A Letters Journal Exploring the Frontiers of Physics

OFFPRINT

The fragility of protein-protein interaction networks

C. M. Schneider, R. F. S. Andrade, T. Shinbrot and H. J. Herrmann

EPL, **95** (2011) 16001

A Letters Journal Exploring the Frontiers of Physics

July 2011

EPL, **95** (2011) 16001 www.epljournal.org
doi: 10.1209/0295-5075/95/16001

The fragility of protein-protein interaction networks

C. M. Schneider[1(a)], R. F. S. Andrade[2], T. Shinbrot[3] and H. J. Herrmann[1,4]

[1] *Computational Physics, IfB, ETH Zurich - Schafmattstrasse 6, 8093 Zurich, Switzerland*
[2] *Instituto de Física, Universidade Federal da Bahia - 40210-210, Salvador, Brazil*
[3] *Biomedical Engineering, Rutgers University - Piscataway, NJ, USA*
[4] *Departamento de Física, Universidade Federal do Ceará - 60451-970 Fortaleza, Ceará, Brazil*

received 15 February 2011; accepted in final form 13 May 2011
published online 15 June 2011

PACS 64.60.aq – Networks
PACS 64.60.ah – Percolation
PACS 89.75.Fb – Structures and organization in complex systems

Abstract – The capacity to resist perturbations from the environment is crucial to the survival of all organisms. We quantitatively analyze the susceptibility of protein interaction networks of numerous organisms to random and targeted failures. We find for all organisms studied that random rewiring improves protein network robustness, so that actual networks are more fragile than rewired surrogates. This unexpected fragility contrasts with the behavior of networks such as the Internet, whose robustness decreases with random rewiring. We trace this surprising effect to the modular structure of protein networks.

Introduction. – Over the past two decades, prodigiously detailed maps of protein interaction networks (PPIs) have been produced [1,2]. These networks in principle present a record of all metabolic processes and their inter-relations, but in practice the number of chemical actors and the complexity of their interactions make the networks difficult to decipher [3]. In this letter, we show that notwithstanding their apparent complexity, it is possible to establish common features of protein networks starting from a few simple principles [4–6].

We begin with the observation, illustrated in fig. 1, that biological protein networks involve both common processes that all cells must use (*e.g.*, enzymes involved in the Krebs cycle, marked with red labels) and what are termed modular processes [7–9] that appear only in special situations (*e.g.*, guidance molecules used only during particular circumstances, such as development, reproduction, or response to heat stress, indicated by blue labels). As we will show, this modular organization produces common, and predictable, network properties shared by all organisms studied. We focus in particular on the fragility of biological networks —a property of manifest importance for survival— to failures by interruption of individual protein function. To this end, we evaluate the extent to which PPI of 20 different organisms ranging from *Bacteria* and plants to *Homo sapiens* (table 1) can be disrupted by either random or targeted failures.

The modular construction shown in fig. 1 consists of a highly interconnected core of proteins, accompanied by satellite clusters with "hub" proteins weakly connected to the core. As a consequence, three predictions can readily be made. First, this type of network can be expected to be vulnerable to failures that interrupt the few hub proteins, but should be comparatively robust against errors that interrupt any of the more numerous proteins attached to "spokes" of these hubs [8]. Thus random errors are unlikely to significantly interrupt function, while malfunction of one or more hub proteins is likely to disrupt the network. Second, through countless generations of failures we expect evolution to have tuned biological networks to be more robust against failures than statistically comparable, but non-biological, networks. Third, through the same reasoning we expect biological networks to be optimal in that alternative interconnections should worsen robustness. As we will show, these predictions are largely correct, but admit unexpected and revealing failures.

To test these predictions, we compare known protein networks with surrogates that are as statistically similar as possible, having the same size and degree distribution as true biological networks, by performing a sequence of randomly chosen switches of connections between pairs of nodes $\{(i,j),(k,\ell)\} \to \{(i,\ell),(k,j)\}$ in a network, so that each node preserves its number of neighbors [10]. The

(a) E-mail: schnechr@ethz.ch

Table 1: List of organisms investigated. Acronyms in column 1 indicate kingdom and phylum the organisms belong to: AA - *Animalia Arthropoda*; AB - *Actinum Bacteria*; AC - *Animalia Chordata*; AN - *Animalia Nematoda*; FA - *Fungi Ascomycota*; PB - *Bacteria Proteo*; PM - *Plantae Magnoliophyta*. The ID (second column) is used to identify organisms in figs. 2 and 4. Columns 3, 4 and 5 define the numbers of nodes (*i.e.* proteins) N, total numbers of edges (connections between proteins) M, and average degree $\langle k \rangle$ (number of connections per protein) in the largest cluster. Shadings correspond to fig. 2.

Organism	ID	N	M	$\langle k \rangle$
Drosophila melanogaster - AA	1	3960	44409	22.4
Gallus gallus - AC	2	3723	54131	29.1
Homo sapiens - AC	3	12299	176316	28.7
Mus musculus - AC	4	9595	123665	25.8
Xenopus tropicalis - AC	5	1870	7374	7.9
Caenorhabditis elegans - AN	6	2113	14261	13.5
Aspergillus fumigatus - FA	7	2364	29288	24.8
Saccharomyces cerevisae - FA	8	5209	66057	25.4
Schizosaccharomyces pombe - FA	9	2458	28822	23.5
Arabidopsis thaliana - PM	10	4205	81957	40.0
Rhodococcus sp - AB	11	5540	57992	20.9
Saccharopolyspora erythraea - AB	12	3715	24691	13.3
Aeromonas hydrophila - PB	13	2765	13849	10.0
Bradyrhizobium japonicum - PB	14	4948	29628	12.0
Citrobacter koseri - PB	15	3477	17288	9.9
Escherichia coli - PB	16	3542	25197	14.2
Nocardia farcinica - PB	17	3277	21359	13.0
Pseudomonas aeruginosa - PB	18	3709	20401	11.0
Serratia proteamaculans - PB	19	3392	16978	10.0
Vibrio cholerae - PB	20	2506	12899	10.3

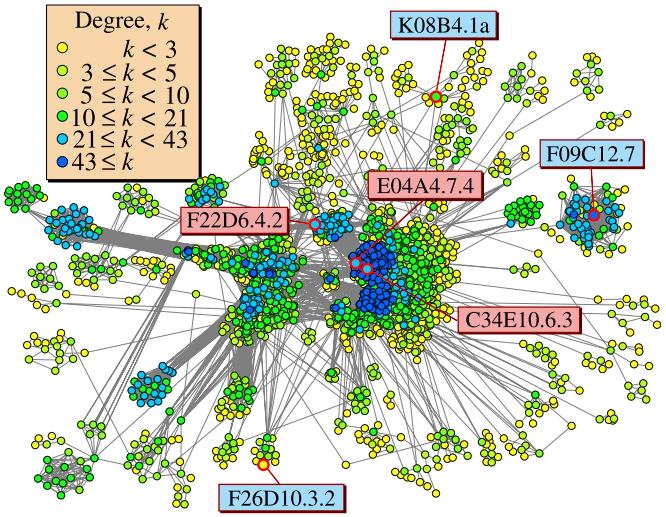

Fig. 1: (Color online) Protein network for *C. elegans* [11]. Modular proteins identified include F09C12.7, an element of major sperm protein, K08B4.1a involved in embryonic development and notch, and F26D10.3.2, which is a heat shock protein. On the other hand, the proteins identified in the central complex are essential to the Krebs cycle: F22D6.4.2 encodes a subunit of NADH dehydrogenase, E04A4.7.4 is better known as cytochrome c 2.1, and C34E10.6.3 is ATP synthase.

randomizing algorithm is repeated T_M times, where T_M ranges from 0 to 10^8. For the organisms that we study, $T_M > 10^7$ ensures that each edge has been swapped on average more than 100 times, effectively destroying any initial correlation in the network. We evaluate correlations between nodes by calculating nearest-neighbor average connectivity [12]

$$\overline{k}_{nn}(k) = \sum_{k'} k' P(k'|k), \tag{1}$$

where $P(k'|k)$ is the conditional probability that a node with degree k is connected with one of degree k'.

Given a network and its surrogates, we evaluate the "robustness" (defined shortly) of the network to random or targeted errors. For biological networks, random failures (RA) [13] take into account single gene changes due to radiation or mutagen exposure and errors in transcription. By contrast, hub malfunctions describe situations in which pathogens or toxins interfere with high-degree hubs of the network. Such a perturbation is termed a "high-degree adaptive attack (HDA)" in the literature [14–16]. To define the robustness of a network against either random or targeted failure, we evaluate the sum of the fractions of the largest connected cluster while removing all nodes,

$$R = \frac{1}{N+1} \sum_{Q=0}^{N} s(Q), \tag{2}$$

where N is the number of nodes in the initial network and $s(Q)$ is the fraction of nodes in the largest connected cluster compared to the initial number of nodes after removing Q nodes. This measure has the advantage over other, *e.g.* percolation [14,17], metrics of robustness in that

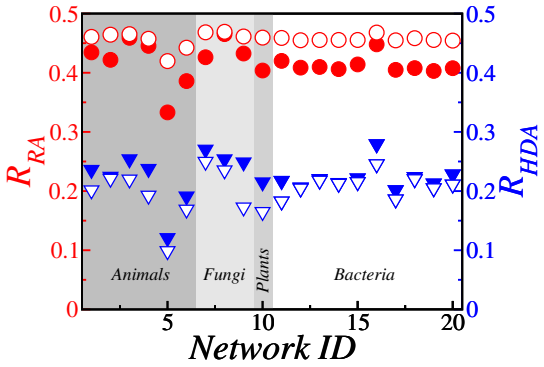

Fig. 2: (Color online) Robustness of the 20 protein networks from table 1 against random and targeted failures. Notice that for every organism studied, the robustness against random failures is *smaller* than surrogates with identical degree distributions, while the robustness against hub malfunction is *larger* than such surrogates. Solid and open symbols correspond, respectively, to biological data and surrogates. Error bars, defining standard deviations over 20 randomized surrogate trials, are smaller than the symbol sizes.

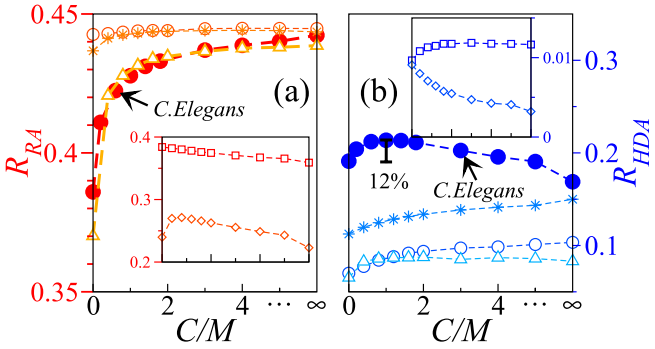

Fig. 3: (Color online) Fundamentally different behaviors of fragile and robust networks. Robustness against random errors, $R_{\rm RA}$, *increases* with an increasing fraction C/M of changed edges for *C. elegans*(filled circles) and other networks such as airline (triangles) [21], citation(stars) [22] and point-of-presence networks (open circles) [23], while by contrast network robustness *decreases* with C/M for the Internet (squares) [24] and corporate ownership network (diamonds) [25]. Note that the improvement in robustness against random errors is significantly larger for *C. elegans* and airline networks, both of which are modular, and is opposite to that of the Internet (inset). Likewise the robustness against hub malfunction, $R_{\rm HDA}$, differs between biological and other networks. $R_{\rm HDA}$ increases with C/M up to 12% until $C/M \approx 1$, after which $R_{\rm HDA}$ decreases for biological networks, in contrast with all other networks except for the ownership network, for which $R_{\rm HDA}$ monotonically increases with C/M. For better visibility some data are shown in the insets having abscissas using the same axes as the main plot; curve fits are included to aid the eye.

it can distinguish between different networks with similar "percolation thresholds", at which a significant number of elements of a network form a single cluster [18]. Even in the case of absence of a percolation threshold, R can distinguish between networks. The normalization $1/(N+1)$ in eq. (2) ensures that the robustness is comparable for different network sizes, and the value of R lies between $1/(N+1)$ and 0.5. The lower limit on R corresponds to entirely isolated nodes, and $R = 0.5$ defines a network where all active nodes remain in a single cluster.

We examine protein networks of 20 different organisms in the *Bacteria* and *Eukarya* domains, identified in table 1. Our measure of robustness is essentially unaffected by the small number of isolated nodes that are detached from the largest cluster, so we neglect these in our analysis. We used the STRING 8.2 "Combined Score" (CS) [19], a measure of the likelihood that two proteins interact in a given network, to impose the criterion that edges $e_{i,j}$ are included in the network only if CS_{ij} is over a threshold value, $\mathrm{CS}_{th} = 70\%$. Smaller values of CS_{th} produce dramatic growth in numbers of edges, masking relevant information with extraneous data, while larger CS_{th} excludes known protein interactions [20].

Results. – Typical results are presented in figs. 2–4, showing the dependence of robustness on random and targeted failures for several network types. As one would expect, for all networks the tolerance to random errors is high (fig. 2, red data), and tolerance to targeted errors is low (fig. 2, blue data). However, unexpectedly we find that all biological networks studied have a significantly *lower* resistance to random failures, and significantly *higher* resistance to hub malfunction than do surrogates, randomized $T_M = 10^8$ times, as described previously. This paradoxical behavior is surprising, and can be analyzed in further detail as shown in fig. 3. In that figure, we plot detailed responses to systematic randomization, using *C. elegans* as an exemplar, compared with several non-biological networks.

For all networks in fig. 3(a), we find that small amounts of random rewiring improve network robustness to random failures. For biological and other modular networks (for example airlines, shown as triangles in the plot), the improvement is much larger than for less obviously modular networks such as citations or access points ("points-of-presence") to the Internet.

Thus a first and unexpected finding of this analysis is that biological proteins and other real networks are less than ideally organized from the point of view of robustness against random failures. A second unexpected finding, shown in fig. 3(b), is that although biological protein networks are more than twice as robust against hub malfunction than any other network tested, modest modifications of the protein interaction structure can improve the network robustness from 2% for *H. sapiens* and 12% for *C. elegans* (fig. 3(b)) up to 28% for *G. gallus*. Apparently, despite the manifest twofold improvement in robustness shown in fig. 3(b) that evolution has produced, life remains among a class of networks that are more fragile to either random or targeted failures than slightly modified surrogates.

This effect, which holds for all of the 20 organisms studied, differs markedly from a second class of networks, shown in the insets to figs. 3(a),(b), that is exemplified by the Internet [24], which was designed for maximal robustness against errors [26], and to a lesser extent corporate ownership networks, that are robust by virtue of similarly numerous inter-relations [25]. Our findings therefore indicate that although the Internet and PPI networks share broad degree distributions, the two types of networks behave fundamentally differently in their overall fragility as measured by comparison with modified surrogates.

To investigate the consistency of these results, we repeated our analyses under various modifications. First, we evaluated the reliability of the data itself by considering both a higher value of the threshold likelihood of protein interactions, $\mathrm{CS}_{th} = 80\%$, and data from a different version of STRING 8.1 [27]. Second, we considered whether the robustness could be an indirect effect of a change in correlation between nodes —for example as high-degree nodes are swapped with low-degree ones. For this purpose, we modified the rewiring to preserve correlations by performing swaps between pairs of nodes $\{(i,j),(k,\ell)\} \to \{(i,\ell),(k,j)\}$ only if the degrees of i and k or j and ℓ were equal. Third, we considered the effect of randomly removing individual edges described again by eq. (2), but with N defined to be the number or edges, rather than nodes. In each of these independent tests, we found the same features commented on for figs. 2 and 3, supporting the two key results that biological networks exhibit more fragility to random errors than similar non-biological networks, and that although biological networks are more than twice as robust against hub malfunction as non-biological networks, they remain less than optimal robust.

To analyze the causes of these unexpected behaviors, we return to the observation made earlier that PPI networks are intrinsically modular. Since modules have many fewer nodes than the central network, it follows that any switch involving a node in a module is highly likely to involve a second connection that is outside of the module. Such a switch will produce two new edges, both of which will connect the module to the central network, so switching connections will typically increase the number of connections from a module to the central network. This in turn will improve the robustness of that module to either random or targeted failures, since such a switch would increase the number of connections that would have to be broken between the module and the rest of the network. We can test this mechanism by constructing model networks with suitable properties, two of which are shown in fig. 4. In that figure, we compare both simple (model A) and more finely tuned model networks (model B) with the biological data that appear as solid symbols in fig. 3. The simple model (left inset) is constructed by creating a central large complex with broad degree distribution. An arbitrary small number

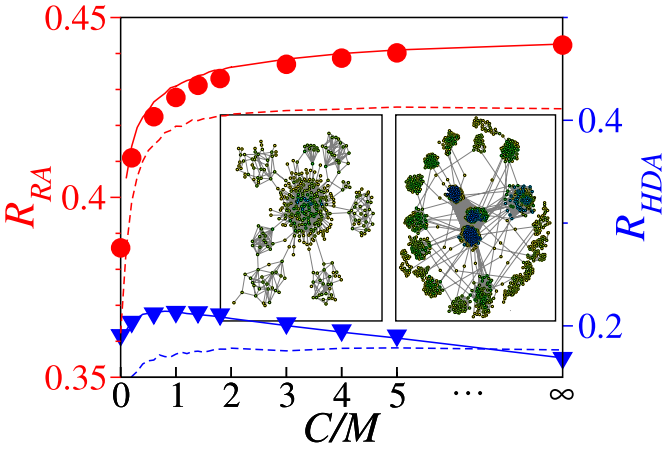

Fig. 4: (Color online) Effects of modularity and correlations on robustness against random (red) and targeted (blue) failures in model (curves) and biological networks (data points, for *C. elegans*). Insets show a simple network (left, model A) without correlations and our tuned model network (right, model B) with correlations. Network robustnesses are shown as dashed and solid lines for model A and model B respectively. While the simple model converges to a constant value R, the tuned model shows increasing and decreasing trends in agreement with the biological data.

(eight, here) of modules, each with different numbers of nodes but the same number of random connections, are added and are attached to random nodes in the central complex by two connections. The robustness in response to random and targeted failures is then evaluated exactly as before, and is plotted in fig. 4 as dashed curves. By contrast, non-modular model networks that we have constructed (not shown) have few vital hubs and so exhibit identical responses to either random or targeted failures, with no dependence on C/M. Evidently, the qualitative behavior of biological responses to random as well as targeted failures can be attributed to the modular structure of biological protein networks. Indeed, it is not difficult to tune the model network to nearly exactly fit the biological data. This is shown in the right inset of fig. 4, where we display a fictitious network whose random and targeted response curves are shown as solid lines in the main plot. This network is constructed by choosing the number of connections of the model to be similar to the biological one. In detail, the nodes are distributed in 20 modules with different densities, in which high-degree nodes are preferentially connected to high-degree nodes. This preferential connectivity is crucial to the reduction in robustness to hub malfunction: for no other structural feature investigated was this reduction seen. These modules are connected preferentially to the largest module with few connections, as we have remarked occurs in biological networks. Back to figs. 2 and 3, they show that for surrogates with large numbers of switches of connections, the robustness of PPI networks to hub malfunction actually *decreases* for all organisms studied.

This behavior can also be reproduced in model networks provided, crucially, that connections are preferentially included between high-degree nodes (see also [28]). In this case, two competing effects arise. The randomization of the modules increases robustness, while the vanishing of the preferential connections decreases the robustness. In case of random errors the second effect is negligible, but for hub malfunction, it leads to the surprising decrease in robustness that we have noted.

In conclusion, we have demonstrated that biological protein networks are unexpectedly fragile against either random or targeted failures. This fragility is measurable by comparison with surrogates with identical network statistics. We find that these behaviors are characteristic of modular networks, in which particular products or processes inhabit isolated modules. As anticipated earlier in this letter, we have confirmed 1) that this modular structure causes biological protein networks to be more vulnerable to targeted than random failures, and 2) that through evolution these networks have become more robust than non-biological networks against hub malfunction. Nevertheless, as we have shown, protein networks are more fragile than extensively rewired surrogates to random errors, while being *less* fragile than the same surrogates to hub malfunction. We find that this final phenomenon is associated with the apparently unique tendency of high-degree nodes in PPI networks to preferentially connect to other high-degree nodes. We speculate that this preferential connectivity may have practical advantages, for example in providing redundant pathways to permit key processes to function after a hub malfunction or genetic deletion [4,5].

* * *

The authors thank Prof. R. M. C. Almeida for helpful discussions on the use of biological data. CMS and HJH acknowledge financial support from the ETH Competence Center "Coping with Crises in Complex Socio-Economic Systems" (CCSS) through ETH Research Grant CH1-01-08-2 and FUNCAP. RFSA thanks CNPq, SECTI-FAPESB-PRONEX, and INCTI-SC for financial support.

REFERENCES

[1] Kohn K. W., *Mol. Biol. Cell*, **10** (1999) 2703.
[2] Kohn K. W. *et al.*, *Mol. Syst. Biol.*, **51** (2006) 1.
[3] Alon U., *Science*, **301** (2003) 1866.
[4] Jeong H. *et al.*, *Nature*, **411** (2001) 41.
[5] Samanta M. P. and Liang S., *Proc. Natl. Acad. Sci. U.S.A.*, **100** (2003) 12579.
[6] Sharan R., Ulitsky I. and Shamir R., *Mol. Syst. Biol.*, **3** (2007) 88.
[7] Ravasz E. *et al.*, *Science*, **297** (2002) 1551.
[8] Rives A. W. and Galitski T., *Proc. Natl. Acad. Sci. U.S.A.*, **100** (2002) 1128.
[9] Hartwell L. H. *et al.*, *Nature*, **402** (1999) C42.
[10] Maslov S. and Sneppen K., *Science*, **296** (2002) 910.
[11] Batageli V. and Mrvar A., *Connections*, **21** (1998) 47.
[12] Pastor-Satorras R., Vázquez A. and Vespignani A., *Phys. Rev. Lett.*, **87** (2001) 258701.
[13] Albert R., Jeong H. and Barabási A. L., *Nature*, **406** (2000) 378.
[14] Holme P. *et al.*, *Phys. Rev. E*, **65** (2002) 056109.
[15] Cohen R. *et al.*, *Phys. Rev. Lett.*, **85** (2000) 4626.
[16] Tanizawa T. *et al.*, *Phys. Rev. E*, **71** (2005) 047101.
[17] Rozenfeld H. D. and ben-Avraham D., *Phys. Rev. E*, **75** (2007) 061102.
[18] Schneider C. M. *et al.*, *Proc. Natl. Acad. Sci. U.S.A.*, **108** (2011) 3823.
[19] `http://string-db.org`.
[20] von Mering C. *et al.*, *Nucleic Acids Res.*, **33** (2005) D433.
[21] Colizza V., Pastor-Satorras R. and Vespignani A., *Nat. Phys.*, **3** (2007) 276.
[22] `vlado.fmf.uni-lj.si/pub/networks/data/cite/default.htm`.
[23] Shavitt Y. and Zilberman N., *INFOCOM IEEE Conference on Computer Communications Workshops (San Diego, USA)* 2010.
[24] `www.netdimes.org`.
[25] Norlen K. *et al.*, *Proceedings of International Telecommunications Society 14th Biennial Conference 2002 (Seoul, Korea)* 2002.
[26] Willinger W. and Doyle J., *Robust design: A Repertoire of Biological, Ecological, and Engineering Case Studies*, edited by Jen E. (Oxford University Press, New York) 2002.
[27] `http://string-db.org/server_versions.html`.
[28] Song C., Havlin S. and Makse H., *Nat. Phys.*, **2** (2006) 275.