Supplementary Materials of “AptRank: an adaptive PageRank model for protein function prediction on bi-relational graphs”

Biaobin Jiang\(^1\), Kyle Kloster\(^2\), David F. Gleich\(^3\), Michael Gribskov\(^1,3,\)*

December 25, 2016

1 Department of Biological Sciences, Purdue University, West Lafayette, IN 47907, USA.
2 Department of Mathematics, Purdue University, West Lafayette, IN 47907, USA.
3 Department of Computer Science, Purdue University, West Lafayette, IN 47907, USA.
* Corresponding Author: gribskov@purdue.edu

1 Supplementary Experiments

1.1 Optimal Set of BirgRank Parameters

To investigate how the four key parameters of BirgRank affect its prediction performance, we set each parameter as 0.1, 0.3, 0.5, 0.7 and 0.9, respectively, and test the prediction of protein functions using the yeast dataset with 50\% of data as training set and the other 50\% as testing set (missing function prediction, see the main text). As we can see in Supplementary Figure 1, different settings of these four parameters did not yield significant differences in performance. Therefore, we empirically set all the four parameters as 0.5 for all subsequent experiments in this study.

Supplementary Figure 1: BirgRank Parametric effect on AUROC

1.2 Effect of Random Ontological Hierarchy

Incorporating the GO functional hierarchy into protein function prediction is a main contribution of this study. To study how significantly the functional hierarchy improve the prediction, we replaced the matrix of hierarchy with a random graph, a hierarchy with shuffled labels, and an identity matrix, respectively. In Supplementary Figure 2, we show that these three replacements of the true hierarchy yield significant reduction in AUROCs in predicting protein functions using the yeast dataset with half data as training and the other half as testing (missing function prediction, see the main text). It demonstrates that the functional hierarchy is able to improve the prediction of protein functions through associating those non-independent labels. Our models provide a novel framework to the classification with non-independent labels.
Supplementary Figure 2: Effect of Randomized Hierarchy. The data of true hierarchy are directly obtained from Figure 3 (A) and (B).

1.3 Prediction of Cellular Components and Molecular Functions

Beside Biological Process (BP) terms, we also performed prediction of Cellular Component (CC) and Molecular Function (MF) terms using the de novo function prediction strategy (see the main text) and the human data collected in 2015. The AUROCs were calculated for different protein categories based on the number of annotated functional terms. We divided all the proteins into 5 categories: the proteins with 1-10, 11-30, 31-100, 101-300, and 300+ annotated terms. As shown in Supplementary Figure 3, the prediction performances of the seven methods in predicting CC and MF are similar with those in predicting BP (Figure 4). Our models, especially BirgRank, still slightly outperforms the runner-up, GeneMANIA-SW (GM-SW), in all the three categories of GO.
2 Comparison of BirgRank and AptRank with Other Methods

In this Section, we investigate the theoretical similarities and differences of our methods and the other four previous methods used in the evaluation.

The linear system of BirgRank in Equation (2) can be expanded into

\[
\begin{cases}
(I - \alpha \tilde{G})X_G = (1 - \alpha)I \\
\alpha R_T^T X_G = (I - \alpha \tilde{H})X_H,
\end{cases}
\]

(1)

where \( \tilde{G}, \tilde{R}_T, \) and \( \tilde{H} = \tilde{H} \) denote the submatrices of the column-stochastic matrix in Equation (2). By solving Equations (1) for \( X_H \), we get

\[
X_H = \alpha(1 - \alpha)(I - \alpha \tilde{H})^{-1} \tilde{R}_T^T (I - \alpha \tilde{G})^{-1}.
\]

(2)

In contrast, ProteinRank [Freschi 2007] uses only the protein-protein association network \( G \) as a one-layer network model — and does not directly take into consideration the functional hierarchy \( H \) — and then computes PageRank using \( R_T \) as the personalization vectors (matrix). ProteinRank constructs a regression model and solves the linear system

\[
X_{\text{ProteinRank}} = (1 - \alpha)(I - \alpha G)^{-1} R_T,
\]

(3)

which can cause poor prediction quality due to the assumption of independence between functions (see Section 3). Our method BirgRank is closely related to ProteinRank: if we plug \( H = I \) into Equation (2), then the resulting BirgRank solution differs from the ProteinRank solution (Equation (3)) only by a scalar coefficient and a slightly different normalization of \( G \).

Similar to ProteinRank, GeneMANIA [Mostafavi et al. 2008] models protein function prediction as a multiclass-multilabel classification problem by integrating multiple heterogeneous network datasets and then using the Label Propagation algorithm [Zhou et al. 2004] as

\[
X_{\text{GeneMANIA}} = (I - L)^{-1} R^T,
\]

(4)

where \( L = D - W \) is the Laplacian matrix, \( W \) is a weighted sum of multiple kernel matrices from heterogeneous network datasets, and \( D \) is a diagonal matrix with \( D_{ij} = \sum_j W_{ij} \). Additionally, GeneMANIA extends the binary matrix \( R_T^T \) to \( R^T \) by introducing negative samples in which \( R_{i,j} = -1 \) if protein \( i \) is known not to have function \( j \). The developers of GeneMANIA further accelerated their algorithm by introducing Simultaneous Weights (hereafter GeneMANIA-SW) [Mostafavi and Morris 2010].

Yu et al. [2013] proposed the Transductive Multilabel Classifier (TMC) by directly applying a Bipartite graph model used in image annotation [Wang et al. 2011] to protein function prediction, without consideration of the functional hierarchy. Instead, they use the cosine similarity of functional annotations to construct a function-function similarity matrix to replace \( H \). The key difference between TMC and BirgRank is that TMC allows information to diffuse from functional terms to proteins, but not proteins to functional terms, as in BirgRank.

Wang et al. [2015] proposed clusDCA by extending their original Diffusion Component Analysis (DCA) method [Cho et al. 2015]. The clusDCA algorithm first uses PageRank to smooth both of the graphs, denoted as \( G \) and \( H \) in this study. Next, it computes Singular Value Decomposition (SVD) for the two smoothed matrices for low-rank matrix approximations. Finally, it attempts to find the optimal projection between the two low-rank matrices.

References


