Protein Subcellular Localization Prediction Based on Support Vector Machines

Chia-Yu Su
Institute of Information Science
Academia Sinica, Taiwan
Outline

■ Introduction

■ SVM models
  ◆ 1-v-r based on general features
  ◆ 1-v-1 based on general features

■ Biological features
  ◆ 1-v-1 based on more specific biological features
  ◆ 1-v-1 based on a new encoding for protein structures

■ Conclusion
Protein Subcellular Localization (PSL) Prediction

- Predict where the protein is located in a cell?
  - C1: cytoplasm
  - C2: inner membrane
  - C3: periplasm
  - C4: outer membrane
  - C5: extracellular

Gram-Negative Bacteria
Importance of PSL Prediction

- **Protein function identification**
  - Modulate and identify protein functions

- **Genome annotation**
  - Annotate genomic features

- **Drug discovery**
  - Give clues to new drug targets
# Current PSL Prediction for Gram-Negative Bacteria

<table>
<thead>
<tr>
<th>Systems</th>
<th>Approaches</th>
<th>Features</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSORTb</td>
<td>Bayesian Network</td>
<td>5 analytical modules</td>
<td>74.8%</td>
</tr>
<tr>
<td>P-CLASSIFIER</td>
<td>SVM</td>
<td>Amino acid subalphabets</td>
<td>89.8%</td>
</tr>
<tr>
<td>CELLO II</td>
<td>SVM</td>
<td>$n$-peptide compositions</td>
<td>90.0%</td>
</tr>
<tr>
<td>PSL101</td>
<td>SVM</td>
<td>Specific biological features</td>
<td>92.7%</td>
</tr>
</tbody>
</table>

The state-of-the-art system is PSL101 with an accuracy of 92.7%.
Outline

- Introduction
- SVM models
  - 1-v-r based on general features
  - 1-v-1 based on general features
- Biological features
  - 1-v-1 based on more specific biological features
  - 1-v-1 based on a new encoding for protein structures
- Conclusion
Multiclass Classification in SVM

- **One-versus-rest (1-v-r) SVM model**
  - Apply a universal set of biological features for different localization classes

- **One-versus-one (1-v-1) SVM model**
  - Different biological features can be used in distinguishing two classes
Outline

- Introduction
- SVM models
  - 1-v-r based on general features
  - 1-v-1 based on general features
- Biological features
  - 1-v-1 based on more specific biological features
  - 1-v-1 based on a new encoding for protein structures
- Conclusion
Multiclass Classification by 1-v-r SVM

- Binary classifiers: for each class $i$, construct a $C_i$ vs. non-$C_i$ binary classifier
  - # of classifiers = 5
- Input features: same features for all binary classifiers
- Class determination:
  - The class with the largest probability ($prob_i$: the confidence of sample predicted as class $i$; $0 \leq prob_i \leq 1$) is chosen as final predicted class

1-v-r SVM Model

- C1: cytoplasm
- C2: inner membrane
- C3: periplasm
- C4: outer membrane
- C5: extracellular
General Biological Features for PSL Prediction

1. Amino acid composition (AA)
2. Dipeptide composition (Dip)
3. Secondary structure elements (SSE)
1. Amino Acid Composition
2. Dipeptide Composition

- **Amino acid composition (AA)** and
- **Dipeptide composition (Dip)**

  - *n*-peptide compositions or their variations have been shown effective in PSL prediction
    - If $n = 1$, then the $n$-peptide composition reduces to the AA
      - Dimension = 20
    - If $n = 2$, then the $n$-peptide composition yields the Dip
      - Dimension = 20*20

N-terminus: Met Phe Leu Ser Val Arg Ala His Phe Ala

C-terminus: $\cdots$
3. Secondary Structure Elements

- Predicted secondary structure elements (SSE) from HYPROSP II server

  - Encoding scheme: compute amino acid compositions of $\alpha$-helix (H), $\beta$-strand (E), and random coil (C)

| Protein Seq. | M P L D L Y N T L T R R K E R F E P M T P D · · |
| Predicted SSEs | C C E E E E C C H H H H H H C C E E E H H · · |

<table>
<thead>
<tr>
<th>$A$</th>
<th>$C$</th>
<th>$D$</th>
<th>...</th>
<th>$Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>↓</td>
<td>↓</td>
<td>↓</td>
<td>...</td>
<td>↓</td>
</tr>
</tbody>
</table>

\[
\begin{align*}
0.12 & 0.02 & 0.03 & \ldots & 0.05 & \leftarrow H \\
0.07 & 0.04 & 0.02 & \ldots & 0.09 & \leftarrow E \\
0.08 & 0.06 & 0.07 & \ldots & 0.03 & \leftarrow C
\end{align*}
\]

$\alpha$-helix \hspace{1cm} $\beta$-strand \hspace{1cm} random coil
Training and Testing in SVM

- Support Vector Machines (SVM)
  - LIBSVM software
  - Kernel: Radial Basis Function (RBF)
  - Parameter selection
    - \( c \) (cost) and \( \gamma \) (gamma) are optimized
  - 10-fold cross-validation
# Gram-Negative Bacteria Data Set

<table>
<thead>
<tr>
<th>Localization sites</th>
<th>No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cytoplasmic (CP)</td>
<td>248</td>
</tr>
<tr>
<td>Inner membrane (IM)</td>
<td>268</td>
</tr>
<tr>
<td>Periplasmic (PP)</td>
<td>244</td>
</tr>
<tr>
<td>Outer membrane (OM)</td>
<td>352</td>
</tr>
<tr>
<td>Extracellular (EC)</td>
<td>190</td>
</tr>
<tr>
<td>All sites</td>
<td>1,302</td>
</tr>
</tbody>
</table>
Performance Evaluation

- **Accuracy** (*Acc*)

\[
Acc_i = \frac{TP_i}{N_i}
\]

\[
Acc = \frac{\sum_{i=1}^{l} TP_i}{\sum_{i=1}^{l} N_i}
\]

- *l = 5* is the number of total localization sites
- *N_i* are the number of proteins in localization site *i*
Results of 1-v-r SVM Model

- Different feature combinations in 1-v-r SVM model
  1. AA
  2. Dip
  3. SSE
  4. AA+Dip
  5. AA+SSE
  6. Dip+SSE
  7. AA+Dip+SSE

<table>
<thead>
<tr>
<th>Feature</th>
<th>AA</th>
<th>Dip</th>
<th>SSE</th>
<th>AA+Dip</th>
<th>AA+SSE</th>
<th>Dip+SSE</th>
<th>AA+Dip +SSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Acc</td>
<td>85.56%</td>
<td>84.87%</td>
<td>83.26%</td>
<td>87.71%</td>
<td>84.95%</td>
<td>83.95%</td>
<td>86.25%</td>
</tr>
</tbody>
</table>

Acc of CELLO II = 90.0%
Outline

- Introduction
- SVM models
  - 1-v-r based on general features
  - 1-v-1 based on general features
- Biological features
  - 1-v-1 based on more specific biological features
  - 1-v-1 based on a new encoding for protein structures
- Conclusion
Multiclass Classification by 1-v-1 SVM

- Binary classifiers: for each pair of classes \textit{i} and \textit{j}, construct a \textit{C}_i vs. \textit{C}_j binary classifier
  - # of classifiers = 5*(5-1)/2 = 10
- Input features: different features can be in different classifiers
- Class determination:
  - Majority votes
  - Average probability
  - In case of a tie in majority votes, the class with the largest average probability is selected as final predicted class

1-v-1 SVM Model

10 classifiers
Accuracy and Feature Combination

- Features used in 1-v-1 SVM:
  - AA, Dip, and SSE

- Use 1-v-1 SVM model in our system, PSL101 (Protein Subcellular Localization prediction by 1-On-1 classifiers)
  - Flexibility of combining different features
  - Better accuracy

\[
\text{Acc of CELLO II} = 90.0\%
\]
Outline

- Introduction
- SVM models
  - 1-v-r based on general features
  - 1-v-1 based on general features
- Biological features
  - 1-v-1 based on more specific biological features
  - 1-v-1 based on a new encoding for protein structures
- Conclusion
Compartment-Specific Biological Features

- Integrate more biological features to improve accuracy
  - For each binary classifier $C_{ij}$, a set of compartment-specific features is incorporated
    - Features unique to $C_i$ or $C_j$
- Select features to mimic protein bacterial secretory pathways
  - Feature selection guided by biological insights
  - A binary classifier $C_{ij}$ distinguishes proteins localized in two different compartments
    - $C1$ vs. $C2$, $C2$ vs. $C3$, etc.
Compartment-Specific Features in Bacterial Secretory Pathways

Figure adapted from Wickner W. and R. Schekman, Protein Translocation Across Biological Membranes, Science 2005
More Compartment-Specific Biological Features

1. Amino acid composition (AA)
2. Dipeptide composition (Dip)
3. Secondary structure elements (SSE)
4. Solvent accessibility (SA) – C5
5. Signal peptides (Sig) – C1
6. Transmembrane $\alpha$-helices (TMA) – C2
7. Transmembrane $\beta$-barrels (TMB) – C4
8. Twin-arginine translocase signal peptides (TAT) – C3
9. Non-classical protein secretion (Sec) – C5
Implication of different biological features to localization classes

<table>
<thead>
<tr>
<th>Fea.</th>
<th>Description</th>
<th>Features $\rightarrow$ Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>SA</td>
<td>Solvent accessibility</td>
<td>Acidic high SA residues $\rightarrow$ C5</td>
</tr>
<tr>
<td>Sig</td>
<td>Signal peptides</td>
<td>Presence of Sig $\rightarrow$ not C1</td>
</tr>
<tr>
<td>TMA</td>
<td>Transmembrane $\alpha$-helices</td>
<td>Presence of TMA $\rightarrow$ C2</td>
</tr>
<tr>
<td>TAT</td>
<td>Twin-arg translocase motifs</td>
<td>Presence of TAT $\rightarrow$ C3</td>
</tr>
<tr>
<td>TMB</td>
<td>Transmembrane $\beta$-barrels</td>
<td>Presence of TMB $\rightarrow$ C4</td>
</tr>
<tr>
<td>Sec</td>
<td>Non-classical protein secretion</td>
<td>Presence of Sec $\rightarrow$ C5</td>
</tr>
</tbody>
</table>
System Architecture of PSL101

Biological features →

1-v-1 binary Classifiers →

PSL101 (Protein Subcellular Localization prediction by 1-On-1 classifiers)

Predicted Localization Site(s)
Feature Selection

- **Motivation**: unlikely to try all possible feature combinations in different classifiers

- **Feature selection**: reduce computational cost
  1. Select at least 1 preferred features for each classifier
     - Choose 1 feature from the preferred list for a classifier
  2. Add up to at most 4 features
     - If adding a new feature improves the accuracy → add the feature into the classifier

- **Example**:
  - Preferred features for C12: Sig, TMA, SA
  - Final selected features for C12: Sig, TMA
Accuracy and Feature Combination

- More biological features used in 1-v-1 SVM:
  - AA, Dip, SSE, SA, Sig, TMA, TMB, TAT, and Sec
- **1.4% improvement** over CELLO II in accuracy!

<table>
<thead>
<tr>
<th>1-v-1</th>
<th>Specific Feature Input</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AA</td>
</tr>
<tr>
<td>C12</td>
<td></td>
</tr>
<tr>
<td>C13</td>
<td>●</td>
</tr>
<tr>
<td>C14</td>
<td>●</td>
</tr>
<tr>
<td>C15</td>
<td>●</td>
</tr>
<tr>
<td>C23</td>
<td>●</td>
</tr>
<tr>
<td>C24</td>
<td>●</td>
</tr>
<tr>
<td>C25</td>
<td>●</td>
</tr>
<tr>
<td>C34</td>
<td>●</td>
</tr>
<tr>
<td>C35</td>
<td>●</td>
</tr>
<tr>
<td>C45</td>
<td>●</td>
</tr>
</tbody>
</table>

Overall accuracy = 91.40%
Outline

- Introduction
- SVM models
  - 1-v-r based on general features
  - 1-v-1 based on general features
- Biological features
  - 1-v-1 based on more specific biological features
  - 1-v-1 based on a new encoding for protein structures
- Conclusion
Refined Encoding Schemes to Encode Protein Structures (SSE)?

- **New:** consider **composition, transition, and distribution** of H, E, and C
  - SSE1 and SSE2 have the same composition but different transition and distribution!

SSE1: `CCCCCCCCCCCCCCCCCCCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH`
SSE2: `CCCCCCCCCCCCCCCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH`
A New Encoding Scheme for SSE

1. Amino acid composition (AA)
2. Dipeptide composition (Dip)
3. Secondary structure elements (SSE) – encoding scheme
2
4. Solvent accessibility (SA)
5. Signal peptides (Sig)
6. Transmembrane $\alpha$-helices (TMA)
7. Transmembrane $\beta$-barrels (TMB)
8. Twin-arginine translocase signal peptides (TAT)
9. Non-classical protein secretion (Sec)
A New Encoding Scheme (EC2) for SSE

1. Composition
   - The number of amino acids of H, E, and C

2. Transition
   - The percent frequency with which H $\leftrightarrow$ E, H $\leftrightarrow$ C, and E $\leftrightarrow$ C

3. Distribution
   - The chain length within which the first, 25, 50, 75 and 100% of the amino acids of a particular property is located respectively
     - $H_{1\%}$, $H_{25\%}$, $H_{50\%}$, $H_{75\%}$, $H_{100\%}$, $E_{1\%}$, $E_{25\%}$, $E_{50\%}$, $E_{75\%}$, $E_{100\%}$, $C_{1\%}$, $C_{25\%}$, $C_{50\%}$, $C_{75\%}$, and $C_{100\%}$
Accuracy and Feature Combination

- Features used in 1-v-1 SVM:
  - AA, Dip, SSE (EC1), SA, Sig, TMA, TMB, TAT, Sec, and SSE (EC2)
- New encoding scheme leads to an improvement of 1.3% in overall accuracy!

<table>
<thead>
<tr>
<th>1-v-1</th>
<th>Specific Feature Input</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AA</td>
</tr>
<tr>
<td>C12</td>
<td></td>
</tr>
<tr>
<td>C13</td>
<td></td>
</tr>
<tr>
<td>C14</td>
<td></td>
</tr>
<tr>
<td>C15</td>
<td>●</td>
</tr>
<tr>
<td>C23</td>
<td></td>
</tr>
<tr>
<td>C24</td>
<td></td>
</tr>
<tr>
<td>C25</td>
<td>●</td>
</tr>
<tr>
<td>C34</td>
<td>●</td>
</tr>
<tr>
<td>C35</td>
<td>●</td>
</tr>
<tr>
<td>C45</td>
<td>●</td>
</tr>
</tbody>
</table>

Overall accuracy = 92.70%
Outline

- Introduction
- SVM models
  - 1-v-r based on general features
  - 1-v-1 based on general features
- Biological features
  - 1-v-1 based on more specific biological features
  - 1-v-1 based on a new encoding for protein structures
- Conclusion
Conclusion

- **SVM models**
  - 1-v-1 outperforms 1-v-r in the flexibility of integrating different features

- **Biological features**
  - Compartment-specific features improve accuracy
  - A refined feature representation leads to $\text{Acc} = 92.7\%$ in PSL101 (vs. 90.0 % in CELLO II)

<table>
<thead>
<tr>
<th>Model</th>
<th>Biological Features</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-v-r SVM</td>
<td>●</td>
<td>87.7%</td>
</tr>
<tr>
<td>1-v-1 SVM</td>
<td>●</td>
<td>88.2%</td>
</tr>
<tr>
<td>1-v-1 SVM</td>
<td>●</td>
<td>91.4%</td>
</tr>
<tr>
<td>1-v-1 SVM</td>
<td>●</td>
<td>92.7%</td>
</tr>
</tbody>
</table>
People

Chia-Yu Su  Allan Lo  Hua-Sheng Chiu
Jia-Ming Chang  Ting-Yi Sung  Wen-Lian Hsu
Thank You!
Questions?