Interdiscip Sci Comput Life Sci (2010) 2: 263–270
DOI: 10.1007/s12539-010-0023-z

Prediction of Non-classical Secreted Proteins Using Informative Physicochemical Properties

Chiung-Hui HUNG[1], Hui-Ling HUANG[1,2], Kai-Ti HSU[1], Shinn-Jang HO[3], Shinn-Ying HO[1,2*]
[1](Institute of Bioinformatics and Systems Biology, National Chiao Tung University, Hsinchu 300, Taiwan)
[2](Department of Biological Science and Technology, National Chiao Tung University, Hsinchu 300, Taiwan)
[3](Department of Automation Engineering, National Formosa University, Yunlin 632, Taiwan)

Received 1 September 2009 / Revised 18 November 2009 / Accepted 18 November 2009

Abstract: The prediction of non-classical secreted proteins is a significant problem for drug discovery and development of disease diagnosis. The characteristic of non-classical secreted proteins is they are leaderless proteins without signal peptides in N-terminal. This characteristic makes the prediction of non-classical proteins more difficult and complicated than the classical secreted proteins. We identify a set of informative physicochemical properties of amino acid indices cooperated with support vector machine (SVM) to find discrimination between secreted and non-secreted proteins and to predict non-classical secreted proteins. When the sequence identity of dataset was reduced to 25%, the prediction accuracy on training dataset is 85% which is much better than the traditional sequence similarity-based BLAST or PSI-BLAST tool. The accuracy of independent test is 82%. The most effective features of prediction revealed the fundamental differences of physicochemical properties between secreted and non-secreted proteins. The interpretable and valuable information could be beneficial for drug discovery or the development of new blood biochemical examinations.
Key words: amino acid index, non-classical secreted protein, SVM prediction.

1 Introduction

Secreted proteins such as cytokines, chemokines and hormones are potential biomarkers for diagnosis or the evaluation of therapeutic efficiency (Damas *et al.*, 2001; Bonin-Debs *et al.*, 2004). They are easy to be detected in body fluids such as serum and urine by non-invasive ways. Discovery of novel human secreted proteins provides possible targets for new drug development and diagnostic technique. With the advances of secretome and proteome researches (Grimmond *et al.*, 2003; Chevallet *et al.*, 2007), the identification of novel secreted proteins has made further progress, but the experimentally confirmed secreted proteins are still limited. Meanwhile, many bioinformatics tools have assisted and accelerated the discovery of unidentified secreted proteins (Klee *et al.*, 2006). After the human genome has been decoded, the prediction tools have became more valuable for their ability to extract information from putative peptide sequences. Many prediction tools were developed to find interest secreted proteins as well (Chen *et al.*, 2003; Bendtsen *et al.*, 2005; Cui *et al.*, 2008; Arnold *et al.*, 2009).

Secreted proteins are secreted from cells into the extracellular space. The secretory process can be classified into two categories: classical and non-classical pathways. In the classical pathway, proteins with signal peptides are processed and transported to the outside of the cells. The signal peptides are usually located in the N-terminal of proteins. In eukaryotic cells, newly translated secreted proteins pass by endoplasmatic reticulum (ER) and Golgi and form secretory vesicles to fuse with the cell membrane (Klumperman, 2000).

In addition to the classical secretory pathway, there is non-classical secretory pathway which is known as leaderless secretion (Nickel, 2003). Many cytokines were proved to be secreted proteins but without signal peptide in their N-terminal, such as FGF-1, FGF-2 and IL-1 (Nickel, 2005). The detailed mechanism how the leaderless proteins were transported was still not clear. Recently, active caspase-1 was found to be a regulator of non-classical secretion (Keller *et al.*, 2008). With the discovery of more and more non-classical secreted proteins, the non-classical pathway seems to play an important role in cell communications. So far, a few groups have attempted to solve this issue and developed tools for the prediction of non-classical proteins. SecretomeP developed by Bendtsen *et al.* is the most well known

*Corresponding author.
E-mail: syho@mail.nctu.edu.tw

tool specifically used for this purpose (Bendtsen *et al.*, 2004b).

Many prediction tools used for subcellular localization such as BaCello, MultiLoc, LOCtree, pTARGET, WoLF PSORT and HSLpred have included the signal peptide or the secreted proteins and regarded as a category of localization (Emanuelsson *et al.*, 2007; Klee and Sosa, 2007). The prediction methods used by these tools include neural network, support vector machine (SVM), Hidden Markov Model (HMM) and k-nearest neighbor (Bendtsen *et al.*, 2004a and 2004b; Pierleoni *et al.*, 2006; Emanuelsson *et al.*, 2007). Most of these tools rely on the signal peptide to determine whether the protein is secreted or not. Apparently, these tools are not suitable for the prediction of leaderless secreted proteins. Besides, the amount of identified non-classical secreted proteins is not sufficient at present for utilizing machine learning approaches. The data sets used for training would involve other secreted proteins with signal peptides consequently.

We believe the existence of differences in the nature of secreted proteins no matter they are with or without signal peptides. Therefore, the data sets we chose were combined with originally leaderless non-classical secreted proteins and signal peptides eliminated classical secreted proteins. The informative physicochemical properties of amino acids indices selected in this study were used as features in designing SVM classifiers. An efficient algorithm inheritable bi-objective genetic algorithm (IBCGA) was used to select significant features which could discriminate the two classes of proteins. The feature sets selected by IBCGA were analyzed carefully to reveal the fundamental differences existed between secreted proteins and non-secreted proteins. In conclusion, we proposed a novel prediction method combining the informative physicochemical properties of amino acid with SVM to solve the prediction problem of non-classical secreted proteins.

2 Methods

2.1 Data sets

The original datasets were downloaded from the website of SecretomeP 1.0 (http://www.cbs.dtu.dk/services/SecretomeP-1.0/datasets.php) (Bendtsen *et al.*, 2004a). There are 3321 positive and 3654 negative samples included in the original data sets. We exclude the sequences with high percentage of similarity to build a classifier which could identify novel secreted proteins according to the physicochemical properties but not the sequence similarity. The sequence identity was reduced to 25% with the PISCES software according to the author's instructions (Wang and Dunbrack, 2003). There are 429 positive and 708 negative samples left after the process of reducing sequence identity. The reduced data sets were further divided into five folds which four fifths were for SVM training and one fifths for the independent test (Table 1). The primary sequences of datasets were transformed into numerical indices with AAindex described in the next section.

Table 1 Statistic of the data set

	Initial	Identity (<25%)	Training	Independent test
Secreted	3321	429	343	86
Non-secreted	3654	708	567	141
total	6975	1137	910	227

2.2 Physicochemical properties

AAindex is a database developed by Kanehisa *et al.* which collects numerical indices representing physicochemical and biochemical properties of amino acids (Kawashima *et al.*, 2008). The 544 properties retrieved from AAIndex 9.0 were reduced to 531 after removing the features containing the value 'NA'. The 531 properties from AAindex were used as initial features to construct SVM classifier for the discrimination between secreted proteins and non-secreted proteins. The original sequences of the datasets were transformed to the numerical indices according to the corresponding values of amino acids of each feature.

The average values of each physicochemical property form a feature vector of the protein sequence. After the transformation of amino acids into numerical indices, these values were normalized to the scale between -1 and 1 for SVM.

2.3 SVM

Support vector machine (SVM) is a commonly used tool to solve the two-class classification problems. The two classes of input datasets were considered as two sets of vectors in the space. SVM constructs a hyperplane which maximizes the margin of two data sets in the high-dimensional space. The redial basis function is used in this work to transform the feature space, defined as follows:

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|), \gamma > 0 \qquad (1)$$

The kernel parameter γ determines how the samples are transformed into a high-dimensional search space. The cost parameter C>0 of SVM adjusts the penalty of total error. The kernel parameter γ and the cost parameter C of SVM must be tuned to get the best performance of the prediction. The SVM is obtained from LIBSVM package version 2.81 (Chang and Lin, 2001).

2.4 Inheritable bi-objective genetic algorithm (IBCGA)

In order to select a minimal number of informative features while maximizing prediction accuracy, we used a previously described inheritable bi-objective genetic

algorithm to solve this problem (Tung and Ho, 2007). IBCGA is an efficient algorithm consisted of an intelligent genetic algorithm which uses orthogonal array crossover to explore the search space efficiently. Moreover, the inheritable mechanism can preserve the features that improve the prediction accuracy during the searching process.

Both feature selection and parameters for tuning SVM were encoded as binary genes for IBCGA. The gene and chromosome are commonly-used terms of genetic algorithm (GA), named GA-gene and GA-chromosome for discrimination in this paper. The GA-chromosome consists of $n = 531$ binary GA-genes bi for selecting informative properties and two 4-bit GA-genes for tuning the parameters C and γ of SVM. If $bi = 0$, the i^{th} property is excluded from the SVM classifier; otherwise, the i^{th} property is included. This encoding method maps the 16 values of γ and C into $\{2^{-7}, 2^{-6}, \cdots, 2^{8}\}$.

The digitized and normalized protein sequences of the training data sets were the input for SVM. The fitness function of SVM is set as the overall accuracy of 5-fold cross validation. The feature selection algorithm of IBCGA is described as follows:

Step 1) (Initialization) Randomly generate an initial population of individuals.

Step 2) (Evaluation) Evaluate the fitness values of all individuals using fitness function.

Step 3) (Selection) Select the winner from two randomly selected individuals to form a mating pool.

Step 4) (Crossover) Select parents from the mating pool to perform orthogonal array crossover on the selected pairs of parents.

Step 5) (Mutation) Apply the swap mutation operator to the randomly selected individuals in the new population. To prevent the best fitness value from deteriorating, mutation is not applied to the best individual.

Step 6) (Termination test) If the stopping condition for obtaining the solution is satisfied, output the best individual. Otherwise, go to Step 2). In this study, the stopping condition is to perform 40 generations.

Step 7) (Inheritance) If $r < r_{end}$, randomly change one bit in the binary GA-genes for each individual from 0 to 1; increase the number r by one, and go to Step 2). Otherwise, stop the algorithm.

In this study, the range of the size of candidate feature set selected by IBCGA is from 10 to 45. 30 independent runs of IBCGA were performed to obtain a robust set of features that can reveal the differences between the two classes of proteins.

2.5 Evaluation of performance

The performance of SVM classifiers was accessed with accuracy, Matthew's correlation coefficient (MCC), sensitivity and specificity, calculated as the following equation:

$$accuracy = \frac{tp + tn}{total} \tag{2}$$

$$MCC = \frac{tp \times tn - fp \times fn}{\sqrt{(tp+fn) \times (tp+fp) \times (tn+fp) \times (tn+fn)}} \tag{3}$$

$$Sensitivity = \frac{tp}{tp + fn} \tag{4}$$

$$Specificity = \frac{tn}{tn + fp} \tag{5}$$

Where *tp*, *tn*, *fp* and *fn* are the number of true positive, true negative, false positive and false negative, respectively.

2.6 Orthogonal experimental design

Statistic design of experiments is a process of planning experiments. Orthogonal experimental design with orthogonal array and factor analysis is an efficient method to analyze the effect of several factors simultaneously (Wu, 1978; Dey, 1985). The factors are the parameters, which affect response variables, and a discriminative value of a factor is regarded as a level of the factor. A "complete factorial" experiment would make measurements at each of all possible level combinations. However, the number of level combinations is often so large that this is impractical, and a subset of level combinations must be judiciously selected to be used, resulting in a "fractional factorial" experiment. Orthogonal experimental design utilizes properties of fractional factorial experiments to efficiently determine the best combination of factor levels to use in design problems.

Orthogonal array is a fractional factorial array, which assures a balanced comparison of levels of any factor. Orthogonal array can reduce the number of level combinations for factor analysis. Each row of an orthogonal array represents the levels of factors in each combination, and each column represents a specific factor that can be changed from each combination. The term "main effect" of one factor designates the effect on response variables that one can trace to a design parameter, which does not bother the estimation of the main effect of another factor. After proper tabulation of experimental results, the summarized data are analyzed using factor analysis to determine the relative level effects of factors.

Factor analysis can evaluate the effects of individual factors on the evaluation function, rank the most effective factors, and determine the best level for each factor such that the evaluation function is optimized. Table 2 shows an illustrative example of orthogonal experimental design using a two-level orthogonal array $\mathrm{L}_M(2^{M-1})$ with M rows and $M-1$ columns. In this example of $M = 8$, there are 7 factors where each corresponds to a physicochemical property and its two levels

Table 2 An illustration example of orthogonal array $\mathbf{L}_8(2^7)$ and factor analysis

t	Factors							Accuracy(%) f_t	Rank
	1	2	3	4	5	6	7		
1	1	1	1	1	1	1	1	28.8	33/128
2	1	1	1	2	2	2	2	18.8	97/128
3	1	2	2	1	1	2	2	28.8	33/128
4	1	2	2	2	2	1	1	17.5	100/128
5	2	1	2	1	2	1	2	20.0	88/128
6	2	1	2	2	1	2	1	41.3	4/128
7	2	2	1	1	2	2	1	33.8	14/128
8	2	2	1	2	1	1	2	20.0	88/128
S_{j1}	93.8	108.8	101.3	111.3	118.8	86.3	121.3		
S_{j2}	115.0	100.0	107.5	97.5	90.0	122.5	87.5		
MED	21.3	8.8	6.3	13.8	28.8	36.3	33.8		
Rank	4	6	7	5	3	1	2		
Better level	2	1	2	1	1	2	1	42.5	1/128

correspond to exclusion and inclusion of the feature in the proposed feature selection. Let f_t denote a function value (prediction accuracy of 10-CV in this study) of the combination t. Define the main effect of factor j with level k as S_{jk} where $j = 1, \cdots, M-1$ and $k = 1$, 2:

$$S_{jk} = \sum f_t \cdot F_t, \quad t = 1, \cdots, M, \tag{6}$$

Where $F_t = 1$ if the level of factor j of combination t is k; otherwise, $F_t = 0$. Since the objective function is to be maximized, the level 1 of factor j makes a better contribution to the function than level 2 of factor j does when $S_{j1} > S_{j2}$. The main effect reveals the individual effect of a factor. After the better one of two levels of each factor is determined, a good combination consisting of all factors with the better levels can be easily reasoned (Ho *et al.*, 2004).

The Rank in Table 2 shows the rank of the combination t in all 128 combinations. In this example, the reasoned combination gets the best accuracy with Rank 1. Notably, the reasoned combination is not guaranteed to be the best one in general cases. The most effective factor j has the largest main effect difference $MED = |Sj1 - Sj2|$. The 6th factor having the largest main effect difference 36.3 is the most effective factor.

2.7 Prediction method

The system flowchart of the prediction method is shown in Fig. 1. The selected m physicochemical properties and the associated parameter set of SVM by using IBCGA are used to predict the test data set. The selected physicochemical properties were analyzed to further understand the special properties that are unique for secreted proteins. The prediction system including four parts described as follows:

1) Numeric Transformation: The sequences in the training data set were transformed into 531-dimensional vectors of numerical values using the AAIndex.

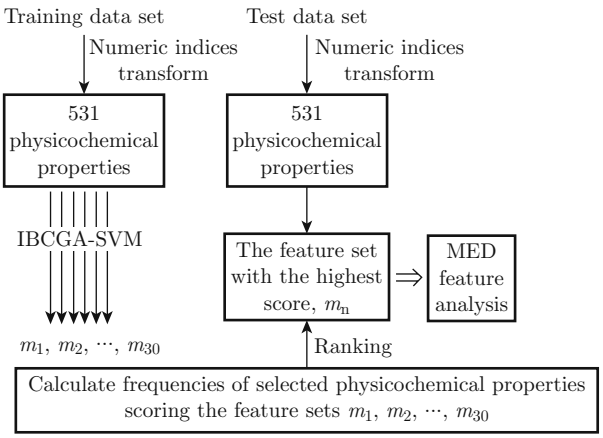

Fig. 1 The system flowchart of the prediction method

2) Feature selection: IBCGA is performed 30 independent runs where each run the training data set is used as the training data set of 5-CV. There are total 30 sets of m physicochemical properties for each of independent data sets.

3) Scoring the feature sets: The frequencies $F(P_i)$ of the selected physicochemical properties in each feature set were added together and then divided by the number of m to obtain the score S_r for each solution.

$$S_r = (\sum_{i=1}^{m} F(P_i))/m \tag{7}$$

4) Independent test: The set of selected physicochemical properties with a maximal value of S_r was used to calculate the performance of the prediction system.

5) Feature analysis: Each feature in the set of selected physicochemical properties with a maximal value of S_r was analyzed by MED analysis to clarify the importance of each feature.

Our method will automatically determine a set of informative physicochemical properties and an SVM-model for prediction of secreted proteins. The robust and informative physicochemical properties were extracted from 30 independent runs of IBCGA to predict the non-classical secreted proteins.

3 Results

3.1 Results of SVM training

The training data sets contain 343 positive and 567 negative samples. The sequence similarity of the training data set is smaller than 25%. We performed 30 independent runs of IBCGA to select robust feature set which could improve the performance of SVM classifier on discriminating the two classes of proteins. The average training accuracy of 30 IBCGA runs were 84.88% and MCC was 0.67 (Table 3). The prediction accuracies for non-secreted protein and secreted protein were 90.16% and 76.17%, respectively.

Table 3 Results of the training and independent test

	Specificity (%)	Sensitivity (%)	Accuracy (%)	MCC
Training Dataset	90.16	76.16	84.88	0.67
Independent test	90.07	68.60	81.94	0.61

3.2 Analysis of the IBCGA selected feature sets

We analyzed the 30 feature sets of independent IBCGA experiments. IBCGA can select m features from 531 physicochemical properties. The result showed that the number of m is between 13 and 43 (data not shown). It is necessary to design a scoring strategy to select the best set from 30 runs. We hypothesize that if a feature is selected by IBCGA repeatedly, it was considered more significant than other features for the classification of the non-secreted and secreted proteins. Based on this hypothesis, we developed an evaluating strategy to choose the best set of features from the 30 sets of features. The frequency of features selected among 30 runs of training experiments was used as the score of each feature to calculate the score of each set of features. The features with high score (>9) are listed in Table 4.

The total score of each set was divided by the number of features of each set to calculate the average score of each set of features. The top 10 feature sets are listed in Table 5. The highest average score S_r was 6.7. The feature set with the highest score contains 20 features which are listed in Table 6.

Furthermore, the feature set with the highest average score was used for the independent test. The result of the independent test was showed in Table 3. The overall accuracy was 81.94%. The accuracies for the non-secreted and secreted proteins were 90.07% and 68.60% respectively and the MCC is 0.61.

3.3 Feature analysis

The contribution of each feature in the feature set with the highest average score was evaluated with two methods. First, each of 20 features was removed to evaluate the reduction of accuracy (Fig. 2). The decreased accuracies were between 11.56% and 12.26%. Besides, the main effect difference (MED) was also used to analyze the importance of each feature (Fig. 3). The reduced accuracies were between 0.49% and 9.84%. The difference between the results of the two methods is due to the combination of features in different number.

Table 4 Results of IBCGA feature selection of 30 runs

Feature ID	Times	Description
420	19	Normalized positional residue frequency at helix termini C" (Aurora-Rose, 1998)
202	17	AA composition of CYT of single-spanning proteins (Nakashima-Nishikawa, 1992)
192	12	Normalized composition of mt-proteins (Nakashima *et al.*, 1990)
221	12	Optimized average non-bonded energy per atom (Oobatake *et al.*, 1985)
196	11	Normalized composition from fungi and plant (Nakashima *et al.*, 1990)
55	10	Normalized hydrophobicity scales for beta-proteins (Cid *et al.*, 1992)
403	10	Normalized positional residue frequency at helix termini N4'(Aurora-Rose, 1998)
13	9	Retention coefficient in HFBA (Browne *et al.*, 1982)
23	9	Free energy of solution in water, kcal/mole (Charton-Charton, 1982)
89	9	Negative charge (Fauchere *et al.*, 1988)
189	9	AA composition of total proteins (Nakashima *et al.*, 1990)
273	9	Weights for beta-sheet at the window position of -4 (Qian-Sejnowski, 1988)

Table 5 Results of feature sets scoring

Run No.	No. of features	Score	Score/No.	Ranking of score/No.
17	20	134	6.70	1
9	21	135	6.43	2
4	15	96	6.40	3
14	19	119	6.26	4
23	20	120	6.00	5
28	22	129	5.86	6
1	28	159	5.68	7
18	26	146	5.62	8
2	13	73	5.62	8
30	30	161	5.37	10

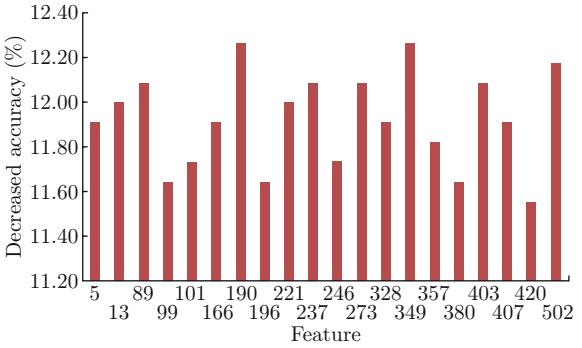

Fig. 2 The significance of each feature selected in run 17 is analyzed by removing each feature orderly to observe the reduction of overall prediction accuracy

Table 6 Selected features of run 17

Feature ID	Times	Description
5	3	Conformational parameter of inner helix (Beghin-Dirkx, 1975)
13	9	Retention coefficient in HFBA (Browne *et al.*, 1982)
89	9	Negative charge (Fauchere *et al.*, 1988)
99	2	Alpha-helix indices for beta-proteins (Geisow-Roberts, 1980)
101	1	Beta-strand indices (Geisow-Roberts, 1980)
166	1	Frequency of occurrence in beta-bends (Lewis *et al.*, 1971)
190	5	SD of AA composition of total proteins (Nakashima *et al.*, 1990)
196	11	Normalized composition from fungi and plant (Nakashima *et al.*, 1990)
221	12	Optimized average non-bonded energy per atom (Oobatake *et al.*, 1985)
237	3	Normalized frequency of turn in alpha+beta class (Palau *et al.*, 1981)
246	7	Surrounding hydrophobicity in turn (Ponnuswamy *et al.*, 1980)
273	9	Weights for beta-sheet at the window position of -4 (Qian-Sejnowski, 1988)
328	4	Relative preference value at N4 (Richardson-Richardson, 1988)
349	7	Information measure for C-terminal turn (Robson-Suzuki, 1976)
357	8	Loss of Side chain hydropathy by helix formation (Roseman, 1988)
380	8	Bitterness (Venanzi, 1984)
403	10	Normalized positional residue frequency at helix termini N4'(Aurora-Rose, 1998)
407	2	Normalized positional residue frequency at helix termini Nc (Aurora-Rose, 1998)
420	19	Normalized positional residue frequency at helix termini C" (Aurora-Rose, 1998)
502	4	Buriability (Zhou-Zhou, 2004)

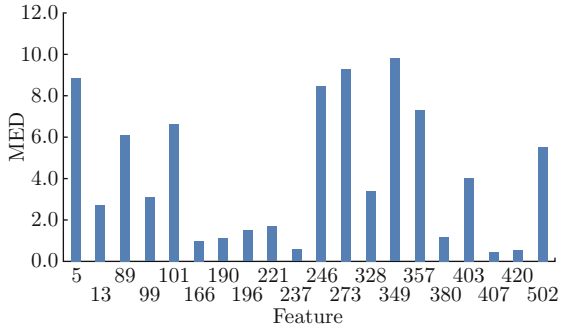

Fig. 3 The significance of each feature selected in run 17 is analyzed by MED analysis

4 Discussion

The advantages of our method are twofold: 1) to predict secreted proteins without high sequence similarity, 2) to extract robust and informative physicochemical properties from the two classes of proteins. Raghava *et al.* have proved that the strikingly high prediction accuracy achieved by BLAST or PSI-BLAST is due to the presence of enough similarity among proteins (Garg and Raghava, 2008). Our algorithm greatly improved the accuracy of SVM classifier when the sequence similarity is smaller than 25%. This advantage of our method is especially crucial for the prediction of non-

classical secreted proteins which are relatively a small fraction of proteome. When the sequence similarity between undiscovered and discovered non-classical secreted proteins is smaller than 25%, the performance of our method will greatly surpass that of other similarity-based tools.

The extracted features from 30 independent IBCGA runs are with strong robustness. Feature 420 and 202 were selected 19 and 17 times respectively. The discrimination ability of only one feature is showed in Fig. 4. Obviously, it is difficult to separate the two classes of proteins according to one feature only. But, one small set of features ($m = 13$) selected by our IBCGA can achieved an accuracy of 82%.

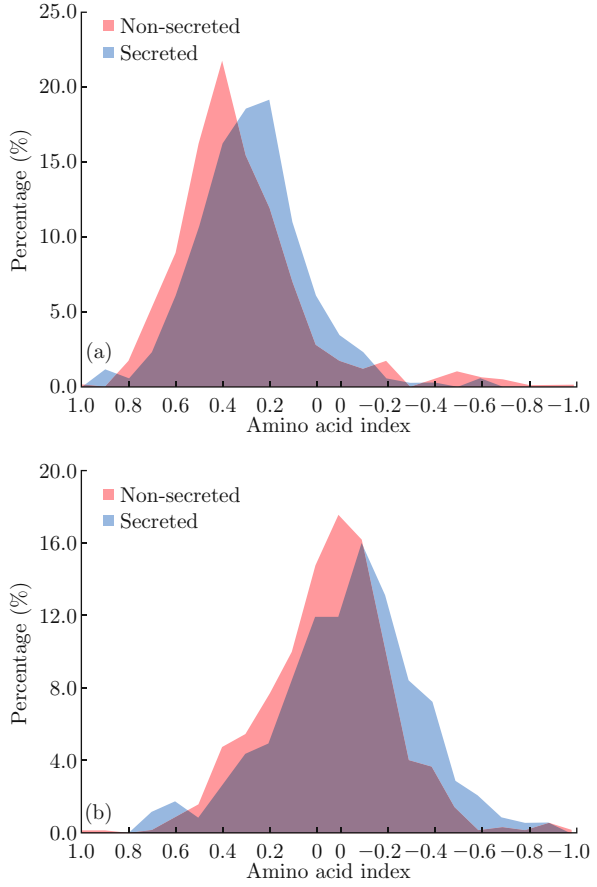

Fig. 4 The distribution of feature (a) 420 and (b) 202. The value is normalized to 1 and −1

Some of the informative physicochemical properties extracted by our algorithm are found to be important in the process of protein secretion (Duong *et al.*, 1996; Tang and Bond, 1998). By the analysis of MED, the feature 349 which describes the Information measure for C-terminal turn is the most important. In contrast to the role that N-terminal signal played in the classical secretory pathway, the C-terminal signal may play a more important role in the non-classical secretory pathway. The extracted properties are also relatively more interpretable for biologists. Some of these informative features need to be investigated more detailedly.

IBCGA is an efficient algorithm to select informative physicochemical properties for SVM classifier. With the IBCGA-selected features, the prediction accuracy of our method is better than the existing method. This method can be also applied to other sequence-based prediction problems.

Acknowledgments The authors would like to thank the National Science Council of Taiwan for financially supporting this research under the contract numbers NSC 96-2628-E-009-141-MY3 and NSC 97-2627-B-009-005.

References

[1] Arnold, R., Brandmaier, S., Kleine, F., Tischler, P., Heinz, E., Behrens, S., Niinikoski, A., Mewes, H.W., Horn, M., Rattei, T. 2009. Sequence-based prediction of type III secreted proteins. PLoS Pathog 5, e1000376.

[2] Bendtsen, J.D., Jensen, L.J., Blom, N., von Heijne, G., Brunak, S. 2004a. Feature-based prediction of non-classical and leaderless protein secretion. Protein Eng Des Sel 17, 349–356.

[3] Bendtsen, J.D., Nielsen, H., von Heijne, G., Brunak, S. 2004b. Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340, 783–795.

[4] Bendtsen, J.D., Binnewies, T.T., Hallin, P.F., Sicheritz-Ponten, T., Ussery, D.W. 2005. Genome update: Prediction of secreted proteins in 225 bacterial proteomes. Microbiology 151, 1725–1727.

[5] Bonin-Debs, Boche, I., Gille, H., Brinkmann, U. 2004. Development of secreted proteins as biotherapeutic agents. Expert Opin Biol Ther 4, 551–558.

[6] Chang, C.C., Lin, C.J. 2001. LIBSVM: A library for support vector machines. http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[7] Chen, Y., Yu, P., Luo, J., Jiang, Y. 2003. Secreted protein prediction system combining CJ-SPHMM, TMHMM, and PSORT. Mamm Genome 14, 859–865.

[8] Chevallet, M., Diemer, H., van Dorssealer, A., Villiers, C., Rabilloud, T. 2007. Toward a better analysis of secreted proteins: The example of the myeloid cells secretome. Proteomics 7, 1757–1770.

[9] Cui, J., Liu, Q., Puett, D., Xu, Y. 2008. Computational prediction of human proteins that can be secreted into the bloodstream. Bioinformatics 24, 2370–2375.

[10] Damas, J.K., Gullestad, L., Aukrust, P. 2001. Cytokines as new treatment targets in chronic heart failure. Curr Control Trials Cardiovasc Med 2, 271–277.

[11] Dey, A. 1985. Orthogonal Fractional Factorial Designs. Wiley, New York.

[12] Duong, F., Lazdunski, A., Murgier, M. 1996. Protein secretion by heterologous bacterial ABC-transporters: the C-terminus secretion signal of the secreted protein confers high recognition specificity. Mol Microbiol 21, 459–470.

[13] Emanuelsson, O., Brunak, S., von Heijne, G., Nielsen, H. 2007. Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc 2, 953–971.

[14] Garg, A., Raghava, G.P. 2008. A machine learning based method for the prediction of secretory proteins using amino acid composition, their order and similarity-search. In Silico Biol 8, 129–140.

[15] Grimmond, S.M., Miranda, K.C., Yuan, Z., Davis, M.J., Hume, D.A., Yagi, K., Tominaga, N., Bono, H., Hayashizaki, Y., Okazaki, Y., RIKEN GER Group, GSL Members, Teasdale, R.D. 2003. The mouse secretome: Functional classification of the proteins secreted into the extracellular environment. Genome Res 13, 1350–1359.

[16] Ho, S.Y., Shu, L.S., Chen, J.H. 2004. Intelligent evolutionary algorithms for large parameter optimization problems. IEEE Transactions on Evolutionary Computation 8, 522–541.

[17] Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., Kanehisa, M. 2008. AAindex: Amino acid index database, progress report 2008. Nucleic Acids Res 36, D202–205.

[18] Keller, M., Ruegg, A., Werner, S., Beer, H.D. 2008. Active caspase-1 is a regulator of unconventional protein secretion. Cell 132, 818–831.

[19] Klee, E.W., Sosa, C.P. 2007. Computational classification of classically secreted proteins. Drug Discov Today 12, 234–240.

[20] Klee, E.W., Finlay, J.A., McDonald, C., Attewell, J.R. Hebrink, D., Dyer, R., Love, B., Vasmatzis, G., Li, T.M., Beechem, J.M., Klee, G.G. 2006. Bioinformatics methods for prioritizing serum biomarker candidates. Clin Chem 52, 2162–2164.

[21] Klumperman, J. 2000. Transport between ER and Golgi. Curr Opin Cell Biol 12, 445–449.

[22] Nickel, W. 2003. The mystery of nonclassical protein secretion. A current view on cargo proteins and potential export routes. Eur J Biochem 270, 2109–2119.

[23] Nickel, W. 2005. Unconventional secretory routes: Direct protein export across the plasma membrane of mammalian cells. Traffic 6, 607–614.

[24] Pierleoni, A., Martelli, P.L., Fariselli, P., Casadio, R. 2006. BaCelLo: A balanced subcellular localization predictor. Bioinformatics 22, e408–e416.

[25] Tang, J., Bond, J.S. 1998. Maturation of secreted meprin alpha during biosynthesis: role of the furin site and identification of the COOH-terminal amino acids of the mouse kidney metalloprotease subunit. Arch Biochem Biophys 349, 192–200.

[26] Tung, C.W., Ho, S.Y. 2007. POPI: Predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties. Bioinformatics 23, 942–949.

[27] Wang, G., Dunbrack, J.R.J. 2003. PISCES: A protein sequence culling server. Bioinformatics 19, 1589–1591.

[28] Wu, Q. 1978. On the optimality of orthogonal experimental design. Acta Math Appl Sinica 1, 283–299.