Optimal topology of gene-regulatory networks: role of the average shortest path

Ahmed F. Abdelzaher
Department of Computer Science
Virginia Commonwealth University
Richmond, VA 23284
abdelzaheraf@vcu.edu

Thang Dinh
Virginia Commonwealth University
Department of Computer Science
Richmond, VA 23284
tndinh@vcu.edu

Michael L. Mayo
Environmental Laboratory
US Army Engineer Research and Development Center
Vicksburg, MS 39183
Michael.L.Mayo@usace.army.mil

Preetam Ghosh [*]
Virginia Commonwealth University
Department of Computer Science
Richmond, VA 23284
pghosh@vcu.edu

ABSTRACT

Gene regulatory networks (GRNs) possess an important structural property; they are sparse and resilient, with a robust topology that affords protection against random “attacks” (e.g., gene deletions). However, such networks exhibit optimal or near-optimal topological features not present in other scale-free networks. This paper utilizes an integer linear program formulation to gauge the **exact structural optimality** of scale-free networks measured using the average shortest path between transcription factors and the regulated genes of a gene-regulatory network sampled from the *Escherichia coli* bacterium. While randomly generated versions of these networks show several cases for improvement, few subnetworks sampled from *Escherichia coli*’s transcriptional network show optimized solutions that differ substantially from their original topology. We therefore conclude that sampled transcriptional subnetworks from *Escherichia coli* exhibit an “optimal” topology not present in alternative networks. Because these analyses do not consider the biology of expression dynamics and are based on topology alone, other communication systems, such as wireless networks, may benefit from a more detailed examination of the role in which the average shortest path affects system function, such as with noise or other signaling disruptions.

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Biology and genetics; G.1.6 [**Mathematics of Computing**]: Numerical Analysis—*Optimization*; G.2.2 [**Discrete Mathematics**]: Graph Theory—*Graph algorithms*

General Terms

Gene-regulatory networks, Graph theory, Linear Programming, Optimization

Keywords

Average shortest path; transcriptional network motif; feed-forward loop; scale-free networks

1. INTRODUCTION

In a gene-regulatory network (GRN), the protein-coding genes of DNA are abstracted as the nodes of a directed graph, with the interconnecting links associated with either a stimulatory (up-regulating), inhibitory (down-regulating), or mixed (both up-regulating and down-regulating) effect on gene expression, and therefore, cellular protein abundance [1]. An exciting capability of GRNs is that they resist deleterious effects from external disruptions–a property termed as ’biological robustness’ [2], such as the ability to mitigate dynamical effects of noise in protein-coding gene expression [3]. While previous research has accredited some of this capability to network topology, there stands an under-explored question: which topological features contribute to this robustness, and what is its underlying mechanism?

Gene-regulatory networks fall within a class of scale-free networks [4], in which their topological degree distribution follows a power-law equation [5]. This feature is associated with loosely connected modules that support fewer genes with larger degree; such genes are distinguishably known as “hubs” or ”hub nodes” [5]. Conversely, the majority of networked genes possess degrees much lower than the average. This aspect attributes resilience to the deleterious effects of genetic mutations, because highly connected master transcriptional regulators occur rarely in these networks; however, targeted suppression of these proteins, possibly by manipulating gene activity, may decouple their influence from the network and lead to a loss in organism-level function. Additionally, connectivity of some transcriptional network motifs, such as the feed-forward loop, to the embed-

[*] Author to whom correspondence should be directed.

BICT 2015, December 03-05, New York City, United States

DOI 10.4108/eai.3-12-2015.2262473

ding network may contribute substantially to its topological properties, such as the average shortest path and clusterring coefficient [6, 7]. Aside from these topological characteristics, feed-forward loops admit dynamical characteristics linked with their topology, such as an ability to generate pulses [8], admit signal delays and irreversible speed-ups [8], or respond differentially to dynamically acute concentration fluctuations [9].

In this paper, we investigate how another topological metric–the average shortest path between any two networked nodes–is affected by different scale-free architectures with identical degree distribution and feed-forward loop abundance. In this context, We show that a gene-regulatory network of the *Escherichia coli* (*E. coli*) bacterium achieves a near-optimal level of robustness against changes to the average shortest path. This finding was not observed for the randomized networks considered here.

2. NETWORK DATASETS

2.1 Transcriptional Regulatory Networks

The *E. coli* bacterium thrives in the mammalian digestive tract, and is a widely-used model prokaryotic organism [10], and its GRN is considered a prototypical scale-free network. We have used the GeneNetWeaver software package [11] to obtain the largest-connected component of *E. coli*'s GRN. We then derived a transcriptional-regulatory network (TRN) from this GRN by pruning the gene-gene interactions, leaving transcription factor (TF)-gene and TF-TF interactions. We term a "regulated gene" as one that does not regulate another, although TFs may regulate both other TFs and regulated genes. In this sense, the transcriptional regulatory network admits a regulation hierarchy with TFs near the root(s) and regulated genes as its leaves. Collectively, the *E coli* TRN so obtained has 23 connected components, 1565 transcription factors and regulated genes, and 3758 directed edges.

A number of subnetworks were sampled randomly from this transcriptional network, each with a number of nodes $n = 5, 10, 15, 20, 25, 30$, using the GeneNetWeaver software package. Networks were sampled in a way that ignored any auto-regulatory edges, as these did not contribute to the optimization metric (see below). Additionally, 10 replicates for each network size, n, were sampled for a total of 300 unique transcriptional-regulatory networks.

2.2 Randomized networks

Each sampled transcriptional-regulatory network was "randomized" using the following algorithm. We began with a number of disconnected nodes n, and $|E|$-many edges from each sampled *E. coli* network. Next, we selected a node pair at random with uniform probability from the set of all possible pair-wise combinations between distinct nodes, and drew a directed edge between them. This node-pair was then removed from the set of available pairs. This procedure was iterated until $|E|$-many edges were drawn.

3. METHODS

3.1 Metrics and criteria for optimal topology

We adopt a version of the average shortest path, $\langle d \rangle$, as a measure for network robustness, which has been widely used as a robustness metric for many complex systems [12]. In this paper, $\langle d \rangle$ is defined as the distance, in number of "hops" (i.e., single movements between adjacent nodes) required to connect a transcription-factor node to a regulated gene, and averaged over all such pairs in a network. We restrict our analyses to the connected paths from transcription factors to genes. With these conditions, $\langle d \rangle$ can be expressed by the equation:

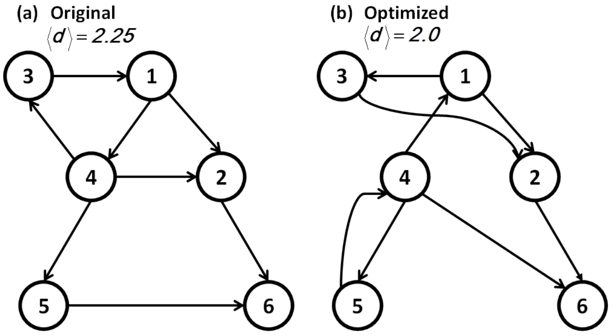

Figure 1: (a) An exemplary random network and (b) it's post-optimization output from the integer linear programming algorithm. The in- and out-degrees of each node in both are identical, and the existence of routes from the TFs (nodes 1-5) to their target regulated gene (node 6) have been preserved. In this example, the optimized network (b) exhibits a smaller average shortest path, $\langle d \rangle$, than the exemplary network (a).

$$\langle d \rangle = \frac{1}{p} \sum_{i}^{n_t} \sum_{j}^{n_g} d_{ij}, \text{ with } d_{ij} < \infty, \tag{1}$$

wherein d_{ij} denotes the shortest path between nodes i and j, p the number of connected TF-gene pairs, n_t the number of TFs, and n_g the number of genes in the network. The type of the regulatory interaction, either up- or down-regulating, was not considered in Eq. (1).

We note that $\langle d \rangle$ is slightly different from other definitions of the average shortest path, wherein p has been taken as the number of connected pairs without regard for their type. However, the p of 1 is distinct from these by its consideration of biology, that it reflects consideration of only a subset of the total number of possible paths by eliminating any contributions to $\langle d \rangle$ between regulated genes.

An optimal topology is one wherein a metric of the topology is at an extremum. As explained in the next section, this criteria is the average shortest path, $\langle d \rangle$, for a given network $G(V, E)$, wherein V is the set of all networked nodes (TFs and regulated genes) and E denotes the set of edges connecting them. An output of the integer linear programming algorithm (explained below) is an "optimized" version of G, $G_0(V, E_0)$, which hosts an identical set of vertices, number of edges, $|E| = |E_0|$, identical degree distribution as G, and the same number (and type) of feed-forward loop transcriptional network motifs, but with potentially different average shortest path, $\langle d_0 \rangle$. To determine whether the topology of G is optimal, we compare $\langle d \rangle$ to that of its optimized coun-

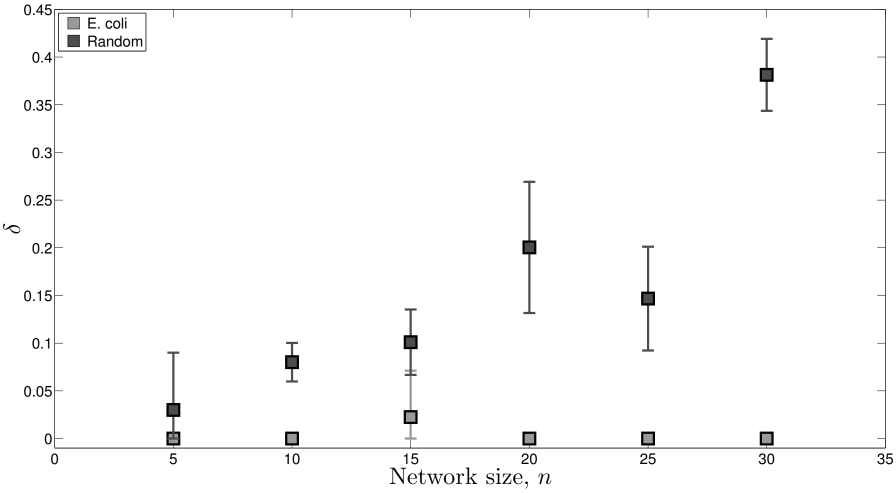

Figure 2: δ (Eq. (2)) evaluated using the integer linear programming method, for transcriptional-regulatory subnetworks sampled from *E. coli* (light grey boxes) of size $n = 5, 10, 15, 20, 25$, and 30, compared against optimal solutions found from randomized versions of these networks (dark grey). If the input network is optimal, then $\delta = 0$.

terpart, $\langle d_0 \rangle$, using the following metric, δ:

$$\delta = \frac{|\langle d \rangle - \langle d_0 \rangle|}{\langle d \rangle}, \tag{2}$$

Therefore, $\delta = 0$ corresponds to the result that G exhibits a perfectly optimal topology.

3.2 Optimizing Topology with Integer Linear Programming

Finding a network with optimal topology is a challenging problem. On the one hand, a brute-force search approach is intractable due to an exponentially large search space on the order of $\mathcal{O}(2^{|V|^2})$-many different graphs. On the other hand, meta-heuristic methods, such as Simulated Annealing [13] and Genetic Programming [14, 15, 16], cannot guarantee an optimal solution. To address this problem, we propose a new approach to finding an optimal graph topologies for small and moderate size networks based on integer linear programming (ILP).

Many graph-theoretical problems are solvable using integer linear programming [17], such as the shortest path, vertex coverage, maximum flow, and minimum cost-flow problems. This is possible, because these problems can be expressed in terms of linear relationships which together form a polytope enclosed by their intersections. From this polytope, it is possible to identify an extremum of a cost function. Moreover, ILP can be useful to identify cases wherein solutions are not feasible with other methods, and their implementation can be facilitated using freely available academic software, such as IBM ILOG CPLEX optimizer [18].

We consider a linear program which identifies a new graph G_0 with minimum $\langle d_0 \rangle$ based on an input G, subject to the following constraints:

1. Connectivity between a TF and a regulated gene "target' must be preserved from G, which ensures invariance of the overall paths, but not necessarily the path-lengths;
2. Degree distributions between G and G_0 are identical;
3. The number and type of feed-forward loop transcriptional network motifs remain invariant between G and G_0, despite any variance in topology.

In particular, we hold the number of feed-forward loops constant during the optimization process, because we have previously identified that path-length metrics may be significantly affected if feed-forward loop transcriptional motifs are "deleted" from the network topology [6]. The detailed equations for the linear constraints have not been included due to space restrictions.

An example of an "optimized" 6-node network is given in Figure 1. The optimal solution (Fig. 1(b)) preserves connections between TF-gene pairs, in- and out-degrees for each node, and the number of feed-forward loop transcriptional motifs, but exhibits a $\langle d \rangle$ smaller by approximately 11%.

4. RESULTS AND DISCUSSION

Figure 2 illustrates results of the integer linear programming based optimization of sampled transcriptional-regulatory network topologies. Here we observe that very few of the subnetworks sampled from *E. coli*, of any size, exhibit an average path length in their optimized topologies that is smaller than already supported by these networks. This result should be contrasted by our attempts to optimize randomized networks, which demonstrate that average shortest path lengths computed for optimized topologies were significantly reduced over their non-optimized input network.

These results indicates that *E. coli* network topologies are already well-suited to minimize the average shortest path between transcription factors and their regulated genes. There may be plausible reasons why an evolved transcriptional-regulatory network may experience pressure to minimize the number of regulatory interactions between the regulating proteins and the terminal genes. Common modes of genetic evolution, such as gene duplication and divergence, alters the degree and connectivity of networked protein-coding genes; while these are manifestly local topological alterations, the path-length encompasses regulatory interactions that transcend a gene's local neighborhood, suggesting that network dynamics manifest with system-level properties might play a role in whether an organism's progeny survives. Although dynamics may play a role in correlating topological characteristics at network scales, there is evidence that node-degrees in both biological and non-biological networks are correlated by a geodesic distance of approximately three steps [19]; however, it is as-yet unclear what general mechanism underlies such a correlative relationship.

5. CONCLUSIONS

In this paper we used an integer linear programming based optimization method to determine a topology for a given (directed) transcriptional-regulatory network under the constraints that: (i) any paths present between transcription factors and genes remain in the optimized network; (ii) that degree distributions of the optimized network remain invariant; and (iii) that the number of feed-forward loop transcriptional motifs remain invariant. By using this method with sampled subnetworks from the transcriptional-regulatory network obtained from the *E. coli* bacterium, we found that *E. coli*'s subnetworks exhibited topologies that already minimized the average shortest path length between any two transcription factor and regulated gene of the network. In contrast, randomized versions of these biological networks were not optimized, in the sense that the integer linear program identified alternative network topologies that further reduced the network's average shortest path length.

Although the integer linear programming formulation reveals differences in topological optimality between the considered TRNs and random networks, it may experience significant difficulties when analyzing larger networks, because the relatively large number of constraints required by this method makes it infeasible. For example, the largest reported example we are aware of involved a 150 node network with over 10^6 constraints [20], which leads to an excessive convergence time. It may therefore be beneficial to develop a heuristic that aims to reduce this computational complexity.

6. ACKNOWLEDGMENTS

This research was funded by the US Army's Environmental Quality and Installations 6.1 basic research program. Opinions, interpretations, conclusions, and recommendations are those of the author(s) and are not necessarily endorsed by the U.S. Army. It was also partially supported by NSF.

7. REFERENCES

[1] Mark E. J. Newman. *Networks: An Introduction.* Oxford University Press, 2010.

[2] Hiroaki Kitano. Biological robustness. *Nature Reviews Genetics*, 5:826–837, 2004.

[3] R. J. Prill, P. A. Iglesias, and A. Levchenko. Dynamic properties of network motifs contribute to biological network organization. *PLoS Biology*, 3:1881, 2005.

[4] Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. 2009.

[5] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Error and attack tolerance of complex networks. *Nature*, 406:378–382, 2000.

[6] Ahmed F. Abdelzaher, Michael L. Mayo, Edward J. Perkins, and Preetam Ghosh. Contribution of canonical feed-forward loop motifs on the fault-tolerance and information transport efficiency of transcriptional regulatory networks. *Nano Communication Networks*, 6(3):133–144, 2015.

[7] B. K. Kamapantula, A. Abdelzaher, P. Ghosh, M. L. Mayo, E. J. Perkins, and S. K. Das. Leveraging the robustness of genetic networks: a case study on bio-inspired wireless sensor network topologies. *J. Ambient Intelligence and Humanized Computing*, 5(3):323–339, 2014.

[8] Shai S. Shen-Orr, Ron Milo, Shmoolik Mangan, and Uri Alon. Network motifs in the transcriptional regulation network of *Escherichia coli. Nature Genetics*, 31:1061–4036, 2002.

[9] M. Mayo, A.F. Abdelzaher, E.J. Perkins, and P. Ghosh. Top-level dynamics and the regulated gene response of feed-forward loop transcriptional motifs. *Physical Review E*, 90(3):032706, 2014.

[10] P. Singleton. *Bacteria in biology, biotechnology, and medicine (5th ed.).* Wiley, 1999.

[11] Thomas Schaffter, Daniel Marbach, and Dario Floreano. Genenetweaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27(16):2263–2270, 2011.

[12] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47–97, 2002.

[13] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[14] Bogdan TomoiagÄČ, Mircea ChindriÅ§, Andreas Sumper, Antoni Sudria-Andreu, and Roberto Villafafila-Robles. Pareto optimal reconfiguration of power distribution systems using a genetic algorithm based on nsga-ii. *Energies*, 6(3):1439, 2013.

[15] Melanie Mitchell. *An Introduction to Genetic Algorithms.* MIT Press, Cambridge, MA, USA, 1998.

[16] S. Ghosh, P. Ghosh, K. Basu, and S. K. Das. Gama : An evolutionary algorithmic approach for the design of mesh-based radio access networks. *Local Area Networks*, pages 374–381, Nov. 2005.

[17] R.K. Ahuja, T.L. Magnanti, J.B. Orlin, and Sloan School of Management. *Network Flows.* Working paper (Sloan School of Management). Alfred P. Sloan School of Management, Massachusetts Institute of Technology, 1988.

[18] IBM ILOG CPLEX Optimizer. `http://www-01.ibm.com/software/integration/optimization/cplex-optimizer/`, 2010.

[19] M. Mayo, A.F. Abdelzaher, and P. Ghosh. Long-range degree correlations in complex networks. *Computational Social Networks*, 2(1):1–13, 2015.

[20] Thang N. Dinh and My T. Thai. Precise structural vulnerability assessment via mathematical programming. *MILCOM*, pages 1351–1356, Nov. 2011.