World Academy of Science, Engineering and Technology
International Journal of Bioengineering and Life Sciences
Vol:3, No:5, 2009

Virulent-GO: Prediction of Virulent Proteins in Bacterial Pathogens Utilizing Gene Ontology Terms

Chia-Ta Tsai, Wen-Lin Huang, Shinn-Jang Ho, Li-Sun Shu, and Shinn-Ying Ho

***Abstract*—**Prediction of bacterial virulent protein sequences can give assistance to identification and characterization of novel virulence-associated factors and discover drug/vaccine targets against proteins indispensable to pathogenicity. Gene Ontology (GO) annotation which describes functions of genes and gene products as a controlled vocabulary of terms has been shown effectively for a variety of tasks such as gene expression study, GO annotation prediction, protein subcellular localization, etc. In this study, we propose a sequence-based method Virulent-GO by mining informative GO terms as features for predicting bacterial virulent proteins.

Each protein in the datasets used by the existing method VirulentPred is annotated by using BLAST to obtain its homologies with known accession numbers for retrieving GO terms. After investigating various popular classifiers using the same five-fold cross-validation scheme, Virulent-GO using the single kind of GO term features with an accuracy of 82.5% is slightly better than VirulentPred with 81.8% using five kinds of sequence-based features. For the evaluation of independent test, Virulent-GO also yields better results (82.0%) than VirulentPred (80.7%). When evaluating single kind of feature with SVM, the GO term feature performs much well, compared with each of the five kinds of features.

***Keywords*—**Bacterial virulence factors, GO terms, prediction, protein sequence.

I. Introduction

THE identification of novel virulence determinants is a key step of the process to understand how pathogenic bacteria interact with their hosts to produce clinical disease [2]. Multiple virulence factors in bacterial pathogens serve separately or are cooperated each other during a course of stages to infect susceptible hosts. The generic mechanisms shared by these bacterial virulence factors and themselves are adequately discussed in a previous review [1]. These bacterial virulence factors may also serve as targets for vaccine and drug development [1; 2; 3].

For this aim at providing an in-depth coverage of the major virulence factors from various best-characterized bacterial pathogens, a reference database for bacterial virulence factors (VFDB) was build [4] and continuously updated [5]. This database contained cumulative information for 16 important bacteria pathogens, virulence-associated genes, protein structural functions, mechanisms and important literatures [4]. A set of virulent proteins in this previous release with others from SWISS-PROT [6] were collected and processed as datasets of virulent proteins in bacterial pathogens to evaluate the existing method VirulentPred [6; 7].

Several mixed-strategy machine learning approaches have been proposed to classify bacterial virulent proteins successfully. While specifying virulence factors to adhesins, a sequence-based prediction method named SPAAN [8] was proposed for prediction of adhesins and adhesin-like proteins. Before being process through five types of attribute modules separately, a given protein sequence has been quantified by these attributes including amino acid frequencies, multiplets frequencies, dipeptide frequencies, charge composition and hydrophobic compositions. A probability of being an adhesin is computed while considering each value resulted from a set of five well-trained neural networks processed respectively.

Recently, VirulentPred [7] used a bilayer cascade support vectors machine (SVM) classifier for prediction of virulent proteins in bacterial pathogens. VirulentPred consists of five separated classifiers trained with single kind of features: 1) amino acid composition, 2) dipeptide composition, 3) high order dipeptide composition, 4) evolutionary information in a form of PSSM profiles and 5) PSI-BLAST based similarity search separately [9], and a summary SVM classifier utilizing their classification turned out to efficiently classify virulent proteins. Although the integrated classifiers perform well, the structure of classifiers or the innate characters of selected feature sets are less interpretable to biologists for advanced analysis.

Gene Ontology (GO) [10] annotation describes functions of genes and gene products as a controlled vocabulary of terms. Recently, GO annotation has been used by many groups for a variety of tasks such as grouping GO terms to improve the assessment of gene set enrichment [11], using GO with probabilistic chain graphs for protein classification [12],

Chia-Ta Tsai is with the Institute of Bioinformatics and Systems Biology, National Chiao Tung University, Hsinchu 300, Taiwan.

Wen-Lin Huang is with the Department of Management Information System, Chin Min Institute of Technology, Miaoli, Taiwan.

Shinn-Jang Ho is with the Department of Automation Engineering, National Formosa University, Yunlin 632, Taiwan.

Li-Sun Shu is with the Department of Information Management, Overseas Chinese Institute of Technology, Taichung 407, Taiwan.

Shinn-Ying Ho is with the Institute of Bioinformatics and Systems Biology, and Department of Biological Science and Technology, National Chiao Tung University, Hsinchu 300, Taiwan (corresponding author to provide phone: 886-3-571-2121, ext: 56909; e-mail: syho @mail.nctu.edu.tw).

prediction of subnuclear localization [13], predicting transcription factor DNA binding preference [14], etc. These applications of GO terms can be referred to the late study [15]. That GO annotation has grown in size and popularity [16] makes effectiveness of the GO-based features increasing. Various efficient sequence-based prediction methods [12; 13; 14; 15; 17; 18] were proposed by utilizing GO terms. The GO terms describe the functions of genes and gene products across species by a graph structure and are categorized into three branches: molecular function, biological process and cellular component [10].

In this study, we propose a sequence-based method Virulent-GO by mining informative GO terms as features for predicting bacterial virulent proteins. The sequences of bacterial pathogens were obtained from SWISS-PROT [6] and VFDB [4]. All the instructive GO terms of these sequences were obtained by using BLAST [19] to obtain its homologies with known accession numbers which are used to query the GOA database [16] consequently. The potential for GO terms to discriminate virulent proteins in bacteria has been demonstrated by distinct differences between virulent and non-virulent proteins. All keywords retrieving from literatures [1] which are associated with categories of virulence factors are also annotated by GO terms. All the GO terms appearing in both sets of instructive GO terms and the GO terms from keywords are denoted as *essential GO terms*. A point of integrative view from the instructive GO term set and the essential GO term set can reveal a few nature of complexity from virulence factors in bacterial pathogens.

The abilities of instructive GO terms combined with various widely-used classifiers, such as k-nearest neighbors, NaïveBayes, decision tree and SVM, to predict bacterial virulent proteins have been evaluated by five-fold cross-validation scheme. After the evaluations of some classifiers, the high-performance method Virulent-GO utilized a well-trained SVM classifier and these informative GO terms to classify bacterial virulent proteins.

Virulent-GO using the single kind of GO term features with an accuracy of 82.5% is slightly better than VirulentPred [7] with 81.8% using five kinds of sequence-based features. For the evaluation of independent test, Virulent-GO also yields better results (82.0%) than VirulentPred (80.7%). When evaluating single kind of feature with SVM, the GO term feature performs much well, compared with each of the five kinds of features.

II. Methods

A. Overview of Constructing Virulent-GO

The design of Virulent-GO is a two-stage approach to classifying virulent proteins in bacterial pathogens utilizing the single kind of GO term features. At the first stage, sequences in the given training dataset are used to obtain their homologies by using BLAST. The accession numbers of homologies were used to query the GOA database to obtain a set of instructive GO terms. All sequences in the training dataset are represented as a vector of instructive GO terms. Additionally, a set of essential GO terms is collected. The flowchart of generating feature vectors of instructive GO terms and essential GO terms is shown in Fig. 1.

At the second stage, a good classifier for utilizing the instructive GO terms is determined by evaluating some widely-used classifiers. The high-performance classifier determined is further evaluated using an independent test dataset. The details are described below.

TABLE I
KEYWORDS OBTAINED FROM LITERATURES

Keyword	Subclassification	Classification	Hits	Elements	N	V
Virulence	N/A	Virulence	1	5	48	315
Adhesion	Adhesion	Membrane Protein	6	91	24	107
Invasion	Invasion	Membrane Protein	1	14	1	2
Colonization	Colonization	Membrane Protein	0	0	0	0
Surface component	Surface component	Membrane Protein	1	13	2	0
Outer membrane protein	Outer membrane protein	Membrane Protein	9	30	395	526
Capsule	N/A	Capsule	6	34	458	398
Immune response inhibitor	Immune response inhibitor	Secretory protein	0	6	0	0
Toxin	Toxin	Secretory protein	13	40	215	443
Exotoxin	Toxin	Secretory protein	1	3	48	315
Transport toxin	Transport of toxin	Secretory protein	3	4	15	64
Cell wall	N/A	Cell wall and outer membrane components	25	219	154	215
Outer membrane components	N/A	Cell wall and outer membrane components	2	12	6	4
Peptidoglycan	N/A	Cell wall and outer membrane components	15	56	96	222
LPS	N/A	Cell wall and outer membrane components	2	5	52	115
Lipopolysaccharide	N/A	Cell wall and outer membrane components	7	38	51	177
Exndotoxin	N/A	Cell wall and outer membrane components	3	7	49	316
Teichoic acid	N/A	Cell wall and outer membrane components	2	6	2	31
Biofilm	Biofilm	Others	0	11	0	0
Iron acquisition	Iron acquisition	Others	0	0	0	0
Siderophore receptor	Iron acquisition	Others	2	5	113	90
ABC transport system	Iron acquisition	Others	7	8	459	255
PhoP/PhoQ two component system	PhoP/PhoQ two component system	Others	0	0	0	0

These keywords are from a tree of hieratical classes that were discussed in these reviews [1; 2]. Hits defined as the amount of intersected GO terms with instructive GO terms set. Elements are defined as GO terms retrieved from GOA database using the specific keyword. N and V represent this keyword the number of annotated proteins in training dataset via GO terms it found.

B. Preparing Datasets

A training dataset and two independent test datasets obtained from VirulentPred [7] were used to evaluate the proposed method Virulent-GO. Protein sequences in these datasets were retrieved from SWISS-PROT [6] and VFDB [4]. These datasets contained virulence factors of bacterial pathogens and scale the redundancy that each sequences shares identities under 40%. After the process of eliminating similar sequences, five species of bacterial pathogens which contain relatively small amount of sequences are used to construct the independent test dataset. The sequences of the other 12 bacterial pathogens were used to construct a training dataset. In addition, a small fraction of SWISS-PROT sequences in the training dataset are randomly selected to construct another independent test dataset. The detailed manipulation of constructing these datasets can be referred to the work VirulentPred [7].

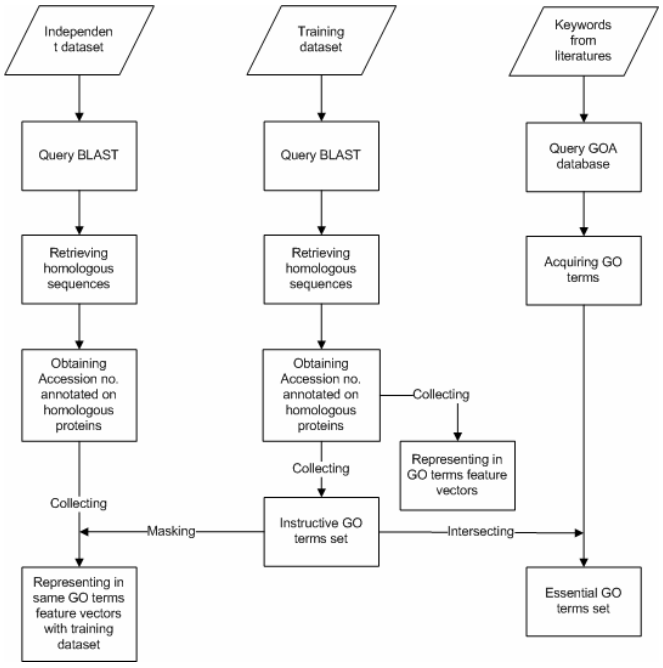

Fig. 1 Flowchart of generating feature vectors of instructive GO terms and essential GO terms

In this study, the two independent datasets were merged for evaluation. The used training dataset consists of 1025 virulent proteins and 1030 non-virulent proteins, and the independent test dataset consists of 181 virulent proteins and 186 non-virulent proteins. The five species of bacterial pathogens are *Campylobacter, Neisseria, Bordetella, Haemophilus* and *Listeria*. On the other hands, the other 12 bacterial pathogens are *Escherichia, Pseudomonas, Salmonella, Streptococcus, Legionella, Bacillus, Staphylococcus, Shigella, Helicobacter, Mycobacterium, Yersinia* and *Vibrio*.

C. Generating Features form GOA Database

The used GO term features of each protein sequence were obtained by using BLAST [9] and to obtain its homology with a known accession number and then querying the GOA database [16]. The parameters for BLAST are $h = 4$ and $e = 0.1$, and retrieving 1396 GO terms to representing training dataset. These proteins are represented as high-dimensional vectors of n binary features, where n is the total number of GO terms in the complete annotation set (a component of 1 if the annotation is hit, and 0 otherwise). The set of GO terms is defined as "instructive GO terms" set which GO terms contained were all annotated in the training dataset. Note that the GO terms that were annotated on independent test dataset were later masked and only represented by instructive GO terms.

For insights the nature of virulence factors of bacterial pathogens, a keyword set is collected and summarized from reviews [1; 2], shown in Table I. These keywords are chosen because of holding the basis of the mechanism of virulence and functions. Each keyword acquired several elements from querying the GOA database and some elements would be overlapped with instructive GO terms, defined as Hit in Table I. The set of essential GO terms has 73 GO terms, shown in Table IV.

To evaluate performance of using only essential GO terms, the training dataset was further processed to generate two other training data sets. One is using only *essential GO terms* and masking out other instructive GO terms to represent whole training dataset and denote as Training Dataset-1. Another is eliminating proteins in Training Dataset-1 if they are annotated without any *essential GO terms* and denoted as Training Dataset-2. This process turns out reducing the size of Training

TABLE II
RESULTS OF INSTRUCTIVE GO ANNOTATION FOR ALL SEQUENCES

Class	Total GO terms n	Number of GO terms			Number of sequences annotated by n GO terms		
		Smallest	Largest	Mean	$n = 0$	$n = 1$	$n > 1$
N	1174	0	34	8.82	4	14	1012
V	599	0	27	6.02	167	21	837
total	1396			7.42	171	35	1849

TABLE III
RESULTS OF ESSENTIAL GO ANNOTATION FOR ALL SEQUENCES IN TRAINNING DATASET

Class	Total essential GO terms g	Number of GO terms			Number of sequences annotated by g essential GO terms		
		Smallest	Largest	Mean	$g = 0$	$g = 1$	$g > 1$
N	65	0	8	1.64	289	261	480
V	60	0	8	2.04	288	206	531
total	73			1.84	577	467	1011

Dataset-2 to 741 non-virulent proteins and 737 virulent proteins.

D. Model Implementations

To implement some typical classifiers in most popular manner, two well-known packages are adopted. Weka is software package collecting machine learning algorithms for data mining task in Java [20]. Three common classifiers are accessed: These are IB*k* (*k*-nearest neighbor classifier), NaïveBayes and J48 (C4.5 decision tree). The IB*k* was performed with $k = 1$, 3 and 5. The NaïveBayes was performed with two different modes that are applied a kernel estimator for numeric attributes or just assumed as normal distribution. The J48 was evaluating by considering confidence factor from 0.1 to 0.5 with a stepwise of 0.05 each and all in a default minimum number, 2 of instances per leaf. The confidence factor was found at 0.15 for maximized accuracy.

Otherwise, an SVM classifier is implement by LIBSVM [21]. By applying grid search toolkits LIBSVM provided, this SVM model was optimized both in cost C and kernel parameter γ corresponded to using Radial Basis Function (RBF) kernel. These two essential parameters are selected from exponent in a range from -7 to 5 with base 2. Note that performing these classifiers is not only to select a best performance one but also demonstrated that the ability of instructive GO terms to classify bacterial virulent proteins properly across classifiers with fine tuned.

E. Performance Evaluation

The leave-one-out cross-validation (LOOCV) is considered to be the most rigorous and objective test for its bias free nature, but this test is very computationally demanding and is often impractical for large datasets The n-fold cross-validation not only provides a bias-free estimation of the accuracy at a much reduced computational cost, but also considered as an reasonable test for evaluating classification performance of an algorithm. In this study, five-fold cross-validation is applied on entire training set to fine tuned parameters of classification models and evaluating its performance [22].

The popular measures to evaluating classification models are Accuracy (ACC), Sensitivity (SN), Specificity (SP) and Matthews Correlation Coefficient (MCC). In this study, virulent proteins and non-virulent proteins are defined as positive and negative respectively. Therefore, TP stands for true positives, TN the true negatives, FP the false positives and FN the false negatives. These measures are defined as below:

$$ACC = \frac{(TP + TN)}{(TP + TN + FP + FN)} \times 100\% \quad (1)$$

$$SN = \frac{(TP)}{(TP + FN)} \times 100\% \quad (2)$$

$$SP = \frac{(TN)}{(TN + FP)} \times 100\% \quad (3)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \quad (4)$$

III. Results

A. Analyzing Instructive GO terms and Essential GO Terms

In integrative views with instructive GO terms set and essential GO terms set, the training datasets that are constructed by non-virulent and virulent protein sequences in bacterial pathogens are well-annotated and informative by these two sets. Non-virulent proteins share more diversity of GO terms (1174) to virulent proteins (599) that is shown in Table II. Proteins which are recognized as non-virulent in bacterial pathogens annotate with more GO terms (8.82) than virulent proteins (6.02). There are 167 virulent proteins annotated with no GO terms from their homology while only 4 non-virulent proteins have no annotated GO term. In contract to the instructive GO terms, the numbers are similar for non-virulent proteins (288) and virulent proteins (289) annotated without any essential GO terms. Although a wider range of essential GO terms (65 to 60 for virulent proteins) is used to annotated non-virulent proteins, the virulent proteins are annotated by more essential GO terms (2.04) than non-virulent proteins (1.64) in average amount. Moreover, a large numbers of virulent proteins were annotated by several GO terms. This trend could be seen from a frequency-distribution in Fig. 2. A clear difference was shown since essential GO terns *g* get larger than 3.

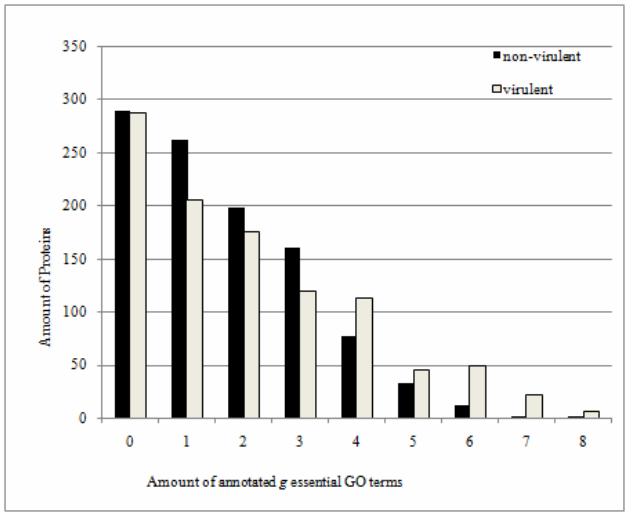

Fig. 2 The number of essential GO terms annotated in each protein is shown in this frequency distribution graph

Most of keywords are successfully accessing to both non-virulent proteins and virulent proteins via retrieving some essential GO terms. Many of they even access hundreds of proteins. The essential GO terms set is constructed across three major branches, and 52 essential GO terms still are shared by both non-virulent proteins and virulent proteins. Thus, a proper classifier should be applied to archive a successful prediction.

Although keywords like "Colonization", "Iron acquisition" and "PhoP/PhoQ two component system" assess to 0 proteins

TABLE IV
THE BASIC INFORMATION OF 73 ESSENTIAL GO TERMS

no	Access No.	Name	Branch	no	Access No.	Name	Branch
1	GO:0000270	peptidoglycan metabolic process	B[a]	38	GO:0009405	pathogenesis	B
2	GO:0000271	polysaccharide biosynthetic process	B	39	GO:0009425	flagellin-based flagellum basal body	C
3	GO:0003796	lysozyme activity	M[b]	40	GO:0009428	flagellin-based flagellum basal body, distal rod, P ring	C
4	GO:0003959	NADPH dehydrogenase activity	M	41	GO:0009626	plant-type hypersensitive response	B
5	GO:0004222	metalloendopeptidase activity	M	42	GO:0009636	response to toxin	B
6	GO:0004595	pantetheine-phosphate adenylyltransferase activity	M	43	GO:0009842	cyanelle	C
7	GO:0004713	protein tyrosine kinase activity	M	44	GO:0009986	cell surface	C
8	GO:0005102	receptor binding	M	45	GO:0015153	rhamnose transmembrane transporter activity	M
9	GO:0005198	structural molecule activity	M	46	GO:0015267	channel activity	M
10	GO:0005215	transporter activity	M	47	GO:0015288	porin activity	M
11	GO:0005515	protein binding	M	48	GO:0015343	siderophore-iron transmembrane transporter activity	M
12	GO:0005524	ATP binding	M	49	GO:0015627	type II protein secretion system complex	C
13	GO:0005576	extracellular region	C[c]	50	GO:0015628	protein secretion by the type II secretion system	B
14	GO:0005618	cell wall	C	51	GO:0015643	toxin binding	M
15	GO:0005887	integral to plasma membrane	C	52	GO:0015762	rhamnose transport	B
16	GO:0006281	DNA repair	B	53	GO:0015774	polysaccharide transport	B
17	GO:0006810	transport	B	54	GO:0015937	coenzyme A biosynthetic process	B
18	GO:0006817	phosphate transport	B	55	GO:0016020	membrane	C
19	GO:0007047	cell wall organization	B	56	GO:0016998	cell wall catabolic process	B
20	GO:0007155	cell adhesion	B	57	GO:0019299	rhamnose metabolic process	B
21	GO:0008270	zinc ion binding	M	58	GO:0019350	teichoic acid biosynthetic process	B
22	GO:0008565	protein transporter activity	M	59	GO:0019534	toxin transporter activity	M
23	GO:0008918	lipopolysaccharide 3-alpha-galactosyltransferase activity	M	60	GO:0019835	cytolysis	B
24	GO:0008919	lipopolysaccharide glucosyltransferase I activity	M	61	GO:0019867	outer membrane	C
25	GO:0009103	lipopolysaccharide biosynthetic process	B	62	GO:0019877	diaminopimelate biosynthetic process	B
26	GO:0009244	lipopolysaccharide core region biosynthetic process	B	63	GO:0030246	carbohydrate binding	M
27	GO:0009245	lipid A biosynthetic process	B	64	GO:0030288	outer membrane-bounded periplasmic space	C
28	GO:0009252	peptidoglycan biosynthetic process	B	65	GO:0031975	envelope	C
29	GO:0009253	peptidoglycan catabolic process	B	66	GO:0042121	alginic acid biosynthetic process	B
30	GO:0009254	peptidoglycan turnover	B	67	GO:0042122	alginic acid catabolic process	B
31	GO:0009273	peptidoglycan-based cell wall biogenesis	B	68	GO:0042243	asexual spore wall assembly	B
32	GO:0009274	peptidoglycan-based cell wall	C	69	GO:0042597	periplasmic space	C
33	GO:0009275	Gram-positive-bacterium-type cell wall	C	70	GO:0042603	capsule	C
34	GO:0009276	Gram-negative-bacterium-type cell wall	C	71	GO:0043190	ATP-binding cassette (ABC) transporter complex	C
35	GO:0009279	cell outer membrane	C	72	GO:0044406	adhesion to host	B
36	GO:0009306	protein secretion	B	73	GO:0045227	capsule polysaccharide biosynthetic process	B
37	GO:0009404	toxin metabolic process	B				

[a] B is the abbreviation of "Biological Process"; [b] M represents for "Molecular Function"; [c] C is for "Cellular Component".

for no GO term own by them could be recognized as a essential GO term, a typical example that is catered to "Iron acquisition", "Siderophore receptor" is querying and access to few proteins. Also, "PhoP/PhoQ two component system" and "ABC transport system" are in the same situation. Besides, there are two keywords retrieved certain GO terms but were failure to intersecting with instructive GO terms. They are "Immune response inhibitor" and "Biofilm".

B. Assessment of Features and Classifiers

To evaluate performance across widely-used classifiers, this study applied four kind of classifiers that are IB*k* (k-nearest neighbor), J48 (Decision Tree), NaïveBayes and SVM. With five-fold cross-validation, this turned out a strong support for the predictive power orientated form instructive GO terms. The accuracy was archived up to 82.5% (SVM), 80.0% (J48) and 79.5% (NaïveBayes). Even a lazy classifier IB*k* like could

TABLE V
DISTRIBUTION OF THE 73 ANNOTATED ESSENTIAL GO TERMS CORRESPONDED TO TRAINING DATASET

no	GO term	Branch	Training Dataset		no	GO term	Branch	Training Dataset	
			N	V				N	V
1	GO:0000270	B	1	0	38	GO:0009405	B	48	315
2	GO:0000271	B	10	22	39	GO:0009425	C	1	8
3	GO:0003796	M	2	7	40	GO:0009428	C	4	1
4	GO:0003959	M	1	0	41	GO:0009626	B	5	8
5	GO:0004222	M	6	11	42	GO:0009636	B	1	1
6	GO:0004595	M	0	1	43	GO:0009842	C	1	2
7	GO:0004713	M	2	3	44	GO:0009986	C	5	36
8	GO:0005102	M	4	19	45	GO:0015153	M	1	0
9	GO:0005198	M	8	22	46	GO:0015267	M	1	0
10	GO:0005215	M	95	28	47	GO:0015288	M	3	1
11	GO:0005515	M	103	112	48	GO:0015343	M	4	9
12	GO:0005524	M	202	64	49	GO:0015627	C	2	30
13	GO:0005576	C	49	166	50	GO:0015628	B	2	30
14	GO:0005618	C	7	28	51	GO:0015643	M	0	1
15	GO:0005887	C	2	0	52	GO:0015762	B	1	0
16	GO:0006281	B	49	5	53	GO:0015774	B	4	10
17	GO:0006810	B	305	203	54	GO:0015937	B	2	1
18	GO:0006817	B	3	3	55	GO:0016020	C	372	328
19	GO:0007047	B	21	27	56	GO:0016998	B	5	12
20	GO:0007155	B	6	57	57	GO:0019299	B	0	2
21	GO:0008270	M	58	38	58	GO:0019350	B	1	4
22	GO:0008565	M	10	50	59	GO:0019534	M	1	1
23	GO:0008918	M	1	0	60	GO:0019835	B	8	37
24	GO:0008919	M	1	2	61	GO:0019867	C	28	80
25	GO:0009103	B	15	38	62	GO:0019877	B	1	0
26	GO:0009244	B	2	0	63	GO:0030246	M	10	8
27	GO:0009245	B	8	2	64	GO:0030288	C	31	17
28	GO:0009252	B	13	1	65	GO:0031975	C	1	0
29	GO:0009253	B	4	8	66	GO:0042121	B	1	20
30	GO:0009254	B	2	0	67	GO:0042122	B	1	0
31	GO:0009273	B	5	2	68	GO:0042243	B	0	2
32	GO:0009274	C	2	1	69	GO:0042597	C	79	42
33	GO:0009275	C	0	1	70	GO:0042603	C	0	1
34	GO:0009276	C	22	14	71	GO:0043190	C	1	0
35	GO:0009279	C	29	75	72	GO:0044406	B	0	1
36	GO:0009306	B	15	67	73	GO:0045227	B	2	6
37	GO:0009404	B	0	3					

make out an accuracy of 78.6%. On the others hand, using instructive GO terms set to classify virulent proteins turned out a better performance (Accuracy 82.5%) compared to several generic features that are amino acids composition (72.1%), dipeptide composition (71.1%), similarity search (52.1%) and PSSM profile (78.1%).

TABLE VI
TRAINING RESULTS OF INSTRUCTIVE GO TERMS AS FEATURE PERFORMED ON MULTIPLE CLASSIFIERS WITH FIVE-FOLD CROSS-VALIDATION

Classifier	ACC (%)	SN (%)	SP (%)	MCC
IB1	78.6	77.5	79.7	0.57
IB3	76.8	73.3	81.6	0.54
IB5	73.4	68.4	82.0	0.49
NaïveBayes - Normal Distribution	78.2	77.0	79.4	0.56
NaïveBayes - Kernel Density Estimator	79.5	74.8	**86.3**	0.60
J48	80.0	80.0	80.1	0.60
SVM	**82.5**	**84.5**	80.6	**0.65**

The five-fold cross-validation scheme is also used to evaluating performance for Training Dataset-1 and Training Dataset-2 by combining with widely-used classifiers to demonstrate the efficiency of essential GO terms. Due to 288 non-virulent and 289 virulent proteins have no essential GO terms annotated, they could be recognized as a same class and lead to a lot of false positives or false negatives. These results could be seen from Table VIII. After excluding these proteins, the accurate rate just a little drop against results from training dataset which is annotated by instructive GO terms. These results are shown in Table VI.

TABLE VII
TRAINING RESULTS FROM APPLYING DIFFERENT FEATURES WITH FIVE-FOLD CROSS-VALIDATION

Feature	Classifier	ACC (%)	SN (%)	SP (%)	MCC
Amino Acid Compositions [a]	SVM	72.1	70.0	74.1	0.44
Dipeptide Compositions (i+1st) [a]	SVM	71.1	70.0	72.3	0.42
Dipeptide Compositions (i+2nd) [a]	SVM	72.0	70.2	73.7	0.44
PSI-BLAST Search [a]	---	52.1	52.5	51.7	---
Position-Specific Scoring Matrix [a]	SVM	78.1	78.1	78.1	0.56
GO terms	SVM	**82.5**	**84.5**	**80.6**	**0.65**

[a] These results are from the previous method, VirulentPred.

TABLE VIII
THE RESULTS OF EVALUATING ON ONLY ESSENTIAL GO TERMS IS INCLUDED WITH FIVE-FOLD CROSS-VALIDATION

Classifiers	Training Dataset-1								Training Dataset-2							
	ACC (%)	SN (%)	SP (%)	MCC	TP	TN	FP	FN	ACC (%)	SN (%)	SP (%)	MCC	TP	TN	FP	FN
IB1	64.9	60.2	69.6	0.30	617	717	408	313	72.3	69.5	75.2	0.45	512	557	225	184
J48	67.6	75.8	59.4	0.36	777	612	248	418	76.4	69.2	83.5	0.53	510	619	227	122
NaïveBayes - Kernel Density Estimator	69.1	51.7	86.4	0.41	530	890	495	140	75.4	69.3	81.5	0.51	511	604	226	137
SVM	70.1	50.8	89.2	0.43	521	919	504	111	78.1	71.2	85.0	0.57	525	630	212	111

C. Evaluating on Training and Independent Test

The virulent-GO is built using only a single SVM classifier comparing to the existing method, VirulentPred, using cascade SVM and obtained comparable results. With a summary SVM classifier to decide the virulence of proteins, VirulentPred enhanced its performance up to accuracy of 81.8%. A comparable result here is achieved by Virulent-GO that its accuracy is 82.5% (Table IX). An accuracy of 82.0% is archived by Virulent-GO on independent test dataset. This result is also improved slightly compared to VirulentPred with 80.7% (Table X).

TABLE IX
TRAINING RESULTS COMPARED WITH EXISTING METHOD WITH FIVE-FOLD CROSS-VALIDATION

Method	SVM	ACC(%)	SN(%)	SP(%)	MCC
VirulentPred	cascade	81.8	82.0	**81.5**	0.64
Virulent-GO	single layer	**82.5**	**84.5**	80.6	**0.65**

TABLE X
INDEPENDENT TEST RESULTS COMPARED WITH EXISTING METHOD

Method	ACC(%)	SN(%)	SP(%)	MCC
VirulentPred	80.7	81.2	80.1	0.61
Virulent-GO	**82.0**	**83.4**	**80.6**	**0.64**

IV. DISCUSSION

This study proposed an efficient method utilizing instructive GO terms to predict virulent proteins in bacterial pathogens. This method performs well across popular classifiers and also has a significantly better performance than applying features like compositions and evolutionary information. Compared to the existing method, VirulentPred, there is a slight better performance in training that may results from bias originated from applying *k*-fold cross-validation. While performing on independent test dataset, the Virulent-GO still has a little improvement.

For some proteins in the dataset, the BLAST program failed even using a loose value 0.1 of *e* to find homology that is annotated with certain GO terms. These proteins with no BLAST-found homology are usually regarded as virulent proteins. Due to the nature of this GO terms mining method, it could imply two hypotheses: some virulent proteins share less conservations to others or the poor understanding of annotation of their homologies. Whatever the exactly explain is, this could be considering as a character for some virulent proteins for now. As increasing the size and popularity of GO terms, the prediction ability of a GO-based classifier can be further enhanced.

The essential GO terms set provides some insights for virulence factors in bacterial pathogens. First, the essential GO terms is built from set keywords relevant to bacterial virulence factors. Secondly, while correlating to training dataset, virulent proteins tend to annotate with essential GO terms in general. After eliminating proteins without annotated essential GO terms, the used feature set still yield a successful classification result. Moreover, some key GO terms are exactly used in essential GO terms. For example, GO:0009405 named pathogenesis is a key specific processes that generate the ability of an organism to cause disease in another, this GO terms contained a dominate ratio in virulent ratio, more detailed information could be found in the GOA database [16]. In the same way, in this primary event of host-pathogen interaction have revealed a wide array of adhesins to a variety of pathogenic microbes [8], the essential GO terms set contained six GO terms about it, they are GO:0004713, GO:0005102, GO:0005515, GO:0007155, GO:0030246 and GO:0044406. GO:0007155 which is named "cell adhesion", in specific, also have been annotated in 57 virulent proteins and 6 non-virulent proteins. Within a thorough analysis of these essential GO terms, it may reveal some characteristics of virulence factors are associated with bacterial pathogens.

Although using only essential GO terms set could successfully predict virulent proteins, it results in large numbers of false positive and false negative due to a small coverage on the training dataset because it is insufficient to cooperate with whole bacteria genome screening for large amount of protein sequences could be annotated with no essential GO terms. The instructive GO terms could provide a reference contracting functions of non-virulent proteins and virulent proteins. An obviously evidence is the significant difference from amount of GO terms that used to annotate non-virulent proteins (1174) against virulent proteins (599) which is shown in Table II. This could infer that virulent proteins share less functions compared with non-virulent proteins in bacterial pathogens. Thus, applying a feature selection scheme for choosing informative GO terms could certainly improve performance of Virulent-GO. This classifier could also be enhanced if some popular features such as composition information on sequences are added.

The decision tree C4.5 (J48) is used and has high prediction performance. The obtained decision tree using instructive GO terms has 80 leaves and 159 nodes. This result implies that building a set of rules using the proposed informative GO terms consisting of instructive GO terms and essential GO terms is plausible. This interpretable rule set could be worth of further developing and analyzing.

V. Conclusion

This study proposes a well-performed method, Virulent-GO, using instructive GO terms to predict virulent proteins in bacterial pathogens against existing methods. By exploring popular classifiers and compared to some features that are in common usage. For the interpretability oriented from instructive GO terms and essential GO terms, this method is suggested that some novel insights of virulence factors could be discovery resulted from analysis both instructive GO terms and essential GO terms.

By cooperating instructive GO terms set with some popular features, the performance could be further improved. Furthermore, the ranking of GO terms in the contribution of prediction and a set of interpretable prediction rules provide valuable information for more understanding in a complex virulence mechanism in bacterial pathogens.

References

[1] B.B. Finlay, and S. Falkow, Common themes in microbial pathogenicity revisited. Microbiology and Molecular Biology Reviews 61 (1997) 136-&.
[2] H.J. Wu, A.H.J. Wang, and M.P. Jennings, Discovery of virulence factors of pathogenic bacteria. Current Opinion in Chemical Biology 12 (2008) 93-101.
[3] R.A. Weiss, Virulence and pathogenesis. Trends in Microbiology 10 (2002) 314-317.
[4] L.H. Chen, J. Yang, J. Yu, Z.J. Ya, L.L. Sun, Y. Shen, and Q. Jin, VFDB: a reference database for bacterial virulence factors. Nucleic Acids Research 33 (2005) D325-D328.
[5] J. Yang, L.H. Chen, L.L. Sun, J. Yu, and Q. Jin, VFDB 2008 release: an enhanced web-based resource for comparative pathogenomics. Nucleic Acids Research 36 (2008) D539-D542.
[6] A. Bairoch, and R. Apweiler, The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Research 28 (2000) 45-48.
[7] A. Garg, and D. Gupta, VirulentPred: a SVM based prediction method for virulent proteins in bacterial pathogens. Bmc Bioinformatics 9 (2008) -.
[8] G. Sachdeva, K. Kumar, P. Jain, and S. Ramachandran, SPAAN: a software program for prediction of adhesins and adhesin-like proteins using neural networks. Bioinformatics 21 (2005) 483-491.
[9] S.F. Altschul, T.L. Madden, A.A. Schaffer, J.H. Zhang, Z. Zhang, W. Miller, and D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25 (1997) 3389-3402.
[10] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, G. Sherlock, and G.O. Consortium, Gene Ontology: tool for the unification of biology. Nature Genetics 25 (2000) 25-29.
[11] A. Lewin, and I.C. Grieve, Grouping Gene Ontology terms to improve the assessment of gene set enrichment in microarray data. Bmc Bioinformatics 7 (2006) -.
[12] S. Carroll, and V. Pavlovic, Protein classification using probabilistic chain graphs and the Gene Ontology structure. Bioinformatics 22 (2006) 1871-1878.
[13] Z.D. Lei, and Y. Dai, Assessing protein similarity with Gene Ontology and its use in subnuclear localization prediction. Bmc Bioinformatics 7 (2006) -.
[14] Z.L. Qian, Y.D. Cai, and Y.X. Li, A novel computational method to predict transcription factor DNA binding preference. Biochemical and Biophysical Research Communications 348 (2006) 1034-1037.
[15] W.L. Huang, C.W. Tung, S.W. Ho, S.F. Hwang, and S.Y. Ho, ProLoc-GO: Utilizing informative Gene Ontology terms for sequence-based prediction of protein subcellular localization. Bmc Bioinformatics 9 (2008) -.
[16] D. Barrell, E. Dimmer, R.P. Huntley, D. Binns, C. O'Donovan, and R. Apweiler, The GOA database in 2009-an integrated Gene Ontology Annotation resource. Nucleic Acids Research 37 (2009) D396-D403.
[17] K. Chan, and W. Lam, Gene ontology classification of biomedical literatures using context association. Information Retrieval Technology, Proceedings 3689 (2005) 552-557.
[18] D.W. Park, H.S. Heo, H.C. Kwon, and H.Y. Chung, Protein function classification based on gene ontology. Information Retrieval Technology, Proceedings 3689 (2005) 691-696.
[19] S. Altschul, T. Madden, A. Schaffer, J.H. Zhang, Z. Zhang, W. Miller, and D. Lipman, Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Faseb Journal 12 (1998) A1326-A1326.
[20] I.H. Witten, and E. Frank, Data Mining: Practical machine learning tools and techniques, Morgan Kaufmann, San Francisco, 2005.
[21] C. Chang, and C. Lin, LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.. 2001.
[22] S. M, Cross-validatory choice and assessment of statistical predictions. Jounral of the Royal Statistical Society 36 (1974) 111-147.