Weighting Protein-Protein Interaction Networks

Gavin Gray Supervisor: Colin Mclean and Douglas Armstrong

1. Introduction

Focusing on proteins in the synapse, we analysed a set of proteins called the “active zone” network. Two graphs were generated, one with unweighted edges and another with weighted edges.

Weighting the edges in the second graph was performed by estimating the posterior probability of interaction for a given pair of proteins using a combination of supervised classification and manually adjusted bayesian inference as described in figure 1. The aim of this project was to determine whether this approach would improve the disease associations of communities detected through use of a spectral modularity community detection algorithm.

2. Methodology

Supervised classification was performed with approximately 1,000,000 training examples and a random forest classifier using Gene Ontology and domain information based features.

Bayesian weight updated was performed with a Naive Bayes modeling using conditionally independent Bernoulli and Beta distributions. The Beta distributions were trained in a supervised manner while the Bernoulli distributions were defined by conservative estimation.

Community detection was performed using a spectral modularity algorithm. Communities detected were compared with Normalised Mutual Information (NMI). Disease associations were found through a disease enrichment algorithm using a hypergeometric test.

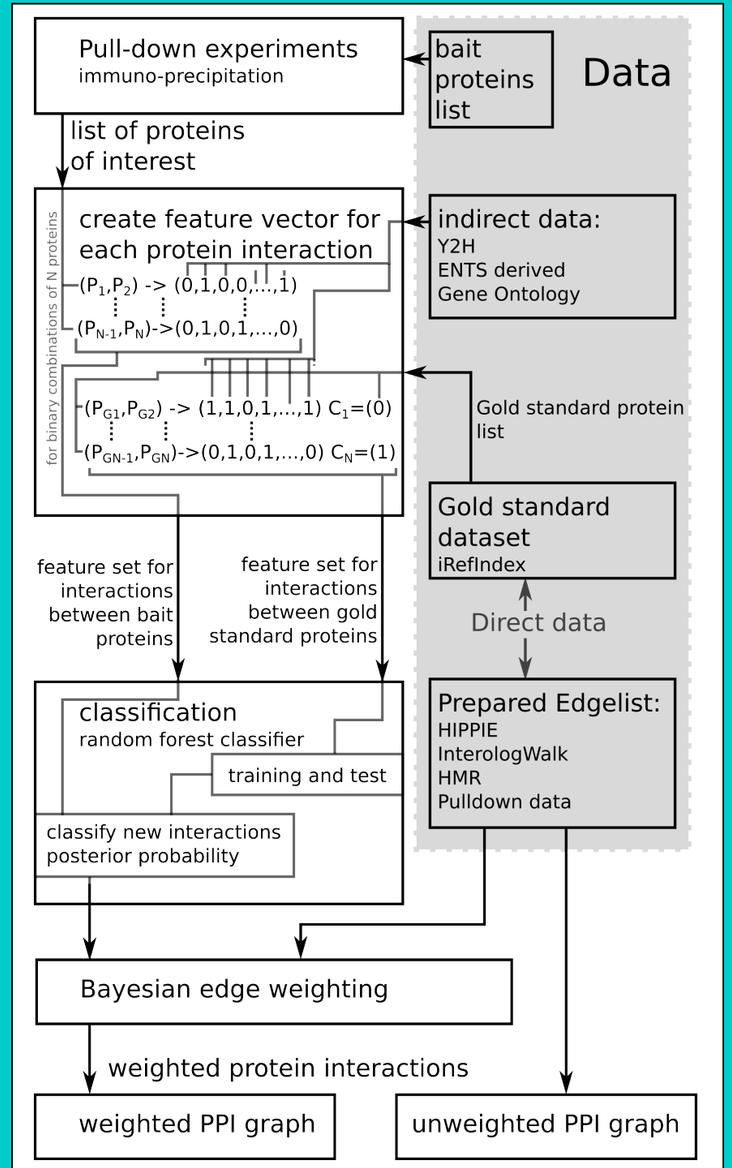

Figure 1: The above flow chart describes the method used to construct both weighted and unweighted networks using a variety of techniques including machine learning and bayesian inference.

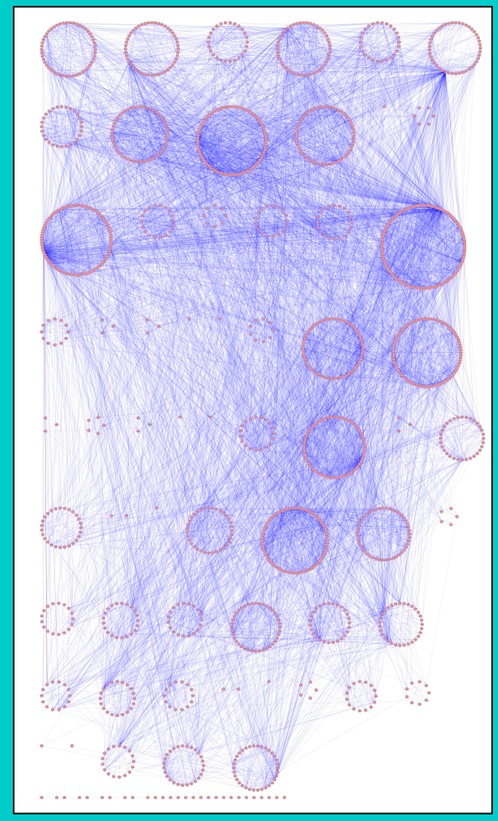

Figure 2. Weighted network produced clustered into communities.

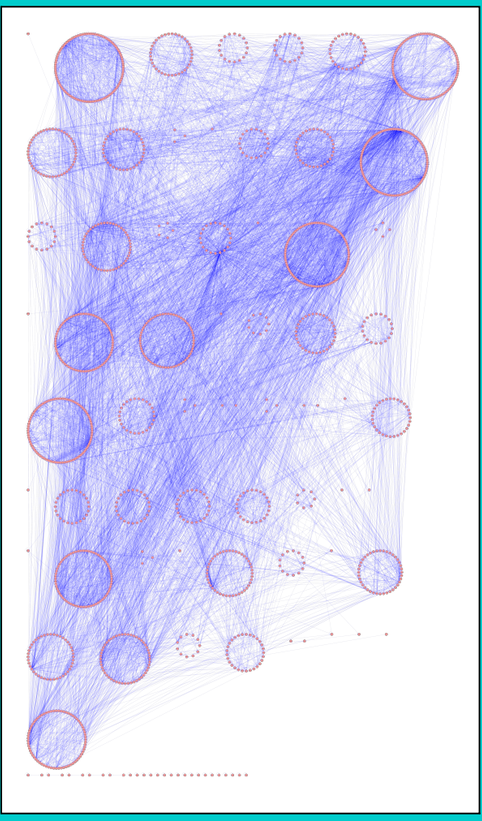

Figure 3. Unweighted network produced clustered into communities.

3. Results

The resulting weights produced a graph which, when separated into communities, is clearly different from the unweighted case. However, there was no clear insight into disease possible.

4. Discussion

A more detailed probabilistic model is required, treating interaction strength as a Beta distributed random variable and attempting to predict it's strength based on a greater array of biological information.