ECS272 Project Proposal

Camille Welcher

February 7, 2016

A Novel Transcriptome Visualization Method

Introduction

As next-generation sequences technology has become increasingly cheap and accessible, many labs are opting to use sequencing technology for their projects. For groups looking to quickly and cheaply produce a working profile of genes in an organism, RNA-seq has become a favored method. An RNA-seq experiment produces hundreds of millions of short sequence fragments, "reads," from the transcribed RNA of an organism. These reads are then assembled into transcripts, full length sequences which represent full genes. These transcripts would eventually be translated into proteins, meaning that their sequence content and abundances are valuable to the understanding of the function of an organism. Characterizing a set of transcripts, or a "transcriptome," is a continuing field of research: actually assembling the reads into full length transcripts, estimating the expression level of those transcripts, and annotating them by identifying the names, structure, and function of those transcripts being major topics. While significant progress has been made in all these pursuits, a key item missing is the ability to rapidly summarize and compare entire transcriptomes. With many studies being published every day with newly assembled and annotated transcriptomes, this absence becomes increasingly clear.

I propose a method to easily compare the quality of annotations, and less directly the quality of assembly, of transcriptomes. This method will combine established techniques from phylogenetics with known visualization methods. The result will be a method to produce easily digestible visual and quantitative summaries of assembled transcriptomes.

Visualization Method

Basic Summary Visualization

The main visualization produced from this technique will use existing phylogenetic trees to help researchers assess their transcriptomes. The most direct way to do this is to map gene annotation identifiers onto their nodes in a known taxonomy, provided by the NCBI (Sayers et al. 2009) or another another curated source. A sunburst partition map can then be built showing the the number of annotations in each subgroup; if many clades unrelated to the origin species are over represented, that suggests contamination or an otherwise poor annotation. A mockup of such a chart drawn using d3.js (Bostock, Ogievetsky, and Heer 2011) is shown in Figure 1.

Phylogenetic Signal Visualization

While the simple method described above might provide an initial summary, it does not provide any statistically relevant quantitative results, and will not be normalized for evolutionary effects. To alleviate these concerns, we can bring in methods traditionally used to gauge evolutionary relationships. One of these is a measurement called phylogenetic signal, which is defined by (Revell, Harmon, and Collar 2008) as "...a measure of the statistical dependence among species' trait values due to their phylogenetic relationships." Here we can use the presence of annotated genes as the trait values to compute phylogenetic signal and visualize the trees using the resulting signal; this also enable quantitative comparison between transcriptomes generated from the same and different species.

Annotation Explorer

This method will be implemented within an existing annotation program maintained by the author called dammit (Scott 2016) which produces all gene alignment and prediction annotations in an easily accessible format. Other than the main summary visualization, the project will aim to produce a multi-paned annotation explorer to navigate through the tens of thousands of annotations produced from a single project. This will have at least four panes:

1. The summary visualization pane, described above;
2. One to view subtrees of the summary and view metadata such as the gene identifiers and signal;
3. One to view gene models and alignments, built on the GenomeD3Plot (Laird, Langille, and Brinkman 2015);
4. One to view full gene information and metadata, using the API provided by mygene.info(Wu, MacLeod, and Su 2013).

Figure 1: *Figure 1:* An example of a potential plot. This one simply shows the entire NCBI taxonomy database.

Because this explorer will be built on web technologies, it will be able to be used either locally on a researcher's computer or set up as a server. Further, the installation of dammit implies that other dependent software is already available, which makes it straight-forward to include an interface for aligning arbitrary sequences to the transcriptome and view the resulting annotated transcripts. A simple search interface over gene names will also be provided.

Outline of Work

Several components of the summary visualization have already been completed by the author. Further work will require a more complete review of methods for phylogenetic signal; collaboration and discussion between the author and Jonathan Eisen's lab is ongoing, providing needed expertise in phylogenetics. The method will need to be evaluated on a number of known high-quality transcriptomes, and eventually showcased on a larger collection of transcriptomes which will be provided by the author's lab. The webserver backend will be coded in a python, while the interface will be HTML and Javascript.

References

Bostock, Michael, Vadim Ogievetsky, and Jeffrey Heer. 2011. "D3: Data-Driven Documents." *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)*. http://vis.stanford.edu/papers/d3.

Laird, Matthew R., Morgan G.I. Langille, and Fiona S.L. Brinkman. 2015. "GenomeD3Plot: A Library for Rich, Interactive Visualizations of Genomic Data in Web Applications." *Bioinformatics* 31 (20): 3348–49. doi:10.1093/bioinformatics/btv376.

Revell, Liam, Luke Harmon, and David Collar. 2008. "Phylogenetic Signal, Evolutionary Process, and Rate." *Systematic Biology* 57 (4): 591–601. doi:10.1080/10635150802302427.

Sayers, E. W., T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, et al. 2009. "Database Resources of the National Center for Biotechnology Information." *Nucleic Acids Research* 37 (Database): D5–D15. doi:10.1093/nar/gkn741.

Scott, Camille. 2016. "dammit: An Open and Accessible de Novo Transcriptome Annotator." *In Prep.*

Wu, Chunlei, Ian MacLeod, and Andrew I. Su. 2013. "BioGPS and MyGene.info: Organizing Online, Gene-Centric Information." *Nucleic Acids Research* 41 (D1): D561–65. doi:10.1093/nar/gks1114.