# Application of Gene Ontology to gene identification

*Hugo Bastos*, Bruno Tavares, Catia Pesquita, Daniel Faria, Francisco Couto*

Department of Informatics, Faculty of Sciences, University of Lisbon, Portugal

* Corresponding author: hbastos@xldb.di.fc.ul.pt

## Abstract

Candidate gene identification deals with associating genes to underlying biological phenomena, such as diseases and specific disorders. It has been shown that classes of diseases with similar phenotypes are caused by functionally related genes. Currently, a fair amount of knowledge about the functional characterization can be found across several public databases, however, functional descriptors can be ambiguous, domain specific and context dependent. In order to cope with these issues the Gene Ontology (GO) project developed a bio-ontology of broad scope and wide applicability. Thus, the structured and controlled vocabulary of terms provided by the GO project describing the biological roles of gene products can be very helpful in candidate gene identification approaches.

The method presented here uses GO annotation data in order to identify the most meaningful functional aspects occurring in a given set of related gene products. The method measures this meaningfulness by calculating an e-value based on the frequency of annotation of each GO term in the set of gene products versus the total frequency of annotation. Then after selecting a GO term related to the underlying biological phenomena being studied, the method uses semantic similarity to rank the given gene products that are annotated to the term. This enables the user to further narrow down the list of gene products and identify those that are more likely of interest.

**Keywords:** Gene identification, Gene Ontology, Protein Functional Annotation, Semantic Similarity, Information Content.

## 1. Introduction

Candidate gene identification is an active research topic that aims at associating genes to underlying biological phenomena, such as diseases and specific disorders. Many approaches have demonstrated their success in this topic but there are also many challenges to address (***1***). To address these challenges various computational methods have been proposed, which can be grouped in two approaches: genome-wide scanning and candidate gene approaches (***2***). The genome-wide scanning approach is normally based on expensive and resource intensive strategies. In contrast, the candidate gene approach is based on the knowledge about the functional characterization of the genes and therefore has been proven to be more effective when analyzing complex biological phenomena. For example, recent studies have shown that many classes of diseases with similar phenotypes are caused by functionally related genes (***3***).

Nowadays, a significant amount of knowledge about the functional characterization of the genes is already available in public databases. However, some of this knowledge is described through free-text statements that are normally ambiguous, domain specific and context dependent. To cope with this, the research community is developing and using bio-ontologies for the functional annotation of genes (***4***). The GO project (***5***) is currently the major effort in this area, having developed a bio-ontology of broad scope and wide applicability that addresses the need for consistent descriptions of gene products in different databases (***6***).

GO provides a structured controlled vocabulary composed of terms that describe gene and protein biological roles, which can be applied to different species (***5***). Since the activity or function of a protein can be defined at different levels, GO has three different aspects: molecular function, biological process and cellular component. This three-way partition is based on the following notions: each protein has elementary molecular functions that normally are independent of the environment, such as catalytic or binding activities; sets of proteins interact and are involved in cellular processes, such as metabolism, signal transduction or RNA processing; and proteins can act in different cellular localizations, such as the nucleus or membrane.

Candidate gene approaches were quick to identify the potential of using GO annotations for the functional characterization of each candidate gene. Onto-Express was one of the first tools to use GO for creating functional profiles that improved the gene expression analysis (***7***). In 2005, at least 13 other tools have been proposed based on the same ontological approach, demonstrating the importance of this topic (***8***). These ontological approaches apply a large number of different statistical tests to identify whether a set of candidate genes represent an enrichment or depletion of a GO category of interest (***9***). Recently, semantic similarity has been proposed to cluster GO terms in order to identify relevant differentially expressed gene sets (***10***).

The method presented in this chapter uses GO annotation data to identify the most meaningful functional aspects in a set of related gene products, for example, identified from a gene expression experiment. It is based on the occurrence of each GO term in that set of gene products and on the global frequency of annotation of that GO term, which are used to calculate an e-value that measures the meaningfulness of that occurrence. The occurring GO terms are then ranked by e-value, and the user can select the term(s) found to be more relevant for the study being conducted by the user. Semantic similarity is then used to rank the proteins in the set that are annotated to the term(s) selected, to further narrow down the list of proteins and help the user identify those that are more likely of interest. This method is available through the web tool ProteInOn (http://xldb.di.fc.ul.pt/biotools/proteinon/).

The remainder of this Chapter is organized as follows: Section 2 provides a general theoretical background, Section 3 describes the method, and Section 4 presents useful notes on performing the method and interpreting the results obtained during its execution. The examples (*see* **Note 1**) presented along the Chapter are derived from a proteomics analysis of cystic fibrosis that was performed using the proposed method (***11***).

## 2. Theoretical background

### 2.1 Gene Ontology

The method here described will perform a functional analysis of gene products based on their GO annotations. The GO project aims at providing a controlled vocabulary for the description of molecular phenomena in which the gene products are involved. In order to achieve that it provides three orthogonal ontologies that describe genes and gene products in terms of their associated biological processes, cellular components and molecular functions (***5***). Each ontology organizes the terms in a Directed Acyclic Graph (DAG), where each node represents a term and the edges represent a relationship between those terms. Each term is identified by an alphanumeric code (e.g. GO:0008150) and its textual descriptors, including its name, definition, and synonyms if they exist. Currently, the relationships between the terms can be of three main types: *is_a*, *part_of* and *regulates*. Due to its broad scope and wide applicability GO is currently the most popular ontology for describing gene and protein biological roles. Each of GO's ontologies describes the biological phenomena associated to gene products at different levels. Catalytic or binding activities are independent of the surrounding environment, and these are the kind of elementary molecular activities that are described by the *molecular function* ontology. On the other hand, activities of sets of proteins interacting and involved in cellular processes, such as metabolism or signal transduction are described by the *biological process* ontology. Proteins can perform their functions in several cellular localizations, such as the Golgi complex or the ribosome, this aspect being then described by the *cellular component* ontology. All the three biological aspects, biological process, molecular function and cellular component are each represented by an individual DAG. While *is_a* and *part_of* relations are only established within each hierarchy, *regulates* relations can occur across ontologies. Figure 1 shows a sub-graph of the GO ontology, where only *is-a* relationships are depicted.

Gene products are not actually incorporated into the Gene Ontology. The latter includes only terms that describethose gene products. However, the GO Consortium, through the Gene Ontology Annotation (GOA) project (***12***), does provide annotations, which are associations between gene products and the GO terms that describe them. A gene product can be annotated with as many GO terms as necessary to fully describe its function. Furthermore, because of GO's true path rule which states that "the pathway from a child term all the way up to its top-level parent(s) must always be true", a gene product which is annotated to a term such as *lipid catabolic process* is also automatically annotated to its parent term *metabolic process*.

Each annotation linking a GO term to a gene product is given an evidence code, which is an acronym that identifies the type of evidence that supports the annotation, i.e. the *IDA* code, which means *Inferred by Direct Assay* is assigned to annotations that are supported by that type of experiment.  Two main types of annotations based on their evidence codes are usually considered: manual annotations and electronic annotations. Manual annotations correspond to annotations made through manual curation, whereas electronic annotations are made through automatic means. Although electronic annotations constitute over 97% of all annotations, many studies choose to disregard them due to a common notion that they are of low quality. However, their use greatly increases GO's coverage and some studies advocate their application (***13***).

## 2.2 GO-based gene product set characterization

Using the GOA database (***12***), we can obtain the list of GO terms annotated to each gene product in a given set, and therefore the global list of GO terms that characterizes that set. However, because GO is organized hierarchically, the number of occurrences of a given GO term in a set of gene products does not accurately reflect its relevance in that set. For instance, all annotated gene products are expected to be annotated (directly or by inheritance) to the root term *biological process*, and therefore the number of occurrences is not a synonym of relevance. Thus, in order to identify the GO terms that are relevant for the characterization of a set of gene products, we need not only the frequency of occurrence of the terms in that set but also a measure of how meaningful it is to observe such a frequency. One such measure is the probability of observing a frequency equal to or greater than the observed frequency in a random set of gene products of the same size as the set of interest, which results in an e-value of the observed frequency.

This e-value can be calculated using the global frequency of annotation of each GO term in the GOA dataset as an estimator of its probability of occurrence, $P(t)$. For each GO term $t$, a protein taken at random from the dataset can be considered a random event with two outcomes: success, if it is annotated to $t$; and failure otherwise. As such, the probability of observing at least $k$ successes in a random set of $n$ gene products is given by a cumulative binomial distribution with probability of success $P(t)$:

$$P(x_t \geq k) = \sum_{i=k}^{n} \left(i^{n}\right) P(t)^{i} \left(1 - P(t)\right)^{n-i}$$

The lower the e-value, the less likely it is that the observed frequency of the term is due to chance, and thus the more meaningful is the term in the set of gene products.

In addition to filtering GO terms by e-value, it is also necessary to exclude terms that are redundant. In this context, a GO term is considered redundant if one of its descendants annotates the exact same gene products in the set. While filtering by e-value naturally excludes many of these cases, there are cases where the ancestor and descendant terms are similar in specificity and thus have similar e-values. Despite being significant according to the e-value, the ancestor term is excluded because it is redundant since its annotations are already implied by the annotations of its descendant and thus does not contribute to the characterization of the set of gene products.

For instance, if the term *actin binding* occurs in 25% of the gene products in a given set, its parent term *cytoskeletal protein binding* will necessarily occur in the same gene products by inheritance and may or may not occur in additional gene products. If it does not, it is considered redundant and is excluded.

## 2.3 Semantic Similarity

Semantic similarity in the context of ontologies can be defined as a numerical value that reflects the closeness in meaning between two ontology terms or two sets of terms annotating two entities. Commonly, the semantic similarity between two gene products annotated with GO terms is called 'functional similarity', since it gives a measure of how similar the gene products functions are.

The following sections focus first on semantic similarity applied to GO terms, and then on semantic similarity for gene products annotated with GO terms.

### 2.3.1 Semantic similarity for GO terms

There are two main approaches for GO term semantic similarity measures: edge-based and node-based. Edge-based approaches use edges and their properties as data sources. Commonly, they rely on counting the number of edges between two terms on the ontology graph, which conveys a distance measure that can easily be converted to a similarity measure (***14***). The shorter the distance between two terms, the more similar they are. Taking the terms in Figure 1, the distance between *lipid biosynthetic process* and *lipid catabolic process* is 2.

Alternatively to such distance metrics, the common path technique can be employed, which is given by the distance between the root node and the lowest common ancestor (LCA) the two terms share (***15***). In this case, the longer the distance between the root and the common ancestor, the more similar are the terms. Taking again Figure 1 to illustrate this technique, *lipid biosynthetic process* and *lipid catabolic process* are more similar than *biosynthetic process* and *metabolic process* since the former have *lipid metabolic process* as a LCA, which is at a distance of 2 from the root, whereas the latter have *biological process* as LCA, which is at a distance of 0 since it is the root. To increase the expressiveness of these measures, several properties of edges such as their type (i.e. *is_a*, *part_of*, etc) and their depth, can be used.

Node-based measures use nodes and their properties as data sources. These measures are better suited for ontologies, such as bio-ontologies, where nodes and edges are not uniformly distributed and where different edges convey different semantic distances. A commonly used node property is the information content (IC), which gives a measure of how specific a term is within a given corpus (***16***). The Gene Ontology is particularly well suited to this, since GO annotations can be used as a corpus. The IC of a term *t* can then be given by:

$$\mathrm{IC}(t) = -\log_2 f(t)$$

where $f(t)$ is the frequency of annotation of term *t*. Consequently, terms that annotate many gene products have a low IC, while terms that are very specific and thus only annotate few gene products have a high IC. Additionally, the IC can be normalized so that it returns more intuitive values.

Semantic similarity measures can use the IC by applying it to the common ancestors that two terms have, under the rationale that two terms are as similar as the information they share. The two most general approaches to achieve this are: the most informative common ancestor (MICA), which considers only the common ancestor with the highest IC (***16***); and the disjoint common ancestors technique (DCA) which considers all common ancestors that do not incorporate any other ancestor (***17***). Popular node-based measures include Resnik's, which only considers the IC of the ancestor (***16***), and Lin and Jiang & Conrath's, which consider the IC of both the ancestor and the terms themselves (***19***,***20***). Consider the subgraph of GO given in Figure 2. Using this subgraph, the Resnik similarity between *transcription factor activity* and *transcription co-factor activity* corresponds to the IC of their MICA, *transcription regulation activity*, which is 0.23.

### 2.3.2 Semantic similarity for gene products

Semantic similarity for gene products is given by the comparison of the sets of GO terms that annotate each gene product within each GO ontology. There are two main approaches that can be used for this: pairwise and groupwise (***20***).

Pairwise approaches are based on combining the semantic similarities between the terms that annotate each gene product. These approaches use only direct annotations, and apply term semantic similarity measures to all possible pairs made between each set of terms. Variations within this type of approach include considering every pairwise combination (*all pairs* technique) or only the best-matching pair for each term (*best pairs* technique). Commonly, the pairwise similarity scores are combined either by average, sum or selecting the maximum to obtain a global functional similarity score between gene products.  Consider the example in Figure 2 where two hypothetical proteins, A and B, and their annotations (direct and inherited) are shown. In this example, the *all pairs* technique would calculate the similarity for all four pairs of directly annotated terms, whereas the best pairs technique would only consider the pairs *transcription factor activity* - *transcription co-factor activity* and *transcription factor binding* - *DNA binding*. The final value would then be given by the maximum, average or sum of these similarities.

Groupwise approaches calculate similarity directly, without applying term similarity metrics. They fall into one of three categories: set, vector, or graph. Set-based measures consider only direct annotations and use set similarity metrics, such as simple overlap. Vector-based measures consider all annotations and represent gene products as vectors of GO terms, and apply vector similarity measures, such as cosine vector similarity. Graph-based measures represent gene products as the subgraphs of GO corresponding to all their annotations (direct and inherited). In this case, functional similarity can be calculated either using graph matching techniques or, because these are computationally intensive, by considering the subgraphs as sets of terms and applying set similarity techniques. A popular set similarity technique used for this is the Jaccard similarity, whereby the similarity between two sets is given by the number of elements they share divided by the number of elements they have in total. The Jaccard similarity can be applied directly to the number of terms (simUI) (***21***), or be weighted by the IC of the terms (simGIC) (***13***) to give more preponderance to more specific terms. Figure 2 illustrates this type of measures, since each node color identifies it as a term that strictly belongs to a single protein's annotations (white or dark grey) or to both (light grey). Using simUI, the similarity between the proteins would be 0.33, whereas using simGIC it would be 0.14.

The semantic similarity measures for GO terms have been developed and employed on various assessment studies. There is not one clear best measure for comparing terms or gene products. While a given measure can be suitable for one task it can perform poorly on another. Lord et al (***22***) were among the first to assess the performance of different semantic similarity measures. In that assessment Resnik's, Lin's, and Jiang and Conrath's measures were tested against sequence similarity using the average combination approach. Pesquita et al. (***13***) also tested several measures against sequence similarity and found simGIC to provide overall better results. However, as stated before some measures perform better in some situations than in others. As an example, simUI was found by Guo et al. (***23***) to be the weakest measure when evaluated for its ability to characterize human regulatory pathways, while Pesquita et al. (***13***) found it to be fairly good when evaluated against sequence similarity.

## 3. Methods

### 3.1 The ProteInOn web tool

ProteInOn is a web tool that integrates GO-based semantic similarity, retrieval of interacting proteins and characterization of gene product sets with meaningful GO terms. It uses GO, GOA, IntAct (***24***) and UniProt (***25***) as data sources.

ProteInOn implements several term and gene product semantic similarities:

- Resnik's measure calculates the similarity between two terms based strictly on the IC of their MICA (***16***).
- Lin's measure reflects how close the terms are to their MICA rather than just how specific that ancestor

is (***18***).

- Jiang and Conrath's measure is based on a hybrid approach derived from edge-based notions with IC as a decision factor (***19***).
- SimUI defines semantic similarity as the fraction between the number of GO terms in the intersection of those graphs and the number of GO terms in their union. This measure accounts for both similar and dissimilar terms in a simpler way than finding matching term pairs (***21***).
- SimGIC also uses the fraction between the number of GO terms in the intersection of those graphs and the number of GO terms in their union, but weights each term by its IC, thus giving more relevance to more specific terms (***22***).

Every semantic similarity measure can be calculated using either the MICA or the DCA approach (***26***). They are also used to calculate gene product similarities using the best pairs technique and the average combination approach.

ProteInOn normalizes the IC to values between 0 and 1 so that all measures also return values between 0 and 1, which can be directly transformed to a percentage of similarity.

ProteInOn is also able to characterize a set of gene products with the top 100 most representative GO terms, by returning a list of the GO terms that annotate the set of gene products ordered by their e-value as discussed previously on section 2.2.

## 3.2 Gene finding approach

This section describes how to use ProteInOn (Fig. 3) for gene finding, by combining several ProteInOn features.

The input for this task is a set of gene products, which is used to generate a list of all GO terms annotated to the given gene products. The tool will calculate the e-value for each of these GO terms and display the sorted list to the user, who may then select the GO terms better related to the biomedical problem being studied. The selected GO terms are then used to select only the gene products from the input set that are annotated with them, and the remaining can thus be disregarded. Finally, the semantic similarity between all remaining gene products is calculated to further support the identification of relevant gene products.

## 3.3 How to use ProteInOn for gene finding

### 3.3.1 Input preparation

Candidate gene approaches are normally based on a list of differentially expressed genes. Our method however, only requires a set of gene products, regardless of their expression values, thus any mechanism able to produce a list of gene products is suitable to generate the input set.

The input set consists of a list of UniProtKB accession numbers for the gene products to be analyzed, which consist of 6 alphanumeric characters, for example: P23508, O00559, Q4ZG55. However, if your gene products are not identified by UniProtKB accession numbers, you can use UniProtKB mapping service (http://www.uniprot.org/mapping/) to convert common gene IDs and protein IDs to UniProtKB accession numbers and vice versa (Fig. 4).

### 3.3.2 Finding relevant Gene Ontology terms

After preparing your input set, you can start using the ProteInOn tool.

1. Go to http://xldb.fc.ul.pt/biotools/proteinon/ and choose *find GO term representativity* from the drop-down menu on *Step 1: Query* (Fig. **3**).
2. On *Step 2: Options*, you will be presented with another drop-down menu where you are able to choose the GO type (*see* **Note 2**) on which you want to focus your analysis: *molecular function*, *biological process* or *cellular component*.
3. In *Step 2: Options* you can also choose to ignore electronic annotations by checking the *ignore IEA* box.
4. On *Step 3: Input* insert the protein list on the input box, taking notice it shouldn't be more than 1000 proteins long and press *Run*.

5. A list of the GO terms (*see* **Note 3**) annotated to the input set of gene products is displayed, with the GO terms ordered by e-value (Fig. 5).
6. This resulting list can be saved in either XML or TSV (under the *Step 3: Input box*) enabling posterior analysis on ProteInOn or other software.
7. You can also save a bar chart that shows the occurrence of the top 10 most representative GO terms (Fig. 6)

### 3.3.3 Gene product semantic similarity based on selected GO terms

You can continue your analysis by calculating the semantic similarity between the most relevant gene products:

1. From the list of ranked GO terms choose up to ten terms by checking their respective check boxes (*see* **Note 4**).
2. Choose the option *compute protein semantic similarity* from the drop-down menu on the *Step 1: Query* box (Fig. 5). On the *Step 2: Options* box select the semantic similarity measure to be used, as previously discussed. Also, the decision about whether or not to use electronic annotations can be controlled here (with the *ignore IEA* checkbox). The input area on the *Step 3: Input* box will be locked, since the gene products that will be used in the current query, correspond to the sub-set of the original query that is annotated with any of the previously selected GO terms.
3. After clicking the *Run* button, a list of gene product pairs (*see* **Note 5**) with their respective functional similarity scores is displayed (Fig. 7). As before, this list can also be saved either in XML or TSV format for posterior or external use.

## 4. Notes

1. Along this chapter, and as example of use and interpretation of ProteInOn results, we present the

analysis of the set of gene products used in a proteomics analysis of cystic fibrosis (***11***). Nevertheless, much of the interpretation is dependent on the context of the experiments which produced the original data and on the biomedical problem being addressed.

2. In our example, a set of 34 UniProtKB accession numbers used in (***11***) was given as input for ProteInOn's *find GO term representatitvity* option. We chose to analyze the *biological process* ontology, since it is evidently the most interesting for candidate gene identification. However, *molecular function* terms can also be of interest, and although the *cellular component* ontology is of limited usefulness in this context, it may serve as a means of validating results.
3. The *find GO term representativity* option results in a list of relevant GO terms ranked by e-value, which includes the number and frequency of occurrences of each term in the gene product dataset, and the information content of the term to provide a measure of its specificity.
4. This choice can be based on e-value alone, however users should take occurrence and IC values into consideration as well. For instance, in some cases there maybe some relatively general terms (with information content below 0.3) whose occurrence is very significant because they are highly overrepresented in the dataset. This is often due to an inherent experimental bias, for example, in an experiment based on cell membrane proteins, it is expected that *binding* and *protein binding* are overrepresented terms. In these cases, the presence of these terms is useful only as an additional validation of the experimental results, and should generally be ignored for the purpose of identifying functional aspects of interest. Another critical aspect in selecting GO terms is the duality between specificity and frequency of occurrence (or representativity). For a given threshold e-value, often there are several related terms with the more general terms occurring in more gene products than the more specific ones. For instance, in our example (Fig. 5) the term *complement activation* and its ancestor *inflammatory response* have similar e-values because despite being less specific the latter occurs in one protein more than the former. Thus, we have to choose between the specificity of the functional aspect considered and its representativity of the dataset, a choice which naturally depends on the context of the problem. To support this we recommend the utilization of a GO browser (http://www.geneontology.org/GO.tools.browsers.shtml) to investigate the relations between GO

terms with similar e-values.

5. In our example, we selected the term *multicellular organismal development,* so ProteInOn could calculate the semantic similarity between the subset of gene products that are annotated with it (Fig. 7). This enables researchers to analyze the topology of this subset of gene products at the functional level, and identify clusters of similar gene products within that subset that merit a more detailed analysis. Most of the semantic similarity values between the gene products annotated to the term *multicellular organismal development* are low, which is not unexpected considering that this term is fairly general. However, by selecting the gene products of the highest scoring pair and using them as input for the *find assigned GO terms* option we find that these gene products have an interesting set of annotations. These are shown in Figure 8 and include terms relevant for cystic fibrosis such as *innate immune response*. In fact one of these proteins, Complement C1s subcomponent, is related to early onset of multiple autoimmune diseases, whereas the other, Keratin, type II cytoskeletal 1, is related to several genetic skin disorders caused by defects in the gene KRT-1. Thus, by applying ProteInOn's gene finding method we were able to identify two relevant candidate genes to cystic fibrosis.

## 5. References

1. Tabor, H. K., Risch, N. J. and Meyers, R. M. (2002) Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nature Reviews Genetics*. **3**, 391-7.

2. Zhu, M. and Zhao, S. (2010) Candidate Gene Identification Approach: Progress and Challenges. Int J Biol Sci **3**,420-427

3. Oti, M and HG Brunner, H.G. (2007) The modular nature of genetic diseases. *Clinical Genetics* **71**, 1–11.

4. Bodenreider O. and Stevens, R. Bio-ontologies: current trends and future directions. (2006) *Briefings in Bioinformatics*. **7**, 256-274.

5. Gene Ontology Consortium, The (2000) Gene Ontology tool for the unification of biology. *Nature Genetics.* **25**, 25–29.

6. Bada, M., Stevens, R., Goble C., Gil, Y., Ashburner, M., Blake, J.A., Cherry, J.M., Harris, M. and Lewis, S.

(2004) A short study on the success of the Gene Ontology. *Web Semantics: Science, Services and Agents on the World Wide Web, 2003 World Wide Web Conference*, **1**, 235-240.

7. Khatri, P. Draghici, S., Ostermeier, G.C. and Krawetz S.A. (2002) Profiling gene expression using onto-express. *Genomics*. **79**, 266-270.

8. Khatri, P and Drăghici, S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*. **21**, 3587-3595.

9. Rivals, I., Personnaz, L., Taing, L. and Potier MC. (2007) Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics* **23**, 401-407

10. Xu, T., Gu, J., Zhou, Y. and Du, L. (2009) Improving detection of differentially expressed gene sets by applying cluster enrichment analysis to Gene Ontology. *BMC Bioinformatics*. **10**, 240.

11. Charro, N., Hood, B.L., Pacheco, P., Azevedo, P., Lopes, C., Almeida A.B., Faria, D., Couto, F.M., Conrads, T.P. And Penque, D. (2010) Serum Proteomics Signature of Cystic Fibrosis Patients: a Complementary 2-DE and LC-MS/MS Approach. *Journal of Proteome Research*. Submitted for publication.

12. Barrell, D., Dimmer, E., Huntley, R.P., Binns, D., O'Donovan, C. and Apweiler, R. (2009) The GOA database in 2009--an integrated Gene Ontology Annotation resource. *Nucleic Acids Research* **37**, D396-D403.

13. Pesquita, C., Faria, D., Bastos, H., Ferreira, A., Falcão, A. and Couto F.M. (2008) Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics* **9**, S4.

14. Rada, R., Mili, H., Bicknell, E. and Blettner, M. (1989) Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man and Cybernetics.***19**, 17-30.

15. Wu, Z. and Palmer, M.S. (1994) Verb semantics and lexical selection. *Proceedings of the 32nd. Annual Meeting of the Association for Computational Linguistics (ACL 1994)*. pp. 133–138.

16. Resnik, P. (1995) Using information content to evaluate semantic similarity in a taxonomy. *In Proc. of the 14th*

*International Joint Conference on Artificial Intelligence*.

17. Couto, F.M., Silva, M.J. and Coutinho, P.M. (2005) Semantic similarity over the gene ontology: Family correlation and selecting disjunctive ancestors. *Proceedings of the ACM Conference in Information and*

*Knowledge Management*.

18. Lin, D. (1998) An information-theoretic definition of similarity. *Proceedings of the 15th International Conference on Machine Learning*. San Francisco, CA: Morgan Kaufmann. pp. 296–304.

19. 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*, Taiwan.

20. Pesquita, C., Faria, D., Falcão, A.O., Lord, P. and Couto, F.M. (2009) Semantic Similarity in Biomedical Ontologies. *PLoS Computational Biology* **5**, e1000443.

21. Gentleman, R. (2005) Visualizing and Distances Using GO. URL http://www.bioconductor.org/docs/vignettes.html .

22. Lord, P., Stevens, R., Brass, A. and Goble, C. (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*. **19**, 1275–1283.

23. Guo X, Liu R, Shriver CD, Hu H, Liebman MN (2006) Assessing semantic similarity measures for the characterization of human regulatory pathways. *Bioinformatics* **22**, 967–973.

24. Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., Bridge1, A., Derow, C., Feuermann, M., Ghanbarian, A.T., Kerrien, S., Khadake, J., Kerssemakers, J., Leroy, C., Menden, M., Michaut, M., Montecchi-Palazzi, L., Neuhauser, S.N., Orchard, S., Perreau, V., Roechert, B., van Eijk, K. and Hermjakob, H. (2010) The IntAct molecular interaction database in 2010. *Nucleic Acids Research*. **38**(Database issue), D525-D531.

25. UniProt Consortium, The (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Research***.** **38**(Database issue), D142-D148.

26. Faria, D. Pesquita, C., Couto F.M. and Falcão A. (2007) ProteInOn: A Web Tool for Protein Semantic Similarity. DI/FCUL TR 07-6, Department of Informatics, University of Lisbon.

**Figure captions:**

**Figure 1** – Sub-graph of GO's biological process ontology.

**Figure 2** – Illustration of graph-based semantic similarity. Full lines are GO edges, dashed lines represent annotation identified with their evidence codes.

**Figure 3** – ProteInOn web tool's homepage

**Figure 4** – Web tool for mapping external database identifiers into UniProt identifiers. Source: http://www.uniprot.org

**Figure 5** – GO terms annotated to the list of proteins used in the proteomics analysis of cystic fibrosis. The terms are ordered by relevance as determined by their e-value. Additionally, the occurrence (absolute and percentage) and the information content of each term is presented.

**Figure 6** – Bar chart of the occurrence of the most representative GO terms within the list of proteins used in the proteomics analysis of cystic fibrosis.

**Figure 7** – Semantic similarity scores for pairs of representative gene products annotated with the GO term *multicellular organismal development*.

**Figure 8** – GO terms from the *biological process* ontology that annotate the selected proteins, Complement C1s subcomponent and Keratin type II cytoskeletal.