A Domain-Based Machine Learning Approach for Function Prediction using CATH FunFams

Jonathan G. Lees*, Sayoni Das, Christine A. Orengo
Institute of Structural and Molecular Biology, UCL, Gower Street, WC1E 6BT, UK
*To whom correspondence should be addressed: jonathan.lees@ucl.ac.uk

1. INTRODUCTION

The rapid increase in the number and diversity of proteins being sequenced and the ever-widening protein function annotation gap, presents new challenges to automated function prediction methods. Here we present FunFam-ML, a domain-based machine learning approach for predicting protein function that incorporates functional family information (CATH-FunFams) and multi-domain architecture data from the CATH-Gene3D resource (1), along with other types of sequence information.

The CATH-Gene3D resource (http://www.cathdb.info) provides a comprehensive classification of structure and sequence domains into 2737 structure-based superfamilies. These domain superfamilies have been further sub-classified into functional families or FunFams by determining an optimal partitioning of the superfamily hierarchical clustering tree on the basis of specificity-determining positions (SDPs) between cluster alignments (2). It has been previously shown that this functional sub-classification helps in making precise function predictions for uncharacterized sequences (2,3).

Domains in query sequences are assigned to CATH domain superfamilies by scanning against HMM models built from all non-redundant domain structures in CATH (at 35% sequence identity), using HMMER3 (4). A recently developed method exploiting dynamic programming (CATH-Resolve-Hits) is used to determine the optimal assignment of domains based on the HMM matches to the CATH models. This has been shown to improve the performance. Following identification of a domain superfamily match, domains are assigned to FunFams using FunFHMMer (2,3) and experimental molecular function and biological process Gene Ontology (GO) (5) terms are inherited.

The FunFam-ML method uses a random-forest approach to predict GO terms for query protein sequences. The results of CAFA 2 (6) international function prediction experiment ranked FunFam-ML among one of the top methods for both ‘molecular function’ and second best for ‘biological process’ function prediction (see Figure 1). We have recently improved our predictor by adding in extra features (such as signal peptide information, disorder information etc.) and extra homology information exploiting functional links between genes. Furthermore, improvements in the underlying domain assignment pipeline, have given improved coverage in the predictions. Initial benchmarks suggest these changes have led to an improvement in the F-max (5-10%, depending on the dataset).

Preliminary analyses incorporating network data are underway and appear to be further improving performance and giving higher confidence for more specific GO terms, which are likely to be of greater use to experimentalists.

2. FIGURE

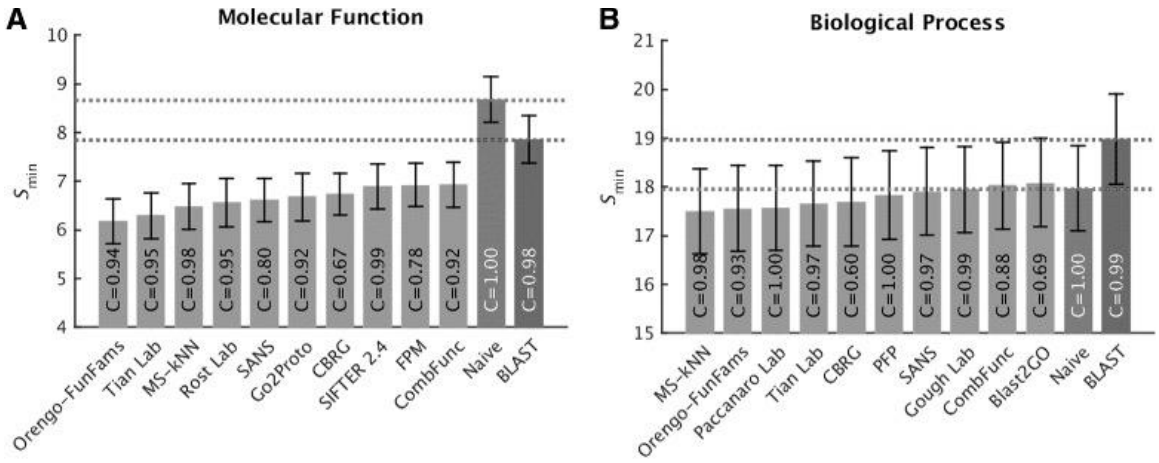

Figure 1: The top 10 function prediction methods in CAFA 2 for predicting GO terms in the Molecular Function Ontology (MFO) (A) and the Biological Process Ontology (BPO) (B) for all targets in the no-knowledge benchmark set in the full evaluation mode, using the minimum semantic distance, S_{min}.The FunFam-ML method described here has been referred to as Orengo-FunFams in this figure. This figure has been adapted from (6).

3. REFERENCES

1. Dawson N.L., Lewis T.E., Das S., Lees J.G., Lee D., Ashford P., Orengo C.A., and Sillitoe I. 2017. CATH: an expanded resource to predict protein function through structure and sequence. Nucleic Acids Research 45(D1):D289-D295.

2. Das, S., Lee, D., Sillitoe, I., Dawson, N. L., Lees, J. G., & Orengo, C. A. 2015. Functional classification of CATH superfamilies: a domain-based approach for protein function annotation. Bioinformatics 31(21):3460-3467.

3. Das, S., Sillitoe, I., Lee, D., Lees, J.G., Dawson, N.L., Ward, J. and Orengo, C.A. 2015. CATH FunFHMMer web server: protein functional annotations using functional family assignments. Nucleic acids research 43(W1):W148-W153.

4. Eddy, S. R. 2009. A new generation of homology search tools based on probabilistic inference. Genome Informatics 23:205-211.

5. Ashburner M., Ball C. A., Blake J. A., Botstein D., Butler H., Cherry J. M., Davis A. P., Dolinski K., Dwight S. S., Eppig J. T. et al. 2000. Gene Ontology: tool for the unification of biology. Nature Genetics 25(1):25-29.

6. Jiang Y., Oron, R. T., Clark T. W., Bankapur R. A., D'Andrea D., Lepore R., Funk S. C. et al. 2016. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. Genome Biology 17(1):1-19.