Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category:

Subject Section

EpistasisRank and EpistasisKatz: interaction network centrality methods that integrate prior knowledge networks

Saeid Parvandeh[1], Brett A. McKinney[1,2*]

[1]Tandy School of Computer Science, University of Tulsa, OK 74104, USA.

[2] Department of Mathematics, University of Tulsa, OK 74104, USA.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: An important challenge in gene expression analysis is to improve hub gene selection to enrich for biological relevance or improve classification accuracy for a given phenotype. In order to incorporate phenotypic context into co-expression, we recently developed an epistasis-expression network centrality method that blends the importance of gene-gene interactions (epistasis) and main effects of genes. Further blending of prior knowledge from functional interactions has the potential to enrich for relevant genes and stabilize classification.

Results: We develop two new expression-epistasis centrality methods that incorporate interaction prior knowledge. The first extends our SNPrank (EpistasisRank) method by incorporating a gene-wise prior knowledge vector. This prior knowledge vector informs the centrality algorithm of the inclination of a gene to be involved in interactions by incorporating functional interaction information from the Integrative Multi-species Prediction (IMP) database. The second method extends Katz centrality to expression-epistasis networks (EpistasisKatz), extends the Katz bias to be a gene-wise vector of main effects and extends the Katz attenuation constant prefactor to be a prior-knowledge vector for interactions. Using independent microarray studies of major depressive disorder, we find that including prior knowledge in network centrality feature selection stabilizes the training classification and reduces overfitting.

Contact: brett-mckinney@utulsa.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Hubs in gene co-expression networks likely play an important role in understanding the regulation of biological processes and phenotypes. Recent studies have investigated the potential for co-expression network hubs to be used to prioritize genes for statistical inference. GeneRank (Morrison, Breitling et al. 2005) used the PageRank algorithm (Page, Brin et al. 1999) to prioritize genes by combining gene co-expression with external information, such as Gene Ontology and protein-protein interactions. The constant damping constant in PageRank was extended to a damping vector in Ref. (Fu, Lin et al. 2006), and the usage of this damping vector to incorporate prior knowledge in GeneRank was discussed in Ref. (Demidenko 2015).

Co-expression network hubs do not explicitly use outcome or phenotype information. This controls the risk of overfitting in classification but also loses important contextual information about connectivity influenced by the phenotype. We developed a gene expression centrality (EpistasisRank) that includes phenotype context by computing statistical interactions (e.g., epistasis) between transcripts (Lareau, White et al. 2015) in an epistasis-expression or differential co-expression network (McKinney, White et al. 2013). Prior to this generalization to expression data, we had developed a SNPrank centrality for epistasis networks in GWAS (McKinney, Crowe et al. 2009, Hu, Andrew et al. 2013). In the current study, we extend the

generalized EpistasisRank method to use a gene-wise interaction prior probability vector, and we develop a new epistasis network centrality based on Katz centrality (Katz 1953) that combines main and interaction effects as well as prior knowledge when ranking the importance of predictors.

2 Methods

2.1 EpistasisRank and EpistasisKatz centrality with gene-wise prior probability

EpistasisRank centrality (ER) operates on a regression-based Genetic Association Interaction (reGAIN) network (Pandey, Davis et al. 2012), which is a weighted *NxN* matrix, *B*, where *N* is the number of genes. The diagonal, B_{ij}, represents the main effect regression coefficient of the gene on the phenotype and the off-diagonal B_{ij} is the interaction effect coefficient between genes on the phenotype. The formula for ER is a system of equations that can be solved through least squares:

$$ER_i = \frac{B_{ii}}{N \cdot \mathrm{Tr}(B)} + d_i \sum_{j \neq i} \frac{B_{ij} \cdot ER_j}{k_j} + \frac{1 - d_i}{N}. \quad (1)$$

In the first term, each gene *i* gets a contribution to network importance from the gene's main effect (B_{ii}), where the trace of *B, Tr(B)*, is a normalization. In the second term, each gene *i* gets a contribution from its interaction partners B_{ij} proportional to the importance of the partners, ER_j, normalized by the degree of gene *j*, k_j (non-zero), from the B matrix. This total interaction contribution is weighted by the prior probability d_i for gene *i* to be involved in interactions. The prior probability vector d_i is the normalized degree of the IMP network. The last term gives all genes a uniform importance proportional to the complement of its inclination for interaction, (1-d_i).

Katz centrality is a two-parameter extension of eigenvector centrality (Supplementary Material). We extend Katz to EpistasisKatz (EK) with prior knowledge as follows

$$EK_i = d_i \sum_{j \neq i} B_{ij} EK_j + B_{ii}. \quad (2)$$

In the first term, each gene *i* is given network importance based on the EK weights, EK_j, of its interaction partners and their B_{ij} reGAIN regression weights. The interaction term is weighted by IMP prior knowledge vector d_i. In standard Katz, this prefactor is a constant that attenuates the centrality contribution of more distant connections. Thus, we extend the attenuation constant in Katz to allow for gene-specific attenuation (d_i), which is the IMP-based prior probability for interactions. In the second term, sometimes referred to as the bias vector, each gene is assigned importance based on its main effect, B_{ii}. In standard Katz, this second term is a vector of repeated constants. This extends Katz to allow a vector of gene-wise constants.

2.2 Data Processing

We identified two gene expression data sets from GEO for major depressive disorder that we refer to as Cambridge (Leday, Vertes et al. 2017) and Japan (Miyata, Kurachi et al. 2016). We *Z*-transformed each dataset based on their respective controls to make the data sets more comparable to each other (Wang, Oh et al. 2016). We were also concerned about the imbalanced case/control ratio in the Cambridge (training) data with its 128 cases and 64 controls. Thus, we under-sampled (Lina 2015) the case samples in the Cambridge data set to obtain a balance of 64 cases and 64 controls. In the Japan (testing) data set, there are 20 cases and 12 controls. In addition, we filtered the top 5,000 genes using coefficient of variation across the two data sets. For prior knowledge, we used the 5,000 genes to query IMP to construct a network based on predicted functional interactions, and then we computed the normalized degree of each gene *i* of the IMP network as the prior knowledge vector d_i.

3 Results

We compared training accuracy and validation accuracy using each centrality method (PageRank, Katz, EK, and ER) for feature selection with and without prior knowledge (Fig. 1). To avoid overfitting, we used nested cross-validation (CV) to prevent feature selection from causing overfitting (Varma and Simon 2006, Le, Simmons et al. 2017). We used xgboost binary classification on boosted decision trees (Chen and Guestrin 2016) for the outer CV loop and centrality feature selection methods in the inner CV loop.

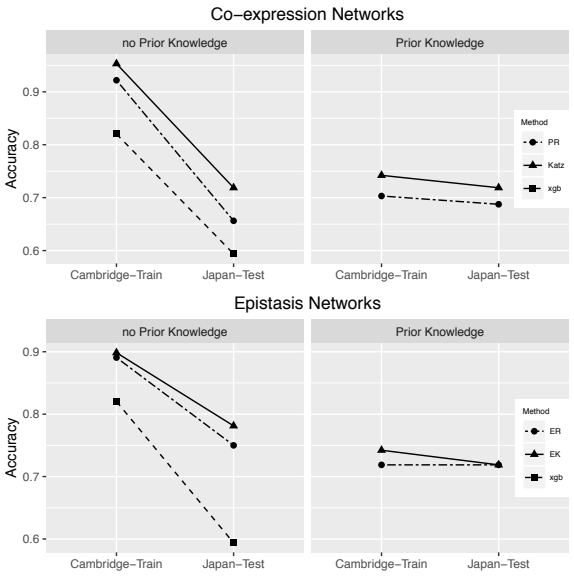

Fig. 1. Training accuracy (Cambridge data) and independent validation accuracy (Japan data) with centrality feature selection without prior knowledge (left panels) and with prior knowledge (right panels). Top: co-expression network centrality feature selection methods, PageRank (PR) and Katz. Bottom row: expression-epistasis network centrality methods, EpistasisRank (ER) and EpistasisKatz (EK). Accuracies computed by xgboosted trees with nested cross-validation. Xgboost accuracies without feature selection also shown (squares).

All centrality feature selection methods improve validation accuracy over xgboost classification without feature selection (Fig. 1). Katz-based centralities have the highest accuracies. Without prior knowledge (left panels of Fig. 1), all feature selection methods show a large drop in validation accuracy relative to the training accuracy (overfitting) despite use of nested CV. Use of prior knowledge to inform centrality (right panels of Fig. 1) yields more stable accuracy across training and validation sets. The training accuracies are lower than without prior knowledge; however, they are more consistent with and a more realistic estimate of the independent validation accuracy.

4 Discussion

In a previous study, we used the Integrative Multi-species Prediction (IMP) database (Wong, Park et al. 2012) to predict functional networks from epistasis network seed genes from SNPrank in GWAS data (McKinney, Lareau et al. 2016). In the current study, we compute the degree centrality (d_i) of each gene *i* from an IMP network and use this as a prior probability vector for the interaction term in the new EpistasisRank (ER) and EpistasisKatz (EK) epistasis-expression centralities for a network from an independent data set. We hypothesize that incorporating functional-connectivity prior knowledge into epistasis-expression network centrality will improve the generalization of classification accuracy

We extended the ER centrality to include a gene-wise vector to integrate prior knowledge. We generalized Katz centrality in EK to include a gene-specific vector, which we use to incorporate the prior probability for interaction effects. We extended the constant bias vector term in Katz to incorporate main effect contributions from the reGAIN matrix. We found prior knowledge led to more stable training accuracy and improved testing validation accuracy in gene expression analysis of major depressive disorder.

Prior knowledge also led to an increase in the number of significantly enriched relevant pathways (Supplement). For example, including prior knowledge led to statistically significant enrichment of Serotonin Receptor and G coupled protein receptor (GPCR) pathways, which are related to mood disorders (Imbrici, Conte Camerino et al. 2013). The ER and EK methods apply to epistasis networks in GWAS as well as gene expression, and the prior probability vector can blend information between heterogeneous data-driven networks as well as prior knowledge from IMP or other prior networks.

The network construction and centrality methods, including `EpistasisRank` and `EpistasisKatz`, are included in our `Rinbix` R package at https://github.com/insilico/Rinbix. The specific feature selection and classification analysis in the current study is reproduced in https://github.com/insilico/PriorKnowledgeEpistasisRank.

Funding

This work was supported in part by the National Institute of Health GM121312 and GM103456 (B.A.M.) and the William K. Warren Jr. Foundation.

Conflict of Interest: none declared.

References

Chen, T. and C. Guestrin (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California, USA, ACM**:** 785-794.

Demidenko, E. (2015). "Microarray enriched gene rank." BioData Mining **8**: 2.

Fu, H.-H., D. K. J. Lin and H.-T. Tsai (2006). "Damping factor in Google page ranking." Applied Stochastic Models in Business and Industry **22**(5-6): 431-444.

Hu, T., A. S. Andrew, M. R. Karagas and J. H. Moore (2013). "Statistical epistasis networks reduce the computational complexity of searching three-locus genetic models." Pac Symp Biocomput: 397-408.

Imbrici, P., D. Conte Camerino and D. Tricarico (2013). "Major channels involved in neuropsychiatric disorders and therapeutic perspectives." Frontiers in Genetics **4**(76).

Katz, L. (1953). "A new status index derived from sociometric analysis." Psychometrika **18**(1): 39-43.

Lareau, C. A., B. C. White, A. L. Oberg and B. A. McKinney (2015). "Differential co-expression network centrality and machine learning feature selection for identifying susceptibility hubs in networks with scale-free structure." BioData Mining **8**(1): 5.

Le, T. T., W. K. Simmons, M. Misaki, J. Bodurka, B. C. White, J. Savitz and B. A. McKinney (2017). "Differential privacy-based evaporative cooling feature selection and classification with relief-F and random forests." Bioinformatics **33**(18): 2906-2913.

Leday, G. G. R., P. E. Vertes, S. Richardson, J. R. Greene, T. Regan, S. Khan, R. Henderson, T. C. Freeman, C. M. Pariante, N. A. Harrison, V. H. Perry, W. C. Drevets, G. M. Wittenberg and E. T. Bullmore (2017). "Replicable and Coupled Changes in Innate and Adaptive Immune Gene Expression in Two Case-Control Studies of Blood Microarrays in Major Depressive Disorder." Biol Psychiatry.

Lina, G. (2015). Data sampling improvement by developing SMOTE technique in SAS. Proceedings of the SAS Global Forum 2015 Conference.

McKinney, B. A., J. E. Crowe, Jr., J. Guo and D. Tian (2009). "Capturing the Spectrum of Interaction Effects in Genetic Association Studies by Simulated Evaporative Cooling Network Analysis." PLOS Genetics **5**(3): e1000432.

McKinney, B. A., C. Lareau, A. L. Oberg, R. B. Kennedy, I. G. Ovsyannikova and G. A. Poland (2016). "The Integration of Epistasis Network and Functional Interactions in a GWAS Implicates RXR Pathway Genes in the Immune Response to Smallpox Vaccine." PLoS ONE **11**(8): e0158016.

McKinney, B. A., B. C. White, D. E. Grill, P. W. Li, R. B. Kennedy, G. A. Poland and A. L. Oberg (2013). "ReliefSeq: A Gene-Wise Adaptive-K Nearest-Neighbor Feature Selection Tool for Finding Gene-Gene Interactions and Main Effects in mRNA-Seq Gene Expression Data." PLOS ONE **8**(12): e81527.

Miyata, S., M. Kurachi, Y. Okano, N. Sakurai, A. Kobayashi, K. Harada, H. Yamagata, K. Matsuo, K. Takahashi, K. Narita, M. Fukuda, Y. Ishizaki and M. Mikuni (2016). "Blood Transcriptomic Markers in Patients with Late-Onset Major Depressive Disorder." PLoS One **11**(2): e0150262.

Morrison, J. L., R. Breitling, D. J. Higham and D. R. Gilbert (2005). "GeneRank: using search engine technology for the analysis of microarray experiments." BMC Bioinformatics **6**: 233.

Page, L., S. Brin, R. Motwani and T. Winograd (1999). The PageRank Citation Ranking: Bringing Order to the Web, Stanford InfoLab.

Pandey, A., N. A. Davis, B. C. White, N. M. Pajewski, J. Savitz and W. C. Drevets (2012). "Epistasis network centrality analysis yields pathway replication across two GWAS cohorts for bipolar disorder." Transl Psychiatry **2**.

Varma, S. and R. Simon (2006). "Bias in error estimation when using cross-validation for model selection." BMC bioinformatics **7**: 91-91.

Wang, L., W. K. Oh and J. Zhu (2016). "Disease-specific classification using deconvoluted whole blood gene expression." Scientific Reports **6**: 32976.

Wong, A. K., C. Y. Park, C. S. Greene, L. A. Bongo, Y. Guan and O. G. Troyanskaya (2012). "IMP: a multi-species functional genomics portal for integration, visualization and prediction of protein functions and networks." Nucleic Acids Res **40**(Web Server issue): W484-490.