Engineering, **2013, 5, 68-72**
http://dx.doi.org/10.4236/eng.2013.510B014 Published Online October 2013 (http://www.scirp.org/journal/eng)

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

Shibiao Wan[1], Man-Wai Mak[1], Sun-Yuan Kung[2]
[1]Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China
[2]Department 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 Ontology 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 classifier 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[1] using the ACs as the searching keys. For proteins without ACs, their AA sequences are presented to BLAST [26] 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 proteins, neither their ACs nor the ACs of their top homologs 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**.

Given a query protein $\mathbf{Q}_i$,

1. if the AC of the query protein is known,
   use the AC to retrieve a set of GO terms $\mathcal{G}_{i,0}$;
   $k = 0$;
   go to Step 3;
   end

2. if $|\mathcal{G}_{i,0}| = 0$ or only the AA sequence of the query protein is known,
   retrieve $k_{\max}$ homologs of the query protein from BLAST using the AA sequence as input;
   $k = 1$;
   while $k \le k_{\max}$,
   use the AC of the k-th homologous protein to retrieve a set of GO terms $\mathcal{G}_{i,k}$;
   if $|\mathcal{G}_{i,k}| \neq 0$,
   go to Step 3;
   else
   $k = k + 1$;
   end
   end
   end

3. Perform semantic similarity measure (Eq. 5) and multi-label SVM classification (Eq. 6 and Eq. 7).

Figure 1. Procedures of retrieving GO terms.

2.2. Semantic Similarity Measure

To obtain the GO semantic similarity between two proteins, we should start by introducing the semantic similarity between two GO terms. The semantic similarity between two categories is based on the information content. As suggested by Resnik [22], the similarity measure of two categories relies on the most specific common ancestor in the GO hierarchy[2]. The semantic similarity between two GO terms x and y is defined as [22]:

$$\mathrm{sim}(x,y) = \max_{c \in A(x,y)} \left[-\log\left(p(c) \right) \right] \tag{1}$$

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:

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

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 $\mathbf{P}i$ and $\mathbf{P}j$, which are annotated by two sets of GO terms $\mathcal{G}_i$ and $\mathcal{G}_j$ retrieved in Section II-A[3], we first computed $S(\mathcal{G}_i, \mathcal{G}_j)$ as follows:

$$S\left(\mathcal{G}_i, \mathcal{G}_j\right) = \sum_{x \in \mathcal{G}_i} \max_{y \in \mathcal{G}_j} \mathrm{sim}(x,y) \tag{3}$$

where sim(x, y) is defined in Equation (2).

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

[2]The 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" relationship is considered for semantic similarity analysis [22].

[3]Strictly speaking, $\mathcal{G}_i$ should be $\mathcal{G}_{i,k_i}$, where k_i is the k_i-th homolog used to retrieve the GO terms in Section II-A for the i-th protein. To simplify notations, we write it as $\mathcal{G}_i$.

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

$$SS\left(\mathcal{G}_i,\mathcal{G}_j\right)=\frac{S\left(\mathcal{G}_i,\mathcal{G}_j\right)+S\left(\mathcal{G}_j,\mathcal{G}_i\right)}{S\left(\mathcal{G}_i,\mathcal{G}_i\right)+S\left(\mathcal{G}_j,\mathcal{G}_j\right)}. \quad (4)$$

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

$$\mathbf{q}_t=\left[SS\left(\mathcal{Q}_t,\mathcal{G}_1\right),\cdots,SS\left(\mathcal{Q}_t,\mathcal{G}_i\right),\cdots,SS\left(\mathcal{Q}_t,\mathcal{G}_N\right)\right]^{\mathbf{T}} \quad (5)$$

where $\mathbf{Q}_t$ is the set of GO terms for the test protein Q_t.

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 $\mathbf{q}_t$. Then, given the t-th query protein $\mathbf{Q}_t$, the score of the m-th SVM is

$$s_m(\mathbf{Q}_t)=\sum_{r\in S_m}\alpha_{m,r}y_{m,r}K\left(\mathbf{p}_r,\mathbf{q}_t\right)+b_m, \quad (6)$$

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

$$\mathcal{M}^*\left(\mathbf{Q}_t\right)=\begin{cases}\bigcup_{m=1}^{M}\left\{m:\ s_m\left(\mathbf{Q}_t\right)>0\right\}, \text{ where } \exists\, s_m\left(\mathbf{Q}_t\right)>0;\\ \arg\max_{m=1}^{M} s_m\left(\mathbf{Q}_t\right), \text{ otherwise.}\end{cases} \quad (7)$$

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][4] 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 $L(\mathbf{p}_i)$ and $M(\mathbf{p}_i)$ as the true label set and the predicted label set for the i-th protein $\mathbf{p}_i$ (i = 1,…, N_{act}), respectively. Then, the overall locative accuracy is:

$$\Lambda_{loc}=\frac{1}{N_{loc}}\sum_{i=1}^{N_{act}}\left|\mathcal{M}\left(\mathbf{p}_i\right)\cap\mathcal{L}\left(\mathbf{p}_i\right)\right|, \quad (8)$$

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:

$$\Lambda_{act}=\frac{1}{N_{act}}\sum_{i=1}^{N_{act}}\Delta\left|\mathcal{M}\left(\mathbf{p}_i\right)\cap\mathcal{L}\left(\mathbf{p}_i\right)\right|, \quad (9)$$

where

$$\Delta\left|\mathcal{M}\left(\mathbf{p}_i\right)\cap\mathcal{L}\left(\mathbf{p}_i\right)\right|=\begin{cases}1 & , if\ \mathcal{M}\left(\mathbf{p}_i\right)\equiv\mathcal{L}\left(\mathbf{p}_i\right)\\ 0 & , \text{otherwise}\end{cases}. \quad (10)$$

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 I** 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

[4]http://bioinfo.eie.polyu.edu.hk/mGoaSvmServer/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.

Label	Subcellular Location	LOOCV Locative Accuracy			
		Plant-mPLoc [12]	iLoc-Plant [14]	mGOASVM [25]	Proposed Predictor
1	Cell membrane	24/56 = 42.9%	39/56 = 69.6%	53/56 = 94.6%	55/56 = 98.2%
2	Cell wall	8/32 = 25.0%	19/32 = 59.4%	27/32 = 84.4%	28/32 = 87.5%
3	Chloroplast	248/286 = 86.7%	252/286 = 88.1%	272/286 = 95.1%	285/286 = 99.7%
4	Chloroplast	72/182 = 39.6%	114/182 = 62.6%	174/182 = 95.6%	175/182 = 96.2%
5	Endoplasmic	17/42 = 40.5%	21/42 = 50.0%	38/42 = 90.5%	40/42 = 95.2%
6	Extracellular	3/22 = 13.6%	2/22 = 9.1%	22/22 = 100.0%	22/22 = 100.0%
7	Golgi apparatus	6/21 = 28.6%	16/21 = 76.2%	19/21 = 90.5%	18/21 = 85.7%
8	Mitochondrion	114/150 = 76.0%	112/150 = 74.7%	150/150 = 100.0%	150/150 = 100.0%
9	Nucleus	136/152 = 89.5%	140/152 = 92.1%	151/152 = 99.3%	150/152 = 98.7%
10	Peroxisome	14/21 = 66.7%	6/21 = 28.6%	21/21 = 100.0%	21/21 = 100.0%
11	Plastid	4/39 = 10.3%	7/39 = 17.9%	39/39 = 100.0%	39/39 = 100.0%
12	Vacuole	26/52 = 50.0%	28/52 = 53.8%	49/52 = 94.2%	50/52 = 96.2%
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. PolyU5264/09E and HKPolyU Grant No. G-YJ86. We also thank Lixin Cheng for his early effort on the perl scripts.

REFERENCES

[1] K. C. Chou and Y. D. Cai, "Predicting Protein Localization in Budding Yeast," *Bioinformatics*, Vol. 21, 2005, pp. 944-950. http://dx.doi.org/10.1093/bioinformatics/bti104

[2] H. Nakashima and K. Nishikawa, "Discrimination of Intracellular and Extracellular Proteins Using Amino Acid Composition and Residue-Pair Frequencies," *Journal of Molecular Biology*, Vol. 238, 1994, pp. 54-61. http://dx.doi.org/10.1002/prot.1035

[3] K. C. Chou, "Prediction of Protein Cellular Attributes Using Pseudo Amino Acid Composition," *Proteins*: *Structure, Function, and Genetics*, Vol. 43, 2001, pp. 246-255. http://dx.doi.org/10.1002/prot.1035

[4] O. Emanuelsson, H. Nielsen, S. Brunak and G. von Heijne, "Predicting Subcellular Localization of Proteins Based on Their N-Terminal Amino Acid Sequence," *Journal of Molecular Biology*, Vol. 300, No. 4, 2000, pp. 1005-1016. http://dx.doi.org/10.1006/jmbi.2000.3903

[5] H. Nielsen, J. Engelbrecht, S. Brunak and G. von Heijne, "A Neural Network Method for Identification of Prokaryotic and Eukaryotic Signal Peptides and Prediction of Their Cleavage Sites," *International Journal of Neural*

Systems, Vol. 8, 1997, pp. 581-599. http://dx.doi.org/10.1142/S0129065797000537

[6] M. W. Mak, J. Guo and S. Y. Kung, "PairProSVM: Protein Subcellular Localization Based on Local Pairwise Profile Alignment and SVM," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 5, No. 3, 2008, pp. 416-422.

[7] S. Wan, M. W. Mak and S. Y. Kung, "Protein Subcellular Localization Prediction Based on Profile Alignment and Gene Ontology," 2011 *IEEE International Workshop on Machine Learning for Signal Processing* (*MLSP*'11), September 2011, pp. 1-6.

[8] K. C. Chou and Y. D. Cai, "Prediction of Protein Subcellular Locations by GO-FunD-PseAA Predictor," *Biochemical and Biophysical Research Communications*, Vol. 320, 2004, pp. 1236-1239. http://dx.doi.org/10.1016/j.bbrc.2004.06.073

[9] S. Wan, M. W. Mak and S. Y. Kung, "GOASVM: A Subcellular Location Predictor by Incorporating Term-Frequency Gene Ontology into the General Form of Chou's Pseudo-Amino Acid Composition," *Journal of Theoretical Biology*, Vol. 323, 2013, pp. 40-48. http://dx.doi.org/10.1016/j.jtbi.2013.01.012

[10] K. C. Chou and H. B. Shen, "Predicting Eukaryotic Protein Subcellular Location by Fusing Optimized Evidence-Theoretic K-Nearest Neighbor Classifiers," *Journal of Proteome Research*, Vol. 5, 2006, pp. 1888-1897. http://dx.doi.org/10.1021/pr060167c

[11] S. Wan, M. W. Mak and S. Y. Kung, "Adaptive Thresholding for Multi-Label SVM Classification with Application to Protein Subcellular Localization Prediction," 2013 *IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP*'13), 2013, pp. 3547-3551.

[12] S. Mei, "Multi-Label Multi-Kernel Transfer Learning for Human Protein Subcellular Localization," *PLoS ONE*, Vol. 7, No. 6, 2012, Article ID: e37716. http://dx.doi.org/10.1371/journal.pone.0037716

[13] S. Wan, M. W. Mak and S. Y. Kung, "GOASVM: Protein Subcellular Localization Prediction Based on Gene Ontology Annotation and SVM," 2012 *IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP*'12), 2012, pp. 2229-2232. http://dx.doi.org/10.1109/ICASSP.2012.6288356

[14] K. C. Chou and H. B. Shen, "Plant-mPLoc: A Top-Down Strategy to Augment the Power for Predicting Plant Protein Subcellular Localization," *PLoS ONE*, Vol. 5, 2010, Article ID: e11335. http://dx.doi.org/10.1371/journal.pone.0011335

[15] H. B. Shen and K. C. Chou, "Virus-mPLoc: A Fusion Classifier for Viral Protein Subcellular Location Prediction by Incorporating Multiple Sites," *Journal of Biomolecular Structure & Dynamics*, Vol. 26, 2010, pp. 175-186. http://dx.doi.org/10.1080/07391102.2010.10507351

[16] Z. C. Wu, X. Xiao and K. C. Chou, "iLoc-Plant: A Multi-Label Classifier for Predicting the Subcellular Localization of Plant Proteins with Both Single and Multiple Sites," *Molecular BioSystems*, Vol. 7, 2011, pp. 3287-3297. http://dx.doi.org/10.1039/c1mb05232b

[17] X. Xiao, Z. C. Wu and K. C. Chou, "iLoc-Virus: A Multi-Label Learning Classifier for Identifying the Subcellular Localization of Virus Proteins with Both Single and Multiple Sites," *Journal of Theoretical Biology*, Vol. 284, 2011, pp. 42-51. http://dx.doi.org/10.1016/j.jtbi.2011.06.005

[18] M. Zhu, L. Gao, Z. Guo, Y. Li, D. Wang, J. Wang and C. Wang, "Globally Predicting Protein Functions Based on Co-Expressed Protein-Protein Interaction Networks and Ontology Taxonomy Similarities," *Gene*, Vol. 391, No. 1-2, 2007, pp. 113-119. http://dx.doi.org/10.1016/j.gene.2006.12.008

[19] Z. Lei and Y. Dai, "Assessing Protein Similarity with Gene Ontology and Its Use in Subnuclear Localization Prediction," *BMC Bioinformatics*, Vol. 7, 2006, p. 491. http://dx.doi.org/10.1186/1471-2105-7-491

[20] X. Wu, L. Zhu, J. Guo, D. Y. Zhang and K. Lin, "Prediction of Yeast Protein-Protein Interaction Network: Insights from the Gene Ontology and Annotations," *Nucleic Acids Research*, Vol. 34, No. 7, 2006, pp. 2137-3150. http://dx.doi.org/10.1093/nar/gkl219

[21] D. Yang, Y. Li, H. Xiao, Q. Liu, M. Zhang, J. Zhu, W. Ma, C. Yao, J. Wang, D. Wang, Z. Guo and B. Yang, "Gaining Confidence in Biological Interpretation of the Microarray Data: The Functional Consistence of the Significant GO Categories," *Bioinformatics*, Vol. 24, No. 2, 2008, pp. 265-271. http://dx.doi.org/10.1093/bioinformatics/btm558

[22] P. Resnik, "Semantic Similarity in a Taxonomy: An Information-Based Measure and Its Application to Problems of Ambiguity in Natural Language," *Journal of Artificial Intelligence Research*, Vol. 11, 1999, pp. 95-130.

[23] P. W. Lord, R. D. Stevens, A. Brass and C. A. Goble, "Investigating Semantic Similarity Measures across the Gene Ontology: The Relationship between Sequence and Annotation," *Bioinformatics*, Vol. 19, No. 10, 2003, pp. 1275-1283. http://dx.doi.org/10.1093/bioinformatics/btg153

[24] D. Lin, "An Information-Theoretic Definition of Similarity," *Proceedings of the* 15*th International Conference on Machine Learning*, 1998, pp. 296-304.

[25] C. Pesquita, D. Faria, A. O. Falcao, P. Lord and F. M. Counto, "Semantic Similarity in Biomedical Ontologies," *PLoS Computational Biology*, Vol. 5, No. 7, 2009, Article ID: e1000443. http://dx.doi.org/10.1371/journal.pcbi.1000443

[26] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. 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*, Vol. 25, 1997, pp. 3389-3402. http://dx.doi.org/10.1093/nar/25.17.3389

[27] S. Wan, M. W. Mak and S. Y. Kung, "mGOASVM: Multi-Label Protein Subcellular Localization Based on Gene Ontology and Support Vector Machines," *BMC Bioinformatics*, Vol. 13, 2012, p. 290.

[28] K. C. Chou and H. B. Shen, "Recent Progress in Protein Subcellular Location Prediction," *Analytical Biochemistry*, Vol. 1, No. 370, 2007, pp. 1-16. http://dx.doi.org/10.1016/j.ab.2007.07.006

