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