Implementation of a Functional Semantic Similarity Measure between Gene-Products

Francisco M. Couto
Mário J. Silva
Pedro Coutinho

DI–FCUL TR–03–29

2003

Departamento de Informática
Faculdade de Ciências da Universidade de Lisboa
Campo Grande, 1749–016 Lisboa
Portugal

Technical reports are available at `http://www.di.fc.ul.pt/tech-reports`. The files are stored in PDF, with the report number as filename. Alternatively, reports are available by post from the above address.

Implementation of a Functional Semantic Similarity Measure between Gene-Products

ABSTRACT

Motivation: Sequence similarity is a common technique to compare gene-products. However, many applications need to compare gene-products based on what they do, not how they are. Most gene-products are being annotated with terms describing their biological function. These terms are defined in biological ontologies. This makes it possible to implement similarity measures between gene-products based on their behavior.

Results: We define FuSSiMeG, a functional similarity measure between gene-products that compares the semantic similarity between the terms in their annotations.

Availability: Software available from `http://xldb.fc.ul.pt/rebil/ssm/`.

Contact: `fcouto@di.fc.ul.pt`

1 Introduction

Given the increasing importance of ontologies in biological settings, mechanisms enabling users to measure the similarity between the concepts represented by the ontologies or between the objects linked to these concepts are required. In computational linguistics, recent research on this topic has emphasized the use of semantic similarity measures. These measures compute distances between terms structured in a hierarchical taxonomy. Two kinds of approaches are prevalent: information content (node based) and conceptual distance (edge based). Information content considers the similarity between two terms the amount of information they share, where a term contains less information when it occurs very often. Conceptual distance is a more intuitive approach. It identifies the shortest topological distance between two terms in the scheme taxonomy. Budanitsky et al. experimentally compared five different proposed semantic similarity measures in WordNet (Budanitsky and Hirst, 2001). The comparison shows that Jiang and Conrath's semantic similarity measure provides the best results overall (Jiang and Conrath, 1997). This semantic similarity measure is a hybrid approach, i.e. it combines information content and conceptual distance with some parameters that control the degree of each factor's contribution. The conceptual distance is based on the node depth and density factors. The node depth factor relies on the argument that similarity increases as we descend the hierarchy, since the relations are based on increasingly finer details. The density factor relies on the argument that when the parent node has several child nodes (high density) they tend to be more similar.

More recently, Lord et al. investigated an information content semantic similarity measure, and its application to annotations found in SwissProt (P.W.Lord *et al.*, 2003). These annotations associate gene-products with functional terms. The authors present results showing that semantic similarity is correlated with sequence similarity, i.e. function is correlated with structure. In our work, we implemented a hybrid semantic similarity measure, which integrates the information content with conceptual distance factors. Based on this measure we propose FuSSiMeG (Functional Semantic Similarity Measure between Gene-Products), which measures the functional similarity between gene-products.

2 FuSSiMeG

2.1 Semantic Similarity between GO terms

To compute the semantic similarity between functional properties, FuSSiMeG implemented Jiang and Conrath's measure in GO (Gene Ontology) (Consortium, 2001). GO provides a structured controlled vocabulary of gene and protein biological roles. The three organizing principles of GO are molecular function, biological process and cellular component. Rison et al. discuss the reasons for choosing GO as the functional scheme in a survey about functional classification schemes (Rison *et al.*, 2000). They describe GO as "representative of the 'next generation' of functional schemes". Unlike other schemes, GO is not a tree-like hierarchy, but a directed acyclic graph (DAG), which permits a more complete and realistic annotation.

Following the Jiang and Conrath's definition, the information content of a GO term t can be quantified as follows:

$$IC(t) = -log(P(t)), \quad (1)$$

where $P(t)$ is the probability of occurring a GO term t. GO provides an association table, which links gene-products to GO terms. We compute $P(t)$ has the number of occurrences of t divided by the total number of occurrences in that table. However, as GO is a hierarchical structure, $P(t)$ has to increase as we ascend the hierarchy. This means that $P(t)$ is larger when t is closer to the root node. Therefore, when a GO term t_1 subsumes a GO term t_2, and t_2 occurs then we

consider that t_1 also occurs. Therefore, $P(t_1)$ will always be greater or equal than $P(t_2)$, i.e. $IC(t_1) \leq IC(t_2)$.

The semantic distance between t_1 and t_2 when t_1 subsumes t_2, is quantified as follows:

$$\Delta(t_1,t_2) = IC(t_2) - IC(t_1). \quad (2)$$

Since, if t_2 occurs then t_1 also occurs, we have $\Delta(t_1,t_2) = 0$ when t_1 occurs, only because t_2 occurs. The assumption is that we have a larger similarity between terms whose occurrences have a stronger correspondence.

The semantic distance between two terms t_2 and t_3, without a subsuming relation, is the sum of their semantic distance to their closest shared ancestor. The closest shared ancestor is the GO term that subsumes both terms and belongs to the shortest path between the two terms. Thus, the semantic distance between the GO terms t_2 and t_3 is quantified as follows:

$$\Delta(t_2,t_3) = \Delta(t_1,t_2) + \Delta(t_1,t_3), \quad (3)$$

where t_1 is the closest shared ancestor of t_2 and t_3.

The distance defined in equations 2 and 3 does not use any conceptual distance factors. Thus, we have to redefine this distance to integrate the node depth and density factors. Considering a GO term t_0 that subsumes a GO term t_n, and the sequence of GO terms $t_0,\ldots,t_n$ representing the path from t_0 to t_n with length n, the semantic distance between t_0 and t_n is redefined as follows:

$$\Delta(t_0,t_n) = \sum_{i=0}^{n-1} D(t_i) \times E(t_i) \times (IC(t_{(i+1)}) - IC(t_i)), \quad (4)$$

where $D(t)$ and $E(t)$ represent the depth and density conceptual distance factors for a GO term t.

$D(t)$ is defined as follows:

$$D(t) = \left(\frac{d(t)+1}{d(t)}\right)^{\alpha}, \quad (5)$$

where $d(t)$ denotes the depth of GO term t in the ontology. The α parameter controls the degree of how much the depth factor contributes in equation 4. When α approaches 0 this contribution becomes less significant, since $D(t)$ will approach 1.

$E(t)$ is defined as follows:

$$E(t) = (1-\beta) \times \frac{\overline{E}}{e(t)} + \beta \quad (6)$$

where $e(t)$ denotes the local density of the GO term t, i.e. the number of edges that start from t. $\overline{E}$ represents the average density in the whole ontology, i.e. the number of edges divided by the number of terms in the ontology. The β parameter controls the degree of how much the density factor contributes in equation 4. When β approaches 1 this contribution becomes less significant, since $E(t)$ will approach 1.

When $\alpha = 0$ and $\beta = 1$ the equations 4 and 2 are equivalent.

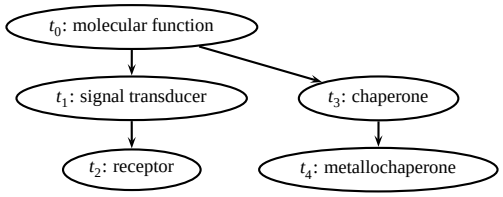

Figure 1: Subgraph of GO

Example 1 *Considering only the subgraph of GO represented in Figure 2.1, and an association table with the following occurrences:*

GO term	*Name*	*Occurrences*
t_0	*molecular function*	*0*
t_1	*signal transducer*	*2*
t_2	*receptor*	*1*
t_3	*chaperone*	*1*
t_4	*metallochaperone*	*2*

Propagating the occurrences through the hierarchy, we reach the following values:

GO term	*Occurrences*	*P(t)*	*IC(t)*
t_0	*6*	*0.4*	*1.32*
t_1	*3*	*0.2*	*2.32*
t_2	*1*	*0.07*	*3.91*
t_3	*3*	*0.2*	*2.32*
t_4	*2*	*0.13*	*2.91*

Since we considered only 5 GO terms with 4 edges, we have $\overline{E} = 4/5$. *Considering* $\alpha = \beta = 0.5$ *the depth and the density factors have the following values:*

GO term	*d(t)*	*D(t)*	*e(t)*	*E(t)*
t_0	*1*	*1.41*	*2*	*0.7*
t_1	*2*	*1.22*	*1*	*0.9*
t_3	*2*	*1.22*	*0*	*0.9*

Using 4 we can calculate $\Delta(t_0,t_2)$ *and* $\Delta(t_0,t_4)$ *as follows:*

$$\begin{aligned}\Delta(t_0,t_2) &= D(t_0) \times E(t_0) \times (IC(t_1) - IC(t_0)) \\ &\quad + D(t_1) \times E(t_1) \times (IC(t_2) - IC(t_1) \\ &= 1.41 \times 0.7 \times (2.32 - 1.32) \\ &\quad + 1.22 \times 0.9 \times (3.91 - 2.32) \\ &= 1.72\end{aligned}$$

$$\begin{aligned}\Delta(t_0,t_4) &= D(t_0) \times E(t_0) \times (IC(t_3) - IC(t_0)) \\ &\quad + D(t_3) \times E(t_3) \times (IC(t_4) - IC(t_4)) \\ &= 1.41 \times 0.7 \times (2.32 - 1.32) \\ &\quad + 1.22 \times 0.9 \times (2.91 - 2.32) \\ &= 1.35\end{aligned}$$

$\Delta(t_0,t_2) > \Delta(t_0,t_4)$ *because the difference between the number of* t_1 *and* t_2 *occurrences is larger than between* t_3 *and* t_4.

Finally, using 3, we can calculate $\Delta(t_2,t_4)$ *as follows:*

$$\begin{aligned}\Delta(t_2,t_4) &= \Delta(t_0,t_2) + \Delta(t_0,t_4)\\ &= 1.72 + 1.35\\ &= 3.07\end{aligned}$$

Since the distance between GO terms is based on the difference between their information content, to normalize this distance we have to divide it by the maximum information content minus the minimum information content. However, the minimum information content is 0, thus we can define the normalized distance as follows:

$$\Delta_n(t_1,t_2) = min\{1, \frac{Delta(t_1,t_2)}{IC(t_0)}\}, \quad (7)$$

where t_1 and t_2 are GO terms. And, t_0 is a term that only occurs once, i.e., $IC(t_0)$ represents the maximum information content possible. All the cases where $\Delta > IC(t_0)$ are considered to have the maximum distance possible 1, because the distance is to large that it is irrelevant to discriminate them.

The distance between GO terms, can be easily converted to a semantic similarity measure as follows:

$$SSM(t_1,t_2) = 1 - \Delta_n(t_1,t_2), \quad (8)$$

where t_1 and t_2 are GO terms. And, we have $0 \leq SSM \leq 1$, because $0 \leq \Delta_n \leq 1$.

2.2 Functional Similarity between Gene-Products

To measure the functional similarity between gene-products, FuSSiMeG uses the assignments provided by GOA (Gene Ontology Annotation) (Camon *et al.*, 2003). GOA is a project that identifies assignments of biomolecules to GO resource GOA provides a vast list of GO assignments for all complete and incomplete proteomes that exist in SwissProt and TrEMBL. Thus, given a gene-product from SwissProt/TrEMBL we define the list of GO terms assigned to the gene-product in GOA as follows:

$$T(g) = \{t : (g,t) \in GOA\}, \quad (9)$$

where *GOA* represents the set of pairs composed by gene-products and GO terms assigned in GOA database.

FuSSiMeG assumes that two gene-products have a functional similarity when they are annotated with similar functional terms. FuSSiMeG measures the similarity between gene-products by the maximum similarity between their assigned terms. FuSSiMeG is defined as follows:

$$\begin{aligned}&FuSSiMeG(g_1,g_2) =\\ &\quad max\{SSM(t_1,t_2) : t_1 \in T(g_1) \wedge t_2 \in T(g_2)\},\end{aligned} \quad (10)$$

where g_1 and g_2 are two gene-products. However, if g_1 and g_1 are both assigned to a frequent GO term, such as "protein", which does not have a large information content, they will have a large functional similarity. Thus, FuSSiMeG improves its assumption so that two gene-products have a functional similarity not only when they are annotated with similar functional terms but also when these terms have significant information content. Thus, the equation 10 is redefined as follows:

$$\begin{aligned}&FuSSiMeG(g_1,g_2) =\\ &\quad max\{SSM(t_1,t_2) \times IC(t_1) \times IC(t_2) : t_1 \in T(g_1)\\ &\qquad \wedge t_2 \in T(g_2)\}.\end{aligned} \quad (11)$$

Since SSM and IC ranges from 0 to 1, FuSSiMeG also ranges from 0 to 1.

3 Results

The results presented in this paper report to an analysis performed on the September 2003 release of GO and on the 11.0 release of GOA SPTr. From the 15944 GO terms available, only 9479 of them have occurrences in the association table. From the 3336167 GOA assignments, 3335363 of them are composed by GO terms having occurrences in the association table. All the 734323 gene-products present in GOA assignments are linked with GO terms having occurrences in the association table. Thus, FuSSiMeG is able to measure the functional similarity between all the gene-products present in GOA.

FuSSiMeG is available on the web (`http://xldb.fc.ul.pt/rebil/ssm/`), where is possible to select the α and β parameters to measure two given GO terms or gene-products. Since the implementation of FuSSiMeG takes insignificant time to measure the similarity of two gene-products, the web page computes FuSSiMeG on the fly.

FuSSiMeG is being applied in some interesting ways:

- FuSSiMeG has been applied to identify a correlation between modular structure and molecular function (Couto *et al.*, 2003a).
- CAC, a text mining method, demonstrates the application of FuSSiMeG to improve the extraction of annotations from biological literature (Couto *et al.*, 2003b).
- FuSSiMeG was adopted by BioCreAtIvE (Critical Assessment of Information Extraction systems in Biology), an international evaluation for the text data mining systems applied to biology (bio, 2003). The ranking of the systems that participate in this evaluation uses FuSSiMeG to measure how close in GO the predictions given by those systems are to the correct annotations.

4 Future Work

To demonstrate the reliability of FuSSiMeG, its results need to be validated. We aim to use FuSSiMeG to compare the functional similarity between gene-products with and without interactions already cataloged in a database, such as DIP (Database of Interacting Proteins). We assume that if two gene-products interact, they should be annotated with a similar biological process, cellular component or molecular function. Thus, if the functional similarity between interacting gene-products is significantly larger than between non-interacting gene-products then the measure is valid.

In the future, we will explore the effects of the three organizing principles of GO in FuSSiMeG. Currently, FuSSiMeG is based on the maximum similarity found in any principle. We think that FuSSiMeG can improve its results if the principles could contribute differently to the measure.

The semantic similarity measure can also consider the type of edges in GO. However, about 99% of the edges are of the 'is-a' type. Thus, FuSSiMeG did not implement this feature until now.

5 Conclusions

In this paper, we present FuSSiMeG a novel functional similarity measure between gene-products. FuSSiMeG provides a new tool to compare gene-products according to what they do and not how they are. Instead of sequence similarity, FuSSiMeG compares the biological activity between gene-products, which represents a surplus value for a vast range of biological applications, such as: organizing the gene-products in families, identifying the common properties of microarray clusters, finding interactions between gene-products, and validating the results of text data mining systems.

FuSSiMeG is publicly available on the web through a simple to use interface. Some projects already use FuSSiMeG to improve or to evaluate their results, which demonstrates its effectiveness.

REFEFENCES

(2003) BioCreAtIvE: Critical Assessment of Information Extraction systems in Biology. `http://www.pdg.cnb.uam.es/BioLINK/BioCreative.eval.html`.

Budanitsky, A. and Hirst, G. (2001) Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. Workshop on WordNet and Other Lexical Resources, in the North American Chapter of the Association for Computational Linguistics (NAACL-2000), Pittsburgh, PA.

Camon, E., *et al.* (2003) The Gene Ontology Annotation (GOA) Project: Application of GO in SWISS-PROT, TrEMBL and InterPro. *Comparative and Functional Genomics*, **4**, 71–74.

Consortium, T. G. O. (2001) Creating the gene ontology resource: design and implementation. *Genome Res*, **11**, 1425–1433.

Couto, F., Silva, M., and Coutinho, P. (2003a) Curating Extracted Information through the Correlation between Structure and Function. In *third meeting of the special interest group on Text Data Mining co-located with 11th Internacional Conference on Intelligent Systems for Molecular Biology (ISMB)*. Brisbane, Australia.

Couto, F., Silva, M., and Coutinho, P. (2003b) Improving Information Extraction through Biological Correlation. In *Data Mining and Text Mining for Bioinformatics Workshop co-located with 7th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*. Dubrovnik-Cavtat, Croatia.

Jiang, J. and Conrath, D. (1997) Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *Proceedings of the 10th International Conference on Research on Computational Linguistics (ROCLING X)*. Taiwan.

P.W.Lord, Stevens, R., Brass, A., and C.A.Goble (2003) Semantic similarity measures as tools for exploring the Gene Ontology. In *Pacific Symposium on Biocomputing*, pages 601–612.

Rison, S., Hodgman, T., and Thornton, J. (2000) Comparison of functional annotation schemes for genomes. *Funct. Integr. Genomics*, **1**, 56–69.