Prospectus for Final Project: *E. Coli* Genomics Paper by Houser et al.

Spencer Woody, SDS 385, November 2, 2016

Hi James,

As we have discussed previously, my plan for the final project is to replicate the results found in the *E. Coli* written by Houser et al using R. My final deliverable will be an Rmd file which is a walkthrough of the paper's methods and contains replications of its summaries and visualizations. I also see potential to use some alternative methods. Houser et al. conducted their computational methods using Python with the SciPy library, and I think it will be useful to have a companion to the paper in the form of R Markdown for those unfamiliar with the methods or who want to replicate or reproduce the results on their own without having to learn Python. Dr. Claus Wilke mentioned another similar dataset that he's willing to offer to me, and I might do work on that as well, but my main focus will be on the paper's data.

In the paper, the authors measure RNA and protein levels of *E. coli* in a starved condition over a span of two weeks. Here's a summary of the genomic methods used in the paper

1. **Clustering of mRNA and protein profiles by relative presence over time**
   The authors perform k-means clustering on mRNA and profiles based on their relative presence at each measured time point. They find that mRNA may be classified into 15 clusters while proteins can be classified into 25, and they used visual inspection to make their choice of k (Note: mRNA and protein profiles are clustered *independently* of each other). From this they induce that mRNA is regulated "in a more uniform manner" compared to proteins. I may try to find a less arbitrary method of finding the optimal number of clusters, and a clustering method which takes the time series nature of the data into account.

2. **Correlations between individual mRNA and protein time courses**
   Next, the authors inspect correlation between relative levels of proteins and their corresponding transcript, using measurements at each time point to calculate correlation. Two sets of correlation histograms are produced: one which shows correlation between a protein's relative level and the *instantaneous* relative level of its transcript (high correlation here implies proportional regulation and fast protein degradation), and one which shows correlation between a protein's relative level and the *cumulative* (i.e., integral) relative level of its transcript (high correlation here implies integral regulation and slow protein degradation).

3. **Compare correlations of mRNAs and proteins located within same operon**
   The authors group together mRNAs and proteins by operon, and calculate correlations between mRNAs and proteins located within the same operon. Proteins within the same operon show weaker correlation with each other than mRNAs because genes within an operon are transribed at the same time, but proteins are regulared less uniformly, which is in line with point 1.

4. **Classification of expression-profile time courses into distinct categories**
   This is the core part of the paper. The authors fit each individual mRNA and protein to a piecewise continuous function (see panel A in the figure below) which has four free time parameters (t_1, t_2, t_3, and t_4) and three free amplitude parameters (the flat regions of the graph, a_1 before t_1, a_2 between t_2 and t_3, and a_3 after t_4). In order to see underlying regulation of cellular processes and mRNA, they then sort mRNA and protein profiles into five categories *based on the estimated amplitude parameters*. The five categories are up-regulated, down-regulated, transiently up-regulated,

transiently down-regulated, and ambiguous. Then term enrichment is performed via gene ontology (GO) on those mRNAs and proteins which are classified as down- or up-regulated to find the associated cellular processes of these components. So there are three steps in this method: fitting, classification, and GO enrichment. Down-regulated components are used in energy-intensive processes, while up-regulated components are used in stress response.

$$\text{category} = \begin{cases} \text{up-regulated} & \text{if } a_1 < a_2 \leq a_3 \\ \text{down-regulated} & \text{if } a_1 > a_2 \geq a_3 \\ \text{transiently up-regulated} & \text{if } a_1 < a_2,\ a_2 > a_3 \\ \text{transiently down-regulated} & \text{if } a_1 > a_2,\ a_2 < a_3 \\ \text{ambiguous} & \text{otherwise} \end{cases} \tag{1}$$

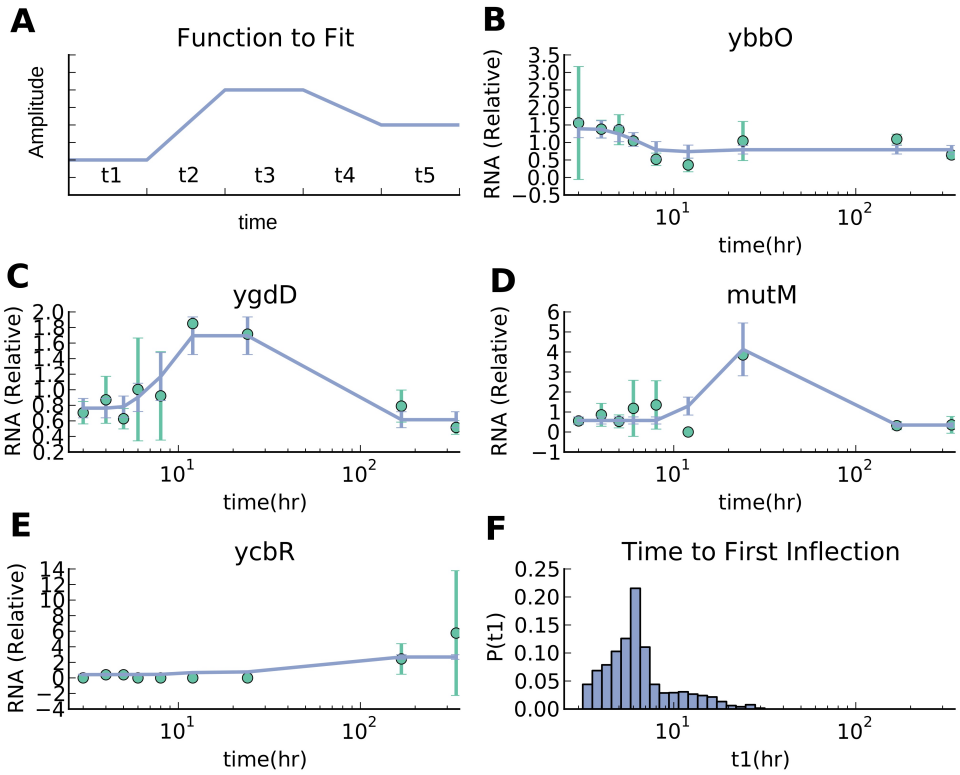

Figure 1: Piecewise continuous function used in profile assignment

Link to paper.

MLA Citation

Houser, John R., Craig Barnhart, Daniel R. Boutz, Sean M. Carroll, Aurko Dasgupta, Joshua K. Michener, Brittany D. Needham, Ophelia Papoulas, Viswanadham Sridhara, Dariya K. Sydykova, Christopher J. Marx, M. Stephen Trent, Jeffrey E. Barrick, Edward M. Marcotte, and Claus O. Wilke. "Controlled Measurement and Comparative Analysis of Cellular Components in E. Coli Reveals Broad Regulatory Changes in Response to Glucose Starvation." *PLOS Computational Biology* 11.8 (2015): n. pag. Web.