Published in final edited form as:

Res Comput Mol Biol. 2015 April ; 9029: 62–64. doi:10.1007/978-3-319-16706-0_9.

Diffusion Component Analysis: Unraveling Functional Topology in Biological Networks

Hyunghoon Cho¹, Bonnie Berger^{1,2}, and Jian Peng^{1,2,3}

¹Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA, USA

²Department of Mathematics, MIT, Cambridge, MA, USA

³Department of Computer Science, University of Illinois at Urbana-Champaign, Champaign, IL, USA

1 Introduction

Complex biological systems have been successfully modeled by biochemical and genetic interaction networks, typically gathered from high-throughput (HTP) data. These networks can be used to infer functional relationships between genes or proteins. Using the intuition that the topological role of a gene in a network relates to its biological function, local or diffusion-based “guilt-by-association” and graph-theoretic methods have had success in inferring gene functions [1, 2, 3]. Here we seek to improve function prediction by integrating diffusion-based methods with a novel dimensionality reduction technique to overcome the incomplete and noisy nature of network data.

A type of diffusion algorithm, also known as random walk with restart (RWR), has been extensively studied in the context of biological networks and effectively applied to protein function prediction (e.g., [1]). The key idea is to propagate information along the network, in order to exploit both direct and indirect linkages between genes. Typically, a distribution of topological similarity is computed for each gene, in relation to other genes in the network, so that researchers can select the most related genes in the resulting distribution or, rather, select genes that share the most similar distributions. Though successful, these approaches are susceptible to noise in the input networks due to the high dimensionality of the computed distributions.

2 Methods

We propose Diffusion Component Analysis (DCA), a novel analytical framework that combines diffusion-based methods and sophisticated dimensionality reduction to better extract topological network information in order to facilitate more accurate functional annotation of genes or proteins. The key idea behind DCA is to obtain informative, but low-dimensional features, which better encode the inherent topological properties of each node in the network. We first run a diffusion algorithm on a molecular network to obtain a distribution for each node that captures its relation to all other nodes in the network. We then

approximate each of these distributions by constructing a multinomial logistic model, parameterized by low-dimensional feature vector(s), for each node. Feature vectors of all nodes are jointly learned by minimizing the Kullback-Leibler (KL) divergence (relative entropy) between the diffusion and parameterized-multinomial logistic distributions. A key differentiating factor of our novel dimensionality reduction from a more conventional approach, such as Principal Component Analysis (PCA), is the use of multinomial logistic models, which more naturally explain the input probability distributions from the diffusion. Moreover, DCA can be naturally extended to integrate multiple heterogeneous networks by performing diffusion on separate networks and jointly optimizing feature vectors. Given the low-dimensional vector representations of nodes, k-nearest neighbor (kNN) voting schemes or support vector machines (SVM) can be used for function prediction.

3 Results

We evaluated the ability of our DCA framework to uncover functional relationships in the interactome of yeast. By combining noise reduction via dimensionality reduction, improved integration of multiple heterogeneous networks (e.g., physical interaction, conserved co-expression), and the use of support vector machines, our DCA framework is able to achieve 71.29% accuracy with five-fold cross-validation on the STRING networks with third level functional annotations from MIPS, which is remarkably 12.31% higher than the previous state-of-the-art diffusion state distance (DSD) [1] method (Figure 1). We also observe improved performance over DSD in a different yeast PPI network, constructed from only physical interactions in the BioGRID database. In addition, we found that conventional approaches to dimensionality reduction, such as principal component analysis or non-negative matrix factorization, fail to achieve similar performance improvements. Our results demonstrate the potential of low-dimensional feature vectors learned by DCA to be plugged into other existing machine learning algorithms to decipher functional properties of and obtain novel insights into interactomes.

References

1. Cao M, et al. New directions for diffusion-based network prediction of protein function: incorporating pathways with confidence. *Bioinformatics*. 2014; 30(12):i219–i227. [PubMed: 24931987]
2. Mostafavi S, et al. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol*. 2008; 9:S1–S4.
3. Milenkovic T, Pržulj N. Uncovering biological network function via graphlet degree signatures. *Cancer Informatics*. 2008; 6:257. [PubMed: 19259413]

Fig. 1. Protein function prediction performance on yeast STRING networks in terms of both accuracy and F1 score—the harmonic mean of both precision and recall—with different levels of functional categories from MIPS

Neighbor majority vote (NMV), Diffusion state distance (DSD), DCA with kNN (DCA), DCA combined with novel network integration with kNN (DCAi) or SVM (DCAi-SVM).