Semantic Similarity over Gene Ontology for Multi-Label Protein Subcellular Localization

Shibiao Wan¹, Man-Wai Mak¹, Sun-Yuan Kung²
¹Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China
²Department of Electrical Engineering, Princeton University, New Jersey, USA
Email: shibiao.wan@polyu.edu.hk, enmwmak@polyu.edu.hk, kung@princeton.edu

Received November 2012

ABSTRACT

As one of the essential topics in proteomics and molecular biology, protein subcellular localization has been extensively studied in previous decades. However, most of the methods are limited to the prediction of single-location proteins. In many studies, multi-location proteins are either not considered or assumed not existing. This paper proposes a novel multi-label subcellular-localization predictor based on the semantic similarity between Gene Ontology (GO) terms. Given a protein, the accession numbers of its homologs are obtained via BLAST search. Then, the homologous accession numbers of the protein are used as keys to search against the gene ontology annotation database to obtain a set of GO terms. The semantic similarity between GO terms is used to formulate semantic similarity vectors for classification. A support vector machine (SVM) classifier with a new decision scheme is proposed to classify the multi-label GO semantic similarity vectors. Experimental results show that the proposed multi-label predictor significantly outperforms the state-of-the-art predictors such as iLoc-Plant and Plant-mPLoc.

Keywords: Protein Subcellular Localization; Semantic Similarity; GO Terms; Multi-Label Classification

1. Introduction

In recent years, protein subcellular localization has gained tremendous attention due to its important roles in elucidating protein functions, identifying drug targets, and so on [1]. Computational methods are required to replace time-consuming and laborious wet-lab methods for predicting the subcellular locations of proteins.

Conventional methods for subcellular-localization prediction can be roughly divided into sequence-based methods [2-6] and annotation-based methods [7-13]. It has been demonstrated that methods based on Gene Ontology are superior [10]. However, most of the existing methods are limited to the prediction of single-location proteins. These methods generally exclude the multi-label proteins or are based on the assumption that multi-location proteins do not exist. In fact, there exist multi-location proteins that can simultaneously reside at, or move between, two or more different subcellular locations. Recently, several multi-label predictors have been proposed, including Plant-mPLoc [14], Virus-mPLoc [15], iLoc-Plant [16] and iLoc-Virus [17]. These predictors use the GO information and have demonstrated superiority over other methods. But these predictors only make use of the occurrences of the GO terms and do not exploit the semantic relationships between GO terms.

Since the relationship between GO terms reflects the association between different gene products, protein sequences annotated with GO terms can be compared on the basis of semantic similarity measures. Actually, the semantic similarity over Gene Ontology has been extensively studied and have been applied in many biological problems, including protein function prediction [18], subnuclear localization prediction [19], protein-protein interaction inference [20] and microarray clustering [21]. The performance of these predictors depends on whether the similarity measure is relevant to the biological problems. Over the years, a number of semantic similarity measures have been proposed, some of which have been used in natural language processing. For example, Resnik [22] proposed the information content of terms in natural language as a similarity measure. Later, Lord et al. [23] introduced this idea into measuring the semantic similarity of GO terms. Lin et al. [24] proposed a method based on information theory and structural information. More recently, Pesquita et al. [25] reviewed the semantic similarity measures applied to biomedical ontologies.

This paper proposes a novel predictor based on the GO semantic similarity for multi-label protein subcellular localization prediction. The predictor proposed is different from other predictors in that 1) it formulates the fea-
ture vectors by the semantic similarity over Gene Ontol-
yogy which contains richer information than only GO
terms; 2) it adopts a new strategy to incorporate richer
and more useful homologous information from more
distant homologs rather than using the top homologs only;
3) it adopts a new decision scheme for an SVM classifi-
er so that it can effectively deal with datasets containing
both single-label and multi-label proteins. Results on a
recent benchmark dataset demonstrate that these three
properties enable the proposed predictor to accurately
predict multi-location proteins and outperform three
state-of-the-art predictors.

2. Method

2.1. Retrieval of GO Terms

The proposed predictor can use either the accession
numbers (AC) or amino acid (AA) sequences of query
proteins as input. Specifically, for proteins with known
ACs, their respective GO terms are retrieved from the
Gene Ontology Annotation (GOA) database using the
ACs as the searching keys. For proteins without ACs,
their AA sequences are presented to BLAST to find
their homologs, whose ACs are then used as keys to
search against the GOA database.

While the GOA database allows us to associate the AC
of a protein with a set of GO terms, for some novel pro-
teins, neither their ACs nor the ACs of their top homo-
logs have any entries in the GOA database; in other
words, no GO terms can be retrieved by using their ACs
or the ACs of their top homologs. In such case, the ACs
of the homologous proteins, as returned from BLAST
search, will be successively used to search against the
GOA database until a match is found. With the rapid
progress of the GOA database, it is reasonable to assume
that the homologs of the query proteins have at least one
GO term [12]. Thus, it is not necessary to use back-up
methods to handle the situation where no GO terms can
be found. The procedures are outlined in Figure 1.

2.2. Semantic Similarity Measure

To obtain the GO semantic similarity between two pro-
teins, we should start by introducing the semantic simi-
larity between two GO terms. The semantic similarity
between two categories is based on the information con-
tent. As suggested by Resnik [22], the similarity measure
of two categories relies on the most specific common
ancestor in the GO hierarchy. The semantic similarity
between two GO terms \( x \) and \( y \) is defined as [22]:

\[
\text{sim}(x, y) = \max_{c \in A(x,y)} \left[ -\log(p(c)) \right]
\]

(1)

where \( A(x,y) \) is the set of ancestor GO terms of both \( x \) and \( y \), and \( p(c) \) is the number of gene products annotated
to the GO term \( c \) divided by the number of all the gene
products annotated to the GO taxonomy.

To further incorporate structural information from the
GO hierarchy, we used Lin’s measures [24] to normalize
the above measure. Then given two GO terms \( x \) and \( y \),
the similarity is calculated as:

\[
\text{sim}(x, y) = \frac{2 \times \max_{c \in A(x,y)} \left[ -\log(p(c)) \right]}{-\log(p(x)) - \log(p(y))}
\]

(2)

Based on the semantic similarity between two GO
terms, we adopted a continuous measure proposed in [21]
to calculate the similarity of two proteins, which are
functionally annotated by a set of GO terms. Given two
proteins \( P_i \) and \( P_j \) which are annotated by two sets of
GO terms \( G_i \) and \( G_j \) retrieved in Section II-A, we first
computed \( S(G_i, G_j) \) as follows:

\[
S(G_i, G_j) = \sum_{x \in G_i} \sum_{y \in G_j} \text{sim}(x, y)
\]

(3)

where \( \text{sim}(x, y) \) is defined in Equation (2).

---

1http://www.ebi.ac.uk/GOA

2The relationships between GO terms in the GO hierarchy, such as “is-a”
ancestor-child, or “part-of” ancestor-child can be obtained from the
SQL database through the link: http://archive.geneontology.org/latest-
termdb/go_daily-termdb-tables.tar.gz. Note here only the “is-a” rela-
tionship is considered for semantic similarity analysis [22].

---

Copyright © 2013 SciRes.
Then, $SS(G_i, G_j)$ is computed in the same way by swapping $G_i$ and $G_j$. Finally, the overall similarity between the two proteins is given by:

$$SS(G_i, G_j) = \frac{S(G_i, G_j) + S(G_j, G_i)}{S(G_i, G_j) + S(G_j, G_i)}.$$  \hspace{1cm} (4)

Thus, for a testing protein $Q$, a GO semantic similarity vector $q$ can be formulated by performing pairwise comparisons with every training protein $\{P_i\}_{i=1}^N$, where $N$ is the number of training proteins. Then, $q$ can be represented as:

$$q = [SS(Q, G_1), \ldots, SS(Q, G_i), \ldots, SS(Q, G_j)]^T$$  \hspace{1cm} (5)

where $Q$ is the set of GO terms for the test protein $Q$.

### 2.3. Multi-Label Multi-Class SVM Classification

To predict the subcellular locations of both single-label and multi-label proteins, a multi-label support vector machine (SVM) classifier is proposed in this paper. Specifically, denote the GO semantic similarity vector of the $t$-th query protein as $q_t$. Then, given the $t$-th query protein $Q_t$, the score of the $m$-th SVM is

$$s_m(Q_t) = \sum_{i \in S_m} a_{m,i} y_{m,i} K(p_i, q_t) + b_m,$$  \hspace{1cm} (6)

where $S_m$ is the set of support vector indexes corresponding to the $m$-th SVM, $a_{m,i}$ are the Lagrange multipliers, $K(\cdot, \cdot)$ is a kernel function; here, the linear kernel is used. $y_{m,i} \in \{-1, +1\}$ are the class labels.

Unlike the single-label problem where each protein has one predicted label only, a multi-label protein could have more than one predicted labels. Thus, the predicted subcellular location(s) of the $t$-th query protein are given by:

$$\mathcal{M}(Q_t) = \bigcup_{w=1}^{M} \{m : s_m(Q_t) > 0\}, \text{ where } \exists s_m(Q_t) > 0;$$  \hspace{1cm} (7)

$$\text{arg max}_{w=1}^{M} s_m(Q_t), \text{ otherwise.}$$

### 3. Results

#### 3.1. Dataset and Performance Metrics

In this paper, the plant dataset used in Plant-mPLoc [14], iLoc-Plant [16] and mGOASVM [27] were used to evaluate the performance of the proposed predictor. The plant dataset was created from Swiss-Prot 55.3. It contains 978 plant proteins distributed in 12 locations. Of the 978 plant proteins, 904 belong to one subcellular location, 71 to two locations, 3 to three locations and none to four or more locations. In other words, 8% of the plant proteins in this dataset are located in multiple locations. The sequence identity of this dataset was cut off at 25%.

To facilitate comparison, the locative accuracy [28] and the actual accuracy were used to assess the prediction performance. Specifically, denote $L(p_i)$ and $M(p_i)$ as the true label set and the predicted label set for the $i$-th protein $p_i (i = 1, \ldots, N_{act})$, respectively. Then, the overall locative accuracy is:

$$\Lambda_{loc} = \frac{1}{N_{loc}} \sum_{i=1}^{N_{loc}} |L(p_i) \cap M(p_i)|.$$  \hspace{1cm} (8)

where $|\cdot|$ means counting the number of elements in the set therein and $\cap$ represents the intersection of sets, $N_{act}$ represents the total number of actual proteins and $N_{loc}$ represents the total number of locative proteins. And the overall actual accuracy is:

$$\Lambda_{act} = \frac{1}{N_{act}} \sum_{i=1}^{N_{act}} \Delta[L(p_i) \cap M(p_i)].$$  \hspace{1cm} (9)

where

$$\Delta[L(p_i) \cap M(p_i)] = \begin{cases} 1, & \text{if } L(p_i) = M(p_i) \\ 0, & \text{otherwise.} \end{cases}$$  \hspace{1cm} (10)

Note that the actual accuracy is more objective and stricter than the locative accuracy [27].

#### 3.2. Comparing with State-of-the-Art Predictors

Table 1 compares the performance of the proposed predictor against three state-of-the-art multi-label predictors on the plant dataset. Plant-mPLoc [14], iLoc-Plant [16] and mGOASVM [27] use the accession numbers of homologs returned from BLAST [26] as searching keys to retrieve GO terms from the GOA database. For a fair comparison with these predictors, the performance of our proposed predictor shown in Table 1 was obtained by using the accession numbers of homologous proteins as the searching keys. Unlike Plant-mPLoc and iLoc-Plant, the ACs of the homologous proteins, as returned from BLAST search, will be successively used to search against the GOA database until a match is found (See Figure 1 for details).

As shown in Table 1, our proposed predictor performs significantly better than Plant-mPLoc and iLoc-Plant. Both the overall locative accuracy and overall actual accuracy of mGOASVM are more than 20% (absolute) higher than iLoc-Plant (97.9% vs 71.7% and 89.6% vs 68.1%, respectively). Our proposed predictor also performs better than mGOASVM in terms of both the overall actual accuracy (89.6% vs 97.4%) and the overall locative accuracy (97.9% vs 96.2%). As for the individual locative accuracy, the individual locative accuracies of our proposed predictor for all of the 12 locations are impressively higher than those of Plant-mPLoc, iLoc-Plant and mGOASVM.

In terms of GO information extraction, Plant-mPLoc, iLoc-Plant and mGOASVM only exploit the occurrences

---

Copyright © 2013 SciRes.

[http://bioinfo.eie.polyu.edu.hk/mGoupSvmServer/mGOASVM.html](http://bioinfo.eie.polyu.edu.hk/mGoupSvmServer/mGOASVM.html)
Table 1. Comparing the proposed predictor with state-of-the-art multi-label predictors based on leave-one-out cross validation (LOOCV). “-” means the corresponding references do not provide the overall actual accuracy.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cell membrane</td>
<td>24/56 = 42.9%</td>
<td>39/56 = 69.6%</td>
<td>53/56 = 94.6%</td>
<td>55/56 = 98.2%</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Cell wall</td>
<td>8/32 = 25.0%</td>
<td>19/32 = 59.4%</td>
<td>27/32 = 84.4%</td>
<td>28/32 = 87.5%</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Chloroplast</td>
<td>248/286 = 86.7%</td>
<td>252/286 = 88.1%</td>
<td>272/286 = 95.1%</td>
<td>285/286 = 99.7%</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Chloroplast</td>
<td>72/182 = 39.6%</td>
<td>114/182 = 62.6%</td>
<td>174/182 = 95.6%</td>
<td>175/182 = 96.2%</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Endoplasmic</td>
<td>17/42 = 40.5%</td>
<td>21/42 = 50.0%</td>
<td>38/42 = 90.5%</td>
<td>40/42 = 95.2%</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Extracellular</td>
<td>3/22 = 13.6%</td>
<td>2/22 = 9.1%</td>
<td>22/22 = 100.0%</td>
<td>22/22 = 100.0%</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Golgi apparatus</td>
<td>6/21 = 28.6%</td>
<td>16/21 = 76.2%</td>
<td>19/21 = 90.5%</td>
<td>18/21 = 85.7%</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Mitochondrion</td>
<td>114/150 = 76.0%</td>
<td>112/150 = 74.7%</td>
<td>150/150 = 100.0%</td>
<td>150/150 = 100.0%</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Nucleus</td>
<td>136/152 = 89.5%</td>
<td>140/152 = 92.1%</td>
<td>151/152 = 99.3%</td>
<td>150/152 = 98.7%</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Peroxisome</td>
<td>14/21 = 66.7%</td>
<td>6/21 = 28.6%</td>
<td>21/21 = 100.0%</td>
<td>21/21 = 100.0%</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Plastid</td>
<td>4/39 = 10.3%</td>
<td>7/39 = 17.9%</td>
<td>39/39 = 100.0%</td>
<td>39/39 = 100.0%</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Vacuole</td>
<td>26/52 = 50.0%</td>
<td>28/52 = 53.8%</td>
<td>49/52 = 94.2%</td>
<td>50/52 = 96.2%</td>
<td></td>
</tr>
</tbody>
</table>

Overall Locative Accuracy: 672/1055 = 63.7% 756/1055 = 71.7% 1015/1055 = 96.2% 1033/1055 = 97.9%

Overall Actual Accuracy: 666/978 = 68.1% 855/978 = 87.4% 876/978 = 89.6%

of GO terms, whereas the proposed predictor discovers the semantic relationships between GO terms, based on which the semantic similarity between proteins (from the GO annotation perspective) can be obtained. The superior performance of the proposed predictor clearly suggests that the semantic similarity over Gene Ontology is conducive to the prediction of multi-label protein subcellular localization.

4. Conclusions and Future Works

This paper proposes a new multi-label predictor based on Gene Ontology semantic similarity to predict the subcellular locations of multi-label proteins. By using the accession numbers of the homologs of the query proteins as the searching keys to search against the GO annotation database, the GO terms of each query protein are retrieved. Then the information of the semantic similarity over GO terms is exploited, which is further utilized to formulate GO semantic similarity vectors for every query protein. The feature vectors are subsequently recognized by support vectors machine (SVM) classifiers equipped with a decision strategy that can produce multiple class labels for a query protein. Experimental results demonstrate that the proposed predictor can efficiently predict the subcellular locations of multi-label proteins. It was also found that the exploitation of the semantic similarity over Gene Ontology is conducive to multi-label protein subcellular localization prediction. There are many different methods [20,22,23] for measuring the GO semantic similarity. The semantic similarity measure used in this paper may not be the best for protein subcellular location. Therefore, as a future work, it is of interest to develop a similarity measure that is more relevant to subcellular localization.

5. Acknowledgements

This work was in part supported by The HK RGC Grant No. PolyUS264/09E and HKPolyU Grant No. G-YJ86. We also thank Lixin Cheng for his early effort on the perl scripts.

REFERENCES


