6.047/6.878 Final Project Proposal

Yang Dai, Aditya Karan, Matthew Feng

October 19, 2018

1 Introduction

While mental illness has become more well-studied there still remains many unanswered questions as to mechanism and relationship between the various diseases. The advent of larger databases of interactions between genes and proteins has opened up the potential use of network analysis to characterize complex interactions across an entire network and similarities across seemingly disparate components diseases.

However, some challenges remain, such as lack of consensus as to what defines a disease cluster and lack of clinical verification of several models. Moreover, most current studies tend to look at differences between classes of disease to show closeness without looking into and attempting to characterize more deeply a specific class of diseases.

We propose to investigate a specific, highly similar set of diseases, heritable mental disorders such as autism, bipolar and schizophrenia, in order to characterize the protein network. For comparison purposes we also add to our analysis several metabolic diseases such as diabetes. For our analysis we plan to construct the PPI network by combining what genes are known to cause these various diseases and layering on dynamical information. From this we can perform tests on our models such as validating whether we see evidence of clinical similarities within diseases that have overlap. Additional analysis on top of these constructed network would be to study perturbations on single nodes within the disease cluster and observe how these perturbations can give rise to various diseases.

2 Specific Aims

We are focused on examining biomolecular pathways involved in the development of highly heritable mental disorders, such as autism, bipolar disorder, schizophrenia. Additionally, we want to examine differences in network topology associated with various diseases, such as potential differences between mental disorders and metabolic disorders, such as diabetes mellitus.

Specifically, we hope to accomplish the following:

Aim 1. Investigate the disease modules of several highly heritable diseases using relationships encoded in the human interactome.

Aim 2. Find the overlap of the disease modules of diseases with similar phenotypic features versus diseases with no shared phenotypic features.

Aim 3. Evaluate disease modules against empirical data (i.e. clinical data).

3 Research Strategy

3.1 Significance

For the past decade, researchers have been steadily improving their understanding and knowledge about the causes underlying many mental illnesses such as autism and schizophrenia. However, much still has to be learned. There is still no definite causal pathways that map genetic mutations to the phenotypical changes exhibited by autism and schizophrenia. As such, diagnosis has struggled as well; from the rapid rise in diagnoses of mental illnesses, it remains unclear whether the illnesses we identify today are simply umbrella terms for multiple different disorders. Thus, we must continue to further our understanding of the pathways that leads to the development of these illnesses if we are to be able to treat them effectively.

The interactome, in spite of its incompleteness, is a valuable tool that has allowed researchers to discover unexpected relations between diseases that did not otherwise have manifest clinical or pathobiological associations [4]. A more thorough and expanded examination of the interactome via augmentations in interaction details may reveal more disease-disease interactions or similarities that would otherwise go unnoticed. In particular, exploration of various different metrics may yield more insightful and useful ways of relating pathways in interaction networks, beyond the typical "shortest-path" formulation.

Furthermore, newly discovered interactions among diseases may reveal drug treatments that may be efficacious beyond the traditional diseases they are known to treat. Similarly, we may discover treatments that are able to alleviate comorbid diseases, thereby increasing the efficiency of prescriptions and potentially reducing the negative side-effects that are bundled with certain drugs.

Finally, the interactome InWeb_InBioMap [2] has incorporated "severalfold more interaction data connecting brain-regulated genes at the protein level than the next largest network"; we hope that this expansion of interaction data will allow us to discover new pathways that may play a role in the development of highly heritable mental disorders such as schizophrenia and autism. With a more solidified knowledge of the pathways regarding such mental illnesses, we will be better equipped to diagnose and develop treatments for these diseases.

3.2 Innovation

Researchers [4] compiled a human interactome consisting of 141,296 physical interactions between 13,460 proteins. The dataset is in the form of an undirected, unweighted graph, which we will use as a starting point for our project. One drawback to the dataset is that it does not provide information on the specifics of the protein-protein interaction (e.g. strength of interaction, activating vs repressing interaction). We plan to augment the dataset through by establishing weights representing the propagation of proteins in a given system based on [5]. Another potential approach is to is to assign a weight to each edge that correlates to the confidence of the interaction between the two proteins. This data can be found in other databases such as InWeb_InBioMap [2]. Last, we can try to construct a signed, directed graph in which an edge represents whether one protein inhibits or activates another protein. We will perform the data augmentation only for edges of the 5 disease modules of interest.

Using this augmented version of the human interactome, we will explore various metrics of each disease module in addition to how the disease modules interact. For example, we can look at the shortest path between all pairs of nodes in the module, the PageRank scores of the nodes, and identify highly connected nodes (hubs). Between disease modules, we can look at the degree of overlap/separation (e.g. shortest distance, min cut, minimum number of edges that need to be removed to disconnect two disease modules).

By looking at previously unexamined network metrics, we hope to find clinically interesting connections between diseases that are already known to share genetic and phenotypic features (schizophrenia, bipolar, autism) and diseases with no shared features (e.g. schizophrenia and diabetes) but have a larger than expected comorbidity.

3.3 Approach

Using OMIM, GWAS, GTEX — we aim to identify the set of proteins that are thought to be highly correlated to the diseases of interest. From this set of genes we can then create a simplified PPI network among the given proteins. We then add weights to the graphs to represent the "strength" of interactions. The weights are determined by a Jacobian matrix derived from differential equations representing the propagation of proteins [5]. Given a weighted graph - we then aim to explore various metrics to define a suitable cluster. The metrics here are more exploratory but may potentially include spectral clustering, identification of bi-connected components/articulation points/ cutoffs based on length of path between the clusters or min cuts. Much of this stage of analysis will be done via tools like NetworkX.

From this we aim to conduct a number of potential applications of these networks. One potential application is evaluating our proposed interaction clusters with clinical data to see if drugs that treat one disease have been shown to be effective in treating systems in another disease (e.g. are there clinical trials showing that a particular heart disease drug has some effect on mitigating

schizophrenia, in concordance with our PPI). To do so — based on our observations of the PPI we would then search literature if there has been any recorded clinical significance between the two components. Another analysis we can look at is perturbing the network to see the spillover effects — if theres a mutation that knocks out a gene — how would that percolate through the network and would this explain the emergence of the diseases of interest. This analysis would be more of a recommendation as to which potential proteins should be furthered studied.

4 Resources

To examine the biomolecular interactions associated with various mental and metabolic disorders, we will use the Online Mendelian Inheritance in Man (OMIM) [1] and GWAS [3] to locate gene mutations that can then be mapped to protein variations. We will then use various public interactomes such as InWeb_InBioMap [2] and the Human Interactome [4] to identify interaction pathways. In particular, we hope to use the additional information encoded in the InWeb interactome to study disease modules, as opposed to only examining network diameters.

We do not expect our project to be computationally intensive; however, should the need arise, we will be able to use Amazon Web Services and Paperspace instances for compute. We have also begun working with Dr. Jose Davila-Velderrain of CSAIL for guidance and mentorship.

Lectures 11, 14, 16, and 22 on network structure, GWAS, and phenotype mining will be particularly relevant to our project.

5 Collaboration

We plan on initially dividing initial work on collection and modeling of data based on the different diseases - each person would take one disease and collect the relevant PPI data for that diseases and create a solo network. Then, combining the various networks together (hopefuly with strong overlap!) we then will split the task where 1-2 people are working on cluster discovery/identification and 1-2 people working on interpreting and predicting outcomes (disease in clinical studies etc.). The split at that point will primarily be determined on which step is more limiting (cluster identification vs. creation or applications).

6 Timeline

10/26	Identify the genes associated with the 5 diseases of interest. Compile human interactome dataset with augmented information.
11/2	Construct disease modules for each disease and find metrics for each disease module.
11/9	Perform comparisons between the disease modules.
11/16	Determine the biologically significant metrics
11/23	Potentially add other diseases for comparison and experiment with different metrics
11/26	Midcourse progress report due
11/30	Link results with what scientists know about the diseases
12/9	Final report due

References

[1] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick. Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33(suppl_1):D514–D517, 2005.

[2] T. Li, R. Wernersson, R. B. Hansen, H. Horn, J. Mercer, G. Slodkowicz, C. T. Workman, O. Rigina, K. Rapacki, H. H. Stærfeldt, S. Brunak, T. S. Jensen, and K. Lage. A scored human protein-protein interaction network to catalyze genomic interpretation. *Nature Methods*, 14:61 EP –, Nov 2016.

[3] J. MacArthur, E. Bowler, M. Cerezo, L. Gil, P. Hall, E. Hastings, H. Junkins, A. McMahon, A. Milano, J. Morales, Z. M. Pendlington, D. Welter, T. Burdett, L. Hindorff, P. Flicek, F. Cunningham, and H. Parkinson. The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog). *Nucleic Acids Res*, 45(Database issue):D896–D901, Jan 2017. 27899670[pmid].

[4] J. Menche, A. Sharma, M. Kitsak, S. D. Ghiassian, M. Vidal, J. Loscalzo, and A.-L. Barabási. Uncovering disease-disease relationships through the incomplete interactome. *Science*, 347(6224), 2015.

[5] M. Santolini and A.-L. Barabási. Predicting perturbation patterns from the topology of biological networks. *Proceedings of the National Academy of Sciences*, 2018.