Chapter 8

Evaluating Computational Gene Ontology Annotations

Nives Škunca, Richard J. Roberts, and Martin Steffen

Abstract

Two avenues to understanding gene function are complementary and often overlapping: experimental work and computational prediction. While experimental annotation generally produces high-quality annotations, it is low throughput. Conversely, computational annotations have broad coverage, but the quality of annotations may be variable, and therefore evaluating the quality of computational annotations is a critical concern.

In this chapter, we provide an overview of strategies to evaluate the quality of computational annotations. First, we discuss why evaluating quality in this setting is not trivial. We highlight the various issues that threaten to bias the evaluation of computational annotations, most of which stem from the incompleteness of biological databases. Second, we discuss solutions that address these issues, for example, targeted selection of new experimental annotations and leveraging the existing experimental annotations.

Key words Gene ontology, Evaluation, Tools, Prediction, Annotation, Function

1 Introduction

Sequencing a genome is now routine. However, knowledge of the gene sequence is only the first step toward understanding it; we ultimately want to understand the function(s) of each gene in the cell. Function annotation using computational methods—for example, function propagation via sequence similarity or orthology—can produce high-probability annotations for a majority of gene sequences, the next step toward understanding. But because computational function annotations often generalize the many layers of biological complexity, we are interested in **evaluating** how well these predictions reflect biological reality. In this chapter, we discuss the evaluation of computational predictions.

First, we highlight issues that make the evaluation of computational predictions challenging, with perhaps the primary challenge being the incompleteness of annotation databases: scoring as "wrong" those computational predictions that are not yet proven or disproven could overestimate the count of "incorrect" predictions, and skew perceptions of computational accuracy [1].

Christophe Dessimoz and Nives Škunca (eds.), *The Gene Ontology Handbook*, Methods in Molecular Biology, vol. 1446,
DOI 10.1007/978-1-4939-3743-1_8,

Second, we discuss solutions that address various aspects of database incompleteness. For example, some solutions directly address the incompleteness of databases by adding new experimental annotations. Yet another solution leverages existing high-quality annotations in a current release of a database, and retrospectively evaluates previous releases of the annotation databases. Intuitively, those annotations that are unchanged through multiple successive database releases may be expected to be of higher quality. Additional solutions include leveraging negative annotations, though sparse but containing valuable information, or performing extensive experimentation for a subset of functions of interest.

1.1 Sources of Gene Ontology Annotations: Curated and Computational Annotations

In practice, functional annotation of a gene means the assignment of a single label, or a set of labels; for example, this might involve using BLAST to transfer the labels from another gene. A particularly valuable set of labels for denoting gene function are those derived from the controlled vocabulary established by the Gene Ontology (GO) consortium [2], with terms such as "oxygen transporter activity," "hemoglobin complex," and "heme transport," as descriptors of a gene's Molecular Function, Cellular Component, and Biological Process.

But just as important as the annotation label itself is the knowledge of the source of the annotation. Based on their source, there are two main routes to produce annotations in the GO, and the GO Consortium emphasizes this distinction using evidence codes [3], as described in Chap. 3 [4].

The first route of annotating requires curator's expertise when assigning: be it examining primary or secondary literature to assign appropriate annotations, manually examining phylogenetic trees to infer events of function loss and gain, or deciding on sequence similarity thresholds for specific gene families to propagate annotations. As curated annotation is time consuming, the curators streamline their efforts, by focusing annotations on the 12 model organisms ([5] and Fig. 1, left). Consequently, fewer than 1% of proteins have this type of annotation in the UniProt-GOA database. Elsewhere, a recent examination of the annotation of 3.3 million bacterial genes found that fewer than 0.4% of annotations can be documented by experiment, although estimates suggest that the actual number might be above 1% [6].

The second route of annotating, **computational prediction of function**, takes high-quality curated annotations propagates them across proteins in nonmodel organisms. Once the pipeline for the computational prediction has been setup—a task which is by no means trivial—it can be relatively straightforward to obtain computational prediction of function across a large number of biological sequences. Chapter 5 [7] contains a detailed introduction to the methods used in computational annotation.

Computational prediction of function propagates annotations to the vast majority of currently annotated genes (Fig. 1, right).

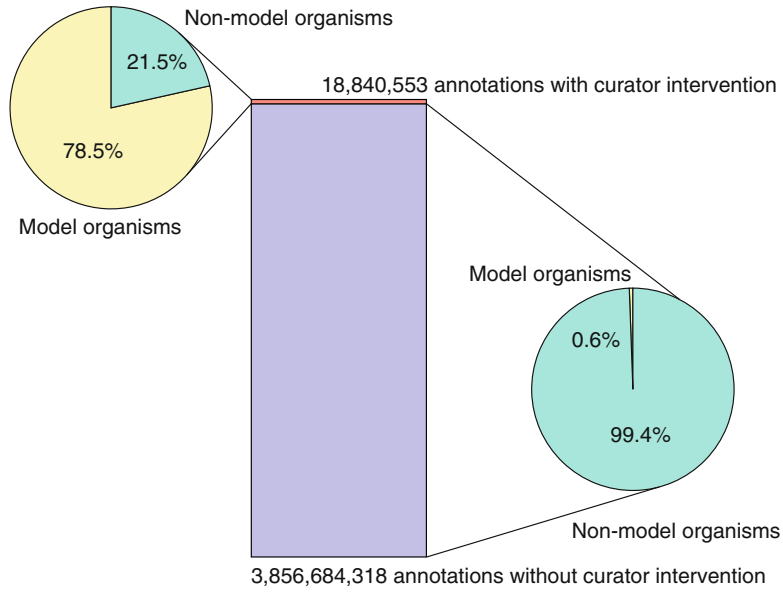

Fig. 1 The distribution of the number of computational annotations obtained *without* curator intervention (evidence code IEA) to all other annotations (evidence codes ISS, IBA, IDA, IMP, ND, IGI, IPI, ISO, TAS, ISA, RCA, IC, NAS, ISM, IEP, IGC, EXP, IRD, IKR). The 12 model organisms are: *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Arabidopsis thaliana*, *Gallus gallus*, *Danio rerio*, *Dictyostelium discoideum*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, and *Escherichia coli* K-12

Over 99 % of all annotations are created in this manner, and they are applied to approximately 76 % of all genes [6]—the remaining 24 % of genes typically have no annotation or are listed as "hypothetical protein." With the exponential growth of biological databases and the labor-intensive nature of manual curation, it is inevitable that automated computational predictions will provide the vast majority of annotations populating current and future databases.

2 Challenges of Assessing Computational Prediction of Function

Computationally predicted annotations are typically assumed to be less reliable than manually curated ones. Manual curation may be thought of as more cautious, as there is typically a single protein being labeled at a time [8], whereas the goal of computational prediction is typically more ambitious: labeling a large number of proteins—possibly ignoring subtle aspects of the biological reality.

Arguably the most accurate method to evaluate computational predictions of functions is to perform comprehensive experiments (e.g., [9]). However, given the number of computational annotations available, experimental evaluation is prohibitively expensive even for a small subset of the available computational annotations.

As a consequence of this discrepancy in numbers, two practical obstacles interfere with the assessment of computational function prediction: the elusiveness of an unbiased gold standard dataset and the incompleteness of the recorded knowledge.

2.1 The Elusiveness of an Unbiased Gold Standard Dataset

A major practical obstacle to the evaluation of computational function prediction methods is the lack of a gold standard dataset—a dataset that would contain complete annotations for representative proteins. Such a dataset should not be used to train the prediction algorithms (refer to Chap. 5 [7]) and can therefore be used to test them. In the current literature, the validation sets mimic the gold standard dataset, but they are biased:proteins that are prioritized for experimental characterization and curation are often selected for their medical or agricultural relevance, and may not be representative of the full function space that the computational methods address. Moreover, with such incomplete validation sets, it is even more difficult to evaluate algorithms specialized for specific functions—e.g., those identifying membrane-bound proteins. The gold standard dataset needs to cover a large breadth of GO terms and also have comprehensive annotations for these GO terms.

In addition to the difficulties of obtaining a gold standard dataset, the complexity of the GO graph (*see* also Chaps. 14 [10] and 2 [11])—a necessary simplification of the true biological reality—poses obstacles to comparison and evaluation. For example, it is not trivial to compare the prediction scores between the parent (more general) and the child (more specific) GO terms: consider the case when computational methods correctly predict annotations using parent terms, but give erroneous predictions for the child terms, i.e., they overpredict. Alternatively, computational predictions might miss to predict some child GO terms, i.e., they underpredict. One way of handling such situations is to use the structure of the GO to probabilistically model protein function, as described in [12].

2.2 Incomplete Knowledge

Underlying the elusiveness of the unbiased gold standard dataset is the main issue: the incompleteness of the annotation databases. When evaluating computational function annotation methods, we typically compare the predictions with the currently available knowledge. We **confirm** the computational annotation when it is available in our validation set, and we **reject** when its negation is available, e.g., via the NOT qualifier in the GO database. If negative annotations are sparse, as is often the case, it is standard practice to consider wrong a prediction when the predicted annotation is absent from the validation set, e.g., [13]. This is formally called the **Closed World Assumption (CWA)**, the presumption that a statement which is true is also **known** to be true. Conversely, under the CWA, that which is not currently known to be true is considered false.

However, the available knowledge—and consequently the validation set—is incomplete; absence of evidence of function does not imply evidence of absence of function [14]. This is formally referred to as the **Open World Assumption (OWA)**, allowing us to **formalize the concept of incomplete knowledge**. As a consequence of the incompleteness of the validation set, we might be rejecting computational predictions that later prove to be correct [1].

To illustrate the challenges related to the evaluation of function prediction, let us focus on one protein, CLC4E_MOUSE (http://www.uniprot.org/uniprot/Q9R0Q8), in particular to two computational annotations assigned to this protein at the time of writing: the OMA orthology database [15] predicted annotation with "integral component of membrane" (GO:0016021) and the InterPro pipeline predicted annotation with "carbohydrate binding" (GO:0030246). There are no available existing high-quality annotations that confirm these computational predictions.

However, if we take a closer look at these annotations, the OMA annotation "integral component of membrane," compared to the experimental annotation (evidence code IDA) of "receptor activity" is consistent with the experimental annotation: in principle, receptors are integral components of membranes. Additionally, the literature contains evidence that this protein indeed binds carbohydrates [16], thereby confirming the InterPro prediction. Therefore, if we revisit the known annotations and make these statements explicitly known to be true, we can confirm them.

Indeed, for the proteins already present in the UniProt-GOA database, we see that curators do revisited them; more than half of the proteins have already been assigned a new GO term annotation after their first introduction into the database (Fig. 2). An extreme example is provided by the Sonic hedgehog entry in mouse

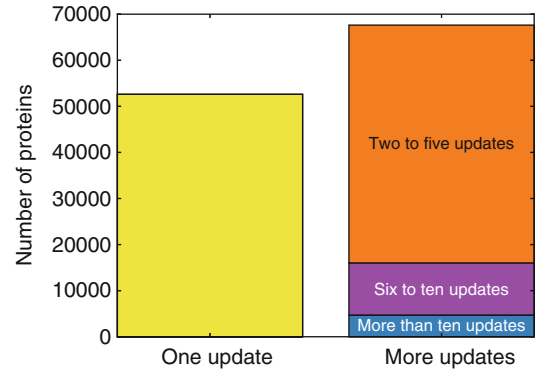

Fig. 2 Distribution of proteins based on the number of times a curator revisits a protein with an annotation from the literature (updates with evidence codes EXP, IDA, IPI, IMP, IGI, IEP). Among the proteins that have a curated annotation based on literature evidence, 56 % are subsequently updated with a new GO term

(http://www.uniprot.org/uniprot/B3GAP8), which has already been revised over a hundred times.

To meaningfully compare computational function annotations, one must account for the Closed World Assumption and have the obstacles it implies in mind. But because of the extent of the gap between the closed and the open world—think of the "unknown unknowns" in the protein function space—a quick-fix solution does not exist. However, numerous ways of tackling the problem were devised, and we turn our attention to those in the subsequent section.

3 Approaches to Test Computational Predictions with Experimental Data

To test computational predictions, experiments have to be conducted. However, the number of proteins that can be experimentally tested are dwarfed by the number of genes identified by genome sequencing, so a very small number of experimental data points must support an enormous number of predicted gene function annotations.

Among the methods to evaluate computational annotations, some are focused on quantifying the available information (e.g., the number and the specificity of annotations) without providing quality judgment (e.g., [17, 18]), while others, the topic of this section, strive to evaluate the quality of the predictions themselves. Addressing some of the complexities of evaluation addressed in the previous section, the latter methods provide good templates for future evaluations of computational methods for function prediction.

3.1 The COMBREX Initiative

The need for experimentally verified annotations is of sufficient scope that it is likely that significant progress can only be made if tackled by the entire scientific community. One such attempt at community building is focused on bacterial proteins: COMBREX (COMputational BRidge to Experiments), along with additional efforts such as the Enzyme Function Initiative [19]. The database (http://combrex.bu.edu) classifies the gene function status of 3.3 million bacterial genes, including 13,665 proteins that have experimentally determined functions [6]. The database contains traceable statements to experimentally characterized proteins, thereby providing support for a given annotation in a clear and transparent manner. COMBREX also developed a tool, named COMBLAST, to associate query genes with the various types of experimental evidence and data stored in COMBREX. COMBLAST output includes a trace to experimental evidence of function via sequence and domain similarity, to available structural information for related proteins, and to association with clinically relevant phenotypes such as antibiotic resistance, and other relevant information. It was used to provide additional annotations for 1474 prokaryotic genomes [20].

Additionally, COMBREX implemented a proof-of-concept prioritization scheme that ranked proteins for experimental testing. For each protein family, distances based on multiple alignments were calculated to help experimentalists easily identify those proteins that might be considered most typical of the family as a whole. The "ideal" COMBREX target is a protein close to many other uncharacterized proteins, and relatively far from any protein of known function, but not so far that it would preclude high-quality predictions of the protein's function for the experimentalist to test.

COMBREX helped fund the implementation of new technology for the experimental characterization of hypothetical proteins from *H. pylori* [21]. A panel of affinity probes was used in a screen to generate initial hypotheses for hypothetical proteins. These hypotheses were then tested and confirmed using traditional in vitro biochemistry. This approach is complementary to other higher throughput methods, such as the parallel screening of metabolite pools [22, 23], and activity-based proteomic approaches to identify proteins of a particular enzymatic class [24, 25].

3.2 CAFA and BioCreAtIvE

CAFA (Critical Assessment of Functional Annotation) is another community-wide effort to evaluate computational annotations, and it promises to uncover some of the most promising algorithms applied to computational function annotation [13]. Such an effort has great utility in establishing success rates of many computational annotation methods based on newly generated curator knowledge. Chapter 10 [26] covers the details of the CAFA evaluation.

Yet another community effort with a more narrow scope, introduced in Chap. 6 [27], BioCreAtIvE (Critical Assessment of Information Extraction systems in Biology) [28] is focused on evaluating annotations obtained through text mining. When evaluating in this setting, the challenges of evaluation within the open/closed world do not exist: methods are evaluated based on the amount of information they can extract from a scientific paper, which in itself has defined bounds. Evaluating the extraction quality of GO annotations for a small set of human proteins showed the extent of the work ahead—text mining algorithms were surpassed by the Precision of expert curators [29]—but also showed the areas that need to be addressed to improve the quality of computational functional annotation using text mining algorithms.

3.3 Evaluating Computational Predictions Over Time Using Successive Database Releases

A strategy to circumvent the problem of the lack of a gold standard is to consider changes in experimental annotations in the UniProt-GOA database [30].

By keeping track of annotations associated with particular proteins across successive releases of the UniProt-GOA database,

one can assess the extent to which newly added experimental annotations agree with previous computational predictions. As a surrogate for the intuitive notion of specificity, the authors defined a reliability measure as the ratio of confirmed computational annotations to confirmed and rejected/removed ones. One computational annotation is deemed confirmed or rejected, depending on whether a new, corresponding experimental annotation supports or contradicts it. Furthermore, if a computational annotation is removed, the annotation is deemed implicitly rejected and thus contributes negatively to the reliability measure. As a surrogate for the intuitive notion of sensitivity, coverage was defined as the proportion of newly added experimental annotations that had been correctly predicted by computational annotations in a previous release.

Overall, this work found that electronic annotations are more reliable than generally believed, to an extent that they are competitive with annotations inferred by curators when they use evidence other than experiments from the primary literature. But this work also reported significant variations among inference methods, types of annotations, and organisms. For example, the authors noted an overall high reliability of annotations obtained from mapping Swiss-Prot keywords associated with UniProtKB entries to GO terms. Nevertheless, there were exceptions: GO terms related to metal ion binding had low reliability in the analysis due to a large number of removed annotations. Similarly, a few annotations related to ion transport were explicitly rejected with the 'NOT' qualifier, e.g., for UniProtID Q6R3K9 ('NOT' annotation for "iron ion transport") and UniProtID Q9UN42 ('NOT' annotation for "monovalent inorganic cation transport").

3.4 Increasing the Number of Negative ('NOT') Annotations

Having a comprehensive set of negative annotations would bridge the gap between CWA and OWA; knowing both which functions *are* and are *not* assigned to a protein will not reject predictions that might later prove to be correct.

While experimentally assigning a function to protein is difficult and time consuming, it may be equally challenging to establish that a protein does *not* perform a particular function. For example, unsuccessfully testing a protein for a particular function may only indicate that it is either more difficult to demonstrate such an activity or that it is not present under the given conditions. Because the number and the combination of environmental conditions to test—e.g., the right partners or the right environmental stimulus—is numerous, obtaining a set of 'NOT' annotations might be feasible only for a subset of functions. Consequently, the negative annotations are few and far in between in annotation databases. For example, the January 2015 release of the UniProt-GOA database contains only 8961 entries that are marked with a 'NOT' qualifier.

There is a small number of reports in the literature stating that a protein does not perform a specific function (e.g., [31]),

and therefore such sporadic reports cannot be the basis for a comprehensive evaluation of computational annotations. Large-scale production of negative annotations do exists; for example, denoting a set of GO terms that are not likely to be assigned to a protein, given its known annotations (e.g., [32]). However, these are also computational *predictions*, they also need to be evaluated.

3.5 Evaluating Computational Predictions for a Specific Subset of GO Terms

The BioCreAtIvE challenge performed annotations without the challenges of the open and closed world of function annotations by focusing on defined "chunks" of information, scientific papers. In the realm of computational predictions, one of the more straightforward ways of avoiding the challenges of the closed world is to limit the scope to function where we have close to complete comprehension. In fact, by narrowing the scope of the function annotation problem, Huttenhower et al. did just that [9].

The authors evaluated the computational predictions, focusing the evaluation on functions related to mitochondrial organization and biogenesis in *Saccharomyces cerevisiae*. They trained their function prediction models only on the annotation data available in the databases, but performed comprehensive experiments for all genes in *S. cerevisiae* to check whether they have function related to mitochondrial organization and biogenesis. This way, they had information for every *S. cerevisiae* gene and were able to evaluate the prediction accuracy without the need for the distinction between the open and the closed world.

3.6 Simulation Studies

Simulation studies are abundantly used to evaluate computational methods that simulate various evolutionary events, as is done, for example, with the simulation framework for genome evolution Artificial Life Framework (ALF) [33]. In a related application of simulation, simulated erroneous annotations were used to study the quality of computational annotations—curated GO annotations obtained using methods based on sequence similarity, in the GO database denoted with the evidence code ISS [34]. First, the authors estimated the level of errors among the ISS GO annotations by checking for the effect of randomly adding erroneous annotations. Second, they obtained a linear model that connected the propensity of (artificially introduced) errors among the annotations with the estimate of Precision. Finally, they used this model to estimate the baseline Precision at the level where there are no introduced errors.

4 Outlook

Experimental annotations are key to evaluate computational methods to predict annotations. Therefore, it is highly desirable that three principles govern experimental testing of gene function: maximal leveraging of existing experimental information, maximal

information gain with each new experiment, and the development of higher throughput approaches.

Maximal leveraging of existing experimental information is easiest to obtain through the use of traceable statements, such as the use of the "with" field in the UniProt-GOA database: the "with" field can record the protein that was used as template to transfer annotation through sequence similarity. However, we could go a step further, toward statements such as: "Gene X has 96.8% sequence identity to the experimentally characterized protein 'HP0050' and therefore this protein is annotated as 'adenine specific DNA methyltransferase'." Traceable statements greatly increase the transparency of a prediction, and allow the users of gene annotations to estimate their confidence in the annotation, regardless of the source—manual curator or an automated computational prediction [35].

In order to increase information gain of new experiments, it would be beneficial to develop and incorporate experimental design principles that help guide the identification of maximally informative targets for function validation. One way to maximize the information gain from the experimental analysis is to choose proteins that generate or improve predictions for many other proteins across many genomes, as opposed to proteins related to few or no other proteins. Alternatively, for function prediction methods that report probabilities, the information gain from an experiment can be quantified as the reduction in the estimated probability of prediction error, summed across all predictions [36].

Development of higher throughput approaches for the testing of protein function is well underway, and we can hope for the same effects as with DNA sequencing. However, at the time of writing, a small number of experimental studies contribute much of the functional protein annotations collected in the databases, thereby biasing the available experimental annotations [8]. Indeed, DNA sequencing did not achieve its dramatic cost reductions and increases in throughput fortuitously, but rather was the result of the systematic investment of hundreds of millions of dollars in technology development over two decades.

Traditionally, the increases of success rates associated with computational function annotation are attributed to methodological refinements. However, we must also quantify the influence of the data available—e.g., more sequences and more function annotations—*independently* of the influence of the algorithms. This information is critical, if only because of the rate of aggregation of new information in the bioinformatics databases. Indeed, an increase in the number of sequenced genomes and an increase in the number of function annotations has a dramatic positive effect on predictive accuracy of at least one computational method of function annotation, phylogenetic profiling [37].

5 Conclusion

There are a plethora of highly accurate, readily available computational function annotation methods available to scientists, and state-of-the-art computational function annotations, such as in the UniProt-GOA database, are easily accessible to all. However, without transparent evaluation and benchmarking, it is still extremely challenging to differentiate among annotations, and annotation methods.

Going forward, the biocuration community will continue to advance along three important lines: increased amounts of biological sequence to be annotated, increased numbers of high-quality experimental annotations, and increased predictive accuracy of computational methods of annotation. In order to achieve the greatest increase in biological knowledge, we will couple the advances made in each of these three areas to reach other, especially coupling advances in the development of new algorithms with robust evaluations of these algorithms based on experimental data, with the purpose of generating new, useful biological hypotheses. Such work will contribute to closing the gap between the Open and the Closed worlds, and greatly increase our understanding of the large number new sequences that are now generated daily.

Acknowledgments

The authors thank Christophe Dessimoz and Maria Anisimova for helpful comments and suggestions. Open Access charges were funded by the University College London Library, the Swiss Institute of Bioinformatics, the Agassiz Foundation, and the Foundation for the University of Lausanne.

References

1. Dessimoz C, Škunca N, Thomas PD (2013) CAFA and the open world of protein function predictions. Trends Genet 29:609–610
2. Ashburner M, Ball CA, Blake JA et al (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29
3. Guide to GO Evidence Codes | Gene Ontology Consortium. http://geneontology.org/page/guide-go-evidence-codes.
4. Gaudet P, Škunca N, Hu JC, Dessimoz C (2016) Primer on the gene ontology. In: Dessimoz C, Škunca N (eds) The gene ontology handbook. Methods in molecular biology, vol 1446. Humana Press. Chapter 3
5. Reference Genome Group of the Gene Ontology Consortium (2009) The Gene Ontology's Reference Genome Project: a unified framework for functional annotation across species. PLoS Comput Biol 5:e1000431
6. Anton BP, Chang Y-C, Brown P et al (2013) The COMBREX project: design, methodology, and initial results. PLoS Biol 11:e1001638
7. Cozzetto D, Jones DT (2016) Computational methods for annotation transfers from sequence. In: Dessimoz C, Škunca N (eds) The gene ontology handbook. Methods in molecular biology, vol 1446. Humana Press. Chapter 5
8. Schnoes AM, Ream DC, Thorman AW et al (2013) Biases in the experimental annotations of protein function and their effect on our understanding of protein function space. PLoS Comput Biol 9:e1003063
9. Huttenhower C, Hibbs MA, Myers CL et al (2009) The impact of incomplete knowledge on evaluation: an experimental benchmark for protein function prediction. Bioinformatics 25:2404–2410
10. Gaudet P, Dessimoz C (2016) Gene ontology: pitfalls, biases, and remedies. In: Dessimoz C, Škunca N (eds) The gene ontology handbook. Methods in molecular biology, vol 1446. Humana Press. Chapter 14
11. Thomas PD (2016) The gene ontology and the meaning of biological function. In: Dessimoz C, Škunca N (eds) The gene ontology handbook. Methods in molecular biology, vol 1446. Humana Press. Chapter 2
12. Clark WT, Radivojac P (2013) Information-theoretic evaluation of predicted ontological annotations. Bioinformatics 29:i53–i61
13. Radivojac P, Clark WT, Oron TR et al (2013) A large-scale evaluation of computational protein function prediction. Nat Methods 10:221–227
14. Thomas PD, Wood V, Mungall CJ et al (2012) On the use of gene ontology annotations to assess functional similarity among orthologs and paralogs: a short report. PLoS Comput Biol 8:e1002386
15. Altenhoff AM, Skunca N, Glover N et al (2014) The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements. Nucleic Acids Res 43(Database issue):D240–D249
16. Yamasaki S, Matsumoto M, Takeuchi O et al (2009) C-type lectin Mincle is an activating receptor for pathogenic fungus, Malassezia. Proc Natl Acad Sci U S A 106:1897–1902
17. Buza TJ, McCarthy FM, Wang N et al (2008) Gene ontology annotation quality analysis in model eukaryotes. Nucleic Acids Res 36:e12
18. del Pozo A, Pazos F, Valencia A (2008) Defining functional distances over gene ontology. BMC Bioinformatics 9:50
19. Gerlt JA, Allen KN, Almo SC et al (2011) The enzyme function initiative. Biochemistry 50:9950–9962
20. Wood DE, Lin H, Levy-Moonshine A et al (2012) Thousands of missed genes found in bacterial genomes and their analysis with COMBREX. Biol Direct 7:37
21. Choi H-P, Juarez S, Ciordia S et al (2013) Biochemical characterization of hypothetical proteins from Helicobacter pylori. PLoS One 8:e66605
22. Proudfoot M, Kuznetsova E, Sanders SA et al (2008) High throughput screening of purified proteins for enzymatic activity. Methods Mol Biol 426:331–341
23. Kuznetsova E, Proudfoot M, Sanders SA et al (2005) Enzyme genomics: application of general enzymatic screens to discover new enzymes. FEMS Microbiol Rev 29:263–279
24. Cravatt BF, Wright AT, Kozarich JW (2008) Activity-based protein profiling: from enzyme chemistry to proteomic chemistry. Annu Rev Biochem 77:383–414
25. Simon GM, Cravatt BF (2010) Activity-based proteomics of enzyme superfamilies: serine hydrolases as a case study. J Biol Chem 285:11051–11055
26. Friedberg I, Radivojac P (2016) Community-wide evaluation of computational function prediction. In: Dessimoz C, Škunca N (eds) The gene ontology handbook. Methods in molecular biology, vol 1446. Humana Press. Chapter 10
27. Ruch P (2016) Text mining to support gene ontology curation and vice versa. In: Dessimoz C, Škunca N (eds) The gene ontology handbook. Methods in molecular biology, vol 1446. Humana Press. Chapter 6

28. Krallinger M, Morgan A, Smith L et al (2008) Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge. Genome Biol 9(Suppl 2):S1
29. Camon EB, Barrell DG, Dimmer EC et al (2005) An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. BMC Bioinformatics 6(Suppl 1):S17
30. Skunca N, Altenhoff A, Dessimoz C (2012) Quality of computationally inferred gene ontology annotations. PLoS Comput Biol 8:e1002533
31. Poux S, Magrane M, Arighi CN et al (2014) Expert curation in UniProtKB: a case study on dealing with conflicting and erroneous data. Database:bau016
32. Youngs N, Penfold-Brown D, Bonneau R et al (2014) Negative example selection for protein function prediction: The NoGO Database. PLoS Comput Biol 10:e1003644
33. Dalquen DA, Anisimova M, Gonnet GH et al (2012) ALF—a simulation framework for genome evolution. Mol Biol Evol 29: 1115–1123
34. Jones CE, Brown AL, Baumann U (2007) Estimating the annotation error rate of curated GO database sequence annotations. BMC Bioinformatics 8:170
35. Bastian FB, Chibucos MC, Gaudet P et al (2015) The Confidence Information Ontology: a step towards a standard for asserting confidence in annotations. Database:bav043
36. Letovsky S, Kasif S (2003) Predicting protein function from protein/protein interaction data: a probabilistic approach. Bioinformatics 19(Suppl 1):i197–i204
37. Škunca N, Dessimoz C (2015) Phylogenetic profiling: how much input data is enough? PLoS One 10:e0114701