

Metabolic network destruction: Relating topology to robustness

Wynand Winterbach^{a,b,*}, Huijuan Wang^b, Marcel Reinders^a, Piet Van Mieghem^b, Dick de Ridder^a

^a The Delft Bioinformatics Lab, Delft University of Technology, P.O. Box 5031, 2600 GA Delft, The Netherlands¹

^b Network Architecture and Services, Delft University of Technology, P.O. Box 5031, 2600 GA Delft, The Netherlands²

ARTICLE INFO

Article history:

Received 15 April 2011

Received in revised form 12 May 2011

Accepted 13 May 2011

Available online 15 June 2011

Keywords:

Metabolic networks

Flux balance analysis

Network topology

Robustness

ABSTRACT

Biological networks exhibit intriguing topological properties such as small-worldness. In this paper, we investigate whether the topology of a particular type of biological network, a metabolic network, is related to its robustness. We do so by perturbing a metabolic system *in silico*, one reaction at a time and studying the correlations between growth, as predicted by flux balance analysis, and a number of topological metrics, as computed from three network representations of the metabolic system.

We find that a small number of metrics correlate with growth and that only one of the network representations stands out in terms of correlated metrics. The most correlated metrics point to the importance of hub nodes in this network, so-called “currency metabolites”. Since they are responsible for interconnecting distant functional modules in the network, they are important points in the network for predicting if reaction removal affects growth. A second set of correlations in contrast is related to “loner” nodes that uniquely connect important pathways and thus correspond to essential steps in metabolism.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

In the last decade, advances in high-throughput biological measurement systems have made it possible to extract large-scale networks from biological systems. Jeong et al. [7] were among the first to study the topologies of metabolic networks, networks of interconversions of small compounds. The metabolic networks of the 43 organisms that they studied gave evidence of a scale-free structure. Characteristic properties of these so-called “small-world” networks are their power-law distributed node degrees and their small average shortest path lengths.

Subsequently, researchers studied the topologies of a number of other types of biological networks [2,12,3].

Much of this work confirmed the results of Jeong et al.: scale-free behavior was everywhere. Even the Internet and some power grids are thought to display scale-free behavior [1]. These latter networks have expanded in a seemingly organic fashion through a process of *preferential attachment* – new nodes are more likely to attach to existing high-degree nodes than to low-degree nodes. This expansion process forms the basis of Barabási and Albert’s [1] random network model. They show that it leads to the characteristic power-law node-degree distribution and small-world properties. Although Kim and Marcotte [8] and Lima-Mendez and Helden [11] argue that biological networks do not develop through simple processes of preferential attachment, the presence of similar topological elements, such as hub nodes, begs the question whether these topological properties confer some benefit or whether certain topologies are inherently suited for particular functionality.

In an effort to understand the relationship between the function of a network and its topological properties, Milo et al. [13] introduced the concept of *motifs*. A

* Corresponding author. Tel.: +31 15 27 81782; fax: +31 15 27 81774.

E-mail addresses: w.winterbach@tudelft.nl (W. Winterbach), h.wang@tudelft.nl (H. Wang), m.j.t.reinders@tudelft.nl (M. Reinders), p.f.a.vanmieghem@tudelft.nl (P. Van Mieghem), d.deridder@tudelft.nl (D. de Ridder).

¹ WWW home page: <http://bioinformatics.tudelft.nl>.

² WWW home page: <http://www.nas.ewi.tudelft.nl>.

motif is a small sub-network (3–5 nodes) whose over-representation may be indicative of its role in maintaining function at a local level. They found that certain motifs occur more often in biological networks than expected by chance and that they may correspond to certain desired behavior such as response acceleration, signal delay and stability. Prill et al. [18] took this idea further and claimed that certain motifs were inherently more prone to display stable behavior than others. By abstracting away from the underlying functionality, they demonstrated that such relations held to some extent over a variety of biological networks. However, Ingram et al. [6] considered gene networks and compared the results of a differential equation model of gene expression to specific motif counts in the gene network but found no correlation. Lima-Mendez and Helden [11] argue that global topological properties cannot explain the function of networks. While they claim that the significance of motif frequencies may have been overestimated (since the frequencies only capture global properties), they do consider a localized approach to be more promising as the key to understanding biological networks lies in understanding local details.

In our work, we take a global approach and investigate to what extent network topology can be related to more systems-level network properties shared by the various network types studied by Barabási et al. An interesting property in this respect is that of *robustness*. Stelling et al. [21] and Kitano [9] define robustness as the ability of a system to maintain its function in the face of perturbations or uncertainty. Biological systems are known to be robust [10] to many forms of perturbation while being highly sensitive to other forms, so-called “highly optimized tolerance” [21]. The question is whether there is something in the topology of these networks that confers robustness to the overall system.

In this paper, we study the relationship between the robustness of a micro-organism (baker’s yeast, *Saccharomyces cerevisiae*) and the topologies of network representations of its metabolic system. Microbial metabolic systems provide a good test bed, since an often assumed functional objective – growth – is easily expressed in terms of fluxes through these systems. Furthermore, good quality metabolic datasets are readily available and resulting flux models can be studied computationally with high efficacy.

To study the link between network topology and robustness, we propose an *in silico* metabolic system perturbation experiment. We define robustness as the ability of the yeast cell to maintain growth under reaction removals. First, we show how its metabolic system can be represented by three different networks. Then, through a number of trials, reactions are removed from the metabolic system until growth ceases. This provides a number of snapshots of partially “destroyed” metabolic systems. For each snapshot, growth and a number of network-wide topological metrics can be computed. By calculating correlations between growth and these metrics, we find that most of the topological metrics are not related to function. The strongest correlations point to the importance of both “hub” nodes (so-called “currency metabolites”) and “loner” nodes.

2. Method

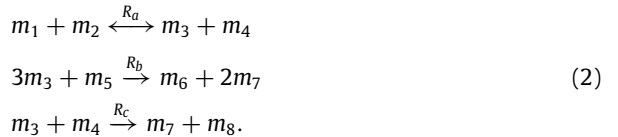
2.1. Computing function

In this work, we define robustness as the maintenance of cell growth under perturbations to the organism’s metabolic system when reactions are removed from the metabolic network. A metabolic system with r reactions and m metabolites is modeled by a set of m differential equations:

$$\frac{dX_i}{dt} = s_{\text{syn}} v_{\text{syn}} - s_{\text{deg}} v_{\text{deg}} - s_{\text{use}} v_{\text{use}} + s_{\text{trans}} v_{\text{trans}} \quad (1)$$

that specify how the concentration X_i of a metabolite i changes in time. v_{syn} is the rate of metabolite synthesis, v_{deg} is the degradation rate, v_{use} is the rate of consumption (by other reactions) and v_{trans} is the rate of transport across the cell boundary (into the cell). v_{syn} , v_{deg} and v_{use} are generally nonlinear functions whose behavior is governed by the kinetic parameters of the enzymes catalyzing the reactions in which they take part and by concentrations of other metabolites. Because the kinetic parameters are not generally known and must be estimated, it is difficult to solve the differential equations directly. s_{syn} , s_{deg} , s_{use} and s_{trans} are *stoichiometric coefficients*³ (reaction rates are measured in $\mu\text{mol gDW}^{-1} \text{h}$, i.e., micromoles per gram of dry weight per hour).

We assume that $s_{\text{trans}} v_{\text{trans}}$ is a constant value b_i , allowing (1) to be written in vector form as $d\mathbf{X}/dt = \mathbf{S} \cdot \mathbf{v} + \mathbf{b}$, with \mathbf{S} the $m \times r$ *stoichiometric matrix*, \mathbf{v} an $r \times 1$ vector of reaction rates (fluxes) and \mathbf{b} the vector of boundary transport reaction rates. We will use a small example to make the form of \mathbf{S} clear (and later to show how networks are derived from \mathbf{S}). Consider the metabolic system:



The corresponding \mathbf{S} matrix is shown in Fig. 1(a). Since each column is labeled by a reaction R_i , we refer to the corresponding flux value in \mathbf{v} as v_i . At steady-state $d\mathbf{X}/dt = \mathbf{0}$, rendering the linear system:

$$\mathbf{S} \cdot \mathbf{v} + \mathbf{b} = \mathbf{0}. \quad (3)$$

Since \mathbf{S} and \mathbf{b} are constant, \mathbf{v} can be determined without any knowledge of enzyme kinetics. Due to the small size of the example, \mathbf{S} is overdetermined (i.e., there are fewer reactions than metabolites). In real biological networks however, stoichiometric matrices are under-determined. Such systems generally have infinitely many solutions but biologists are only interested in biologically significant ones. A common (biological) assumption is that microbial cells attempt to maximize the rate of their biomass production or in other words, growth. Growth can be expressed as a linear combination $\mathbf{c}^T \cdot \mathbf{v}$ of certain key reaction rates in the

³ These are derived from the chemical mass balance coefficients: e.g. $2\text{H}_2 + \text{O}_2 \rightarrow 2\text{H}_2\text{O}$ corresponds to the stoichiometric coefficient vector $[-2 \quad -1 \quad 2]$.

metabolic system. The reaction rates can then be computed by a linear program:

$$\begin{aligned} &\text{Maximize } \mu = \mathbf{c}^T \cdot \mathbf{v} \\ &\text{subject to } \mathbf{S} \cdot \mathbf{v} + \mathbf{b} = \mathbf{0}. \end{aligned} \quad (4)$$

Positive components of \mathbf{v} correspond to forward-acting reactions, whilst negative components correspond to reactions running in reverse. In (4), the components of \mathbf{v} may assume negative and positive values meaning that any reaction can, in principle, occur in either direction. Due to thermodynamics, some reactions are very unlikely to occur in reverse (in the example, only reaction R_a is reversible). These constraints are modeled by restricting rates of non-reversible reactions to be non-negative. Thus for each non-reversible reaction R , the constraint $v_R \geq 0$ is added, rendering the linear system:

$$\begin{aligned} &\text{Maximize } \mu = \mathbf{c}^T \cdot \mathbf{v} \\ &\text{subject to } \mathbf{S} \cdot \mathbf{v} + \mathbf{b} = \mathbf{0} \\ &\quad v_{R_i} \geq 0 \quad \text{for each non-reversible reaction } R_i. \end{aligned} \quad (5)$$

In addition, biological constraints limit the rates of some reactions. These inequalities are simply added to the list of constraints of the linear program. This steady-state framework for computing metabolic fluxes by optimizing some criterion is known as *flux balance analysis*. Orth et al. [16] give a good overview of the framework.

2.1.1. Testing robustness

We test robustness by iteratively removing reactions and recalculating (5) until growth μ drops below a low threshold value ($1 \times 10^{-9} \mu\text{mol gDW}^{-1} \text{h}$). This produces a sequence $T = \{s_1, s_1, s_2, \dots, s_n\}$ which is referred to as the *trial* T . A *step* is a reaction label index: step s_i corresponds to the removal of reaction R_{s_i} . Removal of a reaction is modeled by removing its corresponding column from \mathbf{S} . The steps in a trial are associated with a sequence of linear programs $P_0, P_1, P_2, \dots, P_n$, where P_0 is the unmodified linear program (from which no reaction has been removed) and P_i is the linear program resulting from the removal of the reactions $R_{s_1}, R_{s_2}, \dots, R_{s_i}$ for $i \geq 1$.

Pseudo-code for the algorithm is shown in Algorithm 1. This algorithm computes the results for one trial. The input is a description of the metabolic system σ and a network metric (that takes a network as input and produces an output of type \mathcal{O}). The i th iteration of the loop corresponds to step s_i .

The function “random-reaction” in Algorithm 1 chooses a random enzyme-catalyzed reaction with uniform probability. Reactions that are not mediated by enzymes but occur due to chemical processes such as diffusion are never removed.

2.2. Topology

To be able to calculate topological properties of the metabolic system, the stoichiometric matrix \mathbf{S} should be represented as a network. However, \mathbf{S} cannot be directly represented as a network since a reaction may interact with more than two metabolites and a metabolite may interact with more than two reactions. A natural

Algorithm 1 Destruction trial (σ : metabolic system, metric: network $\rightarrow \mathcal{O}$)

```

X  $\leftarrow$  empty-list() {List of growth values}
M  $\leftarrow$  empty-list() {List of metric values}
P  $\leftarrow$  to-linear-program( $\sigma$ ) {Compute  $P_0$ }
 $\mu$   $\leftarrow$  growth-rate( $P$ )
while  $\mu > 1 \times 10^{-9}$  do {One step  $s_i$  in the current trial}
  R  $\leftarrow$  random-reaction( $\sigma$ ) {Pick  $R_{s_i}$ }
   $\sigma$   $\leftarrow$  remove-reaction( $\sigma, R$ )
  P  $\leftarrow$  to-linear-program( $\sigma$ ) {Compute  $P_i$ }
   $\mu$   $\leftarrow$  growth-rate( $P$ )
  g  $\leftarrow$  network( $\sigma$ )
  m  $\leftarrow$  metric(g)
  X  $\leftarrow$  append-to-list(X,  $\mu$ )
  M  $\leftarrow$  append-to-list(M, m)
end while
return X, M

```

representation of such a system is a *hyper-network* in which a link may connect more than two nodes. The stoichiometric matrix represents a hyper-network where the columns are links and the rows are nodes. The links are directed: negative values in a column represent source nodes and positive values represent target nodes. Let u be a node, and let L be a set of links that have u as their source nodes, then the target nodes of L are the out-neighbors of u . The in-neighbors are defined analogously, with u as the target node.

Note that the stoichiometric matrix derived from the linear programming formulation does not capture the reversibility of reactions (such as R_a in the example) because a reaction R_i is considered to act in reverse when its rate v_i in the linear program solution is negative. We therefore reformulate the linear program such that $\mathbf{v} \geq \mathbf{0}$ (i.e., all fluxes are positive). A reversible reaction R_i is converted to a pair of reactions R_i^+ and R_i^- ; then if \mathbf{c}_i is the column vector in \mathbf{S} corresponding to R_i , \mathbf{c}_i is replaced by two column vectors \mathbf{c}_i^+ and \mathbf{c}_i^- (corresponding to R_i^+ and R_i^- respectively) such that $\mathbf{c}_i^+ = \mathbf{c}_i$ (the forward reaction) and $\mathbf{c}_i^- = -\mathbf{c}_i$ (the reverse reaction); for example, column R_a in Fig. 1(a) is replaced by the columns R_a^+ and R_a^- in Fig. 1(a). Converting \mathbf{S} leads to the stoichiometric matrix \mathbf{S}' in Fig. 1(b). The hyper-network is shown in Fig. 2(a).

The linear program (5) is modified with the new stoichiometric matrix \mathbf{S}' and non-negative flux constraints, giving:

$$\begin{aligned} &\text{Maximize } \mu = \mathbf{c}^T \cdot \mathbf{v} \\ &\text{subject to } \mathbf{S}' \cdot \mathbf{v} + \mathbf{b} = \mathbf{0} \\ &\quad \mathbf{v} \geq \mathbf{0}. \end{aligned} \quad (6)$$

Network theory provides many tools for studying the topological properties of normal networks, whilst there are very few metrics that can be computed on hyper-networks. Thus we considered three possible network representations of the hyper-networks specified by the stoichiometric matrix \mathbf{S}' . First, a hyper-network $H(\mathcal{M}, \mathcal{R})$ can be modeled as a bipartite network $G_B(\mathcal{M} \cup \mathcal{R}, \mathcal{L})$. The nodes in the set \mathcal{M} represent the metabolites in H , whilst the nodes in the set \mathcal{R} represent reaction links in H . Conversion of the hyper-network H in Fig. 2(a) produces

$$S = \begin{matrix} & R_a & R_b & R_c \\ \begin{matrix} m_1 \\ m_2 \\ m_3 \\ m_4 \\ m_5 \\ m_6 \\ m_7 \\ m_8 \end{matrix} & \begin{pmatrix} -1 & 0 & 0 \\ -1 & 0 & 0 \\ 1 & -3 & -1 \\ 1 & 0 & -1 \\ 0 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 1 \end{pmatrix} \end{matrix} \quad \begin{matrix} & R_a^+ & R_a^- & R_b & R_c \\ \begin{matrix} m_1 \\ m_2 \\ m_3 \\ m_4 \\ m_5 \\ m_6 \\ m_7 \\ m_8 \end{matrix} & \begin{pmatrix} -1 & 1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 1 & -1 & -3 & -1 \\ 1 & -1 & 0 & -1 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{matrix}$$

(a) The stoichiometric matrix from (4). (b) The stoichiometric matrix from (6).

Fig. 1. Stoichiometric matrices of the toy problem in Section 2.1.

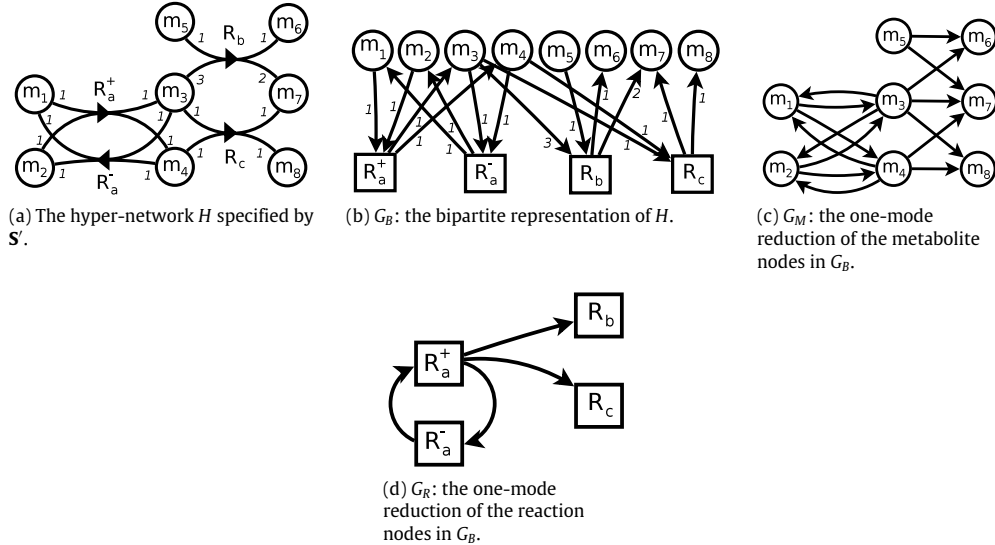


Fig. 2. The hyper-network and networks derivable from S' in (6).

the bipartite network G_B in Fig. 2(b). We refer to this network as the *metabolite-reaction network*⁴ as it contains both metabolite nodes \mathcal{M} and reaction nodes \mathcal{R} .

Although standard network theory techniques can be applied to G_B , its bipartite nature makes some metrics difficult or impossible to compute. For example, the clustering coefficient for any node in a bipartite network is 0. For this reason, we also considered one-mode reductions of G_B . An \mathcal{M} -node (\mathcal{R} -node) *one-mode reduction* $G'(\mathcal{N}, \mathcal{L}')$ of $G_B(\mathcal{M} \cup \mathcal{R}, \mathcal{L})$ is a network that contains only nodes from the set \mathcal{M} (the set \mathcal{R}) such that for each directed link $l = (n_1, n_2) \in \mathcal{L}'$ there is a node $n_3 \in \mathcal{R}$ ($n_3 \in \mathcal{M}$) such that $(n_1, n_3) \in \mathcal{L}$ and $(n_3, n_2) \in \mathcal{L}$ (note that there may be many nodes n_3 that satisfy this condition). We call the \mathcal{M} -node one-mode reduction simply the *metabolite network* G_M (shown in Fig. 2(c)) and likewise the \mathcal{R} -node one-mode reduction simply the *reaction network* G_R (illustrated in Fig. 2(d)).

Note that it is possible to represent the link weights of the hyper-network H in its bipartite representation G_B : such a mapping can be seen in Fig. 2(b). However, there is no obvious way to map these weights to G_M or G_R . For this

paper, we opted to consider only unweighted networks. Furthermore, note that when a reaction is removed from the metabolic system, the corresponding networks G_B , G_M and G_R may become disconnected. For a given network, all metrics are applied to the largest component whilst the small components are ignored.

2.2.1. Topological metrics












For every step of each trial, a number of topological metrics were computed for each of the three network representations (where possible). Since $G_B = G_B(\mathcal{M} \cup \mathcal{R}, \mathcal{L})$ contains two types of nodes, the metrics are applied separately to its reaction nodes \mathcal{R} and metabolite nodes \mathcal{M} , giving two sets of results.

The metrics employed are listed in Table 1. These metrics divide into two groups: those that associate a value $c(G)$ with a network G and those that associate values $\{c(n_1), c(n_2), \dots, c(n_N)\}$ with the nodes $\mathcal{N} = \{n_1, n_2, \dots, n_N\}$ of G . In order to compare this latter group of metrics to growth values, the node values (for a given metric c) have to be reduced to a single value $c^*(G) = f(c(n_1), c(n_2), \dots, c(n_N))$ (where f is function of N arguments that produces a single real value $c^*(G) \in \mathbb{R}$). A simple choice is to let f compute the minimum, mean or maximum values of $\{c(n_1), c(n_2), \dots, c(n_N)\}$ (thereby yielding three metrics). This is the approach that we took.

⁴ This representation is the Petri-net representation [14,4] of the metabolic system.

Table 1

A list of the various network metrics that were calculated on the networks G_B , G_M and G_R . Metrics that are calculated for a network as a whole are marked “scalar” whilst those that are calculated for every node are marked “node”.

<i>Newman's assortativity coefficient [15]</i>	Scalar
How likely is it for nodes with similar degrees to be connected to each other. Calculated for the out-degrees, in-degrees and undirected degrees of nodes, it is computed as $r = \sum_{ij} d_i d_j (b_{ij} - a_i a_j) / \sigma_a^2$, where a_i is the distribution of the degree of node n_i without the link $\{i, j\}$, b_{ij} is the joint probability distribution of the degrees between n_i and n_j without the link $\{i, j\}$ and σ_a^2 is the variance of degrees.	
<i>Transitivity</i>	Scalar
The number of triangles in the network divided by the maximum possible number of triangles in the network. This is computed on undirected versions of the networks.	
<i>Clustering coefficient</i>	Node
For a node n , the number of links spanning n 's neighbors divided by the maximum possible number of links that can span n 's neighbors. The mean clustering coefficient is equal to the transitivity.	
<i>Reciprocity</i>	Scalar
The ratio of reciprocal pairs to all possible reciprocal pairs. A pair of nodes n_1 and n_2 is reciprocal if there are bi-directed links (n_1, n_2) and (n_2, n_1) .	
<i>Betweenness centrality</i>	Node
In a network G , the betweenness centrality $C_B(n)$ of a node n is computed as the fraction of shortest paths in G that include n : $C_B(n) = \sum_{s \neq n \neq t \in \mathcal{N}} \sigma_{s,t}(n) / \sigma_{s,t}$, where \mathcal{N} is the node-set of G , $\sigma_{s,t}(n)$ is the number of shortest paths that include n and $\sigma_{s,t}$ is the total number of shortest paths between the nodes s and t .	
<i>Eigenvector centrality and Largest eigenvalue of the adjacency matrix</i>	Node
For a graph G , the largest eigenvalue is that of the corresponding adjacency matrix \mathbf{A} . The eigenvector centrality of a node n_i is defined as the i th component of the eigenvector corresponding to that eigenvalue.	
<i>In- and out-degrees</i>	Node
The out-degree d_i^{out} of a node is the number of links leaving a node, i.e., $d_i^{\text{out}} = \sum_{(n_i, n_j) \in L(G)} 1$. Likewise, the in-degree d_i^{in} is the number of links entering a node, i.e., $d_i^{\text{in}} = \sum_{(n_j, n_i) \in L(G)} 1$.	
<i>Average in- and out-degrees of incoming and outgoing neighbors</i>	Node
For a network G , the average out-degree of outgoing neighbors of a node n_i is $\sum_{(n_i, n_j) \in L(G)} d_j^{\text{out}} / d_i$ whilst the average in-degree of incoming neighbors is $\sum_{(n_j, n_i) \in L(G)} d_j^{\text{in}} / d_i$ where d_i^{out} is the out-degree of n_i and d_i^{in} is the in-degree of n_i .	
<i>Coreness</i>	Node
A k -core is a subset of nodes in which each node has a degree of at least k . A node has a coreness value of c if it is in a c -core but not in a $c + 1$ -core.	
<i>Dice similarity</i>	Node
If the neighbors of two nodes are the sets X and Y , the Dice similarity of the nodes is $2 X \cap Y / (X + Y)$, i.e., a measure of how similar their neighbor sets are. Since this metric is defined for pairs of nodes, a vector of metrics is associated with each node. We compute the Dice similarity for all outgoing neighbors, all incoming neighbors and also for the combination of these.	
<i>Reciprocal node hop-count</i>	Node
The hop-count between a pair of nodes is equal to the number of links on a shortest path between them. For each node there is a vector of hop-counts to all other nodes, reduced to a single value by taking the mean. Because the networks are directed, there are nodes which are unreachable from other nodes and are thus at an infinite distance. We therefore used reciprocal hop-count values, converting infinite distances to zero distances.	

Some metrics associate vectors of values with each node; thus, if the metric \mathbf{c} associates a vector with a node, the result will be a set of vectors $\{\mathbf{c}(n_1), \mathbf{c}(n_2), \dots, \mathbf{c}(n_N)\}$. The hop-count is such a metric, since it associates a vector of hop-count values $\mathbf{c}(n)$ with a node n containing the hop-counts to all other nodes in the network. We took the approach of first reducing the vectors to scalars—thus we converted $\{\mathbf{c}(n_1), \mathbf{c}(n_2), \dots, \mathbf{c}(n_N)\}$ to

$\{c'(n_1), c'(n_2), \dots, c'(n_N)\}$ where c' is a function that reduces vectors to real values. As above, we performed the reductions by computing the minima, maxima and means of the vectors. Once this initial reduction is performed, we can proceed as before (by reducing the sets of node values to single values). Note that this double reduction scheme can lead to confusing metric names. To take the example of the hop-count again, we could proceed by first

computing the means of the hop-count vectors associated with each node and then we could compute the minimum over these mean values. In this case, we would refer to the minimum of the mean hop-count, or in the naming convention used in the results section, “mean hop-count ∇ ”. Likewise, we refer to the mean of the mean hop-count as “mean hop-count \square ” and the maximum of the mean hop-count as “mean hop-count Δ ”.

In our experiments, many reactions have zero reaction rates (as predicted by the flux balance linear program) in all trials. These reactions contribute links and nodes to the network representations whilst their removal cannot influence growth. We excluded these reactions when constructing $G_B = G_B(\mathcal{M} \cup \mathcal{R}, \mathcal{L})$ by letting \mathcal{R} be the set of all reactions that have non-zero reaction rates in at least one step of one trial and \mathcal{M} the metabolites that interact with the reactions in \mathcal{R} . Note that this is only a global pre-processing step; in each individual trial, reactions are randomly chosen without regard to whether they are active at that time or not.

2.3. Relating growth and topology

For each trial (i.e., sequence of reaction removals) we compute a sequence of growth values (computed from the linear program discussed in Section 2.1) and three sequences of networks, one for each representation. For each network, a set of topological metrics is calculated. This allows us to relate growth to topology.

An obvious first choice for calculating the relationship is, for each individual trial, to compute correlation coefficients ρ between the growth sequence and each of the sequences of topological metrics. However, apparent correlations found by this method may simply be side-effects of the network size decreasing as we remove reactions. We can reduce the impact of this incidental correlation by binning the steps from all of the trials: trial-step pairs whose corresponding networks have similar numbers of nodes and links are placed into the same bin. This process is illustrated in Fig. 3: here one sees network sequences from two trials placed into bins (the bin width here is 1 for both nodes and links). In our experiments, we used a bin width of 2 nodes \times 4 links—i.e., in a bin, node counts can differ by 1 and link counts by 3.

Since a bin contains numerous steps, it is possible to correlate growth with any of the topological metrics. We used the Pearson correlation coefficient to compute, for a given topological metric, a correlation value ρ_i for every bin i . An example of bin correlations is shown in Table 2 (here binning is only performed using link counts, with a bin width of 4 links). The per-bin results for each metric were then averaged, weighted by the number n_i of items in each bin: $\bar{\rho} = (\sum_i n_i \rho_i) / (\sum_i n_i) = 0.27$ in our example. For each topological metric, this yields one value $\bar{\rho}$ indicating the strength of its binned correlation with growth.

For all of the metrics that we studied, there were one or more bins for which correlations could not be computed, since the growth and/or metric values in the bin were constant. In this case, the Pearson correlation coefficient is

not defined. These bins were excluded from the calculation of $\bar{\rho}$. We also required correlations to be:

- *reliable*, i.e., calculated on a sufficient number of data-points, by demanding that at least 90% of all steps fall in bins on which correlations are defined; and
- *consistent*, by requiring that at least 90% of all steps fall in bins whose correlations have the same sign.

Metrics that did not pass this test were not considered.

2.4. Experimental setup

We used the genome-scale metabolic dataset which is available from the UCSD Systems Biology Research Group website [17]. The website provides a minimal aerobic growth environment which was used for our experiments. In this experiment,

- the rate of the ATP maintenance reaction (ATPM) is $1 \mu\text{mol gDW}^{-1} \text{h}$ whilst the acetyl-CoA hydrolase (ACOAH) and the glutamate synthase for NADH (GLUSx) reactions are disabled;
- the reaction rates of reactions that transport O_2 , NH_4^+ , SO_4^{2-} , P_i , H_2O , K , Na and CO_2 are unconstrained.

3. Results and discussion

3.1. Metrics correlate with network size

We initially performed one thousand *in silico* reaction removal trials and for each trial computed the Pearson correlation ρ between the growth values of the trial and the metrics in Table 1 as computed on G_B , G_M and G_R (where applicable). The average metric correlations over 200 random trials for G_M are shown in Fig. 4 (here, we have only aggregated node-wise metrics using the mean, as described in Section 2.2.1). Many metrics stand out as strongly correlated.

We found that most of these correlations are due to the reduction of the number of nodes and/or links in G_B , G_M and G_R associated with each step in a destruction trial of a metabolic system. This growth-size relationship confounds the search for metrics that correlate with growth, since any apparent correlation ρ may be due solely to the correlation between the metric and the number of nodes/links in the network.

Removal of this effect by metric normalization is non-trivial, since the relationship between a given metric and the number of nodes/links in a network is, in general, nonlinear. Furthermore, any technique that reduces this effect, must use topological information; but then this information itself is affected by the changing topology. We therefore devised a “binning” procedure to calculate alternative correlation measures $\bar{\rho}$ in which this effect is reduced (as described in Section 2.3). In the remainder, all results reported employ this binned correlation measure.

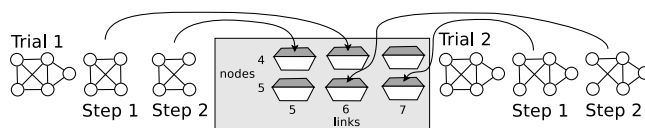


Fig. 3. An example of binning for two steps of two trials.

Table 2

Bins showing Pearson correlations ρ between growth and an unspecified network metric.

Bin number	1	2	3	4	5
# links	2685–2688	2689–2692	2693–2696	2697–2700	2701–2704
Correlation ρ_i	0.342	0.286	0.322	0.236	0.172
# items in bin n_i	889	935	907	959	936

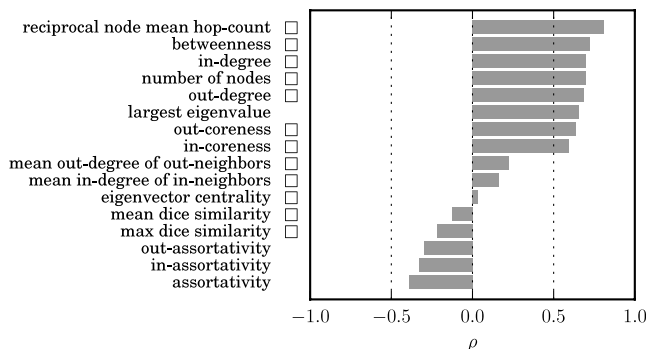


Fig. 4. Network metrics are correlated with network size. This gives the appearance of strong correlations between growth and metrics. These metrics were all calculated for G_M . The symbol \square indicates that node values were reduced to single values by computing their means as discussed in Section 2.2.1.

3.2. Topology is weakly correlated with function

Next, we calculated correlations $\bar{\rho}$ (using the binning procedure) between growth and each metric. The results for $G_B = G_B(\mathcal{M} \cup \mathcal{R}, \mathcal{L})$ are shown in Fig. 5 (recall that there are two sets of results for G_B : one for the metabolite nodes \mathcal{M} and one for the reaction nodes \mathcal{R}) whilst the correlations for G_M are shown in Fig. 6. There are no correlations for G_R that satisfy the reliability and consistency requirements described in Section 2.3. First we discuss these results from a purely topological perspective and then we interpret the biological aspects.

The results show that most metrics do not correlate well with growth. An obvious first explanation for this lack of correlation is that it is possible to remove a reaction without affecting growth (since the reaction may be part of a bypass that is not used when the cell is functioning normally). However, at a deeper level, the low correlations may be explained by the indirect relationship between the flux balance analysis framework (which measures function) and the network (on which topological metrics are measured). In flux balance analysis, growth is the objective function of a linear program in terms of metabolic fluxes, whilst the topologies of the metabolic networks are only functions of the stoichiometric matrix. While the objective function may be changed (perhaps to study a scenario other than growth maximization) the topology remains unchanged. Thus, correlations between the objective function and topological metrics depend to some extent on the objective function.

3.3. The metabolite-reaction network G_B is the best representation

Here we investigate some of the $\bar{\rho}$ correlations observed in Figs. 5 and 6. We generally limit our discussion to metrics for which $|\bar{\rho}| \geq 0.2$.

Metabolite-reaction network G_B . As discussed in Section 2.2.1, correlations for the metabolite nodes \mathcal{M} and the reaction nodes \mathcal{R} were computed separately. First, the results for the metabolite nodes are considered, followed by the reaction node results.

Metabolite nodes \mathcal{M} . There are a number of relatively strong correlations for nodes in \mathcal{M} , mostly falling into two groups:

- For both the metabolite nodes, so-called “hub” nodes provide shortcuts through which shortest paths are routed. Removal of a reaction node that interacts with a hub node may therefore remove a shortcut through which some shortest paths are routed. Thus, the mean reciprocal hop-count is decreased (and the mean hop-count is increased). In the remainder of the paper, all correlations associated with hub nodes are colored light gray.
- So-called “loner” nodes are nodes with low incoming and/or outgoing degrees. Some of these nodes are on important pathways and can cause growth to decrease when they are no longer produced (i.e., when their incoming links are removed) or consumed (i.e., when their outgoing links are removed) by any reactions. As a result, they are often implicated in correlations

using the minimum function (those indicated by ∇). Correlations associated with loner nodes are colored dark gray.

Reaction nodes \mathcal{R} . Only a few reliable, consistent correlations were found for the reaction nodes \mathcal{R} in G_B . Of these, the mean reciprocal hop-count is the only reaction node metric that stands out, owing its presence to the metabolite hubs which provide shortcuts between a large number of reaction nodes.

Metabolite network G_M . The correlation results for G_M are shown in Fig. 6. G_M has more high-degree nodes than G_B and these are at least partially responsible for the strongest correlations. As with its progenitor G_B , the hub nodes in G_M provide shortcuts and thus provide the basis for the strong mean reciprocal node mean hop-count correlations.

The out-degree of out-neighbors correlations are due either to hub nodes themselves or nodes attached to the hub nodes (in particular hydrogen). The Dice similarity correlations are also the result of hub nodes – for example, the maximum mean Dice similarity is the result of a certain node (Asparagine) which is connected to a number of hub nodes; therefore it shares neighbors with many other nodes.

There are no apparent loner-node related correlations amongst the top correlations ($|\bar{\rho}| \geq 0.2$). However, the three correlations immediately following the top correlations (the minimum in-degree, hop-count and out-degree) are due to loner nodes.

Reaction network G_R . The reaction network G_R yielded apparently no reliable, consistent correlations. As G_R is much denser than either G_B or G_M , each reaction removal forces a node to be removed from G_R . This leads to larger changes in G_R relative to the other networks; a property that may in part explain the difficulty of finding a connection between topology and growth in this representation.

Metabolite relationships hold the key to understanding the topology of metabolic systems. The most interesting results are associated with the metabolite nodes. As mentioned in Section 2.1, there are more reactions than metabolites in metabolic systems. A reaction ties together a small number of metabolites while there are metabolites that are involved in many reactions. In other words, metabolites bind the network at a high level and are responsible for global connectivity. This leads us to conclude that the metabolite-reaction network G_B and the metabolite network G_M are the most useful representations for our purposes. The reaction network G_R is less interesting, as no reliable, consistent correlations were found. Reactions are, of course, essential to the metabolic system, but metabolites tell the most interesting story.

Because G_B is the most accurate representation of the metabolic system and because of its strong correlations, we consider the metabolite nodes of G_B to be the most promising entities for studying metabolism.

3.4. The strongest correlations point to currency metabolites

Many of the hub metabolite nodes implicated in the previous section correspond to so-called *currency metabolites*. We know from biology that currency metabolites play a crucial role in metabolism: they are energy carriers or co-factors that are used by many reactions. Holme and

Huss [5] found the currency metabolites of *S. cerevisiae* to be H, H₂O, ATP, ADP, AMP, NAD, NADH, NADP, NADPH, CoA, CO₂, O₂, P_i, PP_i and NH₄⁺ (for this set they used the undirected version of G_M with information taken from the BiGG database).

To validate the role of these metabolites, we repeated our experiments with currency metabolites removed from G_B , G_M and G_R . Note that the metabolites were not removed from the flux balance linear program, as this would lead to incorrect chemical equations and it would change the computed growth. The five most significant $\bar{\rho}$ correlations for each of G_B , G_M and G_R are shown in Fig. 7 (note that the reaction nodes in G_B were omitted, as all $\bar{\rho}$ correlations for these nodes fell below 0.2). Correlations that are neither the direct result of hub nodes nor loner nodes are shown as medium gray bars in the figure.

There are a number of interesting differences in the correlations brought about by currency metabolite removal:

- Most of the strong correlations due to hub nodes have been strongly reduced. The exception is the mean reciprocal mean hop-count correlation in G_M which remains approximately the same, in contrast with the correlation of the same metric in G_B . This hints at second-order network structure (as opposed to first-order hub structure) that is important in routing shortest paths.
- Removal of hub nodes removes shortcuts that route many shortest paths. The shortest paths are therefore more “spread out” through the metabolite network. This leads to a relative increase in node betweenness values and a concomitant increasing influence of arbitrary nodes on the average betweenness. Although this effect is most pronounced for G_M , it is also present for the metabolite nodes of G_B .
- As G_R is less dense due to currency metabolite removal, a number of reliable, consistently correlated metrics could now be found. The majority of reaction nodes have degrees below the mean degree, so that a reaction removal is likely to increase the mean in- and out-degrees. Likewise the minimum and mean Dice similarities are likely to be increased, since the low-degree nodes have low Dice similarities. The correlations are not obviously due to hub nodes or loner nodes.

4. Conclusions

The goal of this study was to determine whether topology and robustness of biological systems are related. To this end, we generated a number of reaction removal sequences or *trials*, each of which resulted in the cessation of growth of our metabolic system. Each step in a trial provided a snapshot of the metabolic system from which growth could be computed as well as topological metrics of the metabolite-reaction network G_B , the metabolite network G_M and the reaction network G_R . This allowed us to calculate a measure of correlation between growth and each of the metrics. In this section, we will summarize some of our findings.

Unambiguously linking robustness to topology is difficult. The term “robustness” is meaningless without context. Since the context of an organism constitutes all its interactions

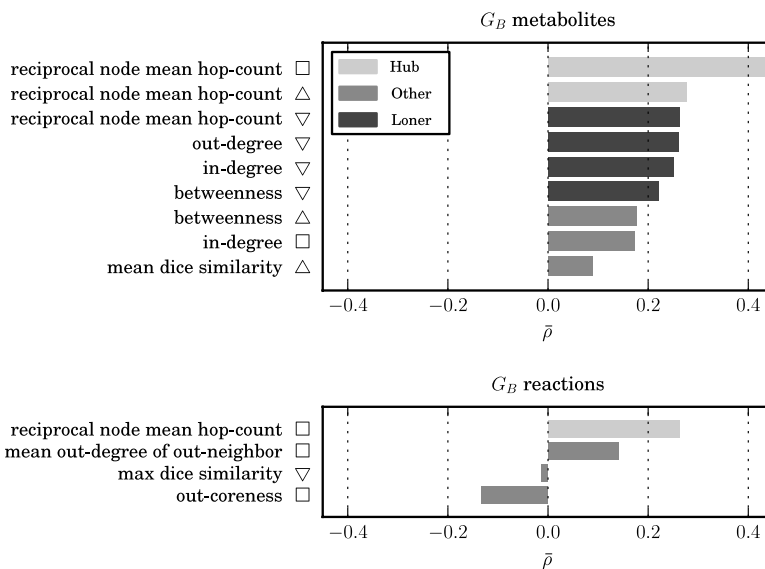


Fig. 5. $\bar{\rho}$ measures between growth and topological metrics for G_B . The symbols ∇ , \square and \triangle indicate that node values were reduced to single values by computing their minima, means and maxima respectively, as discussed in Section 2.2.1. As the legend shows, the light gray bars correspond to hub nodes, the dark gray bars correspond to loner nodes and the medium gray bars correspond to metrics that were either not interpreted or that do not fit the hub/loner distinction.

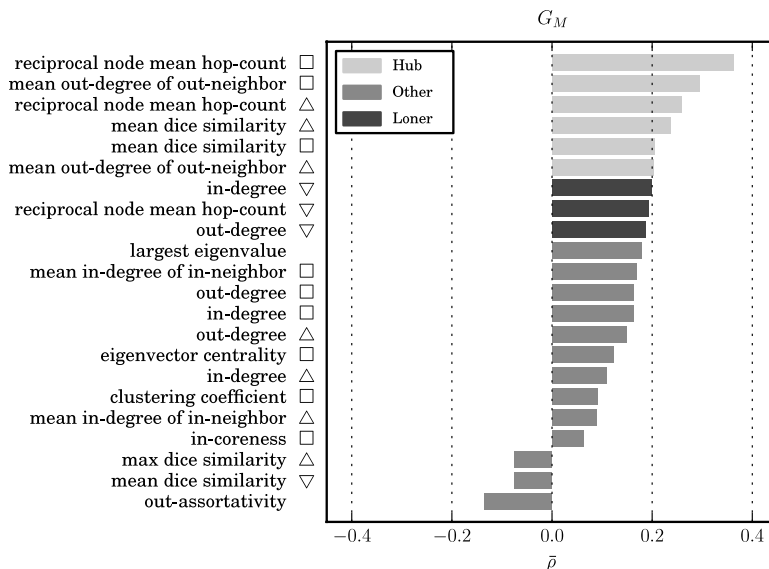


Fig. 6. $\bar{\rho}$ measures between growth and topological metrics for G_M . The labels and colors are explained in Fig. 5.

with its environment, a precise definition may forever elude us. However, every organism engages in a (small) number of vital functions that dominate its struggle for survival. By studying only these functions and their degradation in the face of perturbations, we may discover some of the principles that help organisms to achieve their resilience. However, an unambiguous connection between such functions/perturbations and the topology of the underlying biochemical network is usually hard to define.

As we studied microbial metabolism, we focused on the one function at which the cell must be successful before all else: biomass production, or growth. While this is a simple representation of cellular activity, it has the

advantage of being based on a well-studied theoretical model of metabolism, flux balance analysis, that can easily be modified to work with a perturbation model of reaction removal. Still, although we were able to directly link metabolic networks, functions and perturbations, finding correlations between robustness and topology proved not to be trivial.

There is no obvious way to reduce a metabolic system to a network. This is a consequence of the correspondence between metabolic systems and hyper-networks. Analyzing hyper-networks directly is the ideal approach but these general structures have resisted the theoretical analysis that has produced the useful tools of (classical) network theory. Therefore, conversion is an analytical necessity. We

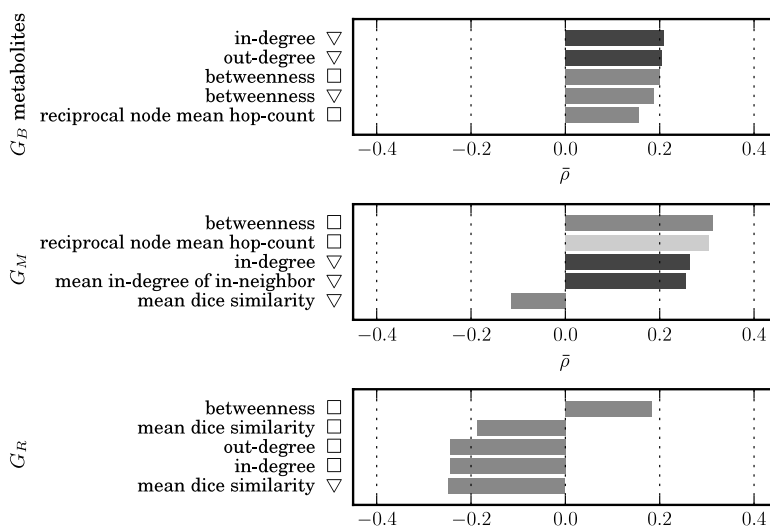


Fig. 7. The most significant $\bar{\rho}$ correlations between growth and topological metrics for networks lacking currency metabolites. The labels and colors are explained in Fig. 5.

described three ways of converting a hyper-network to a network: the metabolite-reaction network G_B , the metabolite network G_M and the reaction network G_R . The multiplicity of representations is a well-known problem that Holme and Huss [5] investigated by matching graph theoretical properties of the three network representations to biological data in order to discover the network representation that “best” captures biological knowledge. We found the correlations were strongest for the metabolite nodes in G_B and for G_M . These findings suggest that metabolite nodes are most important for studying the structure of a metabolic system. In line with this, Holme and Huss found G_M to be the most favorable representation, although we favor G_B since it maintains most of the original metabolic information.

Topology correlates weakly with growth. Many of the topological metrics we calculated did not correlate with growth. We classified those that did correlate into two groups: those caused by hub metabolite nodes and those caused by loner metabolite nodes. They point to the importance of (a) global connectivity (by hub nodes that tie the network together by connecting many reaction nodes); and (b) local connectivity (by metabolites that are produced and/or consumed by few reactions). The role of hub nodes was verified in an experiment where we removed currency metabolites, which led to a large shift in metrics correlated to growth.

5. Outlook

In this work, we studied the relationship between topology and growth. Using our framework as a starting point, one can investigate whether other functions of the metabolic network are related to topology or whether topology plays a role in other biological networks (e.g., gene regulation or protein interaction networks).

Our approach can be refined in a number of ways. On the one hand, flux balance analysis can be done with more sophisticated methods, such as MOMA [19] (Minimization

Of Metabolic Adjustment) and ROOM [20] (Regulatory On/Off Minimization), both of which were designed to better approximate the metabolic behavior of systems from which reactions have been removed. With sufficient enzyme kinetic parameters, one could even attempt to solve the nonlinear differential equations (1). On the other hand, we could remove genes rather than reactions, more in line with the biological perturbations we intend to model. In this case, removing a gene may lead to the removal of multiple reactions, or alternatively a reaction may only be removed if all genes coding for isoenzymes are lost. However, such refinements to the model are unlikely to paint a very different picture since, if there were an effect, a first-order approach (such as ours) would pick up some correlation if it were there.

This work was an exploration of how topology is related to robustness. Although whether topology confers robustness or vice versa remains an open question, a change of perspective points to a number of paths for future investigation.

In a more local approach, one could isolate a small, fixed sub-network such as the citric acid cycle (a central part of metabolism in many organisms). Then our framework could be applied almost unchanged. Metrics would still be computed for entire networks but only the values corresponding to the sub-network under consideration would be compared with growth.

On a more global level, one could consider a number of species related by evolution. The species cannot necessarily be directly compared to each other, since they are specialized for different environments (and thus different contexts). But these differences in specialization enable us to study the connection between robustness and topology, since differences in metabolism are the results of specialization and these differences will be reflected in metabolic networks.

An interesting related approach is the study of the metabolic networks of gene knockout mutants of a given organism. This is essentially our approach with *in silico*

knockouts replaced by *in vivo* knockouts, giving actual flux measurements which are more reliable than fluxes computed by flux balance analysis approaches.

References

- [1] A.L. Barabasi, R. Albert, Emergence of scaling in random networks, *Science* 286 (1999) 509–512.
- [2] D. Bu, Y. Zhao, L. Cai, H. Xue, X. Zhu, H. Lu, J. Zhang, S. Sun, L. Ling, N. Zhang, G. Li, R. Chen, Topological structure analysis of the protein–protein interaction network in budding yeast, *Nucleic Acids Research* 31 (2003) 2443–2450.
- [3] N. Guelzim, S. Bottani, P. Bourguin, F. Kepes, Topological and causal structure of the yeast transcriptional regulatory network, *Nature Genetics* 31 (2002) 60–63.
- [4] S. Hardy, P.N. Robillard, Modeling and simulation of molecular biology systems using petri nets: modeling goals of various approaches, *Journal of Bioinformatics and Computational Biology* 2 (2004) 595–613.
- [5] P. Holme, M. Huss, Substance graphs are optimal simple-graph representations of metabolism, *Chinese Science Bulletin* 55 (2010) 3161–3168. For a list of the currency metabolites, see the arXiv version at <http://arxiv.org/abs/0806.2763>.
- [6] P.J. Ingram, M.P. Stumpf, J. Stark, Network motifs: structure does not determine function, *BMC Genomics* 7 (2006) 108.
- [7] H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, A.L. Barabási, The large-scale organization of metabolic networks, *Nature* 407 (2000) 651–654.
- [8] W.K.K. Kim, E.M. Marcotte, Age-dependent evolution of the yeast protein interaction network suggests a limited role of gene duplication and divergence, *PLoS Computational Biology* 4 (2008) e1000232.
- [9] H. Kitano, Biological robustness, *Nature Reviews Genetics* 5 (2004) 826–837.
- [10] H. Kitano, Computational systems biology, *Nature* 420 (2002) 206–210.
- [11] G. Lima-Mendez, J. Helden, The powerful law of the power law and other myths in network biology, *Molecular BioSystems* 5 (2009) 1482–1493.
- [12] S. Maslov, K. Sneppen, Specificity and stability in topology of protein networks, *Science* 296 (2002) 910–913.
- [13] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, U. Alon, Network motifs: simple building blocks of complex networks, *Science* 298 (2002) 824–827.
- [14] T. Murata, Petri nets: properties, analysis and applications, *Proceedings of the IEEE* 77 (1989) 541–580.
- [15] M.E.J. Newman, Assortative mixing in networks, *Physical Review Letters* 89 (2002) 208701.
- [16] J.D. Orth, I. Thiele, B.O. Palsson, What is flux balance analysis? *Nature Biotechnology* 28 (2010) 245–248.
- [17] B.O. Palsson, Organisms–systems biology research group, 2009. <http://systemsbiology.ucsd.edu>.
- [18] R.J. Prill, P.A. Iglesias, A. Levchenko, Dynamic properties of network motifs contribute to biological network organization, *PLoS Biology* 3 (2005) e343.
- [19] D. Segrè, D. Vitkup, G.M. Church, Analysis of optimality in natural and perturbed metabolic networks, *Proceedings of the National Academy of Sciences of the United States of America* 99 (2002) 15112–15117.
- [20] T. Shlomi, O. Berkman, E. Ruppin, Regulatory on/off minimization of metabolic flux changes after genetic perturbations, *Proceedings of the National Academy of Sciences of the United States of America* 102 (2005) 7695–7700.
- [21] J. Stelling, U. Sauer, Z. Szallasi, F. Doyle 3rd, J. Doyle, Robustness of cellular functions, *Cell* 118 (2004) 675–685.



Wynand Winterbach received his M.Sc. degree in Applied Mathematics in 2005, from Stellenbosch University, South Africa. In 2009, he joined Delft University of Technology as a Ph.D. student in the Faculty of Electrical Engineering, Mathematics and Computer Science. Before his return to the academic world, he worked as a software developer. His research interests center around the connection between the structure of biological networks and their robustness. In particular, he is interested in how insights gleaned from

these networks may be applied to improve the robustness of human-made networks, such as the internet.



Huijuan Wang was born in Harbin, China. She graduated and received her M.Sc. degree (cum laude) in Electrical Engineering at the Delft University of Technology, the Netherlands, in the year 2005. She obtained her Ph.D. degree (cum laude) in 2009 at the same faculty. She is currently an assistant professor in the Network Architecture and Services (NAS) Group at Delft University of Technology.



Marcel Reinders received his M.Sc. degree in Applied Physics and a Ph.D. degree in Electrical Engineering from Delft University of Technology, The Netherlands, in 1990 and 1995, respectively. In 2004 he became a Professor in Bioinformatics within the Mediamatics Department of the Faculty of Electrical Engineering, Mathematics and Computer Science at the Delft University of Technology. The background of Prof. Reinders is within pattern recognition. Besides studying fundamental issues, he applies pattern recognition techniques to the areas of bioinformatics, computer vision and context-aware recommender systems. His special interest goes toward understanding complex systems (such as biological systems) that are severely under-sampled.



Piet Van Mieghem is professor at the Delft University of Technology with a chair in telecommunication networks and chairman of the section Network Architectures and Services (NAS) since 1998. His main research interests lie in the modeling and analysis of complex networks (such as biological, brain, social, infrastructural, etc. networks) and in new Internet-like architectures and algorithms for future communication networks.

Professor Van Mieghem received a Master and Ph. D. degree in Electrical Engineering from the K.U.Leuven (Belgium) in 1987 and 1991, respectively. Before joining Delft, he worked at the Interuniversity Micro Electronic Center (IMEC) from 1987 to 1991. During 1993–1998, he was a member of the Alcatel Corporate Research Center in Antwerp where he was engaged in performance analysis of ATM systems and in network architectural concepts of both ATM networks (PNNI) and the Internet.

He was a visiting scientist at MIT (Department of Electrical Engineering, 1992–1993) and a visiting professor at UCLA (Department of Electrical Engineering, 2005) and at Cornell University (Center of Applied Mathematics, 2009).

He was member of the editorial board of the journal *Computer Networks* from 2005 to 2006. Currently, he serves on the editorial board of the *IEEE/ACM Transactions on Networking* and *Computer Communications*.

He is the author of three books, *Data Communications Networking*, *Performance Analysis of Computer Networks and Systems* and *Graph Spectra for Complex Networks*.



Dick de Ridder received his M.Sc. degree in Computer Science in 1996 and a Ph.D. degree in Applied Physics in 2001, from Delft University of Technology, The Netherlands. In 2005 he became assistant professor in bioinformatics at the Faculty of Electrical Engineering, Mathematics and Computer Science of Delft University of Technology. In the past, he has worked on pattern recognition, neural networks and image processing. Currently, his interest is in integrating various high-throughput measurements and prior knowledge to model the living cell. Specifically, he studies the problem how to integrate data obtained from various measurement devices and databases with wildly varying levels of coverage, reliability, noise etc.