Predicting Epistatic Interactions Using Information and Network Theory

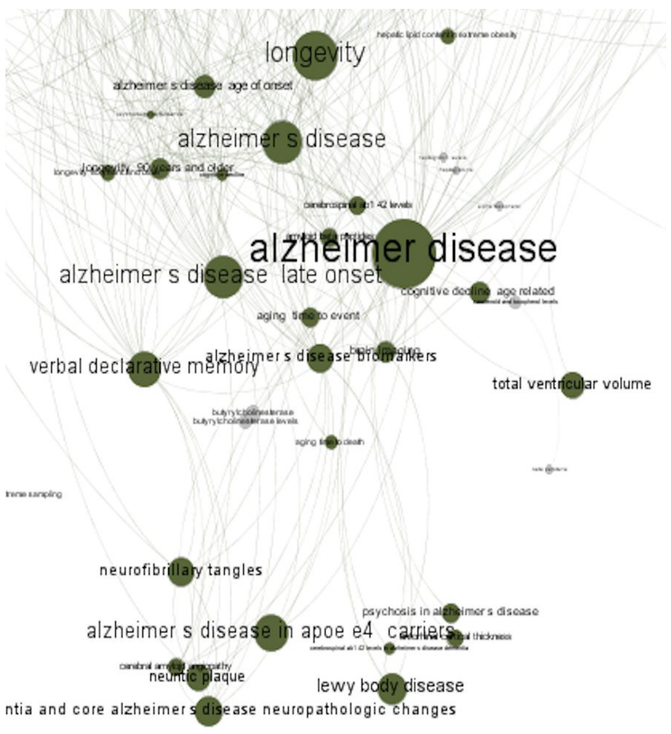
Krishna C. Bathina

bathina@umail.iu.edu, krishnacb.com

Indiana University
School of Informatics, Computing, and Engineering

Motivation

Since the Human Genome Project, there has been a vested interest in discovering genetic bases for disease phenotypes. Most research methods focus on finding the effects of **individual Single Nucleotide Polymorphisms (SNPs)** on a phenotype. While producing many positive results, these **methods typically do not discover *multiple* SNP**, or epistatic, effects on a phenotype. One method from [1,2], uses the ***Information Gain (IG)*** between SNPs as edge weights within a SNP-SNP interaction network. to find important causal SNPs in binary phenotypes. I extend upon this method to work for **continuous phenotypes**.



Conclusion

This method provides a simple and quick way to calculate the epistatic interaction that two SNPs could have on a phenotype.

Steps to further this work:

1. **Make a series of toy data sets** over reasonable distributions. Compare this method with other well established ones.
2. **Choose a disease phenotype** and **apply this method on genomic data from dbGaP**. The results can be annotated and submitted for further study.
3. **Experiment with new network methods**, such as community detection to find a better set of SNPs and dyadicity and heterophilicity to capture the effect of node properties.

References

1. Hu, Ting, et al. "Genome-wide genetic interaction analysis of glaucoma using expert knowledge derived from human phenotype networks." *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*. Vol. 20. NIH Public Access, 2015.
2. Hu, Ting, et al. "Characterizing genetic interactions in human disease association studies using statistical epistasis networks." *BMC bioinformatics*12.1 (2011): 364.
3. Ross, Brian C. "Mutual information between discrete and continuous data sets." *PloS one* 9.2 (2014): e87357.

Methods - Epistatic Detection

1. **Build a phenotype-phenotype network** (figure to the left). Edge weights are the Jaccard index of the common SNPs between any two phenotypes. *Phenotypes with more overlapping SNPs have a larger edge weight.*
2. **Choose a phenotype** and its first degree neighbors.
3. All of the SNPs in the group of phenotypes are used to **build a SNP-SNP network**. Edge weights are proportional to the IG between them.

   Information Gain - Given two SNPs, A and B, and phenotype, P, the IG, is the difference of the joint mutual information of (A,B;P) with the mutual information of both (A,P) and (B,P). The calculation is shown in the section below.

$$IG(A,B;\mathcal{P}) = I(A,B;\mathcal{P}) - I(A;\mathcal{P}) - I(B;\mathcal{P})$$

4. **Permute the original network to form 100 new SNP-SNP networks** by randomizing the phenotype class and recalculating the IG.
5. For each SNP-SNP network, **threshold the edges** from IG = 0 to max(IG), in increments of 0.0001, by only including edges with IG ≥ current threshold.
6. **Calculate network statistics** for all of the thresholded networks for each SNP-SNP network
7. **Run a permutation test** to find which threshold leads to the statistically ($p < 0.05$) largest connected component in the original SNP-SNP network compared to the permuted networks.
8. **Calculate degree, betweenness, and closeness centrality** of the original SNP-SNP network at the statistical threshold to find most important SNPs.
9. **Annotate SNPs** to find existing pathways/functions from past lab and GWAS results.

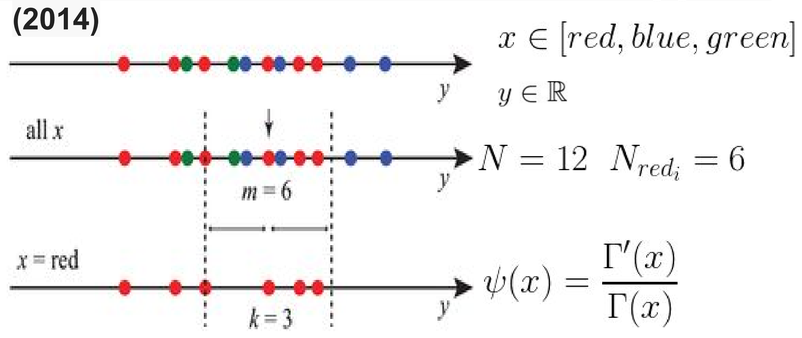
Methods - Mutual Information

- **Mutual Information (MI) of X,Y is represented as**

$$I(X,Y)_D = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$

$$I(X,Y)_C = \int_x\int_y p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dydx$$

- **What about continuous and discrete data? - [3] Ross (2014)**



$$I_i = \psi(N) - \psi(N_{x_i}) + \psi(k) - \psi(m_i)$$

$$I(X,Y) = \langle I_i \rangle = \psi(N) - \langle \psi(N_x) \rangle + \psi(k) - \langle \psi(m) \rangle$$

Data

- Toy dataset - **4000 subjects** and **200 SNPs**
- **Risk variants** were assigned according to **Hardy-Weinberg equilibrium with MAF < 0.5**,
- Phenotype - **mixed linear model** with bilinear term

$$P = \beta_0 + \beta_{0,1}X_0X_1 + \sum_{n=1}^{N}\beta_nX_n + \mathcal{N}(0,1)$$

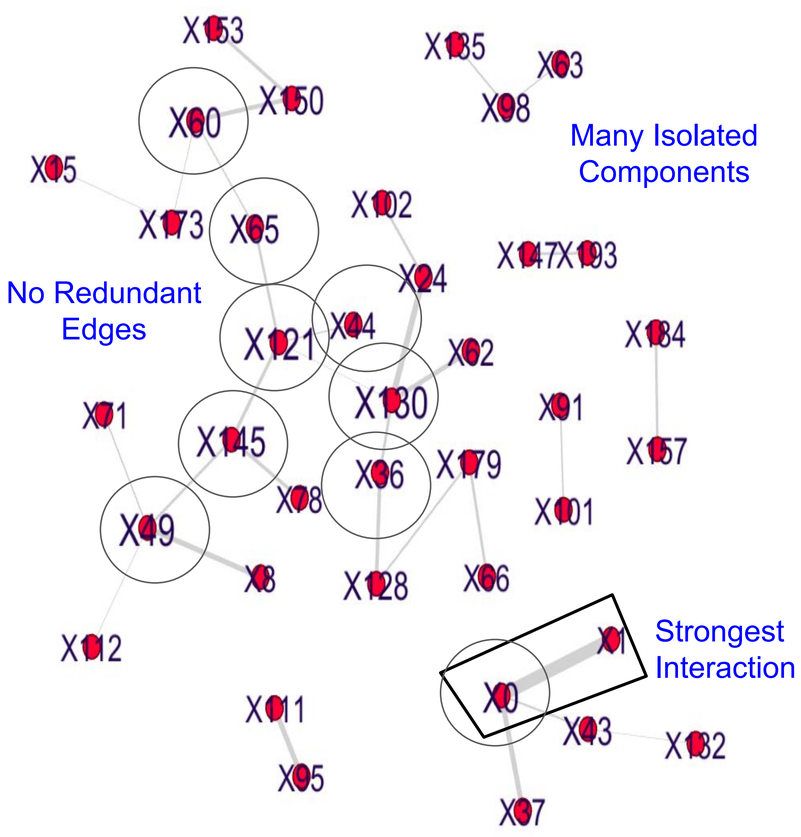
- **P** - continuous value representing the phenotype
- β_0 - intercept
- β_n - effect size of base n
- X_n - number of risk variants (**AA** = 2, **Aa** = 1, **aa** = 0) where A is the risk allele and a is the common allele
- $\beta_{a,b}$ - effect size of epistatic effect between base a and b
- **N(0,1)** - error term

Results

Parameters::

- $\beta_0 = 1$
- $\beta_1 = 1.5$
- $\beta_2 = 1.5$
- $\beta_{1,2} = 2.2$
- $\beta_3 = N(0,0.5)$
- MAF = U(0,0.5)

- Prop. of interactions with negative IG = **0.538**
- Prop. of interactions with no IG = **0.177**
- Statistically sig. cutoff = **0.0216** (p = 0.05)



Nodes to Investigate

Degree Centrality	Betweenness Centrality	Closeness Centrality
X130	X121	X121
X121	X130	X130
X49	X145	X145
X145	X65	X65
X0	X49	X44
X60	X36	X60