ACADEMIA SINICA
Taiwan International Graduate Program

Protein Subcellular Localization Prediction Based on Support Vector Machines

Chia-Yu Su

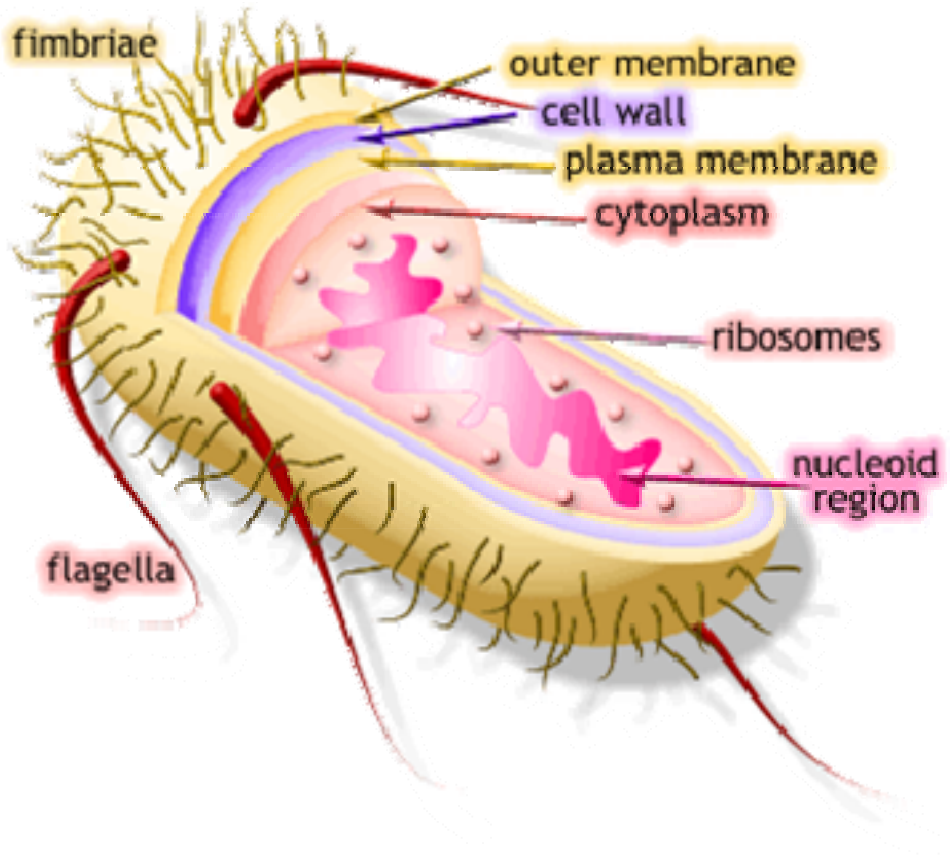
Institute of Information Science

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

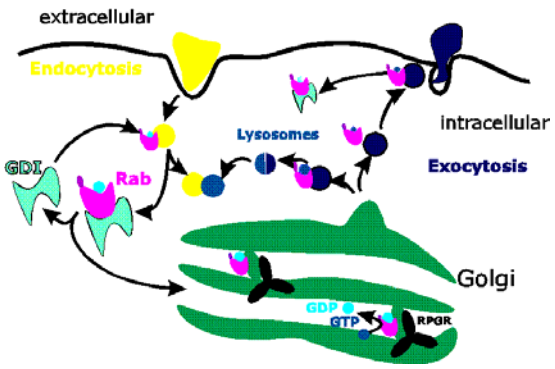


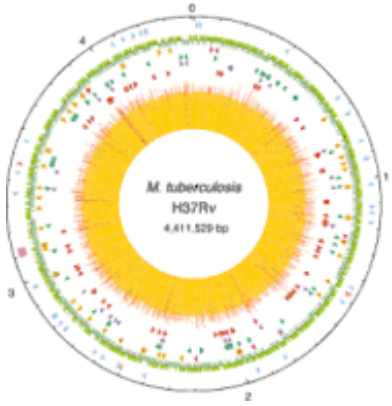
Gram-Negative Bacteria

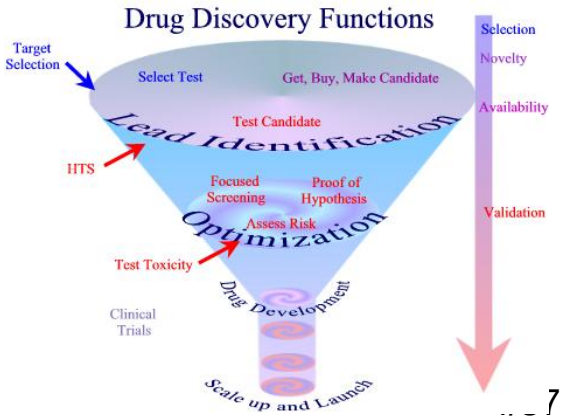
- **Predict where the protein is located in a cell?**
  - C1: cytoplasm
  - C2: inner membrane
  - C3: periplasm
  - C4: outer membrane
  - C5: extracellular

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

Systems	Approaches	Features	Accuracy
PSORTb	Bayesian Network	5 analytical modules	74.8%
P-CLASSIFIER	SVM	Amino acid subalphabets	89.8%
CELLO II	SVM	*n*-peptide compositions	90.0%
PSL101	SVM	Specific biological features	92.7%

The state-of-the-art system

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

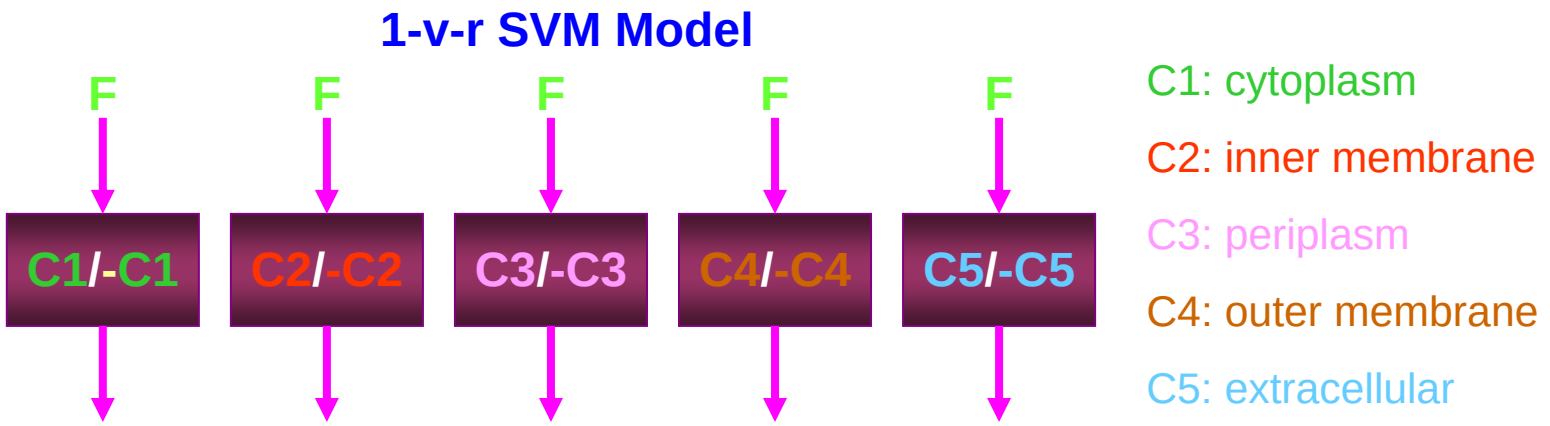
- **One-versus-rest (1-v-r) SVM model**
  - Apply a universal set of biological features for different localization classes
- **One-versus-one (1-v-1) SVM model**
  - Different biological features can be used in distinguishing two classes

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<= $prob_i$ <=1) is chosen as final predicted class

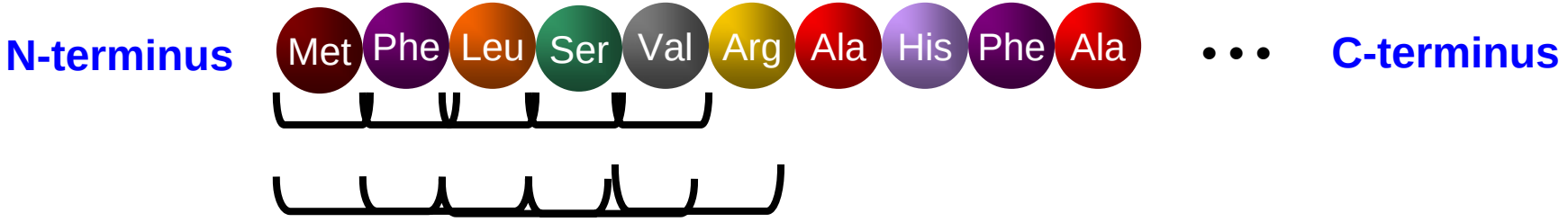


General Biological Features for PSL Prediction

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



3. Secondary Structure Elements

- **Predicted secondary structure elements (SSE) from HYPROSP II server**
  - Encoding scheme: compute amino acid compositions of α–helix (H), β-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 C H H H H H H C C E E E H H · ·

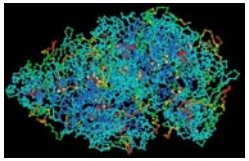
A C D … Y

↓ ↓ ↓ … ↓

0.12	0.02	0.03	…	0.05	←H
0.07	0.04	0.02	…	0.09	←E
0.08	0.06	0.07	…	0.03	←C

α–helix

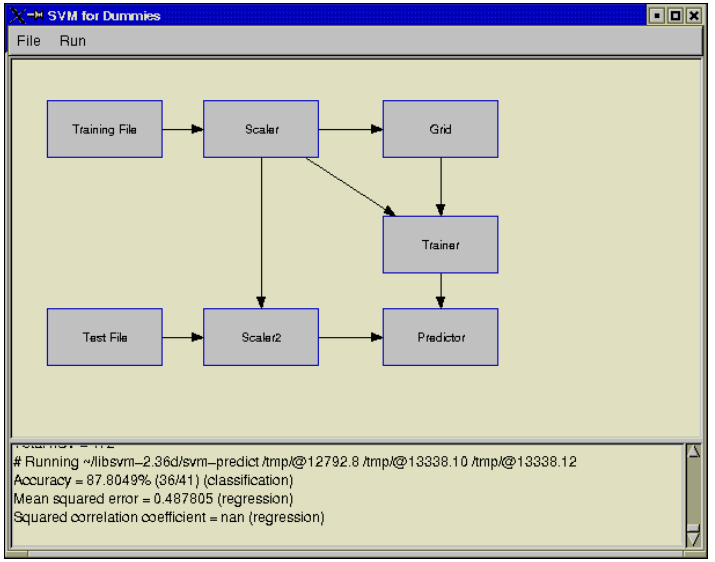
β-strand



random coil

Training and Testing in SVM

- **Support Vector Machines (SVM)**
  - LIBSVM software
  - Kernel: Radial Basis Function (RBF)
  - Parameter selection
    - **c** (cost) and γ (gamma) are optimized
  - 10-fold cross-validation



Gram-Negative Bacteria Data Set

Localization sites	No.
Cytoplasmic (CP)	248
Inner membrane (IM)	268
Periplasmic (PP)	244
Outer membrane (OM)	352
Extracellular (EC)	190
All sites	1,302

Performance Evaluation

- **Accuracy (*Acc*)**

$$Acc_i = TP_i / N_i$$

$$Acc = \sum_{i=1}^{l} TP_i \Big/ \sum_{i=1}^{l} N_i$$

  - $l = 5$ 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

Feature	AA	Dip	SSE	AA+Dip	AA+SSE	Dip+SSE	AA+Dip +SSE
Overall Acc	85.56%	84.87%	83.26%	87.71%	84.95%	83.95%	86.25%

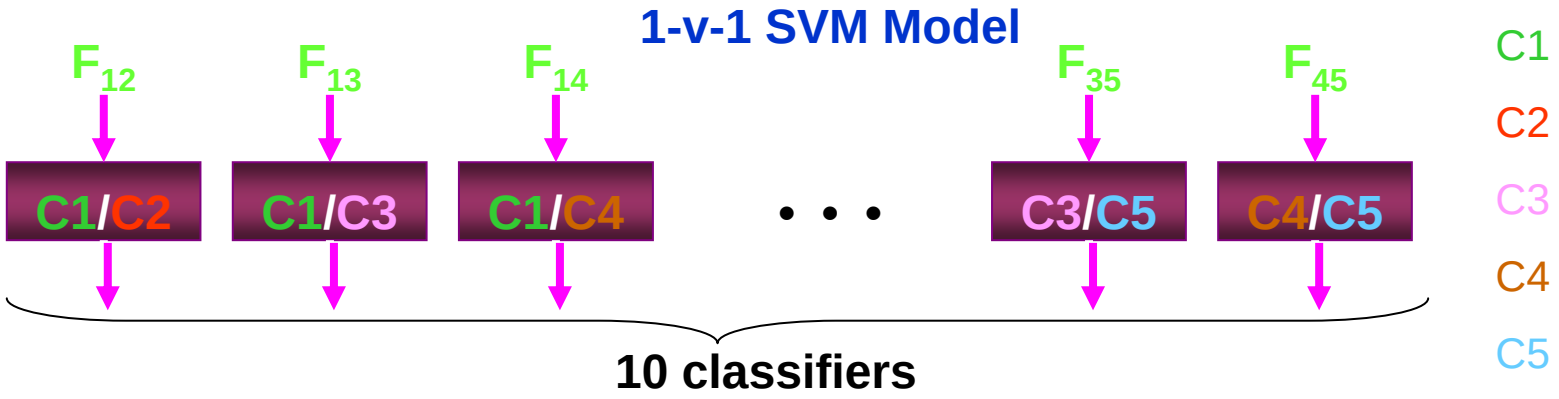
Acc of CELLO II = 90.0%

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 *i* and *j*, construct a C_i vs. 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



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

Acc of CELLO II = 90.0%

1-v-1 binary classifiers	Specific Feature Input		
	AA	Dip	SSE
C12		●	
C13	●		
C14	●	●	
C15	●	●	●
C23		●	
C24	●	●	
C25	●		
C34	●	●	
C35	●	●	
C45	●		●
Overall Accuracy=	88.17%		

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

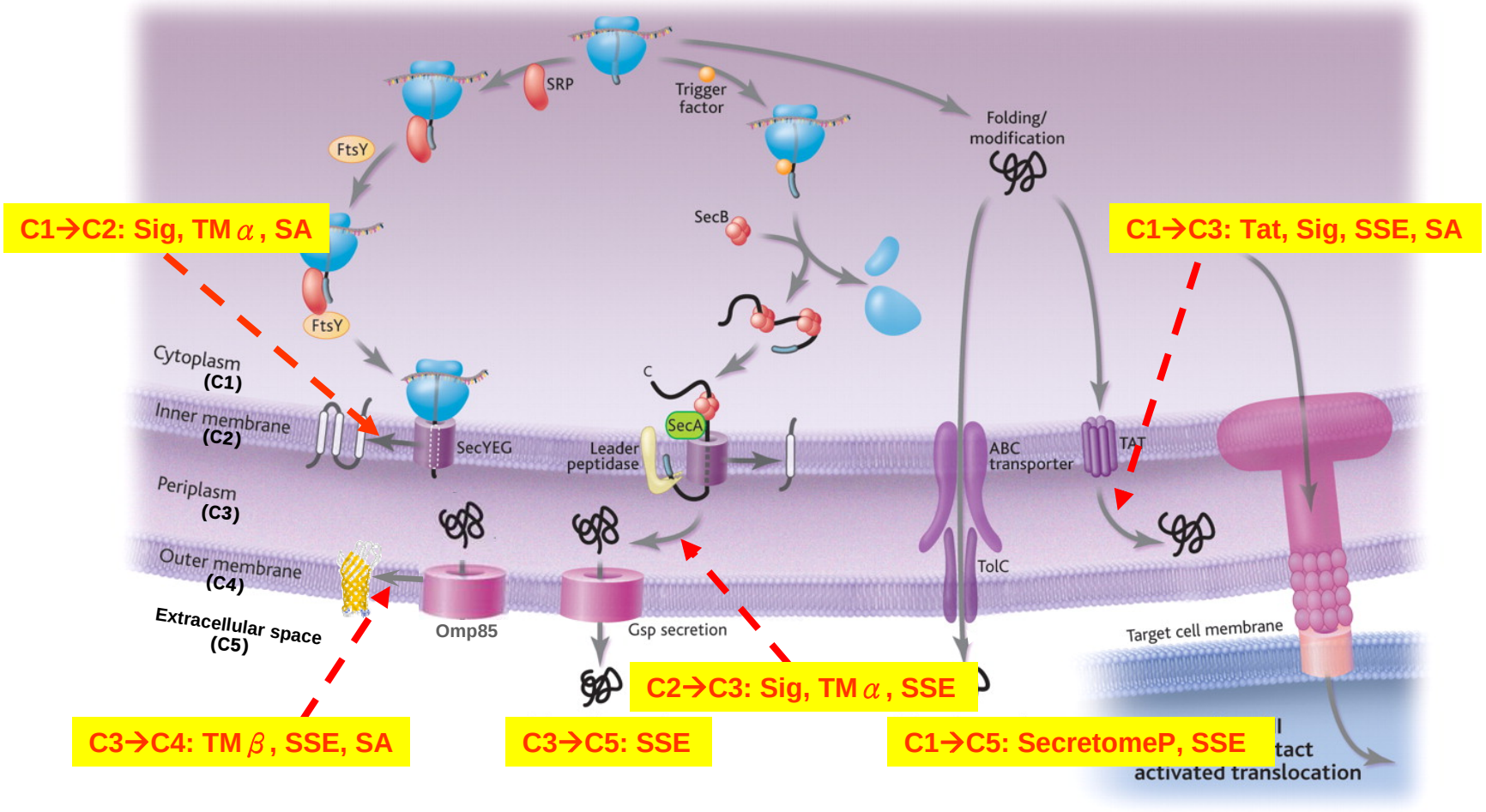


Figure adapted from Wickner W. and R. Schekman, Protein Translocation Across Biological Membranes, Science 2005

More Compartment-Specific Biological Features

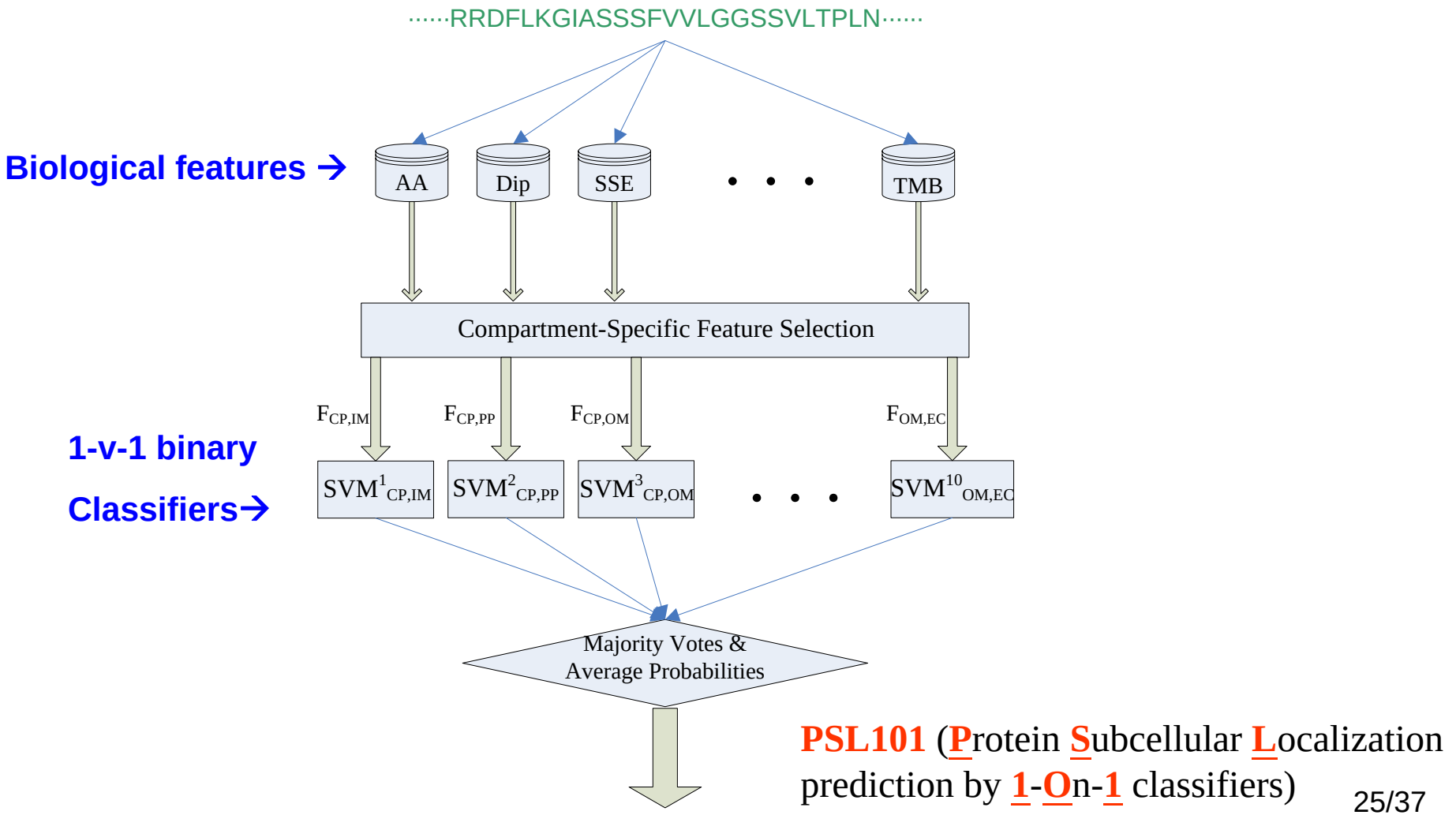
1. Amino acid composition (AA)
2. Dipeptide composition (Dip)
3. Secondary structure elements (SSE)
4. **Solvent accessibility (SA) –** C5
5. **Signal peptides (Sig) –** C1
6. **Transmembrane α-helices (TMA) –** C2
7. **Transmembrane β-barrels (TMB) –** C4
8. **Twin-arginine translocase signal peptides (TAT) –** C3
9. **Non-classical protein secretion (Sec) –** C5

Compartment-Specific Biological Features

- **Implication of different biological features to localization classes**

Fea.	Description	Features → Classes
SA	Solvent accessibility	Acidic high SA residues → C5
Sig	Signal peptides	Presence of Sig → not C1
TMA	Transmembrane α-helices	Presence of TMA → C2
TAT	Twin-arg translocase motifs	Presence of TAT → C3
TMB	Transmembrane β-barrels	Presence of TMB → C4
Sec	Non-classical protein secretion	Presence of Sec → C5

System Architecture of PSL101



PSL101 (**P**rotein **S**ubcellular **L**ocalization prediction by **1**-**O**n-**1** classifiers)

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!**

1-v-1	Specific Feature Input								
	AA	Dip	SA	SSE	Sig	TMA	TAT	TMB	Sec
C12					●	●			
C13	●	●		●	●				
C14		●			●			●	
C15	●	●	●	●					
C23	●				●	●	●		
C24		●		●		●			
C25	●		●			●			
C34	●	●						●	
C35	●	●							
C45	●	●	●	●					
Overall accuracy = 91.40%									

Outline

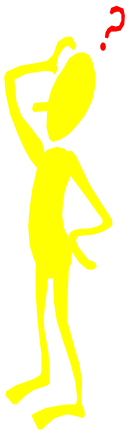
- **Introduction**
- **SVM models**
  - 1-v-r based on general features
  - 1-v-1 based on general features
- **Biological features**
  - 1-v-1 based on more specific biological features
  - 1-v-1 based on a new encoding for protein structures
- **Conclusion**

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: CCCCCCCCCCCCCCCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHH

SSE2: CCCHHHHHHCCCHHHHHHCCCCCHHHHHCCCHHHHHCCCCCHHHHH



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 α-helices (TMA)
7. Transmembrane β-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 ←→ E, H ←→ C, and E ←→ 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!**

1-v-1	Specific Feature Input									
	AA	Dip	SA	SSE(EC1)	SSE(EC2)	Sig	TMA	TAT	TMB	Sec
C12			●			●	●			
C13			●		●	●				
C14		●				●			●	
C15	●		●			●				
C23	●					●	●	●		
C24		●		●			●			
C25	●	●					●			●
C34	●	●							●	
C35	●	●			●					
C45	●		●		●					
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 Acc = **92.7% in PSL101** (vs. 90.0 % in CELLO II)

Model	Biological Features										Acc
	AA	Dip	SSE EC1	SA	Sig	TMA	TMB	TAT	Sec	SSE EC2	
1-v-r SVM	●	●	●								**87.7%**
1-v-1 SVM	●	●	●								**88.2%**
1-v-1 SVM	●	●	●	●	●	●	●	●	●		**91.4%**
1-v-1 SVM	●	●	●	●	●	●	●	●	●	●	**92.7%**

People

Chia-Yu Su

Allan Lo

Hua-Sheng Chiu

Jia-Ming Chang

Ting-Yi Sung

Wen-Lian Hsu

Thank You!

Questions?