# Use of Gene Ontology semantic information in protein interaction data visualization

Raimon Massanet, Pere Caminal and Alexandre Perera

*Abstract*— **The Gene Ontology project is an effort to structure knowledge on biological products and processes by adding semantic information to them. This is done in a systematic way so that this additional information can be automatically processed. In this contribution a protein-protein interaction visualization algorithm is proposed, which combines protein interaction data with Gene Ontology semantic information. The information is integrated using a semantic distance measure defined in ontologies or taxonomies. Multidimensional scaling is applied to this measure and the output complements protein interaction data in building an interaction visualization map.**

## I. INTRODUCTION

The field of proteomics presents some additional complexity as compared to genetics. This higher complexity has two main components.

Firstly there is a great difference in dynamics. Whereas the genome remains relatively stable experiencing small changes at an evolutionary scale, the proteome presents a continuous activity, being able to undergo significant changes in a minute scale. Most external stimuli and metabolism responses entail changes in protein concentration, structure or unbalance of chemical reactions.

On the other hand, there is a difference of size. The number of genes in the human genome is estimated to be around $2 \cdot 10^4$, two magnitude orders smaller than the estimated size of the human proteome. Furthermore, many proteins are modified by other proteins or by environmental factors after being synthesized, giving rise to new proteins possibly not encoded in the genome. Also, proteins can bind to form a *complex* whose chemical properties can be very different from the properties of each constituent. If all these post-synthesis alterations are taken into account, the estimated size of the proteome rises up to an order of $10^7$ [3].

Investigations carried by biologists are usually directed to a single pathway involving a small number of proteins. A pathway is a map of interactions among proteins or other molecules that regulate a biological process. Focusing on a small part of the proteome scientists can study more precisely protein structure, function and relations among proteins. Efforts trying to trace protein interactions at a large scale generate necessarily massive, heterogeneous and continuously growing databases. Some of these databases are: Entrez, by the National Center for Biotechnology Information (NCBI); ENSEMBL and UniProt, by the European Bioinformatics Institute (EBI) or KEGG (Kyoto Encyclopedia of Genes and Genomes) [6], [8], [7].

The community is publishing efforts trying to structure genetic information in order to offer protein interaction data with an organized and fair presentation. Some examples are Biocarta `http://www.biocarta.com`, REACTOME `http://www.reactome.org`, the previously cited KEGG and others. These tools help in visualizing maps of proteins in a way that fits the needs of the users.

One of the most recent contributions from bioinformatics to the fields of genomics and proteomics are ontologies. An ontology is a set of terms related to a certain domain, and hierarchical relations among them. In the field of Artificial Intelligence the use of ontologies is very common for describing the Universe in which an intelligent agent will make decisions. Recently they have been incorporated in the biomedic field providing means to describe the properties of a biological product using a controlled vocabulary which is very close to natural language.

The Gene Ontology [1] is the most used ontology in biology. It contains around 55,000 manual annotations and more than 130,000 electronically inferred. An annotation is a relation between a biological product and an ontology term that describes some aspect of the product. This allows to define the product in a way that can be processed by computer applications.

The Gene Ontology is divided in three orthogonal ontologies that define three different types of semantic information about a biological product. Each of these ontologies is a directed acyclic graph.

The *Biological Process* (BP) ontology defines terms which yield information on the biological function of a product. These terms are closely related to the concept of pathway, described earlier. Terms in the *Cellular component* (CC) ontology describe the physical locations where a product can be found, usually relative to the cell. Finally the *Molecular Function* (MF) ontology contains terms that describe the biochemical properties of a biological product at a molecular level. Their structure allows for graph algorithms to be applied to the ontologies. This could shed light on the structural information about them and the biological products mapped to them.

A meaningful measure that can be obtained from the annotations in the Gene Ontology is the semantic distance between two biological products. This yields information on how similar or different they are, always from three different points of view: their biological function, their

The three authors are with ESAII department, Biomedical Engineering Research Center, Technical University of Catalonia, CIBER in Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN). Pau Gargallo, 5. 08028 Barcelona, Spain. {`Raimon.Massanet, Pere.Caminal, Alexandre.Perera`}`@upc.edu`

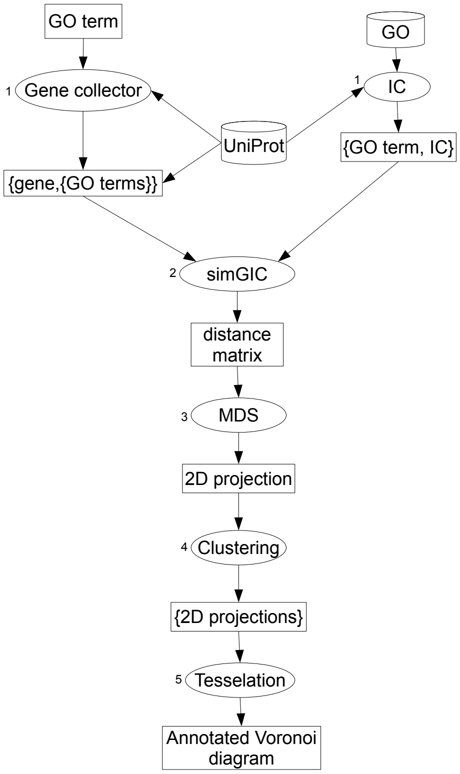

Fig. 1. A diagram of the proposed algorithm. Ellipses represent methods, rectangles represent data and cylinders represent external data sources.

location, and their molecular properties [14], [10], [5], [12]. Semantic distance is a useful measure for classifying biological products according to the semantic knowledge provided by field researchers. However, there is no application, to our knowledge, that uses any measure related to semantic information for data visualization purposes.

In this contribution an algorithm is proposed for organizing and visualizing interaction maps of proteins including semantic information extracted from the Gene Ontology annotations.

## II. Methods

The algorithm proposed shows graphical semantic information about all gene products involved in a biological process represented by a Gene Ontology (GO) term. The algorithm has been implemented using R [13]. The task is performed in five steps, as represented in figure 1.

1) Collection of genes that contain the input GO term as annotation. Collection of all semantic annotations in the ontology and calculation of the Information Content (IC) of every term, which will be needed for calculating semantic distance.
2) Calculation of semantic distance between all pairs of genes obtained before. Semantic measure is computed for each of the three ontologies, yielding three distance matrices.
3) Multidimensional scaling (MDS) of the distances calculated before. As a result, three planar projections of gene products are obtained. The euclidean distance between every pair of projected genes is the same as their semantic distance.
4) Clustering of genes in the MDS plane to obtain semantically similar groups of genes.
5) Division of the MDS plane in sectors according to the group division calculated before. Labeling of each sector with its most frequent GO terms.

In the first step, genes are collected using the *biomaRt* package [2]. All gene products that have the input GO term as annotations are stored. Then all semantic annotations are retrieved and stored for subsequent use. In order to compute the semantic distance between two terms of an ontology, define the ancestors of a term $t$, denoted by $anc(t)$, as all nodes that can be reached from $t$ following reversed links in the directed graph that represents the ontology. Given two terms $t_1$ and $t_2$ their common ancestors are:

$$comm(t_1, t_2) = \{anc(t_1) \cap anc(t_2)\} \tag{1}$$

And their most informative ancestor $mica(t_1, t_2)$ is:

$$mica(t_1, t_2) = \max_t \{IC(t) \mid t \in comm(t_1, t_2)\} \tag{2}$$

$IC(t)$ is the Information Content of a term $t$, as it was defined by Resnik [14]:

$$IC(t) = -\log p(t) \tag{3}$$

Where $p(t)$ is the probability of usage of the term $t$ in a given knowledge corpus, in this case UniProt. In order to compute $p(t)$, the quantity of proteins that have $t$ or any of its descendants is calculated. Then this quantity is divided by the total number of annotations in the corpus, $N$.

Finally the IC is normalized following [12]:

$$IC_u(t) = \frac{IC(t)}{\log_2 N} \tag{4}$$

The IC calculation is independent of the input GO term and could be computed and stored beforehand, so it can be used in future calculations. The output of this step is a set of gene products involved in a biological process and their semantic annotations. Three different sets of annotations are kept for each product, according to the ontology to which they belong.

In the second stage of the algorithm, the semantic distance is calculated for each pair of genes, obtaining three matrices of gene to gene distances — one from each ontology. For computing semantic distance between two gene products, define $ann(GP_A)$ as the set of terms annotated to a gene product $GP_A$. Then compute the $simGIC$ measure, as defined in [12]. To compute $simGIC$, the subgrafs generated by the annotations of each gene product are calculated. Then the ratio between the accumulated informative content in the intersection of the two subgrafs and its union is calculated (see [12] for more details):

$$simGIC(GP_A, GP_B) = \frac{\sum_{t \in \{ann(GP_A) \cap ann(GP_B)\}} IC_u(t)}{\sum_{t \in \{ann(GP_A) \cup ann(GP_B)\}} IC_u(t)} \quad (5)$$

This measure of similarity yields a normalized value within the range $[0, 1]$. To obtain a dissimilarity measure define:

$$distGIC(GP_A, GP_B) = 1 - simGIC(GP_A, GP_B) \quad (6)$$

In the third step, multidimensional scaling (MDS) is applied to the distance matrices in order to obtain three planar projections of the gene products. MDS is a technique to calculate low dimensionality projections from a distance matrix, so that the distance between each pair of projected points is the same as, or very close to, the original distance matrix. It is often used for data visualization purpose as it effectively reduces dimensionality preserving relative distances [4].

In the fourth step of the algorithm, hierarchical clustering is performed in the projected space to group semantically similar gene products [11]. This method works by first assigning each point to its own cluster and merging together the closest two groups at each iteration. It finishes when all points belong to the same cluster. The output of this method is a dendrogram showing which two groups have been merged at each step and what was the distance between them. Plotting these distances a decision can be made on the clusters that best describe the data. The output of this step is a partition of gene products in semantically similar sets.

Finally a Delaunay triangulation is calculated taking as input the centroids of each cluster. Then a Dirichlet tessellation is obtained [9]. A tessellation of an area is a set of planar figures, so that they do not overlap and together they cover the totality of the area. This step has been implemented using the *deldir* package [15]. The output of this step is a division of the planar projection in sectors.

Finally each sector is labeled with the three most frequent GO terms within.

## III. Results

In this section some graphics are shown, obtained as a result of applying the algorithm to the blood coagulation cascade (term "GO:0007596").

Figure 2 shows the grouping of gene products by semantic similarity as described in step 4 for the "molecular function" (MF) ontology. Figure 3 shows a planar tessellation of the projected gene products according to the "biological process" (BP) semantic distance. Most blood coagulation factors lie in the "E" zone, so close to each other that appear as a single spot in the graph. This zone contains gene products that are close to each other, but far away from products outside. This is because the genes within this zone have few GO annotations. However, blood coagulation factors FII (prothrombin) and FIII (Tissue Factor) lie in the "C" zone, which holds a considerable amount of gene products related to GO terms "inflammatory response", "immune response" and "signal transduction". This zone contains products with high annotation rates involved in metabolism response to wounds and other threats.

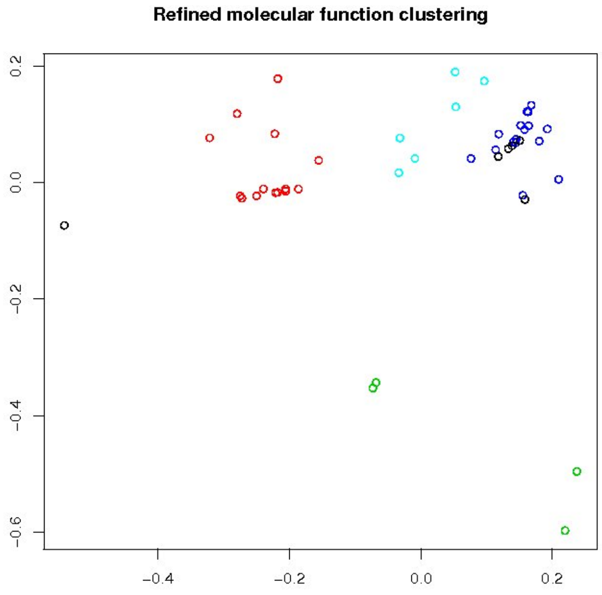

Fig. 2. Representation of gene product grouping by semantic similarity according to "molecular function" distance.

A more significant result can be seen in figure 4, which shows a distribution of gene products involved in the blood coagulation cascade according to their physical localization — "cellular component" (CC) semantic distance. Most blood coagulation factors lie within the "A" zone, which groups proteins that are usually found in regions outside the cell. However, FIII (Tissue Factor) is located in the "C" zone, which contains mostly intracellular gene products. Therefore, looking at the map it can be quickly concluded that most blood coagulation factors are found outside the cell while only Tissue Factor is usually found inside.

## IV. Conclusions

An algorithm has been developed to display protein data including a measure of semantic distance between pairs of proteins taken from the Gene Ontology. This algorithm is based on the calculation of three matrices of semantic distances corresponding to the three orthogonal ontologies contained in GO (biological process, cellular component and molecular function). Multidimensional scaling is used to calculate a planar position for each protein so that distances are conserved between pairs of proteins. This is done for each matrix distance, and three planar protein projections are obtained. Each map graphically shows semantic information relative to each ontology. The algorithm applied to the blood coagulation cascade yields a quick and simplified approximation to the existing knowledge on the biological process and localization of the proteins involved. This algorithm can be integrated with protein interaction visualization tools to achieve a higher organization of the interaction maps.

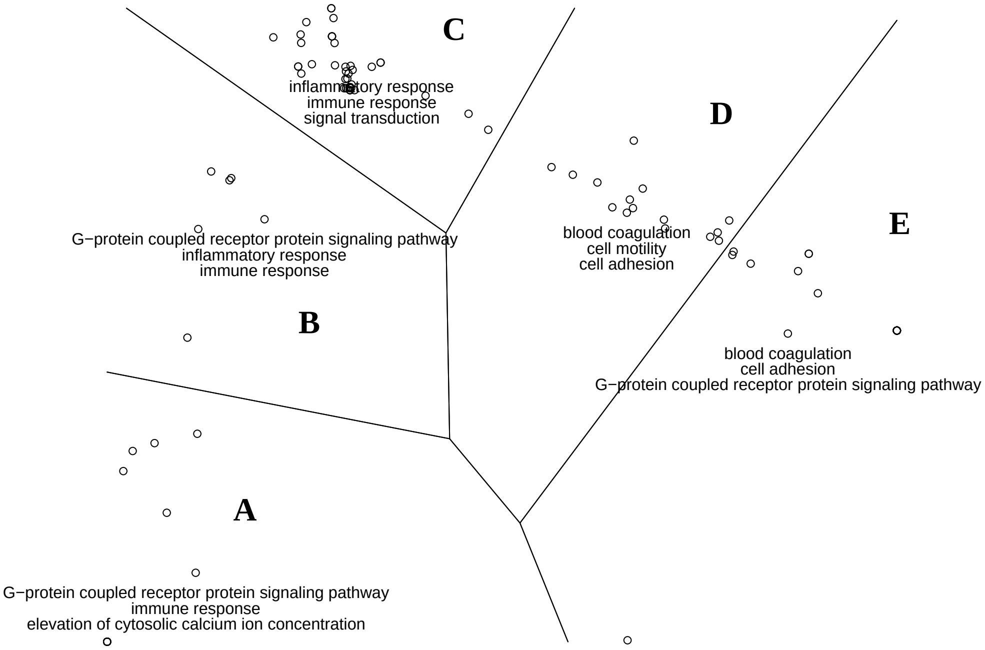

Fig. 3. Tessellation of the MDS plane according to the ”biological process” semantic distance.

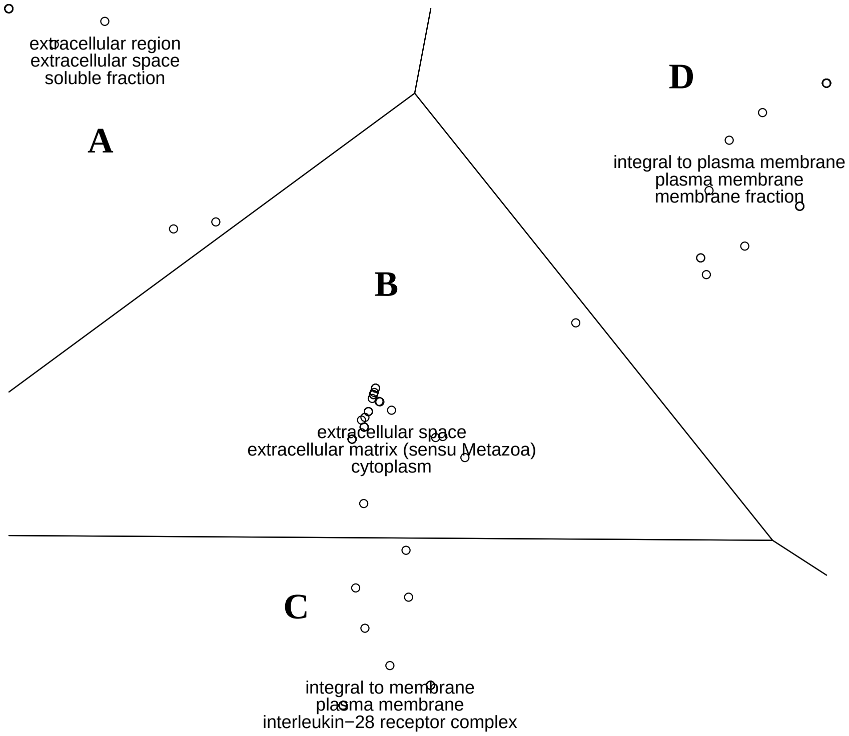

Fig. 4. Tessellation of the MDS space according to the ”cellular component” semantic distance.

## V. Acknowledgments

This work has been partly financed by CICYT TEC2007-63637/TCM and also by the *Ramón y Cajal* programme. CIBER in Bioengineering, Biomaterials and Nanomedicin is an initiative by ISCIII.

## References

[1] Gene Ontology Consortium. Creating the gene ontology resource: design and implementation. *Genome Res*, 11(8):1425–1433, Aug 2001.

[2] Steffen Durinck, Wolfgang Huber, and Sean Davis. *biomaRt: Interface to BioMart databases (e.g. Ensembl, Wormbase, Gramene and Uniprot)*. R package version 1.12.2.

[3] R. Gerzsten and T. Wang. The search for new cardiovascular biomarkers. *Nature*, 451:949–952, 2008.

[4] J. Gower. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53:325–338, 1966.

[5] Jay J. Jiang and David W. Conrath. *Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy*, 1997.

[6] M. Kanehisa and S. Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27–30, Jan 2000.

[7] Minoru Kanehisa, Michihiro Araki, Susumu Goto, Masahiro Hattori, Mika Hirakawa, Masumi Itoh, Toshiaki Katayama, Shuichi Kawashima, Shujiro Okuda, Toshiaki Tokimatsu, and Yoshihiro Yamanishi. Kegg for linking genomes to life and the environment. *Nucleic Acids Res*, 36(Database issue):D480–D484, Jan 2008.

[8] Minoru Kanehisa, Susumu Goto, Masahiro Hattori, Kiyoko F Aoki-Kinoshita, Masumi Itoh, Shuichi Kawashima, Toshiaki Katayama, Michihiro Araki, and Mika Hirakawa. From genomics to chemical genomics: new developments in kegg. *Nucleic Acids Res*, 34(Database issue):D354–D357, Jan 2006.

[9] D. T. Lee and B. J. Schacter. Two algorithms for constructing a delaunay triangulation. *Int. J. Computer and Information Sciences*, 9:219–242, 1980.

[10] D. Lin. *An information-theoretic definition of similarity*, 1998.

[11] F. Murtagh. *Multidimensional clustering algorithms*, 1985.

[12] Catia Pesquita, Daniel Faria, Hugo Bastos, Antonio Ferreira, Andre Falcao, and Francisco Couto. *Metrics for GO based protein semantic similarity: a systematic evaluation*, 2008.

[13] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[14] P. Resnik. Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *Artificial Intelligence Research*, 11:95–130, 1999.

[15] Rolf Turner. *deldir: Delaunay Triangulation and Dirichlet (Voronoi) Tessellation.*, 2007. R package version 0.0-7.