Source of Functional Annotations in UniProtKB and implications for Function Prediction

Charles A Darwin*[1,2], Jane E Doe*[2] and John RS Smith[3]

[1]Life Sciences Department, Kings College London, Cornwall House,Waterloo Road, London, UK
[2]Department of Zoology, Cambridge, Waterloo Road, London, UK
[3]Marine Ecology Department, Institute of Marine Sciences Kiel, Düsternbrooker Weg 20, 24105 Kiel, Germany

Email: Charles A Darwin*- charles@londonzoo.co.uk; Jane E Doe*- jane.e.doe@cambridge.co.uk; John RS Smith - john.RS.Smith@cambridge.co.uk;

*Corresponding author

Abstract

Background: Computational protein function prediction programs rely upon well-annotated databases for training their algorithms. These databases, in turn, rely upon the work of curators applying experimental findings from scientific literature to protein sequence data. However, with the advent of various high-throughput experimental assays, there are a few papers that dominate the protein annotation field. 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

Text for this section.

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 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 highest reliability. In this study, we 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 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 next 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 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.

To see how much a single species– and method– specific large-scale assay affects the entire annotation of that species, we examined the relative contribution of each paper to the entire annotation of the species. The results are summarized in Figure**??**.

Results sub-heading

This is a sub-sub-heading

Sub-sub-sub-headings are made with the

subsubsection command.

pb at end of lines ensures correct paragraph spacing.

Text for this sub-sub-section . . .

Another sub-sub-sub-heading

Text for this sub-sub-section . . .

Another results sub-heading

Text for this sub-section . . .

Yet another results sub-heading

Text for this sub-section. More results . . .

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 . . .

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.