

Model validation of simple-graph representations of metabolism

Petter Holme^{1,2,*}

¹*Department of Physics, Umeå University, 901 87 Umeå, Sweden*

²*Computational Biology, Royal Institute of Technology, 100 44 Stockholm, Sweden*

The large-scale properties of chemical reaction systems, such as metabolism, can be studied with graph-based methods. To do this, one needs to reduce the information, lists of chemical reactions, available in databases. Even for the simplest type of graph representation, this reduction can be done in several ways. We investigate different simple network representations by testing how well they encode information about one biologically important network structure—network modularity (the propensity for edges to be clustered into dense groups that are sparsely connected between each other). To achieve this goal, we design a model of reaction systems where network modularity can be controlled and measure how well the reduction to simple graphs captures the modular structure of the model reaction system. We find that the network types that best capture the modular structure of the reaction system are substrate–product networks (where substrates are linked to products of a reaction) and substance networks (with edges between all substances participating in a reaction). Furthermore, we argue that the proposed model for reaction systems with tunable clustering is a general framework for studies of how reaction systems are affected by modularity. To this end, we investigate statistical properties of the model and find, among other things, that it recreates correlations between degree and mass of the molecules.

Keywords: chemical networks; complex networks; chemical reaction systems; statistical graph methods

1. INTRODUCTION

Metabolism, the set of chemical processes sustaining life in an organism, is a well-studied example of a chemical reaction system. Such systems are fundamental structures on many scales in both living organisms and the rest of the Universe: from the reactions in planetary atmospheres to the geochemical processes under the surface of our planet, and the already mentioned metabolism. One can even include nuclear reactions, occurring in stars and planetary interiors, in the same framework (in this paper, however, we use chemical terminology). Chemical reaction systems can be modelled and analysed on different levels. On a detailed level, one can study the dynamics of the system with differential equations. Such an approach is fruitful for modelling a relatively independent subsystem, a *module* (Del Vecchio *et al.* 2008), such as the citric acid cycle. In a simple model at this level of description using mass-action kinetics, reactions are described by concentrations of the reactants, catalysts and reaction coefficients. The number of parameters (the reaction coefficients) scales as the number of reactions. For real systems, many of these parameters are unknown. (In more elaborate models, including

temperature dependence, the situation gets even more intricate.) The complexity of modelling a large metabolic system in such a framework is staggering. The approach we take disregards all the reaction coefficients and reduces the system to a graph. Such a reduction can be done in several ways. Typically, in more elaborate graph representations (including directed edges, separating reactions, enzymes and substances, etc.), one can encode more information from the original system, but few general graph methods apply, so one would have to construct new ones. For simpler graph representations, more information is lost in the reduction from the reaction system to the graph, but a vast number of analysis methods are applicable. In this work, we choose the second approach and study a very simple class of graphs, conveniently termed *simple graphs*—unweighted, undirected graphs without multiple edges or self-edges. Even in such a framework, one can construct graphs from reaction systems in many ways. In figure 1, we define four types of simple graphs: *substrate–product graphs* connecting the substances reacting (the substrates) with the products of the reaction; *substrate–substrate graphs* linking molecules that can react with each other, or are products of the same reaction; *substance graphs* connecting all substances participating in a reaction; and *reaction graphs* where the vertices (nodes) are reactions and an edge represents a pair of reactions that have a common

*Address for correspondence: Department of Physics, Umeå University, 901 87 Umeå, Sweden (petter.holme@physics.umu.se).

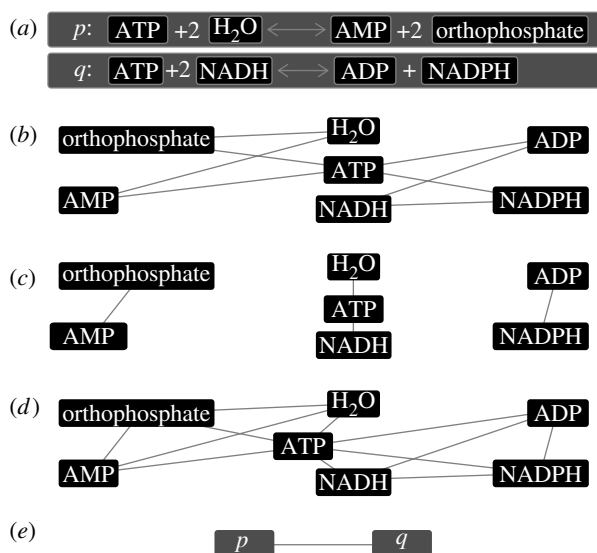


Figure 1. An illustration of different network representations derived from the two hypothetical reactions shown in (a). (b) Substrate–product network, (c) a substrate–substrate network, (d) a substance network (including both the substrate–product and substrate–substrate type edges) and (e) a reaction network where the vertices are reactions connected if they have a substance in common.

substance. These different graph representations accentuate different aspects of the reaction system. For different questions about the system, one may need different graph representations. However, for all these representations, information is lost in the conversion process. This is also true for several types of more sophisticated representations (with different types of vertices for substances and reactions, directed edges, and so on). The reason is that, with respect to the reaction dynamics, the edges are not independent. In order for mass to flow along an edge in a substrate–product graph, molecules of other vertices (other than the two making up the edge) need to be present. Of course, more complex graph types can embody more information about the original system than simple graphs can. The main reason for studying simple graphs is the vast number of analysis tools that can be applied without modification.

One of the main conclusions from graph-based studies of metabolic networks is that they have a modular structure—they can be divided into subgraphs that are more densely connected within, than between, each other (Zhao *et al.* 2006, 2007; Huss & Holme 2007; Goelzer *et al.* 2008). Assuming that the edges of the metabolic network contribute to more or less the same degree to the dynamics of the chemical system, these network modules should be subsystems (of the dynamic reaction system) with some degree of autonomy. This is close to the general idea of biological modularity—a biological module is commonly defined as a subsystem performing some specific, rather well-defined, biological function (Ihmels *et al.* 2002; Han *et al.* 2004; Kitano 2004; Del Vecchio *et al.* 2008). One can, of course, think of modules of non-biological reaction systems as well. Biological modularity is a dynamic concept that lacks a precise, general definition and the link between biological and network modularity is a vast research

question, beyond the full scope of this paper. Nevertheless, to be able to group and categorize modules in a biologically relevant way is important for our understanding of the biological organizations, with implications from more applied issues (such as drug design) to more fundamental ones (such as evolution). One reasonable requirement for choosing a network representation is that it should preserve the network modularity as much as possible. In this paper, we will use this criterion, and a model of chemical reaction systems with a tunable network modularity, to investigate the above-mentioned graph representations.

To outline this paper, we start by defining the model for modular reaction systems, discuss how the network modules can be identified in simple graphs and finally investigate how well the different graph representations can preserve the information about the modules as well as other statistical properties of the model.

2. DEFINITIONS AND MODEL

2.1. The model

In this section, we present the model for modular chemical reaction systems mathematically. To do this, we need to start by introducing a few notations. Let X be a set of N atoms (in the case of nuclear reaction systems ‘atoms’ would mean elementary particles). A molecule $\sigma \in \Sigma$ is a set of atoms and can be represented as a vector $\sigma = (m_1(\sigma), \dots, m_N(\sigma))$. $m_x(\sigma)$ is the multiplicity of atom x in the molecule. A reaction is a pair of sets of molecules (S, P) , where S is called substrates and P products. $M_S(\sigma)$ denotes the multiplicity of σ in S . A set of reactions is called a reaction system.

What are the constraints on the formation of chemical reaction systems? Arguably, the most fundamental restriction is law of mass conservation, stating that the multiplicities of all substances are the same in both S and P — $M_S(\sigma) = M_P(\sigma)$ for all reactions and all $\sigma \in \Sigma$. This means that no atoms are lost, or created, during chemical reactions. A second important constraint of chemical reactions is the stability of molecules—even if, say, an O_4 molecule could form by two oxygen molecules colliding, it is such a transient state that it is meaningless to count as a member of a reaction system. A third major factor shaping the reaction systems of the real world is the reaction dynamics itself. The probability of a hypothetical reaction $A + B \rightarrow C + D$ not only depends on the presence of A , B , C and D , but also on the free energy change of the reaction, and the presence of catalysts (or enzymes, in a biological context) or inhibitors that can lower, or raise, the activation barrier, and thus affect the probability of a reaction. Our model will obey mass conservation, but we will keep it on an abstract and general level and not map the atoms X to real atoms. Thus, the stability of molecules will not be an explicit factor in the model. In this study, reaction coefficients are not assigned to the reactions—if needed, this could be done with an auxiliary model, perhaps including correlations between the molecular mass and

reaction coefficients. Furthermore, to keep the model as simple as possible, we will not attempt to model features other than modularity, observed in reaction systems. One reason is that different kinds of reaction systems may have different network structures. Another reason is that some of these structures still are not completely characterized—the degree distribution, for example, is broad, but not strictly power law (Holme & Huss 2008; in fact, for the most common network representations, it is too broad to follow a power-law distribution). A third reason not to include any structures than network modularity is that our model can easily be extended to include this.

To sketch the algorithm, let N_A be the number of atoms in the system, N_R be the number of reactions, and g be the number of groups (potential network modules). The algorithm starts by going through all groups adding $\gamma N_R/g$ reactions within each group. After that, the remaining $(1-\gamma)N_R$ reactions are added to tie the different modules together. If the parameter γ is large (close to 1), many of the reactions will take place within a group, giving the group high network modularity. In the remainder of this section, we describe the details of these reaction addition procedures.

Consider the group i ($1 \leq i \leq g$). We assign the atoms $(i-1)N_A/g+1, \dots, iN_A/g$ to this group and, furthermore, add the N_A/g possible single-atom molecules to the set of i 's molecules Σ_i . After this, iterate the following $\gamma N_A/g$ times:

- (i) Pick two random substances s_1 and s_2 from Σ_i .
- (ii) Assign reaction multiplicities μ_1 and μ_2 to these (we choose s_1 and s_2 at random, with equal probability, from the interval $[1, \dots, \mu_{\max}]$).
- (iii) Now, construct two potential molecules p_1 and p_2 by, for each atom x , assigning a random fraction of the sum of multiplicities $\mu_1 m_x(s_1) + \mu_2 m_x(s_2)$ to p_1 and the rest to p_2 (to ensure mass conservation).
- (iv) If $p_1 \neq p_2$, $p_1, p_2 \in \Sigma_i$, both p_1 and p_2 are different from s_1 and s_2 , and the reaction $\mu_1 s_1 + \mu_2 s_2 \leftrightarrow p_1 + p_2$ does not exist in the set of reactions of group i R_i , then add this reaction to R_i . Otherwise, go to step (i).
- (v) If the algorithm has been iterated to step (iv) n_{trail} times without the condition that p_1 and p_2 should be in Σ_i fulfilled, then make another iteration from step (i) and add these molecules to Σ_i and the reaction $\mu_1 s_1 + \mu_2 s_2 \leftrightarrow p_1 + p_2$ to R_i .

The purpose of the n_{trail} iterations is to control how dense a group is. The larger the n_{trail} is, the more reactions will each molecule be involved in, i.e. the denser the group will be (as larger n_{trail} means that fewer molecules have to be added to Σ_i). The reactions are assumed to be bidirectional, but this can trivially be changed to directed reactions. To avoid the reactions being too unbalanced with respect to mass, the μ_{\max} parameter should be rather low (one could also think of assigning multiplicities to the product side, but to keep the model as simple as possible we do not do this). We also require there to be exactly two molecular species in

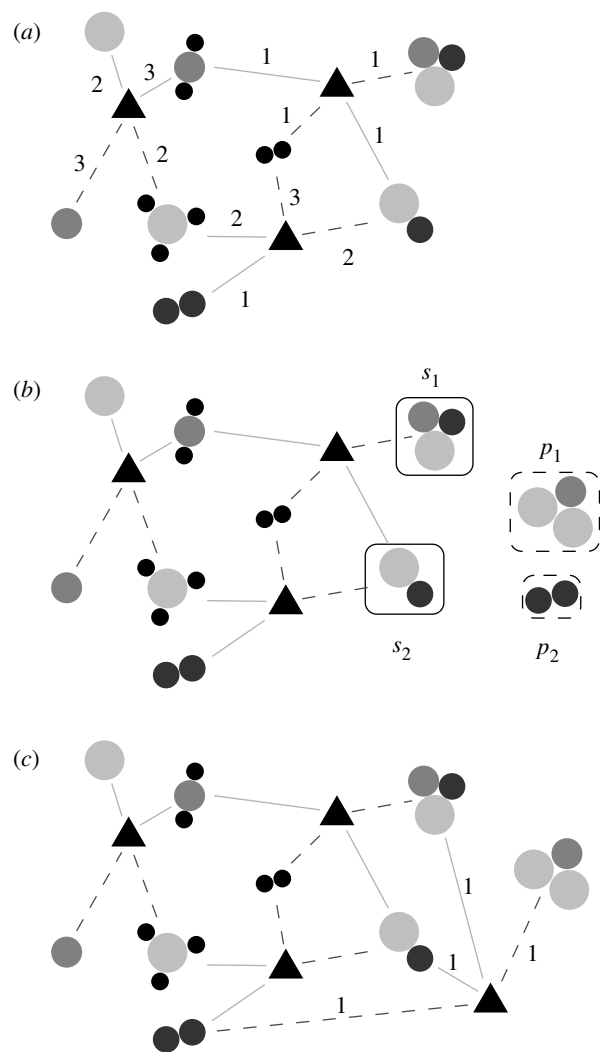


Figure 2. An illustration of the model. The circles represent atoms, grouped into molecules. The triangles represent reactions. The solid lines link substrates with reactions, while the dashed lines indicate products. Numbers on the lines indicate the multiplicities μ_i of the substances in the reactions. Starting from the configuration in (a), two molecules, s_1 and s_2 in (b), are picked at random. From these, two other molecules, p_1 and p_2 , and multiplicities, μ_1 and μ_2 , are generated in such a way that the mass is conserved in the reaction $\mu_1 s_1 + \mu_2 s_2 \rightarrow r_1 + r_2$. Since p_2 is already in the set of metabolites, the resulting reaction system, from adding this reaction, looks similar to (c). An aspect of the model not illustrated here is that the algorithm favours products already present in the substance set. In the program, we treat reactions as bidirectional, but the structure of the algorithm is identical if they are directional.

both S and P . This type of reaction is by far the most common, making up 45 per cent of all human metabolic reactions as studied in Holme & Huss (2008; the second most common order is two substrates and three products, constituting 14% of the reactions or 77 per cent of the reactions in the Earth's atmosphere; Solé & Munteanu 2004). It is trivial to extend the model to sample reactions according to some distribution, empirical or not, of their order. An illustration of these steps (except the requirement that p_1 and p_2 should be in Σ_i to break the iterations) of the algorithm can be seen in figure 2.

To add the cross-group reactions, we do as the steps above, only that we (in step (i)) select substances from the entire set of substances present, not only within one group. Furthermore, we skip step (iv) (as it is very unlikely that we will find p_1 and p_2 in Σ in this case).

All in all, our model has six parameters: the number of groups g ; the number of atoms N_A ; the number of reactions N_R ; the fraction of reactions within the group γ ; the maximum reaction multiplicity μ_{\max} ; and the number of trials to find reactions already connecting existing substances n_{trail} .

2.2. Network modularity

We briefly mention how we calculate network modularity. For a more in-depth account, see Newman (2006). Consider a partition (division) of the vertex set into groups and let e_{ij} denote the fraction of edges between groups i and j . The modularity of this partition is defined as

$$Q = \sum_i \left[e_{ii} - \left(\sum_j e_{ij} \right)^2 \right], \quad (2.1)$$

where the sum is over all groups of vertices. The term $(\sum_j e_{ij})^2$ is the expectation value of e_{ii} in a random multigraph. A prototype measure for the modularity $\hat{Q}(G)$ of a graph G is Q maximized over all partitions, of all sizes. For metabolic networks, it is standard practice to regard degrees as intrinsic properties of the vertices, and therefore measure network structure against a null model $\mathcal{G}(G)$ —random graphs with the (only) constraint that the set of degrees is the same as in G . We also do this, and take

$$\Delta(G) = \hat{Q}(G) - \langle \hat{Q}(G') \rangle_{G' \in \mathcal{G}(G)}, \quad (2.2)$$

where angular brackets denote average over $\mathcal{G}(G)$, as our measure of network modularity (Huss & Holme 2007). We use a random rewiring of the original graph to sample $\mathcal{G}(G)$ (Maslov & Sneppen 2002), and the heuristics proposed in Newman (2006) to calculate \hat{Q} . We note that it is notoriously difficult to compare modularity of networks of different sizes (Kumpula *et al.* 2007), but, in this case, when the networks come from a series from the same network being continuously reduced, this can at most shift the peak of maximum Δ marginally (Huss & Holme 2007).

We note that there are several other ways, apart from maximizing equation (2.1), of dividing networks into clusters. We mention the spectral methods originating from Pothén *et al.* (1990), and information theory approaches (e.g. Ziv *et al.* 2005; Rosvall & Bergstrom 2007), asking how the graph can best be represented as a coarse-grained graph where vertices are the clusters of the original graph in such a way that as little information as possible is lost in the process. (For even more graph clustering methods, see Newman 2006; Rosvall & Bergstrom 2007 and references therein.) The benefit of Q -maximization, as we see it, is that it is close to the notion of clusters being densely connected within and sparsely connected between each other. A cluster is identified with this algorithm as the property that it cannot be divided in any way that

would increase the Q -modularity. This gives the somewhat peculiar feature that very sparse regions of the graph can be considered a cluster even if they are fragmented. For most real-world networks such regions are small and so is this effect. One common way of validating these clustering schemes is to start from a network of a number of separated cliques (fully connected subgraphs), rewire a fraction f of the edges and measure the overlap between the detected clusters and the original cliques (Girvan & Newman 2002). Many of the proposed network-clustering methods perform well in this test, which also means that they are likely to perform quite similarly for the purposes of this work.

In metabolic networks, it is common practice to delete currency metabolites—abundant substances with a high degree that ties together modules and thereby blurs the modular network structure (Huss & Holme 2007; Holme & Huss 2008). In our model, the new vertices introduced with the inter-module reactions do blur the modular structure, but they do not have particularly high degrees. To keep the analysis of the model simple, we do not include currency-metabolite identification.

3. NUMERICAL RESULTS

3.1. An example output

In figure 3a, we display an example output of our algorithm, represented as a bipartite network where edges connect reactions (triangles) with the substances (circles or squares). The parameter γ controlling the network modularity ($\gamma=0.90$) is rather high, giving a visually clear modular structure. The additional molecules introduced during the addition of inter-module reactions are indicated by squares. Figure 3b shows the resulting substance network. The modular structure of the bipartite representation in figure 3a seems to be retained. From this substrate network, we run an algorithm detecting network modules (Newman 2006). This algorithm finds as many clusters as the original network, but misclassifies approximately 2 per cent of the vertices (the additional vertices for the inter-module reactions are not counted; figure 3c). In figure 3d, we display the output from running the modularity detection algorithm on the reaction network and assigning the identity of a module (in the set of reactions) to all vertices participating in this reaction. In metabolic databases, one metabolite can be assigned many functions; in the spirit of such multifunctionality, this kind of categorization might seem reasonable. (Note that the underlying network displayed in figure 3d is, for comparison, the same as the substrate network of figure 3b,c.)

3.2. Network modularity

The parameter γ is the model parameter intended for tuning the network modularity of the reaction system. It is, therefore, highly desirable that the relative modularity Δ should be a monotonous function of γ for sensible network representations and parameter

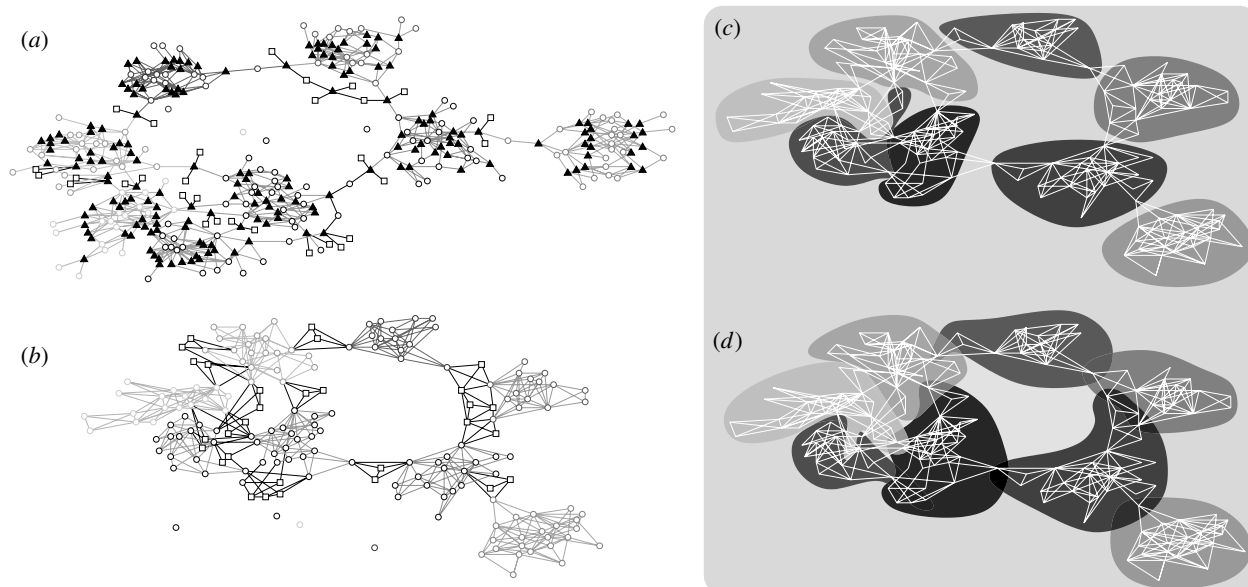


Figure 3. An example of a network realization with model parameters $g=8$, $N_A=32$, $N_R=175$, $\gamma=0.90$, $\mu_{\max}=4$ and $n_{\text{trial}}=20$. (a) A representation of the reaction system as a bipartite network where one edge type represents reactions and the other type represents chemical substances (circles, in-cluster substances; squares, system-wide substances; triangles, reactions). (b) The corresponding substance network. Different colours represent different groups in the construction algorithm. Vertices created on the addition of a fraction γ of long-range reactions are symbolized by squares. (c) The modules detected from the substance network in (b). (d) The modules identified by first running the network-clustering algorithm on the reaction network, then assigning the same cluster identity to all vertices taking part in a reaction of a specific network cluster. In this case, one vertex may belong to several network clusters. The underlying (white) network in (d) is, for comparison, the same as in (c).

values. In figure 4a, we plot Δ as a function of γ for all types of networks discussed, and find such a monotonous relationship. Preferably, one would like the derived networks to span a large range of Δ -values. All three types of networks where the vertices are substances perform comparatively well in this respect, whereas the reaction network seems a little worse (even if the actual Δ -values are larger).

When we tune γ , we keep the size of the reaction system—the number of reactions and their order (number of substrates and products)—constant. However, the size of the derived network can be γ dependent, which complicates the analysis of the network modularity a little. As seen in figure 4b,c, except the reaction network, the networks get smaller and denser as γ is increased. The reaction network has (by construction) N_R vertices constantly. Even if $\Delta=0$ represents a neutral modularity with respect to the null model, one needs to be careful when comparing Δ for networks of different sizes and average degrees. In figure 4d, we plot the two terms, $\hat{Q}(G)$ (the maximal modularity of the network) and $\langle \hat{Q}(G') \rangle_{G' \in \mathcal{G}(G)}$ (the average maximal modularity of $\mathcal{G}(G)$), subtracted in the definition of Δ for the substance networks. The decrease of the second term as a function of γ means that smaller and denser networks have lower network modularity \hat{Q} . If this were true for the networks derived from our reaction system model too, then, if we choose the size of the reaction system such that the size and average degree of the derived networks are independent of γ , the trend of growing \hat{Q} would be even stronger. This indicates that γ works as a tuning parameter for network modularity (in the substance network)

whether one keeps the size of the original reaction system or the derived networks fixed. The same conclusion can be drawn for all the other three types of networks.

3.3. Module predictability

As mentioned above, when γ is large, and the network is reasonably small, one can spot the modules by inspection, as shown in figure 3a. By applying a network-clustering method, one can then recover the group structure, as shown in figure 3c. One desirable property of a network representation of a reaction system is that this procedure is accurate. In other words, the original, modular structure should be recovered reasonably well by the clustering algorithm in the presence of the noise created by the inter-group reactions. We measure this noise tolerance of the network representations, or module predictability (to stress another angle of the issue), by the fraction of overlapping group identities in the best match between the original group identities and the identities detected by the clustering algorithm. In other words, let x_i be vertex i 's group in the original reaction system construction ($x_i \in [1, \dots, g]$) and y_i be vertex i 's identity from the graph clustering algorithm ($y_i \in [1, \dots, h]$, where h is the number of detected groups). Then choose a labelling of the graph-clustering groups such that each group has a unique number in the interval $[1, \dots, h]$, and that the number n_{match} of vertices i with $x_i=y_i$ is as large as possible. Then, we define the predictability $\lambda = n_{\text{match}}/n_g$, where n_g is the number of vertices except for the vertices introduced while adding inter-module reactions.

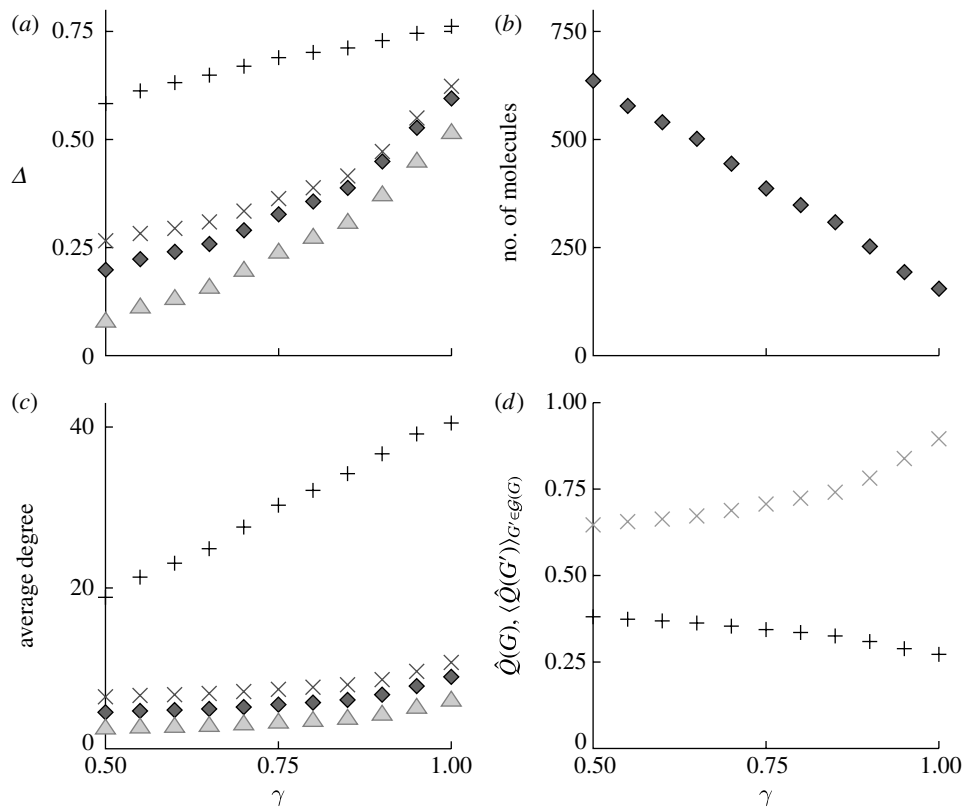


Figure 4. Statistics of the model as a function of the parameter γ . The other model parameter values are $g=10$, $N_A=50$, $N_R=500$, $\mu_{\max}=4$ and $n_{\text{trail}}=100$. The points are averaged over 100 network realizations. Standard errors are smaller than the symbols (diamonds, substrate–product; triangles, substrate–substrate; crosses, substance; plus signs, reaction), and therefore not shown. (a) The relative modularity Δ of the networks as a function of γ . (b) The average number of molecular species for substrate–product networks (the number is, by construction, the same for substrate–substrate and substance networks), while it is exactly N_R for reaction networks. (c) The average degree and (d) the two different terms in the calculation of Δ for the substance networks (crosses, $\hat{Q}(G)$; plus signs, $\langle \hat{Q}(G') \rangle_{G' \in \mathcal{G}(G)}$).

We calculate n_{match} by a simple heuristic,

- (i) Start with a random labelling of the groups.
- (ii) Select a pair of group labels.
- (iii) If n_{match} does not decrease and if these labels are swapped, then swap them.
- (iv) If no improvement has been made during the last n_{rep} steps, then go to step (ii).
- (v) Start over from step (i) with a new random seed unless a new highest n_{match} has been found in step (iv), the last N_{rep} time steps.

For our small system (h is rarely larger than 20), choosing N_{rep} and n_{rep} between 100 and 10^4 gives the same results. In figure 5, we display the result for the three network types with vertices being chemical substances (we do not include the reaction network, although that could, with a few additional assumptions, also be done). The matching is closest for the substrate–product and substance networks. This observation is in line with the conclusions of a study of metabolic networks (Holme & Huss 2008), where these networks gave the most plausible set of currency metabolites and best functional overlap. Finally, we note that one major factor in the decrease of λ as γ is lowered from 1 is that the network-clustering algorithm identifies too many groups (going up to more than twice the value of g).

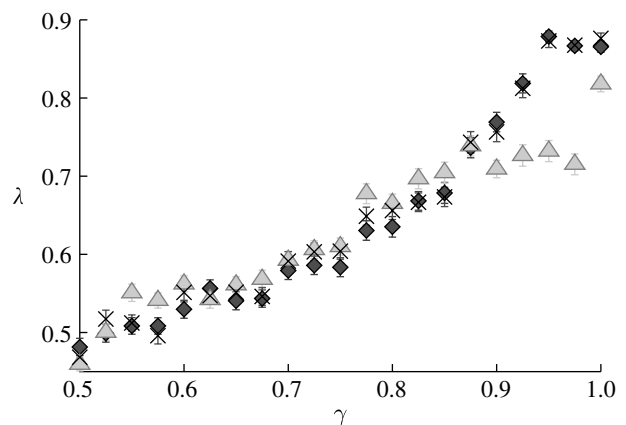


Figure 5. The predictability λ as a function of γ . λ is the fraction of correctly predicted group identities in matchings between the original group identities in the algorithm and the clusters identified by a network-clustering algorithm and the original groups in the network constructions (diamonds, substrate–product; triangles, substrate–substrate; crosses, substance). The parameter values are the same as in figure 4.

It might be the case that another network-clustering method, being more restrictive in the number of detected molecules, can give higher λ -values. In other words, the information of the original graph structure might decay slower with decreasing γ than indicated by figure 5.

3.4. The relationship between molecular mass and degree

As a final analysis of the reaction system model, we include our explicit representation of atoms. One situation where this aspect of our model can prove useful is to explain the relationship between molecular mass and degree. In figure 6a, we plot the degree as a function of molecular mass for human metabolic data from the KEGG database (the same data as used in Holme & Huss (2008)). There is a decaying trend in this relationship (not clear enough to talk of a scaling law, though). In figure 6b, we show the corresponding plot for our model reaction system (all network representations except the reaction networks, since the mass of a reaction is somewhat hard to define). For our model, we assume that all molecules have the same weight. Neither the degree nor the mass have such large ranges of values in the model network as in the real network. The functional form of the model curves does not match the empirical curves; however, there is a clear trend of decay in these data as well. Possibly, the stoichiometric constraints of the model are enough to explain this correlation in real reaction systems. More work is needed to confirm this conjecture; we leave it as a speculation in this paper.

4. DISCUSSION AND CONCLUSIONS

We have investigated four types of network representations of chemical reaction systems. Of these, substrate-product and substance networks are the ones that encode the information about modularity best. This is in line with observations that these two network representations are the ones that best capture the functional organization of metabolism (Holme & Huss 2008). To reach this conclusion, we presented a scheme to generate chemical reaction systems such that the network modularity of the derived networks can be tuned by an input parameter. The model contains molecules with atoms modelled explicitly and reactions ensured to conserve mass. The only other structure that is embedded is network modularity. One can use the model to study the response of various chemical phenomena to network modularity. It can also be a base for more elaborate modelling, including the highly skewed degree distributions of both metabolic networks and those derived from reaction systems in planetary atmospheres. Except some basic stoichiometric restrictions (included via the explicit modelling of the atoms), the model has very few constraints; this, however, can cause observable effects in the network topology. One such effect that we find is that our model displays a negative correlation between molecular mass and degree, a feature that also can be observed in real-world metabolic networks.

In future studies, in addition to making the model more accurate as a generative model of real reaction systems, it would be interesting to modify it to a model of the evolution of chemical reaction systems (Furusawa & Kaneko 2006). In such an approach, it would be desirable that modularity emerges from the evolutionary dynamics (Grönlund & Holme 2004) rather than being imposed by the construction algorithm. Another direction is to use a more dynamic interpretation of

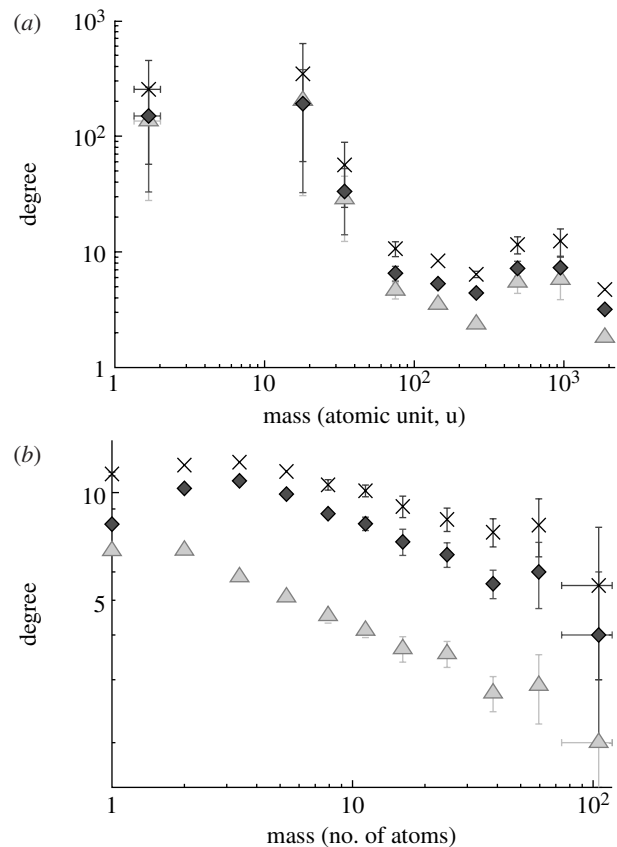


Figure 6. Correlation between molecular weight and degree in networks derived from real and model reaction systems. The points are averages for logarithmic bins. (a) The data from the human metabolic network from the KEGG database. (b) The corresponding plot for our model. The mass of all atoms is set to unity in this case (so mass is just a count of the number of atoms constituting the molecule). Here, $\gamma=0.94$ and the points are averages over 20 networks (diamonds, substrate-product; triangles, substrate-substrate; crosses, substance). Other parameter values are the same as in figure 4.

modularity (Han *et al.* 2004; Gallos *et al.* 2007) and find a generative model for such systems.

P.H. acknowledges economic support from the Swedish Foundation for Strategic Research and the Japan Society of the Promotion of Science, and thanks Mikael Huss and Andreea Munteanu for data, and Mikael Huss for comments.

REFERENCES

- Del Vecchio, D., Ninfa, A. & Sontag, E. 2008 Modular cell biology: retroactivity and insulation. *Mol. Syst. Biol.* **4**, 4 100–204. (doi:10.1038/msb4100204)
- Furusawa, C. & Kaneko, K. 2006 Evolutionary origin of power-laws in a biochemical reaction network: embedding the distribution of abundance into topology. *Phys. Rev. E* **73**, 011 912. (doi:10.1103/PhysRevE.73.011912)
- Gallos, L. K., Song, C., Havlin, S. & Makse, H. A. 2007 Scaling theory of transport in complex biological networks. *Proc. Natl Acad. Sci. USA* **104**, 7746–7751. (doi:10.1073/pnas.0700250104)
- Girvan, M. & Newman, M. E. J. 2002 Community structure in social and biological networks. *Proc. Natl Acad. Sci. USA* **99**, 7821–7826. (doi:10.1073/pnas.122653799)
- Goelzer, A., Brikci, F. B., Martin-Verstraete, I., Noirot, P., Bessières, P., Aymerich, S. & Fromion, V. 2008

- Reconstruction and analysis of the genetic and metabolic regulatory networks of the central metabolism of *Bacillus subtilis*. *BMC Syst. Biol.* **2**, 20. (doi:10.1186/1752-0509-2-20)
- Grönlund, A. & Holme, P. 2004 Networking the seceder model: group formation in social and economic systems. *Phys. Rev. E* **70**, 036 108. (doi:10.1103/PhysRevE.70.036108)
- Han, J.-D. J. *et al.* 2004 Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature* **430**, 88–93. (doi:10.1038/nature02555)
- Holme, P. & Huss, M. 2008 Substance graphs are optimal simple-graph representations of metabolism. E-print. (arxiv:0806.2763).
- Huss, M. & Holme, P. 2007 Currency and commodity metabolites: their identification and relation to the modularity of metabolic networks. *IET Syst. Biol.* **1**, 280–285. (doi:10.1049/iet-syb:20060077)
- Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y. & Barkai, N. 2002 Revealing modular organization in the yeast transcriptional network. *Nat. Genet.* **31**, 370–377. (doi:10.1038/ng941)
- Kitano, H. 2004 Biological robustness. *Nat. Rev. Genet.* **5**, 826–837. (doi:10.1038/nrg1471)
- Kumpula, J., Saramäki, J., Kaski, K. & Kertész, J. 2007 Limited resolution in complex network community detection with potts model approach. *Eur. Phys. J. B* **56**, 41. (doi:10.1140/epjb/e2007-00088-4)
- Maslov, S. & Sneppen, K. 2002 Specificity and stability in topology of protein networks. *Science* **296**, 910–913. (doi:10.1126/science.1065103)
- Newman, M. E. J. 2006 Modularity and community structure in networks. *Proc. Natl Acad. Sci. USA* **103**, 8577–8582. (doi:10.1073/pnas.0601602103)
- Pothen, A., Simon, H. & Liou, K. 1990 Partitioning sparse matrices with eigenvectors of graphs. *SIAM J. Matrix Anal.* **11**, 430. (doi:10.1137/0611030)
- Rosvall, M. & Bergstrom, C. T. 2007 An information-theoretic framework for resolving community structure in complex networks. *Proc. Natl Acad. Sci. USA* **104**, 7327–7331. (doi:10.1073/pnas.0611034104)
- Solé, R. V. & Munteanu, A. 2004 The large-scale organization of chemical reaction networks in astrophysics. *Europhys. Lett.* **68**, 170–176. (doi:10.1209/epl/i2004-10241-3)
- Zhao, J., Yu, H., Luo, J.-H., Cao, Z.-W. & Li, Y.-X. 2006 Hierarchical modularity of nested bow-ties in metabolic networks. *BMC Bioinform.* **7**, 386. (doi:10.1186/1471-2105-7-386)
- Zhao, J., Ding, G.-H., Tao, L., Yu, H., Yu, Z.-H., Luo, J.-H., Cao, Z.-W. & Li, Y.-X. 2007 Modular co-evolution of metabolic networks. *BMC Bioinform.* **8**, 311. (doi:10.1186/1471-2105-8-311)
- Ziv, E., Middendorff, M. & Wiggins, C. H. 2005 Information-theoretic approach to network modularity. *Phys. Rev. E* **71**, 046 117. (doi:10.1103/PhysRevE.71.046117)