

Structured modeling of transcription networks: computational and experimental applications in mammalian systems biology

Daniel E. Zak

University of Delaware, Department of Chemical Engineering

Newark, DE 19713

zak@che.udel.edu

<http://www.che.udel.edu/systems/people/zak>

Chemical engineering for many years has been intimately connected to the biological sciences, from the tradition of biochemical engineering, to the increasing involvement of chemical engineers in *systems biology*. In systems biology, the objective is to understand how all the *parts* of a biological process function together as an integrated *system*, understanding that cannot be obtained by considering the parts alone (Kitano, 2002). Knowing how biological systems function as a whole will lead to intelligent drug design and other practical applications. Given their arsenal of tools from kinetics, transport, and systems engineering, it is not surprising that chemical engineers can approach systems biology problems in a productive way. My research is focused on the unraveling and dynamic modeling of regulatory networks that underlie the fundamental ability of mammalian cells to remodel themselves in response to cues from their local environments. Since cellular 'remodeling' largely is accomplished through variation in gene 'activities' or *expression levels*, and since regulation of gene expression is largely accomplished through regulation of transcription, I am focusing on transcription networks. The specific 'cues' I have considered are *extracellular ligands* (hormones, growth factors, neurotransmitters) that bind to membrane-localized receptors and initiate complex intracellular signaling reactions that culminate in changes in gene expression. My research in this area has involved computational modeling and identification of biological systems, analysis of biological datasets, application of statistical approaches, and experimental techniques to support and validate the computational work.

The *omics* revolution that began with the genome sequencing projects (*genomics*), and has led to system-wide measures of gene expression (*transcriptomics*) and protein activities (*proteomics*), has created new opportunities for the study of biology at the systems level. My research with mammalian transcription networks makes extensive use of omics datasets, particularly genomics and transcriptomics (*microarrays*). A growing literature describes approaches for unraveling transcription networks from microarray data alone (Brazhnik et al., 2002). While ambitious, these attempts suffer from the limited quality and quantity of microarray data, and the sheer complexity of the underlying systems. In a number of simulation studies using an idealized regulatory network, I have shown that typical microarray data is generally not informative enough to reveal the underlying networks on its own (Zak et al., 2001; Zak et al., 2003a; Zak et al., 2003b). Because of these limitations of microarray data, my colleagues and I have developed a *structured* approach to inferring system-wide transcriptional networks that incorporates multiple omics datasets into a dynamic modeling framework (Zak et al., *in press*).

Our structured approach to inferring transcription networks involves a tight integration of biological knowledge, bioinformatics, and statistical and systems engineering techniques. For

example, it is obvious biologically that only a subset of all genes, the transcription factors, can regulate other genes. Yet, given the scale of omics data, it is not always straightforward to identify which genes are transcription factors, and it becomes necessary to work with bioinformatics tools and databases, and to develop new tools to perform automated literature searches, to identify the potential regulators. Similarly, it is biologically plausible that a group of genes will be regulated similarly during a particular response, and thus gene groups can be searched for over-represented regulatory elements as compared to random gene groups of the same size. Searching for over-represented regulatory elements, however, requires integration of several data types (regulatory sequences for genes, databases of regulatory elements), and testing groups of genes for statistically significant enrichment. This is a more complex task than direct model identification, and is facilitated by bioinformatics tools developed by our group (Vadigepalli et al., 2003) and others. On the other hand, a systems engineering perspective improves these biologically inspired approaches. For example, it may seem reasonable biologically that similarly regulated genes will behave similarly over time. Considering the 100-fold gene-gene variability in both gene-specific time constants and gene-specific delay times, however, it is clear from a systems engineering perspective that genes with the same *input* (regulation) may have very different *outputs* (gene expression levels). We have been developing techniques that allow the principled grouping of genes on the basis of similarity of regulation, rather than similarity of expression. The approaches noted above do not eliminate the need for model identification, however, because they can only be used to determine network structures. To determine the *functional* nature of the interactions in the networks (i.e., activation, repression, etc), model identification approaches are necessary, and we have been developing methods that are well-suited to biological data (Zak et al., 2003b).

While an enormous amount of suitable data for the structured identification of transcription networks is available publicly, use of public data alone allows for limited validation of predictions, and constrains both the experimental system and the experimental methods to those defined by other researchers. For this reason, in parallel to my development of the structured framework, I have been actively collecting my own gene expression data for the response of mammalian cells to ligand inputs, as described above. My experimental work has also involved validation of gene expression predictions (RT-qPCR) and validation of predictions of transcription factor activities. I have focused on systems where the intracellular signaling pathways are fairly well-characterized, with the objective of understanding the complex intracellular signaling pathways in terms of their functional transcriptional outputs, to give physiological context to the molecular details.

Acknowledgments

I thank my advisors Babatunde A. Ogunnaike and James S. Schwaber for their support and guidance. I also thank the members of the Ogunnaike research group (University of Delaware) and members of the Daniel Baugh Institute for Functional Genomics and Computational Biology (Thomas Jefferson University) for their support and discussions. Finally, I thank the University of Delaware Department of Chemical Engineering for funding.

References

- Brazhnik et al. (2002) Trends Biotechnol. 20(11):467-72.
- Kitano H. (2002) Science. 295(5560):1662-4
- Vadigepalli et al. (2003) Omics. 7(3): 235-52.
- Zak DE et al. (2001) Proc. 2nd Int. Conf. Systems Biology. 231-238.
- Zak DE et al. (2003a) Genome Res. 13(11):2396-405.
- Zak DE et al. (2003b) Omics. 7(4):373-86.
- Zak DE et al. (*In press*) Computers and Chemical Engineering.