Supporting Information
Predicting effects of structural stress in a genome-reduced model bacterial metabolism

Oriol Güell[1], Francesc Sagués[1] & M. Ángeles Serrano[2]

July 30, 2012

1. Departament de Química Física, Universitat de Barcelona, Martí i Franquès 1, 08028 Barcelona, Spain

2. Departament de Física Fonamental, Universitat de Barcelona, Martí i Franquès 1, 08028 Barcelona, Spain

Contents

1 Topological analysis

Metabolic networks can be represented as (semi)directed or undirected networks. Directed networks allows us to distinguish between reactants and products and between reversible and irreversible reactions. However, an undirected version of the metabolic network of the organisms is often reconstructed in order to compute the degree distribution of metabolites and reactions (Figure S 1).

For metabolites, we show the cumulative probability distribution function ($P(k'_M \geq k_M)$), whereas for reactions we compute the direct probability distribution function ($P(k_R)$). Metabolites display characteristic scale-free degree distributions $P(k) \sim k^{-\gamma}$ with exponents that are rather similar. To check the validity of the null model that the observed empirical metabolite degree distributions have been generated by a power law, we perform goodness of fit tests. We compute the Kolmogorov statistic

$$D = \max_{k \geq k_{min}} \left| P_c(k) - \frac{\sum_{k'=k} k'^{-\gamma}}{\sum_{k'=k_{min}} k'^{-\gamma}} \right|, \quad (1)$$

where k_{min} is the minimum degree beyond which we expect the power law to hold and $P_c(k)$ is the empirical complementary cumulative degree distribution of metabolites. The exponent γ and the minimum degree k_{min} are computed using maximum likelihood methods as described in Ref. (1), resulting in $k_{min} = 2$, $\gamma = 2.44 \pm 0.04$, $D = 0.017$ for the *E. coli* metabolism, $k_{min} = 3$, $\gamma = 2.20 \pm 0.09$, $D = 0.039$ for *S. aureus*, and $k_{min} = 2$, $\gamma = 2.3 \pm 0.1$, $D = 0.070$ for *M. pneumoniae*. According to the Kolmogorov Smirnov (KS) test, the variable $\sqrt{N}D$ follows the Kolmogorov distribution $P_K(K)$, of which 95% confidence level is at $K_{95\%} = 1.35$. Given the size of our samples, we obtain $\sqrt{N}D = 0.63 < 1.35$ for *E. coli*, $\sqrt{N}D = 0.53 < 1.35$ for *S. aureus*, and $\sqrt{N}D = 0.83 << 1.35$ for *M. pneumoniae*. This implies that the null model cannot be ruled out and, consequently, that power laws are a plausible explanation of these metabolite degree distributions. Reactions show a peaked distribution centered at the correspondent average degree of each network. In this case, all the degree distributions have a maximum at the value of 4. Strong similarities in both distributions are evidenced for the three analyzed species.

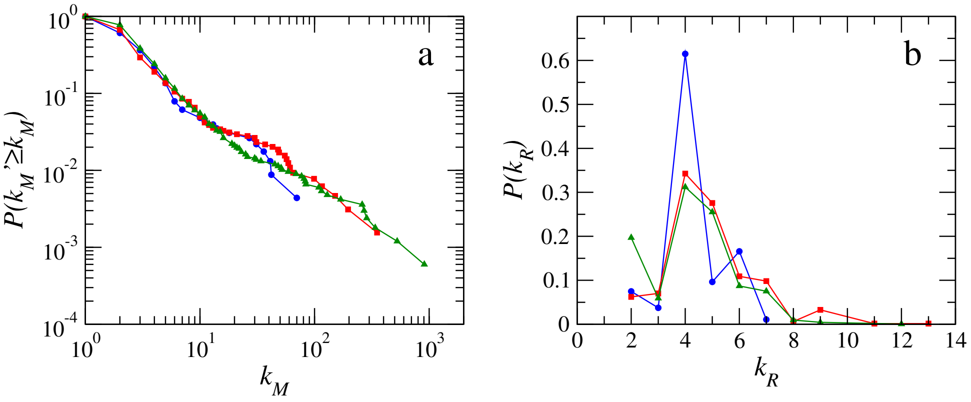

Figure S 1: Topological analysis of *M. pneumoniae* (blue), *S. aureus* (red) and *E. coli* (green). a). Cumulative probability distribution function (CPDF) of metabolite degrees. b). Probability distribution function (PDF) of reaction degrees.

2 Implementation of cascades

Failure cascades in metabolic networks triggered by specific reactions propagate by turning further reactions non-operative. We follow the algorithm in Ref. (2). When a reaction fails, it affects its metabolites such that some may become inviable, in the sense that their concentrations cannot be maintained anymore at stationary values and accumulate or deplete, which in turn affects other reactions. The algorithm that propagates the cascade follows the sequence:

1. Choose the triggering reaction and remove all the links that it shares with its metabolites.

2. Check if the affected metabolites remain viable, $k_{in} \neq 0$ and $k_{out} \neq 0$ (except metabolites exchanged with the environment, which must have just one of the degrees different from zero). If the metabolite is inviable, remove all the edges that it shares with other reactions, that become non-operative.

3. Remove all reactions that became non-operative.

4. Repeat the last two steps until all the reactions that remain are operative.

Reversible reactions are handled as in Ref. (2). They are decoupled in two half-nodes, the forward and the reverse sense of the reaction. A cascade propagating to a metabolite of a reversible reaction fixes it in the forward or reverse direction depending on whether the lone incoming or outgoing link left to the affected metabolite is connected to the forward or reverse half of the reaction. In all cases, when any metabolite of a reversible reaction has this reaction as the lone producing and consuming it, the reaction must be removed to satisfy the viability criterion.

3 Null models of randomized metabolic networks

Null models are constructed by picking at random a pair of links of the network and swapping the end of the links. Nevertheless, there are some aspects that must be considered. The number of in, out and bidirectional links must be preserved, as well as the fact that metabolites must be connected to reactions and reactions with metabolites. Also, there must be neither self-links nor repeated links. The number of potentially realized moves is n_{links}^2 (where n_{links} is the sum of in, out and bidirectional links) and 100 realizations of each network are used for statistical analysis.

4 Metabolic effects of the failure of individual reactions

4.1 General results

In Figure S 2 we plot the damage distributions obtained from single reaction failures for the three considered species. While some do not propagate at all, there are several reactions whose associated damages are very large and trigger cascades which are potentially lethal. Reactions prone to induce vulnerability are listed in Table S 1, together with their associated damages and corresponding values of the predictor index (See main text).

Reaction	Damage	Predictor
9	32	1
10	32	10
138	14	3
145	14	4
141	13	4
178	13	1
187	13	1
1	11	19
77	9	4
142	9	10
55	9	4
165	9	5
105	8	3
73	6	2
160	6	3
161	6	1

Table S 1: Reactions corresponding to largest damages and associated predictor values.

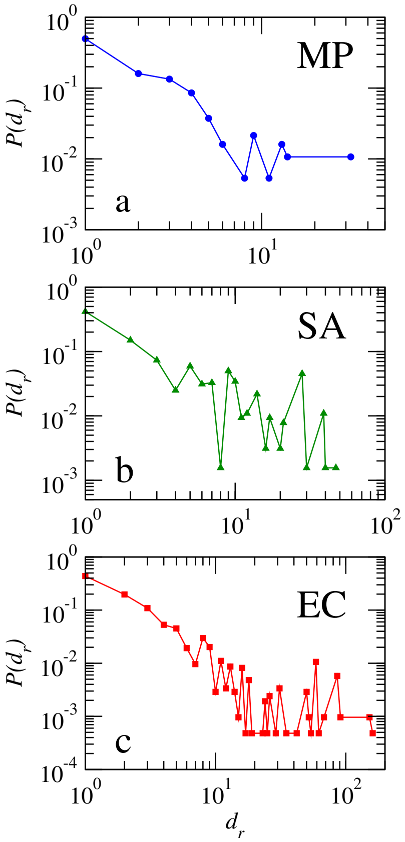

Figure S 2: Probabilities of damage spreading from single reacions. a) *M. pneumoniae* (blue). b) *S. aureus* (red). c) *E. coli* (green). Notice the large number of reactions which do not propagate when they are removed ($d = 1$).

4.2 Motifs triggering large cascades

Cascades propagate through motifs which display characteristic structures shown in Figure S 3. Note that a distinction must be made between motifs which are potential triggers and those that are real triggers with $d > 1$. This distinction comes from the fact that potential triggers involve necessarily reversible reactions. The latter sense might be driven after the removal of the initial reaction (see above for the handling of reversible reactions). If such a process renders the reversible reaction not viable, then the cascade is going to spread and the damage is going to be greater than 1. Otherwise, if the reaction is left viable, the cascade is going to stop in that step and thus the damage is going to have a value of 1. In any case, both real and potential triggers are accounted for in the expression of the predictor (see main text).

An example of the application of Eq. 1 of the main text is shown in Fig 4.

5 Metabolic effects of knocking out gene co-expression clusters

5.1 Hierarchical clustering

This method is used in Ref. (3). It is based on transforming the correlation between genes into average distances between them. Therefore, clusters of genes are genes which are near to each other.

5.2 Infomap

This method is based on the algorithm of Ref. (4) called Infomap. It detects clusters of genes depending on how information flows through the network. Clusters are thus groups of nodes where information flows easily and quickly.

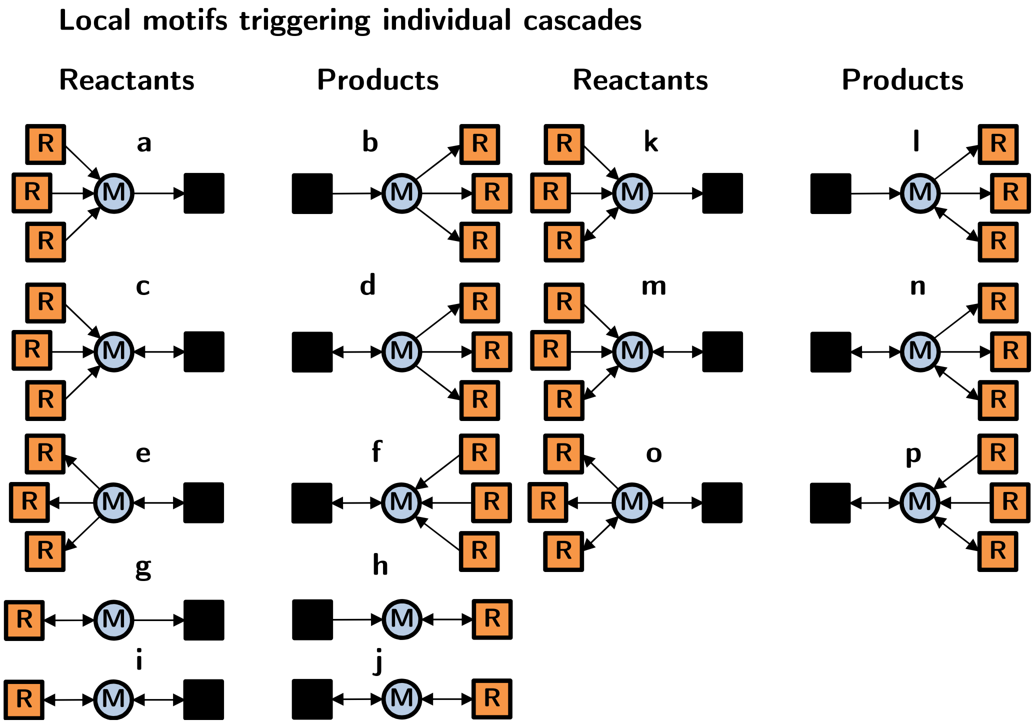

Figure S 3: Motifs of cascade propagation after failure of individual reactions. Cases a-j are going to result into cascades with d larger than 1, while cases k-p correspond to potential transmitters in the sense that they may or may not spread the cascade.

5.3 Recursive percolation

Recursive Percolation is a method that identifies clusters depending on the correlation intensities. The higher the correlation, the more probable is that two nodes belong to the same cluster. This method proceeds by finding the percolation threshold of the selected set of links iteratively until a criterium is satisfied (Figure S 5). In this case, the used criterium was that the distribution of sizes were similar to the corresponding ones for hierarchical clustering and Infomap (Figure S 6).

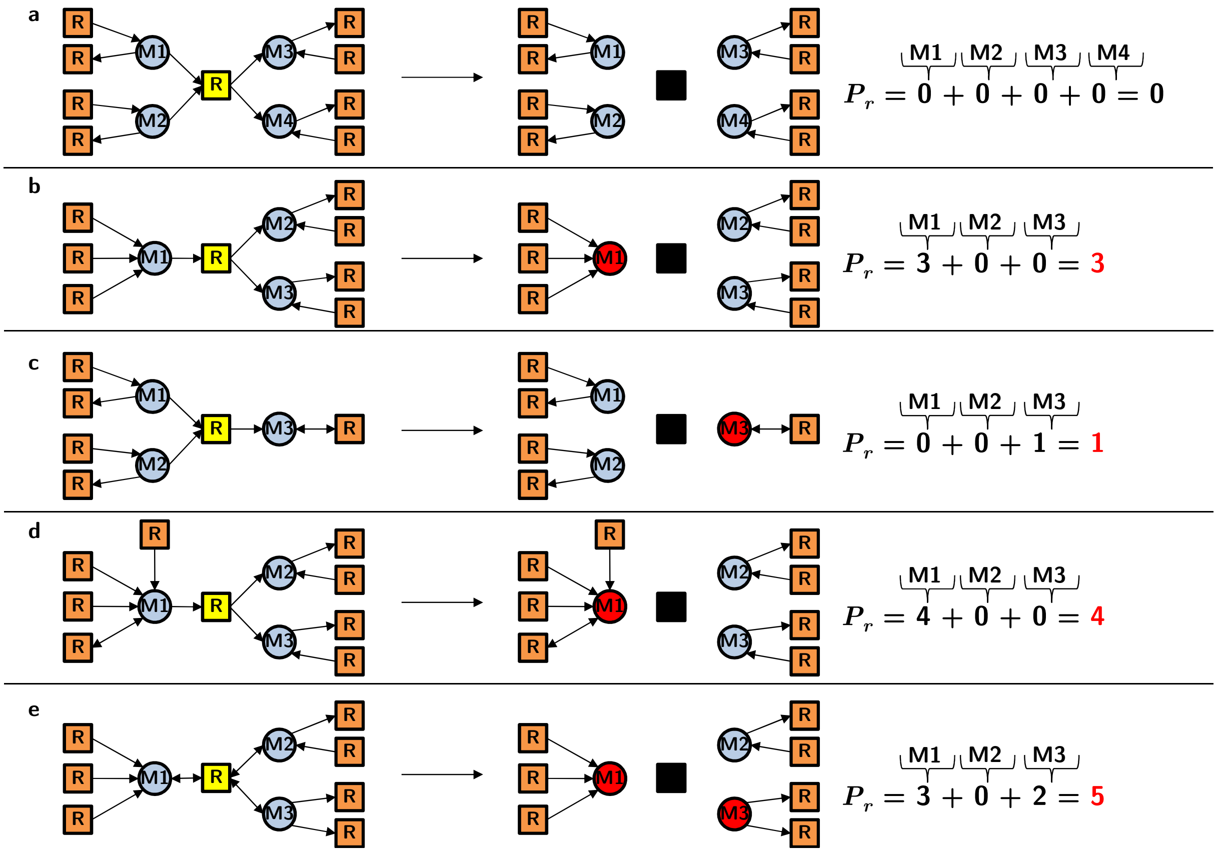
a
b
c
d
e
R
M1
M2
M3
M4
$P_r = \underbrace{0}_{M1} + \underbrace{0}_{M2} + \underbrace{0}_{M3} + \underbrace{0}_{M4} = 0$
$P_r = \underbrace{3}_{M1} + \underbrace{0}_{M2} + \underbrace{0}_{M3} = 3$
$P_r = \underbrace{0}_{M1} + \underbrace{0}_{M2} + \underbrace{1}_{M3} = 1$
$P_r = \underbrace{4}_{M1} + \underbrace{0}_{M2} + \underbrace{0}_{M3} = 4$
$P_r = \underbrace{3}_{M1} + \underbrace{0}_{M2} + \underbrace{2}_{M3} = 5$

Figure S 4: Examples of Eq. 1 of main text to several cases. Triggering reactions are colored yellow, whereas metabolites which spread the cascade are colored red. For clarity, the contribution of each metabolite to the value of P_r is also given.

5.4 General results

The distribution of the size of the clusters is given in Figure S 6 for the three considered methods. Although they all follow a similar power-law distribution, the composition of the clusters is not equal. Recursive Percolation is the method where the probability of finding large clusters is lower, whereas Infomap has the largest probability.

To check if the composition of the clusters is relevant, we compared the obtained damages with a null model. In this case, the null model consists in randomizing the metabolic genes of the clusters, maintaining constant the number of metabolic genes in each cluster. The regulation

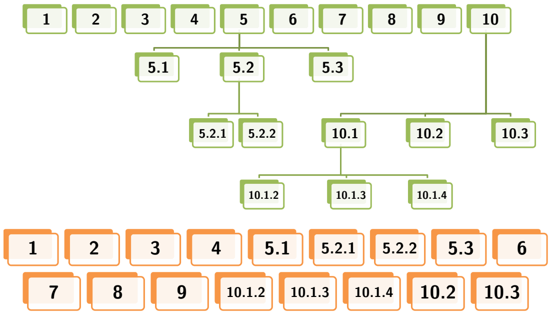

Figure S 5: Example of application of the method Recursive Percolation to a matrix of correlated data. The first step leads to 10 clusters. Among these 10 clusters, the largest are fragmented, leading to more clusters. This is done until the distribution of sizes is similar to that found in the other methods.

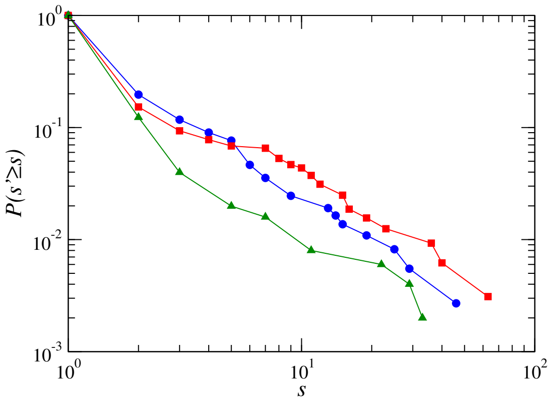

Figure S 6: Size distribution of clusters in terms of their cumulative probability distributions. a) Hierarchical clustering (blue). b) Infomap (red). c) Recursive Percolation (green). They all show similar power-law cluster size distributions.

of the reactions is neither modified. The number of trials is n_{genes}^2 (where n_{genes} is the number of metabolic genes) for each realization and we performed 100 realizations for statistical analysis.

6 Data

Along with this file, we also provide a .xls file as supporting information containing the edge list of *M. pneumoniae*, all the reactions with their associated damages, and the cluster where each gene belongs to for the three different methods.

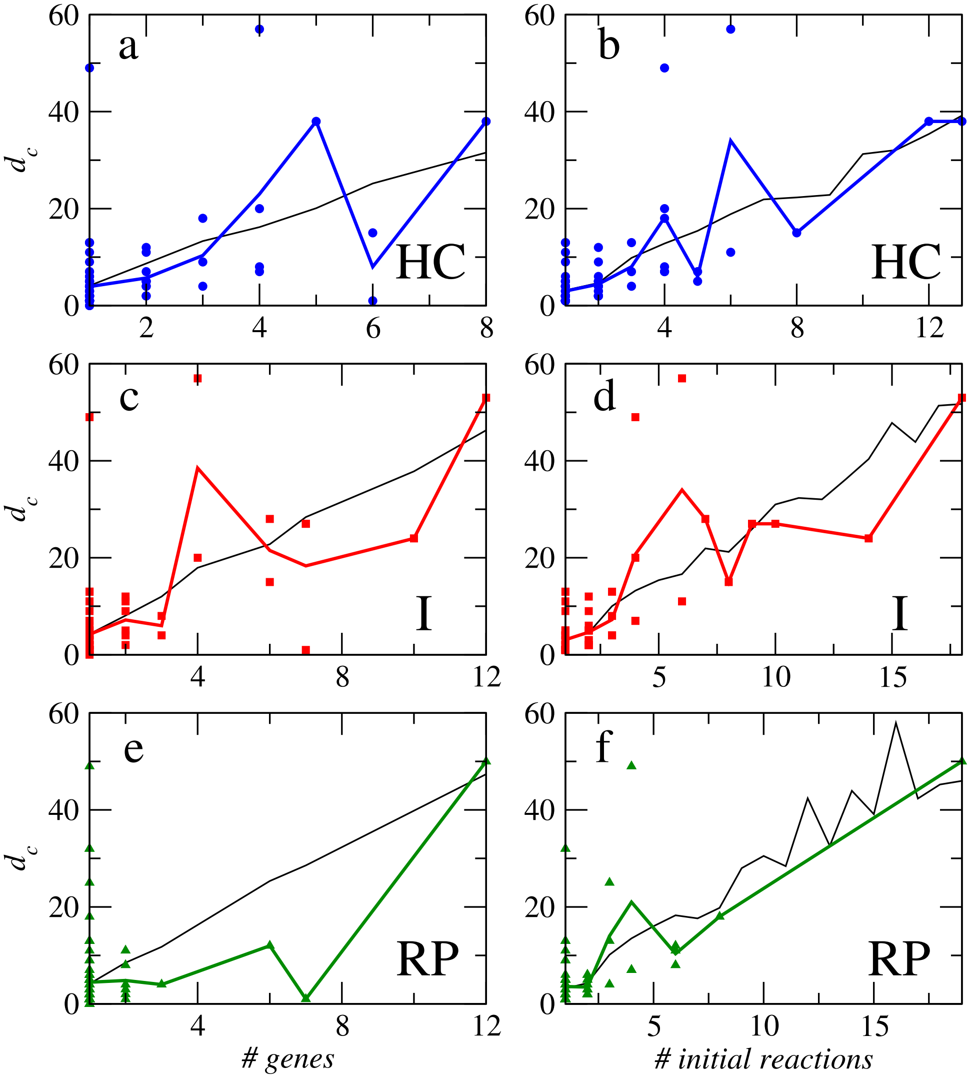

Figure S 7: Damage distributions as a function of the number of genes and reaction failures, similar to Fig. 4 in the main text, but now randomizing the specific genetic contents of each cluster while maintaining the total number of metabolic genes in each cluster.

References

[1] Clauset, A., Shalizi, C. R., Newman, M. E. J. Power-law distributions in empirical data. *SIAM Review* **51**, 661–703 (2009).

[2] Smart, A. G., Amaral, L. A. N., and Ottino, J. Cascading failure and robustness in metabolic networks. *Proc. Natl. Acad. Sci. USA* **105**, 13223–13228 (2008).

[3] Güell, M. *et al.* Transcriptome complexity in a genome-reduced bacterium. *Science* **326**, 1268–1271 (2009).

[4] Rosvall, M. and Bergstrom, C. T. Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci. USA* **105**, 11181123 (2008).