Title

(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

1 Author1 Dept/Program/Center, Institution Name, City, State, Country

2 Author2 Dept/Program/Center, Institution Name, City, State, Country

3 Author3 Dept/Program/Center, Institution Name, City, State, Country

$*$ **E-mail: Corresponding author@institute.edu**

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, with the increasing use of high-throughput experimental assays, a small number of experimental papers dominate the functional protein annotations collected in databases. Here we investigate just how prevalent is the "few papers – many proteins" phenomenon. We hypothesize that the dominance of high-throughput experiments in proteins annotation biases our view of the corpus of functions enabled by proteins.

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 0.06% of papers dominating 20% 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 substantial biases, in several aspects, in what we know about the function of many proteins.

Conclusions: Given the experimental techniques available, protein function annotation bias due to high-throughput experiments is unavoidable. Knowing that these biases exist and understanding their characteristics and extent is important for database curators, developers of function annotation programs, and anyone who uses protein function annotation data to plan experiments.

Author Summary

Introduction

Functional annotation of proteins is a primary challenge in molecular biology today [1–3]. The continuing revolution in sequencing technology means that the conversation has shifted from realizing the $1000 genome to the one-hour genome [**?**]. The ability to rapidly and cheaply sequence genomes is creating a flood of sequence data, which require extensive analysis and characterization before they can be useful. A large proportion of this work involves assigning biological function to these newly determined gene sequences, a process that is both complex and costly [4]. Furthermore, the ability to accurately assign function through computational means is challenging and open problem [5]. 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 [6], 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 [7] 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 references 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 to be a gold standard set of functional annotation [5].

It is important, therefore, to understand any trends and biases that are encapsulated by the UniProt-GOA database, as those impact well-used sister databases and therefore a large number of users worldwide. Furthermore, any biases would impact function prediction algorithms development and training.

One concern surrounding the capture of functional data from papers is the propensity for high-throughput experimental work to become a large fraction of the data in UniProt-GOA, this having few experiments dominate the protein function landscape. In this work we analyzed the relative contribution of papers to the experimental annotations in UniProt-GOA. We found some striking biases, stemming from the fact that a small fraction of papers that describe high-throughput experiments, disproportionately contribute to the pool of experimental annotations of model organisms. Consequently, we show that: 1) annotations coming from high-throughput experiments are mostly less informative than those provided by low-throughput experiments; 2) annotations from high throughput experiments bias the annotations towards a limited number of functions, and, 3) many high-throughput experiments overlap in the proteins they annotate, and in the annotations assigned. Taken together, our findings offer a comprehensive picture of how the current protein function landscape is generated. Furthermore, due to the biases inherent in the current system of sequence annotations, this study serves as a caution to the producers and consumers of biological data from high-throughput experiments.

Results

Articles and Proteins

With the advent of high-throughput experiments it has become possible to conduct large-scale studies of protein functions. Consequently, some studies reveal very specific functional aspects of a large amount of proteins as a result of the particular type of assay or assays used. To understand the impact of large-scale studies on the corpus of experimentally annotated proteins, we looked at the UniprotKB Gene Ontology annotation files, or UniProt-GOA. UniProt-GOA proteins are individually annotated by one or more GO terms. using a procedure described in [6]. Briefly, this procedure consists of six steps which include sequence curation, sequence motif analyses, literature-based curation, reciprocal BLAST [8] searches, attribution of all resources leading to the included findings, and quality assurance. If the annotation source is a research article, the attribution includes its PubMed ID. For each GO term associated with a protein, there is also an *evidence code* 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.” (Taken from the GO site, geneontology.org). Computational evidence codes 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) which is “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 or checked by a person. Different degrees of reliability are associated with the evidence codes, with experimental codes generally considered to be of higher reliability than non-experimental codes. However, the increase in the number of high-throughput experiments used to determine protein functions may introduce biases into protein annotations, due to the inherent capabilities and limitations of high-throughput assays.

To test the hypothesis that such biases exist, and to assess their extent if they do, we compiled the details of all experimentally-annotated proteins in UniProtKB. This included all proteins whose GO annotations have the GO experimental evidence codes EXP, IDA, IPI, IMP, IGI, IEP. We first examined the distribution of articles by the number of proteins they annotate. The results are shown in Figure **??**.

As can be seen in Figure**??**, the distribution of the number of proteins annotated per paper follows a power-law distribution, $f(x) = a\dot{x}^k$. Using the goodness-of-fit based method we found a significant fit to $a = 7; k = 2.59$. We therefore conclude that there is indeed a substantial bias in experimental annotations, in which there are few papers that annotate a large number of proteins.

To better understand the consequences of such a distribution, we divided the annotating articles into four cohorts, based on the number of proteins each article annotates. *Single-throughput* papers are those papers that annotate only a single protein; *low throughput* papers annotate 2-9 proteins; *moderate throughput* papers annotate 10-99 proteins and *high throughput* papers annotate over 99 proteins. The results are shown in Table **??**. High throughput papers are responsible for 21% of the annotations in Uniprot-GOA, even though they comprise 0.07% of the papers. 66% of the papers are single-throughput and low throughput, however those annotate 53% of the proteins in Uniprot-GOA. So while moderate throughput and high-throughput experiments account for slightly under half of the annotations in Uniprot-GOA, the number of experiments in those cohorts is much smaller than low throughput experiments and single-throughput experiments.

What typifies high-throughput papers? Also, how may the log-odds distribution bias what we understand of the protein function universe? To answer these questions, we examined different aspects of the annotations in the four paper cohorts. Also, we examined in higher detail the top 50 annotating papers. (Overall, 62 papers in our study annotated more than 100 proteins).

An initial characterization of the top 50 high-throughput papers is shown 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 (MFO, BPO or CCO) was used for annotation. For some species this means that a single functional aspect (MFO, BPO or CCO) of a species will be dominated by a single experiment.

Term frequency bias

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 the top-50 papers to the entire corpus of experimentally annotated protein in each 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 Figure1.

Terms that were found to be frequent were:

Another type of annotation bias may result due to GO-structure derived redundancy:

the annotation of a single term using a given GO-term and one or more of its parents. However, this type of annotation may not necessarily be wholly redundant. For example, a protein may localize to the nucleolus and the nucleus itself. In the Cellular Compartment ontology, “nucleus” is a parent term for “nucleolus”. In the case of this hypothetical protein, annotating with “nucleolus” and “nucleus” is not an error. However, if the protein was found to localize to the nucleolus only, then annotating with both terms is a redundancy. We termed such a redundancy *propagation bias*. We examined all proteins for possible propagation bias. If a protein was annotated with a GO term and one or more parent terms, the parent terms were removed. We found possible propagation biases in 17 of the 50 papers. The results are summarized in Table**??**.

Ontology type bias

The term frequency bias appears to be a direct result of the ontology bias described here. The proteins annotated by single-protein papers, low-throughput papers, and moderate throughput papers (less than 100) have similar ratios of the fraction of proteins annotated. Twenty-two to twenty-six percent of assigned terms are in the Molecular Function Ontology, and 51-57% are in the Biological Process Ontology and the remaining 17-25% are in the Cellular Component ontology. This ratio changes dramatically with high-throughput papers (over 99 terms per paper). Now only 5% of assigned terms are in the Molecular Function Ontology, 38% in the Biological Process Ontology and 57% in the Cellular Compartment Ontology, ostensibly due to a lack of high-throughput assays that can be used for generating annotations using the Molecular Function Ontology.

Reannotation Bias

Another type of annotation bias is that of protein re-annotation. How many of the top-50 papers actually re-annotate the same set of proteins? On the one hand, independent experimental confirmation does have it merits. On the other hand, repeated similar annotations of the same proteins in the same organisms using similar assays may be simply indicate a confirmation bias. To investigate the extent of repetitive annotations in different papers, we clustered all the proteins annotated by the top-50 papers using CD-HIT [**?**] at 100% sequence identity. We then examined the number of clusters containing 100% identical sequences per model species. The product of the number of proteins divided by the number of clusters is the redundancy percentage. For example, if each of the top-50 papers annotating the proteins in a given species annotated the same protein set, the redundancy percentage would be 100%. The results of confirmation bias analysis are shown in Figure **??** and in Table **??**. As can be seen, the highest percent redundancy is among the ZZZ papers annotating *C. elegans*.

We have determined, therefore, that there is a some degree of repetition between experiments in the proteins they annotate, with some overlap being quite high. However, there is still the need to determine the extent of the repetition of the annotation. We therefore analyzed the 100% sequence identity clusters for overlap in annotation. To do so, we counted the number of identical GO-terms per ontology within each cluster, and divided that by the sum of GO-terms shared between all papers in the cluster. The result is a number between 0 and 1. Zero means no GO-terms are shared, while one means all GO-terms are shared.

The results are shown in Figure **??** and in Table **??**. In *S. cerevisiae*, four papers contribute to the Cellular Component ontology, by annotating 635 proteins which are common between two or more of these papers. Among those proteins, 79.6% of the terms produced are identical.

Quantifying annotation information

A common assumption holds that while high-throughput experiments do annotate more protein functions than low-throughput experiments, the former also tend to be more shallow in the predictions they provide. The information provided, for example, by a large-scale protein binding assay will only tell us if two proteins are binding, but will not reveal whether that binding is specific, will not provide an exact K_{bind}, will not say under what conditions binding takes place, or whether there is any enzymatic reaction or signal-transduction involved. Having on hand data from experiments with different "thorughputness" levels, we set out to investigate whether there is a difference in the information provided by high-throughput experiments vs. low-throughput ones. To answer this question, we first have to quantify the information given by GO terms. One way to do so, is to use the depth of the term in the ontology: the term "enzyme activity" would be less informative than "dehalogenase" and the latter will be less informative than "haloalkane dehalogenase". We therefore counted edges from the ontology root term to the GO-term to determine term information. The larger the number of edges, the more specific – and therefore informative – the annotation. In cases where several paths lead from the root to the examined GO-term, we used the minimal path. We did so for all the annotating papers split into groups by the number of proteins each paper annotates.

Edge counting provides a measure of term-specificity. It is, however, imperfect. The reason is that different areas of the GO DAG have different connectivities, and terms may have different depths unrelated to the intuitive specificity of a term. For example "high-affinity Tryptophan transporter", (GO:0005300) is 14 terms deep, while "anticoagulant", (GO:0008435) is only three terms deep. For this reason, information content, the logarithm of the inverse of the GO term frequency in the corpus is generally accepted as a measure of GO-term information content [?]. Therefore, to account for the possible bias created by the GO-DAG structure, we also used the log frequency of the terms in the experimentally annotated proteins in Uniprot-GOA. However, it should be noted that the log-frequency measure is also imperfect because, as we see throughout this study, a GO-term's frequency may be heavily influenced by the top annotating papers, injecting a circularity problem into the use of this metric in the first place. Since no single metric for measuring the information conveyed by a GO term is wholly satisfactory, we used both in this study.

The results of both analyses are shown in Figure **??** and the accompanying Table **??**. In general, the results from the depth-based analysis and the log-frequency based analysis are in agreement, when compared across groupings based on the number of proteins annotated by the papers. For the Molecular Function ontology, the distribution of edge counts and log-frequency scores decreases as the number of annotated proteins per-paper increases. For the Biological Process ontology, the decrease is significant. However the contributer to the decrease are the high-throughput papers while there is little change in the first three paper cohorts. Finally, there is no significant trend of GO-depth decrease in the Cellular Component Ontology. However, using the information content metric, there is also a significant decrease in information content in the high-throughput paper cohort.

Annotation consistency

Another interesting question was how consistent were the annotations between different experiments?

Evidence and Assertion

There are two complementary ways by which we come to knowledge about a protein's function. The approximately 20 GO evidence codes, discussed above, encapsulate the type results by which the function was inferred, but they do not capture all the necessary information. For example, "Inferred by Direct Assay (IDA)" informs that experimental evidence was used, but does not say which type of experiment was performed. This information is often needed for several reasons. Knowing which experiments were performed can help the researcher establish the reliability and scope of the data. For example, RNA for RNAi experiment does not traverse the blood-brain-barrier, meaning that no data from the central nervour system can be drawn from an RNAi experiment. In addition to evidence terms, the ECO ontology provides *assertion terms* in in which the nature of the assay is given. For example, an enzyme-linked immunosorbent assay (ELISA) provides quantitative protein data *in vitro* while an immunogold assay may provide the same information, and cellular localization information *in vivo*. It is therefore important to know both the assertion and the evidence to understand what sort of information may be gleaned from the assay. However, to understand which types of assertions are made in the top-50 high throughput papers, we performed manually assigned ECO assertion terms to the top-50 papers. The results are shown in Figure **??** and in Table **??**.

Interestingly, the most frequently used assertion in the top experimental papers was not an experimental assertion, but rather a computational one: the term ECO:00053 "computational combinatorial evidence" is defined as "A type of combinatorial analysis where data are combined and evaluated by an algorithm." This is not a computational prediction per-se, but rather a combination of several experimental lines of evidence used in a paper.

The most used experimental assertion term was ECO:000160 "protein separation followed by fragment identification evidence", which captures different types of mass-spectrometry experiments. The next ranking assertion terms were computational: "motif similarity evidence" and "sequence similarity evidence used in automatic assertion". Those were generally combined with the mass-spectrometry experiments to identify protein sequence fragments reconstructed from the mass-spectrometry. Other frequently used experimental techniques used were "RNAi experimental evidence". This type of experiment was mostly with the papers that used RNA interference in studying *C. elegans.*

Discussion

We have identified several annotation biases in UniProt-GOA. These biases stem from the uneven number of annotations produced by different types of experiments. It is clear that results from high-throughput experiments contribute heavily to the function annotation landscape. At the same time, these experiments produce less information per protein than moderate–, low– and single–t̃hroughput experiments as evidenced by the type of terms produced in the Molecular Function and Biological Process ontologies. Furthermore, the number of total GO terms used in the high-throughput experiments is much lower than that used in low and medium throughput experiments. Therefore, while high throughput experiments provide a higher coverage of protein function space per experiment, it is the low throughput experiments that provide information richness.

We have also identified several types of biases that are contributed by high throughput experiments. First, there is the enrichment of low-information content GO-terms.

Taken together, these annotation biases affect our understanding of protein function space. This, in turn, affects out ability to properly understand the connection between predictors of protein function and the actual function – the hallmark of computational function annotation. As a dramatic example, during the Critical Assessment of Function Annotation experiment [?] we have noticed that about 25% of the proteins participating in the challenge were annotated as “protein binding”. This GO-term is not an informative one. Furthermore, it was shown that the major contribution of this term to the CAFA cahallenge data set was due to high-throughput assays. (Over 100 annotated proteins annotated per experiment). Obviously, such a large bias in prior probabilities is can adversely affect programs employing prior probabilities to that always predicts (Need to write more here.)

Several steps can be taken to address this situation. Annotations are derived from high-throughput experiments can be flagged as such in the database. The flagging can then be read by sequence similarity or other search software, and flagged proteins removed or otherwise tagged in the search. In a typical scenario, a researcher will BLAST their query protein to determine its function by sequence similarity. If a target protein is tagged as annotated by a high throughput assay, it would be removed form the search if asked to do so by the user. This filtering can also be done by assay type, number of proteins annotated per experiment, or a combination of the above. This requires that GO-annotated proteins should also be annotated with assertion codes in addition to the evidence codes and GO term-codes; but given the large volume of data in UniprotKB is it hard to expect such massive reannotation with assertion terms undertaken. (Any other ideas?)

Materials and Methods

Acknowledgments

We thank Predrag Radivojac, members of the Friedberg and Babbitt labs for insightful discussions. This research was funded, in part by NSF / ABI XXXXXXX-XXX award to IF and XXXXXXXXXXXX to PCB.

References

References

1. Friedberg I (2006) Automated protein function prediction–the genomic challenge. Brief Bioinform 7: 225–242.

2. Erdin S, Lisewski AM, Lichtarge O (2011) Protein function prediction: towards integration of similarity metrics. Current Opinion in Structural Biology 21: 180 - 188.

3. Rentzsch R, Orengo CA (2009) Protein function prediction the power of multiplicity. Trends in Biotechnology 27: 210 - 219.

4. Sboner A, Mu X, Greenbaum D, Auerbach R, Gerstein M (2011) The real cost of sequencing: higher than you think! Genome Biology 12: 125+.

5. Schnoes AM, Brown SD, Dodevski I, Babbitt PC (2009) Annotation error in public databases: Misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol 5: e1000605+.

6. Dimmer EC, Huntley RP, Alam-Faruque Y, Sawford T, O'Donovan C, et al. (2012) The uniprot-go annotation database in 2011. Nucleic Acids Research 40: D565–D570.

7. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. Nature Genetics 25: 25–29.

8. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, et al. (1997) Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic acids research 25: 3389–3402.

Figure Legends

Tables