Global Information Processing in Gene Networks: Fault Tolerance

Frank Emmert-Streib[1]
Stowers Institute for Medical Research
1000 E. 50th Street
Kansas City, MO 64110, USA
v@bio-complexity.com

Matthias Dehmer
Institute of Discrete Mathematics and Geometry
Vienna University of Technology
Wiedner Hauptstrasse 8-10
A-1040 Vienna, Austria
mdehmer@geometrie.tuwien.ac.at

ABSTRACT

In this paper we study the fault tolerance of gene networks. We assume single gene knockouts and investigate the effect this kind of perturbation has on the communication between genes globally. For our study we use directed scale-free networks resembling gene networks, e.g., signaling or protein-protein interaction networks, and define a Markov process based on the network topology to model communication. This allows us to evaluate the spread of information in the network and, hence, detect differences due to single gene knockouts in the gene-gene communication asymptotically.

Keywords

Information Theory, information processing, gene networks, scale-free networks, robustness.[1]

1. INTRODUCTION

In recent years, networks have been studies numerously [2, 1, 14, 15, 16, 18]. In contrast to classical approaches dealing mainly with random networks [8, 9] the last decade was dedicated to the study of small-world [21, 20] and scale-free [2, 11] networks or combinations thereof [19]. The rapid increase in interest studying complex networks in general can be explained by the fact that many real world phenomena can be modeled within the frame of either of these non-random network topologies. For example, the World-Wide Web, the Internet or biological networks are examples these networks are playing key roles [2, 1, 14, 19].

Recently, ALBERTS et al. [3] studied the error and attack tolerance of complex networks and compared random and scale-free networks, e.g., the World-Wide Web or the Internet. By using purely graph theoretical measures they found that scale-free networks are much more robust against random errors than random networks but more vulnerable against directed attacks. In this paper we extend this analysis regarding two important points. First, we use directed scale-free networks because we are interested in the fault tolerance of gene networks, e.g., transcriptional regulatory, signaling or protein-protein interaction networks, which are directed. Second, we introduce a measure that aims to capture the perturbed communication abilities of directed scale-free networks information theoretically [4, 22] due to structural modifications of the network [5, 7].

[1]Present address: University of Washington, 1705 NE Pacific St, Box 357730, Seattle, WA 98195-5065, USA

BIONETICS '07, December 10-13, 2007, Budapest, Hungary

2. MOTIVATION

In this paper we are interested in the analysis of the effect single gene knockouts, which are a special form of network perturbation, have on the communication abilities of directed scale-free networks resembling gene networks. More precisely, we are interested to study the global effect of these perturbations on a systems level. The information we want to gain from this investigation is twofold. First, we want to learn if there are genes that are more vulnerable regarding single gene knockouts than others and, second, what does this mean in terms of local properties of these genes with respect to the network topology. The second question addresses a connection between global and local properties because we measure the perturbed communication regarding all genes whereas locally we consider only a small neighborhood around genes sending signals.

3. METHODS

3.1 Generate directed scale-free network

The algorithm we use to generate a network that has scale-free in- and out-degree distribution is very similar to an algorithm that has been recently introduced [10].

1. Start with n_0 unconnected nodes.
2. Add one new node to the network.
3. Connect it to e ($\leq n_0$) nodes from the existing network. A node is chosen based on the out-degree distribution
$$p_i^{out} = \frac{k_i^{out}}{\sum_j k_j^{out}}. \quad (1)$$
The direction of a new edge is 'to' the new node.
4. Connect the new node to e nodes in the existing network. A node is chosen based on the in-degree distri-

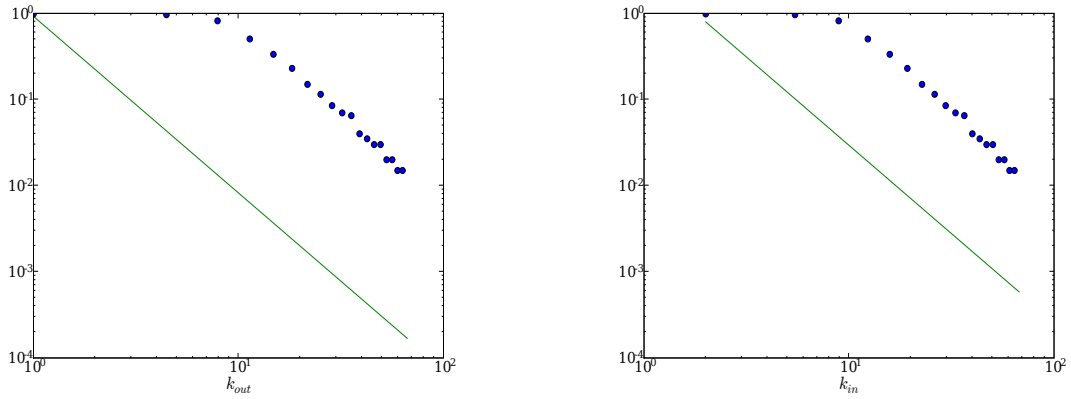

Figure 1: Cumulative Out-degree (left) and in-degree distribution (right) of a scale-free network with $|V| = 200$ nodes and $e_d = 0.03$. The straight line in both figures corresponds to a power-law $P(k) \sim k^{-\gamma}$ with exponent $\gamma = 2.3$.

bution

$$p_i^{in} = \frac{k_i^{in}}{\sum_j k_j^{in}}. \quad (2)$$

The direction of a new edge is 'from' the new node.

5. If the number of nodes in the networks is N stop, otherwise go to step (2).

The algorithm to generate a directed scale-free network has several free parameters. The total number of nodes in the network $|V|$, the number of initial nodes n_0 in the unconnected network and the number of edges e that are added each iteration step. Because e has to be $\leq n_0$ [10] there is a trade-off between the resulting edge density (e_d) and the mean number of nodes that is in the final network unconnected. Choosing $n_0 = e = 1$ would lead to a fully connected network, however, the edge density is very low compared to real networks. Because we are aiming to study gene networks we choose these parameters higher to obtain a more realistic edge density. To ensure connectedness we search in the final network the nodes with zero in- and zero out-degree and connect them with one connection to a node randomly chosen from the remaining network. Because the mean number of unconnected nodes for $|V| = 200$ is of order $O(1)$ these additional edges do not destroy the overall degree distributions for the out- and in-degrees.

We want to point out that the property of a network being *scale-free* does not determine other network properties, e.g., $< k_{in} >, < k_{out} >$ or the edge density. That means, if one has additional requirements one needs to define an appropriate stochastic process that produces a growing network with the desired properties correspondingly. Fig. 1 shows the cumulative out- (left) and in-degree (right) distributions for one scale-free network with $|V| = 200$ nodes and $e_d = 0.03$ (3% of all possible connections). The straight line in both figures corresponds to a power-law $P(k) \sim k^{-\gamma}$ with exponent $\gamma = 2.3$. As previous studies indicate, many real networks have a scaling exponent between 2 and 3 [1, 14].

3.2 Markov process

We define a Markov process by using a given network topology G and the plausible assumption that all possible interactions are equal likely. Plausible in this context does not necessarily mean that this corresponds best to the real situations of, e.g., protein-protein interaction or signaling networks, but to the most simple assumption one can make. Because G is a directed graph it is possible that the resulting Markov process is not ergodic despite the connectedness of G. This means that not all nodes in the network can be reached via a path from other nodes. For simplicity, we assume the Markov process to be of first-order

$$T(X_{n+1} = j | X_n = i_n, ..X_1 = i_1) = T(X_{n+1} = j | X_n = i_n).$$

Definition 1. A Markov process of first-order for a network $G = (V, E)$ is defined by

$$T(X_{n+1} = j | X_n = i) = \begin{cases} \frac{1}{k_i^{out}} & : \quad k_i^{out} > 0 \ \& \ E_{ij} = 1 \\ 0 & : \quad \text{else} \end{cases} \quad (3)$$

for all $i, j \in V$.

3.3 Global effect of the perturbation

In this paper we study the effect single gene knockouts have on the information processing in gene networks. To detect this influence we define a global information theoretic measure. By global effect we define the deviation of the unperturbed (or normal) state from the perturbed state caused by deletion of gene k. Concretely, we measure this effect as the relative entropy also known as Kullback-Leibler (KL) distance D [13, 12] between the stationary distribution of these two states given by

Definition 2.

$$D_{ik} = D(p_{i,k}^{p,\infty} || p_i^{n,\infty}) = \sum_m p_{i,k}^{p,\infty}(m) \log \frac{p_{i,k}^{p,\infty}(m)}{p_i^{n,\infty}(m)}. \quad (4)$$

The stationary distribution for $p_{i,k}^{p,\infty}$ and $p_i^{n,\infty}$ is obtained by

$$p^{\infty} = \lim_{t \to \infty} T^t p_i^0. \quad (5)$$

We want to remark that we start for both distributions from the same initial distribution which depends on i as explained below. Here $p_{i,k}^{p,\infty}$ is the stationary distribution of the perturbed (p) network obtained by knocking out gene k and

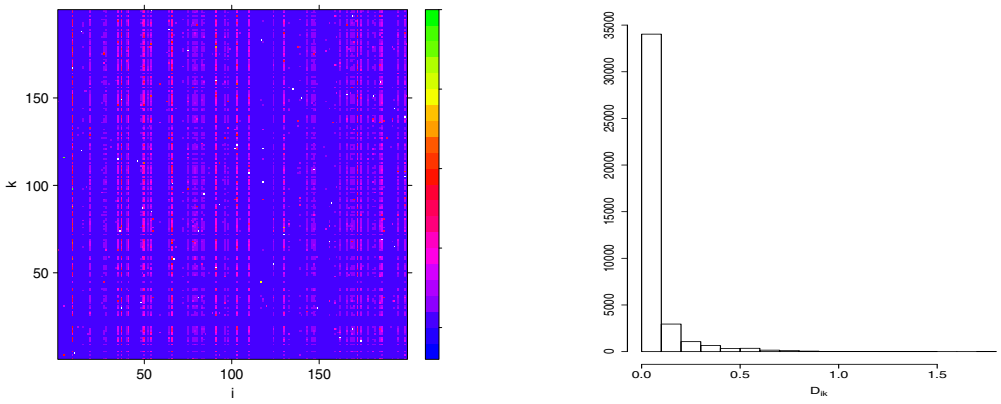

Figure 2: Results for a directed scale-free network with $|V| = 200$ nodes. Left: Color-coded D_{ik} values. Blue corresponds to low and green to high values. Right: Histogram of the D_{ik} values.

starting from the initial distribution $p_i^0 = \delta_{i,m}$[2] $(\forall\, m \in V)$. The interpretation for the unperturbed distribution $p_i^{n,\infty}$ is correspondingly. It is important to use $|V|-1$ (starting from k is excluded) different initial distributions p_i^0 because the underlying Markov process T might not be ergodic, hence, different initial distributions might give different asymptotic results. That means Eq. 4 is a matrix whose components D_{ik} correspond to deletion of gene k and initial distribution $p_i^0 = \delta_{i,m}$. The diagonal elements of D_{ik} are not defined.

4. RESULTS

In Fig. 2 we show the obtained results for a directed scale-free network with $|V| = 200$ nodes. The left figure shows the components of D_{ik} color-coded, whereas blue represents a low and green a high relative entropy. Apparent patterns in this figures are the vertical strips. They indicate that starting from gene i, that has such a vertical strip, a perturbation of almost any other gene will heavily perturb the resulting stationary distribution. This can be interpreted as fragility of gene i regarding general single gene perturbations. The number of fragile genes in dependence of a thresholding parameter β corresponding to the severity of the perturbation is shown in Fig. 3. Here the number of effected genes N_e is defined by

$$N_e(\beta) = \sum_i \Theta\Big(\sum_k \Theta(D_{ik} - \beta) - \frac{1}{2|V|}\Big). \tag{6}$$

The thresholding β parameter is in $[0, 1]$ and $\Theta()$ corresponds to the theta-function. The term $\frac{1}{2|V|}$ indicates that we consider only genes as effected if at least half of all possible knockouts k perturb the system in a way that $D_{ik} > \beta$. From Fig. 3 one can see that about 20 genes are very vulnerable (for $\beta = 0.1$) against gene knockouts. That means, about 10% of all genes seem to be the bottlenecks regarding communication failures in the system.

[2]Kronecker delta.

Now we want to investigate if the number of single gene perturbations that has significant influence on a gene is correlated to the in- and out-degrees of the gene. For this reason we determine

$$N_i = \sum_k \Theta\big(D_{ik} - \beta\big), \tag{7}$$

for $\beta = 0.1$ and compare it with a corresponding in- and out-degree vectors. We use Spearman's rank correlation coefficient [17] and test for correlation between the ranked vectors. For the in- and out-degree vectors we obtain p-values below a significance level of $\alpha = 0.05$ indicating that, e.g., high degrees correspond to many perturbed genes. These results seem plausible considering the following situation: Suppose we have a gene that is connected to all other genes (outgoing edges). It is clear, that an arbitrary knockout of a single gene effects with probability one an outgoing edge of this gene. Hence, this knockout will have an influence on the information processing of this gene. The strength of this influence can not be easily predicted given just this information, however, the point is that we have an influence with probability one. Instead, a gene having very few outgoing connections has a lower probability that a single knockout effects one of its outgoing edges ($Pr = k_{out}/N_p$ with N_p the number of genes that can be perturbed). However, it is possible that the knocked out gene destroys some communication paths (secondary- or even higher-order effect if measured as Dijkstra distance [6]) and, hence, can still have a strong impact on the information processing. It seems to be reasonable to assume that the further away the knockout gene is from the starting gene (in Dijkstra distance [6]) the less the impact will be. This is a strong indicator that information processing on a systems level (that means global) depends crucially on the information processing in a local environment of the gene that sends the information. We want to remark that in our analysis the number N_i, given in Eq. 7, of single gene perturbations that has significant

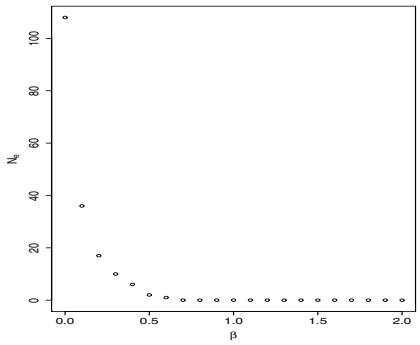

Figure 3: The number of fragile genes in dependence of a thresholding parameter β corresponding to the severity of the perturbation.

influence on a gene is a global measure, whereas the in- and out-degree vectors are local measures.

5. CONCLUSIONS

In this paper we analyzed the influence single gene perturbations have on the global communication abilities of directed scale-free networks resembling gene networks. We defined an information theoretic measure to quantify differences between perturbed and unperturbed communication as relative entropy between the corresponding stationary distributions resulting from Markov processes. We found that there are about 10% of the genes in the network that are very sensitive against single gene perturbations. Further, we showed by a statistical test that there is a correlation between the grade of sensitifity of genes and their in-and out-degrees. This result estabishes a connection between global (the grade of sensitifity against perturbations) and local (in- and out-degrees) properties regarding the communication of information within networks and, hence, suggests that the effects of perturbation detectable globally might be understood by studying the local environment, corresponding to a subgraph, of the gene that sends the information initially. Regarding the fault tolerance of gene networks our results raise the question of the *importance* of genes with high in- and out-degrees. In future studies we will investigate the interplay between local and global effects perturbations have on the communication within the network in more detail to shed further light on this important problem.

6. ACKNOWLEDGMENTS

We would like to thank Chris Seidel and Dongxiao Zhu for fruitful discussions. Matthias Dehmer has been supported by the European FP6-NEST-Adventure Programme, contract n° 028875.

7. REFERENCES

[1] R. Albert and A. Barabasi. Statistical mechanics of complex networks. *Rev. of Modern Physics*, 74:47, 2002.

[2] R. Albert, H. Jeong, and A. L. Barabasi. Diameter of the world wide web. *Nature*, 401:130–131, 1999.

[3] R. Albert, H. Jeong, and A.-L. Barabasi. Error and attack tolerance of complex networks. *Nature*, 406:378–482, 2000.

[4] T. Cover and J. Thomas. *Information Theory*. John Wiley & Sons, Inc., 1991.

[5] M. Dehmer. A novel method for measuring the structural information content of networks. *Cybernetics and Systems*, 2007. accepted.

[6] E. Dijkstra. A note on two problems in connection with graphs. *Numerische Math.*, 1:269–271, 1959.

[7] F. Emmert-Streib and M. Dehmer. Information theoretic measures of UHG graphs with low computational complexity. *Applied Mathematics and Computation*, 190(2):1783–1794, 2007.

[8] P. Erdös and A. Rényi. On random graphs. *Publicationes Mathematicae*, 6:290–297, 1959.

[9] P. Erdös and A. Rényi. On random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, 5:17, 1960.

[10] K. Iguchi, S. Kinoshita, and H. Yamada. Boolean dynamics of kauffman models with a scale-free network. *J Theor Biol.*, 247(1):138–51, 2007.

[11] H. Jeong, B. Tombor, Z. N. Albert, R. Olivai, and A. L. Barabasi. The large-scale organization of metabolic networks. *Nature*, 407:651–654, 2000.

[12] S. Kullback. *Information theory and statistics*. Wiley, 1959.

[13] S. Kullback and R. Leibler. On information and sufficiency. *Ann. Math. Stat.*, 1951.

[14] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.

[15] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabasi. Hierarchical organization of modularity in metabolic networks. *Science*, 297:1551–1555, 2002.

[16] B. Schwikowski, P. Uetz, and S. Fields. A network of proteinâĂŞprotein interactions in yeast. *Nat. Biotechnol.*, 18:1257–1261, 2000.

[17] D. J. Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures*. RC Press, Boca Raton, FL, 3rd edition, 2004.

[18] L. A. Soinov, M. A. Krestyaninova, and A. Brazma. Towards reconstruction of gene networks from expression data by supervised learning. *Genome Biology*, 2003.

[19] V. van Noort, B. Snel, and M. A. Huymen. The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO reports*, 5(3):280–284, 2004.

[20] D. Watts. *Small Worlds: The Dynamics of Networks between Order and Randomness*. Princeton University Press, 1999.

[21] D. Watts and S. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.

[22] R. Yeung. *A first course in information theory*. Springer, 2002.