Modeling quantitative cell-type specific regulatory networks from public, large data compendiums

Aviv Madar[1], Maria Ciofani[1], Kieran Mace[1], Ashish Agarwal[1], Carolina Galan[1], Kim Newberry[2], Richard M. Myers[2], Richard Bonneau[1] and Dan R. Littman[1,3]

Publically available genome-wide datasets (such as RNAseq/Microarrays and ChIPseq) become increasingly available. These datasets contain quantitative information relating to a plethora of biological samples, and have been used in the past to quantitatively model genome-wide regulatory networks. Yet it remains unclear how to use these datasets to model regulatory networks that are active or relevant for a specific biological sample. Here we show that by combinatorial consideration of regulatory interactions (for TFs that are active in the sample of interest) it is possible to learn highly accurate sample-specific regulatory models, even in the absence of sample specific public data.

I. Problem setup

In the past decade and years to come the efforts of many research groups will generate a plethora of genome-wide quantitative measurements of many key processes in biology (e.g. Transcriptome, Metabolome, Proteome, etc.). Most of these datasets are or will soon become publically available[1]. These datasets contain much more information than any one lab can hope to generate alone, even given that each lab has a specific focus (say the lab is interested in a given cell type). This is true as often interactions between biological components happen across cell types, but in different combinations at any given cell type. For example a TF can be expressed in many cell types, but its combinatorial regulatory interactions with other TFs determine which genes will be expressed in any given cell type. Thus, it will be highly desirable to take advantage of the information that is available in the public domain. The problem we address in this talk is how to use public data (here the mouse immune cells Immgen[2] Microarray dataset compendium), which for our example contained no specific information for the immune cell type we were interested in (T helper 17), to learn highly accurate quantitative regulatory network models. We also compare this publically derived network to a quantitative network model that we derived using over a hundred ChIP-seq and RNAseq samples that were specifically tailored to learn the true regulatory network responsible for T helper 17 specification[3].

II. Preparation of Abstracts

Using 10 fold cross-validation to estimate the percent of explained variance in gene expression, and focusing on genes that are differentially expressed in the cell type of interest from a progenitor cell, we show that in general we can learn predictive quantitative regulatory circuits for most genes both for the public and specific datasets, presumably because there is signal for this genes in the specific dataset (as they were chosen to be differentially expressed), and in the public dataset (as they may be differentially expressed across several other cell types). However, for the highly specific T helper 17 genes we can only learn predictive regulatory models from our specific dataset. We also show that the regulatory circuits learned for each TF over either the public or specific dataset have little similarity, presumably because one is a conglomerate of regulatory networks of many cell types (non of which is the one we care for), and the other contains only information for the cell type of interest. However, when learning regulatory circuits that integrate the input of several TFs (key T helper 17 TFs) the similarity between the public and specific network becomes highly significant, presumably because the T helper 17 specific regulatory interactions existed in the public data, but were "buried" under the universe of other possible regulatory interactions each TF has. These cell type specific regulatory interactions became apparent when modeling the integrated regulatory response, as many actual biological systems use a strategy of combinatorial input to determine how to actuate a reponse (particularly for gene expression).

III. Conclusion

The results we present here are significant as they show that with little prior knowledge (e.g. knowing about several TFs that are relevant for your biological sample of interest) one can harvest a public dataset to get highly informative and quite accurate sample-specific regulatory models at a modest cost of time and money.

[1] Molecular Pathogenesis Program and Courant Institute for Mathematical Sciences, New York University, NY 10003. E-mail: madaraviv@nyu.edu

[2] HudsonAlpha Institute for Biotechnology, Huntsville, AL 35806.

[3] Howard Hughes Medical Institute, New York University, New York, NY 10016

References

[1] Tanya B et al. (2011) "NCBI GEO: archive for functional genomics data sets—10 years on". Nucleic Acids Res.

[2] Tracy SPH et al (2008) "The Immunological Genome Project: networks of gene expression in immune cells," Nature Immunology.

[3] Ciofani et al. "The transcription factor network regulating Th17 lineage specification and function" (2012) in preparation.