# PROTEIN SUBCELLULAR LOCALIZATION PREDICTION BASED ON COMPARTMENT-SPECIFIC BIOLOGICAL FEATURES

Chia-Yu Su[1,2], Allan Lo[1,3], Hua-Sheng Chiu[4], Ting-Yi Sung[4], Wen-Lian Hsu[4,*]

[1]*Bioinformatics Program, Taiwan International Graduate Program, Academia Sinica, Taipei, Taiwan*
[2]*Institute of Bioinformatics, National Chiao Tung University, Hsinchu, Taiwan*
[3]*Department of Life Sciences, National Tsing Hua University, Hsinchu, Taiwan*
[4]*Bioinformatics Lab., Institute of Information Science, Academia Sinica, Taipei, Taiwan*
*Email: {cysu, allanlo, huasheng, tsung, hsu}@iis.sinica.edu.tw*

Prediction of subcellular localization of proteins is important for genome annotation, protein function prediction, and drug discovery. We present a prediction method for Gram-negative bacteria that uses ten one-versus-one support vector machine (SVM) classifiers, where compartment-specific biological features are selected as input to each SVM classifier. The final prediction of localization sites is determined by integrating the results from ten binary classifiers using a combination of majority votes and a probabilistic method. The overall accuracy reaches 91.4%, which is 1.6% better than the state-of-the-art system, in a ten-fold cross-validation evaluation on a benchmark data set. We demonstrate that feature selection guided by biological knowledge and insights in one-versus-one SVM classifiers can lead to a significant improvement in the prediction performance. Our model is also used to produce highly accurate prediction of 92.8% overall accuracy for proteins of dual localizations.

## 1. INTRODUCTION

Gram-negative bacteria have five major subcellular localization sites, which are the cytoplasm (CP), the inner membrane (IM), the periplasm (PP), the outer membrane (OM), and the extracellular space (EC). Prediction of protein subcellular localization for Gram-negative bacteria has been extensively studied and several systems have been developed. PSORT I[1] has been a widely used prediction tool. Gardy *et al*.[2] proposed PSORT-B, a multi-modular method combined with a Bayesian network, to improve the performance of PSORT I. Although PSORT-B has a high precision, it only yields an overall prediction recall, also referred to as accuracy, of 74.8%. Yu *et al*.[3] presented an approach called CELLO that utilized support vector machines (SVM) based on *n*-peptide compositions. The overall prediction accuracy of CELLO reaches 88.9% but the accuracy for extracellular proteins is still relatively low, at 78.9%. Recently, Wang *et al*.[4] developed a system called P-CLASSIFIER that used multiple SVM based on amino acid subalphabets. The system attains an overall prediction accuracy of 89.8%.

In this study, we present a method called PSL101 (Protein Subcellular Localization prediction by 1-On-1 classifiers) that incorporates compartment-specific biological features in ten one-versus-one (1-v-1) SVM classifiers to predict protein subcellular localization for Gram-negative bacteria. Given a protein sequence, PSL101 constructs feature vectors extracted from specific input features that are characteristic of a given localization. These features include amino acid composition, di-peptide composition, solvent accessibility, secondary structure, signal peptides, transmembrane $\alpha$-helices, transmembrane $\beta$-barrels, and non-classical protein secretion. Biological knowledge and insights are used to guide our feature selection in the classification of different compartments. The output probability values from ten binary classifiers are integrated by a combination of majority votes and a probabilistic method to determine the final prediction of localization sites. Experiment results show that our method attains an overall prediction accuracy of 91.4%, which has presently the most accurate prediction performance for single-localized proteins. Based on a forward feature selection algorithm, the final feature combinations correlate well with biological insights. We further make use of this method in the prediction of dual-localized proteins and obtain an overall accuracy of 92.8%.

## 2. METHODS

### 2.1. SVM framework

SVM has been widely used in pattern recognition applications on data mining and bioinformatics. Prediction of

* Corresponding author.

protein subcellular localization can be treated as a multi-class classification problem. For multi-class classification, the one-versus-rest (1-v-r) SVM model has demonstrated a good classification performance[5]. However, for any localization site, it is difficult to find a universal set of biological features from the remaining four sites that can be effectively used for 1-v-r SVM model. Based on biological domain knowledge, compartment-specific biological features should be used in distinguishing two localization sites, and this presupposition is later confirmed by our experiment results. Thus, we propose to use ten 1-v-1 SVM classifiers for protein subcellular localization prediction. The system architecture of PSL101 is shown in Fig. 1.

The LIBSVM[6] software is used in our experiments. For all classifiers, we use Radial Basis Function (RBF) kernel and optimize the cost ($c$) and *gamma* ($\gamma$) parameters. The probability estimates by LIBSVM are used for determining the confidence levels of classifications[7].

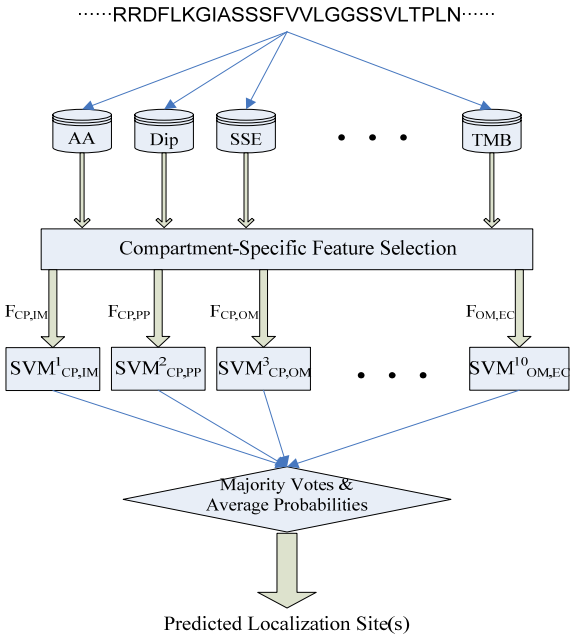


**Fig. 1.** System architecture of one-versus-one SVM models based on compartment-specific features.

## 2.2. Biological input features

In Gram-negative bacteria secretory pathways, proteins localized to a particular subcellular compartment have distinct biological properties. We consider the following nine biological input features to distinguish between proteins translocated to different compartments and construct our classification framework to mimic the translocation process of bacterial secretory pathways.

1. *Amino acid composition (AA).* Protein descriptors based on $n$-peptide compositions or their variations have been shown effective in protein subcellular localization prediction[3,4]. If $n = 1$, then the $n$-peptide composition reduces to the amino acid composition. The feature vector is of dimension 21 (i.e., 20 amino acids types plus a symbol 'X,' for others).
2. *Di-peptide composition (Dip).* The $n$-peptide compositions preserve more global sequence information when $n$ gets larger. For computational efficiency, we choose $n = 2$, the di-peptide composition. This feature vector has dimension 441 ($21 \times 21$).
3. *Solvent accessibility (SA).* Protein structures from different compartments show characteristic differences, particularly at the surface, which is directly exposed to the environment. Proteins in different localization sites have different surface residue compositions. Cytoplasmic proteins have a balance of acidic and basic surface residues, while extracellular proteins have a slight excess of acidic surface residues[8]. Thus, solvent accessibility represented by the amino acid composition of surface residues could be useful to identify extracellular proteins.
4. *Secondary structure elements (SSE).* Transmembrane $\alpha$-helices are frequently observed in inner membrane proteins while transmembrane $\beta$-barrels are largely found in outer membrane proteins[9]. The secondary structure elements are useful for detecting proteins localized in the inner membrane and the outer membrane. We compute the amino acid compositions of three secondary structure elements ($\alpha$-helix, $\beta$-strand, and random coil) based on the predicted results from HYPROSP II[10], a knowledge-based secondary structure prediction approach.
5. *Signal peptides (Sig).* Signal peptides are N-terminal peptides typically between 15 and 40 amino acids long, and they target proteins for translocation through the general secretory pathway[11]. The presence of a signal peptide suggests that the protein does not reside in the cytoplasm. SignalP[12], a neural network and hidden Markov model based method, is used to predict the presence and location of signal peptide cleavage sites in protein sequences. We employ this prediction method to distinguish cytoplasmic and non-cytoplasmic proteins.

6. *Transmembrane α-helices (TMA).* Integral inner membrane proteins are characterized by transmembrane $\alpha$-helices. The presence of transmembrane $\alpha$-helices could imply that the protein is located in the inner membrane. TMHMM[13] is a hidden Markov model based method for the prediction of transmembrane $\alpha$-helices and their topology in proteins. We apply TMHMM to identify potential transmembrane $\alpha$-helical proteins residing in the inner membrane.
7. *Twin-arginine signal peptides (TAT).* The twin arginine translocase (TAT) system exports proteins from the cytoplasm to the periplasm. The proteins translocated by TAT bear a unique twin-arginine motif[14]. The presence of the motif is a useful feature to distinguish periplasmic and non-periplasmic proteins. TatP server[15] uses a combination of two neural networks to predict the presence and location of twin-arginine signal peptide cleavage sites in bacteria. This server is used to detect TAT.
8. *Transmembrane β-barrels (TMB).* A large number of proteins residing in the outer membrane are characterized by $\beta$-barrel structures. Thus, they could be a candidate feature to detect outer membrane proteins. TMB-Hunt[16] is a method that uses a modified *k*-Nearest Neighbor (*k*-NN) algorithm to distinguish protein sequences of transmembrane $\beta$-barrel (TMB) from non-TMB on the basis of amino acid composition. We employ TMB-Hunt to identify potential outer membrane proteins.
9. *Non-classical protein secretion (Sec).* It had been believed for a long time that an N-terminal signal peptide was strictly required to export a protein to the extracellular space. Recent studies, however, have shown that several extracellular proteins can be secreted without a classical N-terminal signal peptide[17]. Identification of non-classical protein secretion, which is not triggered by signal peptides, could be a potential discriminator for cytoplasmic and extracellular proteins. Predictions produced from SecretomeP[18], a non-classical protein secretion method, are applied in our experiments.

## 2.3. Feature selection in SVM classifiers

Since it is unlikely to try all possible feature combinations in different classifiers, heuristics guided by biological insights are used to determine a small subset of input features specific to each classifier. Starting with an empty subset, a forward feature selection algorithm keeps adding the best features that lead to an improvement on the accuracy of the classifiers. The process is terminated if adding the features no longer improves the accuracy.

## 2.4. Class determination

In order for each binary classifier $C_{i,j}$ to distinguish class *i* and *j*, the input feature vector is constructed by concatenating different biological features refined specifically according to the intrinsic characteristics of proteins in localization sites *i* and *j*. We utilize several prediction methods to extract specific features based on biological domain knowledge. For each protein in the testing set, a predicted class and its corresponding probability are returned from each classifier.

In order to determine the predicted localization site of each protein, we combine the predicted results from ten binary classifiers by majority votes. In the case of a tie, the localization site with the highest average probability is assigned as the final prediction of localization site.

# 3. RESULTS AND DISCUSSION

## 3.1. Benchmark data set

To train and test our method, we use a benchmark data set of proteins from Gram-negative bacteria applied in previous works[1-4]. It consists of 1,441 proteins with experimentally determined localizations, in which 1,302 proteins have a single localization site and 139 proteins have dual localization sites. Table 1 lists the number of proteins in different sites in the data set.

**Table 1.** Number of proteins in different localization sites.

| Localization sites | No. |
|---|---|
| Cytoplasmic (CP) | 248 |
| Inner membrane (IM) | 268 |
| Periplasmic (PP) | 244 |
| Outer membrane (OM) | 352 |
| Extracellular (EC) | 190 |
| Cytoplasmic / Inner membrane (CP / IM) | 14 |
| Inner membrane / Periplasmic (IM / PP) | 49 |
| Outer membrane / Extracellular (OM / EC) | 76 |
| All sites | 1,441 |

### 3.2. Evaluation measures

For comparison with other approaches, we follow the same measures used in previous works[1-4] to evaluate the performance of our method. Accuracy (*Acc*) and Matthew's correlation coefficient (*MCC*)[19] defined in Eq. (1) and (2) are used to assess the performance at five localization sites. The overall accuracy is defined in Eq. (3).

$$Acc_i = TP_i / N_i \tag{1}$$

$$MCC_i = \frac{(TP_i)(TN_i) - (FP_i)(FN_i)}{\sqrt{(TP_i + FN_i)(TP_i + FP_i)(TN_i + FP_i)(TN_i + FN_i)}} \tag{2}$$

$$Acc = \sum_{i=1}^{l} TP_i \Big/ \sum_{i=1}^{l} N_i \ , \tag{3}$$

where $l = 5$ is the number of total localization sites, and $TP_i$, $TN_i$, $FP_i$, $FN_i$ ,and $N_i$ are the number of true positives, true negatives, false positives, false negatives, and proteins in localization site $i$, respectively. *MCC*, considering both under-and over-predictions, offers a complementary measure for the prediction performance, where *MCC* = 1 indicates a perfect prediction and *MCC* = 0 indicates a completely random assignment.

Due to different intrinsic characteristics of single localization and dual localization proteins, their prediction results are reported separately.

### 3.3. Results of single localization proteins

In Table 2, we compare the performance of our approach with other approaches using 1,302 single-localized proteins in a ten-fold cross validation test. The overall accuracy of PSL101 reached 91.4%, which is 1.6% better than the state-of-the-art system, P-CLASSIFIER. In addition, PSL101 outperforms P-CLASSIFIER in terms of *MCC*s except for extracellular proteins. The compartment-specific features selected in PSL101 are summarized in Table 3. The experiment results show that our feature selection not only leads to a significant improvement in the overall accuracy but also correlates well with biological insights. For example, PSL101 selects signal peptides and transmembrane $\alpha$-helices as the optimal features to distinguish proteins localized in the cytoplasm (no signal peptides) and the inner membrane (presence of transmembrane $\alpha$-helices).

### 3.4. Results of dual localization proteins

For dual localization classification, we conduct two experiments. In the first experiment, we compare with P-CLASSIFIER in which the dual-localized proteins are tested with classifiers trained on single-localized proteins. The two localization sites receiving highest probability sums from the 10 classifiers are assigned as the dual localization sites of the protein. Instead of giving full marks to dual-localized proteins with at least one site predicted correctly, we choose a less biased criterion to assess the performance: if only one of the dual

**Table 3.** Compartment-specific feature selection.

| 1-v-1 classifiers | AA | Dip | SA | SSE | Sig | TMA | TAT | TMB | Sec |
|---|---|---|---|---|---|---|---|---|---|
| $C_{CP, IM}$ | | | | | ● | ● | | | |
| $C_{CP, PP}$ | ● | ● | | ● | ● | | | | |
| $C_{CP, OM}$ | | ● | | | ● | | | | ● |
| $C_{CP, EC}$ | ● | ● | ● | ● | | | | | |
| $C_{IM, PP}$ | ● | | | | ● | ● | ● | | |
| $C_{IM, OM}$ | | ● | | ● | | ● | | | |
| $C_{IM, EC}$ | ● | | ● | | | ● | | | |
| $C_{PP, OM}$ | ● | ● | | | | | | | ● |
| $C_{PP, EC}$ | ● | ● | | | | | | | |
| $C_{OM, EC}$ | ● | ● | ● | ● | | | | | |

**Table 2.** The comparison of different approaches in the prediction of subcellular localization for Gram-negative bacteria.

| Localization | **PSL101** | | P-CLASSIFIER | | CELLO | | PSORT-B | | PSORT I | |
|---|---|---|---|---|---|---|---|---|---|---|
| | ***Acc* (%)** | ***MCC*** | *Acc* (%) | *MCC* | *Acc* (%) | *MCC* | *Acc* (%) | *MCC* | *Acc* (%) | *MCC* |
| CP | **95.2** | **0.88** | 94.6 | 0.85 | 90.7 | 0.85 | 69.4 | 0.79 | 75.4 | 0.58 |
| IM | **93.7** | **0.95** | 87.1 | 0.92 | 88.4 | 0.92 | 78.7 | 0.85 | 95.1 | 0.64 |
| PP | **87.3** | **0.84** | 85.9 | 0.81 | 86.9 | 0.80 | 57.6 | 0.69 | 66.4 | 0.55 |
| OM | **93.8** | **0.93** | 93.6 | 0.90 | 94.6 | 0.90 | 90.3 | 0.93 | 54.5 | 0.47 |
| EC | **84.2** | **0.83** | 86.0 | 0.89 | 78.9 | 0.82 | 70.0 | 0.79 | – | – |
| Overall | **91.4** | – | 89.8 | – | 88.9 | – | 74.8 | – | 60.9 | – |

sites is predicted correctly, the prediction receives only half mark. Table 4 lists the prediction performance. PSL101 outperforms P-CLASSIFIER except for the class of cytoplasmic/inner membrane, in which there are only 14 proteins in the data set.

In the second experiment, we apply 1-v-1 SVM models directly on dual-localized proteins in a ten-fold cross validation test. Since there are three pairs of dual localization sites: {CP,IM}, {IM,PP}, and {OM,EC}, we use the following three 1-v-1 SVM classifiers: $C_{\{CP,IM\},\{IM,PP\}}$, $C_{\{CP,IM\},\{OM,EC\}}$, and $C_{\{IM,PP\},\{OM,EC\}}$. For each dual-localized protein, ten predicted probabilities, generated from previous 10 classifiers trained on single-localized proteins, comprise the input feature vector (of dimension 10). Since the classifier $C_{\{CP,IM\},\{IM,PP\}}$ has the IM site in common, it requires an additional single localization classifier $C_{CP,PP}$ to distinguish {CP,IM} and {IM,PP}. Thus the final prediction of dual localization sites is determined by a combination of the output probabilities from the 3 dual localization classifiers and the single localization classifier $C_{CP,PP}$. The final prediction of localization sites are determined by a combination of the output probabilities from both dual localization classifiers and the distinct single localization classifiers. To assess the prediction performance, we use the same evaluation measures defined in Eq. (1), (2), and (3). The predicted results are shown in Table 5. The overall accuracy reaches 92.8% for proteins localized in two different localizations. The results indicate that PSL101 performs consistently well in both single and dual localization proteins. Thus, the input feature vector of dimension 10 trained on single-localized proteins is able to capture the important relationships between input biological features and characteristics of localization sites.

## 4. CONCLUSION

In this study, we propose a method to predict protein subcellular localization using multiple 1-v-1 SVM models based on compartment-specific features. Experiment results show that our method attains high overall prediction accuracies of 91.4% and 92.8% for single and dual localization proteins, respectively. The feature combinations generated by a forward feature selection algorithm correlate well with biological insights. Our method provides accurate predictions and suggests useful biological features in protein localization prediction.

**Table 4.** The comparison of the prediction performance for dual localization proteins.

| Localization | **PSL101** | | P-CLASSIFIER | |
|---|---|---|---|---|
| | **Mark** | ***Acc* (%)** | Mark | *Acc* (%) |
| CP / IM | **6.5** | **46.4** | 10.5 | 75.0 |
| IM / PP | **26.5** | **54.1** | 19.0 | 38.8 |
| OM / EC | **73.0** | **96.1** | 64.0 | 84.2 |
| Overall | **106** | **76.3** | 93.5 | 67.3 |

**Table 5.** The performance of dual localization classifiers that use predicted probabilities from 10 single localization classifiers as an input feature.

| Localization | *Acc* (%) | *MCC* |
|---|---|---|
| CP / IM | 64.3 | 0.70 |
| IM / PP | 93.9 | 0.85 |
| OM / EC | 97.4 | 0.96 |
| Overall | 92.8 | – |

## Acknowledgments

We thank Hsin-Nan Lin, Jia-Ming Chang, and Ching-Tai Chen for helpful suggestions and computational assistance. The research was supported in part by the thematic program of Academia Sinica under grant AS94B003 and AS95ASIA02.

## References

1. Nakai K and Kanehisa M. Expert system for predicting protein localization sites in Gram-negative bacteria. *Proteins* 1991; **11:** 95-110.
2. Gardy JL, Spencer C, Wang K, Ester M, Tusnady GE, Simon I, Hua S, deFays K, Lambert C, Nakai K, *et al*. PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acids Res* 2003; **31:** 3613-3617.
3. Yu CS, Lin CJ, and Hwang JK. Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. *Protein Sci* 2004; **13:** 1402-1406.
4. Wang J, Sung WK, Krishnan A, and Li KB. Protein subcellular localization prediction for Gram-negative bacteria using amino acid subalphabets and a combination of multiple support vector machines. *BMC Bioinformatics* 2005; **6:** 174.

5. Garg A, Bhasin M, and Raghava GP. Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. *J Biol Chem* 2005; **280:** 14427-14432.
6. Chang CC and Lin CJ. LIBSVM: A library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/.
7. Wu TF, Lin CJ, and Weng RC. Probability estimates for multi-class classification by pairwise coupling. *J Machine Learning Res* 2004; **5:** 975-1005.
8. Andrade MA, O'Donoghue SI, and Rost B. Adaptation of protein surfaces to subcellular location. *J Mol Biol* 1998; **276:** 517-525.
9. Pautsch A and Schulz GE. Structure of the outer membrane protein A transmembrane domain. *Nat Struct Biol* 1998; **5:** 1013-1017.
10. Lin HN, Chang JM, Wu KP, Sung TY, and Hsu WL. HYPROSP II--a knowledge-based hybrid method for protein secondary structure prediction based on local prediction confidence. *Bioinformatics* 2005; **21:** 3227-3233.
11. Emanuelsson O, Nielsen H, Brunak S, and von Heijne G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol* 2000; **300:** 1005-1016.
12. Bendtsen JD, Nielsen H, von Heijne G, and Brunak S. Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* 2004; **340:** 783-795.
13. Krogh A, Larsson B, von Heijne G, and Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* 2001; **305:** 567-580.
14. Berks BC. A common export pathway for proteins binding complex redox cofactors? *Mol Microbiol* 1996; **22:** 393-404.
15. Bendtsen JD, Nielsen H, Widdick D, Palmer T, and Brunak S. Prediction of twin-arginine signal peptides. *BMC Bioinformatics* 2005; **6:** 167.
16. Garrow AG, Agnew A, and Westhead DR. TMB-Hunt: an amino acid composition based method to screen proteomes for beta-barrel transmembrane proteins. *BMC Bioinformatics* 2005; **6:** 56.
17. Nickel W. The mystery of nonclassical protein secretion. A current view on cargo proteins and potential export routes. *Eur J Biochem* 2003; **270:** 2109-2119.
18. Bendtsen JD, Jensen LJ, Blom N, Von Heijne G, and Brunak S. Feature-based prediction of non-classical and leaderless protein secretion. *Protein Eng Des Sel* 2004; **17:** 349-356.
19. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta* 1975; **405:** 442-451.