Sources of Experimental Function Annotation in UniProt-GoA and implications for Function Prediction

Alexandra Schnoes[1], Alexander Thorman[2] and Iddo Friedberg*[2,3]

[1]Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA, USA
[2]Department of Microbiology, Miami University, Oxford, OH USA
[3]Department of Computer Science and Software Engineering , Miami University, Oxford, OH USA

Email: Alexandra Schnoes - alexandra.schnoes@ucsf.edu; Alexander Thorman - thormanaaw@muohio.edu; Iddo Friedberg*-i.friedberg@muohio.edu;

*Corresponding author

Abstract

Background: Computational protein function prediction programs rely upon well-annotated databases for testing and training their algorithms. These databases, in turn, rely upon the work of curators to capture experimental findings from scientific literature and apply them to protein sequence data. However, due to high-throughput experimental assays, it is possible that a small number of experimental papers could dominate the functional protein annotations collected in databases. Here we investigate just how prevalent is the "few papers – many proteins" bias. We discuss how this bias affects our view of the protein function universe, and consequently our ability to predict protein function.

Results: We examine the annotation of UniProtKB by the Gene Ontology Annotation project (GOA), and show that the distribution of proteins per paper is a log-odd, with X papers dominating X% of the annotations. Since each of the dominant papers describes the use of an assay that can find only one function or a small group of functions, this leads to a substantial bias in what we know about the function of many proteins.

Conclusions: Given the experimental techniques available, the protein function annotation bias is unavoidable. Knowing that this bias exists and understanding its extent is important for database curators, developers of function annotation programs, and anyone who uses protein function annotation data to plan experiments.

Background

Functional annotation of proteins is a cardinal challenge in molecular biology today. The continuing revolution in sequencing technology means that the conversation has shifted from realizing the $1000 genome to the one-hour genome [?]. This ongoing race to extremely rapid and low cost genomic data is creating a flood of sequence data that requires intensive analysis and characterization before it can be useful for the general scientific public. A large proportion of this computational work will involve ascribing functional data to these newly determined sequences, a process that is neither simple nor inexpensive [?]. However, this functional annotation work is critical in order for the potential of these sequence data to be fully realized. Recent work has shown that the ability to accurately ascribe function through computational means is challenging and open problem [?]. In order to aid current annotation procedures and improve computational function prediction algorithms, sources of high-quality, experimentally derived functional data are necessary. Currently, one of the few repositories of such data is the UniProt-GOA database [?], which contains both computationally derived and literature derived functional information. The literature derived information is extracted by human curators who capture functional data from publications, assign the data to its appropriate place in the Gene Ontology hierarchy [?] and label them with appropriate functional evidence codes. The UniProt-GOA database is one of only a small number of databases that explicitly connects functional data, publication reference and evidence codes to specific, experimentally studied sequences. In addition, annotations captured in UniProt-GOA directly impact the annotations in the UniProt/Swiss-Prot database, widely considered the largest gold standard set of functional data [?] available.

It is important, therefore, to understand the some of trends and biases encapsulated by the UniProt-GOA database as it impacts well-used sister databases and function prediction algorithm development and training. One concern surrounding the capture of functional data from papers, is the propensity of high-throughput experimental work to become a large proportion of the data in UniProt-GOA. While high-throughput experimentation has revolutionized our ability to examine function on a large scale, typically the read-outs from these experiments are only very general function descriptions that lack specificity and, therefore, are assigned high in the GO ontology tree (depth in the GO tree indicates functional specificity). These data are arguably less useful than explicit descriptions of enzyme activity or biological role. However, given the prevalence of high-throughput experimentation, there is potentially a ”few-paper, many proteins” bias in which a small number of papers (and thereby, non-specific function annotations) dominate the database content. The goal of our study here was to examine whether

UniProt-GOA contains an annotation bias towards high throughput, low information papers.

Results and Discussion

Papers and proteins

As described in the Background section, with the advent of high-throughput experiments it has become possible to conduct large-scale interrogations of protein functions. Some papers therefore reveal one or more functional aspects of a large amount of proteins which respond to the particular type of interrogation conducted. To understand how prevalent this phenomenon is, we looked at the UniprotKB gene ontology annotation files, or UniProt GOA. UniProtKB is annotated by GO terms both manually and automatically using an exacting procedure described in [**?**]. Briefly, there is a six-step procedure which includes sequence curation, sequence motif analyses, literature-based curation, reciprocal BLAST [**?**] searches, attribution of all resources leading to the included findings, and a quality assurance phase. If the annotation source is a research article, the attribution includes a PubMed ID. For each GO term associated with a protein, there is also an *evidence code* with which is used to explain how the association between the protein and the GO term was made. Experimental evidence codes include such terms as: Inferred by Direct Assay (IDA) which indicates that “a direct assay was carried out to determine the function, process, or component indicated by the GO term” or *Inferred from Physical Interaction* (IPI) which “Covers physical interactions between the gene product of interest and another molecule.” (Quotes are from the GO site, geneontology.org). The computational analysis evidence codes are generally considered less reliable than the experimental ones, and include terms such as *Inferred from Sequence or Structural Similarity* (ISS) and *Inferred from Sequence Orthology* (ISO). However, these are still assigned by a curator. There are also non-computational and non-experimental evidence codes, the most prevalent being *Inferred from Electronic Annotation* (IEA) “Used for annotations that depend directly on computation or automated transfer of annotations from a database”. IEA evidence means that the annotation was not made by a curator, and is not checked manually.

Different degrees of reliability are associated with the evidence codes, with experimental codes considered to be of higher reliability than non-experimental codes. However, due to an ongoing increase in high-throughput experimental methods, we suspected that high-throughput experiments may dominate the protein annotation landscape.

To test our hypothesis, we first examined assignments with two degrees of reliability: experimental evidence codes (EXP, IDA, IPI, IMP, IGI, IEP), and all others. The reason for this partition into two groups is that we wanted to know which, if any, high throughput experiments dominate the protein function annotation landscape. This we compared with a baseline of annotations of all types of evidence, and with those which are non-experimental. The results are shown in Figure**??**.

As can be seen in Figure**??**, the distribution of proteins annotated per paper follows a log-odds ZZZcheckthis ratio. Many proteins are annotated by a few papers. (ZZZNeed to look at the percentage of proteins annotated by the top 10, top 20 papers.) We therefore conclude that there is indeed a bias in experimental annotations, in which there are few papers that annotate a large number of proteins.

Our next question was how much annotation bias these papers introduce. We decided to look who are the top contributing papers to protein annotation in UniProtKB-GOA, and what type of GO terms they contribute. The results to the first question are summarized in Table**??**. As can be seen, almost all of the papers are specific to a single species (typically a model organism) and assay that is used to annotate the proteins in that organism. Since a single assay was used, then typically only one ontology (MF, BP or CC) was annotated.

Annotation bias in model organisms

To see how much a single species– and method– specific large-scale assay affects the entire annotation of a species, we examined the relative contribution of each paper to the entire corpus of experimentally annotated protein in that species. All the species we examined were model organisms, as all the top annotation-contributing papers dealt with model organisms. The results are summarized in Figure**??**.

Annotation quality

One reflection on annotation coverage is the number of GO terms assigned to any given protein. Ostensibly, the larger the number of GO terms that are assigned to a protein, the more comprehensive its annotation, provided that these terms are non-redundant, i.e. not direct parents of each other.

The median number of annotations per protein was 1.09, which means that most of the top-50 papers do not provide more than a single GO-term per annotation. Also, the mean number of annotations per protein was 1.59. The difference between the mean and median reflects that most papers have an annotations-per-protein ration which is close to 1, whereas few have a higher annotation ratio. As shown in Figure **??**, that is essentially the case. Seven papers had an annotations per proteins ratio which is above 2. Those were paper numbers 1, 4, 10, 23, 34 and 43 on the list. We decided to examine these papers to see whether these studies did indeed provide better coverage, and what distinguished them from the other high throughput studies. To follow are brief summaries of the methodologies used in these papers.

Toward a confocal subcellular atlas of the human proteome.

In this microscopy-based study, the authors used specific staining techniques and confocal microscopy to describe the subcellular localization of 4937 proteins, which were each assigned to one or more of ten different subcellular compartments. Since proteins may be assigned to more than a single compartment, there is a mean of 2.23 GO annotations per protein. The most frequent term was GO:0016235 *centrosome* which is defined as "an inclusion body formed by dynein-dependent retrograde transport of an aggregated protein on microtubules" [**?**].

Conclusions

Text for this section ...

Methods

Databases used

Text for this sub-section ...

Another methods sub-heading for this section

Text for this sub-section ...

Yet another sub-heading for this section

Text for this sub-section ...

Authors contributions

Text for this section ...

Acknowledgements

Text for this section ...

References

Figures

Figure 1 - Sample figure title

A short description of the figure content should go here.

Figure 2 - Sample figure title

Figure legend text.

Tables

Table 1 - Sample table title

Here is an example of a *small* table in LaTeX using `\tabular{...}`. This is where the description of the table should go.

My Table		
A1	B2	C3
A2	...	..
A3	..	.

Table 2 - Sample table title

Large tables are attached as separate files but should still be described here.

Additional Files

Additional file 1 — Sample additional file title

Additional file descriptions text (including details of how to view the file, if it is in a non-standard format or the file extension). This might refer to a multi-page table or a figure.

Additional file 2 — Sample additional file title

Additional file descriptions text.