Fifth International Workshop on Computational Intelligence & Applications
IEEE SMC Hiroshima Chapter, Hiroshima University, Japan, November 10, 11 & 12, 2009

A Semantic Similarity Measurement Method Based on Information Quality in the Structure of the Gene Ontology

Junya Hirai, Hideki Katagiri, Ichiro Nishizaki and Tomohiro Hayashida
Graduate School of Engineering, Hiroshima University
Kagamiyama 1-4-1, Higashi-Hiroshima City, Hiroshima, 739-8527 Japan
email:{hirai-junya,katagiri-h,nisizaki,hayashida}@hiroshima-u.ac.jp

***Abstract*—Gene ontology (GO) which described a biological concept of gene has attracted attention as an index for measuring semantic similarity of gene. This paper considers a new method for measuring the semantic similarity of GO through an extension and combination of two existing methods by Resnik and Wang et al. in order to improve their drawbacks of effects of shallow annotation. It is shown that the proposed method is superior to existing methods through experiments with pathway data.**

***Index Terms*—gene ontology，semantic similarity，pathway**

I. Introduction

Although controlled biochemical or biological vocabularies (for example, Gene Ontology (GO) [1]) is used for consistent descriptions of genes in different data sources, automatically measuring the functional similarities of genes based on these annotation data remains a challenge. Currently, researchers use online information retrieval tools (for example, AmiGO [2] and QuickGO [3]), to collect gene annotation data from various databases and manually discover the correlations or similarities of gene products by visually examining their biological functions. However, because the manual discovery of this important knowledge requires significant time and effort, there is a critical need to build automated tools to measure and visualize the functional similarities of gene products based on existing annotation information from heterogeneous data sources.

In past years, some online tools such as eGOn [4], FuSSiMeG [5], and DAVID [6] were developed to measure the functional similarity of genes. However, their similarity measurement methods have drawbacks. Some approaches [6] measure gene functional similarities based on the probability of the appearance of GO terms or the kappa statistics of similar annotation terms correlated with different genes, and ignore the semantic relations (‘is-a’ and ‘part-of’) among these terms in the GO graph. Although other methods (Jiang and Conrath [9], Lin [8], Resnik [7]) were proposed to measure the semantic similarity of terms in a specific taxonomy, these methods were originally developed for the natural language taxonomies and it is unclear whether they are suitable for measuring the semantic similarity of GO terms.

These existing methods (Jiang and Conrath, Lin, Resnik) and their variants (Couto [**?**], Kriventseva [**?**], Lee [**?**]) determine the similarity of two GO terms based on their distances to the closest common ancestor term and/or the annotation statistics of their common ancestor terms. Although recent studies (Guo [14], Sevilla [12], Wang [10]) evaluating these methods showed that Resnik’s method is better than other methods in terms of the correlation with gene sequence similarities and gene expression profiles, none of these evaluation studies provided direct evidences on how well these methods measure the functional similarity of genes. Instead, they pointed out some drawbacks in these existing similarity measurement methods that hinder their ability of determining the functional similarity of genes.

A drawback of Resnik’s method is that it ignores the information quantity in the structure of the ontology by only concentrating on the information quantity of a term derived from the corpus statistics. However, the specificity of a GO term is usually determined by its location in the GO graph and a GO term’s semantics (biological meanings) are inherited from all its ancestor terms. Therefore, using the information quantity as the sole determination factor for the semantic similarity of GO terms is inappropriate. On the other hand, based on human perspectives, if two terms sharing the same parent are near the root of the ontology (terms are more general), they should have larger semantic difference than two terms having the same parent and being far away from the root of the ontology because the later are more specific terms. However, using Jiang’s or Lin’s method, as pointed out by Sevilla et al., if two gene products are well annotated near the root of the ontology (shallow annotation), their semantic similarity will always be measured at very high (close to 1) and their semantic distance will always be computed close to nil, thus providing a misleading result. The effect of shallow annotation is a serious drawback of both Jiang and Lin’s methods.

First, the distances to the closest common ancestor term cannot accurately represent the semantic difference of two GO terms. As discussed previously, if two terms sharing the same parent are near the root of the ontology, they should have larger semantic difference than two terms having the same parent and being far away from the root of the ontology. In addition, one GO term may have multiple parent terms with

different semantic relations. A GO term's semantics (biological meanings) must be the aggregate semantic contributions from all ancestor terms (including this specific term). Second, measuring the semantic similarity of two GO terms based only on the number of common ancestor terms cannot discern the semantic contributions of the ancestor terms to these two specific terms. In fact, a common ancestor of two GO terms may have different contributions to the semantics of these specific terms because their distances to this common ancestor in the GO graph may differ and the semantic relations (edges in the GO graph) leading to this common ancestor may vary as well. Based on human perspectives, an ancestor term farther from a descendant term in the GO graph contributes less to the semantics of the descendant term while an ancestor term closer to a descendant term in the GO graph contributes more to the semantics of this descendant term. Unfortunately, most existing ontology-structure-based methods (Langaas [4], Wang [10]) also have their drawbacks in that they determine the semantic similarity of two GO terms either based on their distances to the closest common ancestor term or based on the number of their common ancestor terms.

Semantics (biological meaning) of GO terms must include a biological meaning from the ancestor term. Wang $et\ al.$ [11] proposed a method considering the position of these ancestor term related to two specific terms as well as the number of common ancestor term in a GO graph to judge semantic similarity of GO terms. However, as well as Lin's method, their method has a serious drawback of the effect of shallow annotation.

The purpose of this study is to propose a new semantic similarity measurement method based on information quantity in the structure of the gene ontology. In order to provide direct evidences on how well the proposed methods measure the functional similarity of genes, some experimental results for pathway data are shown.

II. CONVENTIONAL STUDY

Gene ontology expresses function information of gene, and the annotation is called GO term. A set of GO term composes a hierarchical structure in which there exist three categories: biological process (BP), cell ingredient (CC) and molecule function (MF).

A hierarchy figure of GO term (A part of the stratosphere)

GO : 0008150 Biological process
  GO:0009987 : Cellular process
    GO:0030154 : Cell differentiation
                 ・
                 ・
GO : 0005575 Cellular component
  GO:0005623 : Cell
    GO:0005622 : Intracellular
      GO:0043229 : Intracellular organelle
        GO:0043231:Intracellular membrance-bound organelle
  GO:0043226 : Organelle
    GO:0043229 : Intracellular organelle
      GO:0043231 : Intracellular membrance-bound organelle
  GO:0043227 : Membrance-bound organelle
    GO:0043231 : Intracellular membrance-bound organelle
                 ・
                 ・

A set of GO terms is presented as a directed acyclic graph (DAG) where there exist two kinds of semantic relations 'is-a' and 'part-of' as shown in Fig. 1. DAG is the directed graph that does not have a cycle. The 'is-a' relation is a simple class-subclass relation, where A is-a B means that A is a subclass of B and expressed by a solid line. The 'part-of' relation is a partial ownership relation; C part-of D means that whenever C is present, it is always a part of D, but C need not always be present and expressed in a dashed line.

Gene ontology terms describing function information attracts attention as an index for measuring semantic similarity between the gene. There are some semantic similarity measurement methods considering hierarchical structure [11] and information quantity [7].

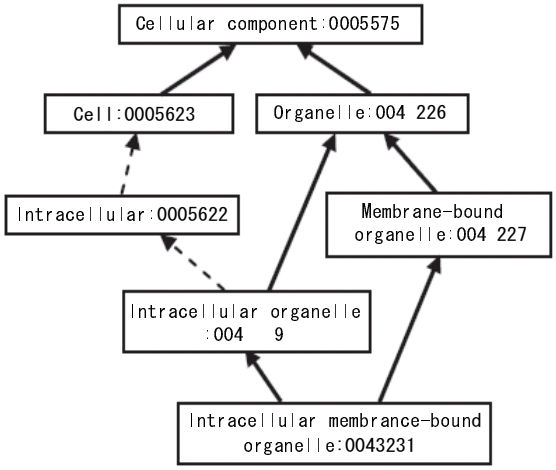

Fig. 1. (A)Intracellular Membrane-bound Organelle:0043231

A. GO term similarity measure

1) Resnik's GO term similarity measure:

Resnik's method which concentrates on similarity concepts for GO terms derived from information theory relies on the notion of the so-called minimum subsumer of two GO terms t and t', which is the lowest common ancestor in the GO hierarchy. Its information quantity IC_{ms}, which can be understood as a measure of similarity between t and t', is given by:

$$sim_R(t,t') = IC_{ms}(t,t') := max_{t \in Pa(t,t')} IC(t) \qquad (1)$$

where $Pa(t,t')$ denotes the set of all common ancestors of GO terms t and t', and $IC(t)$ denotes the information quantity of term t defined as the negative logarithm of the probability of

observing t (c.f. [7]) as follows:

$$IC(t) = -logP_R(t) \tag{2}$$

where $P_R(t)$ is a probability of observing each GO term given by

$$P_R(t) = \frac{freq(t)}{N} \tag{3}$$

where $freq(t)$ is the number of times that term t itself and its offspring are observed in a DAG of each category (BP, CC and MF), and N is the number of times that all of the GO terms of each categories are observed.

As extensions of Resnik's GO term similarity measurement method, there exist Lin's method [8], and Jiang-Conrath's method [9]. Both methods normalize the value obtained by Eq. (1), and only the difference between the two methods is just the way of normalizing: Lin's similarity measure is defined as

$$sim_L(t,t') = \frac{2IC_{ms}(t,t')}{IC(t) + IC(t')} \tag{4}$$

and, Jiang and Conrath's similarity measure is defined as

$$sim_{JC}(t,t') = 1 - min(1, IC(t) - 2IC_{ms}(t,t') + IC(t')) \tag{5}$$

As can be seen in (4) and (5), if the difference between information quantity of some GO term and that of the minimum subsumer is small, the degree of similarity measure is high.

2) Wang's GO term similarity measure:

In Wang's method, S-Value, the index of the relation between GO A and its ancestor of t, is introduced as follows:

$$S_A(t) = \begin{cases} 1 & if\ t = A \\ max\Big\{w_e * S_A(t') | t' \in children\ of(t)\Big\} & otherwise \end{cases}$$

where w_e is a weight of edge e to the child of GO term, and Wang $et\ al.$ set 'is-a'=0.8 and 'part-of'=0.6. Because 'is-a' relation is a class-subclass and 'part-of' relation is a partial ownership relation, the weight of 'part-of' should be smaller than that of 'is-a'. The relationship between GO A and its ancestors T_A is described by using SV defined as

$$SV(A) = \Sigma_{t \in T_A} S_A(t) \tag{6}$$

Then, the semantic similarity measure of GO A and B is defined as

$$S_{GO}(A,B) = \frac{\Sigma_{t \in T_A \cap T_B}(S_A(t) + S_B(t))}{SV(A) + SV(B)} \tag{7}$$

Eq. (7) means that if the ancestors of two GO terms are similar, the value of S_{GO} is high. However, one of drawbacks of this method is to fix the weights of the edges like $w_e = 0.8$ if the edge e expresses relation 'is-a', and $w_e = 0.6$ if the edge e expresses the relation 'part-of'. In this case, when a DAG is not so deep, the effects of GO terms located in the upper part is too strong, and as a result, high degrees of semantic similarity are easy to appear. On the other hand, if these weights are set to smaller values, the differences of degrees of similarity between any genes become small because in general the degrees of similarity are lower overall.

B. Gene similarity measure

Next, we introduce existing gene similarity measurement methods based on the similarity of GO terms. There are maximum method, average method and optimal assignment method which are used in GOSim [13].

1) Maximum pairwise GO term similarity:

The idea of the maximum pairwise GO term similarity is straight forward. Given two genes g and g' annotated with GO terms $t_1, \ldots, t_n$ and $t'_1, \ldots, t'_m$, we define the functional similarity between g and g' as

$$\begin{aligned} sim^{max}_{gene}(g,g') &= max\ sim(t_i, t'_j), \\ & \quad i = 1, \ldots, n,\ \ j = 1, \ldots, m \end{aligned}$$

where sim is some similarity measure (e.g. Resnik method and Wang method) to compare GO terms t_i and t'_j. In general, the value of sim_{gene} is normalized because if genes have a large number of GO terms, the similarity between such genes has a tendency to become large.

$$sim^{max}_{gene}(g,g') \longleftarrow \frac{sim_{gene}(g,g')}{\sqrt{sim_{gene}(g,g) sim_{gene}(g',g')}} \tag{8}$$

2) Average pairwise GO term similarity:

In the average method, the average of the semantic similarity between GO terms of two genes is calculate as follows:

$$sim^{avg}_{gene}(g,g') = \sum_{i=1}^{n}\sum_{j=1}^{m} sim(t_i, t'_j)/(n+m) \tag{9}$$

3) Optimal assignment gene similarities:

The idea of an optimal assignment is to assign each term of the gene having fewer GO terms to exactly one term of the other gene such that the overall similarity is maximized (c.f. Fig. 2). More formally, this can be stated as follows: Let π be some permutation of either an $n\text{-}subset$ of natural numbers $\{1, \ldots, m\}$ or an $m\text{-}subset$ of natural numbers $\{1, \ldots, n\}$. Let $subset$ be a smaller number of gene ontology terms of $m\text{-}subset$ and $n\text{-}subset$. Then the similarity is defined as

$$sim^{opt}_{gene}(g,g') = \begin{cases} max_\pi \sum_{i=1}^{n} sim(t_i, t'_{\pi(j)}) & if\ m > n \\ max_\pi \sum_{j=1}^{m} sim(t_{\pi(i)}, t'_j) & otherwise \end{cases}$$

As well as maximum method, the normalization like Eq. (8) is necessary to prevent that larger lists of terms automatically achieve a higher similarity.

III. New semantic similarity measurement method

On the assumption that assume a directed non-patrol graph T_A includes gene ontology A and its ancestors, we calculate

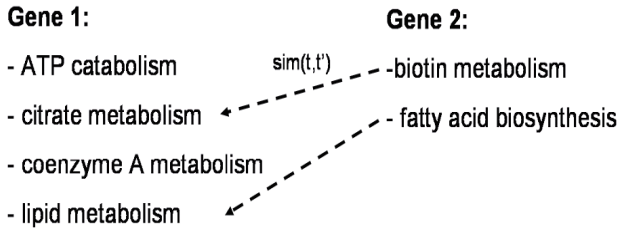

Fig. 2. Optimal assignment gene similarities

probability of observing $P_A(t)$ of gene ontology belonging to T_A by the next expression like expression (3) of Resnik.

$$P(t) = \frac{freq(t)}{N}$$

where N is the total number of gene ontology t belonging to DAG T_A.

We calculate information quantity $IC_A(t)$ of gene ontology t belonging to T_A based on probability of observing $P_A(t)$ by the next expression like expression (2) of Resnik.

$$IC_A(t) = -logP_A(t)$$

Next, the relationship between a gene ontology A and its ancestors t in T_A is defined by

$$R_A(t) = \frac{IC_A(t)}{IC_A(A)} \quad (10)$$

It should be noted there that Eq. (10) generally assigns larger weights to the edges located at lower parts than those at upper parts because this calculation regards GO having low expression frequency as ones including more information quantity.

In a manner similar to Eq. (6), the relationship $SI(A)$ between gene ontology A and its ancestors T_A is defined as

$$SI(A) = \sum_{t \in T_A} R_A(t) \quad (11)$$

A new semantic similarity measure of gene ontology A and B is defined as

$$sim_{GO}(A,B) = \frac{\Sigma_{t \in T_A \cap T_B}(R_A(t) + R_B(t))}{SI(A) + SI(B)}$$

As for the semantic similarity of gene, we use the optimal assignment method described in the previous section because it considers the location of GO as well as the structure of GO.

IV. Experiment

There are Guo *et al.*[14], Wang *et al.*[10], Sevilla and Segula[12] as evaluation methods of the semantic similarity of GO terms: Guo have shown that information quantity-based methods are better than other methods through the ROC curve for the data extracted from KEGG database [15]. Wang have shown that there is high correlation between the calculated similarity of genes and their expression quantities by using experimental data [16]. Sevilla and Segula have shown that there is higher correlation between the semantic similarity calculated by Resnik and expression quantity in comparison with other methods by Lin and Jiang-Conrath. However, none of these methods offers direct evidence on how well these methods measure semantic similarity of gene. In general, the biological functions of genes on the same pathway are more similar than those on the different pathways. This means that if a semantic similarity measure of genes is appropriate, the similarities of genes in the same pathway should be higher than those in the different pathway. From this viewpoint, we check whether the proposed method gives higher similarity to genes in the same pathway than to those in the different pathway, and compare the results with those of other existing methods.

A. Experimental

As experimental data, we use four pathway maps: (a)Purine metabolism, (b)Valine·leucine and isoleucine degradation, (c)Tyrosine metabolism and (d)Pyruvate metabolism, which can be extracted from KEGG database [15].

B. Experimental procedure

We utilize the software R [17] to analyze the data. The experimental procedure is as follows:

1) GO terms are extracted from UniProt [18] by using Hs.numbers of about 500 gene data .
2) The extracted GO terms are classified into three categories (BP, CC, MF).
3) The semantic similarities between genes on the same or different pathway map are calculated by the proposed method and existing methods.

C. Experimental results

Tables I, II and III show the experimental results by the proposed method, Lin's method and Wang's method, respectively. Diagonal values in each table represent the average degrees of gene similarity in the same pathway map, while other values represent those between different pathway maps. It is observed from Table I that the average degrees of similarity between genes in the same pathway map are larger than those between the different pathway maps. On the other hand, in Tables II and III, some diagonal values, which are average values between the different pathway map, are larger than others. In order to make the fact clearer, Figs. 3 and 4 show the difference and ratio of average degrees of similarity in the same and different pathway maps, respectively. From these figures, all the values obtained by the proposed method are larger than those of other methods. Considering that the biological functions of genes are generally similar if the genes are in the same pathway map, the proposed method provides a more proper semantic similarity measure of genes than others.

V. Conclusion

In this paper, we have proposed a new genetic similarity measurement method in order to appropriately measure the functional similarity of genes, which improves some drawbacks involved in some existing methods. We have shown

TABLE I
PROPOSED METHOD

pathway map	(a)	(b)	(c)	(d)
(a)	0.329	0.188	0.208	0.223
(b)		0.355	0.256	0.272
(c)			0.483	0.318
(d)				0.425

TABLE II
RESNIK'S METHOD

pathway map	(a)	(b)	(c)	(d)
(a)	0.225	0.154	0.135	0.154
(b)		0.253	0.185	0.332
(c)			0.241	0.195
(d)				0.242

TABLE III
WANG'S METHOD

pathway map	(a)	(b)	(c)	(d)
(a)	0.622	0.515	0.550	0.578
(b)		0.633	0.857	0.579
(c)			0.715	0.637
(d)				0.672

that the proposed method is better than some existing methods through the comparative experiments using pathway map data. Since the experiments are executed for only four pathway maps, more experiments should be done in order to show the validity of the proposed method. In the future, we will incorporate the relations 'is-a' and 'part-of' into the proposed method and construct a more useful genetic similarity measurement method.

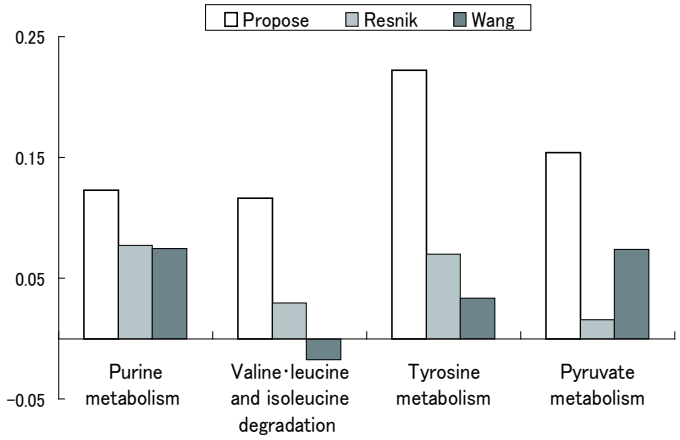

Fig. 3. Difference with same pathway map and different pathway map

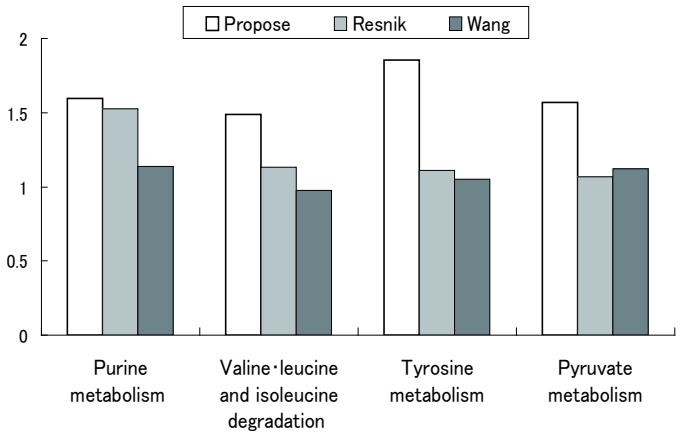

Fig. 4. The same pathway map divided average of each different pathway

REFERENCES

[1] Gene Ontology Home [http://www.geneontology.org]
[2] AmiGO [http://www.godatabase.org]
[3] QuickGO [http://www.ebi.ac.uk/QuickGO/]
[4] M. Langaas, C. Gunther and S. Lydersen, "Statistical hypothesis testing of associa-tion between two reporter lists within the GO-hierarchy", *Technical report*, Department of Mathematical Sciences, Norwegian University of Science and Technology, 2005.
[5] FuSSiMeG [http://xldb.di.fc.ul.pt/rebil/ssm/]
[6] DAVID [http://david.abcc.ncifcrf.gov/]
[7] P. Resnik, "Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language", *J. Artificial Intelligence Res.*, vol-11, pp.95-130, 1999.
[8] D. Lin, "An information-theoretic definition of similarity", *Proceedings of the 15th International Conference on Machine Learning*, San Francisco, vol-1, pp.296-304, 1998.
[9] J. Jiang and D. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy", *Proceedings of the International Conference on Research in Computational Linguistics*, Taiwan, 1998.
[10] H. Wang, F. Azuaje, "Gene Expression Correlation and Gene Ontology-Based Similarity: An Assessment of Quantitative Relationships", *In Proceedings of the 2004 IEEE Symposium on Computational In-telligence in Bioinformatics and Computational Biology*, pp.25-31, 2004.
[11] J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu and C. Chen, "A new method to measure the semantic similarity of GO terms", *BIOIMFOMATICS*, vol.23, pp.1274-1281, 2007.
[12] J. L. Sevilla, V. Segura, et al., "Correlation between Gene Expression and GO Semantic Similarity", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol-2, pp.330-338, 2005.
[13] H. Fröhlich, N. Speer, A. Poustka and T. BeiBbarth, "GOSim - an R-package for computation of information theoretic GO similarities between terms and gene products", *BMC Bioinformatics*, 2007.
[14] X. Guo, R. Liu, C. D. Shriver, H. Hu and M. N. LiebmanGuo, "Assessing semantic similarity measures for the characterization of human regulatory pathways", *Bioinformatics*, vol-22, pp.967-973, 2006.
[15] KEGG [http://www.genome.jp/kegg/]
[16] M. Eisen, P. L. Spellman, P. O. Brown, and D. Botsein," Cluster analysis and display of genome-wide expression patterns, " *Proc. Natl. Acad. Sci. USA*, vol. 95, pp. 14863-14868, 1998.
[17] C. Higuchi, K. Ishii, "Bioinformatics Data Analysis Using R and Bioconductor", *Kioritz Publication*, 2007.
[18] UniProt [http://www.uniprot.org/]