http://www.scirp.org/journal/ns Natural Science, 2018, Vol. 10, (No. 5), pp: 168-177

An Improved Approach for Rapidly Identifying Different Types of Gram-Negative Bacterial Secreted Proteins

Lezheng Yu[1], Fengjuan Liu[1], Lixiao Du[1], Yizhou Li[2]

[1]School of Chemistry and Life Science, Guizhou Education University, Guiyang, China; [2]College of Chemistry, Sichuan University, Chengdu, China

Correspondence to: Lezheng Yu, xinyan_scu@126.com

Keywords: Gram-Negative Bacteria, Secreted Protein, Position-Specific Scoring Matrix, Signal Peptide, Support Vector Machine

Received: March 26, 2018 **Accepted:** May 15, 2018 **Published:** May 18, 2018

ABSTRACT

Protein secretion plays an important role in bacterial lifestyles. In Gram-negative bacteria, a wide range of proteins are secreted to modulate the interactions of bacteria with their environments and other bacteria via various secretion systems. These proteins are essential for the virulence of bacteria, so it is crucial to study them for the pathogenesis of diseases and the development of drugs. Using amino acid composition (AAC), position-specific scoring matrix (PSSM) and N-terminal signal peptides, two different substitution models are firstly constructed to transform protein sequences into numerical vectors. Then, based on support vector machine (SVM) and the "one to one" algorithm, a hybrid multi-classifier named SecretP v.2.2 is proposed to rapidly and accurately distinguish different types of Gram-negative bacterial secreted proteins. When performed on the same test set for a comparison with other methods, SecretP v.2.2 gets the highest total sensitivity of 93.60%. A public independent dataset is used to further test the power of SecretP v.2.2 for predicting NCSPs, it also yields satisfactory results.

1. INTRODUCTION

As a universal and important biological process, protein secretion may occur in all organisms. In this process, Gram-negative bacterial secreted proteins should cross two lipid bilayers including the cytoplasmic membrane (CM) and the outer membrane (OM), while Gram-positive bacterial secreted proteins just need cross the CM [1]. Therefore, the secretion process of the former is more complex than that of the latter, and more secretion systems are existing in Gram-negative bacterial cells.

Up to now, at least nine secretion systems have been discovered from Gram-negative bacteria, which are named from the type I (T1SS) to the type IX secretion system (T9SS) on the basis of the OM secretion

mechanisms [2]. Proteins released via the T1SS are called type I secreted proteins (T1SPs), and other types of proteins are known by analogy with this. According to the presence of N-terminal signal peptides or not, secreted proteins can be simply classified into two groups: classically secreted proteins (CSPs) (e.g., T2SPs, T5SPs, T7SPs, T8SPs and T9SPs) and non-classically secreted proteins (NCSPs) (e.g., T1SPs, T3SPs, T4SPs and T6SPs) [3]. They are normally secreted into the extracellular environment or directly injected into host cells, but also anchored to the OM at times, even as a part of cell-surface appendages such as flagella and pili [4]. They are essential for the virulence of bacteria and lead to various diseases [5] [6], so it is crucial to study them for the pathogenesis of diseases and the development of drugs. Unfortunately, researchers pay more attention to the structure and function of different secretion systems, rather than their secretory products [7]. Moreover, there have been a number of computational approaches designed to identify type-specific Gram-negative bacterial secreted proteins, such as T3SPs [9]-[14] or T4SPs [15] [16] [17] [18] [19], but only a few for distinguishing different types of secreted proteins simultaneously.

Based on our previous research [20], this work is intended to further improve the efficiency of recognition among different types of Gram-negative bacterial secreted proteins. Firstly, two different substitution models are developed based on AAC, PSSM and N-terminal signal peptides. Then, a SVM-based multi-classifier is constructed by the "one to one" algorithm, which is called SecretP v.2.2 in this paper. When using a test set to assess the actual performance of SecretP v.2.2, it achieves an overall sensitivity of 93.60% for distinguishing six different types of Gram-negative bacterial secreted proteins. Furthermore, a public independent dataset is used to evaluate the prediction performance of SecretP v.2.2 in identifying different types of NCSPs, and the prediction results are comparable to those of the previous version SecretP v.2.1.

2. MATERIALS AND METHODS

2.1. Data Sets

To make a comprehensive comparison in method, all data sets used in this study are exactly the same as those in our previous work [20]. The training and test sets consisted of six types of Gram-negative bacterial secreted proteins, including T1SPs, T2SPs, T3SPs, T4SPs, T5SPs and T7SPs. Here, "T1SP" represents the type I secreted protein, and the remaining are named by analogy with it. A public independent dataset of 89 NCSPs was constructed by Kampenusa and Zikmanis [21], which contains 32 T1SPs, 41 T3SPs and 16 T4SPs. The detailed data processing has been described in our previous work [20], and all data sets used in this study are listed in **Table 1**.

Table 1. All data sets used in this study.

Type	Training set	Test set	Independent dataset
T1SP	112	25	32
T2SP	99	29	-
T3SP	182	28	41
T4SP	62	22	16
T5SP	164	35	-
T7SP	48	33	-
Total	667	172	89

2.2. Feature Extraction

2.2.1. Amino Acid Composition

Amino acid composition (AAC) represents the occurrence frequencies of the twenty common amino acids in a protein sequence, and each protein is described as a 20-dimensional vector by this method.

2.2.2. Position-Specific Scoring Matrix

Position-specific scoring matrix (PSSM) is commonly used to describe the evolutionary information of amino acid residues in protein sequences, and it has been repeatedly proved that when adding PSSM into a protein substitution model for protein classification, the prediction performance of the method will significantly improve [22] [23]. So PSSM was also chosen to represent protein samples in this study. The PSSM for each protein sequence was firstly generated by using PSI-BLAST [24] against the Swiss-Prot database, with three iterations and an *E*-value cut-off of 0.001. In this way, a matrix consisting of L rows and 20 columns is created, where L is the length of a query sequence, and 20 columns represent occurrence or substitution of each type of twenty common amino acids. Because the lengths of proteins are not equal, an equation was then used to make all PSSM matrix size-uniformed, as described in our earlier study [23]. Finally, a 20-dimensional vector is also obtained for each protein sequence.

2.2.3. N-Terminal Signal Peptides

As a critical factor for distinguishing CSPs from NCSPs, N-terminal signal peptides in protein sequences were predicted by the SignalP 4.1 server [25], and represented by the D-scores.

2.3. Model Construction

Support vector machine (SVM) has been shown as a powerful machine learning algorithm in computational biology [20] [23] [26] [27] [28] [29]. Here, the LIBSVM program (http://www.csie.ntu.edu.tw/~cjlin/libsvm/) was employed to build different SVM models. As the default kernel function of LIBSVM, the radial basis function (RBF) was chosen here, and a grid search approach was used to optimize the regularization parameter C and the kernel width parameter γ. Though there have been several different validation methods in statistical prediction, the jackknife test is deemed the most rigorous and objective [30], and it was also adopted for this study. It has been confirmed that the "one to one" algorithm is more effective than the "one to rest" algorithm [20], so the "one to one" algorithm was also selected to solve the multi-class classification problem. Meanwhile, different weights were assigned to reduce the data imbalance, which are inversely proportional to the corresponding rates between any two types of secreted proteins in the training set.

2.4. Performance Evaluation

In order to evaluate the performance of different types of SVM models, sensitivity and accuracy are used here, and they are defined by the following equations.

$$\text{Sensitivity} = TP/(TP + FN) \tag{1}$$

$$\text{Accuracy} = (TP + TN)/(TP + TN + FP + FN) \tag{2}$$

where TP, TN, FP and FN represent true positive, true negative, false positive and false negative, respectively.

3. RESULTS

3.1. Parameters Optimization

Based on the features described in Section 2.2 and the "one to one" algorithm, three different multi-classifiers are developed in this study, and each one of them contains 15 SVM models. The substitution

model of the first multi-classifier consists of AAC and PSSM, as called PsePSSM by Shen and Chou [31], and each protein sequence is represented by a 40-dimensional vector. While the substitution model of the second multi-classifier is constructed by combining AAC, PSSM and N-terminal signal peptides, and a 41-dimensional vector is used to describe each protein sequence. With the adding of N-terminal signal peptides, the predictive ability of SVM models for discriminating CSPs from NCSPs can effectively improve, but reduce for identifying different types of CSPs or NCSPs. So 6 SVM models (13, 14, 34, 25, 27, 57) from the first multi-classifier and 9 SVM models (12, 15, 17, 23, 24, 35, 37, 45, 47) from the second multi-classifier, are selected to construct the third hybrid multi-classifier, which is called SecretP v.2.2 in this study. Here, model "12" represents that this model was constructed by using the training sets of T1SPs and T2SPs, and the remaining are known by analogy with it.

All prediction results of the three multi-classifiers were presented in Supplementary **Tables S1-S3**, respectively. As shown in these Supplementary Tables, models constructed by both of CSPs and NCSPs (e.g., 45 and 12) tend to achieve higher accuracies, while those constructed by only CSPs or NCSPs appear to get lower accuracies (e.g., 27 and 34). Comparing the results listed in Supplementary **Table S1** and **Table S2**, it is clear that with the adding of N-terminal signal peptides, the performance of models composed by CSPs and NCSPs (e.g., 23 and 45) slightly improve, while those composed by only CSPs or NCSPs (e.g., 27 and 34) cut down. In view of these factors, SecretP v.2.2 is proposed as described in the previous paragraph, and chosen as the final predictor for distinguishing different types of Gram-negative bacterial secreted proteins.

3.2. Performance on the Independent Data Sets

In order to compare the prediction performance of SecretP v.2.2 with other methods, including the first and the second multi-classifiers described in Section 3.1, and SecretP v.2.1, the test set shown in **Table 1** is used here. All statistical results of the four methods are listed in **Table 2**. From this table, it is clear

Table 2. Prediction results of the four methods obtained by analyzing the test set.

Type	T1SP	T2SP	T3SP	T4SP	T5SP	T7SP	Total
No. of sequences	25	29	28	22	35	33	172
The first multi-classifier							
Correct hit	18	24	28	20	35	30	155
Sensitivity (%)	72.00	82.76	100.00	90.91	100.00	90.91	90.12
The second multi-classifier							
Correct hit	23	24	27	17	35	29	155
Sensitivity (%)	92.00	82.76	96.43	77.27	100.00	87.88	90.12
SecretP v.2.1							
Correct hit	22	23	28	18	35	29	155
Sensitivity (%)	88.00	79.31	100.00	81.82	100.00	87.88	90.12
SecretP v.2.2							
Correct hit	23	25	28	20	35	30	161
Sensitivity (%)	92.00	86.21	100.00	90.91	100.00	90.91	93.60

that SecretP v.2.2 gets the highest total sensitivity of 93.60%, while other three methods achieves the same total sensitivity of 90.12%, but the detailed prediction results of them are different. This indicates that it is a right decision to choose SecretP v.2.2 as the final predictor in this study.

As described in Section 2.1, a public independent dataset is selected to further evaluate the predictive power of SecretP v.2.2 for identifying different types of NCSPs. The comparison results of the four methods for this dataset are listed in Table 3. As shown in Table 3, 86 of the 89 NCSPs are correctly identified by SecretP v.2.2 and SecretP v.2.1, but only 82 are correctly identified by the first and the second multi-classifiers. For the detailed results, SecretP v.2.2 wrongly predicted 2 T1SPs as T5SPs and 1 as a T7SP, while SecretP v.2.1 wrongly predicted 2 T1SPs as T2SPs and 1 as a T5SP [20]. Therefore, the prediction performance of SecretP v.2.2 for identifying NCSPs is comparable to that of SecretP v.2.1.

4. DISCUSSION AND CONCLUSION

A large number of secreted proteins have been discovered from Gram-negative bacteria in recent years, and they are classified into different types according to diverse secretion systems. These proteins play an important role in the interactions between bacteria and host cells, so more and more works have been done for them.

Many computational methods have been proposed to identify secreted proteins so far, but only a very few for distinguishing different types of secreted proteins simultaneously. To address this, SecretP v.2.1 has been developed in our previous work [20]. As an upgraded version of SecretP v.2.1, SecretP v.2.2 is also proposed for this purpose here. The same training and test sets are used to build the two methods, and both of them are constructed based on SVM and the "one to one" algorithm. The biggest difference between them is the feature sets of protein sequences. The substitution model of SecretP v.2.1 contains AAC

Table 3. Prediction results of the four methods obtained by analyzing the independent dataset.

Type	T1SP	T3SP	T4SP	Total
No. of sequences	32	41	16	89
The first multi-classifier				
Correct hit	25	41	16	82
Sensitivity (%)	78.13	100.00	100.00	92.13
The second multi-classifier				
Correct hit	29	40	13	82
Sensitivity (%)	90.63	97.56	81.25	92.13
SecretP v.2.1				
Correct hit	29	41	16	86
Sensitivity (%)	90.63	100.00	100.00	96.63
SecretP v.2.2				
Correct hit	29	41	16	86
Sensitivity (%)	90.63	100.00	100.00	96.63

and auto covariance (AC), and each protein is translated into a 45-dimensional numerical vector. While the substitution models of SecretP v.2.2 consist of AAC, PSSM, with or without N-terminal signal peptides, and a 40-dimensional or 41-dimensional vector is used to represent a protein sequence. The dimension of numerical vectors for SecretP v.2.2 is slightly less than that for SecretP v.2.1, which results in shorter transit times. Moreover, though AC can reflect the neighboring effects between amino acid residues in a protein sequence, it has been confirmed that the parameter *lg* in the equation of AC is not sensitive enough for the classification of different types of secreted proteins [20] [28]. Conversely, PSSM can effectively describe the evolutionary information of amino acid residues in protein sequences, and N-terminal signal peptides are very useful for distinguishing CSPs from NCSPs and NSPs [27]. So comparing with SecretP v.2.1, SecretP v.2.2 seems to be a more reasonable predictor for distinguishing different types of Gram-negative bacterial secreted proteins, and the final results also support this view.

With a comprehensive comparison between SecretP v.2.2 and SecretP v.2.1, several conclusions could be drawn from this study. 1) The evolutionary information of protein sequences can effectively improve the total power of predictors for protein classification. 2) Though N-terminal signal peptides are originally used to distinguish CSPs from non-secreted proteins (NSPs), they also play an important role in the classification of CSPs and NCSPs. 3) The effective feature selection can not only improve the prediction performance of classifiers, but also cut down the dimension of numerical vectors to reduce operation time. 4) The "one to one" algorithm is really good at solving the multi-class classification problem.

Overall, as an improved approach for rapidly and accurately identifying different types of Gram-negative bacterial secreted proteins, SecretP v.2.2 is established in this work, which could be a beneficial supplement for future secretome studies.

ACKNOWLEDGMENTS

This work was supported by grants from The National Natural Science Foundation of China (21305096), The Fund of Science and Technology Department of Guizhou Province (J[2014]2134) and The Development Program for Youth Science and Technology Talents in Education Department of Guizhou Province (KY[2016]219).

REFERENCES

1. Desvaux, M., Hébraud, M., Talon, R. and Henderson, I.R. (2009) Secretion and Subcellular Localizations of Bacterial Proteins: A Semantic Awareness Issue. *Trends in Microbiology*, **17**, 139-145. https://doi.org/10.1016/j.tim.2009.01.004
2. Chagnot, C., Zorgani, M.A., Astruc, T. and Desvaux, M. (2013) Proteinaceous Determinants of Surface Colonization in Bacteria: Bacterial Adhesion and Biofilm Formation from a Protein Secretion Perspective. *Frontiers in Microbiology*, **4**, 303. https://doi.org/10.3389/fmicb.2013.00303
3. Bendtsen, J.D., Kiemer, L., Fausbøll, A. and Brunak, S. (2005) Non-Classical Protein Secretion in Bacteria. *BMC Microbiology*, **5**, 58. https://doi.org/10.1186/1471-2180-5-58
4. Wang, G., Chen, H., Xia, Y., Cui, J., Gu, Z., Song, Y., Chen, Y.Q., Zhang, H. and Chen, W. (2013) How Are the Non-Classically Secreted Bacterial Proteins Released into the Extracellular Milieu? *Current Microbiology*, **67**, 688-695. https://doi.org/10.1007/s00284-013-0422-6
5. Blocker, A., Komoriya, K. and Aizawa, S. (2003) Type III Secretion Systems and Bacterial Flagella: Insights into Their Function from Structural Similarities. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 3027-3030. https://doi.org/10.1073/pnas.0535335100
6. Ding, Z., Atmakuri, K. and Christie, P.J. (2003) The Outs and Ins of Bacterial Type IV Secretion Substrates. *Trends in Microbiology*, **11**, 527-535. https://doi.org/10.1016/j.tim.2003.09.004
7. Konkel, M.E., Kim, B.J., Rivera-Amill, V. and Garvis, S.G. (1999) Bacterial Secreted Proteins Are Required for

the Internalization of Campylobacter Jejuni into Cultured Mammalian Cells. *Molecular Microbiology*, **32**, 691-701. https://doi.org/10.1046/j.1365-2958.1999.01376.x

8. Buttner, D. and Bonas, U. (2003) Common Infection Strategies of Plant and Animal Pathogenic Bacteria. *Current Opinion in Plant Biology*, **6**, 312-319. https://doi.org/10.1016/S1369-5266(03)00064-5
9. Mudrak, B. and Kuehn, M.J. (2010) Specificity of the Type II Secretion Systems of Enterotoxigenic Escherichia Coli and Vibrio Cholerae for Heat-Labile Enterotoxin and Cholera Toxin. *Journal of Bacteriology*, **192**, 1902-1911. https://doi.org/10.1128/JB.01542-09
10. Arnold, R., Brandmaier, S., Kleine, F., Tischler, P., Heinz, E., Behrens, S., Niinikoski, A., Mewes, H.W., Horn, M. and Rattei, T. (2009) Sequence-Based Prediction of Type III Secreted Proteins. *PLoS Pathogens*, **5**, e1000376. https://doi.org/10.1371/journal.ppat.1000376
11. Yang, Y., Zhao, J., Morgan, R.L., Ma, W. and Jiang, T. (2010) Computational Prediction of Type III Secreted Proteins from Gram-Negative Bacteria. *BMC Bioinformatics*, **11**, S47. https://doi.org/10.1186/1471-2105-11-S1-S47
12. Wang, Y., Zhang, Q., Sun, M.A. and Guo, D. (2011) High-Accuracy Prediction of Bacterial Type III Secreted Effectors Based on Position-Specific Amino Acid Composition Profiles. *Bioinformatics*, **27**, 777-784. https://doi.org/10.1093/bioinformatics/btr021
13. Yang, Y. (2012) Identification of Novel Type III Effectors Using Latent Dirichlet Allocation. *Computational and Mathematical Methods in Medicine*, **2012**, Article ID: 696190. https://doi.org/10.1155/2012/696190
14. Sui, T., Yang, Y. and Wang, X. (2013) Sequence-Based Feature Extraction for Type III Effector Prediction. *International Journal of Bioscience, Biochemistry and Bioinformatics*, **3**, 246-251. https://doi.org/10.7763/IJBBB.2013.V3.206
15. Yang, X., Guo, Y., Luo, J., Pu, X. and Li, M. (2013) Effective Identification of Gram-Negative Bacterial Type III Secreted Effectors Using Position-Specific Residue Conservation Profiles. *PLoS One*, **8**, e84439. https://doi.org/10.1371/journal.pone.0084439
16. Yang, Y. and Qi, S. (2014) A New Feature Selection Method for Computational Prediction of Type III Secreted Effectors. *International Journal of Data Mining and Bioinformatics*, **10**, 440-454. https://doi.org/10.1504/IJDMB.2014.064894
17. McDermott, J.E., Corrigan, A., Peterson, E., Oehmen, C., Niemann, G., Cambronne, E.D., Sharp, D., Adkins, J.N., Samudrala, R. and Heffron, F. (2011) Computational Prediction of Type III and IV Secreted Effectors in Gram-Negative Bacteria. *Infection and Immunity*, **79**, 23-32. https://doi.org/10.1128/IAI.00537-10
18. Zou, L., Nan, C. and Hu, F. (2013) Accurate Prediction of Bacterial Type IV Secreted Effectors Using Amino Acid Composition and PSSM Profiles. *Bioinformatics*, **29**, 3135-3142. https://doi.org/10.1093/bioinformatics/btt554
19. Wang, Y., Wei, X., Bao, H. and Liu, S.L. (2014) Prediction of Bacterial Type IV Secreted Effectors by C-Terminal Features. *BMC Genomics*, **15**, 50. https://doi.org/10.1186/1471-2164-15-50
20. Yu, L., Luo, J., Guo, Y., Li, Y., Pu, X. and Li, M. (2013) In Silico Identification of Gram-Negative Bacterial Secreted Proteins from Primary Sequence. *Computers in Biology and Medicine*, **43**, 1177-1181. https://doi.org/10.1016/j.compbiomed.2013.06.001
21. Kampenusa, I. and Zikmanis, P. (2008) Distinctive Attributes for Predicted Secondary Structures at Terminal Sequences of Non-Classically Secreted Proteins from Proteobacteria. *Central European Journal of Biology*, **3**, 320-326. https://doi.org/10.2478/s11535-008-0026-5
22. Restrepo-Montoya, D., Pino, C., Nino, L.F., Patarroyo, M.E. and Patarroyo, M.A. (2011) NClassG+: A Classifier for Non-Classically Secreted Gram-Positive Bacterial Proteins. *BMC Bioinformatics*, **12**, 21.

https://doi.org/10.1186/1471-2105-12-21

23. Luo, J., Yu, L., Guo, Y. and Li, M. (2012) Functional Classification of Secreted Proteins by Position Specific Scoring Matrix and Auto Covariance. *Chemometrics & Intelligent Laboratory Systems*, **110**, 163-167. https://doi.org/10.1016/j.chemolab.2011.11.008
24. Altschul, S.F. and Koonin, E.V. (1998) Iterated Profile Searches with PSI-BLAST—A Tool for Discovery in Protein Databases. *Trends in Biochemical Sciences*, **23**, 444-447. https://doi.org/10.1016/S0968-0004(98)01298-5
25. Petersen, T.N., Brunak, S., von Heijne, G. and Nielsen, H. (2011) SignalP 4.0: Discriminating Signal Peptides from Transmembrane Regions. *Nature Methods*, **8**, 785-786. https://doi.org/10.1038/nmeth.1701
26. Guo, Y., Yu, L., Wen, Z. and Li, M. (2008) Using Support Vector Machine Combined with Auto Covariance to Predict Protein-Protein Interactions from Protein Sequences. *Nucleic Acids Research*, **36**, 3025-3030. https://doi.org/10.1093/nar/gkn159
27. Yu, L., Guo, Y., Zhang, Z., Li, Y., Li, M., Li, G, Xiong, W. and Zeng, Y. (2010) SecretP: A New Method for Predicting Mammalian Secreted Proteins. *Peptides*, **31**, 574-578. https://doi.org/10.1016/j.peptides.2009.12.026
28. Yu, L., Guo, Y., Li, Y., Li, G., Li, M., Luo, J., Xiong, W. and Qin, W. (2010) SecretP: Identifying Bacterial Secreted Proteins by Fusing New Features into Chou's Pseudo-Amino Acid Composition. *Journal of Theoretical Biology*, **267**, 1-6. https://doi.org/10.1016/j.jtbi.2010.08.001
29. Wu, J., Li, M.L., Yu, L.Z. and Wang, C. (2010) An Ensemble Classifier of Support Vector Machines Used to Predict Protein Structural Classes by Fusing Auto Covariance and Pseudo-Amino Acid Composition. *The Protein Journal*, **29**, 62-67. https://doi.org/10.1007/s10930-009-9222-z
30. Chou, K.C. and Shen, H.B. (2007) Recent Progress in Protein Subcellular Location Prediction. *Analytical Biochemistry*, **370**, 1-16. https://doi.org/10.1016/j.ab.2007.07.006
31. Shen, H.B. and Chou, K.C. (2007) Nuc-PLoc: A New Web-Server for Predicting Protein Subnuclear Localization by Fusing PseAA Composition and PsePSSM. *Protein Engineering, Design & Selection*, **20**, 561-567. https://doi.org/10.1093/protein/gzm057

SUPPLEMENTARY

Table S1. Parameter statistics of different SVM models for the first multi-classifier.

Type	C	λ	Weight	Accuracy (%)
12	128	0.03125	10:11	96.2085
13	8	0.125	3:2	95.9184
14	0.5	0.5	1:2	95.4023
15	2	0.03125	10:7	93.1159
17	2	0.03125	3:7	96.2500
23	32	0.03125	9:5	93.2384
24	2	0.5	3:5	93.1677
25	32	0.03125	8:5	92.7757
27	8	0.03125	1:2	83.6735
34	2	0.5	1:3	86.4754
35	2	0.5	10:11	95.0867
37	0.5	0.125	5:19	95.6522
45	8	0.03125	8:3	96.4602
47	2	0.5	7:9	97.2727
57	2	0.125	2:7	97.6415

Note: "12" represents that this model was constructed by using the training sets of T1SPs and T2SPs, and the remaining are known by analogy with it. Different weights were assigned to reduce the data imbalance, which are inversely proportional to the corresponding rates between any two types of secreted proteins in the training set.

Table S2. Parameter statistics of different SVM models for the second multi-classifier.

Type	C	λ	Weight	Accuracy (%)
12	32	0.03125	10:11	97.1564
13	8	0.03125	3:2	96.9388
14	32	0.03125	1:2	98.8506
15	2	0.03125	10:7	94.2029
17	2	0.5	3:7	97.5000
23	2	0.125	9:5	95.3737
24	2	0.03125	3:5	91.9255
25	32	0.03125	8:5	92.0152
27	8	0.03125	1:2	82.9932

Continued

34	8	0.125	1:3	86.0656
35	2	0.5	10:11	96.8208
37	8	0.03125	5:19	96.9565
45	2	0.125	8:3	98.2301
47	2	0.5	7:9	96.3636
57	32	0.03125	2:7	96.6981

Note: "12" represents that this model was constructed by using the training sets of T1SPs and T2SPs, and the remaining are known by analogy with it. Different weights were assigned to reduce the data imbalance, which are inversely proportional to the corresponding rates between any two types of secreted proteins in the training set.

Table S3. **Parameter statistics of different SVM models for the third multi-classifier.**

Type	C	λ	Weight	Accuracy (%)
12	32	0.03125	10:11	97.1564
13	8	0.125	3:2	95.9184
14	0.5	0.5	1:2	95.4023
15	2	0.03125	10:7	94.2029
17	2	0.5	3:7	97.5000
23	2	0.125	9:5	95.3737
24	2	0.03125	3:5	91.9255
25	32	0.03125	8:5	92.7757
27	8	0.03125	1:2	83.6735
34	2	0.5	1:3	86.4754
35	2	0.5	10:11	96.8208
37	8	0.03125	5:19	96.9565
45	2	0.125	8:3	98.2301
47	2	0.5	7:9	96.3636
57	2	0.125	2:7	97.6415

Note: "12" represents that this model was constructed by using the training sets of T1SPs and T2SPs, and the remaining are known by analogy with it. The third multi-classifier also contains 15 SVM models, 6 (models 13, 14, 34, 25, 27, 57) of which are from the first multi-classifier, and 9 (models 12, 15, 17, 23, 24, 35, 37, 45, 47) from the second multi-classifier. Different weights were assigned to reduce the data imbalance, which are inversely proportional to the corresponding rates between any two types of secreted proteins in the training set.