Reasoning on Gene Ontology Networks Predicts Novel Protein Annotations

Ilya Novikov[1], Angela Wilkins[2], Olivier Lichtarge[1,2,*]
[1]Department of Biochemistry and Molecular Biology, [2]Department of Molecular and Human Genetics, Baylor College of Medicine, 1 Baylor Plaza, Houston, TX, 77030

*To whom correspondence should be addressed: lichtarge@bcm.edu

1. INTRODUCTION

Careful examination of the existing evidence underpins formulation of novel hypotheses and experimental design. However, each year the state of the ground truth becomes more difficult to approximate due to the rapid proliferation of primary literature and associated databases. With over a million new articles added to MEDLINE annually, mastering even the subfield literature has become difficult. To address this problem, algorithmic approaches have been developed to automatically collect, integrate, and even reason on data. For instance, in 2014 the KnIT method condensed the entirety of the human kinome literature into a single graph, and then reasoned on it to predict novel kinases that target a critical cancer suppressor protein (1). While these methods are powerful, they rely on primary literature as their main source of data, and therefore ignore structured data contained in the multitude of domain databases.

One such database is the Gene Ontology (GO). GO provides functional annotations for proteins by assigning them labels, referred to as GO terms. GO terms describe a protein's function, role, and location within the cell. Compared to data automatically extracted from primary sources, data in GO has two advantages. First, principal label assignment in GO is carried out by human curators, so error rate is low. Second, GO terms are related to each other through a hierarchical ontology that reflects the specificity and biological context of each term. The term ontology serves as an additional layer of information provided by the domain experts. Clearly, the Gene Ontology contains a trove of curated, context-sensitive biological data.

Here we present two related approaches for extracting, modeling, and reasoning on GO data. First, in a low-level approach, we extract the term annotations as a term-protein matrix, and use Non-Negative Matrix Factorization (NNMF) to suggest novel term-protein annotations. We show that this representation is stable in ten-fold cross-validation, and that it has predictive power in time-stamped trials. In the second approach, we use Resnik semantic similarity to build a pairwise distance matrix for annotated proteins, and then convert the distance matrix into a protein-protein network. To reason on this representation, we label protein nodes with known functions, and then use graph diffusion to propagate the labels to the unlabeled nodes. We show that this graph representation is stable and self-consistent, and then use time-stamp validation to show that it, too, has predictive power. Together these data show that GO can be readily distilled into a form amenable to automated reasoning, and that it has predictive power.

2. METHODOLOGY AND RESULTS

In our first approach, we extract all protein-term pairs from an organism's annotation corpus in the form of matrix A, where $A(i,j) = 1$ if protein i is annotated with term j, and 0 otherwise (note that we explicitly propagate terms to the root during matrix construction). To predict novel protein-term pairs from A, we employ NNMF (2), a machine learning method for matrix completion. NNMF factorizes the original matrix A, and then reconstructs it as an approximation

A', where all of the zero pairs now have a weight assigned to them. The weights reflect the likelihood that the protein-term pairs actually exist in the original matrix *A*.

To test this model, we constructed protein-term matrices for 19 of the species in the CAFA challenge. To show that the matrices are self-consistent, we iteratively left out 10% of the non-zero pairs, and found that NNMF recovers them with AUC > 0.85 in 95% of the species (mean AUC = 0.88). Next, we tested the predictive power of this approach in a time-stamp trial. We constructed the matrices using annotations from August 2016, and then applied NNMF to predict novel protein-term pairs. Comparing the predictions to the February 2017 annotation corpus, we found that the predictions were were enriched with actual annotations added after August 2016 (mean AUC = 0.89 in the 19 species). These data show that reasoning on even the relatively naïve representation of the Gene Ontology can yield novel functional predictions. (We submitted predictions, based on the January 2017 GO NNMF to the CAFA challenge.)

Next, we developed a higher-level GO term similarity model that abstracts not only the protein-term annotations, but also the term-term relationships. To build this model for an organism, we measure Resnik GO term similarity (3) for every possible pair of proteins, and then threshold the similarity scores to create a sparse protein-protein network. We discovered that in these networks proteins with similar GO terms, and thus similar functions, preferentially connect to each other. We hypothesized that this property can be used for functional inference, and tested it with a time-stamped trial. Using GO from August 2016, we constructed a proteomic GO term similarity network for *Bacillus subtilis*. In the network, we labeled all proteins associated with a particular GO term, and then used graph information diffusion (4) to propagate the label. We predicted that unlabeled nodes with the highest post-diffusion label content should be assigned the GO term. Comparing predictions to the actual annotations made after August 2016, we found that graph diffusion predicted novel annotations for 58 GO terms with mean AUC = 0.90. These data show that GO term similarity networks have predictive power.

3. CONCLUSION

We have shown that data contained in the Gene Ontology can be compressed into self-consistent representations that retain high information density, and can be reasoned upon to discover novel functional annotations. These approaches generalize to any entity annotated with GO terms, and can be used to create networks that integrate genes, proteins, RNAs, drugs, and diseases. Furthermore, predictions produced by reasoning on GO can be directly combined with other inference methods.

4. REFERENCES

1. Spangler S and Wilkins A. 2014. Automated Hypothesis Generation Based on Mining Scientific Literature. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*: 1877-1886.
2. Lee D and Seung S. 2001. Algorithms for Non-negative Matrix Factorization.*Adnvaces in Neural Information Processing Systems* 13: 556-562.
3. Zhou D, Bousquet O, Weston J, and Scholkopf B. 2004. Learning with local and global consistency. *Adnvaces in Neural Information Processing Systems* 16: 321-328.
4. Resnik P. 1999. Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research* 11: 95-130.