Automating Genomic Context Analysis with a Probabilistic Model of Protein Function and Relatedness

Jeffrey M. Yunes, Patricia C. Babbitt*
UC Berkeley - UCSF Graduate Program in Bioengineering, 1700 4th St, San Francisco, CA 94158, United States
Department of Bioengineering and Therapeutic Sciences, 1700 4th St, San Francisco, CA 94158, United States
*To whom correspondence should be addressed: babbitt@cgl.ucsf.edu

1. INTRODUCTION

Prediction of protein function continues to be an active area of research. While the performance of automated function prediction methods continues to improve, assessments of their performance have begun to identify useful sources of data, better delineate the characteristics of the problem, and identify limitations with current methods (1). Here, we demonstrate a method inspired by the practice of using genomic context to suggest function of a query protein. The method uses a probabilistic graphical model (PGM) that carefully incorporates rare, valuable experimental data, with copious amounts of noisy higher order data.

2. METHODS

First, we build a protein network with edges of sequence similarity (SS) and functional associations (Figure left). This network is quickly constructed by broadly BLASTing (2) sequence queries to collect homologs, querying STRING (3) to get functionally associated proteins and their corresponding edges, and then using DIAMOND (4) to fill in the all-by-all sequence similarity edges of the network.

Second, we construct a tractable probabilistic graphical model based on this network (Figure right). The network is first converted to a minimum spanning tree, directed, and pruned to query proteins or proteins with experimental annotation. When the edge between two proteins is due to sequence similarity, we associate the corresponding molecular function (MF) terms from Gene Ontology (5). When the edge between two proteins is due to functional association (i.e., from STRING), we associate their corresponding biological process (BP) terms. Finally, we add edges between MF terms and BP terms within a single protein, allowing belief to flow between the two aspects of protein function, and reducing the influence of assignments with unlikely combinations of terms.

Third, we learn the parameters. Since the model is directed, the parameters have a global probabilistic interpretation: each variable in the probabilistic model adds a factor representing the conditional probability of that variable given its parents. We can then learn these parameters from looking at all pairs of neighboring proteins that have experimental annotations.

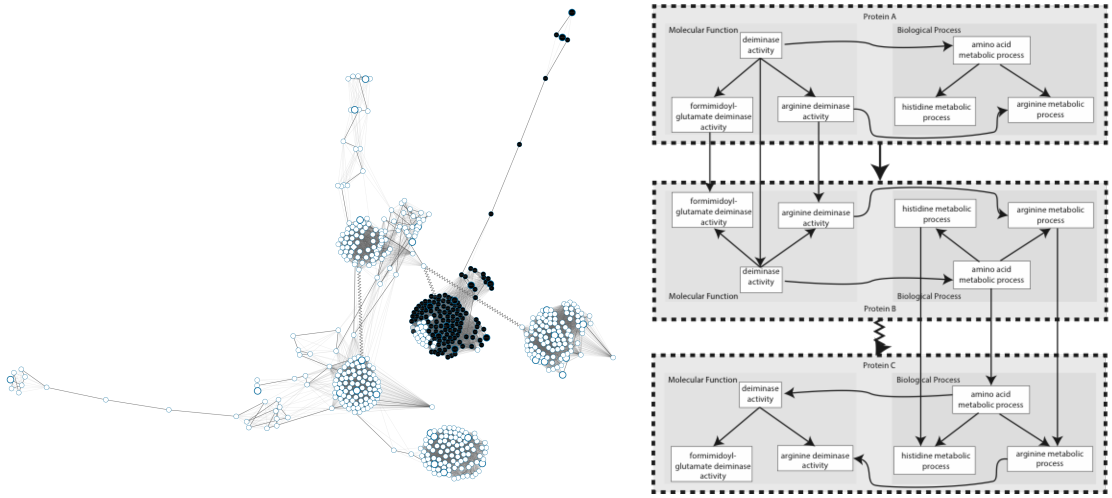
Finally, we perform inference with belief propagation, summarize the data, and evaluate the predictions. Proteins used for evaluation were the 1000 proteins most recently assigned an experimental catalytic annotation. We used performance metrics that are revealing and critical, including normalized and weighted misinformation and remaining uncertainty (6).

2. RESULTS

PGMs are well suited to the problem of protein function prediction for several reasons. First of all, propagation of functional information introduces random error. Second, it is also valuable to provide ranked probabilities and identify uncertainty in the model. Third, PGMs can naturally take advantage of the large amount of unlabeled and partially unlabeled protein data (semi-supervised learning), where other methods cannot handle such high degrees of sparseness. Fourth, PGMs can seamlessly do structured output prediction, and therefore produce predictions that are consistent with, and take advantage of, the structure of the Gene Ontology.

We found that our model, which models the MF and BP of each protein, and only propagates belief where appropriate, outperforms methods that only model a single aspect of protein function (7), and propagates belief regardless of type of similarity that the edge represents. As a result, this model uses functional associations, such as genomic context, to inform a prediction for MF, even in the absence of any characterized homologs.

3. FIGURES


Protein A
Molecular Function
Biological Process
deiminase activity
amino acid metabolic process
formimidoyl-glutamate deiminase activity
arginine deiminase activity
histidine metabolic process
arginine metabolic process
Protein B
Protein C

Left: Sequence similarity - functional association protein network for query Q818C8, shared according to probability of GTPase activity. Each node is a protein. Straight edges represent sequence similarity. Zigzag edges represent functional association, according to STRING. Edges not transparent are in the MST. yFiles Organic Layout. **Right: PGM for three proteins** one pair of which are close homologs, another pair of which are functionally associated.

4. REFERENCES

1. Jiang, Y. *et al*. 2016. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biology.* 17.

2. Altschul, S. F. *et al*. 1990. Basic Local Alignment Search Tool. *J. Mol. …* **215,** 403–410.

3. Szklarczyk, D. *et al.* 2015. STRING v10: Protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43,** D447–D452.

4. Buchfink, B. *et al*. 2015. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12,** 59–60.

5. Ashburner, M. *et al.* 2000. Gene Ontology: Tool for The Unification of Biology. *Nat. Genet.* **25,** 25–29.

6. Clark, W. T. & Radivojac, P. 2013. Information-theoretic evaluation of predicted ontological annotations. *Bioinformatics* **29,** 1–9.

7. Mitrofanova, A. *et al*. 2011. Prediction of protein functions with gene ontology and interspecies protein homology data. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* **8,** 775–784.