# Profiling of Semantically Annotated Proteins

Hollunder J[1], Mironov V[2*], Antezana E[2], Hoehndorf R[3], Kuiper M[2]

[1]Department of Plant Systems Biology, Flanders Institute for Biotechnology and Department of Plant Biotechnology and Bioinformatics,
Ghent University, Technologiepark 927, B-9052 Ghent Belgium

[2]Department of Biology, Norwegian University of Science and Technology (NTNU), Høgskoleringen 5, N-7491 Trondheim, Norway

[3]Department of Physiology, Development and Neuroscience, University of Cambridge, Downing Street, Cambridge CB2 3EG, UK

[*]Corresponding author {vladimir.n.mironov@gmail.com}

The first two authors contributed equally

**Abstract.** We have exploited semantic annotations of biological entities to develop a novel approach to infer new knowledge. We demonstrate this in four use cases based on the Gene Expression Ontology, an applied ontology that we developed to serve the needs of researchers involved in the analysis of genes and proteins implicated in transcriptional control of pathways/diseases. We have found that semantic annotations associated with biological entities in various commonly used data sources support the identification of related entities, thereby emulating associations that can be inferred from sequence or other structural similarities between these entities. We demonstrate how those semantic annotations can be used to make inferences about the respective biological entities.

**Keywords:** ontology, annotation, semantic similarity, gene expression, pattern identification, hypothesis generation

## 1 Gene Expression Ontology

The Gene Expression Ontology (GeXO) [1] is an application ontology that integrates fragments of GO and the Molecular Interaction ontology (MI) with data from GOA, IntAct, KEGG, SwissProt, and NCBI Gene. It also includes information on predicted orthology relations among the proteins. The knowledge in GeXO covers three biological species: human, mouse, and rat. GeXO comprises 168,417 terms of which 39,680 correspond to proteins. In the present study we attempted to assess the global implicit informational value contained in GeXO.

## 2 Semantic profiles of protein terms

Protein features were extracted from GeXO in the form 'predicate-object.' The features form a matrix with 39680 rows corresponding to proteins and 132360 columns to features. The types of features we used are summarized in Table 1.

**Table 1.** Ten sets of features of proteins in GeXO, with their subject name space, predicate and object name space.

| namespace | predicate | namespace | count |
|---|---|---|---|
| NCBIGene | codes_for | UniProtKB | 21184 |
| GO | contains | UniProtKB | 1025 |
| IntAct | has_agent | UniProtKB | 30538 |
| UniProtKB | has_function | GO | 2619 |
| GO | has_participant | UniProtKB | 339 |
| UniProtKB | has_source | NCBITaxon | 3 |
| UniProtKB | is_a | SSB | 5 |
| UniProtKB | member_of | KEGG | 2045 |
| UniProtKB | orthologous_to | UniProtKB | 14413 |
| GeXO | transformation_of | UniProtKB | 60189 |

This feature matrix was used to compute semantic similarity among all the proteins in the data set on the basis of the Jaccard index weighted by the information content as described in [2]. We evaluated the quality of the computed semantic similarities using ROC analysis.

For classifying false and true positives we used KEGG clusters of orthologous proteins as positive sets. The KEGG cluster annotations were removed from the data set prior to the analysis. The results in Figure 1 demonstrate the very high predictive value of semantic annotations (the results with the full data are given just as a reference). To exclude an impact of sequence information on the analysis, we removed orthology information from the data set, which affected the results only slightly. We concluded that semantic annotations are able to reveal protein similarity even in the absence of sequence information.

## 3 Patterns in semantic profiles

To identify recurrent patterns in the data, we used the DASS tool [3], which finds *closed sets* in the data. Closed sets have the property that there is no superset that occurs more frequently in the data set. A data set consists of a set of sets of *elements* (referred to as *host sets*).

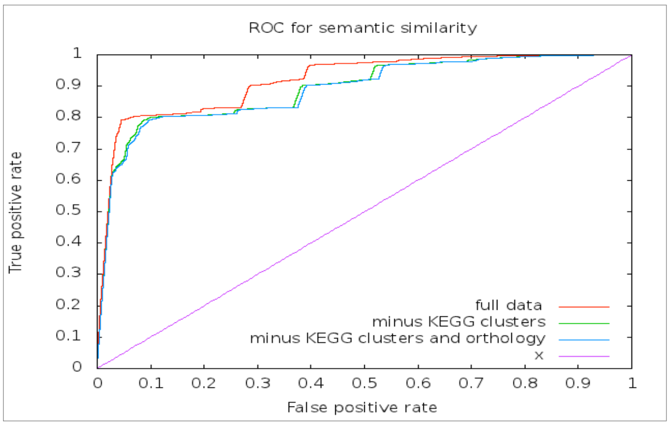

**Fig. 1**. **ROC analysis of semantically annotated proteins:** KEGG clusters of proteins were used for classifying true *vs.* false positives. Analysis was performed on **(1)** the full set of annotations, AUC 0.92; **(2)** the same as **(1)** but without KEGG cluster annotations, AUC 0.89; **(3)** the same as **(2)** but without orthology annotations, AUC 0.89.

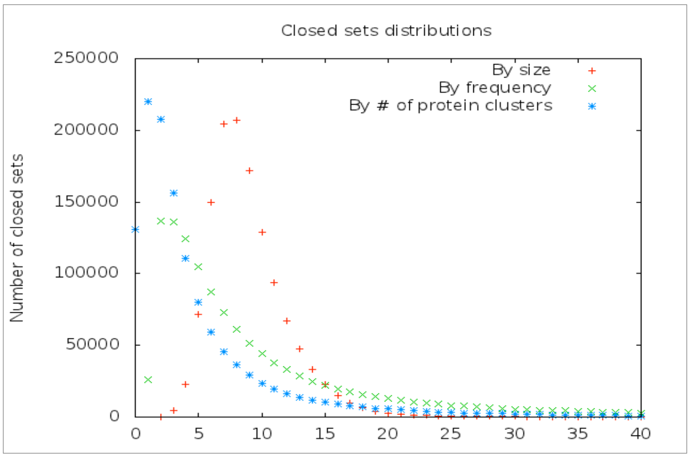

**Fig. 2. Distribution of closed sets:** The distribution of closed sets by **(1)** the number of features in the set, **(2)** the number of times the set occurs in the data, **(3)** the number of KEGG protein clusters associated with the set. All the distributions had a long tail excluded from the plot. Very similar trends were observed for closed sets with p-values below 0.05.

Figure 2 provides the distribution of *closed sets* according to the size (number of *elements*), or frequency (number of occurrence in the data set). The vast majority of *closed sets* fall within a narrow range of the size and frequency. The sets of low frequency are likely to be highly predictive due to their specificity. To have a more precise view on the predictive value of the set we focused on KEGG clusters, which define functionally distinct protein types, associated with *closed sets*. Figure 2 gives the distribution of *closed sets* according to the number of associated clusters. The highest number of sets was found to be associated with a single KEGG cluster, thus confirming the high predictive value of the *closed sets*. It is worth noting the very high number of *closed sets* without any associated KEGG cluster. In combination with the results of the ROC analysis this suggests that the *closed sets* could be used to classify the proteins associated to those *closed sets*.

## 4 Use cases

To demonstrate the high predictive value of the closed sets, we extracted a subset containing 106 transcription factors (TFs) of 40 distinct types known or suspected to be involved in the response to the hormon gastrin.

The 106 TFs were subjected to clustering with a number of approaches on the basis of associated closed sets. The resulting clusters were used as templates for screening the total GeXO data set to identify hypothetical TFs and target genes (TGs). For an initial validation of the identified candidates, we downloaded additional information from UniProtKB (lookup for the term: *gastric* in http://uniprot.org) and mapped it on the identified candidates. We identified more than 1700 potential candidates including more than 460 genes with transcriptional activity (non-deep analysis, automated screening by extracting information from http://www.uniprot.org/uniprot/ with a customized Python script). Furthermore, we identified 53 known TGs and 28 genes linked to the term *gastric,* whereas 11 TGs and 7 gastric genes occur in more than two clustering solutions and represent (partially) supported hypotheses. Thus, these results show that we can use the *closed sets* concept for predicting TGs and regulators involved in *response to gastrin*. Additionally, we identified more than 400 novel candidates occurring in more than two clustering solutions. Evidently, not all of these candidates are directly involved in this response, but they represent a good basis for further (more detailed) analyses as well as possible wet-lab experiments.

## References

1. Venkatesan, A., Mironov, V., Kuiper, M.: **Towards an integrated knowledge system for capturing gene expression events.** ICBO, 3rd International Conference on Biomedical Ontology (2012)
2. Hoehndorf, R., Schonfield, P. N., Gkoutos, G. V.: **PhenomeNET: a whole-phenome approach to disease gene discovery.** Nucl. Acids Research, 39, e119 (2011)
3. Hollunder, J., Friedel, M., Beyer, A., Workman, C. T., Wilhelm, T., **DASS: efficient discovery and p-value calculation of substructures in unordered data.** Bioinformatics, 23, 77 (2007)