Eukaryotic Protein Subcellular Localization Based on Local Pairwise Profile Alignment SVM

Jian Guo 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
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 is often time-consuming and laborious.
- This motivates the prediction of subcellular locations through amino acid sequences.

```
MITILEKISAIESEMARTQ
KNKATSAHGLLKLANKA
VIERRELISPKGGGGGTG
EAGFEVAKTGDARVVF
VIEHLNDEDVVQIVKKV
```

Subcellular Location Predictor

Subcellular Location
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

\[ S^{(1)} = \text{KNKATSAHLLKAN...} \]
\[ S^{(2)} = \text{KAKATSLHLGLLKAN...} \]
\[ S^{(3)} = \text{KNKATSAHLALLKAN...} \]

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

Pairwise similarity scores
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 \( S^{(j)} \)

\[
S^{(i)} = K \quad N \quad K \quad A \quad T \quad S \quad A \quad H \quad L \quad G \quad L \quad L \quad K \quad A \quad N \ldots
\]
\[
S^{(j)} = K \quad N \quad K \quad A \quad A \quad S \quad A \quad H \quad L \quad H \quad L \quad L \quad K \quad S \quad N \ldots
\]

Penalty Applied
Feature Extraction by Sequence Alignment

- Five possible kernels:

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

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

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

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

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

\[\text{Dot product: } \langle \bar{\zeta}^{(i)}, \bar{\zeta}^{(j)} \rangle\]

\[T \text{ is the number of training sequences with known subcellular location}\]
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 (comparing matrices).

\[
S^{(i)} = \text{KNKAT} \ldots \quad S^{(j)} = \text{KAKAT} \ldots
\]

\[
\phi(\mathcal{S}): \mathcal{S} \rightarrow \{P, Q\}
\]

\[
\zeta(\phi(S^{(i)}), \phi(S^{(j)}))
\]
Feature Extraction by Profile Alignment

- PSSM (Position-Specific Scoring Matrix):
  - The \((i,j)\)-th entry represents the likelihood score of amino acid in the \(j\)-th position of the query sequence being mutated to amino acid type \(i\) during the evolution process.

\[
\begin{align*}
\text{Score}(V \rightarrow H \mid \text{pos} = 1) &= -3 \\
\text{Score}(V \rightarrow H \mid \text{pos} = 8) &= -4
\end{align*}
\]

20 Amino Acid

| Position | 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| 1 | 5 | -3| 2 | 0 | -3| -3| -2| -2| -1| 1 |
| 3        | I | -2| -3| -4| -4| -2| -3| -4| -4| -4| 5 | 2 | 7 | 1 | 0 | 7 | -3| -1| -3| -2| 2 |
| 4        | K | -1| 2 | 0 | -1| -4| 1 | 1 | -2| -1| -3| -3| 0 | -1| -3| -2| -3| 0 | -1| -3| -2| -3 |
| 5        | E | -1| 0 | -1| 1 | -4| 2 | 6 | -3| 0 | -4| -3| 0 | -1| -3| -2| -3| 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| -2| -1| -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 |

Different scores
Feature Extraction by Profile Alignment

- PSFM (Position-Specific Frequency Matrix):
  - The \((i,j)\)-th entry represents the chance of having amino acid type \(i\) in position \(j\) of the query sequence.

```
A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V
0  0  0  0  0  0  0  0  0  0  100 0  36 0  0  0  0  0  0  64
0  0  0  0  0  0  0  0  0  0  77 23 0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  100 0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0 100 0  0  0  0  0  0  0
0 100 0  0  0  0  0  0  0  0  0  0  0  36 0  0  0  0  0  0  64
0  0  0  0  0  0  0  0  0  0  23 100 0  0  0  0  0  0  0  41
0  0  0  0  0  0  0  0  0  0  0  0  36 0  0  0  0  0  0  64
0  0  0  0  0  0  0  0  0  0  0  0  100 0  0  0  0  0  0  41
0  0  0  0  0  0  0  0  0  0  0  0  0  36 0  0  0  0  0  0  23
0  0  0  0  0  0  0  0  0  0  0  0  0  64 36 0  0  0  0  100
0  0  0  0  0  0  0  0  0  0  0  0  0  36 0  0  0  0  0  100
```

\(P(\text{AA} = 'H'|\text{pos} = 1) = 0.36\)
Feature Extraction by Profile Alignment

\[
M(u, v) \ni \zeta(\phi(S^{(i)}), \phi(S^{(j)}))
\]
Feature Extraction by Profile Alignment
Profile Alignment Kernels

Denote the operation of PSI-BLAST search as

\[ \phi^{(i)} \equiv \phi(S^{(i)}) \rightarrow \{ P^{(i)}, Q^{(i)} \} \]

PSSM  PSFM

The 5 profile alignment kernels are defined as

\[ K_1^{pro} (\phi(S^{(i)}), \phi(S^{(j)})) = \zeta(\phi^{(i)}, \phi^{(j)}) \]

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

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

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

\[ K_5^{pro} (\phi(S^{(i)}), \phi(S^{(j)})) = \sum_{i=1}^{T} \zeta(\phi^{(i)}, \phi^{(t)}) \zeta(\phi^{(j)}, \phi^{(t)}) \]
# Sequence Kernels Vs. Profile Kernels

<table>
<thead>
<tr>
<th>Sequence Kernels</th>
<th>Profile Kernels</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K_{seq}^1(S^{(i)}, S^{(j)}) = \zeta(S^{(i)}, S^{(j)})$</td>
<td>$K_{pro}^1(\phi(S^{(i)}), \phi(S^{(j)})) = \zeta(\phi^{(i)}, \phi^{(j)})$</td>
</tr>
<tr>
<td>$K_{seq}^2(S^{(i)}, S^{(j)}) = \max_{1 \leq l \leq T} \zeta(S^{(i)}, S^{(l)}) \zeta(S^{(j)}, S^{(l)})$</td>
<td>$K_{pro}^2(\phi(S^{(i)}), \phi(S^{(j)})) = \max_{1 \leq l \leq T} \zeta(\phi^{(i)}, \phi^{(l)}) \zeta(\phi^{(j)}, \phi^{(l)})$</td>
</tr>
<tr>
<td>$K_{seq}^3(S^{(i)}, S^{(j)}) = \left(\zeta(S^{(i)}, S^{(j)}) + 1\right)^d$</td>
<td>$K_{pro}^3(\phi(S^{(i)}), \phi(S^{(j)})) = \left(\max_{1 \leq l \leq T} \zeta(\phi^{(i)}, \phi^{(l)}) \zeta(\phi^{(j)}, \phi^{(l)}) + 1\right)^d$</td>
</tr>
<tr>
<td>$K_{seq}^4(S^{(i)}, S^{(j)}) = \sum_{t=1}^{T} \zeta(S^{(i)}, S^{(t)}) \zeta(S^{(j)}, S^{(t)})$</td>
<td>$K_{pro}^4(\phi(S^{(i)}), \phi(S^{(j)})) = \sum_{t=1}^{T} \zeta(\phi^{(i)}, \phi^{(t)}) \zeta(\phi^{(j)}, \phi^{(t)})$</td>
</tr>
</tbody>
</table>
Training 1-vs-Rest SVM Classifier

\[ D = \{S^{(1)}, S^{(2)}, \ldots, S^{(T)}\} \rightarrow \text{Pairwise Sequence Alignment} \rightarrow \zeta(S^{(i)}, S^{(j)}) \]

\[
\max_{\alpha_c} \sum_i \alpha_{c,i} - \sum_i \sum_j y_{c,i} y_{c,j} \alpha_{c,i} \alpha_{c,j} K_{seq}(S^{(i)}, S^{(j)})
\]

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

\[ y_{c,i} \alpha_{c,i}, b_c \]

\[ i \in SV_c, c = 1, \ldots, C \]

\[ \alpha_{c,i}, b_c \]
Classification by 1-vs-Rest SVM

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

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

• 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

$$f_c(S) = \sum_{i \in SV_c} y_{c,i} \alpha_{c,i} K^\text{pro}_k (\phi(S^{(i)}), \phi(S)) + b_c$$

• Prediction is based on

$$y(S) = \arg \max_{c=1}^C f_c(S)$$
Prediction by Profile Alignment SVM

\[ f_c(S) = \sum_{i \in S_{V_c}} y_{c,i} \alpha_{c,i} K^\text{pro}_k (\phi(S^{(i)}), \phi(S)) + b_c \]

\[ y(S) = \arg \max_{c=1}^C f_c(S) \]
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.

\[
K_{1}^{\text{seq}}(S^{(i)},S^{(j)}) = \zeta(S^{(i)},S^{(j)}) \\
K_{2}^{\text{seq}}(S^{(i)},S^{(j)}) = \max_{1 \leq l \leq T} \zeta(S^{(i)},S^{(l)}) \zeta(S^{(j)},S^{(l)}) \\
K_{3}^{\text{seq}}(S^{(i)},S^{(j)}) = \left(\zeta(S^{(i)},S^{(j)}) + 1\right)^d \\
K_{4}^{\text{seq}}(S^{(i)},S^{(j)}) = \left(\max_{1 \leq l \leq T} \zeta(S^{(i)},S^{(l)}) \zeta(S^{(j)},S^{(l)}) + 1\right)^d
\]

\[
f_{c}(S) = \sum_{i \in SV_c} y_{c,i} \alpha_{c,i} K_{k}^{\text{seq}}(S^{(i)},S) + b_{c}, \quad k = 1, \ldots, 4
\]
Complexity Analysis

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

For $S$, needs to be aligned with all training sequences $seq_5$.

For $S$, needs to be aligned with all training sequences $seq_5$.

For $S$, needs to be aligned with all training sequences $seq_5$.

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

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

• One possible solution to reduce the complexity of K5 is select the relevant features (rows) from the kernel matrix.


Session: TB408
Thur. 13:30 – 15:10
Theoretical Modeling and Analysis II
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

<table>
<thead>
<tr>
<th></th>
<th>Sequence Alignment Kernel</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$K_1^{\text{seq}}$</td>
</tr>
<tr>
<td><strong>Accuracy</strong></td>
<td>87.0%</td>
</tr>
<tr>
<td>% of negative eigenvalues in $K^{\text{seq}}$</td>
<td>0</td>
</tr>
<tr>
<td>Meeting Mercer’s condition</td>
<td>Yes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Profile Alignment Kernel</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$K_1^{\text{pro}}$</td>
</tr>
<tr>
<td><strong>Accuracy</strong></td>
<td>45.9%</td>
</tr>
<tr>
<td>% of negative eigenvalues in $K^{\text{pro}}$</td>
<td>8.5</td>
</tr>
<tr>
<td>Meeting Mercer’s condition</td>
<td>No</td>
</tr>
</tbody>
</table>

### 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

Each graph represents the Receiver Operating Characteristic (ROC) curves for different subcellular locations: Cytoplasm, Extracellular, Mitochondria, and Nuclear. The curves indicate the trade-off between sensitivity and specificity for various kinetic constants, denoted as $K_1^{\text{pro}}$, $K_2^{\text{pro}}$, $K_3^{\text{pro}}$, $K_4^{\text{pro}}$, and $K_5^{\text{pro}}$. These constants are used to model the protein interactions within the cellular compartments.
Results: Comparison with Existing Methods

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

<table>
<thead>
<tr>
<th>Subcellular Location</th>
<th>ESLpred Acc(%)</th>
<th>PairSeqSVM $K_{5}^{seq}$ Acc(%)</th>
<th>MCC</th>
<th>PairProSVM $K_{5}^{pro}$ Acc(%)</th>
<th>MCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cytoplasm</td>
<td>85.2</td>
<td>83.2</td>
<td>0.78</td>
<td>100.0</td>
<td>1.00</td>
</tr>
<tr>
<td>Extracellular</td>
<td>88.9</td>
<td>84.3</td>
<td>0.89</td>
<td>98.2</td>
<td>0.97</td>
</tr>
<tr>
<td>Mitochondria</td>
<td>68.2</td>
<td>61.7</td>
<td>0.73</td>
<td>96.9</td>
<td>0.97</td>
</tr>
<tr>
<td>Nuclear</td>
<td>95.3</td>
<td>97.8</td>
<td>0.83</td>
<td>99.5</td>
<td>1.00</td>
</tr>
<tr>
<td><strong>Overall</strong></td>
<td><strong>88.0</strong></td>
<td><strong>87.1</strong></td>
<td>0.83</td>
<td><strong>99.1</strong></td>
<td>0.99</td>
</tr>
<tr>
<td>Weighted Average</td>
<td>–</td>
<td>0.83</td>
<td>–</td>
<td>0.81</td>
<td>–</td>
</tr>
</tbody>
</table>
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