CAFA: a Challenge Dedicated to Understanding the Function of Biological Macromolecules

Predrag Radivojac[1], Casey S. Greene[2], Sean D. Mooney[3], and Iddo Friedberg[4]

[1]Indiana University, Bloomingon, Indiana, USA; [2]University of Pennslylvania, Philadelphia, Pennsylvania, USA; [3]University of Washington, Seattle, Washington, USA; [4]Iowa State University, Ames, Iowa, USA

The accurate annotation of protein function is key to understanding life at the molecular level and has important biomedical and pharmaceutical implications. However, with its inherent difficulty and expense, experimental characterization of function cannot scale to accommodate the vast amount of sequence data already available. The computational annotation of protein function has therefore emerged as a problem at the forefront of computational and molecular biology. Furthermore, it presents unique machine learning challenges related to integrating noisy incomplete and multimodal data, developing models in the structured-output framework, as well evaluating methods in the open-world domain.

We propose to present the first large-scale community-wide effort whose goal is to help understand the state of affairs in computational protein function prediction and drive the field forward. We are holding a series of challenges, every three years, which we named the Critical Assessment of Functional Annotation (CAFA). CAFA was first held in 2010-2011 (CAFA1) and included 23 groups from 14 countries who entered 54 computational function prediction methods that were assessed for their accuracy [1]. CAFA2 was held in 2013-2014, and more than doubled the number of groups (56) and participating methods (126) [2]. CAFA3 is an ongoing challenge relevant for the workshop participants [3]. The challenge is associated with the Function Special Interest Group (Function-SIG) meeting that is held annually at the Intelligent Systems for Molecular Biology (ISMB) conference. Although several repetitions of the challenge would likely give accurate trajectory of the field, there are valuable lessons already learned from the first two CAFA efforts. Unexpected developments have already led to adjustments in the past and upcoming challenges.

CAFA is an atypical machine learning challenge primarily connecting three communities: experimental biologists, computational scientists and biocurators (Figure 1A). It is run as a timed-challenge (Figure 1B). At time t_0, a large number of experimentally unannotated proteins are made public by the organizers and the predictors are given several months (coordinated with educational semesters), until time t_1, to upload their predictions to the CAFA server. At time t_1 the experiment enters a waiting period of at least several months, during which the experimental annotations are allowed to accumulate in public databases. These newly accumulated annotations are collected at time t_2 and are expected to provide experimental annotations for a subset of original proteins. The performance of participating methods is then analyzed between time points t_2 and t_3 and presented to the community at time t_3. It is important to mention that unlike some machine learning challenges, CAFA organizers do not provide training data that is required to be used. CAFA, thus, evaluates a combination of biological knowledge, the ability to collect and curate training data, and the ability to develop advanced computational methodology.

We propose to discuss several of the unique challenges that characterize CAFA: (i) its scientific component, where computational methods are critical for the prioritization of biological experiments

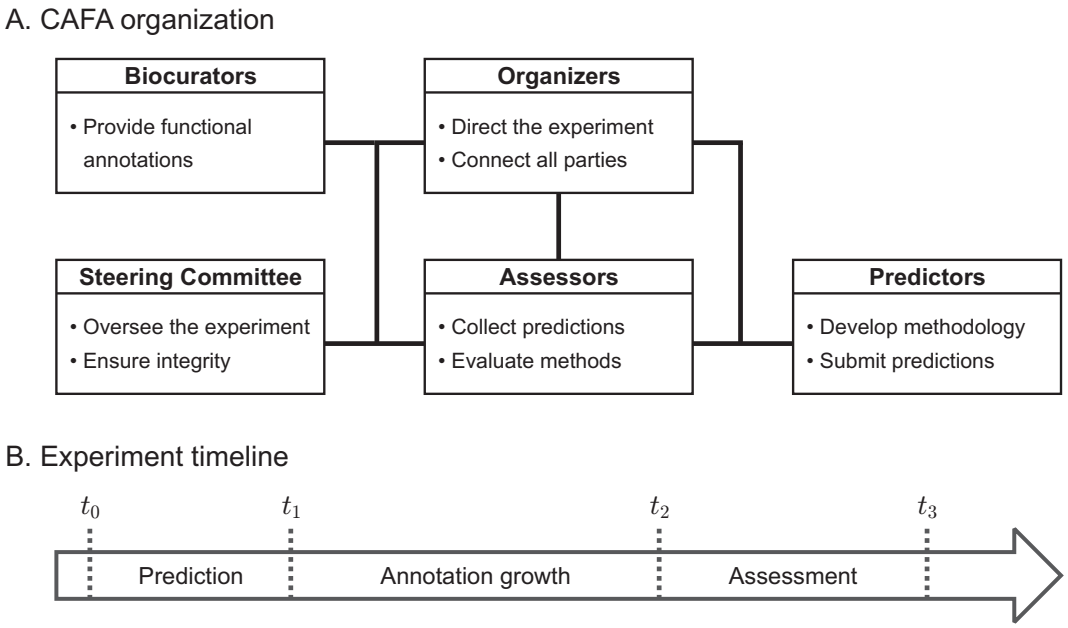

Figure 1: Organization of the CAFA experiment. (A) Five groups of participants in the experiment with their main roles. Organizers, assessors and biocurators cannot participate as predictors. (B) Timeline of the experiment.

as well as for learning trends in the molecular world; (ii) its machine learning component related to the development and evaluation in the positive-unlabeled structure-output domain (class labels correspond to the consistent subgraphs of a large directed acyclic graph such as Gene Ontology or Human Phenotype Ontology); and (iii) its educational component in computer science and biology, involving class-participation in the challenge as well as providing opportunities for students to present their results (supported so far by NSF, NIH, DOE, and IEEE). We also intend to discuss several other aspects of the challenge including the selection of ontologies that define output spaces, establishing standards of truth, and selection of assessment metrics. We will emphasize that the evaluation in this domain involves assessing similarity between graphs by weighting graph nodes by assigned (conditional) information content and is therefore an interesting problem that extends beyond molecular biology.

Finally, there are other educational opportunities that stem from the CAFA effort. These include web programming for the community site, software engineering for distributing production-grade software to participants, classroom teaching as in [1], computational applications to the domain science of molecular biology, and the development of leadership skills for students managing the challenge.

References

[1] P. Radivojac et al. A large-scale evaluation of computational protein function prediction. *Nat Methods*, 10(3):221–227, 2013.

[2] Y. Jiang et al. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biol*, 17(1):184, 2016.

[3] http://biofunctionprediction.org/.