Growth Processes for Synthetic Regulatory Networks

Caitlin McCollister
Department of Electrical Engineering & Computer Science
University of Kansas
cemccoll@ku.edu

Introduction

I arrived at the Bioinformatics department at University of North Carolina, Charlotte, in June 2010 as a participant in the Distributed Research Experiences for Undergraduates program. With no previous direct experience in bioinformatics and limited background knowledge of genetics, I had done some background reading before I arrived, which left me with some perplexed questions about the nature of gene expression.

If living organisms worked from just a static library of genetic code executed like a computer program, how would they progress through the various phases of life, or resist changes from a dynamic environment? There must be more dynamic mechanisms at work.

Borrowing possibly analogous terms from process control field, is there some mechanism for coordinated self-monitoring of the overall system—the organism? Do the cells in one type of tissue or organ "detect" when the system state has changed, and "signal" others to respond in certain ways? Perhaps every cell, tissue, and organ is an independently acting "agent" and the apparent "system behavior" is actually an emergent property?

How have these systems come into existence? How long does it take for beneficial changes to dominate a population, or for less useful characteristics to fall out of the genetic record? For a given gene, is it possible to estimate the extent of its influence? Is it possible to infer backwards which genes could be deliberately modified to cause a desired change in the organism?

Background

I started my work by reading several academic papers recommended by my research mentor so we could discuss them together. Given my interest in applications of graph theory and probability, we tended toward graph-theoretic models and explanations of genetic processes.

"Stochastic mechanisms in gene expression" by McAdams and Arkin [1] was a relatively early publication in the domain of genetic regulatory networks. The authors present gene expression as an ongoing chemical process inside the cell. Certain genes called promoters contain the code to produce signal proteins that encourage the transcription of specific target genes. Because the signal protein has to travel some distance to reach its target site and faces the possibility of degradation along the way, there is a variable delay for enough of the signal protein to accumulate at the target site and take effect. After the target gene does respond to the promoter signal, the resulting transcript from that section of DNA either will be successfully translated or will degrade before it ever encounters a ribosome. The extent to which a target gene is being actively expressed at any moment is due to the concentration of its signaling proteins, a cumulative result of variable production and degradation rates over time.

The “Genomic Analysis. . .” paper by Jothi et al [2], published years later, is representative of the currently dominant research interests involving gene regulatory networks. Having reached a consensus on the small-scale processes that actually constitute gene expression, such as those described by McAdams and Arkin, the same research community is now concerned with the heavily intertwined logical relationships among transcription factors and target genes. Most research groups have adopted a simplified Boolean representation of regulatory relationships in their models rather than time-dependent measurements of protein concentrations, apparently due to computational challenges.

The authors of that paper in particular present a selection of genes from the yeast genome as the nodes of a graph whose edges indicate the presence of a regulatory relationship. The authors define a procedure in which they partition the transcription factors in the yeast genome into a three-layer hierarchy. However, they make the nature of their study clear: it is not their end goal to demonstrate that it is possible to classify the nodes into several distinct layers based on incoming and outgoing links, which they claim can likely be done for nearly any graph of similar size. Rather, they emphasize that there is a correlation between a gene’s position in the computed hierarchy and the variability of its expression level across individual members of a population.

Genes in one layer of the yeast regulatory network take on a wide range of expression levels over time and across different individuals, even in clonal populations. This suggests that there is a distinct set of genes within the genome that are not as tightly regulated as they could be, despite an ample period of evolutionary opportunity. However, it appears that strongly determined sets of biological traits are not necessarily beneficial to a species. Spontaneous variations in physical traits, not dependent on genetics, could allow enough individuals to survive and repopulate, even when environmental challenges arise faster than the species could be expected to evolve in response.

Current State of the Field

A reasonable next step would be to gather more experimental data to investigate the specific regulatory interactions of some well-studied species. By gathering high resolution time series data of the expression levels for a few candidate genes, we could hopefully detect which genes in a grouping activated first, which might suggest a causal relationship. To investigate the validity of those apparent directed relationships, deliberate knock-out studies allow us to see how one process affects another while other factors remain controlled.

Results from such a research workflow are already abundant in the literature. It is still a perplexing task, however, to connect features of the network with actual biological traits. Some of these mutations affect seemingly unrelated systems in the organism and have lethal results. In other cases, rather large mutations, even lengthy duplications or deletions, occur without any detected change at all.

What has become clear so far is the existence of certain network motifs: abstract patterns describing the number and direction of links within a small subgraph, whatever specific genes constitute the nodes in any particular instance of the motif. A few distinct motifs occur far more often than we would expect from random chance, suggesting an underlying modular structure in the genetic code. Furthermore, motifs may be found connected to or embedded within other motifs, resembling the nested logic gates in a digital electronic circuit.

What do we stand to gain from further study of genomics in the framework of graphs and networks? If we find it is impossible to specify an evolving network process that produces networks consistent with experimentally observed ones, we would need to reconsider the entire network model. On the other hand, if we can show by example that the expression data we see in the lab can be generated under a framework whose mechanisms are entirely specified by us, this would encourage further research into those mechanisms.

Research Aspirations

Given computer systems with sufficient processing power and storage capacity, we could try to simulate the accumulation of many years' worth of random mutations. By operating on networks that have been artificially assembled, either from scratch or from some recombination of real-world networks, we may gain some insight into the natural evolution and possible artificial manipulation of regulatory networks.

Numerous techniques for generating the synthetic datasets have been proposed and prototyped, ranging from near-random processes to well-specified sets of transformation rules. Certainly the objective is to generate networks that are as true as possible to their biological counterparts. They should also be varied enough, relative to each other, that we get "interesting" results.

One might try by starting with just one node and adding edges and nodes iteratively until some threshold is reached. Alternatively, one could take advantage of available datasets and construct new graphs by selecting and combining subnetworks to obtain roughly the desired number of nodes or edges.

What is an appropriate metric by which to judge the quality of synthetic biological networks? Intuitively, the network's degree distribution would be an obvious place to look. However, the degree distributions observed in biological networks are especially difficult to imitate directly through algorithmic processes.

Adami and Hintze outline a parameterized method for growing synthetic networks which they claim, with the right choice of parameter values, can generate networks with nearly arbitrary degree distribution [3]. I have written a basic implementation of this algorithm using the software package Mathematica.

Algorithm

The method begins by creating a single-node undirected network and then repeatedly runs a growth step in which each of three types of events may occur or not occur. Node events, edge events and duplication/fusion events have corresponding, independent probabilities that they will occur in a given iteration. These three probabilities are configurable parameters.

Three more parameters specify the conditional probabilities for outcomes of these events. For node events, a node is added to the network with probability p and a node is removed with probability $(1 - p)$. For edge events, the procedure attempts to add an edge with probability q, and attempts to remove one with probability $(1 - q)$. Edges, unlike nodes, can only be added if an unconnected pair of nodes is found, and can only be removed if an edge already exists between two nodes.

In a duplication/fusion event, there is a probability r that a node is selected and duplicated, which also creates edges between the new node and all nodes to which the original was connected. There is also the probability $(1 - r)$ for a node fusion, in which two selected nodes become a single node with edges to all nodes to which either of the originals had connections.

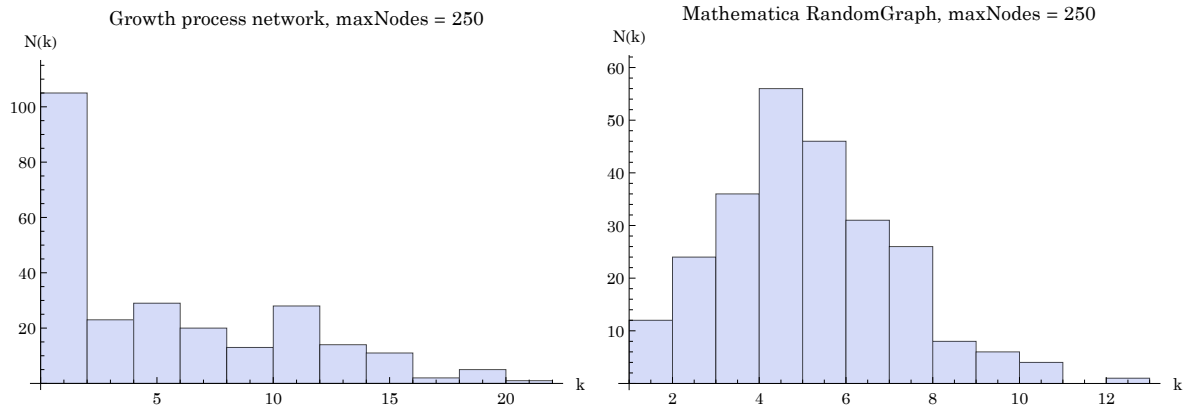
Implementation

My goal was to investigate the variety of network degree distributions I could generate compared to graphs generated using Mathematica's RandomGraph function [4]. The RandomGraph function's parameters are the desired number of nodes and an edge probability with which to decide whether to place an edge in each possible location. I calculated the edge probability for the random graph after I generated the growth process network in order to obtain a graph with a similar number of nodes and edges:

$$\text{randEdgeProb} = \frac{2 * \text{myGraphEdges}}{\text{myGraphNodes}\,(\text{myGraphNodes} - 1)}$$

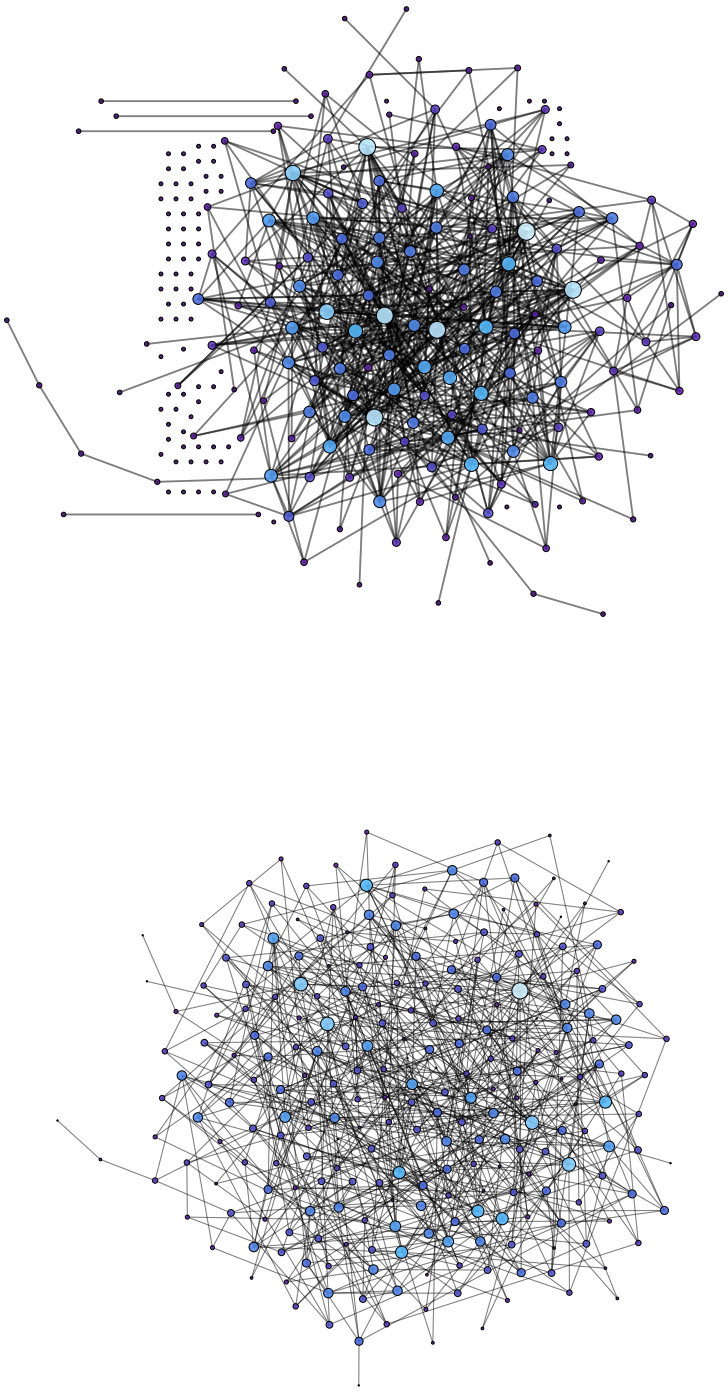
The variable myGraphEdges represents the actual number of edges in the growth process network. The number of edges in a complete graph of n vertices is $n(n - 1)/2$, so to obtain the edge probability to use in the random graph, we divide myGraphEdges by that quantity.

I generated histogram plots of the degree distributions in the growth process network and the random network. The random network displays an approximately binomial distribution, which is not surprising given the nature of its creation.



The degree distribution of the growth process network does vary significantly depending on the choice of values for the constant probabilities used in the algorithm. I chose a set of values from the many that were suggested in [3], which has resulted in a much different degree distribution. If we were to carry out many trials with these parameters and analyze some sort of averaged distribution, I suspect it could be consistent with a negative binomial distribution.

The full source code for the Mathematica notebook can be found on the website along with sample output from one run of the program.



Mathematica plots of the growth process network (top) and the random network (bottom). Larger and lighter colored nodes have a proportionally higher degree than smaller, darker colored nodes.

Conclusion

When we view the genetic regulatory system as a network of promoters and targets, it is easy to see how certain individual mutations could accumulate over time and drive the evolution of an entire species. We are only now beginning to appreciate the subtleties of modular regulatory network architectures. Rather than standing perplexed by a lack of correspondence between genome size and "organism complexity" in the natural world, it may be time for human scientists to give some of our fellow life forms a bit more credit. Even tiny organisms as common as brewer's yeast look quite complex indeed when we acknowledge the significance of long-term species adaptability and resilience.

References

[1] McAdams HH, Arkin A. Stochastic mechanisms in gene expression. Proc Natl Acad Sci U S A. 1997 Feb 4;94(3):814-9.

[2] Jothi R, Balaji S, Wuster A, Grochow JA, Gsponer J, Przytycka TM, Aravind L, Babu MM. Genomic analysis reveals a tight link between transcription factor dynamics and regulatory network architecture. Mol Syst Biol. 2009;5:294.

[3] Hintze A, Adami C. Modularity and anti-modularity in networks with arbitrary degree distribution. Biol Direct. 2010 May 6;5:32.

[4] Wolfram Research, Inc. "Combinatorica Package Tutorial." http://reference.wolfram.com/mathematica/Combinatorica/tutorial/Combinatorica.html