Lethality and Entropy of Protein Interaction Networks

Thomas Manke [1] **Lloyd Demetrius** [1 2]
manke@molgen.mpg.de ldemetr@fas.harvard.edu
Martin Vingron [1]
vingron@molgen.mpg.de

[1] Max Planck Institute for Molecular Genetics, Ihnestr.73, 14195 Berlin, Germany
[2] Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA02138, USA

Abstract

We characterize protein interaction networks in terms of network entropy. This approach suggests a ranking principle, which strongly correlates with elements of functional importance, such as lethal proteins. Our combined analysis of protein interaction networks and functional profiles in single cellular yeast and mulit-cellular worm shows that proteins with large contribution to network entropy are preferentially lethal. While entropy is inherently a dynamical concept, the present analysis incorporates only structural information. Our result therefore highlights the importance of topological features, which appear as correlates of an underlying dynamical property, and which in turn determine functional traits. We argue that network entropy is a natural extension of previously studied observables, such as pathway multiplicity and centrality. It is also applicable to networks in which the processes can be quantified and therefore serves as a link to study questions of structural and dynamical robustness in a unified way.

Keywords: network topology, network robustness, lethality

1 Introduction

Recent experimental efforts have highlighted the pervasiveness of molecular networks in biological sciences. They have triggered an increased interest in the relation between the structure and the function of biological networks which control the information flow and regulation of cellular signals [1].

Already at the present level of experimental resolution it is apparent that protein interaction networks and regulatory networks have topological properties which cannot be explained by random graph models. One is led to the assumption that certain evolutionary mechanisms have shaped the network structures which can be observed today. A realistic model of biological networks should invoke a selection principle as the driving force of natural evolution. However, since selection acts on the functional properties of an organism one has to better understand which functional aspects are correlated with topological features.

One particularly important functional characterization of organismal fitness is the resilience against random changes in the environment [10, 15], which, at the molecular level, amount to changes in the network parameters. In a recent work [4] we introduced the concept of network entropy to quantify this resilience against perturbations and showed that, at the structural level, networks with higher entropy disintegrate less rapidly under random node removal. This notion has been converted into a simple evolutionary principle which states that organisms are selected preferentially, according to the entropy of their underlying network.

Such a system-level description incorporates and goes beyond other evolutionary network models, which tend to emphasize local connectivity properties and attachment rules based on node degree. Other global approaches have been formulated in terms of wiring costs [3]. Various groups have

argued that simple copy mechanisms can, for certain parameter ranges, give rise to network topologies observed in nature [2, 9, 13, 14]. Our formalism does not aim to describe one or the other topological feature, but rather considers all of them as correlates of an underlying functional characterization according to which networks can be ranked with respect to their resilience against change. This approach naturally extends to situations where the directionality and weights of edges are known.

How can this global network characterization help to make predictions about individual proteins in the context of their interaction network? In this work we define a natural ranking of proteins according to their relative contribution to network robustness. We demonstrate that this topological ranking is strongly correlated with the functionally defined lethality of a protein.

2 Methods and Results

As we showed earlier [4], the global quantity entropy can be decomposed into contributions from individual nodes

$$H = \sum_i \pi_i H_i \ , \tag{1}$$

where the sum is over all nodes and π_i are the components of the stationary distribution for a given process [1] and $H_i = \sum_j p_{ij} \log p_{ij}$ the Shannon entropy at node i. For the lack of better knowledge we define p_{ij} to be the process that maximizes H (for unweighted and undirected networks) [4].

This interpretation gives rise to the simple hypothesis that network elements with a higher contribution to the overall entropy should have a bigger effect on the networks functionality when being removed. If only the network topology is known the entropic ranking reflects the impact of node removal on network integrity, i.e. removal of nodes with high entropic contribution causes the network to disintegrate more rapidly.

In terms of actual perturbation experiments, we postulate that proteins with higher entropic contribution to the cellular network should more frequently have an associated phenotype when impaired (knock-out/knock-down). Previously this question has been phrased in various invocations of network centrality: degree [7], path length [16], and more recently betweeness [6]. Here we provide a rationale for why these a-hoc measures are all convenient proxies for network resilience.

We study both a single-cellular organism (*Saccharomyces cerevisiae*) and the multi-cellular worm (*Caenorhabditis elegans*) for which both binary interaction data (yeast-2-hybrid) as well as functional profiles are now available in large scale. Specifically, we use the binary interaction data from several yeast-two hybrid screens from MIPS [12] and [11], for yeast and *C.elegans*, respectively. The phenotype data is using gene deletion experiments and RNAi methodology for *C.elegans* [8] and gene knock-out [5] for yeast. For the latter we use the gene disruption data from MIPS (Oct. 2004). We also binned all different phenotypes from [8] into a single classification (phenotype/no phenotype). Notice that in *C.elegans* a large fraction of gene disruptions with phenotypic effect are actually lethal and our subsequent analysis does not change significantly if we restrict ourselves to the classification (lethal/viable).

For both organisms the protein interaction data together with our entropic analysis provides a topological ranking of proteins, which we systematically compare to the simple functional characterization. In the case of yeast we recorded 3741 proteins with an interaction and a knock out phenotype (681 lethal). For *C.elegans* the intersect is 2023 proteins, out of which 399 show a phenotypic effect when disrupted.

In Figure 1 we show that the number of top ranking proteins contains an unusually large number of those whose perturbations result in a phenotype. This can be quantified by the p-value for observing a given overlap or bigger by chance using Fisher's exact test. It is apparent that the topological ranking based on entropic contribution, $\pi_i H_i$, is strongly correlated with the observed phenotypic effects.

[1] The stationary distribution π of a process P is defined by $\pi P = \pi$.

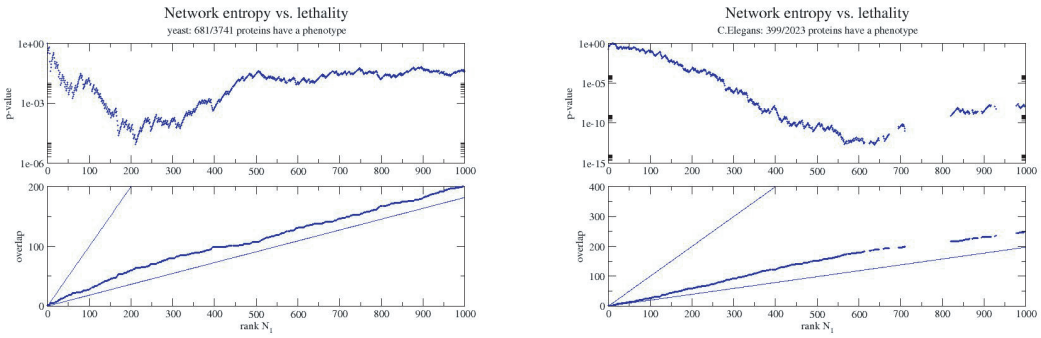

Figure 1: We illustrate the predictive power of network entropy for lethal genes in yeast and *C. elegans*. The bottom row gives the overlap k-top ranking proteins with known lethal proteins and the top row converts this overlap into a p-value, the probability to observe such an overlap by chance.

3 Discussion

In this section we will discuss the above result in the light of possible errors. First and foremost, the networks used in our work is based are known to suffer large error rates, both in terms of real interactions which were not observed (false negatives) and in terms of wrongly annotated interactions (false positives). To test whether such errors could impede our results, we arbitrarily increased the number of interactions by adding 50% more edges. Similarly we mimicked false positive errors by removing randomly 50% of all observed interactions.

In Figure 2 it is shown that the predictive power of the entropic ranking is hardly affected by such drastic changes. On the other hand, a fully randomized network does not show any correlations, as expected. This shows that the observed enrichment is very robust against rather drastic alterations to the data.

Another source of errors concerns the annotation of a gene disruption to a phenotype, which is usually done for one given condition and the observed correlation is strictly with respect to this single condition. On the other hand, the static network representation of yeast-two hybrid data only reports possible interactions and discards all dynamical dependencies. One should expect to observe different phenotypes for different conditions and our implicit assumption is that the classification into lethal/viable proteins is robust and also representative for many other conditions. We cannot exclude the possibility that additional proteins will prove lethal under different (untested) conditions. Such questions can be addressed only after a dynamical picture of the proteome is available.

The observed correlation between lethality and topological ranking could be a secondary effect caused by other correlation and/or experimental bias. For example, it has been suggested that the experimental yeast-two-bybrid data could be biased towards interactions within certain cellular compartments. If lethality had a similar bias then the observed correlation would be merely a reflection of this bias (similar for abundant proteins). We find that nuclear proteins are indeed more likely to be lethal and also tend to contribute more to the overall network entropy. However, this alone cannot explain all the observed correlations as we have check by taking the compartmental heterogeneity into account. For each number, N_1, of top-ranking proteins, we have drawn the same number of random proteins, but with the same distribution over different compartmental categories. Repeating this process, we have estimated the p-values by counting and still find a significant values $p \leq 10^{-3}$ for $100 < N_1 < 300$ (data not shown).

Finally, we want to stress that in our theoretical analysis we made a gross simplification which assumes bidirected and unweighted links between proteins, while cellular information flows almost always have a clear direction and different rates. This is not a limitation of our method, but rather

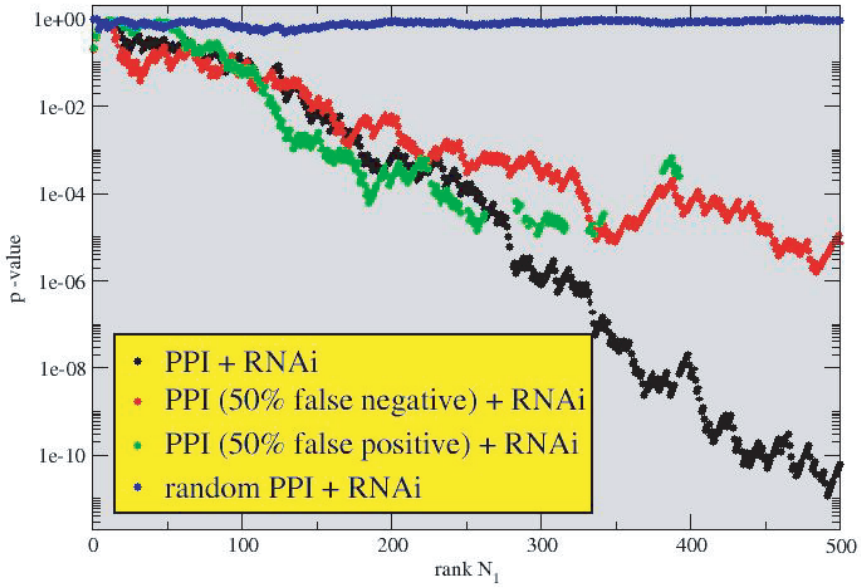

Figure 2: The correlation between the entropic ranking and lethality is not significantly altered by drastic changes to the network topology.

the only feasible way to interpret current large scale data. These errors and assumptions can only reduce the predictive power of the entropic assignments. The fact that there is still a significant enrichment for two very distinct organisms can be taken as a sign that network topology alone is a major determinant of behavioral properties.

Previous works have revealed correlations of several different centrality measures with lethality and with each other [6, 7, 16]. Our notion of network entropy is inherently a dynamical concept, which encodes the resilience of the network against changes in its parameters. If reduced to topological questions, network entropy also entails correlations with structural network observables such a degree and inbetweeness. It can therefore serve as a motivation for these simple proxies and account for the observed correlations. We argue that network entropy is the appropriate generalization of these structural concepts and is better suited to describe networks where dynamical processes can be quantified, for example through transition rates.

Acknowledgments

T.M. acknowledges funding by European Community Contract No. QLRI-CT-2001-00015 for "TEMBLOR" under the specific RTD programme "Quality of Life and Management of Living Resources".

References

[1] Alm, E. and Arkin, A.P., Biological networks, *Curr. Opin. Struct. Biol.*, 13(2):193–202, 2003.

[2] Chung, F., Lu, L., Dewey, T.G., and Galas, D.J., Duplication models for biological networks, *J. Comput. Biol.*, 10(5):677–687, 2003.

[3] Colizza, V., Banavar, J.R., Maritan, A., and Rinaldo, A., Network structures from selection principles, *Phys. Rev. Lett.*, 92(19):198701-1–198701-4, 2004.

[4] Demetrius, L. and Manke, T., Robustness and network evolution – an entropic principle, *Physica A*, 346(3-4):682–696, 2004.

[5] Giaever, G. *et al.* Functional profiling of the Saccharomyces cerevisiae genome, *Nature*, 418(6896):387–391, 2002.

[6] Hahn, M.W. and Kern, A.D., Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks, *Mol. Biol. Evol.*, 22(4):803–806, 2005.

[7] Jeong, H., Mason, S.P., Barabási, A.L., and Oltvai, Z.N., Lethality and centrality in protein networks, *Nature*, 411(6833):41–42, 2001.

[8] Kamath, R.S., Fraser, A.G., Dong, Y., Poulin, G., Durbin, R., Gotta, M., Kanapin, A., Le Bot, N., Moreno, S., Sohrmann, M., Welchman, D.P., Zipperlen, P., and Ahringer, J., Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi, *Nature*, 421(6920):231–237, 2003.

[9] Kim, J., Krapivsky, P.L., Kahng, B., and Redner, S., Infinite-order percolation and giant fluctuations in a protein interaction network, *Phys. Rev. E Stat. Nonlin. Soft. Matter Phys.*, 66(5 Pt 2):055101-1–055101-4, 2002.

[10] Kitano, H., Biological robustness, *Nat. Rev. Genet.*, 5(11):826–837, 2004.

[11] Li, S. *et al.* A map of the interactome network of the metazoan *C. elegans*, *Science*, 303(5657):540–543, 2004.

[12] Mewes, H.W., Frishman, D., Güldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Münsterkötter, M., Rudd, S., and Weil, B., MIPS: A database for genomes and protein sequences, *Nucleic Acids Res*, 30(1):31–34, 2002.

[13] Pastor-Satorras, R., Smith, E., and Solé, R.V., Evolving protein interaction networks through gene duplication, *J. Theor. Biol.*, 222(2):199–210, 2003.

[14] Rzhetsky, A. and Gomez, S.M., Birth of scale-free molecular networks and the number of distinct DNA and protein domains per genome, *Bioinformatics*, 17(10):988–996, 2001.

[15] Stelling, J., Gilles, E.D., and Doyle, F.J. 3rd, Robustness properties of circadian clock architectures, *Proc. Natl. Acad. Sci. USA*, 101(36):13210–13215, 2004.

[16] Yu, H., Greenbaum, D., Xin Lu, H., Zhu, X., and Gerstein, M., Genomic analysis of essentiality within protein networks, *Trends Genet.*, 20(6):227–231, 2004.