Network-Based Prediction of Polygenic Disease Genes Involved in Cell Motility

Extended Abstract

Miriam Bern*
Biology Department
Reed College
Portland, Oregon
mirbern@reed.edu

Alexander King*
Biology Department
Reed College
Portland, Oregon
aleking@reed.edu

Derek A. Applewhite
Biology Department
Reed College
Portland, Oregon
applewhd@reed.edu

Anna Ritz
Biology Department
Reed College
Portland, Oregon
aritz@reed.edu

CCS CONCEPTS

• **Applied computing → Biological networks**; **Systems biology**; • **Mathematics of computing → Graph algorithms**; • **Computing methodologies → Semi-supervised learning settings**;

KEYWORDS

Semi-supervised learning, functional interaction network, schizophrenia, autism, cell motility

ACM Reference Format:
Miriam Bern, Alexander King, Derek A. Applewhite, and Anna Ritz. 2018. Network-Based Prediction of Polygenic Disease Genes Involved in Cell Motility: Extended Abstract. In *ACM-BCB'18: 9th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, August 29-September 1, 2018, Washington, DC, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3233547.3233697

known disease genes and cell motility genes in the network. Top schizophrenia candidates include many Protein Phosphatase 1 subunits and Lysyl Oxidase, which are promising genes for follow-up experimental validation. Candidate genes predicted by our method suggest testable hypotheses about these genes' role in cell motility regulation, offering a framework for generating predictions for experimental validation.

1 INTRODUCTION

Schizophrenia and autism are examples of polygenic diseases caused by a multitude of genetic variants. Recently, both diseases have been associated with disrupted neuron motility and migration patterns, suggesting that aberrant cell motility is a phenotype for these neurological diseases [2, 8]. Abnormal neuronal development is central to both schizophrenia and autism, which critically implicates these cell motility perturbations in the disease mechanisms. However, despite the genetic characterization of these diseases by large-scale genome-wide association studies, extracting causality for symptoms and pathophysiology from these data remains challenging due to the large number of genes implicated and the additive effect the mutations have on the cellular processes [7].

We present a network-based machine learning approach to identify genes implicated in both a disease of interest (e.g., schizophrenia or autism) and a disease phenotype (e.g., aberrant cell motility). We use a brain-specific functional interaction network to identify which genes are most centrally implicated in a polygenic disease based on functional similarity. Our algorithm identifies genes that are near

*These authors contributed equally to this work.

ACM-BCB'18, August 29-September 1, 2018, Washington, DC, USA

ACM ISBN 978-1-4503-5794-4/18/08.
https://doi.org/10.1145/3233547.3233697

2 METHODS

Given a functional interaction network represented as a weighted, undirected graph $G = (V, E)$, a set of curated positive nodes $C \subseteq V$, and a set of curated negative nodes $\overline{C} \subseteq V$, we model a random Gaussian field on G given the labeled nodes [9]. Let $f : V \mapsto [0, 1]$ be a function where $f(v) = 1$ if $v \in C$, $f(v) = 0$ if $u \in \overline{C}$, and $f(v)$ over unlabeled nodes is "smooth" with respect to the topology of G. That is, we wish to choose values for the unlabeled nodes that minimizes

$$\min_f \frac{1}{2} \sum_{(u,v)\in E} w_{uv}(f(u) - f(v))^2,$$

where w_{uv} is the weight of edge (u, v). This equation can be calculated efficiently using an iterative method that is known to converge, and has been implemented in a method called SinkSource that has been applied to molecular interaction networks [6]. Because diseases like schizophrenia and autism are based on multiple mutations rather than a single mutation, our method builds on SinkSource and corrects for low-degree genes connected to labeled nodes that are scored disproportionately high or low. Our method (a) adds a user-defined λ-weighted edge to all nodes that connects to a global negatively-labeled "sink" and (b) partitions the positives and negatives across k "layers," where each layer contributes to the node's final score. We use cross validation to compare our method with different values of k and λ.

Our problem formulation involves two sets of positively-labeled nodes (disease genes and cell motility genes). We run the machine learning method twice: once for the disease $\mathcal{D}$ (e.g., schizophrenia or autism) to get $f_{\mathcal{D}}$ and once for the biological process $\mathcal{P}$ (e.g., cell motility) to get $f_{\mathcal{P}}$. In both cases, we use the disease negatives. We rescale the functions so the largest value is 1 and define a combined score $g(v) = f_{\mathcal{D}}(v)f_{\mathcal{P}}(v)$. Prioritizing nodes based on this score implies that high-scoring nodes v must have large $f_{\mathcal{D}}(v)$ and $f_{\mathcal{P}}(v)$.

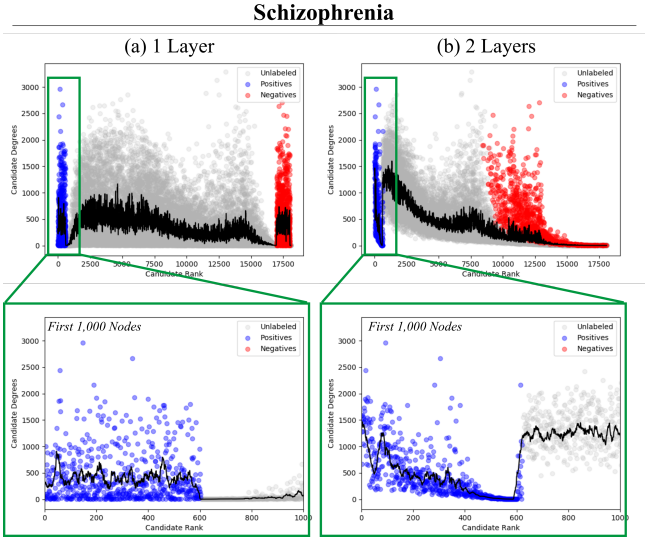

Figure 1: (a) One-layer and (b) two-layer node rankings (x-axis) by node degree (y-axis) for schizophrenia positives. Nodes are colored as unlabeled (gray), positives (blue), or negatives (red). Black line denotes moving average (15 nodes), and bottom panel shows first 1,000 nodes.

3 RESULTS AND DISCUSSION

We used a brain-specific functional interaction network from HumanBase [4], compiled three sets of positive genes from existing literature (702 schizophrenia-associated genes, 594 autism-associated genes, and 542 cell motility-associated genes) and compiled one set of 1,189 genes that were likely associated with non-neurological diseases (the negative set from [5]). We show that our method improves over SinkSource and a positive-only version of SinkSource in k-fold cross validation (data not shown).

We found that the top unlabeled nodes ranked by the one-layer version of our method had very low degree in the network (Figure 1(a)). The first 1,000 nodes show a stark drop in degree when the first unlabeled nodes are ranked (bottom row). These top-ranked, low-degree unlabeled nodes tended to be connected to positives. The multi-layer version of our method corrects for this effect; the top unlabeled nodes in the two-layer version have larger degree (Figure 1(b)).

We compared the accuracy of the multi-layer version for different numbers of layers. For each layer and each positive set, we selected the value of λ that achieved the highest accuracy in terms of area under the curve (AUC, Figure 2). The two-layer method for schizophrenia positives had an average AUC of 0.698 compared to the one-layer average AUC of 0.605 ($p = 7.07 \times 10^{-18}$, Wilcoxon rank-sum test) and the three-layer average AUC of 0.663 ($p = 1.27 \times 10^{-12}$). This trend is consistent for the two-layer method for autism positives ($p = 7.07 \times 10^{-18}$ vs. one-layer, $p = 1.55 \times 10^{-15}$ vs. three-layer) and cell motility positives ($p = 2.07 \times 10^{-17}$ vs. one-layer, $p = 4.42 \times 10^{-13}$ vs. three-layer). In terms of AUC, using two layers with $\lambda = 10$ did the best job ranking the hidden positives in the k-fold cross validation across different positive sets.

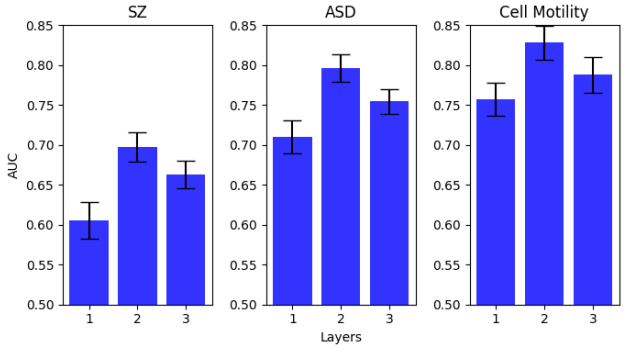

Figure 2: Five-fold cross validation performance (AUC across 50 iterations) of our method for different numbers of layers across three positive sets (schizophrenia, autism, and cell motility).

We investigated the top-ranking genes for schizophrenia in the two-layer method. Many protein phosphatases appear in the list, including Protein Phosphatase 1 subunits (PPP1R12C, PPP1R12A, and PPP1CB), and a Protein Phosphatase 2 subunit (PPP2R2A); these are all labeled as cell motility positives. Protein Phosphatase 1 is a protein necessary for cell division and regulates muscle contractility among many other functions. Interestingly, inhibition of PPI prolongs memory after a learning scenario, suggesting a link between PPI and learning and memory [3]. Predicted genes that are known schizophrenia positives but are not cell motility positives are also valuable candidates for follow-up experimental validation using cell motility assays. The first gene that appears as a cell motility unlabeled node is Lysyl Oxidase (LOX) at rank 30. LOX has been associated with metastasis in certain cancers due to its role in hypoxic conditions [1], and cell motility is one component of invasive cell migration. Our work provides a methodology for investigating biological processes that may be disrupted in polygenic diseases.

Acknowledgements. This work was supported by a Computing Research Association (CRA-W) Collaborative REU (CREU) and NSF awards MCB-1716964 and ABI-1750981 (to PI AR).

REFERENCES

[1] Janine T Erler and others. 2006. Lysyl oxidase is essential for hypoxia-induced metastasis. *Nature* 440, 7088 (2006), 1222.
[2] Yongjun Fan and others. 2013. Focal adhesion dynamics are altered in schizophrenia. *Biological psychiatry* 74, 6 (2013), 418–426.
[3] David Genoux and others. 2002. Protein phosphatase 1 is a molecular constraint on learning and memory. *Nature* 418, 6901 (2002), 970.
[4] Casey S Greene and others. 2015. Understanding multicellular function and disease with human tissue-specific networks. *Nature genetics* 47, 6 (2015), 569.
[5] Arjun Krishnan and others. 2016. Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nature neuroscience* 19, 11 (2016), 1454.
[6] TM Murali and others. 2011. Network-based prediction and analysis of HIV dependency factors. *PLoS computational biology* 7, 9 (2011), e1002164.
[7] Stephan Ripke and others. 2014. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 511, 7510 (2014), 421.
[8] Jerzy Wegiel and others. 2010. The neuropathology of autism: defects of neurogenesis and neuronal migration, and dysplastic changes. *Acta neuropathologica* 119, 6 (2010), 755–770.
[9] Xiaojin Zhu and others. 2003. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)*. 912–919.