Journal of Theoretical Biology 312 (2012) 105–113

Contents lists available at SciVerse ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Ranking Gene Ontology terms for predicting non-classical secretory proteins in eukaryotes and prokaryotes

Wen-Lin Huang *

Department of Management Information System, Asia Pacific Institute of Creativity, No. 110 XueFu Rd., Tou Fen, Miaoli, Taiwan, ROC

HIGHLIGHTS

- ▸ Use Gene Ontology (GO) terms as the only one type of input features.
- ▸ Identify two small sets of 436 and 158 GO terms for eukaryotes and prokaryotes.
- ▸ The Sec-GO method performs better (96.7%) than SPRED (82.2%) in eukaryotes.
- ▸ The Sec-GO method performs better (94.5%) than NClassG+ (90.0%) in prokaryotes.

ARTICLE INFO

Article history:
Received 8 January 2012
Received in revised form
30 May 2012
Accepted 28 July 2012
Available online 8 August 2012

Keywords:
Secretion prediction
Support vector machine
Signal peptide

ABSTRACT

Protein secretion is an important biological process for both eukaryotes and prokaryotes. Several sequence-based methods mainly rely on utilizing various types of complementary features to design accurate classifiers for predicting non-classical secretory proteins. Gene Ontology (GO) terms are increasing informative in predicting protein functions. However, the number of used GO terms is often very large. For example, there are 60,020 GO terms used in the prediction method Euk-mPLoc 2.0 for subcellular localization. This study proposes a novel approach to identify a small set of m top-ranked GO terms served as the only type of input features to design a support vector machine (SVM) based method Sec-GO to predict non-classical secretory proteins in both eukaryotes and prokaryotes. To evaluate the Sec-GO method, two existing methods and their used datasets are adopted for performance comparisons. The Sec-GO method using $m=436$ GO terms yields an independent test accuracy of 96.7% on mammalian proteins, much better than the existing method SPRED (82.2%) which uses frequencies of tri-peptides and short peptides, secondary structure, and physicochemical properties as input features of a random forest classifier. Furthermore, when applying to Gram-positive bacterial proteins, the Sec-GO with $m=158$ GO terms has a test accuracy of 94.5%, superior to NClassG+ (90.0%) which uses SVM with several feature types, comprising amino acid composition, di-peptides, physicochemical properties and the position specific weighting matrix. Analysis of the distribution of secretory proteins in a GO database indicates the percentage of the non-classical secretory proteins annotated by GO is larger than that of classical secretory proteins in both eukaryotes and prokaryotes. Of the m top-ranked GO features, the top-four GO terms are all annotated by such subcellular locations as GO:0005576 (Extracellular region). Additionally, the method Sec-GO is easily implemented and its web tool of prediction is available at iclab.life.nctu.edu.tw/secgo.

1. Introduction

Both eukaryotic and prokaryotic cells have highly evolved secretion processes. The primary route for protein secretion from eukaryotic cells is called the classical or endoplasmic reticulum (ER)/Golgi-dependent secretory pathway (Nickel, 2003; Radisky et al., 2009). Secreted eukaryotic proteins typically contain short N-terminal signal peptides that direct them to the translocation apparatus of the ER. However, several secretory proteins that lacks signal peptides, such as fibroblast growth factors (FGF-1 and FGF-2), HMGB1, interleukins (IL-1β), hydrophilic acylated surface protein B (HASPB) and galectins, are exported by distinct non-classical secretion pathways (Nickel, 2003; Prudovsky et al., 2003; Radisky et al., 2009).

Secretion is not unique to eukaryotes; it is also present in bacteria. Bacterial secretion of proteins is via highly complex translocation machineries that actively move proteins to be secreted across the

* Tel.: +886 37 605673.
E-mail addresses: wenlinhuang2001@yahoo.com.tw, wenlinhuang2001@gmail.com

0022-5193/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jtbi.2012.07.027

bacterial cytoplasmic membrane (*i.e.* the classical secretion pathway) (Bendtsen and Wooldridge, 2009). However, many proteins secreted via alternative routes (i.e. the non-classical secretion pathway) are involved in pathogenesis (Bendtsen and Wooldridge, 2009). Six secretion systems which transport proteins across the cytoplasmic membrane have been identified in Gram-positive bacteria, secretion (Sec), twin-arginine translocation (Tat), flagella export apparatus (FEA), fimbrilin-protein exporter (FPE), hole-forming (holin), and WXG100 secretion system (Wss) (Desvaux and Hébraud, 2006). Numerous bacterial proteins that are released via the Sec and Tat secretion pathways can be secreted without N-terminal signal peptides and are also called non-classically secreted proteins, such as proteins released via Wss in Gram-positive bacteria (Bendtsen et al., 2005a; Desvaux and Hébraud, 2006).

Several sequence-based methods using hybrid feature types have been developed to predict proteins secreted via non-classical pathways (Bendtsen et al., 2004, 2005b; Garg and Raghava, 2008; Hung et al., 2010; Kandaswamy et al., 2010; Yu et al., 2010) (Table 1). SecretomeP uses neural networks (NNs) with various sequence-derived features comprising the number of atoms, number of positively charged residues, low complexity regions, transmembrane helices, propeptide cleavage sites and subcellular localization to predict non-classical secretion in mammals (Bendtsen et al., 2004). In addition to these feature types, SecretomeP also integrates additional feature types, such as amino acid composition (AAC), secondary structure and disordered regions, and uses an artificial NN (ANN) to predict non-classical secretory proteins from Gram-positive bacteria and Gram-negative bacteria (Bendtsen et al., 2005b).

The SRTpred method uses a hybrid approach to integrate a PSI-BLAST module and support vector machine (SVM), which uses AAC and dipeptide composition as input features (Garg and Raghava, 2008). The SPRED method uses an information gain algorithm with the ranking method to select the 50 top-ranked features from 119 sequence-based features including frequencies of tri-peptides and short peptides, the secondary structure, and physicochemical properties (PCPs) (Kandaswamy et al., 2010). The SecretP2.0 method fuses AAC, auto-covariance, and pseudo-AAC (PseAAC) with SVM to predict bacterial l secretory proteins (Yu et al., 2010). A novel method NClassG+ utilizes SVM with various sequence transformation vectors, frequencies, di-peptides, physicochemical factors, and the position specific weighting matrix (PSSM) to predict non-classically secreted Gram-positive bacterial proteins (Restrepo-Montoya et al., 2011).

These methods combine various types of complementary features in designing accurate classifiers. Conversely, one SVM-based method uses a single type of PCP features to predict non-classical secretory proteins (Hung et al., 2010), where the set of informative PCPs is identified by utilizing a high-performance feature selection algorithm (Ho et al., 2004). Due to different design aims, feature selection, classifiers and datasets used, determining which feature type is the most effective in classification is extremely difficult. However, this study aims to propose a novel and highly-effective feature type to predict non-classical secretory proteins in eukaryotes and prokaryotes.

The Gene Ontology (GO) is a controlled vocabulary used to describe the biology of a gene product in any organism (Ashburner et al., 2000). The GO annotations have three structured and controlled vocabularies (*i.e.* ontologies) that characterize individual gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner. The Plant-associated Microbe Gene Ontology (PAMGO) Consortium (Torto-Alalibo et al., 2009) has developed standardized terms for describing biological processes and cellular components that play important roles in the interactions between microbes and plant and animal hosts, including bacterial secretion processes (Tseng et al., 2009). Hence, the GO annotations have a high potential in improve prediction performance when identifying non-classically secreted proteins in eukaryotes and prokaryotes.

Notably, GO annotation has been used successfully to solve in various sequence-based prediction problems and to extract many other important features of proteins, such as protein subcellular localization (Chou and Shen, 2010a; Chou et al., 2011, 2012; Shen and Chou, 2010), enzyme classification (Chou and Cai, 2004), membrane protein type (Chou and Cai, 2005b), and protein–protein interaction (Chou and Cai, 2005a). However, proteins are often represented as high-dimensional vectors of n binary features, where n is the total number of GO terms in a complete annotation set (a component of 1 is assigned when the annotation is hit and 0 is assigned otherwise). For example, 60,020 GO terms are used in Euk-mPLoc 2.0 (Chou and Shen, 2010a) and Gneg-mPLoc (Shen and Chou, 2010), both of which use PseAAC and GO terms with ensemble classifiers to predict proteins in multiple subcellular locations.

Additionally, each gene product is generally annotated by only few GO terms, which results in long and sparse vectors and renders the clustering algorithm problematic (Popescu et al., 2006). Therefore, this study proposes a novel approach, namely Sec-GO, to identify a small set of m top-ranked GO term features for non-classical secretory protein prediction where $m \ll n$. The n GO terms are ranked according to their scores—a score is the difference in the occurrence frequencies of the GO term between positive and negative datasets. The number m is determined using the number of GO terms with scores exceeding the mean of n scores.

To evaluate the proposed Sec-GO method, two existing methods, SPRED (Kandaswamy et al., 2010) and NClassG+ (Restrepo-Montoya et al., 2011) as well as their datasets, ES_SPRED and PS, respectively, are adopted for performance comparisons. An additional mammalian dataset, ES from ES_SPRED, is established to have a sequence identity of 25%. Using this ES dataset, the Sec-GO method identifies $m=501$ GO terms and obtains a test accuracy of 96.8%. Additionally, the Sec-GO method using $m=436$ top-rank GO terms yields an independent test accuracy of 96.7% on ES_SPRED, better than that of SPRED which has an accuracy of 82.2%. Compared with the NClassG+ method, which has a test

Table 1
Sequenced-based methods with the hybrid feature types and classifiers for predicting non-classical secretory proteins.

Method	Feature types	Classifier
SecretomeP 2.0 (2004, 2005)	Number of atoms, number of positively charged residues, low complexity regions, transmembrane helices, propeptide cleavage sites and subcellular localization	Neural networks
SRTpred (2008)	AAC, AAC order, and PSI-BLAST similarity search	Artificial neural network and SVM
SPRED (2010)	Frequencies of tri-peptides and short peptides, secondary structure and PCPs	Random forest classifier
SecretP 2.0 (2010)	AAC and PCP	SVM
NClassG+ (2011)	AAC, di-peptides, PCP and PSSM	SVM
Sec-GO (this study)	GO terms	SVM

accuracy of 90.0%, the proposed Sec-GO method yields a higher accuracy of 94.5% on the Gram-positive bacterial dataset PS when $m=158$ GO term features are used. Analysis of the distribution of secretory proteins in a GO database shows that the number of non-classical secretory proteins annotated by GO is larger than that of classical secretory proteins in both eukaryotes and prokaryotes. Of the $m=501$ and 158 GO terms, the top-four GO terms are all annotated by such subcellular locations as GO:0005576 (Extracellular region) and GO:0005737 (Cytoplasm).

The high prediction performance of Sec-GO arises mainly from increasingly informative GO terms utilized with the proposed GO identification approach. A web-based prediction server of Sec-GO can be found at iclab.life.nctu.edu.tw/secgo. Using 19 experimentally verified non-classical secretory mammalian proteins, this prediction server Sec-GO can obtain 18 true predictions, which is higher than those by SecretomeP (12) and SRTpred (5), and SPRED (15).

2. Materials and methods

According to a recent comprehensive review (Chou, 2011), to establish a useful statistical predictor for a protein system, the following procedures must be used: (1) construct or select a valid benchmark dataset to train and test the predictor; (2) formulate the protein samples with an effective mathematical expression that truly reflects their intrinsic correlation with the attribute to be predicted; (3) introduce or develop a powerful algorithm (or engine) for prediction; (4) perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; and (5) establish a user-friendly web-server for the predictor that can be accessed by the public.

A block diagram is used to simplify the steps of the Sec-GO method (Fig. 1). The block diagram consists of the following parts: datasets; homolog-based GO extraction; scores of GO terms; constructing the training model; and prediction of query secretory sequences. The models with the highest predictive accuracy are tested on an independent dataset. Parameters and training features that provide the best predictive performance are used to implement the web-based system. Each part is described in detail as follows.

2.1. Datasets

This work uses one eukaryotic secretory dataset ES and one prokaryotic secretory dataset PS (Restrepo-Montoya et al., 2011) to evaluate the proposed Sec-GO method. This work establishes the ES dataset with a sequence identity of 25% from ES_SPRED (Kandaswamy et al., 2010) using the PISECS culling application (Li and Godzik, 2006) to avoid homolog bias and overestimation. The ES dataset has 372 non-classical (positive dataset) and 1011 classical (negative dataset) secretory proteins in mammals (Table 2). The ES_SPRED dataset is adopted for performance comparison with the SPRED method (Kandaswamy et al., 2010), in which protein sequences are taken from the Swiss-Prot (Bairoch and Apweiler, 2000) protein sequence database according to the annotation information in the CC (comment or notes) and ID (identification) fields. Proteins in the dataset are collected confidently using the following criteria: (1) only the sequences annotated with ‘‘mammalian’’ in their ID field are collected; (2) sequences with uncertain annotation labels such as ‘‘probable’’, ‘‘potential’’ or ‘‘by similarity’’ are removed; (3) sequences annotated with keywords ‘‘extracellular’’ are collected as the positive dataset; (4) signal peptides are removed from the positive dataset, and (5) sequences annotated in cytoplasm and/or nucleus subcellular locations are taken as the negative dataset.

The PS dataset with a 25% sequence identity has 501 non-classical (positive dataset) and 696 classical (negative dataset) secretory proteins in Gram-positive bacteria (Restrepo-Montoya et al., 2011). These bacterial proteins in the datasets are adopted by strictly applying the following criteria: (1) the positive dataset only comprises proteins whose annotation in Swiss-Prot (Bairoch and Apweiler, 2000) has the words ‘‘signal’’, ‘‘secreted’’, ‘‘extracellular’’, ‘‘periplasmic’’, ‘‘periplasm’’, ‘‘plasma membrane’’, ‘‘integral membrane’’ or ‘‘single pass membrane’’; (2) the sequence portions corresponding to the translocation mechanism (first region between position 1 up to a varying point that ranges between amino acids 21 and 55) are manually removed based on the annotation reported in Swiss-Prot, and (3) the negative protein dataset comprises proteins whose annotation contains the words ‘‘cytoplasm’’ or ‘‘cytoplasmic’’. The PS dataset is divided into two datasets, PSL for learning (i.e. training) and PST for independent testing (Restrepo-Montoya et al., 2011). The learning

Table 2
The numbers of proteins in the datasets ES and PS. The number g of (g) in the training datasets represents the number of sequences which are correctly annotated by BLAST, where (h, e)=(5, 10^{-9}) are used for the ESL and PSL.

Class	Numbers of secretory proteins					PST
	ESL	(g)	EST	PSL	(g)	
Positive	186	(11)	186	419	(156)	82
Negative	506	(302)	505	433	(366)	263
Total	692	(313)	691	852	(392)	345

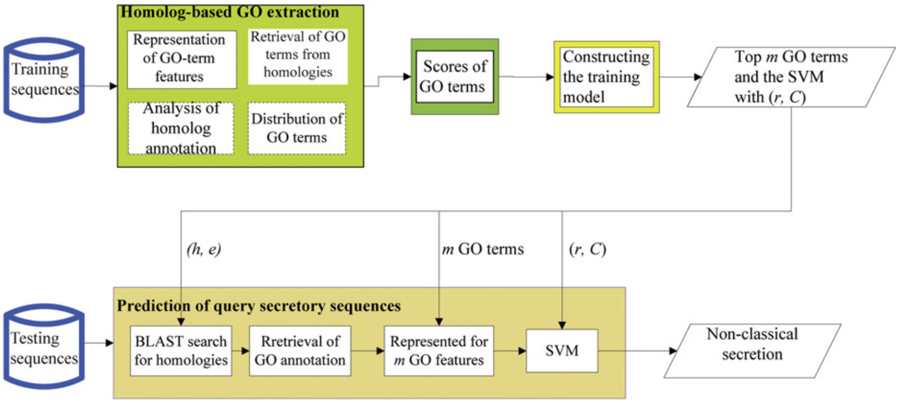
Homolog-based GO extraction
Training sequences
Representation of GO-term features
Retrieval of GO terms from homologies
Analysis of homolog annotation
Distribution of GO terms
Scores of GO terms
Constructing the training model
Top m GO terms and the SVM with (r, C)
(h, e)
m GO terms
(r, C)
Prediction of query secretory sequences
Testing sequences
BLAST search for homologies
Rretrieval of GO annotation
Represented for m GO features
SVM
Non-classical secretion

Fig. 1. Block diagram of the proposed Sec-GO method. The five main blocks are datasets, homolog-based GO extraction, scores of GO terms, construction of training model, and prediction of query secretory sequences.

dataset is done with the purpose of estimating scores of GO terms, identifying a small set of GO terms and finding the best parameters of a SVM to train the complete dataset. Similarly, the ES dataset is divided into two parts, ESL for training and EST for independent testing. Table 2 lists the numbers of proteins in each dataset.

2.2. Homolog-based GO extraction

To identify novel non-classical secretory proteins, the proposed Sec-GO method only uses sequences of query proteins without their accession numbers. To extract GO terms from a GO database, the accession number is indispensable. Thus, this work first uses BLAST (Altschul et al., 1990; Altschul et al., 1997) to find good homologies of a query protein from the Swiss-Prot database (version 55.3). The accession numbers of the homologies are then applied to the GO database (released on August 25, 2011) at www.ebi.ac.uk/GOA/ to retrieve annotated GO terms. Finally, the query protein sequence is represented for GO-term feature vectors.

2.2.1. Representation of GO-term features

According to Eq.(6) of a recent comprehensive review (Chou, 2011), the general form of Chou′s PseAAC can be formulated a query protein P as

$$P=[\Psi_1 \quad \Psi_2 \quad \cdots \quad \Psi_i \quad \cdots \quad \Psi_n]^T \tag{1}$$

where T is a transpose operator, operator, while the subscript n reflects the dimension of the vector and its value as well as the components ψ_1, ψ_2, … will be defined by a series of feature extractions as elaborated below. Thus, when using the general formulation of PseAAC to incorporate the GO information, there are n GO terms, i.e.

$$P=[\Psi_1^G \quad \Psi_2^G \quad \cdots \quad \Psi_i^G \quad \cdots \quad \Psi_n^G]^T \tag{2}$$

The variable $\psi_i^G=1$ if a hit is found against the ith GO term for the protein P; otherwise, $\psi_i^G=0$, $i=1, 2, \ldots, n$. For example, the query protein sequence in Euk-mPLoc 2.0 (Chou and Shen, 2010a) are represented for $n=60{,}020$ GO-term feature vector.

2.2.2. Retrieval of GO terms from homologies

Two threshold parameters of BLAST, the E-value and the number of homologies h, have influence in the quality of homolog GO term retrieval. This work finds $(h, e)=(5, 10^{-9})$ are good trade-off values after using the k-nearest-neighbor classifier (k-NN) to test from $e\in\{10^0, 10^{-1}, 10^{-2}, \ldots, 10^{-10}\}$ and $h\in\{1, \ldots, 5\}$, where prediction accuracies are 90.6% and 85.7% for the ESL and PSL datasets, respectively (Fig. 2). The parameter value $e=10^{-9}$ may cause that two proteins are thought to be homologous when they are only $>60\%$ pair-wise identical. The protein sequences may be very few; thus, the k-NN classifier can obtain a prediction accuracy $>85\%$.

2.2.3. Analysis of homolog annotation

Some secretory proteins in the ES and PS datasets are derived according to the sequences annotated in such subcellular locations as "extracellular", "periplasm", "cytoplasm" and "nucleus" (Table 3). Therefore, the analysis of the subcellular location annotations of homologies is important when using BLAST to find homologies of query secretory proteins and retrieve their corresponding GO terms. First, this work finds all homologies for each training sequence in datasets by using BLAST with $(h, e)=(5, 10^{-9})$. This work then investigates the annotation information of homologies (subcellular location) in the "General annotation" (Comments) field. If one homology is annotated in the same subcellular location as the query sequence, a true prediction is obtained; otherwise, a false prediction is obtained.

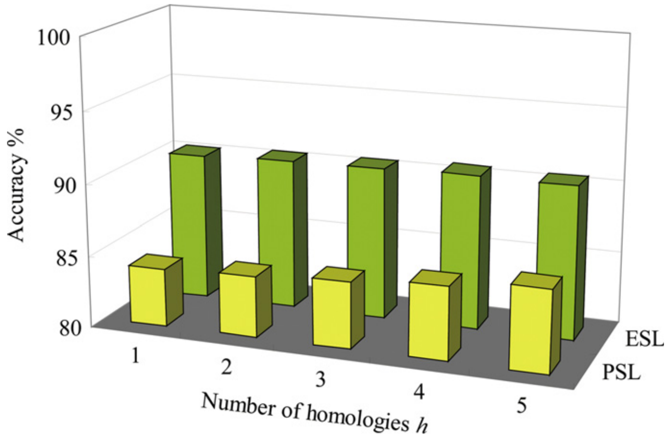

Fig. 2. Accuracy performance with different number of homologies. The highest accuracies 90.6% and 85.7% are obtained by using k-NN with $k=1$ and $h=5$ for ESL and PSL, respectively.

Table 3
Subcellular locations annotated in Swiss-Prot protein database.

Class	ESL	PSL
Positive	Extracellular	Extracellular, secreted, periplasm
Negative	Cytoplasm, nucleus	Cytoplasm

Take one secretory protein HES1_MOUSE (P35428) as an example. This protein sequence, which is annotated in the "nucleus" location, corresponds to five homologies, P35428, Q04666, Q14469, Q3ZBG4 and Q5PPM5 by using BLAST with $(h, e)=(5, 10^{-9})$, where the pair-sequence identity exceeds 90%. Homolog P35428 is removed out due to all query proteins are regarded as unknown in this work. The remaining four homologies are annotated in the subcellular location "nucleus" through retrieving the Swiss-Prot database, such that a true prediction is obtained for HES1_MOUSE (P35428). This prediction rule is overestimated because some secretory proteins may be annotated in multiple subcellular locations (Chou and Shen, 2010a).

The number g of (g) in Table 2 represents the number of true predictions obtained by using the above-mentioned subcellular location annotation. The percentages of true predictions are 45.2% (=313/692) and 46.0% (=392/852) for the ESL and PSL datasets, respectively. Roughly, 50% of the training datasets can be correctly distinguished between classical and non-classical secretions. These analytical results reveal that this homolog search using BLAST plays an important role in accurate prediction of non-classical secretion; however, the homolog search cannot be used alone when designing prediction methods. This fact motivates this work to design a computational method based on the homology search rather than only using this homology search to identify non-classical secretion.

2.2.4. Distribution of GO terms

Tables 4 and 5 show the sizes of the complete sets of all GO terms from the ESL and PSL datasets, where $n=2741$ and 1039, respectively. The percentage of training proteins obtained using no annotated GO term (*i.e.*, $t=0$) is 6.79% (=47/692) and 21.2% (=181/852), where 6.5% (=47/692) and 18.3% (=181/852) of proteins without homologies are included in the ESL and PSL datasets, respectively. Specifically, the percentage 14.52% (=27/186) in the positive dataset is much larger than 3.95% (=20/506) in the ESL dataset. The same situation occurs for the PSL dataset, *i.e.*, a large gap exists between 38.66% (=162/419) and 4.39% (=19/433) for the positive and negative datasets, respectively. Analytical results indicate that a possibility exists that a non-classical secretion annotated by GO is larger than that of a classical secretion in both eukaryotes and prokaryotes.

Additionally, the mean number of GO terms obtained for individual proteins is much smaller than n, i.e.,$13.8 \ll 2741$ and $6.2 \ll 1039$ for the ESL and PSL datasets, respectively (Tables 4 and 5). This finding indicates that feature vectors are long and sparse, making the clustering rather problematic (Popescu et al., 2006). Therefore, this work aims to confine the huge n-dimensional search space by identifying m characteristic GO terms with discriminative ability to solve this clustering problem, where $m \ll n$.

2.3. Scores of GO terms

This proposed Sec-GO method uses m top-ranked GO term features with scores based on the difference between occurrence frequencies for each GO term, described as follows.

Table 4
Analysis results of GO annotation for the dataset ESL. The number t is the number of GO terms obtained from one training sequence. The mean number of GO terms is *Mean*=13.8.

Class	n=2741 GO terms			
	t=0	$1 \le t <$ *Mean*	*Mean* $< t$	*Mean*
Positive	27	95	64	13.2
Negative	20	304	182	14.1
Total	47	399	246	13.8

Table 5
Analysis results of GO annotation for the dataset PSL. The number t is the number of GO terms obtained from one training sequence. The mean of GO terms is 6.2 and denoted as *Mean*.

Class	n=1039 GO terms			
	t=0	$1 \le t <$ *Mean*	*Mean* $< t$	*Mean*
Positive	162	14	243	4.4
Negative	19	113	301	8.0
Total	181	127	544	6.2

Step 1: Let n be the number of GO terms obtained by all training proteins. Calculate the occurrence numbers in each class for each GO term. For example, the occurrence numbers of GO:0005737 in the positive and negative classes of the ESL dataset are 19 and 334, respectively.
Step 2: Calculate the occurrence frequency for each GO term by dividing the occurrence numbers using the total occurrence numbers of n GO terms in each class. For example, the total occurrence numbers of n GO terms in the positive and negative classes are 2461 and 7117, respectively. Therefore, occurrence frequencies of GO:0005737 are 0.007720 and 0.046930, respectively.
Step 3: For each GO term, the score is the absolute value of the difference between the occurrence frequencies in the positive and negative classes. For example, the score of GO:0005737 is 0.039210 ($=|0.007720-0.046930|$).
Step 4: Normalize scores of all n GO terms into the range [0, 1], denoted as $\{s_1, s_2, \ldots, s_n\}$. The normalized score of the term GO:0005737 is 0.628.

2.4. Constructing the training model

Designing accurate prediction methods aims to find feature vectors with high discrimination abilities for classifying positive and negative samples. The proposed computational method Sec-GO performs well by using an SVM with a set of m top-ranked GO term features. The number m is determined by analyzing n score of GO terms. First, this work estimates the mean score and then all n scores of GO terms in decreasing order. Second, the n GO terms are divided into increment-based sets according to the mean score. Finally, the SVM is used with each set of m top-ranked GO terms to evaluate which set of GO terms obtains the highest accuracy. For example, the mean score of ESL dataset is 0.009; eight sets of m GO terms are then generated, where m=23, 42, 77, 116, 178, 271, 372 and 501 (Fig. 3a). All m=501 GO terms have scores >0.009. The highest accuracy, 96.2%, is obtained by using m=501 GO terms by applying the SVM on ESL dataset (Fig. 4a). For the PSL dataset, m=158 top-ranked GO terms, which are input into the SVM, yield the highest accuracy of 93.7%, where m=25, 43, 70, 106, 136 and 158 (Figs. 3 and 4(b)).

The SVM classification problem in this work is solved by utilizing an SVM with a radial basis kernel function $\exp(-\gamma\|x^i-x^j\|^2)$ from LIBSVM (Chang and Lin, 2001), where x^i and x^j are training samples, and γ is a kernel parameter. The parameters γ and a cost parameter C are to be tuned in using the SVM. In this study, the best parameter values are determined from $\gamma \in \{2^{-7}, 2^{-6}, \ldots, 2^8\}$ and $C \in \{2^{-7}, 2^{-6}, \ldots, 2^8\}$.

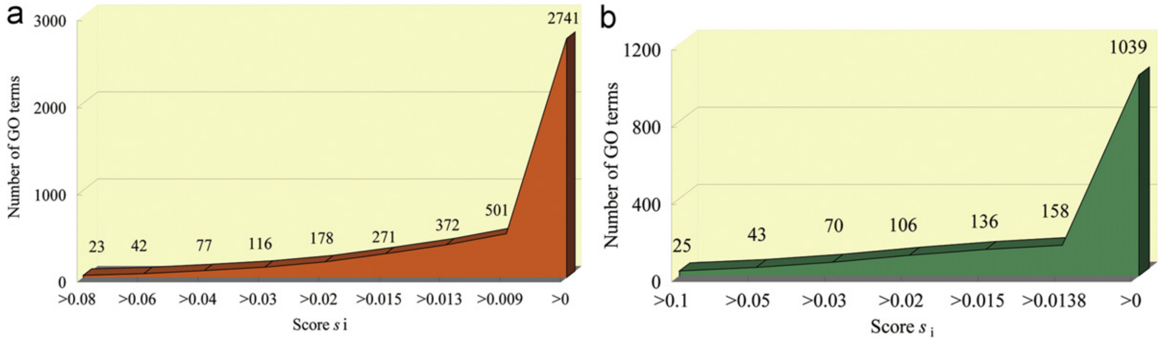

Fig. 3. Distribution of scores for the datasets ESL and PSL: (a) the mean of n scores is 0.009 for ESL, where n=2741 GO terms are obtained for ESL. Eight sets of m GO terms are generated according to their scores, where m=23, 42, 77,116, 178, 271, 372 and 501. Each of the m=501 GO terms has more than the mean scores (0.009). (b) For PSL, n=1039 GO terms are obtained and the means of n scores is 0.0138. Six sets of m GO terms are generated according to their scores, where m=25, 43, 70,106, 136, and 158. For each of 158 GO terms, the score is larger than the mean of n scores (0.0138).

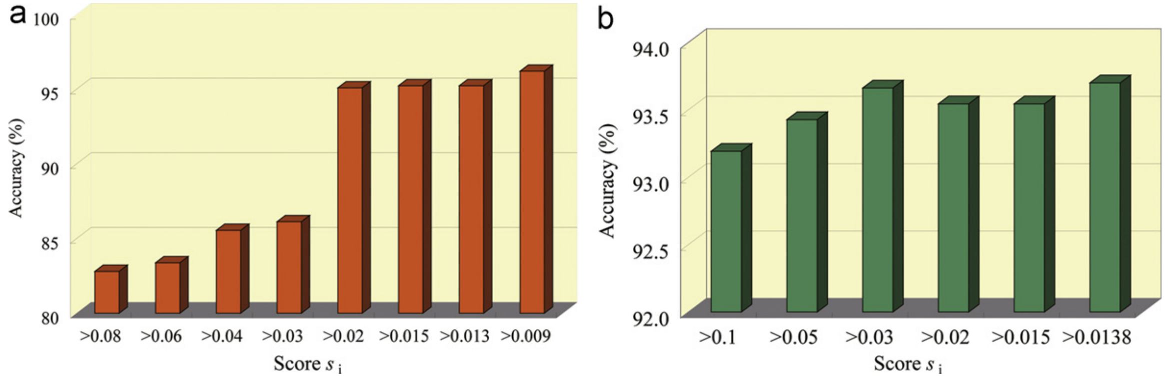

Fig. 4. Accuracy performance for all sets of m GO terms: (a) the highest accuracy 96.2% is obtained by using $m=501$ GO terms with SVM on ESL dataset, where each of the 501 GO terms has a score ≥ 0.009 (the mean score). (b) For PSL, the $m=158$ GO terms with SVM can yield the best accuracy 93.7%. Each of the 501 GO terms has a score ≥ 0.0138 (the mean score).

Among the independent dataset test, sub-sampling or N-fold (e.g., 5- or 10-fold) cross-validation test, and jackknife test, which are often used for examining the accuracy of a statistical prediction method (Chou and Zhang, 1995), the jackknife test was deemed the least arbitrary that can always yield a unique result for a given benchmark dataset (Chou, 2011). Therefore, the jackknife test has been increasingly recognized and widely adopted to various prediction methods (Cai et al., 2005; Chou and Cai, 2005a; Chou and Shen, 2006a, 2006b, 2007, 2010a, 2010b; Chou et al., 2011, 2012; Shen and Chou, 2010). When using the jackknife test like Euk-mPLoc 2.0 (Chou and Shen, 2010a) and Gneg-mPLoc (Shen and Chou, 2010), the high dimensional ($n=60{,}020$) GO term feature vector is actually necessary for a universal representation of the protein sequences. Conversely, this work calculates scores of GO terms and thus identifies a small set of m GO terms. Consequently, 10-fold cross-validation (10-CV) is adopted as done by many investigators with SVM as the prediction engine, where $m \ll n$.

In addition to the overall accuracy (ACC), sensitivity (SN) and specificity (SP) are also widely used performance evaluation parameters, and are derived as follows:

$$ACC=(q_c+s_c)/(q_c+s_c+u_c+o_c) \tag{3}$$

$$SN=q_c/(q_c+o_c) \tag{4}$$

$$SP=s_c/(u_c+s_c) \tag{5}$$

where q_c is the number of sequences of class c that are predicted correctly, s_c is the number of sequences not in class c that are predicted correctly, u_c is the number of under-predicted sequences, and o_c is the number of over-predicted sequences. Additionally, the value of the Matthews correlation coefficient (MCC) is also recorded (Matthews, 1975), belonging to $[-1,1]$. A coefficient of 1.0 is a perfect prediction; 0 represents an average random prediction, and -1.0 represents an inverse prediction. The MCC value is defined as follows (Ferreira and Azevedo, 2007; Matthews, 1975):

$$\mathrm{MCC}_c=q_cs_c-u_co_c/\sqrt{(q_c+u_c)(q_c+o_c)(s_c+u_c)(s_c+o_c)},\quad c=1,2 \tag{6}$$

2.5. Prediction of query sequences

The proposed SVM-based classifier uses an m-dimensional feature vector as input to implement the prediction system. Parameter settings of the SVM are determined in the training phase. The block diagram shows that the input of this prediction system is a secretory protein sequence P with FASTA format. The output is the predicted non-classical secretion. The prediction procedure is as follows.

Step 1: For the query sequence P, BLAST with $(h, e)=(5, 10^{-9})$ is utilized with the Swiss-Prot database to obtain homologies with known accession numbers for secretory proteins in eukaryotes and prokaryotes.
Step 2: The obtained accession numbers are used to retrieve their annotated GO terms from the GO database. Let the total number of GO terms obtained be j.
Step 3: Confirm whether each j GO term exists in the m features of GO terms. If the jth GO term is in the set of m GO terms, then element $\psi_i^G=1$ (Eq. (2)); otherwise, $\psi_i^G=0$, where $i=1, 2, \ldots, m$.

For the query protein sequence, when no of the j GO terms belongs to the set of m top-ranked GO terms or no homologies are obtained, this Sec-GO method simple predicts the query protein with non-classical secretion under concerning the possibility of non-classical secretion annotated by GO is larger than that of classical secretion in both eukaryotes and prokaryotes (Tables 4 and 5). For example, only 6.8% ($=(45+2)/692$) of the proteins in the ESL dataset are represented for zero vectors, whereas 45 proteins have no homologies and two proteins are not annotated by the m top-ranked GO terms.

3. Results and discussion

3.1. Comparison of other feature types

This proposed Sec-GO method is implemented by using SVM with a set of m top-ranked GO term features. Figs. 3 and 4 show two sets of $m=501$ and 158 GO terms are identified; their correlated training accuracies are 96.2% and 93.7% for the ESL and PSL datasets, respectively. To evaluate the effectiveness of the GO terms used in Sec-GO, the SVM with three additional feature sets—20 AAC features (SVM-AAC), 531 PCP features (SVM-PCP), and PSSM (SVM-PSSM)—are individually evaluated in terms of the prediction accuracy of 10-CV using two ESL and PSL datasets. The 531 PCP features are derived from the 20 physicochemical properties of AAindex (Kawashima and Kanehisa, 2000) by averaging over the entire protein sequence (Huang et al., 2007; Hung et al., 2010). The PSSM profile along with compositional attributes and similarity search based information are generated using PSI-BLAST (Garg and Raghava, 2008).

The best values of parameters γ and C in the SVM-based classifiers were determined using a step-wise approach from

$\gamma \in \{2^{-7}, 2^{-6}, \ldots, 2^{8}\}$ and $C \in \{2^{-7}, 2^{-6}, \ldots, 2^{8}\}$. Tables 6 and 7 show the Sec-GO prediction method obtains test accuracies of 96.8% and 94.5%, which are better than those of SVM-AAC (82.5% and 81.3%), SVM-PCP (76.4% and 65.4%) and PSSM-based classifiers (88.3% and 85.3%) for the ESL and PSL datasets, respectively.

3.2. Ten top-ranked GO features

Tables 8 and 9 list the ten top-ranked GO features with their scores in descending order. All top-four GO terms in the two datasets are annotated by subcellular locations comprising GO:0005576 (extracellular region), GO:0005634 (nucleus), GO:0005737(cytoplasm), GO:0005615 (extracellular space), GO:0016020 (membrane) and GO:0005886 (plasma membrane). Additionally, the term GO:0005515 (protein binding), ranked five in Table 8, belongs to the molecular function branch and can interact selectively and non-covalently with any protein or protein complex (a complex of two or more proteins that may include other non-protein molecules).

Five GO terms in Table 9 belong to the molecular function branch GO:0016787 (hydrolase activity), GO:0000166 (nucleotide binding), GO:0016740 (transferase activity), GO:0005524 (ATP binding) and GO:0016798 (hydrolase activity, acting on glycosyl bonds). Of the two hydrolase activities, GO:0016787 catalyzes the hydrolysis of various bonds; however, GO:0016798 catalyzes only the hydrolysis of glycosyl bond, where hydrolase is the systematic name of any enzyme in EC class 3. Furthermore, the term GO:0009405 (pathogenesis), which has a rank of 10, belongs to the branch of biological process ontology. This term activates the set of specific processes that generate the ability of an organism to cause disease in another organism.

3.3. Performance evaluation

Due to different design aims and datasets used, determining which prediction method is the most efficient or best by re-implementing these prediction methods is difficult. Therefore, in comparing the prediction performance of SRTpred, SecretomeP 2.0, SecretP 2.0, and NClassG+ (Tables 6, 7, 10, and 11), their prediction performance is obtained directly from the studies by Kandaswamy et al. (2010) and Restrepo-Montoya et al. (2011). Similarly, comparing the computational and training time of these methods is also difficult.

The Sec-GO method yields an independent test accuracy of 95.0%, compared with that of an existing method NClassG+ (90.0%) (Restrepo-Montoya et al., 2011) in predicting non-classically secreted Gram-positive bacterial proteins (Table 7). Additionally, Sec-GO with MCC=0.911 is also better than NClassG+ with MCC=0.71. The Sec-GO method applying to the ES_SPRED dataset identified $m=436$ GO terms and achieved a test accuracy of 96.1% and MCC of 0.845 for mammalian proteins, better than SPRED (Kandaswamy et al., 2010) with an accuracy of 82.2% and MCC of 0.504 when using the top 50 of the frequencies of tri-peptides and peptides, secondary structure, and physicochemical property features (Table 10). The Sec-GO method uses 5-fold training for comparison to SPRED.

Table 8
Ten top-rank GO terms for the dataset ESL. The abbreviations M, B and C represent the three branches molecular function, biological process and cellular component, respectively.

Rank	GO term	Score	Branch	Name
1	GO:0005576	1.000	C	Extracellular region
2	GO:0005634	0.698	C	Nucleus
3	GO:0005737	0.628	C	Cytoplasm
4	GO:0005615	0.509	C	Extracellular space
5	GO:0005515	0.314	M	Protein binding
6	GO:0006350	0.295	B	Transcription
7	GO:0045449	0.280	B	Regulation of transcription
8	GO:0003677	0.243	M	DNA binding
9	GO:0005829	0.183	C	Cytosol
10	GO:0005179	0.163	M	Hormone activity

Table 6
Performance comparison on SN (%), SP (%), ACC (%) and MCC in the dataset ES.

Type	Features			ESL 10-fold training				EST independent test			
	Size	Type	(C,γ)	SN	SP	ACC	MCC	SN	SP	ACC	MCC
SPRED	–	–	–	–	–	–	–	–	–	–	–
SVM-AAC	20	AAC	$(2^{1}, 2^{-2})$	54.8	92.7	82.5	0.526	50.5	91.1	80.2	0.460
SVM-PCP	20	PCP	$(2^{4}, 2^{2})$	14.5	99.0	76.4	0.294	48.3	89.1	78.1	0.408
SVM-PSSM	400	PSSM	$(2^{7}, 2^{1})$	94.1	72.0	88.2	0.698	93.9	73.1	88.3	0.694
Sec-GO	501	GO	$(2^{3}, 2^{-4})$	98.9	95.2	96.2	0.910	98.9	96.0	96.8	0.923

–: Not available.

Table 7
Performance comparison on SN (%), SP (%), ACC (%) and MCC in the dataset PS.

Type	Features			PSL 10-fold training				PST independent test			
	Size	Type	(C, γ)	SN	SP	ACC	MCC	SN	SP	ACC	MCC
NClassG+[a]	–	[b]	–	–	–	–	–	87	97	90	0.71
SecretomeP 2.0[a]	–	[b]	–	–	–	–	–	86	88	88	0.76
SecretP 2.0[a]	46	[b]	–	–	–	–	–	32	99	83	0.50
SVM-AAC	20	AAC	$(2^{-3}, 2^{-3})$	69.4	92.8	81.3	0.642	63.4	82.1	77.7	0.428
SVM-PCP	531	PCP	$(2^{-7}, 2^{-3})$	30.1	99.5	65.4	0.414	31.7	76.8	66.1	0.083
SVM-PSSM	400	PSSM	$(2^{8}, 2^{-2})$	92.4	91.5	92.0	0.840	78.4	78.4	83.5	0.634
Sec-GO	158	GO	$(2^{4}, 2^{-5})$	97.1	90.5	93.7	0.877	81.7	98.5	94.5	0.844

–: Not available.

[a] Taken from Restrepo-Montoya et al. (2011).

[b] Refer to Table 1.

The prediction server based on Sec-GO is available at http://iclab.life.nctu.edu.tw/secgo. This work uses 19 human proteins that are experimentally verified as non-classical secretory proteins by (Kandaswamy et al., 2010) to evaluate prediction performance of this web server.

These secreted sequences that lack signals are not found in any of the above datasets, ES and ES_SPRED, used to train or test Sec-GO. For comparison, this work applies SecretomeP 2.0 (Bendtsen et al., 2004), SRTpred (Garg and Raghava, 2008) and SPRED (Kandaswamy et al., 2010) to evaluate these 19 proteins. Prediction results (Table 11) show that Sec-GO correctly predicts 18 proteins, which is better than SPRED (15), SecretomeP 2.0 (12) and SRTpred (5).

3.4. Discussion

For those who do not really understand GO, it may be of help to carefully read an incisive analysis elaborated in Chou and Shen (2006b). For reader's convenience, some of its key points are summarized as follows:

(1) Since the GO database was established according to the molecular function, biological process, and cellular component, when a protein already has a GO annotation, why does one need to predict its biological function, pathway, and subcellular localization, which are merely different names for molecular function, biological process, and cellular component, respectively?
Although the GO database is constructed based on protein function and cellular component, for those proteins with the annotation of "subcellular location unknown" in the Swiss-Prot database, most (> 99%) of their corresponding GO terms in the GO database are also annotated with "cellular component unknown" (Chou and Shen, 2006b).

(2) Is it merely a procedure for converting GO annotations from one format into another?
Even for those proteins whose secretions are clearly annotated in the Swiss-Prot database, their corresponding GO terms in the GO database do not always directly indicate their corresponding non-classical secretion function.

(3) Is the high success rate obtained via the proposed GO approach merely due to a trivial utilization of annotations in the GO database?

Table 9
Ten top-rank GO terms for the dataset PSL. The abbreviations M, B and C represent the three branches molecular function, biological process and cellular component, respectively.

Rank	GO term	Score	Branch	Name
1	GO:0005737	1.000	C	Cytoplasm
2	GO:0005576	0.667	C	Extracellular region
3	GO:0016020	0.502	C	Membrane
4	GO:0005886	0.392	C	Plasma membrane
5	GO:0016787	0.382	M	Hydrolase activity
6	GO:0000166	0.271	M	Nucleotide binding
7	GO:0016740	0.260	M	Transferase activity
8	GO:0005524	0.250	M	ATP binding
9	GO:0016798	0.242	M	Hydrolase activity, acting on glycosyl bonds
10	GO:0009405	0.197	B	Pathogenesis

Importantly, only GO numbers of a query protein, not its GO annotations, are used, which is similar to using all other predictors in identifying subcellular localization, in that only the sequence of a query protein, not its Swiss-Prot annotation, is used. Accordingly, when BLAST is used to find homologies for non-classical secretion prediction, only 45.2% and 46.0% of true predictions are obtained according to the homologies' subcellular location annotations for the ESL and PSL datasets, respectively.

Table 11
Prediction result using 19 experimentally verified non-classical secretory proteins in eukaryotes for SecretomeP 2.0, SRTpred, SPRED and Sec-GO.

Swiss-Prot ID	Protein annotation	SecretomeP 2.0[a]	SRTpred[a]	SPRED[a]	Sec-GO
P05230	Heparin-binding growth factor 1	+	+	+	+
P09038	Heparin-binding growth factor 2	−	+	+	+
P01584	Interleukin 1 beta	+	+	+	+
P01583	Interleukin 1 alpha	+	−	+	+
P17931	Galectin-3	+	−	+	−
P14174	Macrophage migration inhibitory factor	+	−	+	+
P26447	Protein S100-A4	+	−	+	+
P09211	Glutathione S-transferase P	+	−	+	+
Q06830	Peroxiredoxin-1	+	−	+	+
Q14116	Interleukin 18	+	−	+	+
P27797	Calreticulin	−	+	+	+
P62805	Histone H4	−	−	+	+
P29034	Protein S100-A2	−	−	+	+
P09382	Galectin-1	−	−	+	+
P10599	Thioredoxin	−	−	+	+
P26441	Ciliary neurotrophic factor	+	+	−	+
P19622	Homeobox protein engrailed-2	+	−	−	+
Q16762	Thiosulfate sulfurtransferase	+	−	−	+
P09429	High mobility group protein B1	−	−	−	+
Total number of correctly predicted proteins		12	5	15	18

+: Proteins correctly predicted as non-classical secretory proteins.
−: Proteins incorrectly predicted as non-classical secretory proteins.
[a] Data taken from Kandaswamy et al. (2010).

Table 10
Performance comparison on SN (%), SP (%), ACC (%) and MCC in the dataset ES_SPRED.

Type	Features			ES_SPREDL 5-fold training				ES_SPREDT Independent test			
	Size	Type	(C, γ)	SN	SP	ACC	MCC	SN	SP	ACC	MCC
SPRED	–	–	–	–	–	–	–	88.3	81.9	82.2	0.504
SVM-AAC	20	AAC	$(2^0, 2^2)$	83.8	84.2	84.0	0.680	92.2	36.2	42.6	0.193
SVM-PCP	20	PCP	$(2^{-3}, 2^0)$	12.8	99.3	56.1	0.242	77.7	69.3	69.0	0.241
SVM-PSSM	400	PSSM	$(2^8, 2^{-4})$	85.3	77.9	81.2	0.628	43.1	99.5	84.9	0.583
Sec-GO	436	GO	$(2^{-3}, 2^{-7})$	99.3	98.0	98.6	0.973	97.7	96.5	96.7	0.861

–: Not available.

The high success rate obtained by the Sec-GO method via the GO approach is primarily due to the fact that proteins mapped in the GO database space are clustered in a way that reflects their biological functions, and this is by no means due to a trivial procedure of converting an annotation from one format into another format, as is often claimed by some researchers. Therefore, the Sec-GO approach significantly enhances the prediction success rate for those proteins that lack significant sequence homology to proteins with known biological functions, as demonstrated by Chou and Shen (2010b).

4. Conclusions

This work uses Sec-GO, a novel prediction method, by ranking and identifying *m* top-ranked GO terms as the only input features to design an SVM-based classifier. Prediction results show that the set of newly developed GO term features is effective in predicting non-classical secretory proteins. Moreover, the top-ranked GO features can be utilized effectively by combining them with other problem-dependent features for an individual SVM or ensemble classifiers to improve prediction accuracy. Conversely, the identified top-rank GO terms can be analyzed by using their GO annotations with their scores to further discover novel non-classical secretory proteins in eukaryotes and prokaryotes.

Acknowledgments

The author would like to thank Chyn Liaw and Chia-Ta Tsai for helping design the system. Additionally, thank all of three reviewers for great suggestions. Finally thank the National Science Council of Taiwan, ROC for financially supporting this research under the contract numbers NSC 100–2221-E-243-004.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at: http://dx.doi.org/10.1016/j.jtbi.2012.07.027.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSIBLAST:a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G., 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet., 25–29.

Bairoch, A., Apweiler, R., 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. 28, 45–48.

Bendtsen, J.D., Wooldridge, K.G., 2009. Bacterial Secreted Proteins: Secretory Mechanisms and Role in Pathogenesis. Caister Academy Press, Norfolk, UK.

Bendtsen, J.D., Kiemer, L., Fausboll, A., Brunak, S., 2005a. Non-classical protein secretion in bacteria. BMC Microbiol. 5, 58.

Bendtsen, J.D., Kiemer, L., Fausboll, A., Brunak, S., 2005b. Non-classical protein secretion in bacteria. BMC Microbiol. 5, 58.

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

Cai, Y.D., Zhou, G.P., Chou, K.C., 2005. Predicting enzyme family classes by hybridizing gene product composition and pseudo-amino acid composition. J. Theor. Biol. 234, 145–149.

Chang, C.C., Lin, C.J., 2001. LIBSVM: A Library for Support Vector Machines. Software available at: 〈www.csie.ntu.edu.tw/∼cjlin/libsvm〉.

Chou, K.C., 2011. Some remarks on protein attribute prediction and pseudo amino acid composition. J. Theor. Biol. 273, 236–247.

Chou, K.C., Zhang, C.T., 1995. Review: prediction of protein structural classes. Crit. Rev. Biochem. Mol. Biol. 30, 275–349.

Chou, K.C., Cai, Y.D., 2004. Using GO-PseAA predictor to predict enzyme sub-class. Biochem. Biophys. Res. Commun. 325, 506–509.

Chou, K.C., Cai, Y.D., 2005a. Predicting protein–protein interactions from sequences in a hybridization space. J. Proteome Res. 5, 316–322.

Chou, K.C., Cai, Y.D., 2005b. Using GO-PseAA predictor to identify membrane proteins and their types. Biochem. Biophys. Res. Commun. 327, 845–847.

Chou, K.C., Shen, H.B., 2006a. Predicting eukaryotic protein subcellular location by fusing optimized vidence-theoretic K-nearest neighbor classifiers. J. Proteome Res. 5, 1888–1897.

Chou, K.C., Shen, H.B., 2006b. Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization. Biochem. Biophys. Res. Commun. 347, 150–157.

Chou, K.C., Shen, H.B., 2007. Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. J. Proteome Res..

Chou, K.C., Shen, H.B., 2010a. A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mPLoc 2.0. PLoS One 5, e9931.

Chou, K.C., Shen, H.B., 2010b. Cell-PLoc 2.0: an improved package of web-servers for predicting subcellular localization of proteins in various organisms. Nat. Sci. 2, 1090–1103.

Chou, K.C., Wu, Z.C., Xiao, X., 2011. iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins. PLoS One, 6.

Chou, K.C., Wu, Z.C., Xiao, X., 2012. iLoc-Hum: using accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. Mol. Biosyst. 8, 629–641.

Desvaux, M., Hébraud, M., 2006. The protein secretion systems in Listeria: inside out bacterial virulence. FEMS Microbiol. Rev. 30, 774–805.

Ferreira, P., Azevedo, P., 2007. Evaluating deterministic motif significance measures in protein databases. Algorithms Mol. Biol. 2, 16.

Garg, A., Raghava, G.P.S., 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.

Ho, S.Y., Chen, J.H., Huang, M.H., 2004. Inheritable genetic algorithm for biobjective 0/1 combinatorial optimization problems and its applications. IEEE Trans. Syst. Man Cybern. Part B 34, 609–620.

Huang, W.L., Tung, C.W., Huang, H.L., Hwang, S.F., Ho, S.Y., 2007. ProLoc: prediction of protein subnuclear localization using SVM with automatic selection from physicochemical composition features. BioSystems 90, 573–581.

Hung, C.H., Huang, H.L., Hsu, K.T., Ho, S.J., Ho, S.Y., 2010. Prediction of non-classical secreted proteins using informative physicochemical properties. Interdiscip. Sci. Comput. Life Sci. 2, 263–270.

Kandaswamy, K.K., Pugalenthi, G., Hartmann, E., Kalies, K.-U., Moller, S., Suganthan, P.N., Martinetz, T., 2010. SPRED: a machine learning approach for the identification of classical and non-classical secretory proteins in mammalian genomes. Biochem. Biophys. Res. Commun. 391, 1306–1311.

Kawashima, S., Kanehisa, M., 2000. AAindex: amino acid index database. Nucleic Acids Res. 28, 374.

Li, W., Godzik, A., 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658–1659.

Matthews, B.W., 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta 405, 442–451.

Nickel, W., 2003. The mystery of nonclassical protein secretion. Eur. J. Biochem. 270, 2109–2119.

Popescu, M., Keller, J.M., Mitchell, J.A., 2006. Fuzzy measures on the gene ontology for gene product similarity. IEEE/ACM Trans. Comput. Biol. Bioinf., 3.

Prudovsky, I., Mandinova, A., Soldi, R., Bagala, C., Graziani, I., Landriscina, M., Tarantini, F., Duarte, M., Bellum, S., Doherty, H., Maciag, T., 2003. The non-classical export routes: FGF1 and IL-1{alpha} point the way. J. Cell Sci. 116, 4871–4881.

Radisky, D.C., Stallings-Mann, M., Hirai, Y., Bissell, M.J., 2009. Single proteins might have dual but related functions in intracellular and extracellular microenvironments. Nat. Rev. Mol. Cell. Biol. 10, 228–234.

Restrepo-Montoya, D., Pino, C., Nino, L., Patarroyo, M., 2011. NClassG+: a classifier for non-classically secreted Gram-positive bacterial proteins. BMC Bioinf. 12, 21.

Shen, H.-B., Chou, K.-C., 2010. Gneg-mPLoc: a top-down strategy to enhance the quality of predicting subcellular localization of Gram-negative bacterial proteins. J. Theor. Biol. 264, 326–333.

Torto-Alalibo, T., Collmer, C., Gwinn-Giglio, M., 2009. The Plant-Associated Microbe Gene Ontology (PAMGO) Consortium: community development of new Gene Ontology terms describing biological processes involved in microbe–host interactions. BMC Microbiol. 9, S1.

Tseng, T.-T., Tyler, B., Setubal, J., 2009. Protein secretion systems in bacterial–host associations, and their description in the Gene Ontology. BMC Microbiol. 9, S2.

Yu, L., Guo, Y., Li, Y., Li, G., Li, M., Luo, J., Xiong, W., Qin, W., 2010. SecretP: identifying bacterial secreted proteins by fusing new features into Chou′s pseudo-amino acid composition. J. Theor. Biol. 267, 1–6.