Region-specific Function Prediction: automatically inferring function labels for protein regions

Da Chen Emily Koo, Noah Youngs, Richard Bonneau*
NYU Departments of Biology and CS, NY, NY, 10003
Simons Center for Computational Biology, Flatiron Institute,
162 5th Avenue, NYC, NY, 10010, USA
*To whom correspondence should be addressed: rb133@nyu.edu

1. INTRODUCTION

Due to the nature of experimental annotation, most of the protein function prediction methodologies operate at the protein-level, where functions are assigned to and transferred between full-length proteins based on shared sequence and structure features. However, most proteins function by specifically interacting with other proteins or other molecules and a majority of functions can be localized to specific regions. These regions include active and binding sites on the proteins and thus not all functions associated to the same protein should be shared at the region-specific level. Furthermore, proteins and domains are usually grouped into families based on sequence and fold similarity, but gene duplication can often lead to functional divergence that may not be detectable at that level (1). Most domain-centric function prediction methods treat the domains as a property of the protein rather than as its own functional entity and depend heavily on accurate domain family assignments to infer relationships between domains and functions. In addition, regions that are unassigned are left out of functional evaluation.

We will present an explicit region-specific function prediction methodology to automatically infer function labels of specific protein regions based on functional annotations available at the protein-level and the features annotated directly to the regions themselves, rather than solely as a member of a domain family. This is based on transfer learning methods for relating sentence/location and document level annotations (2) with additional model components to account for known domain labels from manually curated databases. Together with improved methods of generating predicted negative annotations (3,4), we train the model using features and functional annotations both from single species and from multiple different species. This is based on the assumption that combining data from different species will improve the predictive capability of the approach by increasing the occurrence of positive examples, which is especially important for predicting labels of rare and less studied functions. The results are benchmarked using temporal hold-out for mouse, human and yeast. We report the predictive performance when evaluated at the whole protein level and also at the region-specific level using manually curated domain annotations from InterPro and other site-specific data such as DNA binding sites. In the future, we want to simultaneously train this model with protein-protein networks constructed from large-scale experimental data such as gene expression and PPI and structural features (5) to increase the power of both protein-level and region-level predictions.

2. FIGURES

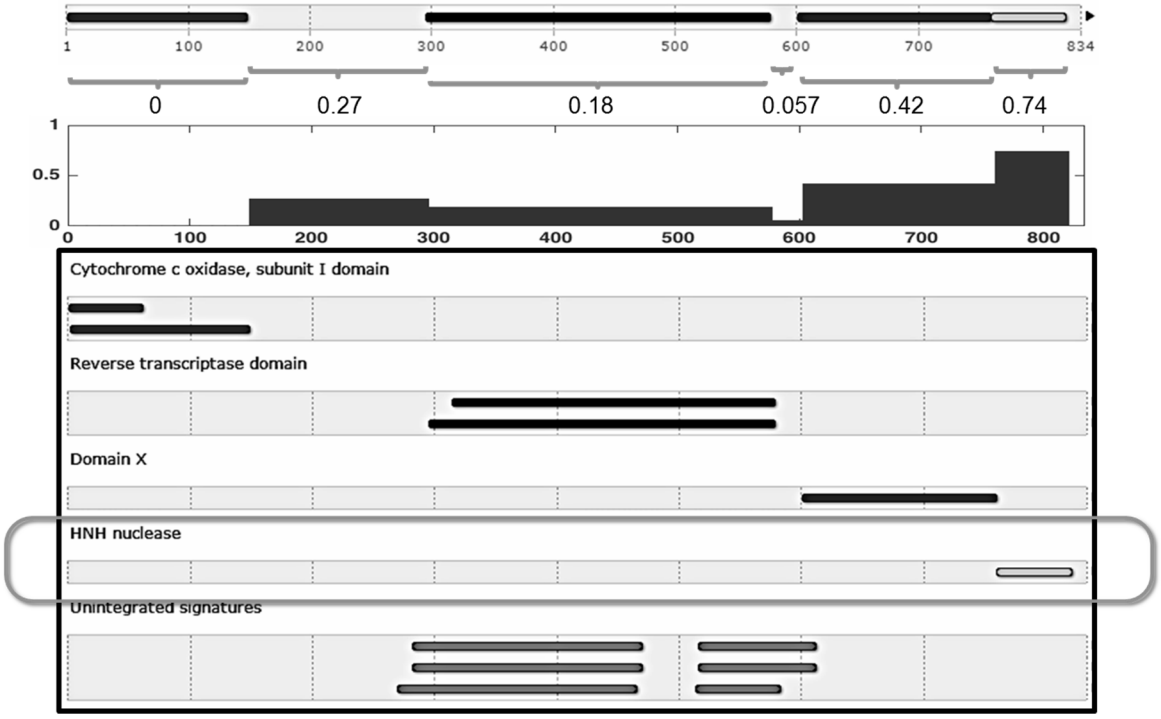

Figure 1: Predicted probabilistic scores along P03875 for GO:0004518 nuclease activity http://www.ebi.ac.uk/interpro/protein/P03875

3. REFERENCES

1. Shulman-Peleg, A., Nussinov, R. & Wolfson, H. J. 2004. Recognition of functional sites in protein structures. *J. Mol. Biol.* **339,** 607–633.

2. Kotzias, D., Denil, M., de Freitas, N. & Smyth, P. 2015. From Group to Individual Labels Using Deep Features. in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15* 597–606

3. Youngs, N., Penfold-Brown, D., Bonneau, R., & Shasha, D. 2014. Negative Example Selection for Protein Function Prediction: The NoGO Database. *PLoS Computational Biology*, 10(6).

4. Youngs N, Penfold-Brown D, Drew K, Shasha D, Bonneau R. 2013. Parametric Bayesian priors and better choice of negative examples improve protein function prediction. *Bioinformatics*. 1;29(9):1190-8.

5. Drew K, Winters P, Viktors Berstis, Keith Uplinger, Jonathan Armstrong, Michael Riffle, Erik Schweighofer, Bill Bovermann, David R. Goodlett, Trisha N. Davis, Dennis Shasha [6], Lars Malmström, Richard Bonneau. 2011. The proteome folding project: proteome-scale prediction of structure and function. *Genome Res*. 21(11):1981-94.