Concept behind the project

The initial concept behind the project was to try and predict protein interactions given the co-expression based on expression data downloaded from GEO. Finally the predicted network and the know networks will be plotted with networkX to try and visualise the interactions.

GEO data

The dataset used here are a co-expression experiment in *E. coli* (data available at GEO accession GSE40811). This data was used as a known protein interaction network was available at the start of the project.

Note that due to the size this data is not included in the GitHub repo.

Performed work

The initial step here was to download the data. Once this was done, I implemented a data reader that takes all (n=97) the co-expression data files in a directory and merges these in to a pandas data frame. In addition to these files the annotation file was loaded and the gene annotations were captured. For the gene names the loci positional code were used as this is what was used in the known gold standard. Finally, in data loading I read in the gold standard and generate a adjacency matrix for all proteins with known interactions.

Once all the data was loaded and put in to two pandas data frames the spearman correlation of each row in the raw data was computed. This was used as a measurement of how likely it is that each gene is co-expressed in the data at hand. The spearman correlation values were then moved in to five bins using the sklearn preprocessing KBinsDiscretizer library. This was done as naïve Bayesian models often tend to work better on binned data than continues data.

Finally, I tried to fit the gold standard to the binned correlation data. Despite attempting several ideas to do this I continuously ran in to issues with the dimensions of the data as well as the fitted data being completely unable to predict anything. To try and resolve this both a Multinomial naïve Bayesian and a gaussian model were used. However, neither of them was able to find any predicted links at all.

To try and get some plotting in with networkX as this was part of the original proposal I opted to plot the know network only. As no links could be predicted from the input data no network could be plotted for this of course.

Discussion

While the project unquestionably was a failure from the point of predicting networks, it is hard to say exactly in what step the issue lies. There are several possible points in which the script might run in to issues. First of it is of course impossible to rule out that I have done something wrong in the code as if I knew what this would be, I would of course have fixed it. However, there are also possibilities that the data is not enough to predict the network as the true network is generated from a multitude of sources. It is also possible that the known network should have had a number of true negatives added to it to allow for prediction. Rather than doing his I opted to use the once with no known interaction as true negatives. Something I think should work but can't truly guaranteed. More work could also have been allocated to ensure that the true network could be fitted to the input data. Sadly, this was not done as the loading and pre-processing part of the data took far longer than I had anticipated in part due to my plan proving somewhat unrealistic in simply reading the data to a pandas data-frame and in some part due to my own inexperience with pandas. In the end I do theorise that it is the dimension between the known network and the raw data that cannot be fitted, however I sadly do not know how to adjust this currently.

Conclusions

While the project as a whole did not work out as planed I did get a chance to learn a great deal about both the sklearn package and even more so pandas which will be very helpful in the future. Sadly in regards to the project idea it is hard to draw any conclusions from this work other than the importance of allotting more than the expected time as unexpected issues are almost guaranteed to arise throughout the project. On the idea of using a naïve Bayesian model based on co-expression I do still believe that this could be possible, however this study fails to address this in any meaningful way.