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],
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, NaiveBayes, 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

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

These keywords are from a tree of hierarchical classes that were discussed in these reviews [1; 2]. Hits defined as the amount of intersected GO terms with informative 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].

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...
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 IBk ($k$-nearest neighbor classifier), NaïveBayes and J48 (C4.5 decision tree). The IBk 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 $\gamma$ 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:

\[ ACC = \frac{(TP + TN)}{(TP + TN + FP + FN)} \times 100\% \]  

\[ SN = \frac{(TP)}{(TP + FN)} \times 100\% \]  

\[ SP = \frac{(TN)}{(TN + FP)} \times 100\% \]  

\[ MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (FN + FP) \times (TN + FN)}} \]  

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 per 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 terms get larger than 3.

![Fig. 2 The number of essential GO terms annotated in each protein is shown in this frequency distribution graph](image)

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...
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 IV

<table>
<thead>
<tr>
<th>No</th>
<th>Access No.</th>
<th>Name</th>
<th>Branch</th>
<th>No</th>
<th>Access No.</th>
<th>Name</th>
<th>Branch</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GO:0000270</td>
<td>peptidoglycan metabolic process</td>
<td>B</td>
<td>38</td>
<td>GO:0009405</td>
<td>pathogenesis</td>
<td>B</td>
</tr>
<tr>
<td>2</td>
<td>GO:0000271</td>
<td>polysaccharide biosynthetic process</td>
<td>B</td>
<td>39</td>
<td>GO:0009425</td>
<td>flagellin-based flagellum basal body</td>
<td>C</td>
</tr>
<tr>
<td>3</td>
<td>GO:0003796</td>
<td>lysozyme activity</td>
<td>M</td>
<td>40</td>
<td>GO:0009428</td>
<td>flagellin-based flagellum basal body, distal rod, P ring</td>
<td>C</td>
</tr>
<tr>
<td>4</td>
<td>GO:0003959</td>
<td>NADPH dehydrogenase activity</td>
<td>M</td>
<td>41</td>
<td>GO:0009626</td>
<td>plant-type hypersensitive response</td>
<td>B</td>
</tr>
<tr>
<td>5</td>
<td>GO:0004222</td>
<td>metalloendopeptidase activity</td>
<td>M</td>
<td>42</td>
<td>GO:0009636</td>
<td>response to toxin</td>
<td>B</td>
</tr>
<tr>
<td>6</td>
<td>GO:0004595</td>
<td>adenylyltransferase activity</td>
<td>M</td>
<td>43</td>
<td>GO:0009842</td>
<td>cyanelle</td>
<td>C</td>
</tr>
<tr>
<td>7</td>
<td>GO:0004713</td>
<td>protein tyrosine kinase activity</td>
<td>M</td>
<td>44</td>
<td>GO:0009986</td>
<td>cell surface</td>
<td>C</td>
</tr>
<tr>
<td>8</td>
<td>GO:0005102</td>
<td>receptor binding</td>
<td>M</td>
<td>45</td>
<td>GO:0015153</td>
<td>rhamnose transmembrane transporter activity</td>
<td>M</td>
</tr>
<tr>
<td>9</td>
<td>GO:0005198</td>
<td>structural molecule activity</td>
<td>M</td>
<td>46</td>
<td>GO:0015267</td>
<td>channel activity</td>
<td>M</td>
</tr>
<tr>
<td>10</td>
<td>GO:0005215</td>
<td>transporter activity</td>
<td>M</td>
<td>47</td>
<td>GO:0015288</td>
<td>porin activity</td>
<td>M</td>
</tr>
<tr>
<td>11</td>
<td>GO:0005515</td>
<td>protein binding</td>
<td>M</td>
<td>48</td>
<td>GO:0015343</td>
<td>siderophore-iron transmembrane transporter activity</td>
<td>M</td>
</tr>
<tr>
<td>12</td>
<td>GO:0005524</td>
<td>ATP binding</td>
<td>M</td>
<td>49</td>
<td>GO:0015627</td>
<td>type II protein secretion system complex</td>
<td>C</td>
</tr>
<tr>
<td>13</td>
<td>GO:0005576</td>
<td>extracellular region</td>
<td>C</td>
<td>50</td>
<td>GO:0015628</td>
<td>protein secretion by the type II secretion system</td>
<td>B</td>
</tr>
<tr>
<td>14</td>
<td>GO:0005618</td>
<td>cell wall</td>
<td>C</td>
<td>51</td>
<td>GO:0015643</td>
<td>toxin binding</td>
<td>M</td>
</tr>
<tr>
<td>15</td>
<td>GO:0005887</td>
<td>integral to plasma membrane</td>
<td>C</td>
<td>52</td>
<td>GO:0015762</td>
<td>rhamnose transport</td>
<td>B</td>
</tr>
<tr>
<td>16</td>
<td>GO:0006281</td>
<td>DNA repair</td>
<td>B</td>
<td>53</td>
<td>GO:0015774</td>
<td>polysaccharide transport</td>
<td>B</td>
</tr>
<tr>
<td>17</td>
<td>GO:0006810</td>
<td>transport</td>
<td>B</td>
<td>54</td>
<td>GO:0015937</td>
<td>coenzyme A biosynthetic process</td>
<td>B</td>
</tr>
<tr>
<td>18</td>
<td>GO:000817</td>
<td>phosphate transport</td>
<td>B</td>
<td>55</td>
<td>GO:0016020</td>
<td>membrane</td>
<td>C</td>
</tr>
<tr>
<td>19</td>
<td>GO:0007155</td>
<td>cell adhesion</td>
<td>B</td>
<td>56</td>
<td>GO:0016908</td>
<td>cell wall catabolic process</td>
<td>B</td>
</tr>
<tr>
<td>20</td>
<td>GO:0019929</td>
<td>rhamnose metabolic process</td>
<td>B</td>
<td>57</td>
<td>GO:0019299</td>
<td>teichoic acid biosynthetic process</td>
<td>B</td>
</tr>
<tr>
<td>21</td>
<td>GO:0008270</td>
<td>zinc ion binding</td>
<td>M</td>
<td>58</td>
<td>GO:0019350</td>
<td>toxic transporter activity</td>
<td>M</td>
</tr>
<tr>
<td>22</td>
<td>GO:0008565</td>
<td>protein transporter activity</td>
<td>M</td>
<td>59</td>
<td>GO:0019534</td>
<td>cytolysis</td>
<td>B</td>
</tr>
<tr>
<td>23</td>
<td>GO:0008898</td>
<td>lipopolysaccharide</td>
<td>M</td>
<td>60</td>
<td>GO:0019835</td>
<td>dianimoipilate biosynthetic process</td>
<td>B</td>
</tr>
<tr>
<td>24</td>
<td>GO:0008919</td>
<td>lipopolysaccharide glucose/transferase activity</td>
<td>M</td>
<td>61</td>
<td>GO:0019867</td>
<td>outer membrane</td>
<td>C</td>
</tr>
<tr>
<td>25</td>
<td>GO:0009103</td>
<td>lipopolysaccharide biosynthetic process</td>
<td>B</td>
<td>62</td>
<td>GO:0019877</td>
<td>carbohydrate binding</td>
<td>M</td>
</tr>
<tr>
<td>26</td>
<td>GO:009244</td>
<td>lipopolysaccharide core region biosynthetic process</td>
<td>B</td>
<td>63</td>
<td>GO:0030246</td>
<td>envelope</td>
<td>C</td>
</tr>
<tr>
<td>27</td>
<td>GO:0009245</td>
<td>lipid A biosynthetic process</td>
<td>B</td>
<td>64</td>
<td>GO:0030288</td>
<td>algicin acid catabolic process</td>
<td>B</td>
</tr>
<tr>
<td>28</td>
<td>GO:0009252</td>
<td>peptidoglycan biosynthetic process</td>
<td>B</td>
<td>65</td>
<td>GO:0031975</td>
<td>algicin acid biologoco process</td>
<td>C</td>
</tr>
<tr>
<td>29</td>
<td>GO:0009253</td>
<td>peptidoglycan catabolic process</td>
<td>B</td>
<td>66</td>
<td>GO:0042121</td>
<td>algicin acid catabolic process</td>
<td>B</td>
</tr>
<tr>
<td>30</td>
<td>GO:0009254</td>
<td>peptidoglycan turnover</td>
<td>B</td>
<td>67</td>
<td>GO:0042212</td>
<td>algicin acid catabolic process</td>
<td>B</td>
</tr>
<tr>
<td>31</td>
<td>GO:0009273</td>
<td>peptidoglycan-based cell wall biogenesis</td>
<td>B</td>
<td>68</td>
<td>GO:0042243</td>
<td>asexual spore wall assembly</td>
<td>B</td>
</tr>
<tr>
<td>32</td>
<td>GO:0009274</td>
<td>peptidoglycan-based cell wall</td>
<td>C</td>
<td>69</td>
<td>GO:0042597</td>
<td>periplasmic space</td>
<td>C</td>
</tr>
<tr>
<td>33</td>
<td>GO:0009275</td>
<td>Gram-negative-bacterium-type cell wall</td>
<td>C</td>
<td>70</td>
<td>GO:0042603</td>
<td>ATP-binding cassette (ABC) transporter complex</td>
<td>C</td>
</tr>
<tr>
<td>34</td>
<td>GO:0009276</td>
<td>Gram-negative-bacterium-type cell wall</td>
<td>C</td>
<td>71</td>
<td>GO:0043190</td>
<td>adhesion to host</td>
<td>M</td>
</tr>
<tr>
<td>35</td>
<td>GO:0009279</td>
<td>cell outer membrane</td>
<td>C</td>
<td>72</td>
<td>GO:004406</td>
<td>capsule polysaccharide biosynthetic process</td>
<td>B</td>
</tr>
<tr>
<td>36</td>
<td>GO:0009306</td>
<td>protein secretion</td>
<td>B</td>
<td>73</td>
<td>GO:0045227</td>
<td>capsule polysaccharide biosynthetic process</td>
<td>B</td>
</tr>
<tr>
<td>37</td>
<td>GO:0009404</td>
<td>toxin metabolic process</td>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A. B is the abbreviation of “Biological Process”; M represents for “Molecular Function”; C is for “Cellular Component”.

Digital Open Science Index, Bioengineering and Life Sciences Vol:3, No:5, 2009 waset.org/Publication/3594

International Scholarly and Scientific Research & Innovation 3(5) 2009 246
TABLE V

<table>
<thead>
<tr>
<th>no</th>
<th>GO term</th>
<th>Branch</th>
<th>Training Dataset</th>
<th>no</th>
<th>GO term</th>
<th>Branch</th>
<th>Training Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GO:0000270</td>
<td>B</td>
<td>1</td>
<td>0</td>
<td>38</td>
<td>GO:00009405</td>
<td>B</td>
</tr>
<tr>
<td>2</td>
<td>GO:0000271</td>
<td>M</td>
<td>2</td>
<td>7</td>
<td>40</td>
<td>GO:0009428</td>
<td>C</td>
</tr>
<tr>
<td>3</td>
<td>GO:0003959</td>
<td>M</td>
<td>1</td>
<td>0</td>
<td>41</td>
<td>GO:0009626</td>
<td>B</td>
</tr>
<tr>
<td>4</td>
<td>GO:0004222</td>
<td>M</td>
<td>6</td>
<td>11</td>
<td>42</td>
<td>GO:0009636</td>
<td>B</td>
</tr>
<tr>
<td>5</td>
<td>GO:0004595</td>
<td>M</td>
<td>0</td>
<td>1</td>
<td>43</td>
<td>GO:0009842</td>
<td>C</td>
</tr>
<tr>
<td>6</td>
<td>GO:0004713</td>
<td>M</td>
<td>2</td>
<td>3</td>
<td>44</td>
<td>GO:0009986</td>
<td>C</td>
</tr>
<tr>
<td>7</td>
<td>GO:0005102</td>
<td>M</td>
<td>4</td>
<td>19</td>
<td>45</td>
<td>GO:0015153</td>
<td>M</td>
</tr>
<tr>
<td>8</td>
<td>GO:0005198</td>
<td>M</td>
<td>8</td>
<td>22</td>
<td>46</td>
<td>GO:0015267</td>
<td>M</td>
</tr>
<tr>
<td>9</td>
<td>GO:0005215</td>
<td>M</td>
<td>95</td>
<td>28</td>
<td>47</td>
<td>GO:0015288</td>
<td>M</td>
</tr>
<tr>
<td>10</td>
<td>GO:0005515</td>
<td>M</td>
<td>103</td>
<td>112</td>
<td>48</td>
<td>GO:0015343</td>
<td>M</td>
</tr>
<tr>
<td>11</td>
<td>GO:0005524</td>
<td>M</td>
<td>202</td>
<td>64</td>
<td>49</td>
<td>GO:0015627</td>
<td>B</td>
</tr>
<tr>
<td>12</td>
<td>GO:0005576</td>
<td>M</td>
<td>49</td>
<td>166</td>
<td>50</td>
<td>GO:0015628</td>
<td>B</td>
</tr>
<tr>
<td>13</td>
<td>GO:0005618</td>
<td>M</td>
<td>2</td>
<td>3</td>
<td>51</td>
<td>GO:0015643</td>
<td>M</td>
</tr>
<tr>
<td>14</td>
<td>GO:0005887</td>
<td>C</td>
<td>2</td>
<td>0</td>
<td>52</td>
<td>GO:0015762</td>
<td>B</td>
</tr>
<tr>
<td>15</td>
<td>GO:0006281</td>
<td>B</td>
<td>49</td>
<td>5</td>
<td>53</td>
<td>GO:0015774</td>
<td>B</td>
</tr>
<tr>
<td>16</td>
<td>GO:0006810</td>
<td>B</td>
<td>305</td>
<td>203</td>
<td>54</td>
<td>GO:0015937</td>
<td>B</td>
</tr>
<tr>
<td>17</td>
<td>GO:0007047</td>
<td>B</td>
<td>21</td>
<td>27</td>
<td>56</td>
<td>GO:0016098</td>
<td>B</td>
</tr>
<tr>
<td>18</td>
<td>GO:0007155</td>
<td>B</td>
<td>6</td>
<td>57</td>
<td>57</td>
<td>GO:0019299</td>
<td>B</td>
</tr>
<tr>
<td>19</td>
<td>GO:0008270</td>
<td>M</td>
<td>58</td>
<td>38</td>
<td>58</td>
<td>GO:0019350</td>
<td>M</td>
</tr>
<tr>
<td>20</td>
<td>GO:0008565</td>
<td>M</td>
<td>10</td>
<td>50</td>
<td>59</td>
<td>GO:0019354</td>
<td>B</td>
</tr>
<tr>
<td>21</td>
<td>GO:0008918</td>
<td>M</td>
<td>1</td>
<td>0</td>
<td>60</td>
<td>GO:0019355</td>
<td>B</td>
</tr>
<tr>
<td>22</td>
<td>GO:0009819</td>
<td>M</td>
<td>1</td>
<td>2</td>
<td>61</td>
<td>GO:0019867</td>
<td>C</td>
</tr>
<tr>
<td>23</td>
<td>GO:0009913</td>
<td>B</td>
<td>15</td>
<td>38</td>
<td>62</td>
<td>GO:0019877</td>
<td>B</td>
</tr>
<tr>
<td>24</td>
<td>GO:0009244</td>
<td>B</td>
<td>2</td>
<td>0</td>
<td>63</td>
<td>GO:0030246</td>
<td>M</td>
</tr>
<tr>
<td>25</td>
<td>GO:0009245</td>
<td>B</td>
<td>8</td>
<td>2</td>
<td>64</td>
<td>GO:0030288</td>
<td>C</td>
</tr>
<tr>
<td>26</td>
<td>GO:0009252</td>
<td>B</td>
<td>13</td>
<td>1</td>
<td>65</td>
<td>GO:0031975</td>
<td>M</td>
</tr>
<tr>
<td>27</td>
<td>GO:0009253</td>
<td>B</td>
<td>4</td>
<td>8</td>
<td>66</td>
<td>GO:0042121</td>
<td>B</td>
</tr>
<tr>
<td>28</td>
<td>GO:0009254</td>
<td>B</td>
<td>2</td>
<td>0</td>
<td>67</td>
<td>GO:0042122</td>
<td>B</td>
</tr>
<tr>
<td>29</td>
<td>GO:0009273</td>
<td>B</td>
<td>5</td>
<td>2</td>
<td>68</td>
<td>GO:0042243</td>
<td>B</td>
</tr>
<tr>
<td>30</td>
<td>GO:0009274</td>
<td>C</td>
<td>2</td>
<td>1</td>
<td>69</td>
<td>GO:0042597</td>
<td>C</td>
</tr>
<tr>
<td>31</td>
<td>GO:0009275</td>
<td>C</td>
<td>0</td>
<td>1</td>
<td>70</td>
<td>GO:0042603</td>
<td>C</td>
</tr>
<tr>
<td>32</td>
<td>GO:0009276</td>
<td>C</td>
<td>22</td>
<td>14</td>
<td>71</td>
<td>GO:0043190</td>
<td>C</td>
</tr>
<tr>
<td>33</td>
<td>GO:0009279</td>
<td>C</td>
<td>29</td>
<td>75</td>
<td>72</td>
<td>GO:0044406</td>
<td>B</td>
</tr>
<tr>
<td>34</td>
<td>GO:0009306</td>
<td>B</td>
<td>15</td>
<td>67</td>
<td>73</td>
<td>GO:0045227</td>
<td>B</td>
</tr>
<tr>
<td>35</td>
<td>GO:0009404</td>
<td>B</td>
<td>0</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 VI

<table>
<thead>
<tr>
<th>Classifier</th>
<th>ACC (%)</th>
<th>SN (%)</th>
<th>SP (%)</th>
<th>MCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>IB1</td>
<td>78.6</td>
<td>77.3</td>
<td>79.7</td>
<td>0.57</td>
</tr>
<tr>
<td>IB3</td>
<td>76.8</td>
<td>73.3</td>
<td>81.6</td>
<td>0.54</td>
</tr>
<tr>
<td>IB5</td>
<td>73.4</td>
<td>68.4</td>
<td>82.0</td>
<td>0.49</td>
</tr>
<tr>
<td>NaïveBayes - Normal Distribution</td>
<td>78.3</td>
<td>77.0</td>
<td>79.4</td>
<td>0.56</td>
</tr>
<tr>
<td>NaïveBayes - Kernel Density Estimator</td>
<td>79.5</td>
<td>74.8</td>
<td>86.3</td>
<td>0.60</td>
</tr>
<tr>
<td>SVM</td>
<td>82.5</td>
<td>84.5</td>
<td>80.6</td>
<td>0.65</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Feature</th>
<th>Classifier ACC (%)</th>
<th>SN (%)</th>
<th>SP (%)</th>
<th>MCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amino Acid Compositions</td>
<td>SVM</td>
<td>72.1</td>
<td>70.0</td>
<td>74.1</td>
</tr>
<tr>
<td>Dipeptide Compositions (i+1st)</td>
<td>SVM</td>
<td>71.1</td>
<td>70.0</td>
<td>72.3</td>
</tr>
<tr>
<td>Dipeptide Compositions (i+2nd)</td>
<td>SVM</td>
<td>72.0</td>
<td>70.2</td>
<td>73.7</td>
</tr>
<tr>
<td>PSI-BLAST Search</td>
<td>---</td>
<td>52.1</td>
<td>52.5</td>
<td>51.7</td>
</tr>
<tr>
<td>Position-Specific Scoring Matrix</td>
<td>SVM</td>
<td>78.1</td>
<td>78.1</td>
<td>78.1</td>
</tr>
<tr>
<td>GO terms</td>
<td>SVM</td>
<td>82.5</td>
<td>84.5</td>
<td>80.6</td>
</tr>
</tbody>
</table>

* These results are from the previous method, VirulentPred.
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>
<thead>
<tr>
<th>Classifiers</th>
<th>Training Dataset-1</th>
<th>Training Dataset-2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ACC (%)</td>
<td>SN (%)</td>
</tr>
<tr>
<td>SVM</td>
<td>81.8</td>
<td>82.0</td>
</tr>
<tr>
<td>NaïveBayes - Kernel Density Estimator</td>
<td>82.0</td>
<td>83.4</td>
</tr>
</tbody>
</table>

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 contains a dominant 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 informative 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 informative GO terms and essential GO terms, this method is suggested that some novel insights of virulence factors could be discovery resulted from analysis both informative GO terms and essential GO terms.

By cooperating informative 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