BIOINFORMATICS MESSAGE FROM THE ISCB Vol. 30 no. 14 2014, pages 2091–2092
doi:10.1093/bioinformatics/btu117

Advance Access publication March 3, 2014

The automated function prediction SIG looks back at 2013 and prepares for 2014

Mark N. Wass[1,*], Sean D. Mooney[2], Michal Linial[3], Predrag Radivojac[4] and Iddo Friedberg[5,6]

[1]School of Biosciences, University of Kent, Canterbury, Kent CT2 7NH, UK, [2]The Buck Institute for Research on Aging, Novato, CA 94945, USA, [3]The Alexander Silberman Institute of Life Sciences, The Hebrew University of Jerusalem, 91904, Jerusalem Israel, [4]Department of Computer Science and Informatics, Indiana University, Bloomington, IN 47405, USA, [5]Department of Microbiology, Miami University and [6]Department of Computer Science and Software Engineering, Miami University, Oxford, OH 45056, USA

Contact: m.n.wass@kent.ac.uk or mark@wass.com

INTRODUCTION

The mission of the Automated Function Prediction Special Interest Group (AFP-SIG) is to coalesce the community of computational biologists, experimental biologists and biocurators who are addressing the challenge of protein function prediction, thereby sharing ideas and creating collaborations. The AFP-SIG holds annual meetings alongside the Intelligent Systems for Molecular Biology, the leading conference of the International Society for Computational Biology. The AFP–SIG also runs the ongoing Critical Assessment of Functional Annotation (CAFA) challenge (Radivojac *et al.,* 2013).

ABOUT THE CAFA CHALLENGE

The problem

There are many proteins in databases for which the sequence is known but the function is not known. The gap between what we know and what we do not know is growing. A major challenge in the field of bioinformatics is to predict the role that proteins play in biological processes and disease and the mechanism by which these functions are performed. As the community develops novel algorithms to address this task, it is important that we are able to assess how well each of these function prediction algorithms performs under certain contexts.

A solution

CAFA is a challenge designed to provide a large-scale assessment of the computational methods dedicated to predicting protein function by comparing community predictions with experimental associations that accumulate over time. Briefly, the CAFA organizers release a large number of protein sequences to the community (>100 000 in CAFA 2). The participants then predict the function of these proteins by associating them with Gene Ontology (GO) (Ashburner *et al.,* 2001) terms or (new in 2013) Human Phenotype Ontology terms (Robinson *et al.,* 2008). Following the prediction deadline, a ≥6-month interval between the submission and assessment times allows some proteins to acquire new experimental annotations. These proteins comprise the benchmark used to assess the 'blind' predictions made by the participating groups.

DISCUSSIONS IN AUTOMATED FUNCTION PREDICTION 2013

Automated Function Prediction (AFP) 2013 consisted of a series of invited talks and competitively selected presentations and posters from extended abstracts submitted by participants. Keynote speakers invited this year were Alex Bateman from the European Bioinformatics Institute; Patricia Babbitt from the University of California, San Francisco; Keith Dunker from Indiana University; and Anna Tramontano from the University of Rome, 'La Sapienza'.

Alex Bateman discussed Pfam (Punta *et al.,* 2012), a database of protein function families of which he was a founder, and UniProt (The UniProt Consortium, 2013), the leading protein sequence database. Specifically, he discussed the uses and misuses of Pfam as a touchstone for the prediction of protein function, and the convoluted relationships between evolutionary relatedness and functional similarity.

Patricia Babbitt described how her research uses protein similarity networks to investigate protein function particularly for classification in the Structure Function Linkage Database, which uses automated clustering via similarity networks to initially group enzyme superfamilies. This is followed by manual analysis of functional residues to refine the classification of subfamilies.

Keith Dunker discussed functional aspects of intrinsically disordered proteins. He highlighted that many disordered proteins currently lack functional annotation and questioned if their functions are even catalogued in ontologies such as GO. Where annotations are known, he discussed how ordered protein functions are typically related to catalysis, membrane transport and small molecule transport, whereas disordered proteins are more likely to have functions relevant to signaling, regulation, recognition and control. The basic idea being that structured proteins do things and disordered proteins regulate the things that ordered

*To whom correspondence should be addressed.

proteins do. He also discussed how disorder leads to tissue-specific gene expression and enables rewiring of protein–protein interaction networks in different cellular compartments.

Anna Tramontano discussed her experiences as an organizer of the Critical Assessment of protein Structure Prediction (CASP) (Moult *et al.,* 2013). CASP is one of the oldest critical assessment challenges in bioinformatics, established to understand and improve protein structure prediction programs. Prof. Tramontano has been prominent in this field for many years, and has also assessed method performance in several CASP meetings. She related her experience in CASP as an assessor, the person who scores the methods based on their performance using various metrics, and provided insights for future rounds of CAFA.

Predrag Radivojac and Sean Mooney discussed lessons learned from the previous CAFA challenge that took place during 2010–2011, and unveiled the upcoming CAFA 2 experiment. CAFA 2 extends on CAFA1 by adding the Cellular Component and Human Phenotype Ontologies to the set previously used for predicting function. A new challenge in CAFA 2 is the prediction of new functions for proteins that already have some (incomplete) functional annotation to assess whether partial knowledge about a protein's function can be successfully used to computationally predict the missing annotation.

Other interesting discussions during the meeting included a talk from Christos Ouzounis (University of Toronto, Canada, and the Institute of Applied Biosciences, Thessalonica, Greece) entitled 'We still haz a job: genome annotashuns', which discussed the problems associated with error propagation within databases. Using the example of a typographical error 'putaive' (instead of 'putative'), he demonstrated how this is a problem for functional annotation, as there are 94 proteins with this annotation in the NCBI protein database with many of these proteins being homologs. Rachael Huntley from European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI), UK, discussed common misconceptions associated with GO. She highlighted how GO annotations should be used and many of the useful features available including inter-ontology links and taxon-specific GO terms.

AWARDS

In 2013, awards were given for the first time for best poster and talk. Noah Youngs from New York University received the Best Talk award for his presentation: 'Negative Example Selection in Protein Function Prediction'. Noah proposed that negative annotations should be assigned to proteins by looking for GO terms that are unlikely to co-occur, and that these can be used to provide negative examples for training sets. Runner-up for best talk award was Joachim Bargsten from Wageningen University, The Netherlands, who discussed 'Integrated Network- and Sequence-based Protein Function Prediction Across the Plant Kingdom'.

Nives Skunca from Swiss Federal Institute of Technology (ETH) Zurich received the Best Poster award for her poster entitled 'Assessing protein function predictions in light of the Open World Assumption'. Her work focused on how false-positive predictions cannot be confirmed, as predictions that are assessed as false-positive results at one time point may be found to be correct later on. Runner-up for the best poster award was Romain Studer from University College London who discussed 'Identification of functional sites in protein structures by combining evolutionary and physical features'.

In summary, the AFP-SIG 2013 provided a venue for extensive discussion of function prediction ranging across new methods, assessments by CAFA, issues with such assessments and the universal resources that are widely used by the field including the UniProt and Pfam databases and GO. AFP-SIG 2014 will take place before Intelligent Systems for Molecular Biology in Boston in July 2014, and will feature the results of CAFA 2 in addition to the many areas relevant to function prediction.

ADDITIONAL INFORMATION

Join us in Boston, July 11–12, for the 2014 AFP-SIG. For more information visit http://biofunctionprediction.org.

ACKNOWLEDGMENTS

The authors acknowledge all AFP meeting participants and CAFA registrants. They gratefully acknowledge the International Society of Computational Biology for its continuous support of the AFP meeting. Special thanks to Steven Leard for extraordinary support and professionalism. The authors are also grateful for the ongoing guidance of the members of the CAFA steering committee: Patricia Babbitt, Steven Brenner, Christine Orengo and Burkhard Rost.

Funding: SDM was supported, in part, by NIH R01 LM009722 and NIH U54-HG004028 awarded by the National Institutes of Health. IF was supported, in part, by the National Science Foundation under Grant Number NSF/ABI 1146960. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or the National Institutes of Health.

Conflict of Interest: none declared.

REFERENCES

Radivojac,P. *et al.* (2013) A large-scale evaluation of computational protein function prediction. *Nat. Methods*, **10**, 221–227.

Ashburner,M. *et al.* (2001) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Robinson,P.N. *et al.* (2008) The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am. J. Hum. Genet.*, **83**, 610–615.

Punta,M. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.

The UniProt Consortium. (2013) Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res.*, **41**, D43–D47.

Moult,J. *et al.* (1995) A large-scale experiment to assess protein structure prediction methods. *Proteins*, **23**, ii–v.