JOURNAL OF THE ROYAL SOCIETY Interface

An entropic characterization of protein interaction networks and cellular robustness

Thomas Manke, Lloyd Demetrius and Martin Vingron

J. R. Soc. Interface 2006 **3**, 843-850
doi: 10.1098/rsif.2006.0140

JOURNAL OF THE ROYAL SOCIETY Interface

J. R. Soc. Interface (2006) **3**, 843–850
doi:10.1098/rsif.2006.0140
Published online 11 July 2006

An entropic characterization of protein interaction networks and cellular robustness

Thomas Manke[1,*], Lloyd Demetrius[1,2] and Martin Vingron[1]

[1]*Max Planck Institute for Molecular Genetics, Ihnestr. 73, 14195 Berlin, Germany*
[2]*Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA*

The structure of molecular networks is believed to determine important aspects of their cellular function, such as the organismal resilience against random perturbations. Ultimately, however, cellular behaviour is determined by the dynamical processes, which are constrained by network topology. The present work is based on a fundamental relation from dynamical systems theory, which states that the macroscopic resilience of a steady state is correlated with the uncertainty in the underlying microscopic processes, a property that can be measured by entropy. Here, we use recent network data from large-scale protein interaction screens to characterize the diversity of possible pathways in terms of network entropy. This measure has its origin in statistical mechanics and amounts to a global characterization of both structural and dynamical resilience in terms of microscopic elements. We demonstrate how this approach can be used to rank network elements according to their contribution to network entropy and also investigate how this suggested ranking reflects on the functional data provided by gene knockouts and RNAi experiments in yeast and *Caenorhabditis elegans*. Our analysis shows that knockouts of proteins with large contribution to network entropy are preferentially lethal. This observation is robust with respect to several possible errors and biases in the experimental data. It underscores the significance of entropy as a fundamental invariant of the dynamical system, and as a measure of structural and dynamical properties of networks. Our analytical approach goes beyond the phenomenological studies of cellular robustness based on local network observables, such as connectivity. One of its principal achievements is to provide a rationale to study proxies of cellular resilience and rank proteins according to their importance within the global network context.

Keywords: network entropy; protein interactions; cellular robustness

1. INTRODUCTION

Recent experimental efforts have highlighted the pervasiveness of molecular networks in the biological sciences (Alm & Arkin 2003; Proulx *et al.* 2005. While a large number of molecular interactions and associations have been mapped qualitatively, we are yet to understand the relation between the structure and the function of biological networks that control the information flow and regulation of cellular signals.

One particularly important functional characterization is the resilience of an organism against external and internal changes (Kitano 2004; Stelling *et al.* 2004), which, at the molecular level, amounts to perturbations in the network parameters. In recent experiments, this resilience has been studied in direct response to gene deletions or RNA interference (Giaever 2002; Kamath 2003). It has been demonstrated that a large number of such network perturbations does not result in any phenotypic variation under a given experimental condition. In other words, different networks show the same apparent phenotype. This observation has led to a simple classification of proteins into 'viable' and 'lethal', according to whether the organism survives the removal of this component or not. In the following, we also refer to the latter as 'essential' proteins.

If network topology characterizes behavioural complexity, one may ask if there is any topological correlate for lethality. The seminal works of Barabasi and colleagues (Barabasi & Albert 1999; Albert *et al.* 2000; Jeong *et al.* 2001) have revived and spawned various efforts (Rapoport 1963; de Solla Price 1965) to characterize the *structural* properties of networks and relate their topological features to experimentally observed resilience. These phenomenological descriptions have highlighted certain commonalities in network structures and provided considerable insight into

*Author for correspondence (manke@molgen.mpg.de).

The electronic supplementary material is available at http://dx.doi.org/10.1098/rsif.2006.0140 or via http://www.journals.royalsoc.ac.uk.

Received 2 May 2006
Accepted 2 June 2006

the possible mechanisms of network evolution. However, the central observables, such as degree, invoked in these structural models, do not derive from any systematic theory, and the basis for their applicability to the characterization of functional resilience has been difficult to elucidate.

Here, we present a systematic approach to this issue based on methods from statistical mechanics and ergodic theory. This provides a natural conceptual framework to derive macroscopic parameters that characterize certain structural and functional properties of the network. The key idea and underlying assumption of our work is that biological processes typically operate at steady state, where characteristic macroscopic observables (the 'phenotype') remain constant for relatively long times. This, however, does not imply that the underlying microscopic variables (such as protein activities and concentrations) are static, but rather that their complex and continuous interplay results in a stable phenotype that can be experimentally observed. Indeed, it is the diversity and uncertainty of the microscopic processes that determine the resilience of macroscopic steady states against perturbations. In the context of the ergodic theory of dynamical systems, this uncertainty is quantified by the dynamical entropy (Kolmogorov–Sinai invariant). The significance of this concept for studies of biological systems resides on a fluctuation theorem for networks, an analogue of the fluctuation–dissipation theorem in statistical mechanics (Demetrius *et al.* 2004). According to this theorem, changes in *network entropy* are positively correlated with changes in the resilience of the macroscopic system against microscopic perturbations. As a great simplification and in recognition of our ignorance about the actual molecular events, we assume that the microscopic processes on the network are Markovian. This leads to characterization of network entropy as a measure of the diversity of molecular interactions that define the system. In recent work (Demetrius & Manke 2004), we applied the fluctuation theorem to a class of biological networks and demonstrated that, at the structural level, networks with higher entropy disintegrate less rapidly under random node removal. Such topological resilience is commonly characterized in terms of an increase in the average shortest path length or the decrease in the fractional size of the largest connected network component when a fraction of nodes is deleted (Albert *et al.* 2000).

Our formalism does not aim to specifically describe one or the other topological features, such as degree or the shortest path lengths, but rather considers them as correlates of an underlying functional property, according to which networks can be ranked with respect to their resilience against random perturbations. This approach naturally extends to situations where the network is described in terms of structure and dynamics and where the directionality and weights of edges are known. How can this global entropic characterization of the network help to make predictions about individual proteins in the context of their interaction network?

To answer this question, we first recall the definition of the dynamical entropy for a Markov process, $P=(p_{ij})$, which is given by (Billingsley 1965):

$$H=-\sum_{ij}\pi_i p_{ij}\log p_{ij}. \tag{1.1}$$

Here, p_{ij} denotes the transition probabilities and π_i are the components of the stationary distribution (see §2 for more details). It should be noted that there are many other ways to investigate complex dynamical systems through microscopic modelling, such as differential equations, Boolean dynamics or cellular automata, to name a few. Our simple stochastic description of dynamical uncertainty is based on random walks on the network and has a long tradition in the analysis of diffusive systems (Berg 1993). This approach serves the larger goal of deriving macroscopic properties (such as diffusion laws and thermodynamic relations) from the microscopic dynamics of a test particle. In our context, the dynamical entropy of a Markov process characterizes the diversity of possible pathways and (through the fluctuation theorem) is related to the system's response to perturbations.

In order to rank individual network elements, we use the decomposition of network entropy into contributions from all individual proteins

$$H=\sum_i \pi_i H_i, \tag{1.2}$$

where H_i is the Shannon entropy associated with protein i. This decomposition suggests that network elements with a higher contribution to the overall entropy have a larger effect on the network's resilience and functionality, when removed. If only the network topology is known, the entropic-ranking reflects the impact of node removal on network integrity, i.e. removal of proteins with higher entropic contribution causes a larger change in topological and functional integrities. In terms of functional perturbation experiments, we will test the hypothesis that proteins with higher entropic contribution to the cellular network more frequently have a lethal phenotype when they are impaired (knockout/knockdown). Previously, this question has been addressed in terms of various notions of network centrality, such as degree (Jeong *et al.* 2001), shortest path length (Yu *et al.* 2004) and more recently betweenness (Hahn & Kern 2005). While these concepts provide useful insights into principles of network organization, they do not derive from any general theory, hence, the range of their applicability can only be tested empirically. On the contrary, our entropic formalism is embedded in a theoretical framework, which analytically characterizes robustness. It also provides a rationale why these *ad hoc* measures are sometimes convenient proxies for network resilience and how they could be extended.

2. METHODS

(*a*) *Definition of network entropy*

The concept of network entropy was introduced by Demetrius *et al.* (2004) and applied to network modelling by Demetrius & Manke (2004). To be self-contained, we review the basic ingredients of this formulation. Every (weighted and directed) network

can be specified by its adjacency matrix $A=(a_{ij})$. The Perron–Frobenius eigenvalue $\lambda(A)$ is the largest eigenvalue of A and the corresponding eigenvector v is strictly positive for each strongly connected graph component.[1]

The dominant eigenvalue is a topological invariant of the adjacency matrix and is known to satisfy a variational principle, which is formally analogous to the minimization of the free energy in statistical mechanics (Arnold *et al.* 1994). In order to characterize this principle, we introduce the notion of a stochastic process $P=(p_{ij})$ which is said to be compatible with the adjacency matrix A, if $\sum_j p_{ij}=1$ and $p_{ij}=0 \Leftrightarrow a_{ij}=0$. This requirement amounts to the idea that microscopic variables, e.g. protein activities or concentrations, can only change in response to changes of their interaction partners. In other words, information can only flow along the interacting proteins. Here, we ignore the possibility that certain protein modifications may trigger global changes, for example, through changes in cellular parameters, such as pressure or salinity.

The stationary distribution, π, is defined as the left-hand eigenvector associated with the largest eigenvalue 1 of the stochastic matrix P:

$$\pi P = \pi. \tag{2.1}$$

The stationary distribution, π, characterizes the long-time invariant behaviour of the Markov process described by matrix P. If we assume ergodicity, i.e. irreducibility of A and P, then the components π_i satisfy $\pi_i>0$ and denote the relative frequency with which the random walk on the network visits node i.

Now, the variational principle for λ can be written as

$$\log\lambda = \sup_P \left[-\sum_{i,j}\pi_i p_{ij}\log p_{ij} + \sum_{i,j}\pi_i p_{ij}\log a_{ij}\right], \tag{2.2}$$

with respect to all compatible processes $P=(p_{ij})$. It has been shown (Arnold *et al.* 1994) that, for strongly connected networks, the supremum is attained for a unique matrix $\hat{P}=(\hat{p}_{ij})$, where

$$\hat{p}_{ij} = \frac{a_{ij}v_j}{\lambda v_i}. \tag{2.3}$$

With this choice of the matrix P, equation (2.2) becomes

$$\log\lambda = -\sum_{i,j}\pi_i\hat{p}_{ij}\log\hat{p}_{ij} + \sum_{i,j}\pi_i\hat{p}_{ij}\log a_{ij}, \tag{2.4}$$

which can be verified by explicit computation. Equation (2.4) can be written in the form

$$\log\lambda = H(\hat{P}) + \Phi(\hat{P}). \tag{2.5}$$

This links the topological invariant λ to the network entropy $H(\hat{P})$ and the 'potential' $\Phi(\hat{P})=\sum_{i,j}\pi_i\hat{p}_{ij}\log a_{ij}$. The relation given in equation (2.2) is an analogue of the Gibbs variational principle in statistical mechanics (see Arnold *et al.* 1994). It should be noted that, for Boolean matrices, the second term in equation (2.4) vanishes ($\Phi=0$) and equation (2.2) corresponds to a maximum entropy principle for the most parsimonious choice of p_{ij}. Accordingly, we can now write

$$H(\hat{P}) = \log\lambda = -\sum_{i,j}\pi_i\hat{p}_{ij}\log\hat{p}_{ij} = \sum_i \pi_i H_i, \tag{2.6}$$

where H_i is the standard Shannon entropy defined for each node i and π_i are the components of the stationary distribution, as defined by equation (2.1).

(*b*) *Fluctuation theorem*

The key motivation for our work is a set of theorems from statistical physics and dynamical systems theory, which relate observables at steady state to the relaxation properties of a perturbed system (measured away from steady state). To give an equilibrium example, it is well known that fluctuations at equilibrium determine the return rate to the equilibrium state (Kubo 1966). There are many extensions of this fundamental relation. In a recent work (Demetrius *et al.* 2004), a fluctuation theorem was derived which invokes the entropy as a measure of microscopic variability and relates it to the macroscopic resilience of the system. This result can be formally described as follows. Consider a perturbation in some microscopic variable. Such changes will generally result in deviations of a steady-state observable from its unperturbed value. Let $P_\epsilon(t)$ denote the probability that the sample mean deviates by more than ε from its unperturbed value at time t. As t increases, $P_\epsilon(t)$ converges to zero and we define the fluctuation decay rate, R, as

$$R = \lim_{t\to\infty}\left[-\frac{1}{t}\log P_\epsilon(t)\right]. \tag{2.7}$$

Large values of R entail small deviations of observables from the steady-state condition and small values of R correspond to large fluctuations around its mean value. Thus, R characterizes the insensitivity of a macroscopic observable in the face of changes in the underlying parameters.

The fluctuation theorem (Demetrius *et al.* 2004) asserts that changes in R are positively correlated with changes in network entropy:

$$\Delta H \Delta R > 0. \tag{2.8}$$

Here, entropy, H, is defined at steady state, while R determines the behaviour away from the steady state. The fluctuation theorem implies that an increase in entropy entails a greater insensitivity of an observable to dynamic or structural perturbations of the network.

(*c*) *Protein–protein interactions and lethality data*

We study both a single-cellular organism (budding yeast) and a multicellular worm (*Caenorhabditis elegans*), for both of which binary interaction data (yeast-two hybrid) as well as functional profiles are now available in large scale.

[1]In a strongly connected (irreducible) graph component every node can be reached from every other node.

For *Saccharomyces cerevisiae,* we retrieved a bidirected interaction network of 3854 proteins with 11 912 yeast-two hybrid interactions[2] from the Munich Information Center for Protein Sequences (MIPS) database (Mewes 2002). From the same source, we also downloaded lethality data from gene disruption experiments (Giaever 2002) (6203 open reading frames (ORFs) = 5033 viables + 1170 lethals). The intersect of these two datasets contained 3741 proteins, for which we had both interaction data and a phenotype annotation. Of these 3741 proteins, 681 are annotated as having a lethal phenotype when removed. For 3534 ORFs, we also know one of the eight compartmental localizations, as defined by Huh *et al.* (2003).

For *C. elegans*, we used the yeast-two hybrid data from the work of Li *et al.* (2004), which we retrieved from wormbase (Chen 2005). This constituted of 2800 proteins and 8740 interactions. It should be noted that the interaction data emerges from the same principle technique for both yeast and worm. However, the phenotype data for worm is determined differently, using the RNAi technology, as employed by Kamath (2003). We use the data from their supplementary materials and focus on the distinction between 17 491 viable proteins and 1170 lethals. After intersection with the protein interaction data, we have 2023 proteins out of which 322 have a lethal phenotype when their production is impaired through RNA interference.

(*d*) *Statistical analysis*

Each topological observable gives rise to a ranking scheme of proteins. Unlike previous works, which have studied the correlation of protein essentiality with various measures of protein 'centrality', we rank proteins according to their contribution to overall network entropy. There are several different ways to assess whether a given ranking scheme of proteins carries any information about their functional classification. In earlier work (Jeong *et al.* 2001), this question was addressed by simply comparing the fraction of essential proteins in arbitrary bins of highly and lowly connected proteins. More systematically, the same question can be addressed using the Fisher's exact test for any number of top-ranking proteins. This is further explained and utilized in the electronic supplementary material. In the main text, we follow yet another approach and test the null hypothesis that the entropy values for essential and non-essential proteins are drawn from the same underlying distribution against the alternative hypothesis that essential proteins are derived from a distribution function not smaller than that of non-essential proteins (Kolmogorov–Smirnov tests).

(*e*) *Network resampling*

For some of our analyses, we require an ensemble of random networks where all proteins retain their precise node degree. To this end, we apply an edge-swapping algorithm introduced by Maslov & Sneppen (2002). The basic idea is to start from a given network and switch two randomly chosen links if they do not introduce self-loops or multiple edges. To avoid autocorrelations from measurements on similar networks, 50 000 such edge swaps are taken between subsequent measurements. In this way, we generate 1000 random networks from which the distribution of relevant observables can be determined.

[2]While much of the experimental data is generated in an asymmetric fashion, we always interpret interactions as bidirected.

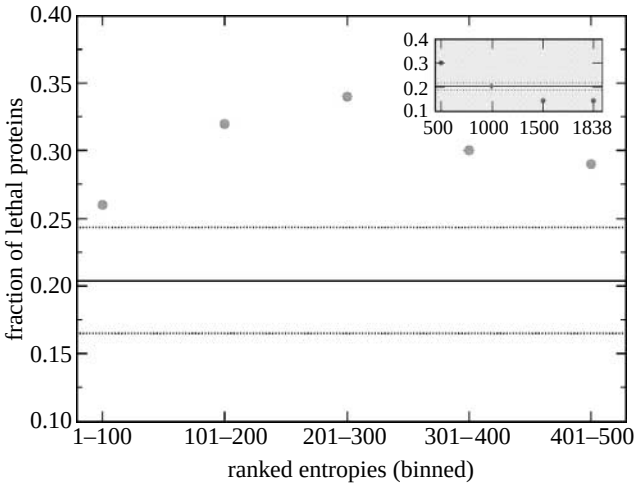

Figure 1. In the main figure, we define five classes of proteins (100 proteins in each) according to their rank with respect to entropic contribution in the interaction network of *C. elegans*: 1–100, 101–200, …. In all these high-ranking cases, the fraction of essential proteins is significant. The expectation from 100 random proteins is shown as horizontal lines (± 1 s.d.). The inset shows the same analysis for taking larger bins (ranks 1–500, 501–1000, 1001–1500, 1500–1838). Again, we can see an enrichment for high-ranking proteins, while there is an under-representation of essential proteins for proteins with small entropic contributions (for ranks greater than 1000). In table 1, we give a detailed statistical account of this simple observation.

3. RESULTS

(*a*) *Lethality correlates with entropic contribution*

The basic question, which we address in this paper, is whether the entropic characterization can predict the phenotypic outcome of a gene disruption based on the overall position of the protein in the interaction network. Given the qualitative character and the large error rates of present experimental techniques, the actual goal is more modest, as we can at most expect to see an enrichment of essential proteins in ranked lists of proteins, which we deem important based on their contribution to network entropy.

As a first simple step, we bin proteins into several classes of high and low entropic contributions and determine the fraction of essential proteins among them. Figure 1 shows that proteins with high contribution to overall entropy are preferentially essential compared to random expectations, while proteins with small contributions are less frequently essential than expected. This simple observation can

Table 1. Summary of our statistical tests for several background models and network observables (entropic contribution, $\pi_i H_i$; degree, k_i and inbetweeness, I_i). (Small p-values signify that the distribution of the observed values is significantly larger for essential than for non-essential proteins (Kolmogorov–Smirnov test). For each background model, we denote N_e as the number of essential proteins and N_n as the number of non-essential proteins in the background. The basic tests for yeast and *C. elegans* (rows 1 and 4) should be compared with the more sophisticated background models, in which (*a*) the network links have been randomly rewired while maintaining the individual degree of a protein (rows 2 and 5), (*b*) the compartmental distribution of the background has been chosen to match that of essential proteins in yeast (row 3) and (*c*) where 50% of links have randomly been added/removed to model false negative/positive errors (rows 6 and 7). As a sanity check, we also performed these tests on a network with randomized node labels, for which we do not expect any significant deviations (last row).)

	(N_e, N_n)	$p(\pi_i H_i)$	$p(k_i)$	$p(I_i)$
yeast	(715, 2807)	4.2×10^{-5}	1.6×10^{-3}	1.6×10^{-3}
yeast, randomized links	(715, 2807)	4.8×10^{-4}	1.6×10^{-3}	6.5×10^{-3}
yeast, compartmental bias	(583, 560)	2.2×10^{-3}	6.2×10^{-3}	1.5×10^{-2}
C. elegans	(372, 1466)	9.1×10^{-11}	4.9×10^{-9}	3.9×10^{-9}
C. elegans, randomized links	(372, 1466)	1.8×10^{-9}	4.9×10^{-9}	4.3×10^{-6}
C. elegans, +50% random links	(396, 1611)	1.3×10^{-5}	8.4×10^{-6}	6.3×10^{-4}
C. elegans, −50% random links	(236, 826)	4.6×10^{-4}	4.6×10^{-4}	6.7×10^{-4}
C. elegans, randomized labels	(368, 1480)	0.956	0.977	0.796

be made more quantitative in several statistical tests as shown in table 1 and in the electronic supplementary material. This demonstrates that the network property 'entropic contribution' contains a significant amount of information about the functional property, 'essentiality'. It can also be seen that the entropic contribution, $\pi_i H_i$, accounts for more essential proteins than traditional observables, such as the degree, k_i or the in-betweeness, I_i. This improvement is due to an improved ranking of lowly connected proteins as is further discussed in §3*d*.

We want to stress that this observation is not equivalent to strong positive predictive power, which is generally low (less than 40%). This is not surprising, given the insufficiencies in the current interaction data and functional screens, as well as the simplicity of our model. For a more detailed received operating characteristic (ROC)-curve analysis, we refer the reader to the electronic supplementary material.

(*b*) *Effect of errors in protein interaction data*

The observed correlation between entropic ranking and lethality assignments should be seen in the light of possible errors. The large-scale interaction data we have used are subject to sizeable error rates, both in terms of missed interactions (false negatives) and predicted interactions, which do not occur in physiological conditions (false positives). To estimate whether the observed enrichment of essential proteins with high entropic values is robust against such errors, we randomly added (deleted) 50% of all interactions and re-analysed the modified networks. This gave rise to a new ranking of proteins, which we compared to the fixed assignment of essential/non-essential proteins. As shown in table 1, such drastic changes resulted in larger p-values, but still significantly better than random assignment. We conclude that the observed enrichment is robust against the rather large error rates commonly associated with yeast-two hybrid data.

(*c*) *Possible biases in experimental data*

Several biases could affect the lethality analysis presented above. Notably, the observed correlation could be simply a secondary effect caused by other correlations between experimental interaction data and functional data for essential proteins. For example, in yeast, we find that there is a strong enrichment of essential proteins with nuclear location (see table S1 in the electronic supplementary material). If the network data obtained from yeast-two hybrid screens are equally biased with respect to compartmental location, the observed enrichment may simply be a secondary effect. We explicitly account for this possibility by choosing a more sophisticated background model, in which the non-essential proteins are selected randomly, while still respecting the compartmental distribution of essential proteins (foreground). As expected, we see that this results in an increase of the p-values compared to the simple test (table 1, first and third row). However, essential proteins still show significantly higher values of $\pi_i H_i$, which cannot be accounted for by compartmental bias.

This analysis is further elaborated upon in the electronic supplementary material, where we also investigated a possible correlation of $\pi_i H_i$ with protein abundance. Our analysis for abundance data in yeast (Huh *et al.* 2003) shows that the entropic contribution of a protein is not correlated with its abundance (see figure 7 in the electronic supplementary material). Therefore, we conclude that our enrichment analysis is not significantly biased toward more or less abundant proteins.

(*d*) *The role of protein connectivity*

Many structural network observables have been suggested in the past and investigated with respect to their functional implications. While none of these measures provides a powerful tool to *predict* functional properties, some have helped to highlight certain common structural features and possible evolutionary

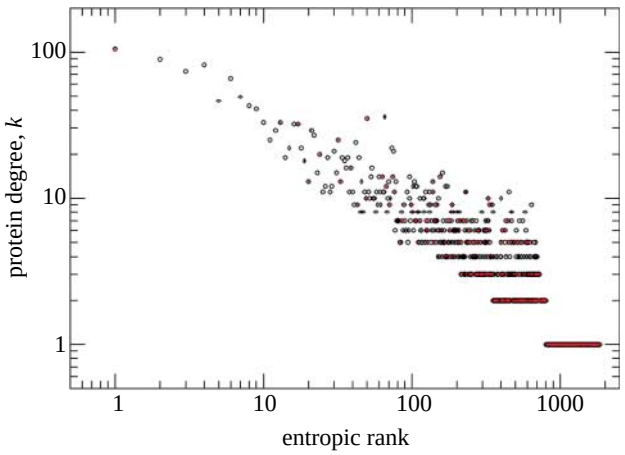

Figure 2. Here, we show (for *C. elegans*) that connectivity and network entropy are correlated, but distinct from one another. For the process defined in equation (2.3), proteins with high degree tend to have high entropic contribution. On the other hand, there are also lowly connected proteins with high contribution to network entropy and hence dynamical stability. A small red dot was added to highlight essential proteins.

mechanisms ('design principles'). Since the degree of a protein has arguably received the highest attention, we wish to use this section to highlight similarities and differences of our approach with degree-based methods. In the electronic supplementary material, we extend this discussion to other structural observables, such as in-betweeness, also (Freeman 1977).

The dynamical properties of the effective process defined in equation (2.3) entail that nodes with high in-degree tend to have larger values of π_i, and nodes with high out-degree tend to have large H_i. For undirected networks (such as protein interaction networks), out- and in-degrees coincide and result in a correlation of large degrees with what we call large entropic contribution, $\pi_i H_i$. This expectation is confirmed by an explicit calculation of these quantities in the protein interaction network of *C. elegans*, figure 2. This figure shows that there is a strong correlation between the degree k and $\pi_i H_i$, especially when both are large. From this perspective, one can now better understand the apparent success and relevance of degree-based structural measures in studies of functional properties. We would like to advocate the view that the degree and other structural observables can be considered as correlates of underlying dynamical properties, such as the stability of a dynamical process to random perturbations.

Having outlined the similarities between the entropic characterization and the degree-based method, we would now like to address in more detail their differences. The local connectivity of a protein is an important, but not the only, determinant of its entropic rank. First, we wish to illustrate some of the differences through concrete examples. Gei-16 (T17H7.5) is an essential protein required for embryonic development and morphogenesis in *C. elegans*. Based on yeast-two-hybrid interaction data (Li 2004), it is top-ranked by both the entropic scheme and the degree-based method (105 interaction partners). At the other extreme, consider the heat-shock protein 4 (hsp-4; F43E2.8), which is essential, but only interacts (in the yeast-two hybrid screen) with two other proteins: a signalling protein (Y51H4A.17) and lin-5 (T09A5.10) which is required for mitosis and cytokinesis, and inhibition of which is also lethal according to Kamath (2003). Most notably, from a structural perspective, these two interaction partners are themselves highly connected (16 and 8 partners, respectively). On the contrary, the viable lipid transporter ckc-1 (T27A10.3) also has only two interaction partners, but their connectivities are equally low (degree 1 and degree 2). The entropic characterization can account for such differences in the neighbourhood structure by ranking hsp-4 much higher (rank 360) than ckc-1 (rank 780).

These examples only illustrate that our analysis yields a different ranking of proteins within their context of global network, especially for lowly connected nodes. We will now investigate systematically how these differences translate into different correlations with the lethality assignment. In particular, we ask whether essential proteins have higher entropic values not only with respect to non-essential proteins, but also compared to non-essential proteins with the same degree distribution. To this end, we repeat the above analysis for a subset of random network realizations, in which all proteins retain their precise degree (see table 1, rows 2 and 5). The significant increase in the p-value (rows $1 \rightarrow 2$, rows $4 \rightarrow 5$) illustrates that the entropic observable carries more information about essential proteins than degree, information which is lost during the randomization process.

Within the framework of multivariate statistics, this question may also be addressed by a partial correlation analysis. To this end, we consider degree and in-betweeness as possible confounding variables and account for their effect on the observed correlation between entropic contribution and essentiality. This results in only a small decrease of the correlation and confirms that entropic contribution carries additional information, when degree and in-betweeness are controlled. Moreover, we have performed a multiple logistic regression including all three variables (entropic contribution, degree and in-betweeness) and find that the entropic contribution is the most important factor, while the addition of degree or in-betweeness does not improve the logistic model. We refer the reader to the electronic supplementary material for further details and a complementary analysis of the same issue from the perspective of enriched essential proteins in top-ranking lists.

We conclude that our entropic observable accounts more effectively for differences in the network neighbourhood, which is especially important for weakly connected nodes. In terms of positive predictive power, the improvement is marginal, but the comparison serves us to illustrate differences between the chosen observables. None of the presented methods (including our own) has a strong predictive power over functional properties, such as 'lethality'. We do not believe that current interaction data already allow for a meaningful

competition between those simple models at the level of prediction accuracy, but rather at the level of motivated hypotheses and their ability to generalize to more quantitative data.

4. CONCLUSION AND DISCUSSION

In summary, we have shown that the entropic characterization of protein interaction networks can account for a significant fraction of proteins, whose removal results in a lethal phenotype.

In our framework, proteins are ranked according to their contribution to network entropy, which is a measure of microscopic uncertainty (pathway diversity) and is correlated with the macroscopic robustness of a dynamical system defined on the network. If only structural information is provided, the ranking of a protein depends on its overall position within the network. This notion is in sharp contrast to other characterizations of network elements based, for example, on their local connectivity (degree). Our analysis has shown that the entropic characterization is also able to account for functional differences of proteins with low or equal degree.

We employed several statistical methods and background models to assess the correlations between the entropic-ranking scheme and phenotypic-lethality data, and have carefully tested the observed correlations against a number of possible errors. It is important to re-iterate that, no method based on the current large-scale interaction data can provide strong predictive power for functional and context-dependent properties, such as lethality. As it is a common practice in this field, we take the observed correlation as an indicator of a possible biological signal, which cannot otherwise be explained and warrants further investigation. Here, we introduced a new conceptual framework, which provides a rationale to understand macroscopic resilience in the light of microscopic uncertainty, as characterized by entropy, rather than structural network observables. From this perspective, the observed enrichment of essential proteins in ranked lists of proteins has a natural and clear interpretation: proteins with higher contribution to cellular resilience are more often essential. Heuristic constructs, such as node degree, emerge as effective descriptors of functional properties, but our work also illustrates where one can go beyond such structural measures. Retrospectively, and as illustrated by the examples in the previous section, one might be tempted to account for the entropic contribution by introducing some 'effective degree'. We consider it an advantage that our approach presents a natural way to introduce and to extend structural concepts like degree. Moreover, and in contrast to degree-based methods, our approach is extendable to networks where more quantitative data are available.

In the following, we want to point to possible limitations of our approach. First, the phenotypic assessment of a gene disruption is usually done for one given condition and the observed correlation is strictly with respect to this single condition. It has been remarked that, the so-called viable proteins may actually play a significant role in untested environments and their disruption could cause lethal phenotypes. An exhaustive study of all possible conditions is clearly beyond experimental capabilities. Therefore, we take the present lethality data as representative for other conditions and implicitly assume that the classification of lethal and viable proteins is at least robust against environmental changes.

A related problem concerns the static representation of interaction data, which discards all dynamical dependencies. Just as many genes are expressed only under specific conditions, we also should think of different network realizations of an underlying blueprint, which experimental interaction screens try to establish.

Since the concept of entropy is based on the notion of dynamical diversity of the microscopic processes underlying the macroscopic cellular states, we believe that this approach will ultimately be more fruitful than network characterizations, which are solely based on topology. Cellular robustness depends on the dynamical properties and interconnections of many diverse molecular networks. We should, however, stress that, in the present application, we relied exclusively on structural information for only a part of the complete cellular network, namely protein–protein interactions. Furthermore, we characterized the microscopic diversity through a Markov process that maximizes the entropy based on a Boolean adjacency matrix, rather than quantitative information about transition rates. Needless to say, actual processes may be different from this representative one. To the extent that real processes resemble the one defined in this work, we can now better understand the importance of structural network observables as correlates of dynamical properties. We expect that structural properties will become less useful concepts for processes that deviate from the one with maximal entropy. Our approach is a first attempt to bridge these two domains and to address structural and dynamical questions in a single framework.

This situation can be likened to thermodynamics, where some properties of large systems can be effectively described by a number of macroscopic parameters, regardless of our ignorance about the microscopic processes. For equilibrium systems, this simplification is made explicit through relations between the Gibbs distribution over microstates and various macroscopic properties that can be derived from it (Gibbs 1901). Formally, our work builds on an extension of the Gibbs formalism, which also applies to non-equilibrium systems at steady state (Demetrius 1997; Ruelle 2004). This assumption is often made for many classes of biological systems (e.g. in metabolic flux analysis) and is the basis of empirical phenotype classification. It does not hold for systems at developmental branch points, where macroscopic observables can change dramatically in response to environmental and developmental signals. If these assumptions hold, our approach should also apply to other complex networks, and there is hope that some systemic properties can be elucidated without having to resort to microscopic details.

We thank K. Sneppen for constructive comments and useful suggestions. M.V. wishes to acknowledge support through the Max Planck Research Award. This work was supported by European Community Contract No. QLRI-CT-2001-00015 for 'TEMBLOR' under the specific RTD programme 'Quality of Life and Management of Living Resources' and by a grant from the German National Genome Research Network (NGFN).

REFERENCES

Albert, R., Jeong, H. & Barabasi, A. L. 2000 Error and attack tolerance of complex networks. *Nature* **406**, 378–382. (doi:10.1038/35019019)

Alm, E. & Arkin, A. P. 2003 Biological networks. *Curr. Opin. Struct. Biol.* **13**, 193–202. (doi:10.1016/S0959-440X(03)00031-9)

Arnold, L., Gundlach, V. & Demetrius, L. 1994 Evolutionary formalism for products of positive random matrices. *Ann. Probab.* **4**, 859–901.

Barabasi, A. L. & Albert, R. 1999 Emergence of scaling in random networks. *Science* **286**, 509–512. (doi:10.1126/science.286.5439.509)

Berg, H. C. 1993 *Random walks in biology*. Princeton, NJ: Princeton University Press.

Billingsley, P. 1965 *Ergodic theory and information*. New York, NY: Wiley.

Chen, N. *et al.* 2005 WormBase: a comprehensive data resource for Caenorhabditis biology and genomics. *Nucleic Acids Res.* **33**, D383–D389. (doi:10.1093/nar/gki066)

de Solla Price, D. J. 1965 Networks of scientific papers. *Science* **149**, 510–515.

Demetrius, L. 1997 Directionality principles in thermodynamics and evolution. *Proc. Natl Acad. Sci. USA* **94**, 3491–3498. (doi:10.1073/pnas.94.8.3491)

Demetrius, L., Gundlach, V. M. & Ochs, G. 2004 Complexity and demographic stability in population models. *Theor. Popul. Biol.* **65**, 211–225. (doi:10.1016/j.tpb.2003.12.002)

Demetrius, L. & Manke, T. 2004 Robustness and network evolution—an entropic principle. *Physica A* **346**, 682–696. (doi:10.1016/j.physa.2004.07.011)

Freeman, L. 1977 A set of measures of centrality based upon betweeness. *Sociometry* **40**, 35–41. (doi:10.2307/3033543)

Giaever, G. *et al.* 2002 Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387–391. (doi:10.1038/nature00935)

Gibbs, J. W. 1901 *Elementary principles in statistical mechanics*. New York, NY: Dover.

Hahn, M. W. & Kern, A. D. 2005 Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol. Biol. Evol.* **22**, 803–806. (doi:10.1093/molbev/msi072)

Huh, W.-K., Falvo, J. V., Gerke, L. C., Carroll, A. S., Howson, R. W., Weissman, J. S. & O'Shea, E. K. 2003 Global analysis of protein localization in budding yeast. *Nature* **425**, 686–691. (doi:10.1038/nature02026)

Jeong, H., Mason, S., Barabasi, A. & Oltvai, Z. 2001 Lethality and centrality in protein networks. *Nature* **411**, 41–42. (doi:10.1038/35075138)

Kamath, R. S. *et al.* 2003 Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* **421**, 231–237. (doi:10.1038/nature01278)

Kitano, H. 2004 Biological robustness. *Nat. Rev. Genet.* **5**, 826–837. (doi:10.1038/nrg1471)

Kubo, R. 1966 Fluctuation—dissipation theorem. *Rep. Prog. Phys.* **29**, 255–284. (doi:10.1088/0034-4885/29/1/306)

Li, S. *et al.* 2004 Armstrong map of the interactome network of the metazoan *C. elegans*. *Science* **303**, 540–543. (doi:10.1126/science.1091403)

Maslov, S. & Sneppen, K. 2002 Specificity and stability in topology of protein networks. *Science* **296**, 910–913. (doi:10.1126/science.1065103)

Mewes, H. *et al.* 2002 MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* **30**, 31–34. (doi:10.1093/nar/30.1.31)

Proulx, S., Promislow, D. & Philips, P. 2005 Network thinking in ecology and evolution. *Trends Ecol. Evol.* **20**, 345–353. (doi:10.1016/j.tree.2005.04.004)

Rapoport, A. 1963 *Mathematical models of social interaction*. New York, NY: Wiley.

Ruelle, D. 2004 *Thermodynamic formalism*. Cambridge, MA: Cambridge Mathematical Library.

Stelling, J., Sauer, U., Szallasi, Z., Doyle, F. J. & Doyle, J. 2004 Robustness of cellular functions. *Cell* **118**, 675–685. (doi:10.1016/j.cell.2004.09.008)

Yu, H., Greenbaum, D., Lu, H. X., Zhu, X. & Gerstein, M. 2004 Genomic analysis of essentiality within protein networks. *Trends Genet.* **20**, 227–231. (doi:10.1016/j.tig.2004.04.008)