Letter

Cell PRESS

CAFA and the Open World of protein function predictions

Christophe Dessimoz[1,2,3], Nives Škunca[3,4], and Paul D. Thomas[5]

[1] University College London, Gower St, London WC1E 6BT, UK
[2] European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, CB10 1SD, UK
[3] Swiss Institute of Bioinformatics, Universitätstr. 6, 8092 Zurich, Switzerland
[4] ETH Zurich, Computer Science, Universitätstr. 6, 8092 Zurich, Switzerland
[5] University of Southern California, Keck School of Medicine, Los Angeles, CA 90089, USA

The plummeting cost of DNA sequencing means that vast amounts of gene sequences are becoming available across all domains of life. Translating this information into useful biological and biomedical knowledge requires an understanding of the biological functions of these genes. Unlike sequencing, however, discovering the function(s) of a gene remains painstaking work that is largely restricted to a handful of model species. Thus, biological analyses, particularly of non-model organisms, increasingly rely on computational inference, or 'prediction', of function. As such, it is critical to understand the strengths and weaknesses of different prediction methods.

Towards this end, the first 'critical assessment of protein function annotation' (CAFA) experiment was recently reported, a community-wide effort to evaluate computational function prediction methods [1]. Following an open call, 23 participating teams submitted 54 algorithms for assessment that predicted gene ontology (GO) functional terms [2] for a common set of ~50 000 proteins that then lacked experimentally corroborated annotations. In the 11 months following the submission deadline, GO curators – continuing to work independently from the CAFA organisers – examined relevant literature and assigned functional annotations to 866 of those proteins. These 866 proteins became the gold standard reference set for evaluating the performance of all CAFA submissions.

We applaud this effort and fully support the aims of CAFA, but we are concerned that the primary CAFA evaluation metric fails to account for the 'Open World' assumption underlying GO annotations [3]: the functional annotations of most proteins are incomplete and, consequently, absence of evidence of function does not amount to evidence of absence of function. This omission leads to a systematic overestimation of false-positive prediction rates, which may significantly affect the results and conclusions reported in the CAFA study.

We agree that the gold standard reference set can confirm predictions (i.e., count true positives) but, because it does not exhaustively represent all functions of the target sequences, we argue that the reference set cannot falsify predictions (i.e., count false positives). To illustrate this point, consider the first target in the reference dataset: CLC4E_MOUSE, which was assigned the molecular functions 'receptor activity' and 'protein binding' based on GO annotations accrued in the Swiss-Prot database [4] in the 11 months following the close of the competition (Supplementary Table 1 in [1]). InterProScan [5] also predicts a function of 'carbohydrate binding', which in CAFA would be considered a false positive because it does not appear in the gold standard set. However, this prediction is actually correct, based on experimental evidence of alpha-mannose binding [6] that is not yet recorded in Swiss-Prot GO annotations. This example is not atypical, because the Swiss-Prot database is maintained by expert curators who, owing to resource limitations, process entries according to defined priorities. Even with vastly more resources, the database would remain incomplete because most functional information has yet to be discovered through direct experiments in the first place.

To quantify the extent of spurious false positives that results from disregarding the Open World assumption, we simulated the CAFA experiment by considering a different, older set of data for which we now have the benefit of hindsight: successive releases of the UniProt-GOA database [7] dating back to 2007. Analogous to predictions submitted to CAFA, we first retrieved all gene products with computational (predicted) annotation but no experimental annotation in the 2007-01-19 release. Analogous to the way the CAFA gold standard set was built, we then established which of these gene products accumulated new experimental annotations between the 19 January 2007 and 16 January 2008 releases. For these targets, and following the CAFA protocol, we counted as false positives all electronic annotations in the 19 January 2007 release that were not confirmed by an experimental annotation in the 16 January 2008 release. However, a considerable proportion of these purported false-positive predictions were in fact confirmed by experimental annotations in subsequent UniProt-GOA releases (22 January 2012, 11 January 2011, 7 February 2012, 7 January 2013), thereby contradicting the initial assessment (Figure 1): in our analysis, ~14.7% of the predictions initially deemed as false positives were later confirmed to be correct. This is necessarily an underestimate of the error rate in the CAFA definition of false-positive predictions, and questions the ranking of methods reported by CAFA.

Corresponding author: Dessimoz, C. (c.dessimoz@ucl.ac.uk).
Keywords: gene ontology; protein function prediction; protein databases; community assessment; Open World assumption; computational methods.

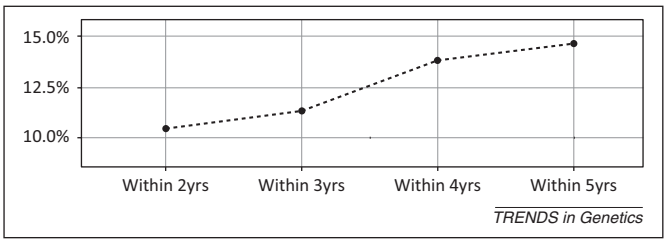

Figure 1. Proportion of 'false positives' that are spurious (i.e., predictions not confirmed in the CAFA reference set but confirmed in a later release).

To compare prediction methods meaningfully, future CAFA sequels* should consider the Open World assumption and tackle head-on the complications associated with it. For instance, explicit annotations of absence of function – identified by the keyword 'NOT' in the qualifier field of GO annotations – should be required to falsify predictions. Currently, however, only about 2500 of the 530 000 (0.48%) experimentally confirmed molecular function and biological process annotations in UniProt-GOA are negative ones – in part because they have often been perceived as less useful than their positive counterparts. But, to improve and evaluate function prediction, negative annotations are invaluable and more of them are needed [8]. Another way of addressing the problem would be to limit the scope of function prediction to specific aspects of function that can be thoroughly assessed in experiments after the submission deadline (e.g., particular enzymatic activities) [9]. For these restricted functional aspects, the more straightforward 'Closed World' assumption would apply, but the conclusions drawn might not hold in general. Neither solution constitutes a 'quick fix'. Indeed, progress in assessing protein function prediction is likely to require a substantial coordinated effort and broad support from the community. In that sense, the CAFA group is already well positioned to help drive the field forward.

Disclaimer statement

N.Š. participated in the CAFA experiment and was a co-author on the CAFA paper [1].

References

1 Radivojac, P. *et al.* (2013) A large-scale evaluation of computational protein function prediction. *Nat. Methods* 10, 221–227
2 Gene Ontology Consortium (2010) The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res.* 38, 331–335
3 Thomas, P.D. *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
4 UniProt Consortium (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* 40, 71–75
5 Hunter, S. *et al.* (2012) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.* 40, 306–312
6 Yamasaki, S. *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
7 Dimmer, E.C. *et al.* (2011) The UniProt-GO annotation database in 2011. *Nucleic Acids Res.* 40, 565–570
8 Skunca, N. *et al.* (2012) Quality of computationally inferred gene ontology annotations. *PLoS Comput. Biol.* 8, e1002533
9 Huttenhower, C. *et al.* (2009) The impact of incomplete knowledge on evaluation: an experimental benchmark for protein function prediction. *Bioinformatics* 25, 2404–2410

http://dx.doi.org/10.1016/j.tig.2013.09.005 Trends in Genetics, November 2013, Vol. 29, No. 11

* The next CAFA challenge has just been announced (http://biofunctionprediction.org/node/20).