ABSTRACT

INVESTIGATING COMPLEXITY IN TRANSCRIPTOME EXPRESSION, REGULATION, AND EVOLUTION USING MATHEMATICAL MODELING

By

Nicholas Louis Panchy

To date, gene expression has been characterized in over one thousand species across more than a million experimental conditions. With this wealth of data, it is possible to investigate the role that differential expression has in key biological processes, such as development, stress response, and cell division. However, the complexity of the transcriptome makes the analysis of expression challenging, as a single genome can contain thousands of genes as well as millions of potential regulatory interactions shaped by more than a billion years of evolution. To address this complexity, we can use the language of mathematics to create models of gene expression, regulation, and evolution that define the system in a testable format. In the following chapters, I will present research that applies mathematical modeling to the identification and regulation of cyclically expressed genes as well as the evolution of transcriptional regulators following whole genome duplication.

Cyclically expressed genes were studied in two systems. First, I investigated day-night cycling or 'diel' genes in Chlamydomonas reinhardtii. Diel genes were identified de novo using two models of cyclic expression that jointly classified half of all genes in C. reinhardtii as diel expressed. To understand the regulation of diel expression, I clustered diel genes according their peak of expression, or 'phase', and searched for cis-regulatory elements enriched (CREs) in the promoters of each cluster. While I found putative CREs corresponding to each cluster, using these CREs to predict diel expression using machine learning performed poorly compared to previous models of expression regulation. Therefore, I changed systems to Saccharomyces cerevisiae and studied cyclic expression during the cell cycle. Here, I applied machine learning models to predict cell-cycle expression using regulatory interactions from four different data sets. These models out performed the previous model of cyclic expression when using regulatory interactions defined by chromatin-immunoprecipitation, transcription factor knockout

experiments, and position weight matrices. Further gains in performance were obtained by combing interactions across data sets and using co-regulation by pairs of regulators involved in feed-forward loops. The most important interactions for predicting cell-cycle expression included not only known cell-cycle regulators but also two groups of transcription factors not previously identified as being involved in cell-cycle regulation.

The evolution of transcriptional regulation was studied in Arabidopsis thaliana, which has undergone several rounds of whole genome duplication (WGD), after which transcriptions factors (TFs) are preferentially retained. Here, I applied maximum likelihood estimation to infer the most likely ancestral expression and regulatory state of pairs of duplicate TFs prior to WGD. Comparing this ancestral state to the existing TF duplicates, I found that one duplicate, the "ancestral' copy, tends to retain the majority of ancestral expression state and CREs, while the other 'non-ancestral' copy loses ancestral expression and CREs, but also gains novel CREs instead. Modeling the evolution of TFs pairs using as system of ordinary differential equations, I demonstrated that the partitioning of ancestral states amongst duplicates is not random, but occurs because the loss of ancestral expression occurs orders of magnitude faster in the first copy than in the second. This suggests that TFs duplicate pairs are preferentially maintained such that one copy is 'ancestral' and the other is not. Taken as a whole, the research in this dissertation demonstrates how mathematical modeling can be applied to studying the expression, regulation and evolution of the transcriptome.

Dedicated to…

…my family at home

…my family in the Shiu Lab

…my family from Providence

Thank you all for your love, support, and patience