GOclasses: molecular function as viewed by proteins

Daniel Faria, Catia Pesquita, Francisco M. Couto and André O. Falcão*

Dept. of Informatics, Faculty of Sciences, University of Lisbon, Portugal

ABSTRACT

Motivation: The Gene Ontology (GO) is now the standard for describing gene and protein functions. By studying the protein function space as described by the GO, we can improve the quality of functional predictions and automated annotations and ultimately, the quality of the GO structure.

Results and Conclusions: We present an approach for the large scale study of the protein function space, and propose a set of strategies to mine that space with the goal of detecting erroneous or inconsistent annotations as well as underlying relationships in the GO.

1 INTRODUCTION

'What does this gene do?' is a question biologists have posed time and again over the last decade. In our quest to address it, experimental determination of function gave way to function prediction based on sequence alignments, as the number of published genetic sequences grew beyond the resources for the former. We also adopted ontologies for functional annotation, as we stumbled upon the barriers of human language and realized that traditional descriptions of gene product function were subjective and unamenable to computation.

The Gene Ontology (GO) (The Gene Ontology Consortium 2000) has become the standard for describing gene product function in a cellular context, and is used extensively to annotate gene and protein databases. While the annotation effort is far from over, we can now for the first time have a glimpse of what the protein function space looks like by studying the topology of the GO annotation space. In doing so, we will expand our knowledge on protein functions, which will in turn allow us to better identify and correct missannotated proteins, and to improve the structure of the GO itself. Furthermore, by studying the relationship between the protein function space and the protein sequence space, we can improve the quality of function predictions and annotate new proteins more accurately.

In this context, Lord et al. (2003) have applied semantic similarity measures to GO as a means to compare proteins on a functional level, and have correlated it with sequence similarity. Several other authors have since developed new semantic similarity measures for comparing proteins (Chabalier et al. 2007, Pesquita et al. 2008, Pozo et al. 2008, Schlicker et al. 2006, Sheehan et al. 2008).

On a distinct approach, Cross and Yi (2008) have proposed the application of Formal Concept Lattices to GO annotations as a means of clustering and mapping proteins on the functional space.

Of particular relevance for this paper, Schlicker and Albrecht (2008) have introduced the concept of GO annotation class (or GOclass) as the set of GO term annotations shared by one or more proteins. The authors noted that there were fewer combinations of GO term annotations than there were proteins, and adopted this concept to reduce the number of calculations necessary to calculate semantic similarity between all UniProt (The UniProt Consortium, 2008) proteins.

In this paper, we delve further into the concept of GOclass, and apply it to the *molecular function* annotations of UniProt proteins, as a means to study the protein function space on a large scale.

2 METHODS

2.1 Data Sources

The GOclasses used in this work were derived from the ProteInOn database (Faria et al. 2007), which integrates the GO, GOA (Barrel et al. 2009) and UniProt database. The update of ProteInOn used in this work was dated of September 26th, 2008, and included the most recent releases of its component databases available at that date.

The calculations of the annotation frequency and information content of each term were made as previously described (Faria et al. 2007), as was the implementation of the simGIC measure (Pesquita et al. 2008).

2.2 GOclasses

In accordance with Schlicker and Albrecht (2008), we define a GOclass as the set of GO terms constituted by the direct and inherited annotations of a given protein or set of proteins, irrespective of their evidence codes. However, in this work the concept of GOclass is only applied to *molecular function* annotations and will only be used in that context.

The following terminology will be used throughout the paper:

* To whom correspondence should be addressed.

- An annotation (or a term of GOclass) is redundant if it is implied by another (more specific) annotation of the same protein (or another term of the same GOclass).
- The set of non-redundant terms of a GOclass is the minimum set of terms necessary to completely identify a given GOclass.
- An annotation is incomplete if it isn't sufficiently specific to suitably describe the real function of the protein, likely due to lack of knowledge about what that function is. For instance, the non-redundant annotation of a protein with the term *binding* is always incomplete, because a protein must necessarily bind to something, and thus be annotated with *'something' binding*. By extension, a GOclass is incomplete if it contains at least one incomplete annotation.
- An annotation is erroneous if it describes a functional aspect the protein doesn't have in reality. A protein is erroneously assigned to a given GOclass if any of its annotations are erroneous.
- A GOclass is inconsistent if it contains terms that don't belong to the set of terms most commonly used to describe the protein function it represents, or if it doesn't contain all the terms in that set, as the result of divergence of annotation criteria.
- A GOclass specifies another GOclass if it includes one or more terms that are descendents of terms of the other GOclass and if all other terms are equal between both GOclasses.

3 RESULTS AND DISCUSSION

3.1 The Protein Function Space

There are 33,346 *molecular function* GOclasses in UniProt, which is less than 1% of the total number of proteins with *molecular function* annotations (3.8 million). The true number of distinct protein functions is likely even smaller, considering that many of the GOclasses are artifacts resulting from incomplete, erroneous or inconsistent annotations.

Remarkably, nearly half the GOclasses (15,094) are singletons (*i.e.* occur in only one protein), which is somewhat surprising considering the ubiquity of functional inference based on sequence alignments. Although some of these singletons are likely artifacts, the number of singleton GOclasses is probably a good estimation of the true number of functional singletons in nature, precisely because of the ubiquity of functional inference. The fact that these are singletons suggests that their function was determined rather than inferred, and thus that their annotations are likely more reliable. Indeed the fraction of manually curated annotations in these singletons is 17.6%, whereas the global fraction of manually curated *molecular function* annotations is only 0.6%. Another interesting aspect of the singletons is the fact that they do not contain terms significantly more specific than the remaining GOclasses (with the exception of 443 GOclasses that include terms that are themselves singletons). This means that most singletons are unique due to unusual combinations of functional aspects which are not unusual in themselves.

On the other end of the spectrum, the 80 GOclasses with more proteins represent 50% of all proteins. The majority of these are GOclasses with a single (often generic) term, such as the four most populous classes: {*transporter activity*}, {*structural molecule activity*}, {*transcription factor activity*} and {*DNA binding*}. However, there are exceptions such as the fifth: {*cytochrome-c oxidase activity*; *electron carrier activity*; *iron ion binding*; *copper ion binding*; *heme binding*}.

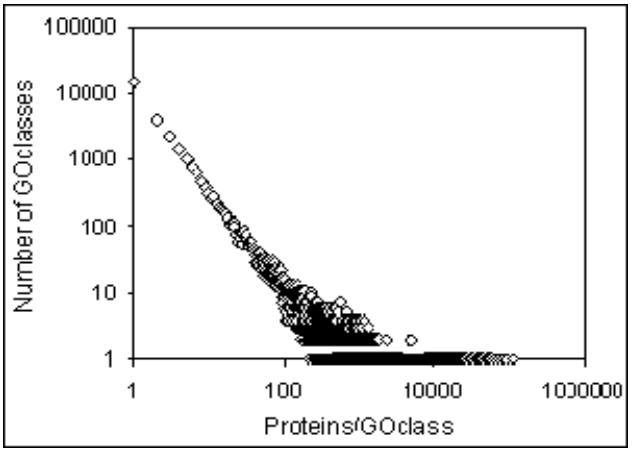

Fig. 1. Distribution of the number of proteins per GOclass in a log-log scale. As evidenced by the linear behaviour, the distribution follows a power law until around 300 proteins per GOclass. Beyond that range the average behaviour of the distribution deviates slightly from the power law, although it is not evident due to the typical fluctuations in the data.

As can be seen in Figure 1, the distribution of the number of proteins per GOclass follows closely a power law. This behaviour is common in natural phenomena (Clauset et al., 2009), and was to be expected in the case of this data due to the ramifying evolution of life. As organisms diverged genetically and phenotypically, the protein functions widespread through nature became very few and the unique functions became many. However, despite the fluctuations observed in the tail of the distribution (Figure 1), which are typical due to the low frequency of GOclasses in that range (Clauset et al., 2009), there is a noticeable deviation from the power law behaviour. That means that GOclasses with many proteins are occurring more frequently than would be expected if the data followed a power law distribution. This is likely due to the fact that many GOclasses in that range

are incomplete and do not correspond to detailed protein functions, as is clearly the case of the classes {*transporter activity*} and {*structural molecule activity*}. Thus, it is crucial to indentify the GOclasses that are populous because they describe a function that is widespread in nature, such as {*transcription factor activity*} and {*cytochrome-c oxidase activity*; …} and exclude those that are populous because they are incomplete.

One of the interesting properties of power law distributions is their scale invariance (Clauset et al., 2009). This means that despite the increasing number of proteins in our databases, we can expect the distribution of proteins per GOclass to keep the same behaviour, including the scaling exponent that defines the slope in the logarithmic scale.

3.2 *Molecular Function* as viewed by Proteins

The number of non-redundant terms per GOclass ranges from 1 to 14, with 57% of the *molecular function* terms having a GOclass for themselves, and 68 GOclasses having 10 or more terms (see Figure 2). On average, it takes 3 *molecular function* terms to describe the function of a protein, a number that is likely to increase as annotations become more complete.

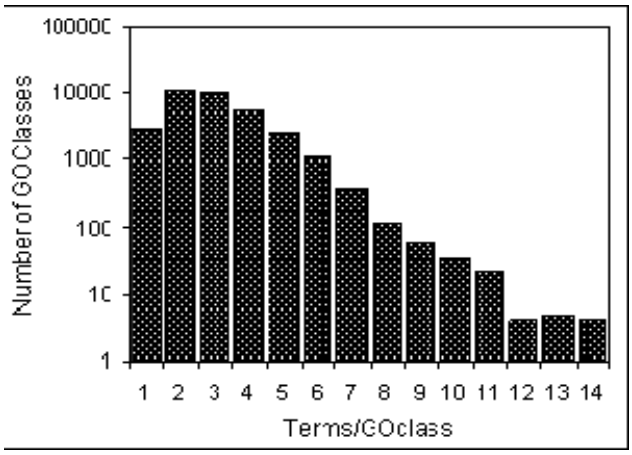

Fig. 2. Distribution of the number of non-redundant terms per GOclass.

Why do some protein functions require so many terms to be suitably described? Do these functions include so many distinct aspects, or are there implicit relationships between their terms that should be formalized in the GO graph? Furthermore, are all proteins with the same function being described consistently with the same sets of terms?

The following are strategies that can be employed to mine the annotation space in order to address these questions, which will allow us to improve the quality of our annotations and the structure of GO to better reflect the protein function space.

3.2.1 Information Content

The information content (IC) provides a measure of a term's specificity based on its frequency of occurrency. Using the IC, we define as primary the most specific term belonging to a GOclasses and as secondary all remaining non-redundant terms.

As can be seen in Figure 3, most GOclasses have a fairly specific primary term, with an average IC of 57%. However the large majority also have at least one secondary term that is more general, with an average IC of 26%. For instance, there are 957 GOclasses (totalling 82,211 proteins) which have the term *binding* as secondary term and 362 GOclasses (totalling 54,422 proteins) with *catalytic activity* as a secondary term. These are cases which obviously need our attention, yet further analysis is necessary to determine if they correspond to incomplete or inconsistent annotations, or even if there is an underlying relationship that is not explicit in the GO.

By searching for other GOclasses that contain the primary term of our GOclass of interest, we can see if there are GOclasses that specify our GOclass regarding the general secondary term. If there are, then that secondary term may be a case of incomplete or inconsistent annotation, with the former being more likely if there are several populous GOclasses that specify our GOclass of interest and the latter being more likely if there is only one populous GOclass.

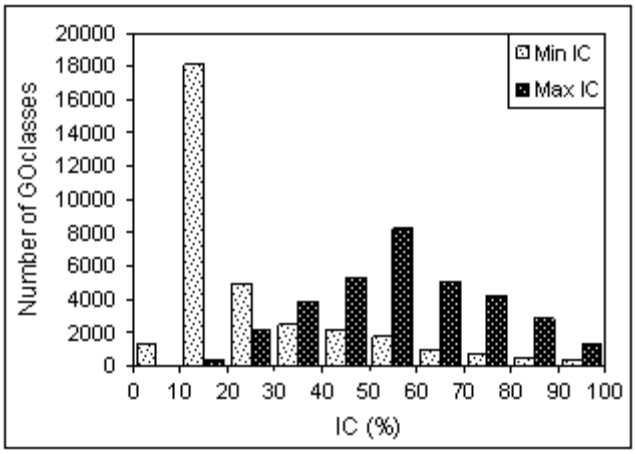

Fig. 3. Distribution of the minimum and maximum information content (IC) per GOclass (as determined by its non-redundant terms), in discrete intervals of 10% IC.

3.2.2 Conditional Probability

We can estimate the conditional probability of a secondary term occurring when the primary term occurs by determining the number of GOclasses and the corresponding number of proteins in which the terms occur together and dividing by the total number of proteins in which the primary term occurs. If the conditional probability is high, then there may be an underlying relationship between the primary and sec-

ondary terms, which can be considered for formalization in the GO or at least put forth as a guideline for annotation. The GOclasses that contain the primary term but not the secondary term should then be analyzed, to assess if they are cases of inconsistent annotation or exceptions which prevent the underlying relationship from being formalized.

3.2.3 Semantic Similarity

Calculating the semantic similarity between GOclasses can also help us identify cases of inconsistent or incomplete annotation. While it is obvious that there are similar functions within nature, very high simGIC values between two GOclasses mean that these classes differ only on general terms, and thus merit a detailed analysis.

Although 91% of all pairwise combinations of GOclasses have simGIC values below 10%, there are 4,583 pairs with semantic similarity values between 90 and 100% and 12,544 with values between 80 and 90% (not counting singleton GOclasses which were excluded due to computational restraints).

3.2.4 A Case Study

As an example, let us select the most populous GOclass with more than one term, class: {*cytochrome-c oxidase activity*; *electron carrier activity*; *iron ion binding*; *copper ion binding*; *heme binding*}, which has 63,658 proteins and as primary the term *cytochrome-c oxidase activity*, with an IC of 25%.

There are 62 other GOclasses that contain this term, totalling 89,166 proteins. The most populous of these GOclasses (with 14,298 proteins) has the same terms as our case study except for term *copper ion binding* and has a semantic similarity of 96% with our case study.

The conditional probability calculations reveal that *iron ion binding* has a 91% probability of being annotated when *cytochrome-c oxidase activity* is, *electron carrier activity* has a 90% probability, *heme binding* has an 89% probability and *copper ion binding* has a 73% probability.

These findings would impell us to seek further information on term *cytochrome-c oxidase activity*, upon which we would discover that the function it describes includes implicitly the remaining terms found in our case study. Whether the relationships between these terms should be formalized in GO is beyond the scope of this paper, however, this information can at least help us correct incomplete or inconsistent annotations.

4 CONCLUSIONS

We have presented an approach to study the protein function space as described by GO *molecular function* annotation based on the concept of GOclass. The result that the distribution of proteins per GOclass follows closely a power law suggests that despite the possible lack of quality of many annotations, the topology of the protein function space is similar to the annotation space.

Furthermore, we propose a set of strategies to mine the annotation space with the goal of identifying erroneous, incomplete or inconscistent annotations, or even underlying relationships in the GO. These strategies will allow us to improve the quality and consistency of our annotations (particularly automated annotations) as well as improve the structure of the GO to better reflect the functional concepts present in nature.

ACKNOWLEDGEMENTS

This work was supported by the FCT through the Multiannual Funding Programme, and the doctoral grants SFRH/BD/29797/2006 and SFRH/BD/42481/2007.

REFERENCES

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 Res.*, **37**, D396–D403.

Chabalier,J., Mosser,J. and Burgun,A. (2007) A transversal approach to predict gene product networks from ontology-based similarity. *BMC Bioinformatics*, **2**, 235.

Clauset,A., Shalizi,C.R. and Newman, M.E.J. (2009) Power-law distributions in empirical data. *SIAM Review*, accepted for publication.

Cross,V. and Yi,W. (2008) Multiple views for ontology-based formal concept lattices. *NAFIPS 2008, Annual Meeting of the North American*, 1–6.

Faria,D., Pesquita,C., Couto,F.M. and Falcao,A.O. (2007) ProteInOn: A Web Tool for Protein Semantic Similarity, *DI/FCUL TR 07-06*, Dpt. Informatics, Univ. Lisbon.

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**, 10, 1275–1283.

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

Pozo,A.D., Pazos,F. and Valencia,A. (2008) Defining functional distances over gene ontology. *BMC Bioinformatics*, **9**, 50.

Schlicker,A., Domingues,F.S., Rahnenfhrer,J., and Lengauer,T. (2006) A new measure for functional similarity of gene products based on gene ontology. *BMC Bioinformatics*, **7**, 302.

Schlicker,A. and Albrecht,M. (2008) FunSimMat: a comprehensive functional similarity database. *Nucleic Acids Res.*, **36**, D434–D439.

Sheehan,B., Quigley,A., Gaudin,B. and Dobson,S. (2008) A relation based measure of semantic similarity for gene ontology annotations. *BMC Bioinformatics*, **9**, 468.

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

The UniProt Consortium (2008) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **36**, D190-D195.