5461 Project

Calvin Roth and Noah Siem

May 2019

1 Abstract

A little-studied problem is how to associate genes with proteins when there is a notable but unknown lag time between the gene expression level and protein expression level. We develop a method that uses the time dependent pearson correlation to try and solve this issue in the Streptomyces Coelicolor bacteria. With the best parameters we got a classification accuracy of 60% which is much higher than what a random classifier would do(33%).

2 Introduction

Surges in informatic capabilities in the past decades have allowed for growth in the analysis of large biological datasets. These datasets include studies of the proteome, genome, and metabolome of common model organisms and have grown to include many more. These datasets have allowed for advances in gene function prediction, transcriptional network mapping, analysis of protein-protein interaction networks, and characterization of gene clusters. This study aims to assist in the characterization of secondary metabolite - natural product - assembly between multi-omic levels, i.e. from gene to protein to metabolite. This involves analysis of time course data that must be able to account for lag period in biomolecular production. The model organism of choice in this multi-omic study is Streptomyces coelicolor, a well-characterized bacterium being examined by the Carlson Lab in the Department of Chemistry at the University of Minnesota, Twin Cities. Streptomyces coelicolor is the natural producer for many antibiotics used in medicine. The pathways examined were biosynthetic gene clusters for actinorhodin, desferrioxamine B, and calcium dependent antibiotic. These were chosen because they were fairly populated gene and protein-wise as determined using Biocyc. Actinorhodin is an antibiotic synthesized from a polyketide backbone. Actinorhodin can be used as a pH indicator as it exhibits visible color change. Desferrioxamine B also has medicinal purpose in that it binds to iron and aluminum, so it is commonly used in treatment for iron overdose. calcium dependent antibiotic is an acidic lipopeptide. The goal of this study was to go beyond expression levels and examine a diverse set of data so that lag time in biomolecule production form the organisms transcript to proteins to metabolites. This means examining the time required for translation of mRNA into proteins and time for biosynthesis of metabolites from these proteins. The Carlson currently has a multi-omic data set consisting over time series - 9 day - transcriptome, proteome, and metabolome data

for Streptomyces coelicolor. The lab had begun to examine the correlation between protein abundance and metabolite production, noting that the correlation was not always clear and can include a time lag. The ability to incorporate multiple datasets especially with respect to time is an interesting problem that is relatively novel.

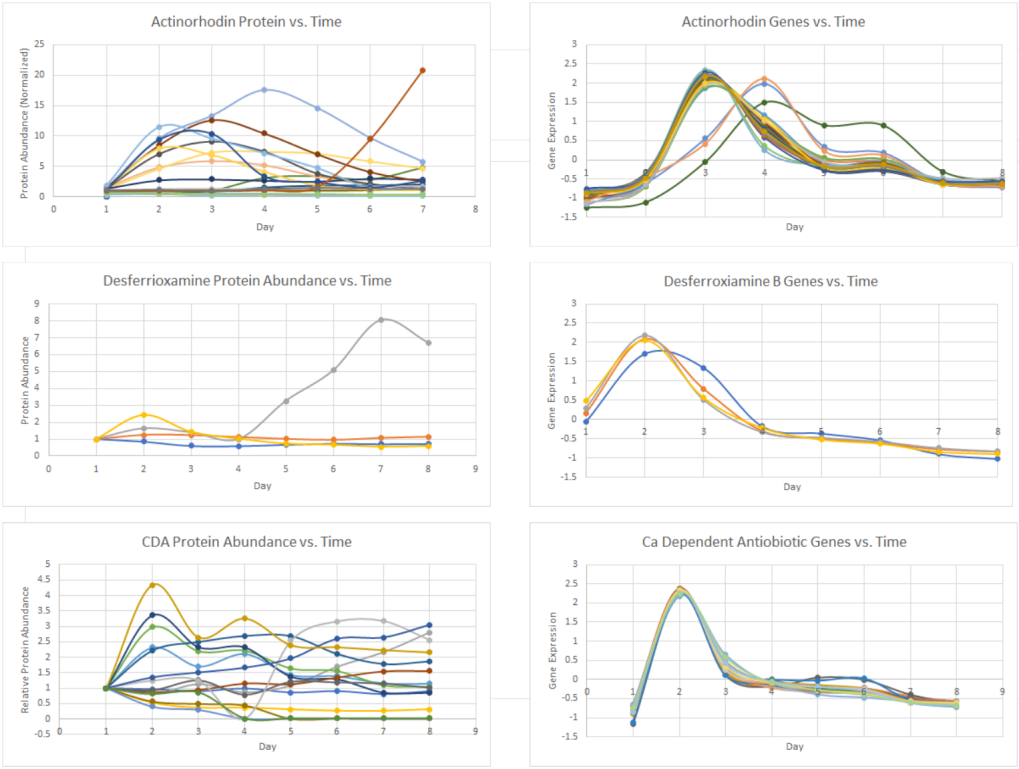
One challenge faced by the Carlson lab has been comparison of data across genes and proteins since they lack the computational tools to normalize this data. In the proteomic data, peptide levels were characterized in ratio the day one levels. While this allows for simplification of the time series data across its duration and condenses the data range, it does not guarantee relation across protein. In genetic expression levels from RNA-sequencing, there is natural variation of data due to uncontrollable variation between experiments. This required normalization between series to allow for comparison. A z-score standardization was used to normalize all transcriptome data. Assuming a data distribution - such as Gaussian - when normalizing data can bias results since it assumes a certain shape and curve. An important consideration for future examination of multi-omic time series data such as this is whether an alternate normalization would be more effective. Quantile normalization, for example, does not assume any data distribution and still arranges data so that it can be aptly compared across trials. Additionally, for the proteomic data, while normalizing the peptide levels based on ratios of the first days levels does allow for easier comparison, additional normalization could have been beneficial. The data was linked in the initial concentration of each peptide was set to 1 abundance unit. However, certain proteins - especially in the calcium dependent pathway - were expressed at much lower ratios than those of actinorhodin. This likely impact the quality of results obtained.

3 Data

One challenge faced by the Carlson lab has been comparison of data across genes and proteins since they lack the computational tools to normalize this data. In the proteomic data, peptide levels were characterized in ratio the day one levels. While this allows for simplification of the time series data across its duration and condenses the data range, it does not guarantee relation across protein. In genetic expression levels from RNA-sequencing, there is natural variation of data due to uncontrollable variation between experiments. This required normalization between series to allow for comparison. A z-score standardization was used to normalize all transcriptome data. Assuming a data distribution - such as Gaussian - when normalizing data can bias results since it assumes a certain shape and curve. An important consideration for future examination of multi-omic time series data such as this is whether an alternate normalization would be more effective. Quantile normalization, for example, does not assume any data distribution and still arranges data so that it can be aptly compared across trials. Additionally, for the proteomic data, while normalizing the peptide levels based on ratios of the first days levels does allow for easier comparison, additional normalization could have been beneficial. The data was linked in the initial concentration of each peptide was set to 1 abundance unit. However, certain proteins - especially in the calcium dependent pathway - were expressed at much lower ratios than those of actinorhodin. This likely impact the quality of results obtained.

The data consists of gene and protein expression data which was collected once a day for

9 days and was composed of 55 genes and 40 proteins split into the previously mentioned three group pairs. For comparability, the data analyzed(and given to us by Carlson lab) is the gene relative gene expression data and relative protein data. Relative meaning that the data value for a gene at a certain time is the ratio of the current level of expression compared to what the level was when measurements were first taken in such that if on day i the expression level for a gene went down relative value is negative and it will be positive if it went up compared to day 1. Furthermore the data is already clustered into three pairs of genes and proteins by function. The three functional groupings genes/proteins associated with the Actinorhodin protein, genes/proteins associated with Desferroxiamine B, and ones associated with Ca Dependent Antiobiotics.



To be clear, what entries are genes and what are proteins is clearly distinguishable. This labelling was also done by the Carlson Lab team. One issue with this data is that it is incomplete by which I mean sometimes a certain protein may for any number of technical reasons fail to get a valid reading for that entry and is not recorded. For have data of uniform length data which is useful for analysis we take the value of a reading like this to be the average of it's two adjacent readings. So if prot 1 on day 2 had an expression level of 2, day 3 had no recorded value, and day 4 had a level of 1 then we assign a value of 1.5 to day three for p. This is the most logical solution to fill in gaps. For reasons that will be explained in the methods it is helpful to use the limited expression data to generate a polynomial using a natural spline cubic interpolation.

4 Methods

The goal of this project is the development of a method that can look past the a time lag between gene expression and protein data. In real world data like this a protein may not change the moment the gene expression level changes. The way we test this to pool all the genes together in one set and all the proteins together in another with the goal of what for a given gene what proteins is it most associated with. In the ideal case a gene a certain group will be highly associated with proteins from the same functional group of that gene. If our model can accurately do this, we could use the same method to investigate understudied genes and proteins to find out what their function is and give a picture of what how biologic genes and proteins do they interact with. The method we employed is a reasonable attempt to associate different signals s_1, s_2 is one based on the time dependent Pearson correlation

$$p_n(s_1, s_2, \tau) = \frac{\sum_{t=0}^{n-1}(s_1(t) * s_2(t+\tau))}{n\sigma_x\sigma_y}$$

where n is the number of points we sample at, t is the time we are at, and σ_i is the standard deviation of the points of s_i that are being considered. Given a maximum lag time and the number of steps to take we build an array of corerlations between genes and proteins where the entry is the highest score obtained from any lag time in the acceptable range(We sampled from 0 to Max-Lag with 0.25 day increments). After finding the best correlations we can come up with for each gene we select the top 10 proteins for that gene. The accuracy for that gene is then the number of proteins that are from the same group as this gene divided by 10. And the total accuracy of one trial is the average accuracy of each each in the data set. This immediately presents two parameters that we could think about tuning: the number of sampled points and the lag. This motivates why we would like to interpolate our data. If we didn't do so the points would be required to be an integer and in the range $[0, 8)$ which gives the space of different lags you can consider is very small and rigid. But using an interpolation method this parameter space is now free to use any floating lag time we want. Proteomic data obtained was pre-normalized based on abundance levels on the first day of the experiment. All RNA-seq data that was obtained for genomic expression data was normalized in Excel using z-scores to fit each series to a mean of 0 and a standard deviation of 1. This standardizes the data for simplified comparison between independently measured data sets. Proteomic and genomic data was obtained from the Carlson lab. The data was pre-grouped into three sets based on pathway: actinorhodin, desferrioxamine B, and calcium dependent antibiotic. Proteomic data was gathered once a day over a nine day period and abundance was measured. The protein data was normalized in ratio to the protein levels of the first day to allow for comparison between entities. Additionally, gene expression data was gathered once a day for the same period of time.

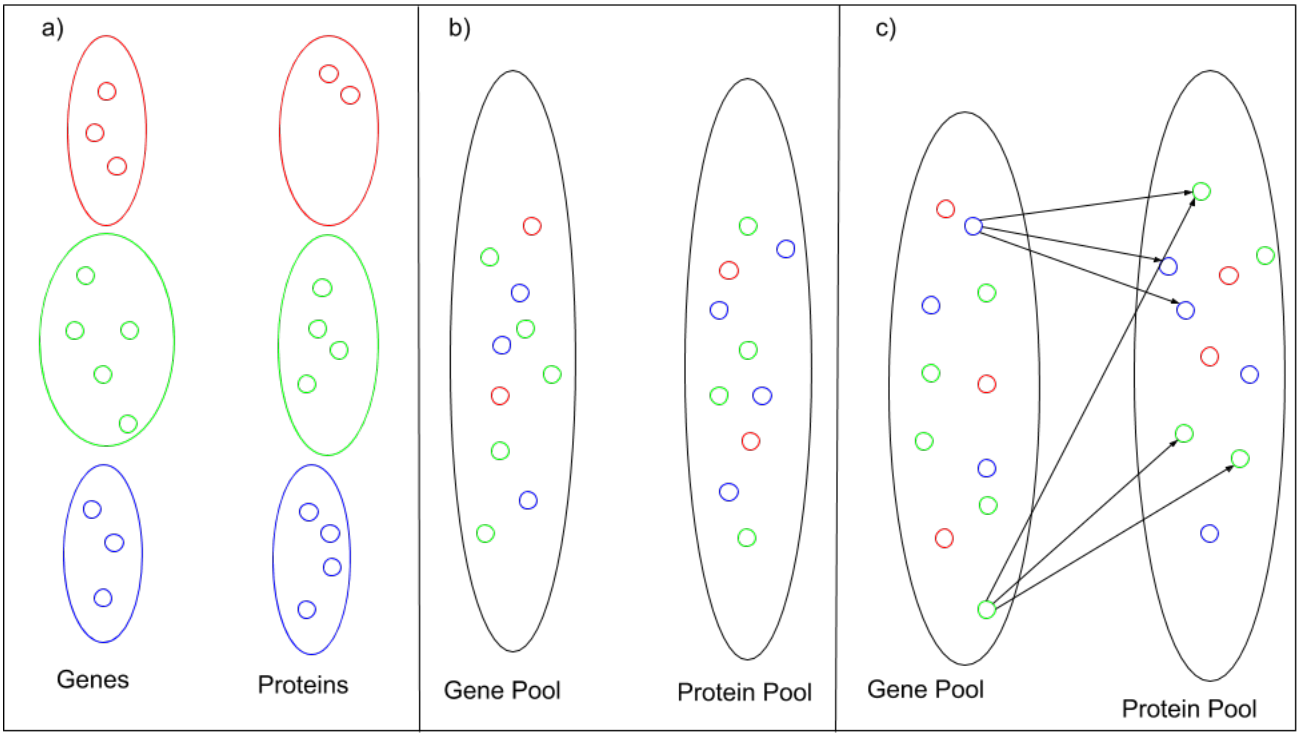
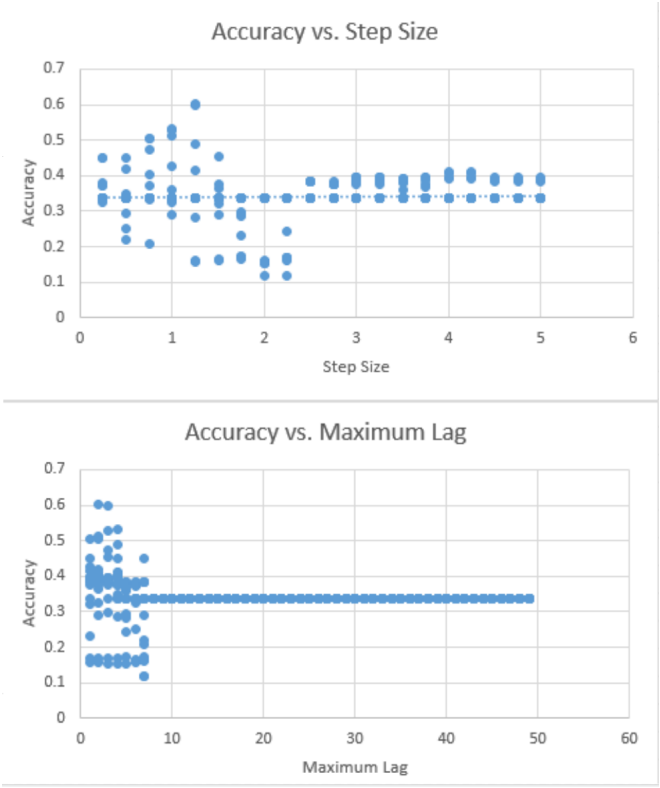


Figure 1: General overview of the steps this method. A) Genes and proteins start in the associated their known associated clusters with genes represented by the small circles on the left and proteins by small circle on the right. B) The genes are pooled together as well as the genes. C) Each gene finds with w.r.t. to any time lag what proteins are most similar to it. We then select the top k to be proteins to be used for error analysis

5 Results

We tested the accuracy of this set up for a wide variety of possible parameters. The parameters that we varied were the maximum lag time allowed in a trial and the number of samples we take of expression data. We found that the most accurate set up to choose is to sample every 1.25 days and allow a maximum lag of .5 days which identified the top 10 most similar proteins with an accuracy of 60%. This does conform to what you would think is a sensible configuration. If the step size is too large we are sample only a few points that are far apart but if we sample too often we may overfit the data. As for lag time the interesting thing to note is that allowing a large range of lag times than a more strict sampling of times actually hurts the results. The implication of this this is that allowing a big range of possible lag times is more likely to find unassociated peaks and valleys that give the pearson correlation a high score. The next best results all had similar parameters such 1.25 step size and 0.75 lag time for 2nd place and 3rd place was a set size of 1 and a lag time of one. This gives the impression that the best parameter wasn't just random noise but instead actually is a good setup for determining accuracy.



6 Discussion

Through this investigation we have arrived at several conclusions. First, our method to classify genes to proteins with an unknown time lag works reasonably well reaching an accuracy rate of as high of 60% for three possible choices. But to get this accuracy for unknown data some degree of prior knowledge of the system is required specifically a rough idea of how much time lag there is between proteins and genes. We suggest that an evaluation of this date for a much larger data set would give a better idea on how viable this method is in practice. Hopefully more data points gives us more accuracy. Aside from lack of data, the other key issue in the data is that in one group there can be a lot of variance which might be giving us our low accuracy scores. Both of these are interesting problems to study in the future.