Label Propagation and Positive-Unlabeled Learning for Protein Function Prediction

Noah Youngs1, Duncan Penfold-Brown3, Kevin Drew5, Emily Koo2, Dennis Shasha4, Richard Bonneau1,4

1. Department of Computer Science, New York University, New York, USA
2. Department of Biology, New York University, New York, USA
3. Social Media and Political Participation Lab, New York University, New York, USA
4. Center for Genomics and Systems Biology, Department of Biology, New York University, New York, USA
5. Department of Biology, University of Texas at Austin, Austin, USA

Protein Function Prediction

- Large portions of many genomes remain without experimentally-evaluated functional annotations.
- Accurate computational predictions will save time and money by guiding experimentalists.
- Predictions can also provide insights into groups of genes that do not currently have annotations.
- There is an ever-increasing amount of different types of data that can be leveraged to provide predictions.

Label Propagation Procedure: Inputs and Outputs

Tertiary Structure as a Similarity Feature

Domain Tertiary Structure to Similarity Score via Structural Alignment

Sequence to Structure  Domain Similarity to Protein Similarity

Positive-Unlabeled Learning

- PU-learning operates in scenarios where labeling negative examples is prohibitive.
- One branch of algorithms works by first finding a set of likely negative examples.
- These negative examples are used as normal in a two-class machine learning algorithm
- Scores outputted by the algorithm can also be used as priors for function prediction learning

Results and Publications

- Algorithmic improvements for Label Propagation appear in Bioinformatics:
- Novel PU-Learning methods appear in PLOS Computation Biology:
- Negative examples are available for download: http://bonneaulab.bio.nyu.edu/noGD.html
- Function predictions on hundreds of thousands of sequences in 27 species were performed as part of the CAFA challenge. Results pending in July.
- Further work incorporating larger amounts of structural data and multi-species prediction via homology links is pending as well.