

Evolution and Statistics of Biological Regulatory Networks

by

Juhi Kiran Chandalia

Submitted to the Department of Physics
in partial fulfillment of the requirements for the degree of

Master of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2005

© Massachusetts Institute of Technology 2005. All rights reserved.

Author *Juhi Kiran Chandalia*

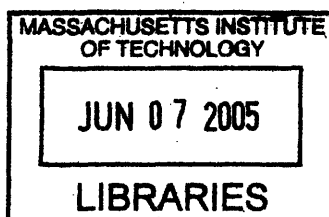
Department of Physics
May 20, 2005

Certified by *[Signature]*

Leonid Mirny
Assistant Professor
Thesis Supervisor

Accepted by *[Signature]*

Thomas J. Greytak
Professor of Physics
Associate Department Head for Education



ARCHIVES

Evolution and Statistics of Biological Regulatory Networks

by

Juhi Kiran Chandalia

Submitted to the Department of Physics
on May 20, 2005, in partial fulfillment of the
requirements for the degree of
Master of Science

Abstract

In this thesis, I study the process of evolution of the gene regulatory network in *Escherichia coli*. First, I characterize the portion of the network that has been documented, and then I simulate growth of the network. In this study, I assume that the network evolves by gene duplication and divergence. Initially, the duplicated gene will retain its old interactions. As the gene accumulates mutations, it gains new interactions and may or may not lose the old interactions. I investigate evidence for the duplication-divergence model by looking at the homology and regulatory networks in *E. coli* and propose a simple duplication-divergence model for growth. The results show that this simple model cannot fully account for the complexity in the real network fragment as measured by conventional metrics.

Thesis Supervisor: Leonid Mirny
Title: Assistant Professor

Acknowledgments

First, I would like to thank Professor Leonid Mirny for his support, understanding and wealth of ideas for research directions. I wish him luck in his future career. Thank you also to Professors van Oudenaarden and Kardar for educating me about the interface between physics and biology.

I have appreciated the camaraderie of my group members. Thank you Dr. Michael Slutsky for advice about physics and life and for just listening now and then. I had a great time in Israel; thank you for showing me around. Dr. Grisha Kolesov, thank you so much for all of your computer help, the late night company and the edits in this document. I really appreciate it. Dr. Victor Spirin, I enjoyed your company on the weekends. Thanks for the breadsticks and jam and tea with lemon. Ivan Adzhubey, thank you for maintaining our computers. And thanks especially for sacrificing your weekend to get the systems up and running at a critical point for me. Vincent Berube and Carlos Gomez-Uribe, thank you for making the lab a more fun and lively place! Last but not least, Zeba Wunderlich, thank you for keeping us clean, organized and fed. I appreciated the sympathetic ear and biological expertise. Good luck with the rest of your PhD.

Thank you Avni, Ateev, Aparna and Rahul for providing a home away from home.

Mom and Dad, thank you for supporting me even when you did not know how. Dad, thank you for showing the kind of strength and resilience I would like to have.

Finally, thank you to my second family. You guys have loved and supported me through some clutch periods in my life: Dr. Mihai Ibanescu, Hassan Ahmad, David Hu, Annie Lo, Junlin Ho, Karen Robinson, and Dr. Ozlem Uzuner. Thanks for believing in me and taking me seriously. I needed it.

Contents

1	Introduction	13
2	Biological Background	17
3	Network Characterization	21
3.1	Introduction	21
3.2	Previous Work	22
3.3	Data Set	23
3.4	Searching Techniques (Methods)	23
3.5	Control Networks	24
3.6	Motif Searches	24
3.6.1	Feed Forward Loops	25
3.6.2	Transcription Vee	27
3.6.3	Target Vee	28
3.6.4	Squares	28
3.7	Specificity	29
3.8	Conclusions from Network Parsing	30
4	Network Growth	31
4.1	Review of Current Models	31
4.2	Our Model	32
4.2.1	Limitations	33
4.2.2	Metrics	34
4.2.3	Network Growth Model	39
4.2.4	Network Growth Model	43

4.2.5	Network Growth Model	46
4.2.6	Results	51
5	Conclusions and Future Work	63

List of Figures

2-1	Possible outcomes as a result of duplication. a. Duplication of a transcription factor. b. Duplication of a target gene. c. Duplication of a transcription factor-target gene pair. Figure reproduced from [2].	18
3-1	The thirteen possible three-node motifs.	25
3-2	A possible scheme for producing a feed forward loop. The transcription factor duplicates (shown by the patterned objects) and inherits the down-regulation of the gene. The original transcription factor regulates the new transcription factor and the target gene. The solid wavy line indicates a homology link. . .	26
4-1	Calculation of clustering coefficient. a. The addition of the dashed line increases the average clustering coefficient. b. The addition of the dashed line decreases the average clustering coefficient.	36
4-2	The in- and out-degree distributions for the real network.	37
4-3	The in- and out-degree distributions for the real network.	41
4-4	The in- and out-degree distributions for the real network.	44
4-5	One step in the network growth process. a. The adjacency matrices of the seed network and the network after one time step (duplication of node 3) b. The seed network and the network resulting after duplication of node 3. . . .	48
4-6	Effect of row density (ρ) on final number of links. $P_o = 0.2$ and each data point represents 50 network realizations.	49
4-7	Self-averaging parameter χ as a function of network size. Each data point represents 100 realizations.	52
4-8	Transcription factor fraction for different values of link retainment probability P_o in a connected network.	54

4-9	Transcription factor fraction versus size of network. P_o indicates link retention probability, ($\rho = 0.7$).	54
4-10	χ as a function of number of nodes for unconnected networks. Each point represents 100 realizations.	55
4-11	Link totals as a function of nodes. P_o is varied as indicated and $P_n = [0, 10^{-3} 10^{-2}]$. Each point represents 100 realizations of the given network parameters. We have fixed the starting density to be $\sim \frac{30}{43} = 0.698$ for all curves in this plot.	56
4-12	Links as a function of size of giant component. The size of the network here is 2800 nodes. The scatter plot shows 100 different network realizations ($P_o = 0.2, P_n = 10^{-3}$)	57
4-13	Total number of links versus the size of the giant component in the network for a. 691 nodes and b. 5600 nodes.	58
4-14	Clustering coefficient versus initial row density for $P_o = 0.2$ and nodes= 691.	59
4-15	Average clustering coefficient versus size of network. Each P_o has three curves plotted for it. These correspond to different values of P_n . We can see the clustering coefficient is largely independent of the generation term.	60
4-16	The strip shows the allowed parameter space found by clustering coefficient constraints. There are probably effects at the extreme values that will blur the validity of this line such as the generation term, nonlinearities in dependence of clustering coefficient on initial row density.	61

List of Tables

3.1	Specificity between transcription factors and gene families.	29
3.2	Interactions categorized by duplication scenarios.	30
4.1	Information on number and size of components in the real network.	38
4.2	Information about homology families in the real network.	38
4.3	Information on number and size of components in the real network.	42
4.4	Information about homology families in the real network.	42
4.5	Information on number and size of components in the real network.	45
4.6	Information about homology families in the real network.	46
4.7	Summary of desired network characteristics. A non-redundant set of network characteristics would be: the number of nodes, the clustering coefficient, the number of links (or average degree) and the number of transcription factors (or modified out-degree).	51
4.8	Clustering coefficient as a function of P_o ($P_n = 10^{-3}$).	60

Chapter 1

Introduction

Life is a mystery. There are many plausible theories for how life emerged from inanimate matter, e.g., the RNA world and iron-sulfur world theory. Likewise, there are many theories for the evolution of life, where evolution is defined as the “cumulative change in the genetic composition of a population.” [12] Here, we take a small step toward illuminating the process of evolution in the model organism *Escherichia coli* (*E. coli*), one of the simplest prokaryotes.

Genes are encoded in the genetic blueprint material, DNA. Recent advances in sequencing technology have made sequencing whole genomes a manageable task. At the time of writing, even the human genome has been fully sequenced. Though we now have blueprints for many different organisms, we still do not know how even the simplest organisms work. And there are many problems buried under this simple statement. We may know the amino acid sequence for a protein; however, we can not predict the three-dimensional structure of a protein from the sequence alone. Given the structure of a protein, we are not able to predict its function. Predicting phenotype from genotype is fueling the enormous interest in computational and physical approaches to biology.

One crucial step is to understand how genes interact. This understanding will help us with knowledge of an organism now, telling us which genes to target for new drugs, etc. Understanding of the network of gene interactions fits in well with the systems-level approach to biology and the recent explosion of interest in networks. Knowledge of the network of gene interactions will also help us understand the evolution of an organism.

At the population level, evolution works through a process of genetic drift, mutation and

selection. Genetic drift is the random sampling of the versions of the genes passed on from one generation. Genetic drift promotes homogeneity by fixing a certain allele of a gene in a population. Mutation counters the reduction in the variation of a population caused by random sampling. Genetic material accumulates mutations through time, which leads to greater variation in a population. Selection needs the raw material of variability to give an advantage to those genotypes that are better adapted to the environment. As environmental conditions change, different organisms will exhibit different fitness levels. We may think of fitness as an energy, and selection as the process of finding a minimum energy in the fitness landscape. Even so, evolution is not infallible. Selection may find a local minimum instead of a true minimum in the fitness landscape. In addition, there is stochasticity in the process of evolution. Population genetics models show that even advantageous genes have a high probability of being eliminated due to genetic drift.

In this work we are interested mainly in the interplay between mutation and selection. Most mutations will be deleterious or neutral; the alleles with the mutation will be eliminated quickly. The chances of an advantageous mutation occurring and then fixing in a population are extremely small. However, there is a mechanism by which the chances of an advantageous mutation fixing in the population greatly increase. This mechanism is called the duplication-divergence model of evolution. If a gene duplicates, there is a redundancy in gene function. With this overlap in gene function, there is a much greater chance for the organism to develop an advantageous mutation. Without the flexibility of having two genes that overlap in function, it is difficult to evolve the genetic material that we have. Initially, the duplicated gene will retain all of its old interactions because it is an exact copy of its parent. As the gene mutates, it will gain new interactions and may or may not lose the old interactions. The duplication-divergence model of evolution is a theory of how networks evolve.

There is evidence for this duplication-divergence model of evolution. Galas *et al.* [11] show that many organisms evolved through whole or partial chromosome duplication. Maslov *et al.* [17] has shown that interactions between homologous genes vary smoothly with sequence identity. This gives additional evidence for the duplication-divergence model. In this work we try to quantify the effects of duplication and divergence in *E. coli*.

We also examine the question of whether we are observing convergent or divergent evolution at the network level. Divergent evolution means that the evolutionarily-favored structures we observe occurred by duplication of an ancient structure. On the other hand,

convergent evolution means that different genes with unique ancestors have evolved to carry out the same function. At the phenotypic level, we know that structures such as eyes and wings have evolved convergently. The wings of bats and birds, though they perform the same function, have grown from different digits. Likewise, the eyes of many animals have similar structures but use different proteins for the lens material. At the network level this would mean that structures do not occur due to duplication of an ancient structure.

Finally, we propose a model for network growth based on the duplication-divergence model of evolution. We include biologically relevant constraints that were missing in previous work, and we are able to generate networks similar to the real networks across commonly used metrics. We determine the viability of the duplication-divergence model as the main evolutionary process, and we are able to predict if the real network fragment is self-averaging or if it is dependent on initial conditions and history.

Chapter 2

Biological Background

The central dogma of biology states that genetic information flows from a DNA strand to RNA molecules through transcription and then to proteins through translation. The base pair sequence for the protein is encoded in the DNA. This sequence is transcribed to an RNA sequence principally through the action of the enzyme RNA polymerase and possibly other enzymes. The transcript is then translated into an amino acid sequence by ribosomes. The amino acid chains then fold into proteins, the workhorses of the cell.

The sequence preceding a coding region of DNA corresponds to the RNA polymerase binding site. In addition to RNA polymerase, other proteins are often needed to initiate transcription. These proteins are called transcription factors and have their own binding sites. In some cases, transcription factors are used to aid transcription (activation); other times, they are used to prevent transcription (repression).

In prokaryotes, genes that are prefaced by the same DNA binding sites and promoter are in the same operon. Whole operons are transcribed to a single mRNA strand; when the mRNA strand is translated, the different genes are translated into distinct proteins. The proteins produced from an operon are regulated as a unit.

Central to the duplication-divergence model of evolution is gene duplication. Initially, the duplicated gene will interact with the same genes as its parent (Figure 2-1). It can be argued that the redundant copy of the gene is not under evolutionary pressure and may acquire new functions. If the gene is a transcription factor, it will continue to respond to the same signals and regulate the same genes as the ancestral gene. This is because it has the same active sites and recognizes the same transcription factor-binding sites upstream of

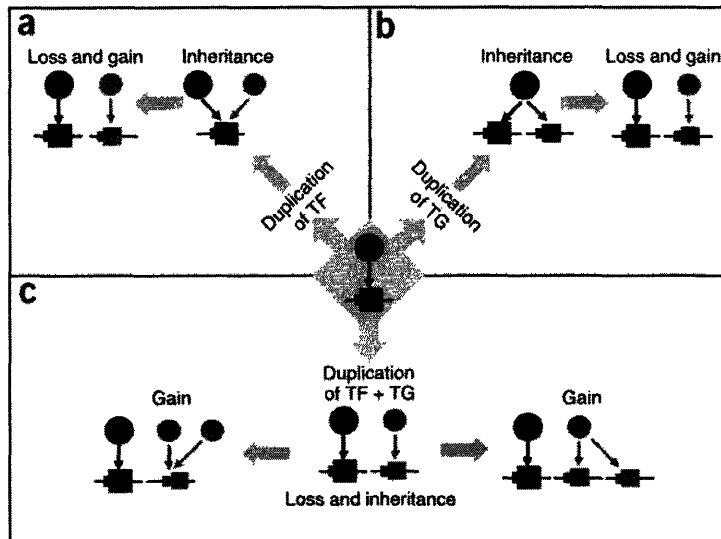


Figure 2-1: Possible outcomes as a result of duplication. a. Duplication of a transcription factor. b. Duplication of a target gene. c. Duplication of a transcription factor-target gene pair. Figure reproduced from [2].

its target gene. As the genes diverge, the duplicated transcription factor can mutate in two ways: the duplicated transcription factor can start responding to a different signal (as in development/pattern formation) or it can recognize a different transcription factor-binding site and start regulating a new target gene (2-1 a). If the duplicated gene is a target gene (along with its upstream region), either the upstream region can mutate and recruit a new transcription factor or the sequence of the gene can mutate so the duplicated gene performs a new function or both can happen at the same time (2-1 b). If a transcription factor-target gene pair duplicates, either member of the pair may gain new interactions or lose old ones.

To determine the history of duplication in the network, we must determine homology between different proteins. Homologous genes are genes that are derived from a common ancestor. In this work we determine homology by sequence similarity between genes. The assumption is that proteins with similar sequences have an ancestral gene in common. The stronger the similarity, the more recent the duplication event. Because we assume mutations occur at a roughly constant rate, sequence dissimilarity gives us a measure of how distant the duplication event is. Of course, genes that are not at all related will have dissimilar sequences. We employ a cutoff after which genes are said to be genetically unrelated.

Another method of determining genetic similarity is through structural analysis [2]. Proteins are comprised of domains of differing structure. Proteins with the same structural

domains can be assumed to be homologous. The most accurate method of determining homology would be some combination of sequence and structural techniques. For simplicity, we confine ourselves to sequence matching.

Chapter 3

Network Characterization

3.1 Introduction

A network is composed of nodes and links. The links connect the nodes and may be directed or undirected. To fully characterize a network, we must examine many properties of the network. These properties range from coarse-grained global properties to more finely-grained measures. Measures such as clustering, connectivity, in-degree and out-degree distributions capture global properties but do not offer evidence for or against the duplication-divergence model. To investigate evidence for the duplication-divergence model of evolution, we must examine properties on a small scale. One goal of this work is to find such appropriate small-scale metrics.

We investigate two types of networks. In the first type of network, a gene regulatory network, genes are nodes and interactions are links. In the regulatory network, the links are directional because the interactions are directional (A influences B but B does not necessarily influence A). We do not distinguish between repressive and activating interactions because we are interested in the topology of the network. The directionality of the network also divides the nodes into two categories, transcription factors and target genes. Transcription factors are those genes that regulate other genes, those nodes with arrows emanating from them; target genes are those that are regulated, nodes with arrows ending on them. Genes may fall into both of these categories if they both regulate other genes and are regulated by other genes.

The second type of network we use is a homology network. Again, genes are nodes, but links connect genes that are homologous to one another. In this network, links are

undirected (if A is homologous to B, B must necessarily be homologous to A). Thus, the adjacency matrix for the homology network is symmetric. The diagonal has all ones because every gene is homologous to itself. Together, the two data sets comprise a graph with links of two colors (regulatory and homology).

Our small-scale metric will be the enumeration of motifs in the above-mentioned networks. Motifs are small, recurring patterns of links. To examine the presence of evolutionary pressure, we initially searched for motifs in the regulatory network. Many motifs are over-represented compared to a suitably randomized control network. This suggests that there is some pressure for these motifs to appear. We assume that overrepresented motifs appear in part because evolutionary forces promoted their retention. Since we are assuming that evolution works through a process of duplication, divergence and then selection, our hypothesis is that these motifs will contain more homology links than average. We are combining the information in the gene regulatory network and the homology network to trace the evolution of the network. In some sense, the regulatory network contains the present while the homology network contains information about the past.

In this part of the thesis, our goal will be to both quantify the effects of the duplication-divergence method of evolution and examine the likelihood of divergent versus convergent evolution. We will do this by binning the possible types of interactions into the different duplication-divergence scenarios outlined in the previous section. An interaction may be the result of duplication and inheritance of ancestral interactions or of duplication and divergence in interaction.

3.2 Previous Work

Previous literature (Conant *et al.* [4]) states that much of the regulatory network has evolved independently, i.e., we observe convergent evolution in *E. Coli*. However, there are some serious simplifications in this work. Conant *et al.* motifs involving operons. This approach ignored much of the regulatory network.

We will use some of the techniques in [4] to examine our data set. In particular, we will use a metric called a motif family. A motif family is a collection of motifs that are derived from one common ancestral motif. In terms of our data, this means that corresponding genes in each motif of the family are homologous. Large motif families support divergent

evolution because this means many motifs have arisen from a single ancestral motif. Small and numerous motif families support convergent evolution because diverse genes have been recruited to a particular motif.

In [1] and [2], we see strong evidence for the duplication-divergence model of evolution. By using the homology network, each regulatory interaction has been binned into possible outcomes of a duplication and divergence process. They have found that over a third of regulatory interactions are a result of gene duplication and inheritance of interactions, and more than half of regulatory interactions are a result of gene duplication and divergence in interactions.

3.3 Data Set

We use two distinct data sets for our motif searching. The first data set comes from Milo *et al.* [3]. This data lists regulatory interactions for operons, not single genes. For many motifs it is more appropriate to look at operons than single genes. The data set comprises 423 operons involved in 578 interactions. We refer to this data set as the wrapped operon data.

The second data set was obtained from the supplementary information of [2]. This data was for single genes, not operons. There were a few misnamings among the 694 genes, so the final data set for single genes contained 691 genes and 1190 regulatory interactions. We refer to this data set as the unwrapped operon data.

3.4 Searching Techniques (Methods)

Homology was determined by sequence similarity using BLAST (Basic Local Alignment Search Tool). A local copy of BLAST was used to compare each gene sequence against every other gene sequence in *E. coli* K12. A BLAST score determines how similar two sequences are by using a simplified Smith-Waterman algorithm. The BLAST score is order dependent and depends on sequence length. There is a penalty employed to score mismatched bases and gaps. We used a cutoff BLAST score of 50. Genes with alignment scores larger than 50 are considered to be homologous.

The network is represented as an adjacency matrix, with column entries indicating regulators of the column and row entries indicating regulatees of the row. Endpoints of motifs

were determined by matrix and dot multiplication of the adjacency matrices [18].

3.5 Control Networks

It is necessary to study a randomized version of a network in order to have a control for observed network characteristics. Without a control network, we will not be able to tell if features in the network are statistically significant. Features that are not statistically significant do not necessarily show any evidence of evolutionary pressure. There are two different levels to the randomization control. First, we have a control that involves only the regulatory network. This control helps us determine which regulatory features are significant.

To create the control network, we have used an algorithm that takes two outgoing links in the real network and swaps the ending nodes for the two links [17]. This process is continued for a sufficiently long time until the network has been randomized. The control networks constructed by this algorithm have the same number of nodes and links as the real network. In addition, this algorithm retains the same number of incoming and outgoing links at each node. This provides a yet more stringent control.

The second type of control network randomizes the homology network. This control network tells us if the number of homology links observed in regulatory structures are more than would be expected at random. To obtain the control network, we randomly rename all of the genes. This preserves the regulatory network but randomizes the homology links between genes and, particularly, alters the homology links within regulatory structures of interest. Similar to Teichmann *et al.* [2], we randomized the transcription factors and regulatory targets separately because the transcription factors tend to be more homologous to each other than regulated genes in our network fragment.

3.6 Motif Searches

Intramotif analyses will quantify evidence for the duplication-divergence model of evolution. We have concentrated on several motifs for our intramotif searches. We analyzed the most prevalent of the thirteen possible three node motifs (Figure 3-1): feed forward loops (motif 5), transcription vees (motif 4) and target vees (motif 1). We also performed analyses on other motifs of interest evolutionarily or functionally. The three node motifs are summarized in Figure 3-1.

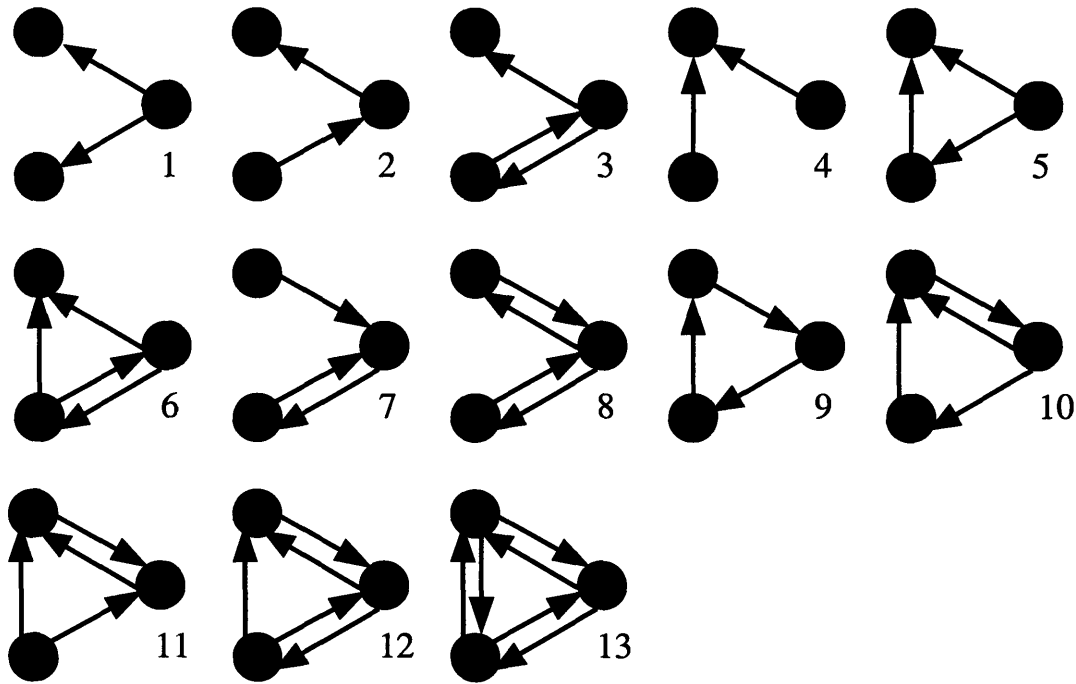


Figure 3-1: The thirteen possible three-node motifs.

3.6.1 Feed Forward Loops

A feed forward motif consists of three genes, two transcription factors and one target gene. One transcription factor regulates both the target gene and the second transcription factor, which also regulates the target gene (Figure 3-1, motif 5). A lot of excitement has been generated about feed forward loops because they could act as a sign-sensitive delay element [20].

We enumerate the number of motifs in the network. To do this analysis we have used the data compiled by Milo *et al.* [3], the wrapped operon data. In this motif, we consider nodes to be operons, not single genes. We have found 42 feed forward motifs in the real network. In the regulatory control network, we have found 8 ± 3 occurrences of the feed forward motif in 150 realizations of the network. This is in close agreement with published results of 7 ± 3 [3].

The actual number of feed forward loops is more than 10 standard deviations away from the expected number. There are two points here. First, we see clearly that feed forward loops are overrepresented in the real network compared to a suitably randomized network. Second, we can surmise that there is something in this structure that makes it evolutionarily favorable because it is so over-represented.

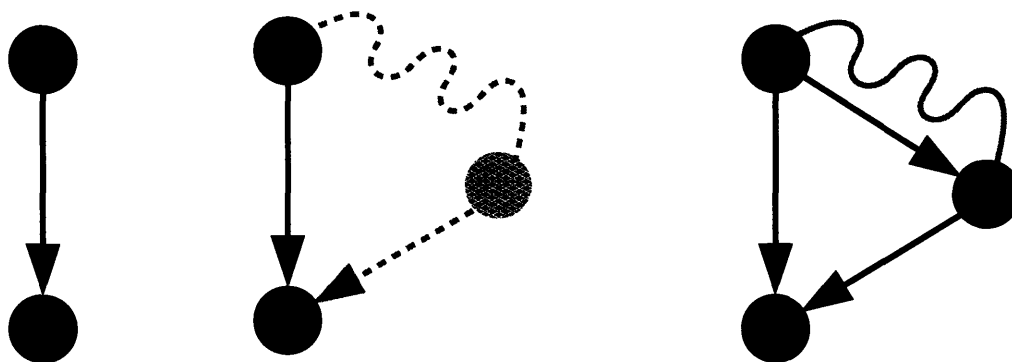


Figure 3-2: A possible scheme for producing a feed forward loop. The transcription factor duplicates (shown by the patterned objects) and inherits the down-regulation of the gene. The original transcription factor regulates the new transcription factor and the target gene. The solid wavy line indicates a homology link.

The next step is to determine if the emergence of the feed forward loops is consistent with the duplication-divergence model of evolution. An explanation for the formation of this motif within the duplication-divergence framework is that a transcription factor regulating a target gene duplicates, giving rise to a second transcription factor that also regulates the gene. Then, the original transcription factor begins to regulate the duplicate (Figure 3-1). To test this explanation, we examine how many feed forward loops have a homology link between the two transcription factors. Here, we used genes as nodes in order to better use the homology data. The 42 motifs in the wrapped operon data become 208 motifs in the unwrapped operon data.

Teichmann *et al.* [2] claimed that no feed forward loops were formed by duplication of the original transcription factor within the motif. We have found 7 out of 208 such feed forward loops based on two pairs of transcription factors. The first transcription factor pair is *rob* and *marA*. This pair is responsible for 5 of the 7 motifs with the homology link. Teichmann *et al.* discards the pair of genes, *rob* and *marA*, as evidence of the transcription factor duplication model outlined above because the genes are structurally dissimilar. By looking at Teichmann's structural database, we find that *marA* consists of two homeodomain-like structures; likewise, *rob*, consists of two homeodomain-like structures, but has an added putative bacterial effector-binding domain. However, it is important to remember that we are tracking homology of DNA-binding domains. The structure (and sequence) of other domains does not concern us. Thus, the two proteins are homologous in structure with respect to DNA-binding domains.

The other transcription factor pair, *exuR* and *uxuR*, is responsible for 2 of the 7 motifs with the homology link. We find again that the DNA-binding domains do indeed match. It is the other domains that do not match or are unknown.

Though our approach has yielded some support for the transcription factor duplication model, there is little real evidence that this is the primary process for the formation of feed forward loops.

To examine support for either a convergent or divergent evolutionary process, we must examine the number of motif families. We have found 192 motif families for the 208 feed forward motifs. This indicates that an overwhelming majority of motif families have only one member, i.e., most motifs are derived from separate ancestors. Thus, we have found that feed forward motifs presumably arose through convergent evolution.

3.6.2 Transcription Vee

The transcription vee motif (motif 4) is composed of two transcription factors regulating one gene. This motif could have arisen from duplication of a transcription factor regulating a gene. There are 227 transcription vee motifs in the wrapped operon data. In the regulatory control network, we find 261 ± 3 transcription vee motifs.

When we count single genes as nodes, there are 810 transcription vee motifs composed of 821 distinct interactions. (Though each vee motif is composed of two regulatory interactions, interactions can be present in several vee motifs. To avoid double-counting, we give the number of interactions involved in a motif.) Of 810 motifs, 77 of the motifs contain a homology link between the two transcription factors, and they comprise 142 interactions. When we compare our results to the homology control network, the original numbers of motifs and interactions remains unchanged (because we are not changing the regulatory network). But the number of motifs with a homology link and associated interactions decreases dramatically. Out of 150 realizations with homology matrix randomized, we observe 28 ± 22 vee motifs comprised of 54 ± 40 interactions (versus 77 motifs composed of 142 interactions). We have significantly more motifs and associated interactions in the real network as compared to the randomized network.

Teichmann *et al.* [2] counted only the number of regulatory interactions contained in the transcription vee motifs with the homology link. They found 128 interactions based on structural homology while, as stated above, we found 152 interactions based on sequence

homology. Though both approaches to homology yielded similar numbers of interactions, only 109 interactions were identical.

For the transcription vee motif, we have found 6 different motif families with more than one member. However, existence of these families does not support divergent evolution because two out of three genes are the same in these motif comparisons. For instance, many of the motif families involve *marA* in one motif and *rob* in another. These two genes seem to be interchangeable in many motifs; four of the six families occur for this reason. The fifth family involves *nuoLMN*, all of which are homologous and in the same operon. The sixth has arisen because of a homology link between *pflB* and *tdcE*.

3.6.3 Target Vee

The target vee (motif 1) comprises one transcription factor regulating two genes. This motif may have arisen by duplication of the target gene. There are 4777 Vee1 motifs in the wrapped operon data [3]. In the regulatory control network, we find 4810 ± 6 .

Teichmann *et al.* have found “duplication of the target gene with inheritance contributed to 272 interactions (22%) in *E. coli*” [2]. We have found 185 interactions (for unwrapped operons). This match is not as bad as they have included 44 tRNA interactions that we have discarded.

There are 61 motifs that are not homologous to any other motif, and there are twelve different motif families with more than one member. The families consist mainly of a transcription factor regulating a group of homologous genes, often genes in the same operon. The motifs consist of the transcription factor and permutations of the regulated genes. Only one family includes motifs that differ by two genes, *arcA* and *narL*, with permutations of the *nuoLMN* operon. There is little evidence of a whole motif duplicating. It seems that genes duplicate separately.

3.6.4 Squares

The square motif is a four-node motif composed of two homologous transcription factors regulating two homologous genes with no crosstalk. The squares are significant in that they may have come about by simultaneous duplication of a transcription factor and regulated gene followed by divergence so that each transcription factor only regulates its own gene, i.e., no crosstalk. Teichmann *et al.* notes that this motif is prominent in sugar regulatory

Transcription Factor	Genes Regulated	Size of Family
envY	cirA fepA fhuA	4
flhD	dmsA fdnG nuoG torA	5
fur	flgE flgF flgG	3
glcC	glcD glcF	3
narL	ompC ompF	3

Table 3.1: Specificity between transcription factors and gene families.

pathways. They quote 74 interactions while we have 72. However, the actual interactions appear to be quite different.

3.7 Specificity

Using the hypergeometric distribution, we may quantify the specificity of a transcription factor for a gene homology family. N represents the total number of families, M represents the total number of links, n is the number of hits in a family and m is the total number of hits of interest in a family.

$$P = \frac{\binom{n}{m} \binom{N-n}{M-m}}{\binom{N}{M}}$$

Using a p-value cutoff of 10^{-4} , we have found 5 transcription factor-gene family pairs. The five transcription factors are summarized in Table 3.1.

Interactions due to:	Number of Interactions
Transcription factor inheritance	130
Target gene inheritance	163
Either type of inheritance	22
Square Motifs	72
Either type of divergence	694
Innovation	109

Table 3.2: Interactions categorized by duplication scenarios.

3.8 Conclusions from Network Parsing

Can we determine conclusively if the *E. coli* gene network is a function of convergent or divergent evolution? We do not find a lot of evidence for divergent evolution. Each gene duplicates separately; three genes that comprise a small regulatory motif will not duplicate together. The lack of evidence of intermotif homologies leads us to believe that networks grow incrementally, not by duplication of motifs. But the case for convergent evolution is clear only for the feed forward motif. The other motifs were not overrepresented by comparison to a randomized network so it is not clear if they are evolutionarily important. However, a larger than expected proportion of these motifs had homology links. This supports the duplication-divergence model of evolution.

Table 3.2 summarizes the origin of the regulatory interactions in our data set. These results indicate clearly that evolution has proceeded through a process of incremental duplication and divergence. Only 109 out of 1190 interactions include genes that have no homologs in our network fragment. This means that only 9% of our network fragment has arisen by pure innovation. Roughly one third of the regulatory interactions have arisen due to duplication and inheritance of ancestral regulations. More than half of the interactions are due to duplication and subsequent divergence of function.

Chapter 4

Network Growth

4.1 Review of Current Models

Early network models were based on the Erdos-Renyi models of random networks, where a random graph is defined as N nodes connected by n edges chosen from the possible $\frac{N(N-1)}{2}$ edges. These graphs are homogeneous; there is no real structure within the graphs. In recent years it has been found that real networks display more complexity than is found in random graphs. This additional structure in real networks can be captured by many different metrics. Here, we will briefly introduce the measures most relevant for distinguishing between different network models.

Degree Distribution: The degree of a node is given by the number of links, k , connected to the node. In a directional network there is an in-degree and out-degree for every node. Both the average degree, $\langle k \rangle$, and the degree distribution, $P(k)$, are informative. Random network models are characterized by a Poisson degree distribution, $P(k) \simeq e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}$. Significantly, the tail end of the degree distribution has an exponential drop off.

Clustering Coefficient: Clustering in a network refers to how many of the k_i neighbors of a particular node i are connected to each other. The more the connections between the neighboring k_i nodes, the higher the clustering coefficient, C_i . If L_i is the number of links that exist out of the possible $\frac{k_i(k_i-1)}{2}$ neighboring links, the clustering coefficient of node i is $C_i = \frac{2L_i}{k_i(k_i-1)}$. The clustering coefficient of a random graph is simply $\frac{\langle k \rangle}{N}$.

It has been found that random graphs fail to emulate many real networks in two main ways. Many real networks have a degree distribution that follows a power law instead of the

exponential drop off of the tail of a Poisson distribution. In addition, most real networks display a clustering coefficient larger than predicted by random network models.

Thus, new network models have been proposed, notably the scale-free model by Barabasi *et al.* [16]. In the scale-free model the degree distribution follows a power law. The term “scale-free” refers to the fact that the power law distribution is self-similar; any portion of the distribution can be made to look like the full power law distribution by a rescaling of the axes. Tied closely to the notion of a scale-free network model is the mechanism for growth of such a network. There are two key points to the growth mechanism: first, the network grows by the addition of new nodes; second, new nodes are attached preferentially to existing nodes with high degree.

The scale-free model is well suited to describe such networks as the world wide web and networks of scientific collaborators [16]. However, the scale-free model is not a good description of many biological networks. Biological networks are distinguished from other networks by their propensity for duplication. Webs like the Internet may grow by creation of innovative nodes; biological networks tend to add nodes through duplication. Duplicated nodes will have links similar to their ancestral node. Thus, the notion of preferential attachment is not appropriate for biological networks.

There has been work in determining biologically appropriate models of evolution. However, even for networks in which duplication is important, we must have different models for different biological networks. For instance, directionality is not indicated for protein-protein interaction networks [10]; but it is a key ingredient in models of gene regulatory networks. There are few models that incorporate directionality, re-wiring and duplication [11]. Here we propose an intermediate model that includes relevant biological constraints and enough free parameters to replicate the process of duplication and divergence.

4.2 Our Model

We use a model of growth based explicitly on duplication and divergence of genes. At each time step we add one node to the network. Each node of the network has an equal probability to be duplicated. We determine the linkage of the new node at the time of duplication. The duplicated node retains each link of the ancestral gene, both incoming and outgoing, with probability P_o . The duplicated node makes innovative links, again both

incoming and outgoing, with probability P_n . In addition to the link retention and link innovation probabilities, we have initial conditions as a free parameter.

Many models begin growing a network from two connected nodes. We start our growth process with a seed network. We vary the number of nodes and number and placement of links in the seed network. As we will see, freedom to determine the initial conditions gives us more control over the final network than possible in previous models.

In many network models, all nodes are equal. We distinguish between transcription factors and target genes through the directionality of the network. Explicitly, we stipulate that regulated genes may not become transcription factors. This constraint supersedes the action of the probabilities. Previous work has not addressed this because models were non-directional for simplicity. This restriction is biologically sound because modification of a DNA binding domain, divergence of a transcription factor, is much more likely than the emergence of a DNA binding domain, non-regulating gene acquiring a DNA binding domain. By our three free parameters, we see a wealth of different network characteristics. We have focused on producing a network that is statistically similar to the real *E. coli* gene regulatory network over commonly used metrics. A summary of relevant network characteristics is given in Section 4.2.2.

4.2.1 Limitations

The observed network fragment is a sample of the larger, total *E. coli* network. It is basically the sugar regulatory system. We do not know if this is a biased or representative sample of the total network. Perhaps the rest of the network is more sparse or more linked. This is a concern because our results may not be applicable for the whole network. As we can not estimate this, we make the assumption that we have a representative portion of the network and grow our network to the size of the real network fragment.

The divergence process in this model has condensed together many steps. For instance, the formation of a new regulatory link involves mutation of a binding site, possibly formation of a new binding site of the regulated gene along with mutation of the transcription factor sequence. We allow a duplicated transcription factor to retain an old interaction and evolve a new one that does not include the ancestral transcription factor, though literature indicates that “gene duplication enables the copies to become specialized in distinct subsets of the ancestral functions” [19].

Assuming divergence of a gene happens at the time of duplication is an approximation; divergence and formation of new links can occur at any time. The probabilities of keeping old and making new links remain constant over time in our model. It would be possible to make these probabilities time or context-dependent. For instance, if we have duplicated a node with high degree, we could make it more likely for this node to make more links.

We disregard the effect of operons upon the growth process. This would mean that some links are correlated with each other, disappearance of one link implies disappearance of all links associated with the same transcript. We assume each gene's interactions evolve independently of every other gene's interactions.

4.2.2 Metrics

There are two types of network characteristics, those that we meet *a priori* and those that we would like to meet. Built into the system is the number of nodes in the network. We grow our network to 691 nodes, the size of the real network fragment. The second network characteristic that we fix is the number of transcription factors in the final network. By choosing the number of transcription factors in and size of the seed network, we are able to determine the distribution of the number of transcription factors in the final network.

The network properties that we would like to emulate are the average degree and the average clustering coefficient. By varying our three free parameters, the initial filling fraction in the seed network, the probability of link innovation and the probability of link retention, we hope to find a parameter space that will give us real network characteristics.

The four network characteristics outlined here, the number of nodes, the number of transcription factors, the average degree and the average clustering coefficient, give four independent measures on the network. The values for these measures on the real network along with descriptions of the measures are given below.

Clustering Coefficient

Clustering in a network, as mentioned above, refers to how likely it is for neighbors of a particular node to be neighbors of each other. The more likely this is, the more clustered the network and the higher the clustering coefficient. For more detailed comparisons between networks, we may examine the clustering coefficient distribution.

The clustering coefficient for an undirected network is well-defined. However, for a directed network, there are two possible clustering coefficients, the out-clustering coefficient and the in-clustering coefficient. As an approximation, one may symmetrize the adjacency matrix of the digraph and then compute the clustering coefficient as if for a non-directional graph.

The clustering coefficient for a network is found by matrix operations on the symmetrized adjacency matrix A as follows:

$$\langle C \rangle = \sum_i \frac{\sum_j [(A^2 - \text{diag}(A^2)) \cdot A]_{i,j}}{\text{diag}(A^2)_{i,i} \star (\text{diag}(A^2)_{i,i} - 1)}$$

For a given network, it is interesting to see that the addition of a random link does not affect the average clustering coefficient in a known way. The addition could result in an increase or a decrease of the average clustering coefficient depending on the structure of the network (Figure 4-1).

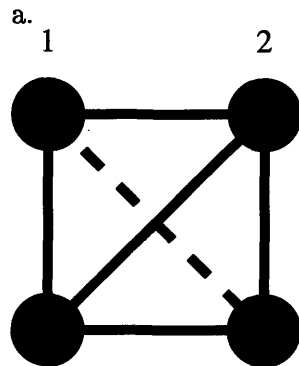
The average clustering coefficient for the real graph of 691 genes is 0.188. This clustering coefficient ignores auto-regulation and symmetrizes the adjacency matrix.

Degree Distribution

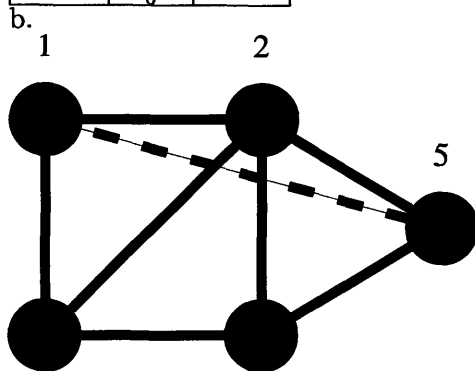
The concept of degree was introduced in the previous section. The degree of a node is given by the number of links connected to the node. In a directed network there is an in-degree and out-degree for every node. Both the average degree and the degree distribution are informative. Another variant of the degree metric that is useful for digraphs is the modified out-degree, the out-degree averaged over only those nodes with outgoing links.

For our network growth procedure, it is important to know how the addition of new link alters the average degree. The average degree for a directed network is $\frac{L}{N}$. So, $\frac{dL}{dN} = \frac{1}{N}$, and the addition of a link to a network increases the average degree distribution by $\frac{1}{N}$. For an undirected network, the average degree distribution is $\frac{2L}{N}$ (because each link would be counted twice, once at each end).

The average degree of the real network is $\frac{1190}{691} = 1.72$ and the modified out-degree is $\frac{1190}{118} = 10.1$. Figure 4-2 plots the real in- and out-degree distributions.



Node	C_{old}	C_{new}
1	1	1
2	$\frac{2}{3}$	1
3	$\frac{2}{3}$	1
4	1	1
$\langle C \rangle$	$\frac{6}{5}$	1



Node	C_{old}	C_{new}
1	1	$\frac{2}{3}$
2	$\frac{3}{4}$	1
3	$\frac{2}{3}$	$\frac{2}{3}$
4	$\frac{2}{3}$	$\frac{2}{3}$
5	1	$\frac{2}{5}$
$\langle C \rangle$	$\frac{49}{60}$	$\frac{48}{60}$

Figure 4-1: Calculation of clustering coefficient. a. The addition of the dashed line increases the average clustering coefficient. b. The addition of the dashed line decreases the average clustering coefficient.

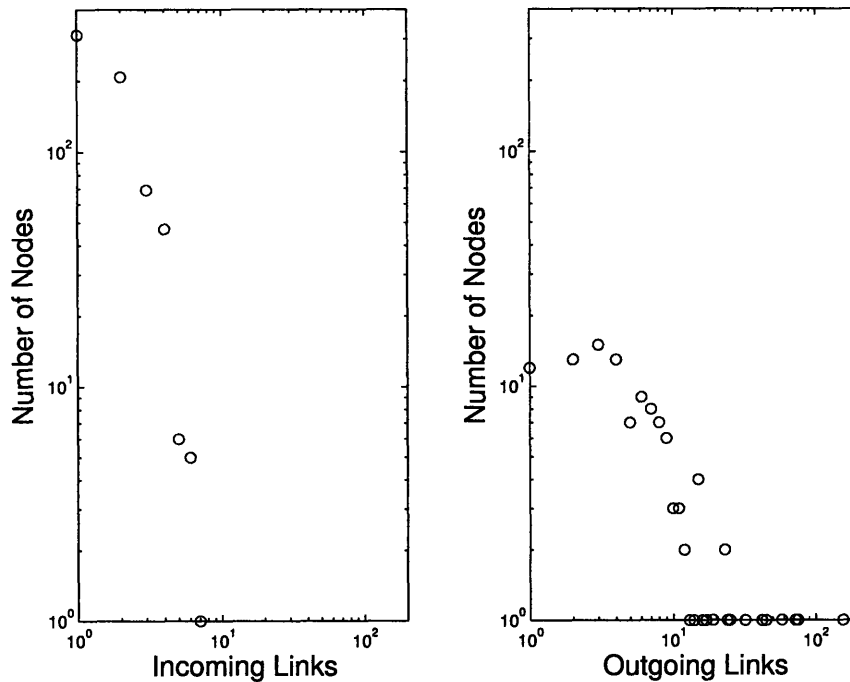


Figure 4-2: The in- and out-degree distributions for the real network.

Number of Transcription Factors

We have tried to fix the number of transcription factors in the grown networks to 118. This was done by specifying the size of and types of nodes in the seed network. This will be discussed at length in the next section.

Distribution of Component Sizes

The real network is not completely connected; there are disconnected fragments of the real network. Table 4.1 gives the information on components of the real network. The giant component is the largest component of the network. We use the size of the component as a measure of the grown networks. An important point is that we do not know if the component distribution of the real network fragment is a vestige of sampling or if it is truly informative.

Homology Families

A homology family is a collection of genes that are homologous to each other. The homology information is summarized in Table 4.2. There are 357 out of 691 genes that are not homologous to anything else in the network.

Size of component	Number of components
539	1
20	1
16	1
15	1
9	1
7	2
6	2
5	3
4	3
3	5
2	12
1	0

Table 4.1: Information on number and size of components in the real network.

Members in a family	Number of families
25	1
23	1
18	1
17	1
13	1
7	1
6	1
5	6
4	8
3	23
2	49
1	357

Table 4.2: Information about homology families in the real network.

Self-Averaging Parameter

To characterize the variability of possible simulated networks, we define the measure $\chi = \frac{\sqrt{\langle L^2 \rangle - \langle L \rangle^2}}{\langle L \rangle}$ [7], the standard deviation over the mean. It is not possible to determine χ analytically; this must be determined computationally. A network is self-averaging if χ approaches zero as the number of nodes increases.

4.2.3 Network Growth Model

Nodes

The average final number of transcription factors is determined by the number of transcription factors in the seed matrix. The probability distribution for the number of transcription factors can be found as a modified Polya's Urn problem. We pick the size of the seed network by determining the final node fraction we would like to have.

We grow the network from the initial seed matrix by randomly picking an existing node and duplicating it at each time step. The initial conditions are given by (a, b) where a gives the number of transcription factors and b gives the number of non-regulating genes. The corresponding urn state is a white balls (transcription factor) and b black balls (non-regulating genes). We choose a ball from the urn at random and replace it and add another ball of the same type to the urn. We do this n times and end up with $a + b + n$ balls in the urn. The probability of having $(a + i, b + n - i)$ for our final state, where i ranges from 0 to n is given by:

$$P(a + i, b + n - i) = \frac{(b + n - i + 1)!(a + b - 1)!n!(a + i - 1)!}{(b - 1)!(a + b + n - 1)!(n - i + 2)!i!(a - 1)!}$$

For the case $(a, b) = (1, 5)$, the above equation reduces to:

$$P(a + i, b + n - i) \propto (n - i + 6)(n - i + 5)(n - i + 4)(n - i + 3)$$

a quartic polynomial. In general, $P(a + i, b + n - i)$ will be a polynomial of order $(a - 1) + (b - 1)$. Thus, even if the ratio $\frac{a}{b}$ is held constant, the initial conditions (a, b) and (ca, cb) will yield different probability distributions.

The smallest seed network to yield the greatest number of networks with final transcription factor ratio of $\frac{118}{691} = 0.1708$ has $(a, b) = (9, 40)$. We have used $(a, b) = (8, 35)$ with a

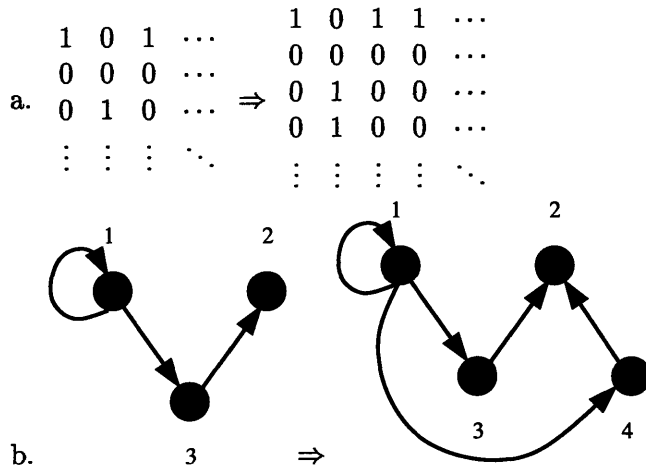


Figure 4-3: One step in the network growth process. a. The adjacency matrices of the seed network and the network after one time step (duplication of node 3) b. The seed network and the network resulting after duplication of node 3.

final transcription factor ratio of $\frac{119}{891}$ to use a smaller seed network. Though other smaller seed networks offer a greater number of final networks within a 0.14 to 0.20 transcription factor ratio range, the distribution around 0.17 is highly skewed. We prefer to have a symmetric distribution, i.e., have a local maximum as close to 0.17 as possible. On the other hand, the larger the seed matrix, the more likely the final network will have a transcription factor ratio close to the desired ratio. The trade-off here is that we want to specify the initial conditions stringently enough to end up near the desired transcription factor ratio but we want to have enough time for the specificity of the initial conditions to relax and for the randomness in the growth process to fully sample the possible networks.

Initial Conditions

The size and transcription factor/target gene composition of the seed network are fixed by desired final network characteristics. However, the number of links in the seed network is still a free parameter. It is important to note that the only rows in our adjacency matrix that are occupied are those that correspond to transcription factors because only transcription factors have outgoing links. Our free parameter is the row occupancy or density of our seed matrix. Figure 4-4 shows the effect of initial row density ρ and P_n on the number of links for $P_o = 0.2$.

For a low generation term, i.e. small probability of making new links, the number of links is related linearly to the initial filling fraction of the seed matrix (seen in the lower

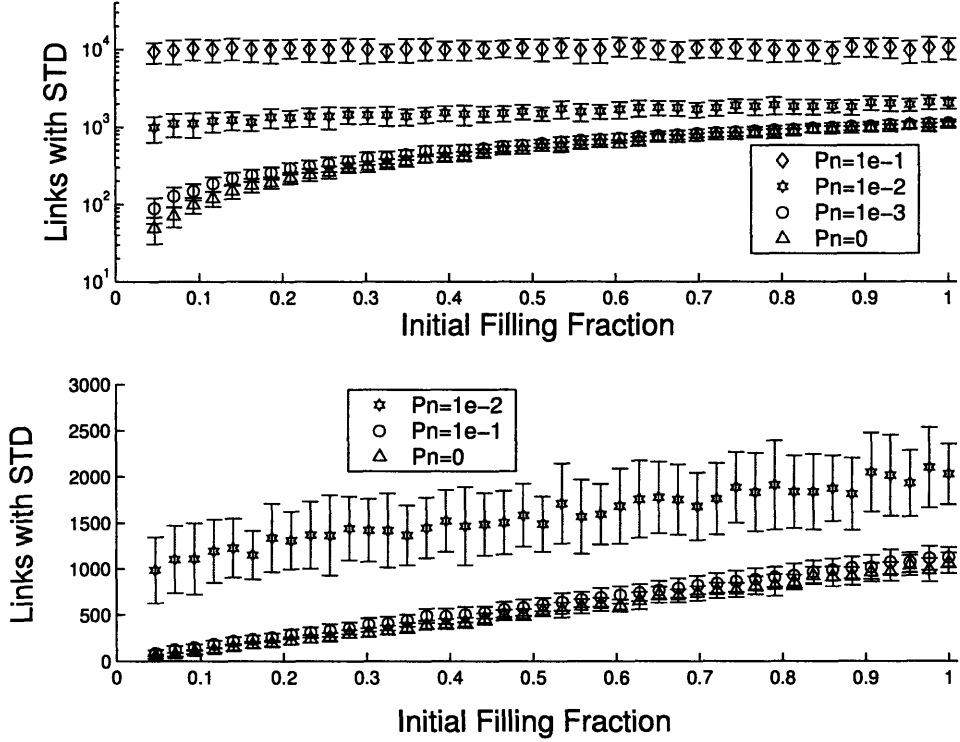


Figure 4-4: Effect of row density (ρ) on final number of links. $P_o = 0.2$ and each data point represents 50 network realizations.

plot of 4-4). As the generation term increases, the number of links becomes independent of the initial density and becomes dependent only on the generation term (apparent from the diamonds in 4-4). These different dependences of link number indicate two regimes of the final network.

Links

The number of links is dependent on all three free parameters. We find the number of links, L , as a function of network nodes, N , analytically for a simplified model of network growth. We have an undirected network with probability of link retainment, P_o , and probability of innovative link formation P_n . k is the degree of each node and n_k is the fraction of genes that have degree k .

$$\Delta N = \sum_k n_k ([1 - (1 - P_o)^k] + [(1 - P_o)^k P_n (N - k)])$$

where $\sum_k n_k = 1$. Following the convention of Ref. [7], we will define ν to be $\sum_k n_k [1 -$

$(1 - P_o)^k$]. Then the above expression reduces to:

$$\Delta N = \nu + (1 - \nu)P_n N - \sum_k n_k [(1 - P_o)^k P_n k]$$

We may regard ν as a constant because it is independent or only weakly dependent on N . If we relax the stipulation that the network remain connected, $\Delta N=1$ because we will not discard a node even if it is disconnected from the rest of the network. In a model where no innovative links are made, the main component of the network will continue growing as if the isolated nodes were not there. In a system where innovative links are made, the once isolated nodes may now become subnetworks/components.

The change in the number of links per time step is

$$\Delta L = \sum_k n_k (k P_o + (N - k) P_n)$$

The first term in the above expression signifies the links due to retainment of old links and the second term signifies new, innovative links made. Using the fact that $N \sum_k n_k k = 2L$, the above equation reduces to

$$\Delta L = \frac{2L(P_o - P_n)}{N} + N P_n$$

Therefore,

$$L' = \frac{dL}{dN} = \frac{\Delta L}{\Delta N} = \frac{2L(P_o - P_n)}{N} + N P_n$$

By defining $L = \tilde{L} e^{f(N)}$, we have

$$L' = \tilde{L}' e^{f(N)} + \tilde{L} e^{f(N)} f'(N) = \tilde{L}' e^{f(N)} + L f'(N)$$

and by rearranging

$$L' - L f'(N) = \tilde{L}' e^{f(N)} = L' - \frac{2L(P_o - P_n)}{N} = N P_n$$

where $f'(N) = \frac{2(P_o - P_n)}{N}$. Thus,

$$\tilde{L}' = N P_n e^{-f(N)}$$

Calculating the integrating factor

$$e^{-f(N)} = C_o N^{-2(P_o - P_n)}$$

and substituting

$$\tilde{L}' = C_o P_n N^{1-2(P_o - P_n)}$$

Integrating and defining the constant of integration C_1 to simplify later expressions

$$\int \tilde{L}' dN = \frac{C_o P_n}{2 - 2(P_o - P_n)} N^{2-2(P_o - P_n)} + C_o C_1$$

Multiplying by $e^{f(N)} = \frac{1}{C_o} N^{2(P_o - P_n)}$, we obtain the expression for L

$$L = \frac{P_n}{2[1 - (P_o - P_n)]} N^2 + C_1 N^{2(P_o - P_n)}$$

For small N , L is similar to the case of duplication with no new, innovative links. In that case, $L \propto N^{2P_o}$ and the density of the network, $\frac{L}{N^2}$, decays as $\frac{N^{2P_o}}{N^2} = \frac{1}{N^{2(1-P_o)}}$. We refer to this as the diluting regime. For a growth process with a generation term, the number of links grows as N^2 for large N . The average density will remain constant for large N , $\frac{L}{N^2} \propto \frac{N^2}{N^2} = \text{constant}$. We will call this the equilibrated regime.

It is interesting to note that the constant C_1 is roughly the same for the growth process without the generation term, $L(P_n = 0) \approx C_1 N^{2P_o}$. Thus, the number of links is basically the sum of the terms due to link retention [$C_1 N^{2(P_o - P_n)}$] and link innovation [$\frac{P_n}{2[1 - (P_o - P_n)]} N^2$].

Though we now know how the number of links scales with number of nodes in the network, we still do not know anything about the spread of values possible around the average. The fluctuations about the mean give us a measure of the self-averaging of the network growth process. As stated above, to characterize the variance of possible simulated networks, we define the measure $\chi = \frac{\sqrt{\langle L^2 \rangle - \langle L \rangle^2}}{\langle L \rangle}$, the standard deviation over the mean. It is not possible to determine χ analytically; this must be determined computationally.

It is probable that networks with $P_n \neq 0$ are self-averaging for a larger range of P_o . We may think of this situation as, for a given final network density, the seed network density can be lower and the generation term will be larger. For a larger generation term, there is

Network Parameters	Values
Nodes	691
Links	1190
Transcription Factors	118
Average Degree	1.72
Modified Out-Degree	10.1
Clustering Coefficient	0.188

Table 4.3: Summary of desired network characteristics. A non-redundant set of network characteristics would be: the number of nodes, the clustering coefficient, the number of links (or average degree) and the number of transcription factors (or modified out-degree).

more erasing of memory.

4.2.4 Results

Node Segregation

We have found that the constraint that a non-regulating gene will not become a transcription factor is not only biologically relevant but also important for the model. The constraint was necessary because the in-degree and out-degree distributions were symmetrical throughout parameter space of the initial filling fraction and two different probabilities for making and keeping links. This is a problem because there is an order of magnitude difference between the maximum in-degree (7) and the maximum out-degree (110). This occurred because there was a high rate of conversion from non-regulating gene to transcription factor. Ultimately, the majority of the genes were transcription factors, which was also a problem. Without this stipulation, the matrix began to have uniform density, whereas the transcription factor rows are meant to be more dense than the non-regulating gene rows.

Connectivity

It has been found [7] that for $P_o \leq 0.5$, a connected network is self-averaging, i.e., the parameter χ tends to zero as the network grows. This means that the simulated network does not depend on initial conditions or the evolution process. This result is for networks with zero probability of making new links. Figure 4-5 reproduces their results with our growth model.

However, it is important to consider whether this connectivity constraint is biologically correct. In [7] there was no innovative link formation ($P_n = 0$) so an isolated node had

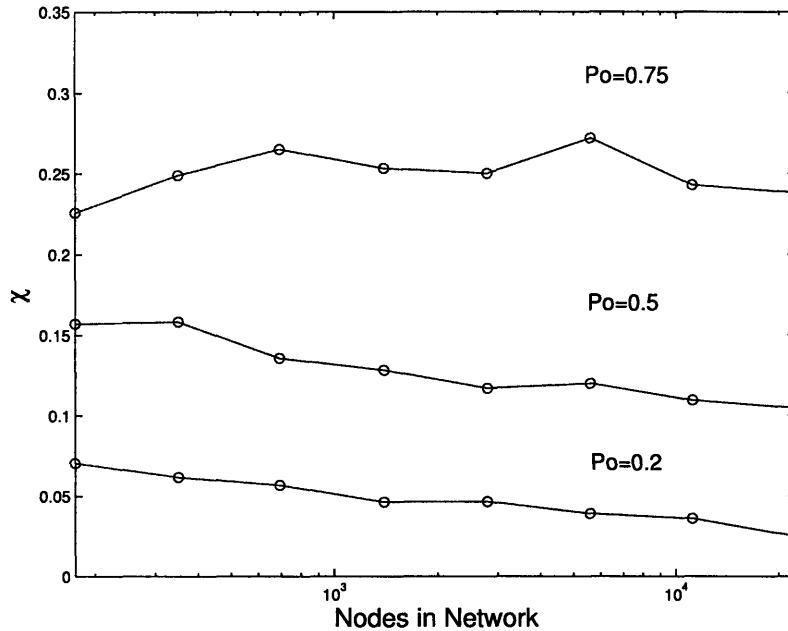


Figure 4-5: Self-averaging parameter χ as a function of network size. Each data point represents 100 realizations.

no possibility of later connecting to other nodes. Thus, it made sense to discard unlinked nodes. In our model an unlinked node may make connections with other nodes because of the possibility of innovative link formation. So, nodes that are initially isolated still have the possibility to affect the structure of the network.

Another problem with the connectivity stipulation is that the Polya's Urn model of gene duplication breaks down. This is because we must discard a possible duplication that results in an isolated node. We will have a bias to duplicate nodes that are more connected. Though it is not necessary for Polya's Urn model to describe the gene duplication process, it is our goal to fix the number of transcription factors. Because transcription factors tend to have more links than average, we will obtain more transcription factors than desired in the final network. The effect of connectivity on transcription factor fraction can be seen from Figures 4-6 and 4-7.

In Figure 4-6 we plot the theoretical Polya's Urn distribution for the fraction of transcription factors along with histograms of transcription factor fractions for two different values of P_o . For high link retainment probability ($P_o = 0.75$), the network remains connected as a matter of course; there are few failed duplication events. The histogram of transcription factor fractions parallels the theoretical curve. For low link retainment probability

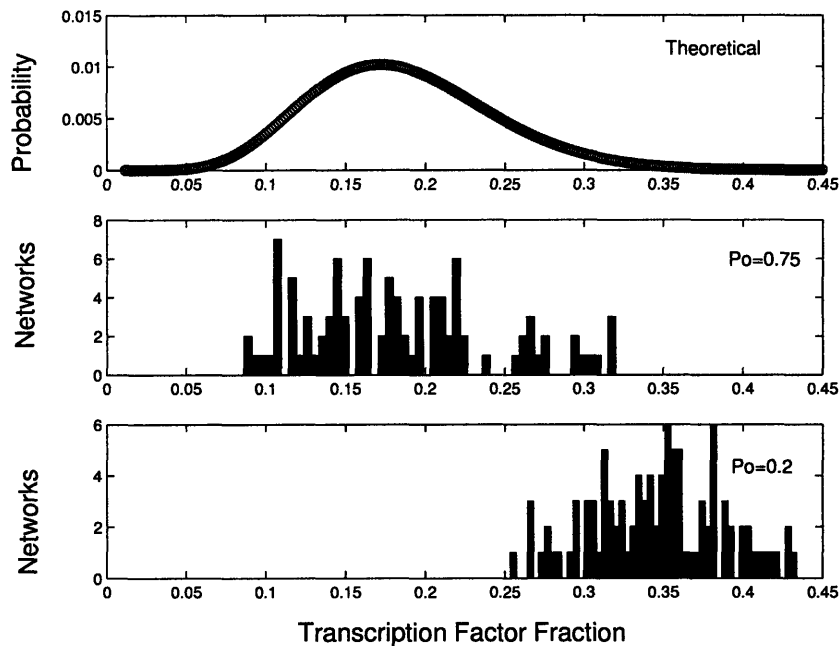


Figure 4-6: Transcription factor fraction for different values of link retention probability P_o in a connected network.

($P_o = 0.2$), there are many failed duplication events. The histogram of transcription factor fractions is biased to a value higher than predicted by the theoretical model. In addition, the distribution has smaller variance than the theoretical prediction.

Figure 4-7 shows the fraction of transcription factors for varying P_o for varying network size. This illustrates that transcription factors are favored for duplication for smaller link retention probabilities. This is a result of the stipulation that the network remain connected. The effect is more pronounced for networks of larger size.

Figures 4-6 and 4-7 and the fact that the real network fragment contains many distinct components of varying size (Table 4.1) have prompted us to relax the condition that our network remain connected. We want to preserve the fraction of nodes that are transcription factors because this is an important network parameter that we want to emulate. Relaxing the condition that the network remain connected, we observe slightly different network characteristics.

Self-Averaging

In Figure 4-8 we plot χ as a function of the number of nodes for networks without the connectivity constraint. We have also included curves for networks with innovative link

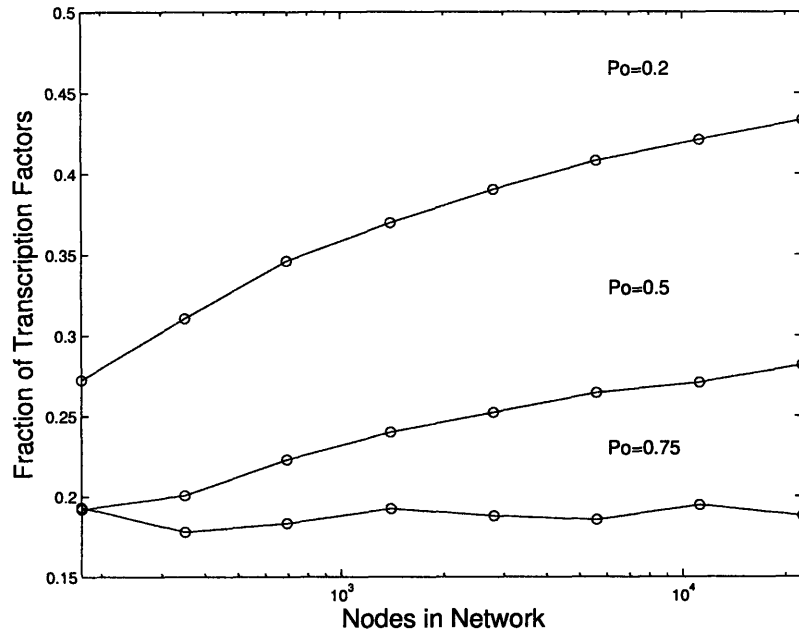


Figure 4-7: Transcription factor fraction versus size of network. P_o indicates link retention probability, ($\rho = 0.7$).

formation. It is clear that the χ does not go to zero as N grows; for connected networks the parameter χ clearly tends toward zero for link retention probabilities $P_o \leq 0.5$ (Figure 4-5). Startlingly, for networks in which the connectivity constraint is relaxed, the parameter χ follows no observable trend! This result is extremely interesting because if our growth model is not self-averaging, it casts doubt on the possibility of obtaining growth parameters from the observed state of the system.

We distinguish between two types of self-averaging. The first type of self-averaging is when the properties of the simulated network do not depend on initial conditions. The second type is when there is no dependence on the growth process. The parameter χ does not distinguish between the two types of self-averaging. Indeed, when the network is connected, the decaying value of χ indicates that variability in the sample space is due to initial conditions, i.e., initial steps in the growth process.

We already know that the distribution of the number of transcription factors in connected networks for smaller P_o is much tighter than exhibited by the Polya's Urn model (Figure 4-6). It is possible that this higher variability in transcription factor number could explain the deviation of χ in Figure 4-8 from the results in Figure 4-5 and [7]. But we will explore further network characteristics before we draw our conclusions. We will also see if it is possible to

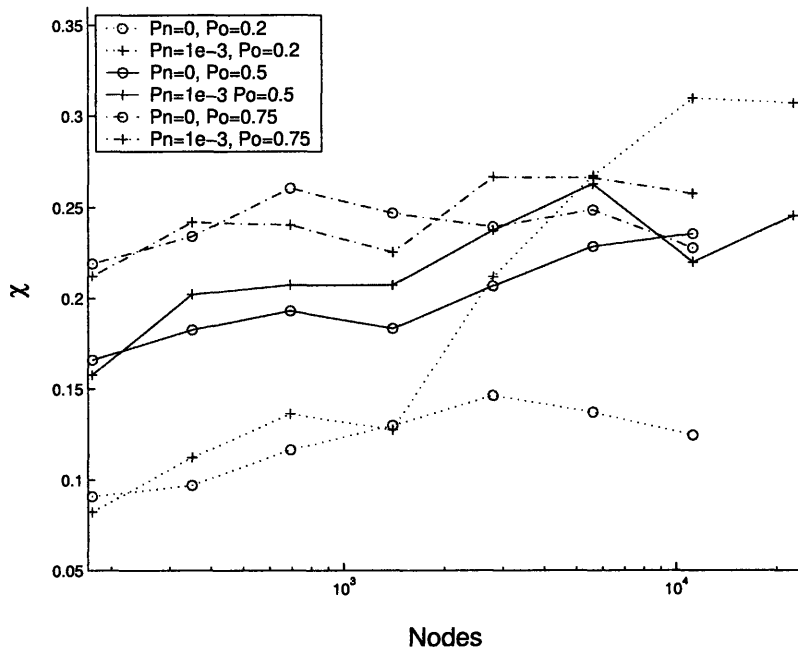


Figure 4-8: χ as a function of number of nodes for unconnected networks. Each point represents 100 realizations.

distinguish the contributions to χ from the initial conditions and from the growth process.

First, we will look at the number of links in the networks for varying P_o and P_n (Figure 4-9). For $P_o = 0.2$ (the dashed curves), the effect of the generation term is large. The three dashed curves correspond to $P_n = [0, 10^{-3}, 10^{-2}]$. There is a clear deviation in the number of links from what is expected without innovative links (largest versus smallest curve). We can see the transition from the dilution regime to the regime of constant density. On the other hand, for $P_o = 0.75$ the effect of the generation term is less pronounced though still visible. The standard deviations for the data points are given, but we still cannot determine anything quantitative about why χ does not tend to zero for $P_o \leq 0.5$.

One hypothesis for the changed character of χ is that the disconnectedness of the network may affect the self-averaging. In Figure 4-10, we plot the number of links L in the total network versus the number of nodes in the giant component. There is a strong correlation between number of links and size of giant component. Furthermore, by examining the simulation results we found that much of the network outside of the main component consists of isolated nodes with no links whatsoever. These isolated or possibly even sparsely connected nodes contribute very little to the total number of links, and they form a disproportionately large fraction for many networks. Thus, we may assume that the large χ for

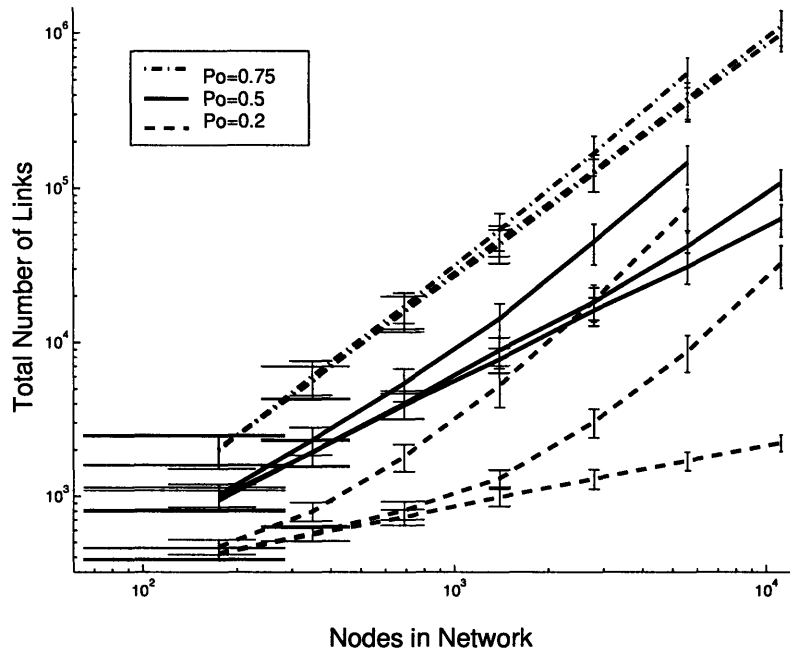


Figure 4-9: Link totals as a function of nodes. P_o is varied as indicated and $P_n = [0, 10^{-3} 10^{-2}]$. Each point represents 100 realizations of the given network parameters. We have fixed the starting density to be $\sim \frac{30}{43} = 0.698$ for all curves in this plot.

the disconnected networks is due to their very disconnectedness.

However, networks with larger generation terms tend to become more connected simply because there are more links in the network. If we believe the large variability in grown networks is due to their disconnectedness, we may expect networks with larger generation terms to have smaller χ . Figure 4-8 has shown that this is not the case. Networks with larger generation terms do not have smaller χ . We explore this further with Figure 4-11.

Figure 4-11 condenses a lot of information into one plot. Each color represents one value of P_n and each symbol represents one value of P_o . By comparing 4-11 a and b, it is clear that for given parameters, larger networks are more connected. This is due to the growing importance of the generation term. In 4-11 a there is a clear correlation between number of links and size of main component, though this correlation grows weaker with increasing P_n and P_o because the network becomes more connected. In the larger network (Figure 4-11 b) there is much less dependence on main component size due to the growing importance of the constant density (link innovation) term, but the variability in link number still does not decrease!

Our conclusion is that the persistence of a high value of χ is due to two different reasons,

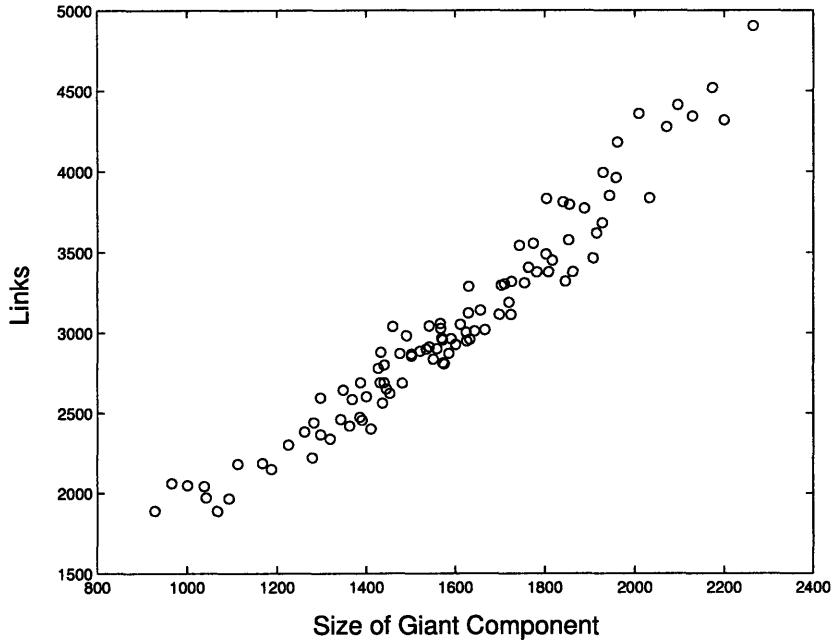


Figure 4-10: Links as a function of size of giant component. The size of the network here is 2800 nodes. The scatter plot shows 100 different network realizations ($P_o = 0.2, P_n = 10^{-3}$)

one for small network size and one for large network size. The disconnectedness of the network leads to a large χ for small networks; the generation term leads to a large χ at larger network sizes.

Clustering Coefficient

The clustering coefficient gives a first approximation of the structure within the networks. Figure 4-12 shows the average clustering coefficient is linearly dependent on initial conditions when the network density is diluting. It also shows that the clustering coefficient is largely independent of the generation term in the diluting region. We may understand this as randomly added links not changing the clustering coefficient greatly. On the other hand, for $P_n = 0.1$, the generation term is so large that its effect overtakes the effect of the initial conditions. This regime is easy to quantify because the number of links is also independent of initial conditions (equilibrated regime). In this regime the number of links and average clustering coefficient are roughly constant as a function of initial density.

Figure 4-13 indicates the clustering coefficient as a function of network size. Each group of curves indicates $P_n = [0, 10^{-3}, 10^{-2}]$. Again, we see that the clustering coefficient depends only upon P_o and initial row density, not P_n . Using Figures 4-12 and 4-13, we determine

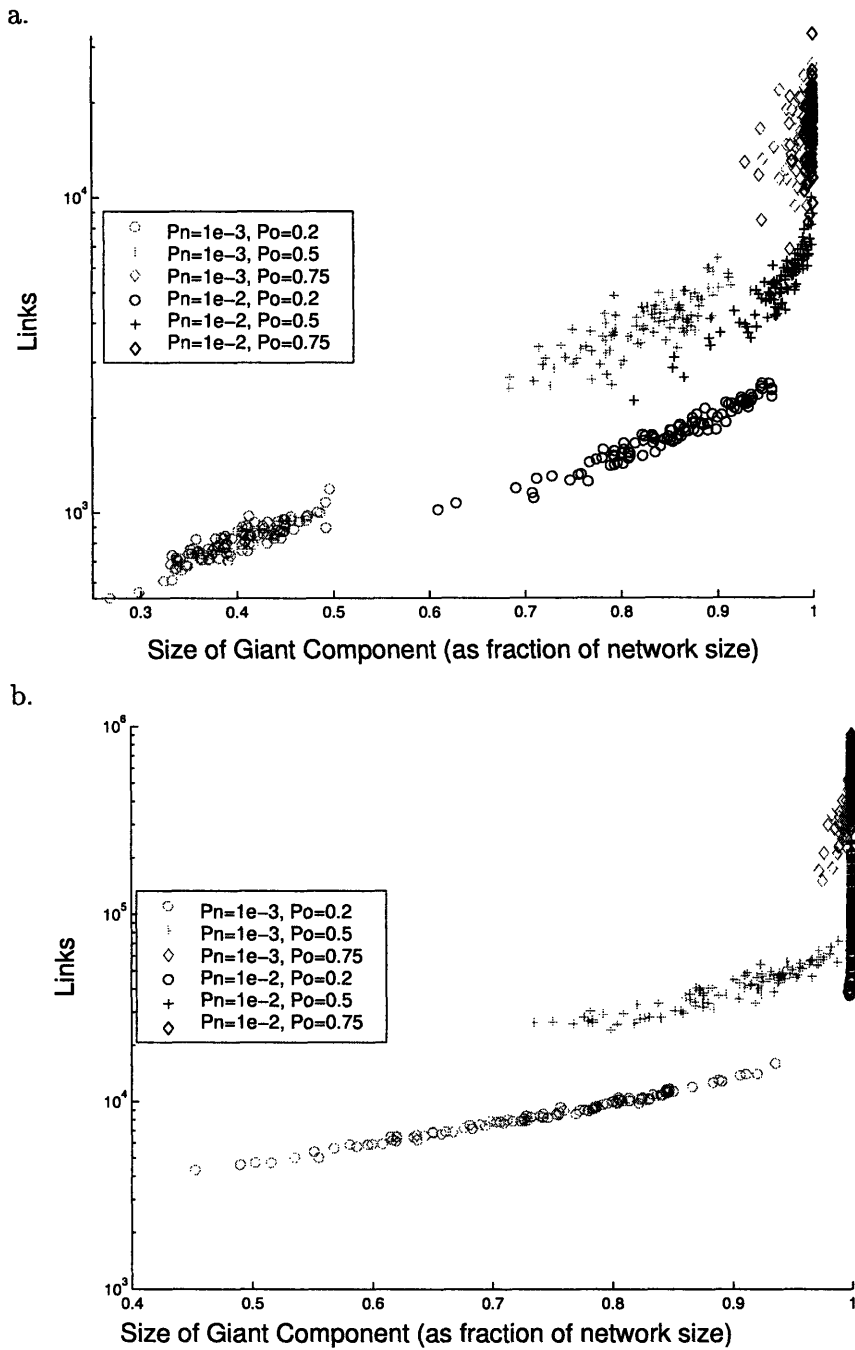


Figure 4-11: Total number of links versus the size of the giant component in the network for a. 691 nodes and b. 5600 nodes.

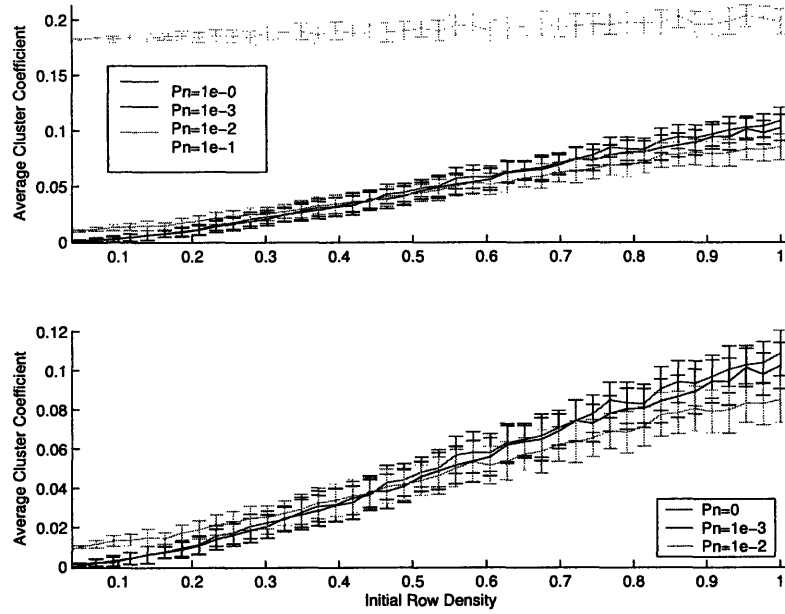


Figure 4-12: Clustering coefficient versus initial row density for $P_o = 0.2$ and nodes= 691.

P_o	C	STD of C
0.2	0.0685	0.0107
0.35	0.1114	0.0183
0.5	0.1820	0.0307
0.6	0.2466	0.0402
0.75	0.3623	0.0679

Table 4.4: Clustering coefficient as a function of P_o ($P_n = 10^{-3}$).

the parameter space in P_o and initial row density that will yield a clustering coefficient of 0.188. We still have the freedom to choose P_n to obtain the right number of links in the network. It will be interesting to see if the parameters put us in the diluting or equilibrated regime.

Parameter Values

We have fit the data summarized in Table 4.4 with a quadratic polynomial and the data in Figure 4-12 by a linear equation. Using these fits, we have determined the parameter space in which the clustering coefficient has the desired value of 0.188 (Figure 4-14).

Given the constraints determined by the clustering coefficient, the simulated networks contain more than the observed number of links ($L = 1190$). No point in this parameter space is consistent with observed network parameters. Though we still retain one free parameter

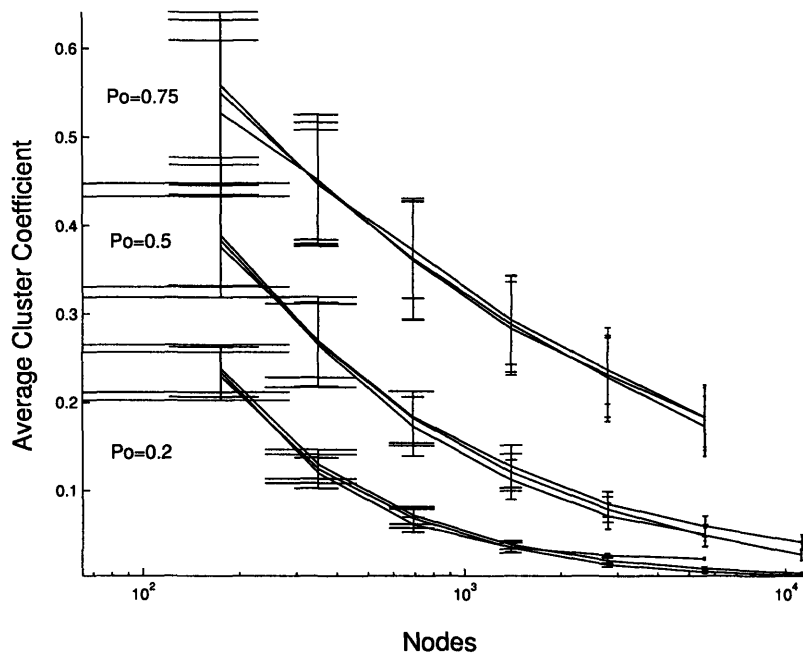


Figure 4-13: Average clustering coefficient versus size of network. Each P_o has three curves plotted for it. These correspond to different values of P_n . We can see the clustering coefficient is largely independent of the generation term.

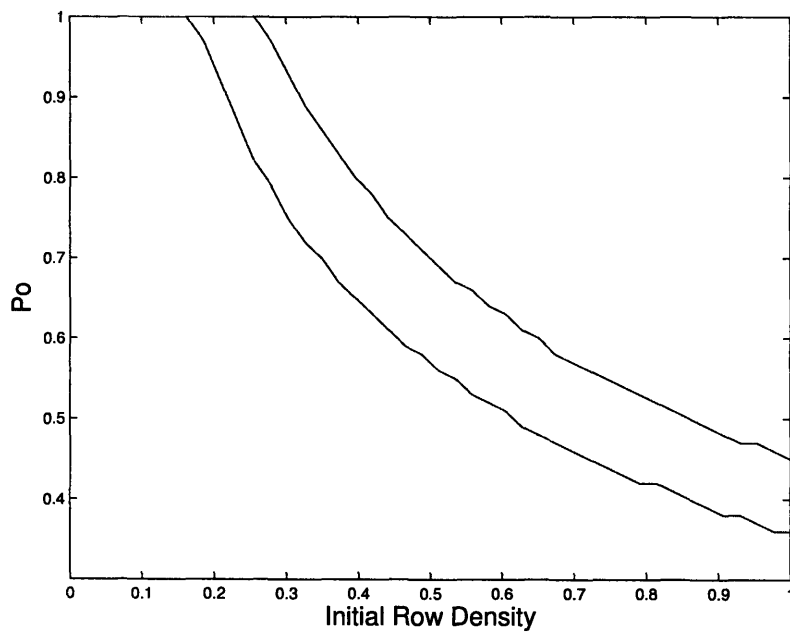


Figure 4-14: The strip shows the allowed parameter space found by clustering coefficient constraints. There are probably effects at the extreme values that will blur the validity of this line such as the generation term, nonlinearities in dependence of clustering coefficient on initial row density.

(P_n) , the generation term cannot save the model because it only adds more links to the simulated networks. There is no free parameter to remove links from the networks.

Chapter 5

Conclusions and Future Work

We have found that duplication and divergence is a key process in the evolution of the gene regulatory network in *E.coli*. Duplication with inheritance and duplication with divergence account for more than 90% of the observed regulatory network fragment in *E. coli*. Though over half of the genes in the real network have no homologs in the network, only 9% of the regulatory interactions involve only genes without homologs.

From these results, we have proposed a simple duplication-divergence model of evolution and found that this model can not account for the structure in the real network. The constraints dictated by the observed clustering coefficient ($\langle C \rangle = 0.188$) define an allowed parameter space in ρ and P_o (leaving P_n unconstrained). We have found that this parameter space is inconsistent with the number of links in the observed network ($L = 1190$) because the simulated networks have too many links. The free parameter P_n can not rescue the situation because it can only increase the number of links in the simulated networks. Intuitively, for a given number of links, the real network is much more clustered than the simulated networks. Further work will determine if a more complicated model can produce the observed network or if there is something inherently missing in a model based solely on duplication and divergence.

In agreement with [7], we have found that simulations of connected networks with $P_o \leq 0.5$ are self-averaging even though we started from different initial conditions and simulated directed networks. It is not immediately obvious that the different cases would have yielded similar results. Furthermore, our calculations of $L(N, P_o, P_n)$ for undirected networks give a good intuitive framework for our simulations of directed networks.

Also, it appears that a duplication-divergence model with innovative link formation does not have a self-averaging regime as we have found no self-averaging in the simulations for $P_n \neq 0$. This raises some concern about the possibility of extracting network parameters from the observed network. We have identified two causes for the lack of self-averaging: lack of connectivity for the diluting regime and the generation term for the equilibrated regime of network growth. Understanding why the generation term leads to large variability in grown networks is an interesting future direction.

Guided by the evidence of convergent evolution in our network from Chapter 3, it is our conclusion that selection accounts for the extra structure within the real network. Future work may try to realize selection in our model as a constraint upon network features.

Bibliography

- [1] Madan Babu, M. & Teichmann, S. A. Evolution of Transcription Factors and the Gene Regulatory Network in *E. coli*. *Nucleic Acids Res.* 31, 1234-1244 (2003).
- [2] Teichmann, S. A. & Madan Babu, M. Gene Regulatory Network Growth by Duplication. *Nat. Genet.* 36, 492-496 (2004).
- [3] Milo, R. et al. Network Motifs: Simple Building Blocks of Complex Networks. *Science* 298, 824-827 (2002).
- [4] Conant G. C. & Wagner, A. Convergent Evolution of Gene Circuits. *Nat. Genet.* 34, 264-266 (2003).
- [5] Barabasi, A. L., Albert, R., Jeong, H. Emergence of Scaling in Random Networks. cond-mat/9910332 (1999).
- [6] Dorogovstev, S. N., Mendes, J. F. F. Evolution of Networks. cond-mat/010614 (2002).
- [7] Ispolatov, I., Krapivsky, P. L. & Yuryev, A. Duplication-Divergence Model of Protein Interaction. q-bio/0411052 (2004).
- [8] Krapivsky, P. L., Rodgers, G. J. & Redner, S. Degree Distributions of Growing Networks. cond-mat/0012181 (2001).
- [9] Sole, R. V. et al. A Model of Large-Scale Proteome Evolution. cond-mat/0207311 (2002).
- [10] Vazquez, A. et al. Modeling of Protein Interaction Networks. cond-mat/0108043 (2001).
- [11] Chung, F., Lu, L., Dewey, T. G., Galas, D. J. Duplication Models for Biological Networks. cond-mat/0209008 (2002).

- [12] Hartl, D. L. *A Primer of Population Genetics* (Sinauer Associates, Sunderland, MA, 2000).
- [13] Gillespie, J. H. *Population Genetics : A Concise Guide* (Johns Hopkins Univ. Press, Baltimore, MD, 2004).
- [14] Raval, A. Some Asymptotic Properties of Duplication Graphs. cond-mat/0307717 (2003).
- [15] Shen-Orr, S. S., Milo, R., Mangan, S. & Alon, U. Network Motifs in the Transcriptional Regulation Network of *Escherichia coli*. *Nat. Genet.* 31, 64-68 (2002).
- [16] Albert, R. & Barabasi, A. L. Statistical Mechanics of Complex Networks. *Reviews Modern Phys.* 74, 47-97 (2002).
- [17] Maslov, S., Sneppen, K. & Eriksen. A. K. Upstream Plasticity and Downstream Robustness in Evolution of Molecular Networks. q-bio/0310028 (2003).
- [18] Itzkovitz, S. et al. Subgraphs in Random Networks. *Phys. Rev. E* 68, 026127 (2003).
- [19] Lynch, M. & Katju, V. The Altered Evolutionary Trajectories of Gene Duplicates. *Trends in Genetics* 20, 544-549 (2004)
- [20] Mangan, S. & Alon, M. Structure and Function of the Feed-Forward Loop Network Motif. *PNAS* 100, 11980-11985 (2003).