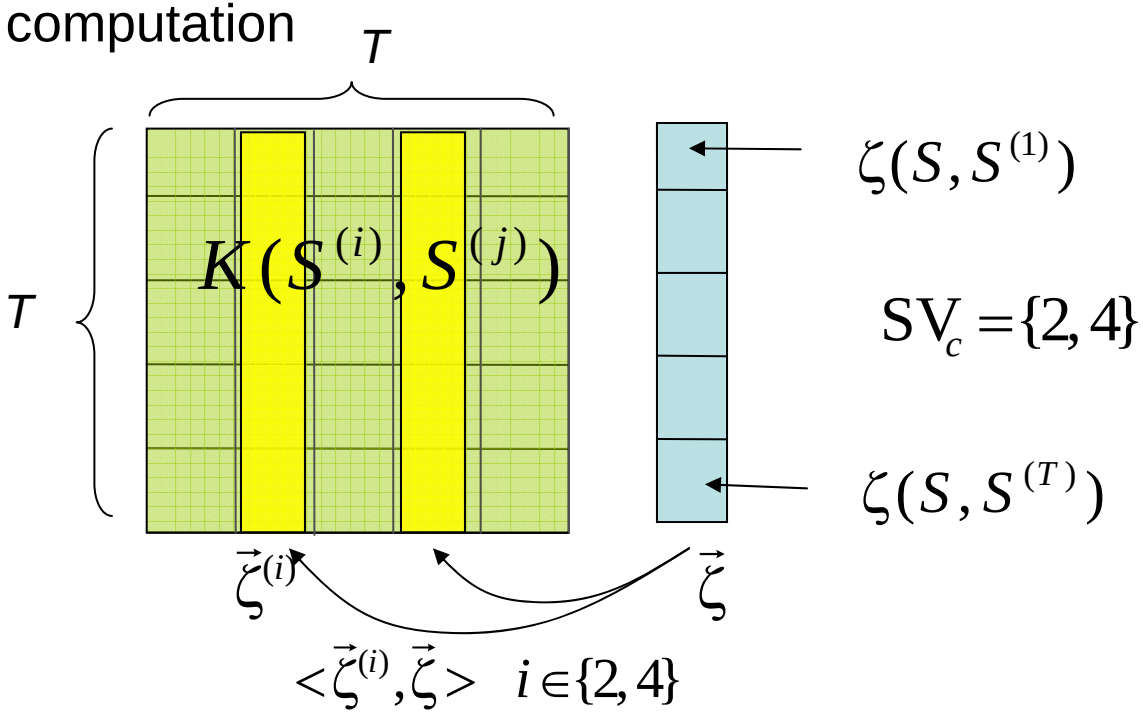
- Prediction is based on

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



Feature Selection

S needs to be aligned with **all** training sequences => Lots of computation

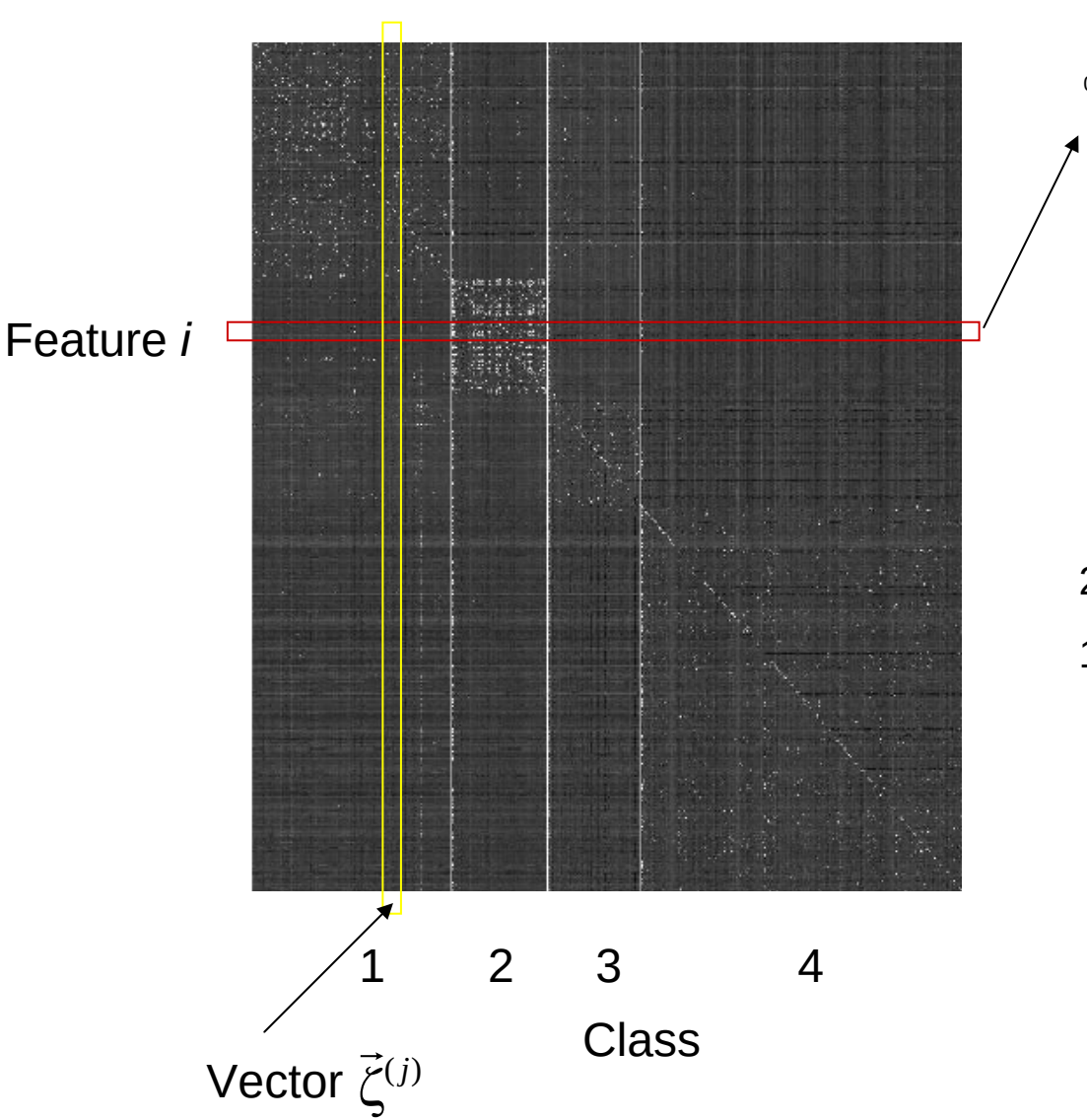


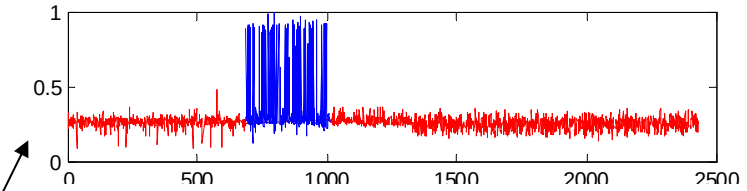
Major cause of computation burden

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

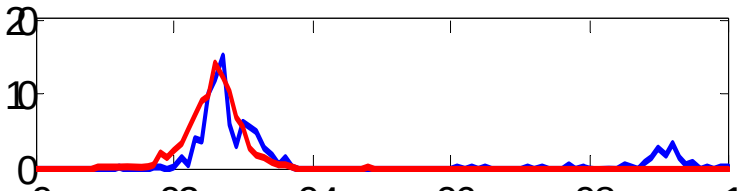
$$= \sum_{i \in SV_c} y_{c,i} \alpha_{c,i} < \vec{\zeta}^{(i)}, \vec{\zeta} > + b_c$$

Feature Selection





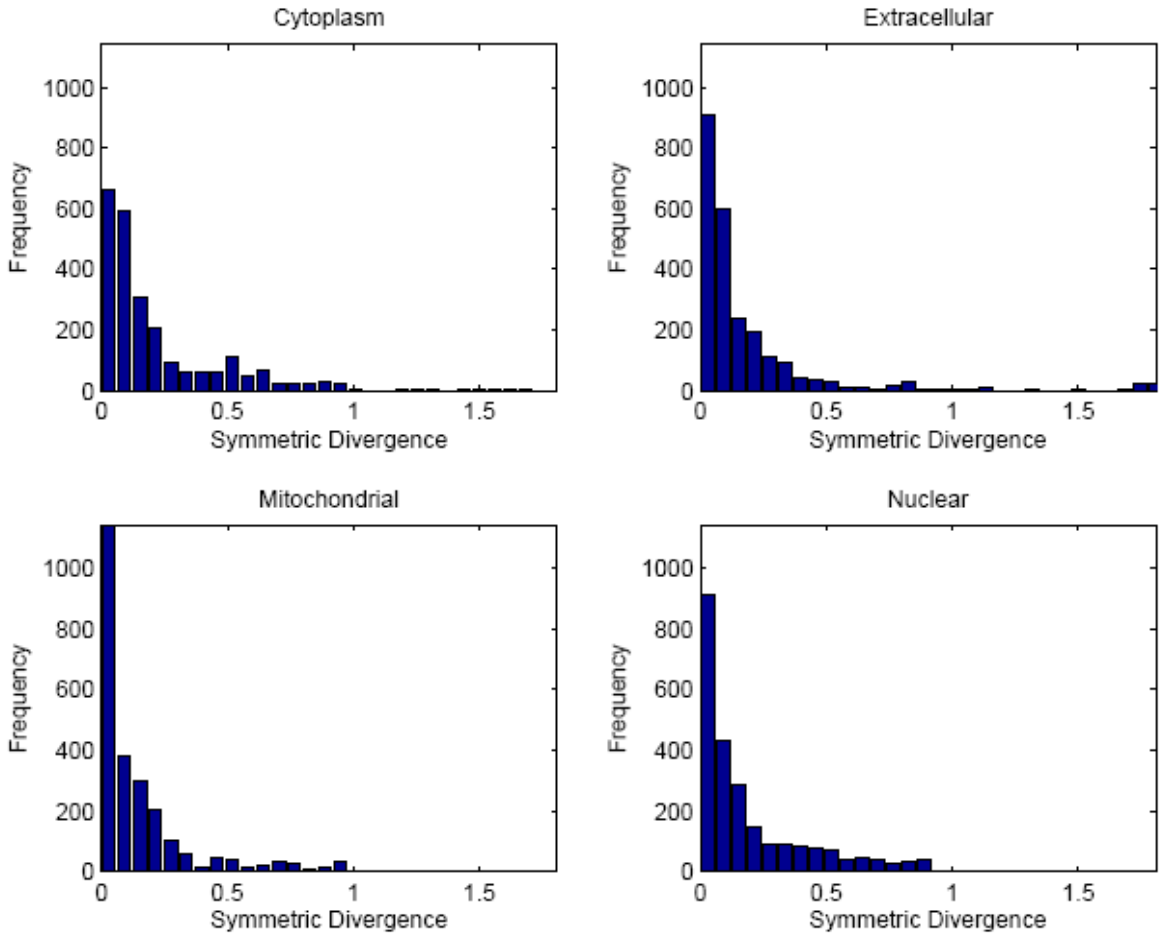
Compute Density Functions of Positive and Negative Classes



Compute Symmetric Divergence

$$D(\gamma_p^{(m)}, \gamma_n^{(m)})$$

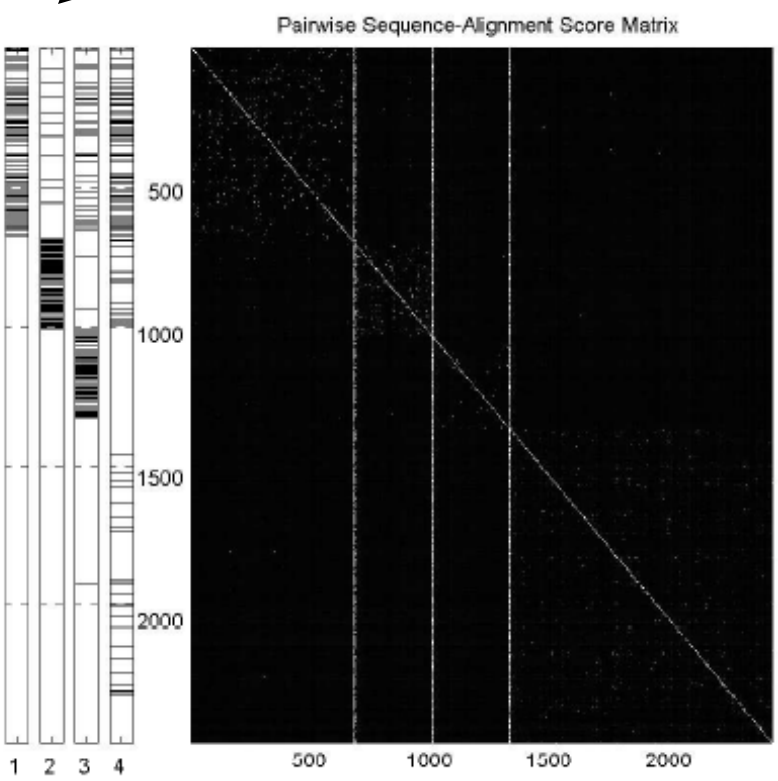
Feature Selection



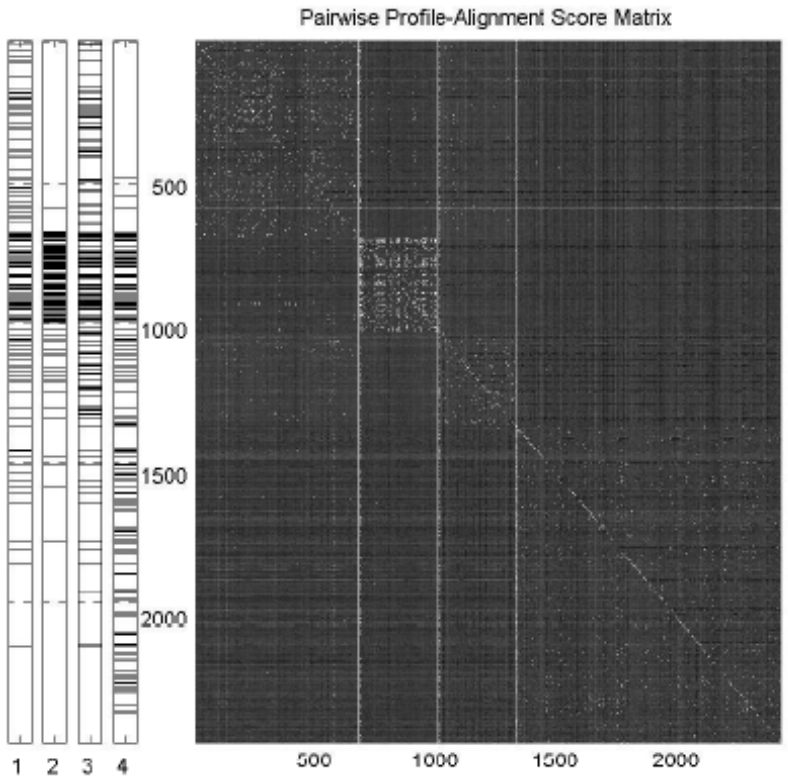
Histogram of $D(\gamma_p^{(m)}, \gamma_n^{(m)})$ for 4 classes

Feature Selection

$$\mathcal{M} = \left\{ m : D(\gamma_p^{(m)}, \gamma_n^{(m)}) > \mu_D + k\sigma_D \quad \forall\, 1 \leq m \leq T \right\}$$



Sequence

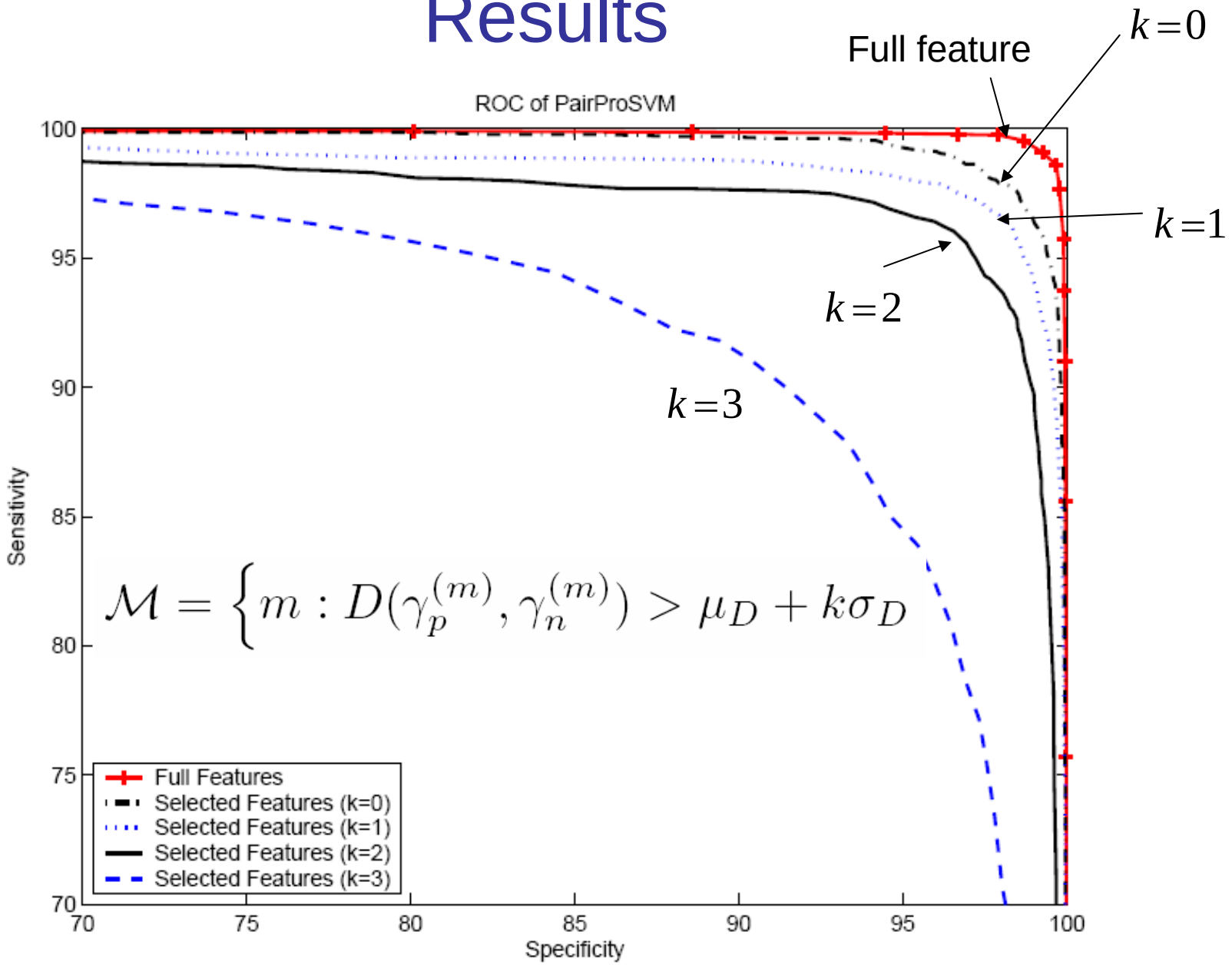


Profile

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.
- 5-Fold cross validation was used to obtain the accuracy.

Results


Full feature
$k=0$
$k=1$
$k=2$
$k=3$
ROC of PairProSVM
Sensitivity
Specificity
100
95
90
85
80
75
70
70
75
80
85
90
95
100
Full Features
Selected Features (k=0)
Selected Features (k=1)
Selected Features (k=2)
Selected Features (k=3)
$\mathcal{M} = \{m : D(\gamma_p^{(m)}, \gamma_n^{(m)}) > \mu_D + k\sigma_D$

Results

$$\mathcal{M} = \left\{ m : D(\gamma_p^{(m)}, \gamma_n^{(m)}) > \mu_D + k\sigma_D \quad \forall\, 1 \le m \le T \right\}$$

$|\mathcal{M}|$

Classifier	Feature	k	Feature Dimension	Accuracy (%)	Rec. Time (sec.)
Linear SVM	ζ'	N/A	1942	87.9	89.9
RBF-SVM	ϕ'	0	730	85.0	30.0
RBF-SVM	ϕ'	1	245	80.8	6.1
RBF-SVM	ϕ'	2	94	75.2	3.1
RBF-SVM	ϕ'	3	39	64.9	2.1
Linear SVM	ζ	N/A	1942	99.4	18.3
RBF-SVM	ϕ	0	595	97.4	21.3
RBF-SVM	ϕ	1	244	96.4	4.2
RBF-SVM	ϕ	2	110	95.5	1.4
RBF-SVM	ϕ	3	37	87.1	1.1

Results

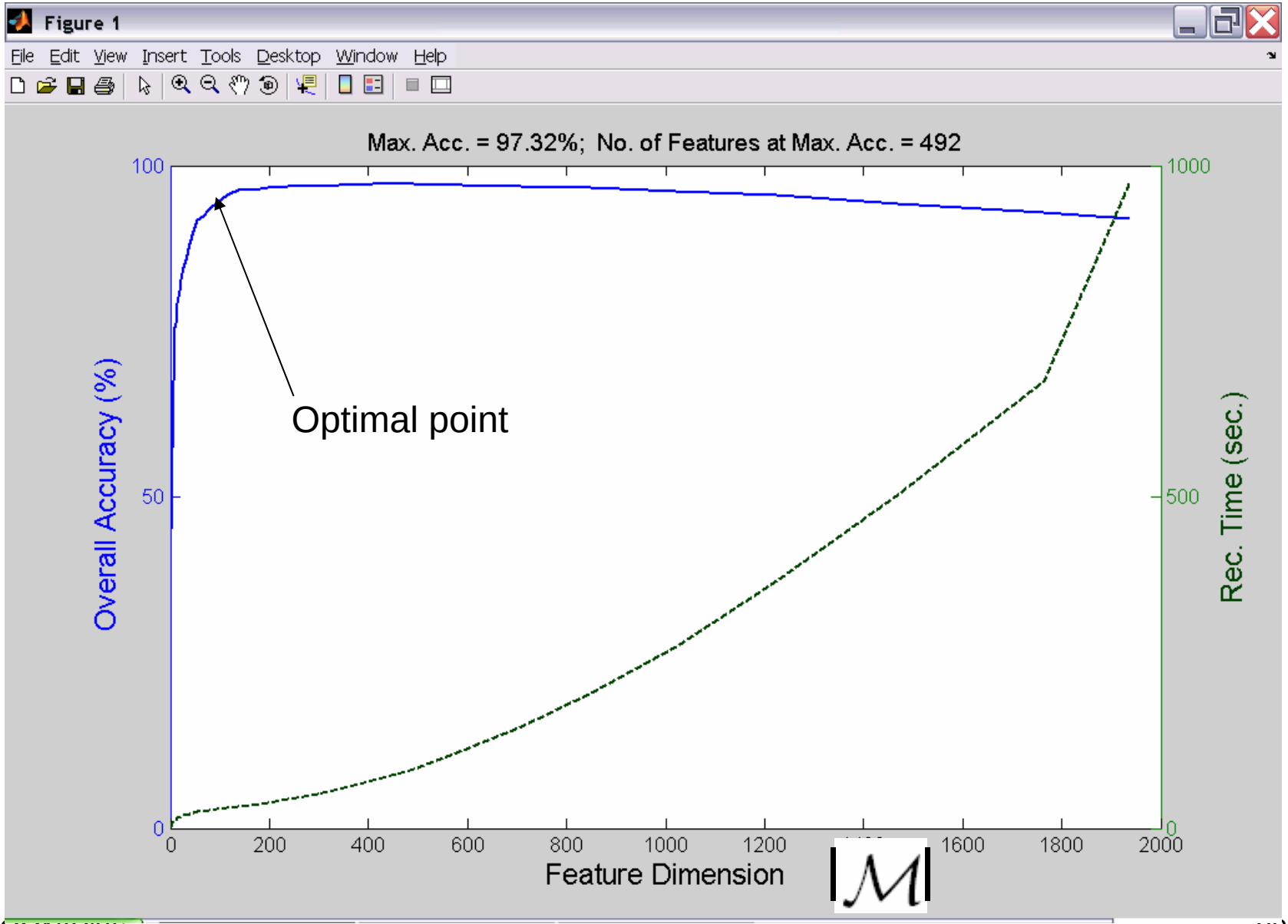


Figure 1
File Edit View Insert Tools Desktop Window Help
Max. Acc. = 97.32%; No. of Features at Max. Acc. = 492
Optimal point
Overall Accuracy (%)
Rec. Time (sec.)
Feature Dimension
100
50
0
1000
500
0
0
200
400
600
800
1000
1200
1600
1800
2000

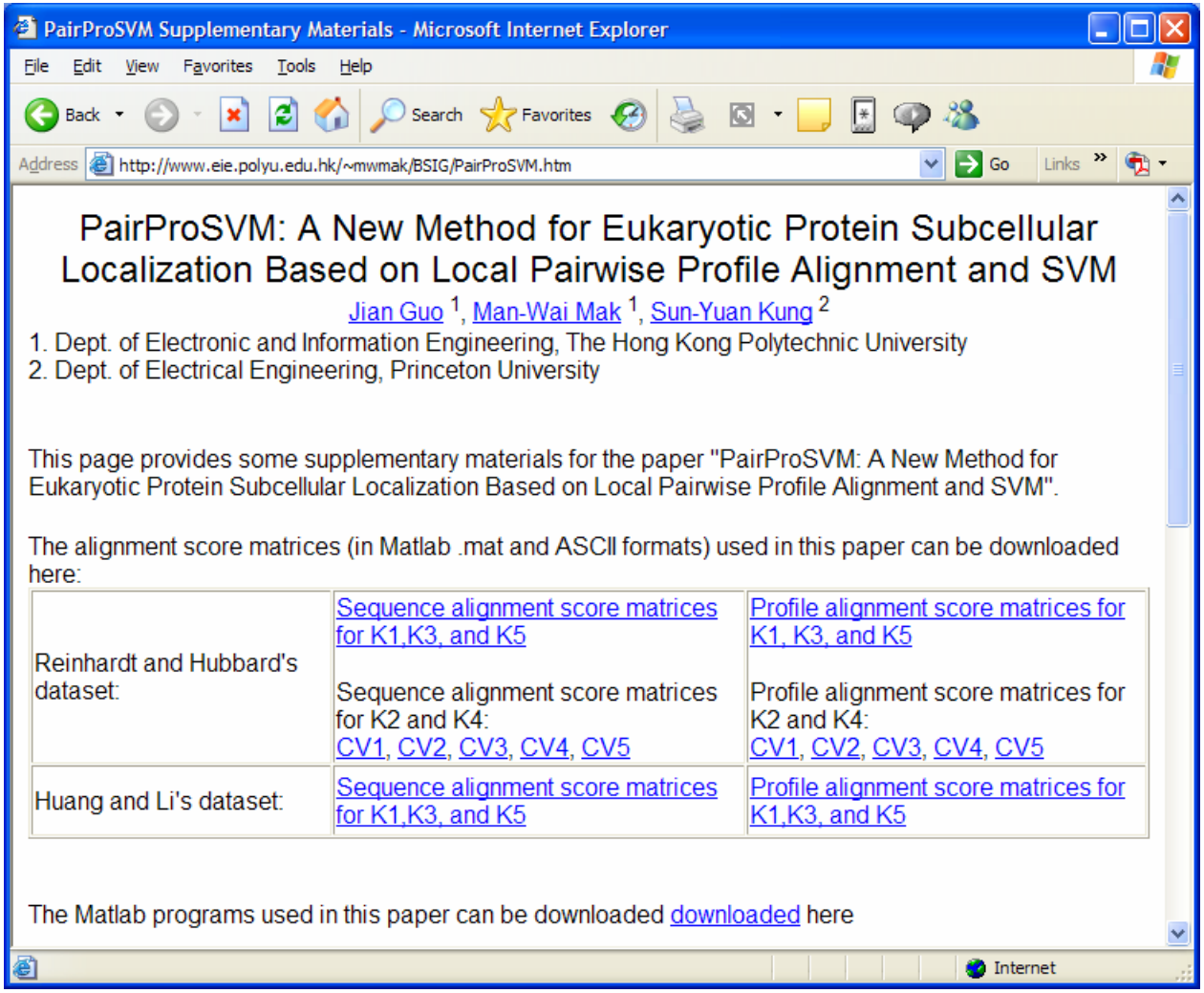
Conclusions

- Experimental evaluation on a benchmark protein sequence dataset shows that FDA-based selection schemes can reduce the feature dimension from thousands to hundreds, making subsequent classification much easier

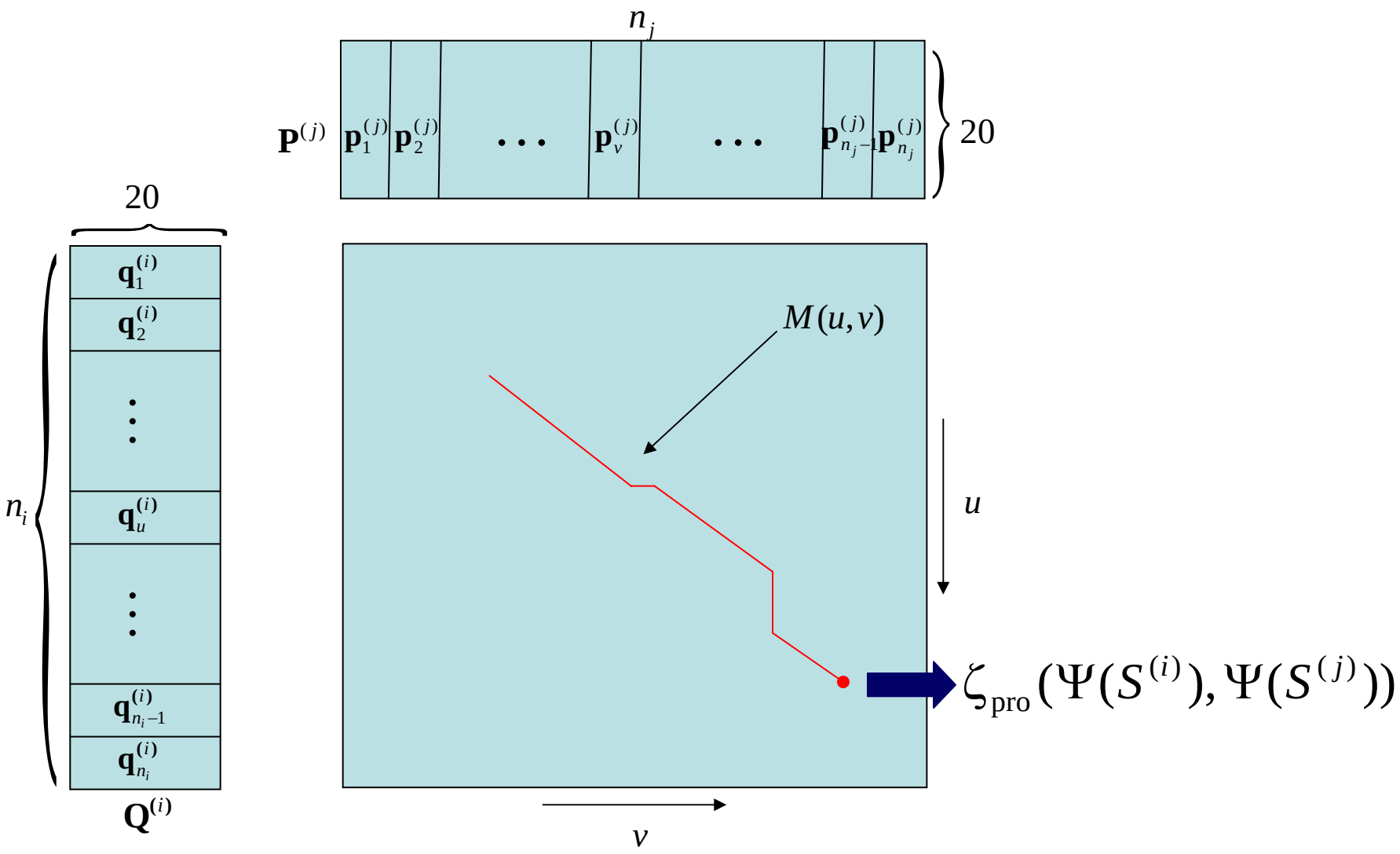
- With just a small reduction in recognition accuracy, a substantial speed up in recognition time can be achieved.

Further Information

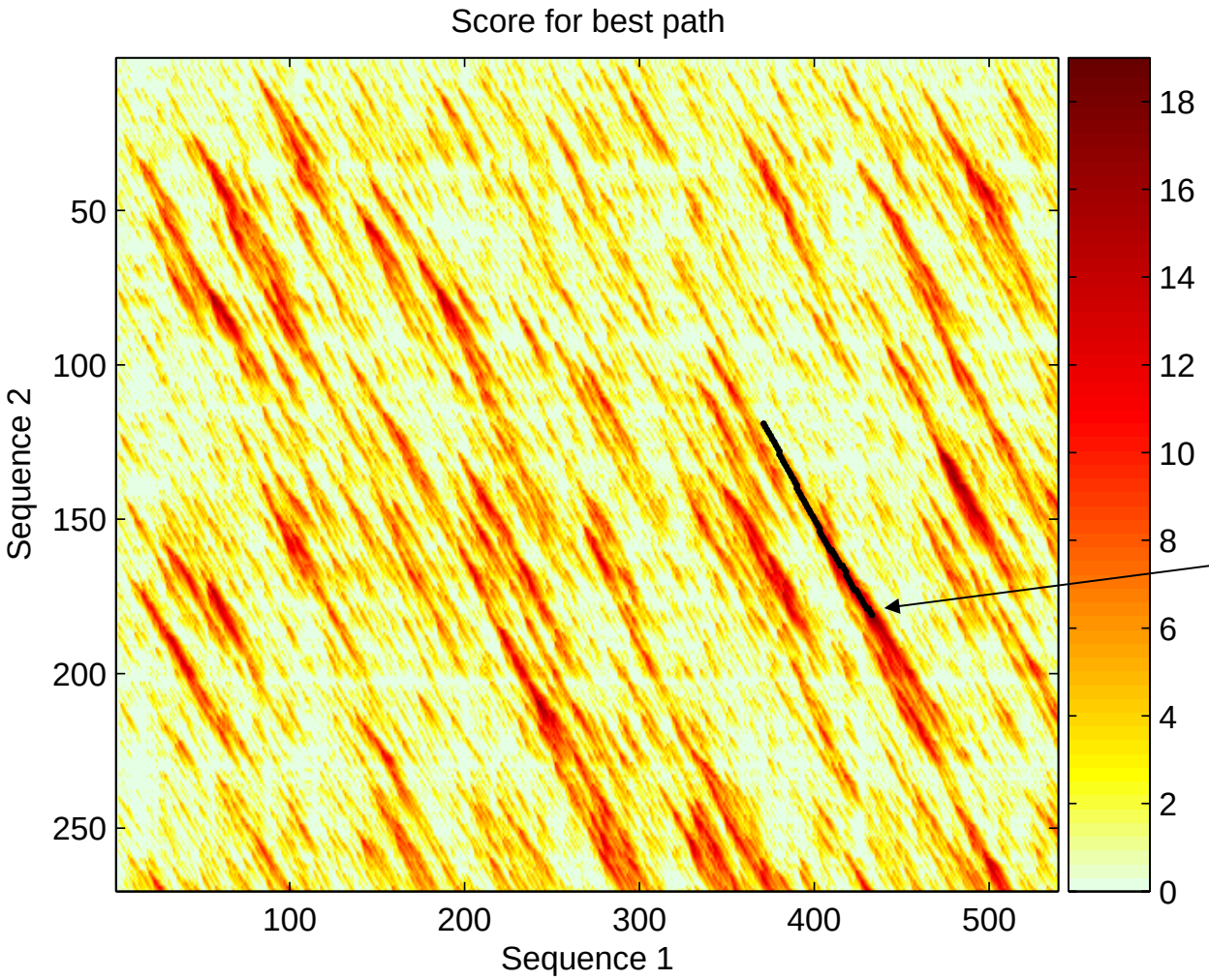
http://www.eie.polyu.edu.hk/~mwmak/BSIG/PairProSVM.htm



Feature Extraction by Profile Alignment



Feature Extraction by Profile Alignment



$$\zeta_{\text{pro}}(\Psi(S^{(i)}), \Psi(S^{(j)}))$$

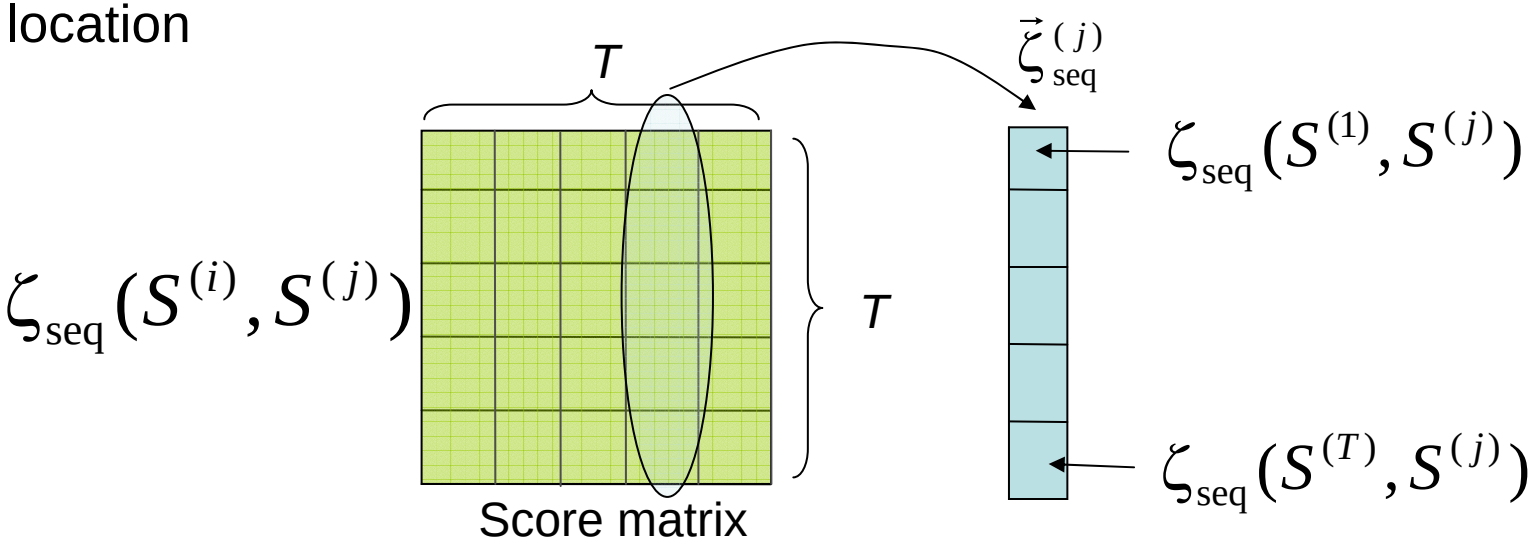
Feature Extraction by Sequence Alignment

- Sequence Alignment kernel:

$$K_{\text{seq}}(S^{(i)}, S^{(j)}) = < \vec{\zeta}_{\text{seq}}^{(i)}, \vec{\zeta}_{\text{seq}}^{(j)} >$$

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

T is the number of training sequences with known subcellular location



Classification by 1-vs-Rest SVM

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

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

- Prediction is based on

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