Substance graphs are optimal simple-graph representations of metabolism

Petter Holme\textsuperscript{1,2} and Mikael Huss\textsuperscript{3}
\textsuperscript{1}Computational Biology, Royal Institute of Technology, 100 44 Stockholm, Sweden
\textsuperscript{2}Department of Physics, Umeå University, 901 87 Umeå, Sweden
\textsuperscript{3}Systems Biology, Genome Institute of Singapore, 138 672 Singapore, Singapore

One approach to studying the system-wide organization of biochemistry is to use statistical graph theory. Even in such a heavily simplified method, which disregards most of the dynamic aspects of biochemistry, one is faced with fundamental questions, such as how the chemical reaction systems should be reduced to a graph retaining as much functional information as possible from the original reaction system. In such graph representations, should the edges go between substrates and products, or substrates and substrates, or both? Should vertices represent substances or reactions? Different definitions encode different information about the reaction system. In this paper we evaluate four different graph representations of metabolism, applied to data from different organisms and databases. The graph representations are evaluated by comparing the overlap between clusters (network modules) and annotated functions, and also by comparing the set of identified currency metabolites with those that other authors have identified using qualitative biological arguments. We find that a “substance network,” where all metabolites participating in a reaction are connected, is relatively better than others, evaluated both with respect to the functional overlap between modules and functions and to the number and identity of identified currency metabolites.

I. INTRODUCTION

Metabolism, the set of all chemical processes in an organism that are necessary for the maintenance of life, can be studied at different levels — from the quantum chemistry of reactions via small reaction pathways and larger feedback loops to system-wide organization. For the smaller-scale problems, sets of reactions are often modeled by systems of ordinary differential equations (ODEs), either in order to perform explicit simulations or to make use of steady-state properties of the systems such as in “flux mode analysis” (17). However, to investigate the system-wide, large-scale organization for whole organisms, such modeling frameworks become obstructively complex. An alternative paradigm is graph methods, where the chemical reaction system is reduced to a graph of nodes (or vertices) pairwise connected by links (edges). In this paper, we define a metabolic network as any simple graph (unweighted, undirected graph without multiple edges or self-edges) derived from a metabolic reaction system. In the process of constructing a metabolic network, one has to discard much of the existing information about reaction systems (information about reaction coefficients, localization etc.). In return, one obtains an object of study — a graph — for which numerous methods exists to characterize its large-scale organization.

As illustrated in Fig. 1, there are several ways to represent chemical reaction systems, such as an organism’s set of metabolic reactions, as simple graphs. The most common way, Fig. 1b, is to link substrates to products in a substrate–product network (sometimes called just “substrate network” (22)). Furthermore, one can consider linking products to products and substrates to substrates into what we call a substrate–substrate network (Fig. 1c), or linking all metabolites participating in the same reaction with each other into a substance network (Fig. 1d). As a final representation, instead of linking all substrates participating in the same reaction, one can make a reaction network (Fig. 1e) of reactions connected if they have a substance in common. These different representations can potentially capture different aspects of the reaction system — the graph distances in a substrate–product network reflect the number of reactions atoms traverse on the way between two substances, and so on. For specific questions about the system, one of these representations might be more useful than the others. Most questions appropriate for a graph-based approach, however, hinge on the assumption that the wiring mirrors the functional organization of the network. To evaluate the different representations, we use three empirical criteria for investigating how well the networks capture biological functionality. First, we investigate the assumption (1) that a tightly knit group of metabolites are relatively likely to

\textbf{FIG. 1} Panel a shows two reactions that together form a minimal reaction system. b displays the substrate–product network derived from the reaction system in a. Here, the vertices are chemical substances that are connected if they are substrate and product of a reaction. c shows the substrate–substrate network where substances are connected if they react with each other. d illustrates the substance network where substances are connected if the occur in the same reaction. e shows the reaction network derived from a, where two reactions are connected if they share a substance.
be involved in the same biological processes, and investigate how well network clusters match annotated functions. Second, we investigate the quality of currency metabolites (ubiquitous substances, like water and carbon dioxide, participating in a large number of reactions; a more precise definition follows below) derived from the network. Finally, making the assumptions that metabolic networks show modular structure, we require that the network representation should yield a clear modularity after currency metabolites have been removed.

II. RESULTS

In this paper we use the reaction systems of ten datasets (eight different organisms from the KEGG and BiGG databases, accessed April 2008). Sizes and basic statistics of the derived metabolic networks are given in Appendix 1. We use a previously described algorithm (6) to decompose the networks into modules. This algorithm essentially maximizes network modularity by successively removing high-degree metabolites. Its output is twofold: on one hand, it returns a set of clusters (or modules), and on the other hand, it identifies a set of currency metabolites. The resulting modules should be enriched in molecules having similar annotation, and the resulting currency metabolites should correspond to substances that have been described as currency metabolites by previous authors. These will be our main criteria for evaluating the different graph representations. Also, Refs. (4; 5; 7; 11; 15; 21; 23) argue that network modularity is a principle of metabolic organization. Thus, we also expect of a good graph representation that the modular structure should be clear after the deletion of the currency metabolites.

A. Functional matching

What is the nature of the dense clusters of non-currency metabolites detected by our decomposition algorithm? The edges in all four representations represent some kind of functional relationship between the connected vertices. Although these functional relationships are slightly different, one can expect the vertices of a cluster to have a stronger functional coupling to each other than to vertices of other clusters. In other words, a cluster should be functionally relatively independent, which is one aspect of the biological notion of a module. To investigate the relationship between clusters and modules, we measure a matching score (defined in the Methods section), displayed in Fig. 2a. A positive value of the matching score represents an overlap greater than expected from a situation where the annotated functions were randomly spread out over the vertices (i.e., relative to a randomized set of the functions). Positive values of the functional overlap cannot be larger than unity. Examining the scores for different organisms in Fig. 2a, we notice that all scores are significantly positive, meaning that there is a tendency for functionality to be concentrated to the network clusters. Comparing the different representations, we see that the reaction networks seem to be particularly poor in capturing the functional aspects of the reaction system. The other three representations appear to have similar performance, with the substance network having especially good performance in the human BiGG and the M. genitalium data sets.

B. Currency metabolites

Hubs in biomolecular networks are usually abundant molecular species (6) which do not put much constraint on the metabolic flows. Accordingly, they tend not to have much regulatory (or other higher-order) functionality. The highest-degree vertices typically connect different network clusters, thus blurring the modular structure of the network. This property (illustrated by the human substance network from the KEGG database in Fig. 3) and the high degree are the two components in our operational definition of currency metabolites (5): If vertices are deleted from the network in order of highest degree, then the set of currency metabolites is the set of vertices that, if deleted, gives the highest relative modular-
TABLE I A comparison between currency metabolites detected by our algorithm (all detected in at least one of eight organisms using substance networks) and currency metabolites defined by Wagner and Fell (20), Schuster et al. (18), and Ma and Zeng (10). Only metabolites which were considered currency metabolites in at least two of the four studies are shown. In Ref. (20), the analysis was done in two versions, one excluding and one including the six first metabolites (ATP to NADH). In Refs. (10) and (18), the sets of currency metabolites were considered context-dependent and thus not completely fixed, hence these lists are approximations. A list of currency metabolites for all studied organisms and graph types in this work can be found in Appendix 2.

|          | this work | Wagner & Fell | Schuster et al. | Ma & Zeng |
|----------|-----------|---------------|-----------------|-----------|
| ATP      | ATP       | ATP           | ATP             | ATP       |
| ADP      | ADP       | ADP           | ADP             | ADP       |
| NADPH    | NADPH     | NADPH         | NADPH           | NADPH     |
| NADP⁺    | NADP⁺     | NADP⁺         | NADP⁺           | NADP⁺     |
| NAD⁺     | NAD⁺      | NAD⁺          | NAD⁺            | NAD⁺      |
| NADH     | NADH      | NADH          | NADH            | NADH      |
| Pᵢ       | Pᵢ        | Pᵢ            | Pᵢ              | Pᵢ        |
| PPᵢ      | PPᵢ       | PPᵢ           | PPᵢ             | PPᵢ       |
| CO₂      | CO₂       | CO₂           | CO₂             | CO₂       |
| H₂O      | H₂O       | H₂O           | H₂O             | H₂O       |
| NH₃      | NH₃       | NH₃           | NH₃             | NH₃       |
| SO₄      | SO₄       | SO₄           | SO₄             | SO₄       |
| H⁺       | H⁺        | H⁺            | H⁺              | H⁺        |
| O₂        | O₂        | O₂            | O₂              | O₂        |

FIG. 3 Illustration of the network structural role of currency metabolites. The figure is based on the human substance network from the KEGG database. The modules are indicated in different colors (areas of the circles are proportional to the size of the clusters). The width of the lines are proportional to the number of edges leading between two different clusters, or from a currency metabolite to a cluster. In panel b, the currency metabolites are removed from the picture and the modular structure is more visible.

C. Modular structure

In Fig. 2c we plot the relative modularity $\Delta$ in a fashion corresponding to Fig. 2a. In general, the substance and reaction networks show the highest $\Delta$-value. This suggests that these representations may be superior in capturing modular structure of metabolic networks. $\Delta$ (defined in Methods) is a modification of Newman and Girvan’s $Q$-modularity (14) for the algorithm to detect currency metabolites. $Q$ is known to be biased by size and degree sequence (3; 8). While $\Delta$ removes enough of these effects to work in the currency metabolite algorithm (where we only compare the marginal effect on $\Delta$ on the removal of vertices), the size differences between the different representations might affect the conclusions. If a cautious conclusion is to be drawn, the large difference between $\Delta$ of the substance network compared with the substrate–product and substrate–substrate representations suggests that the substrate network representation is likely to be better at capturing the modular structure of the network, although the reaction network (which performs poorly for the other two criteria) is even better than the substance network. This high-modularity criterion is the weakest one, as it does not relate to any external information. Nevertheless, the outcome of the test is at least not an argument against the substance network representation.

D. Functional organization — A continuum between a core and a modular periphery

From a biological viewpoint, maybe the most important conclusion from Fig. 2a is that the functions are concentrated to the network clusters — $\mu$ is significantly larger than zero for all graph representations and organism. But how is the functionality distributed among the clusters? One description of metabolism is that currency metabolites form the core

ity. Note that this definition does not specify the graph representation. Early papers (10; 18; 20) have defined currency metabolites based on biological reasoning. Although these sets of currency metabolites are somewhat disparate (Table I), we recognize the relevance of these biological arguments. A quality criterion of the graph representations would then be that the currency metabolites defined by these representations should have considerable overlap with these other studies. A first check is that the number of currency metabolites is of the same order, between 11 and 14, as proposed in the other studies (10; 18; 20). In Fig. 2b we plot the number of currency metabolites for ten organisms of various complexity, from two different databases. None of the representation produces a very consistent set of currency metabolites. This is not expected either — consider, for example, oxygen, which appears as a currency metabolite in the human metabolic network according to our algorithm; in the metabolism of anaerobic bacteria, it will most likely not be a currency metabolite (cf. Ref. (16). The representation whose number of currency metabolites agrees best with the range 11–14 in humans is the substance network.
A. Network modularity and currency metabolites

Real-world complex networks, including metabolic networks, can be described as having both randomness and some regularities, or network structure. The network structure contains information both about how dynamic systems associated with the network behave, and about the evolutionary history of the network itself. A key to accessing this information is to design quantities to measure network structure. One such type of network-structural quantities are measures of network modularity — how well a graph can be decomposed into sub-

III. METHODS

A. Network modularity and currency metabolites

E. Discussion

In this paper we have discussed four different types of graph representations of metabolism — substrate–product networks, substrate–substrate networks, substance networks and reaction networks. The first three representations are more or less the only ways, using simple rules and available database information, to reduce a reaction system to a graph of interlinked substances. For reaction graphs one can imagine other varieties (e.g., connecting reactions sharing substrates), but in this work we consider reaction graphs mostly to illustrate the possibility.

We use quantitative arguments for assessing which one of these representations that is the most informative about the functional organization of metabolic reaction networks. The criteria are that the network modules should have a high overlap with annotated functions, identified currency metabolites should be similar to those of other studies, and the modular structure should be clear. With these criteria, the substrate graph representation emerges as a relatively clear winner. The other representations may, however, be preferable for more specific questions. For instance, a substrate–product graph stresses the interconversion between substances, making it possible to trace paths between different substances. Substrate–substrate graphs highlight reactivity between metabolites. A substance graph embodies information about spatial and temporal co-localization of substances, while the reaction networks show how the reactions are connected. An investigator may be able to select a suitable representation depending on the intended usage of the network, be it bottleneck identification, lethality prediction, or one of any number of possible applications.

Analyzing the substrate networks, we can describe the large-scale functional organization of metabolism as centered around a core of currency metabolites with other clusters forming a periphery of gradually increasing functional specificity, decreasing multi-functionality and decreasing coupling to currency metabolites. A full verification of this picture has to wait for future functional annotations of higher precision.

In future work, it would be interesting to compare the different graph representations in a dynamic setting with some process acting over the vertices, or in a task-oriented setting where the networks are used to address a specific problem. However, we believe that our work presented here, based on simple, static graphs (for example, relating metabolites cellular and pathogenic functions) is a necessary first step to understand the basic differences between the graph types.
networks that are densely connected within and sparsely connected between each other. In Ref. (14) a measure of modularity of a partition of a graph into subgraphs was proposed:

\[ Q = \sum_i \left[ e_{ii} - \left( \frac{\sum_j e_{ij}}{\sum_j e_{jj}} \right)^2 \right], \quad (1) \]

where the sum is over subgraphs and \( e_{ij} \) is the fraction of edges that leads between vertices of cluster \( i \) and \( j \). The term \( \left( \sum_j e_{ij} \right)^2 \) is the expected value of \( e_{ii} \) in a random multigraph (without any correlations). One way of detecting (or defining) clusters in networks is to maximize \( Q \) over all partitions, of all sizes. This is a computationally hard optimization problem (2) that a large number of papers has proposed heuristics for solving (see Ref. (13) and references therein). We use a recent and competitive algorithm proposed in Ref. (13).

The maximal value \( \hat{Q} \) of \( Q \) for all partitions is a prototype measure of the modularity of a network. However, as hinted above, network structure needs to be measured relative to a measure of the modularity of a network. This is a resampling technique, randomly rewiring the edges of the original network (12). In this work we use averages over 100 independent samples of this ensemble.

The maximal value \( \hat{Q} \) of \( Q \) for all partitions is a prototype measure of the modularity of a network. However, as hinted above, network structure needs to be measured relative to a measure of the modularity of a network. This is a resampling technique, randomly rewiring the edges of the original network (12). In this work we use averages over 100 independent samples of this ensemble.

B. Functional matching

Metabolites in KEGG are annotated with one or several pathways which we equate with functions. In BiGG, similar functional (pathway) annotations are instead assigned to reactions, but one can indirectly assign functions to a metabolite by simply selecting all functions listed for reactions where the metabolite participates. Every metabolite \( i \in V \) is thus associated with a set of functions \( f_i \in F \) (\( F \) is the set of all functions). We can use this to evaluate the assumption that network modules correspond to biological functions. Let \( \phi_{CF}(c, f) \) denote the fraction of vertices with function \( f \) belonging to cluster \( c \in C(G) \) is the set of detected clusters of \( G \); let \( \phi_{F}(f) \) denote the fraction of vertices with function \( f \); and let \( \phi_{C}(c) \) denote the fraction of vertices in cluster \( c \). Then, if the functions were randomly distributed, in an infinite system, the expectation value of \( \phi_{CF} \) would be \( \phi_{C}(c)\phi_{F}(f) \). From this we derive a prototypical score function for the match between network modules and metabolite function:

\[ v = \sum_{c \in C} \sum_{f \in F} |\phi_{CF}(c, f) - \phi_{C}(c)\phi_{F}(f)|, \quad (3) \]

where \(|\cdots|\) denotes absolute values of numbers and cardinality (number of elements) of sets. High values of \( v \) mean that functions are concentrated to network clusters. If the number of vertices and the sizes of clusters go to infinity, \( v = 0 \) signals neutrality. Note that this definition does not require the vertex set to be partitioned into a set of distinct clusters (similarly it allows several functions to be assigned to one vertex). When we apply the currency metabolite definition algorithm to the product network, we still delete vertices (and all the reactions they participate in), but define the currency metabolites with respect to the modularity of the reaction network. This means that the network clusters will be sets of reactions. When calculating \( v \) for the reaction network, we let a metabolite belong to the network clusters of all the reactions in which it participates. However, \( v \) does not capture all the aspects we desire — since fluctuations give a positive contribution to \( v \) (due to the absolute values), the finite sizes of the reaction systems will give a positive bias. To make zero represent neutrality, we rather measure

\[ \mu = \frac{v - \langle v' \rangle}{v - \langle v' \rangle}, \quad (4) \]

where \( v' \) is a random configuration of functions with the properties that \( F \) and \( \phi_{F}(f) \) (for all \( f \in F \)) is the same as in the original system, that no vertex is assigned the same function more than once, and \( \langle \cdots \rangle \) denotes the average over 100 samples of this ensemble, and \( v' \) represents the ensemble’s maximal value.
Acknowledgments

PH acknowledges economic support from the Swedish Foundation for Strategic Research. The authors thank Lei-Han Tang and Jing Zhao for insightful comments.

References

[1] U. Alon. Biological networks: The tinkerer as an engineer. *Science*, 301:1866–1867, 2003.
[2] U. Brandes, D. Delling, M. Gaertler, R. Görke, M. Hoefer, Z. Nikoloski, and D. Wagner. On finding graph clusterings with maximum modularity. *Lecture Notes in Computer Science*, 4769:121–132, 2007.
[3] R. Guimerà, M. Sales-Pardo, and L. A. Nunes Amaral. Modularity from fluctuations in random graphs and complex networks. *Phys. Rev. E*, 70:025101, 2004.
[4] P. Holme, M. Huss, and H. Jeong. Subnetwork hierarchies of biochemical pathways. *Bioinformatics*, 19:532–538, 2003.
[5] M. Huss and P. Holme. Currency and commodity metabolites: Their identification and relation to the modularity of metabolic networks. *IET Systems Biology*, 1:280–285, 2007.
[6] K. Kaneko. *Life: An introduction to complex systems biology*. Springer, Berlin, 2006.
[7] A. Kreimer, E. Borenstein, U. Gophna, and E. Ruppin. The evolution of modularity in bacterial metabolic networks. *Proc. Natl. Acad. Sci. USA*, 105:6976–6981, 2008.
[8] J. Kumpula, J. Saramäki, K. Kaski, and J. Kertész. Limited resolution in complex network community detection with potts model approach. *Eur. Phys. J. B*, 56:41, 2007.
[9] D.-S. Lee, J. Park, K. A. Kay, N. A. Christakis, Z. N. Oltvai, and A.-L. Barabási. The implications of human metabolic network topology for disease comorbidity. *Proc. Natl. Acad. Sci. USA*, 105:9880–9885, 2008.
[10] H. Ma and A.-P. Zeng. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics*, 19:270–277, 2003.
[11] H.-W. Ma, X.-M. Zhao, Y.-J. Yuan, and A.-P. Zeng. Decomposition of metabolic network based on the global connectivity structure of reaction graph. *Bioinformatics*, 20:1870–1876, 2004.
[12] S. Maslov and K. Sneppen. Specificity and stability in topology of protein networks. *Science*, 296:910–913, May 2002.
[13] M. E. J. Newman. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA*, 103:8577–8582, 2006.
[14] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113, 2004.
[15] E. Ravasz, A. Somera, D. Mongru, Z. N. Oltvai, and A.-L. Barabasi. Hierarchical organization of modularity in metabolic networks. *Science*, 297:1551–1555, 2002.
[16] J. Raymond and D. Segrè. The effect of oxygen on biochemical networks and the evolution of complex life. *Science*, 311:1764–1767, 2006.
[17] S. Schuster, D. Fell, and T. Dandekar. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature Biotechnol.*, 18:326–332, 2000.
[18] S. Schuster, T. Pfeiffer, F. Moldenhauer, I. Koch, and T. Dandekar. Exploring the pathway structure of metabolism: decomposition into subnetworks and application to mycoplasma pneumoniae. *Bioinformatics*, 18:351–361, 2002.
[19] A. E. Vinogradov. Modularity of cellular networks shows general center-periphery polarization. *Bioinformatics*, page btn555, 2008.
[20] A. Wagner and D. A. Fell. The small world inside large metabolic networks. *Proc. Biol. Sci.*, 268:1803–1810, 2001.
[21] J. Zhao, G.-H. Ding, L. Tao, H. Yu, Z.-H. Yu, J.-H. Luo, Z.-W. Cao, and Y.-X. Li. Modular co-evolution of metabolic networks. *BMC Bioinformatics*, 8:311, 2007.
[22] J. Zhao, H. Yu, J. Luo, Z. W. Cao, and Y. Li. Complex networks theory for analyzing metabolic networks. *Chinese Science Bulletin*, 51:1529–1537, 2006.
[23] J. Zhao, H. Yu, J.-H. Luo, Z.-W. Cao, and Y.-X. Li. Hierarchical modularity of nested bow-ties in metabolic networks. *BMC Bioinformatics*, 7:386, 2006.
Appendix 1:  
Network sizes

Substance graphs are optimal simple-graph representations of metabolism

Petter Holme and Mikael Huss

In this file we present the sizes of the networks and give a brief explanation to why the reaction networks are denser than other networks.
Table of some fundamental quantities — the number of vertices $N$, edges $M$, currency metabolites $c$ and relative modularity $\Delta$ — for the various types of networks, for different organisms and different databases (KEGG, unless otherwise stated). $\Delta$ and $c$ are compared visually in Fig. 3 of the paper.

| organism                  | N   | M    | $\Delta$ | c   | $N$   | M    | $\Delta$ | c   |
|---------------------------|-----|------|----------|-----|-------|------|----------|-----|
| human                     | 2310| 6785 | 0.118    | 9   | 2303  | 609,377| 0.236    | 3   |
| human (BiGG)              | 1507| 6527 | 0.197    | 7   | 3311  | 1,307,061| 0.405   | 2   |
| S. cerevisiae (KEGG)      | 1577| 4426 | 0.132    | 2   | 1449  | 229,483| 0.235    | 5   |
| S. cerevisiae (BiGG)      | 641 | 2693 | 0.160    | 10  | 841   | 153,200| 0.482    | 5   |
| C. elegans                | 1587| 4250 | 0.144    | 1   | 1378  | 206,869| 0.203    | 6   |
| M. musculus               | 2255| 6578 | 0.141    | 8   | 2214  | 562,374| 0.231    | 7   |
| R. norvegicus             | 2098| 5922 | 0.140    | 1   | 1956  | 459,000| 0.213    | 7   |
| D. melanogaster           | 1897| 5273 | 0.141    | 2   | 1781  | 361,124| 0.220    | 5   |
| M. genitalium             | 473 | 1030 | 0.184    | 3   | 328   | 7104  | 0.381    | 0   |
| E. coli                   | 1864| 5262 | 0.127    | 3   | 1800  | 330,247| 0.263    | 5   |
Why reaction networks are denser than other representations

To sketch an explanation of this phenomenon for the case of substance and reaction networks (the explanations for the other two types are similar) consider a bipartite network with two types of vertices representing reactions or substances, with edges linking substances to reactions in which they participate. In this network, the degrees of reactions are very narrowly distributed (between two and eight in the human network from KEGG), whereas the degrees of substances are large (between one and 836 in the human KEGG-network). Now, consider a vertex \( i \) and assume that the network is sparse and locally tree-like (no cycles within two steps from \( i \)). The expected degree of a vertex in the projected network (the substance network if the vertex is a substance, or reaction network if the vertex is a reaction) is

\[
k = \sum_{k'} k' p_{k'} \sum_{K} p_K K(K - 1) = \langle k' \rangle (\mu_2(K) - \langle K \rangle)
\]

where \( k' \) is the degree (in the bipartite representation) of the kind of vertices projected to, and \( K \) is the degree of the kind of vertices projected away from. Angular parentheses represent averages and \( \mu_2 \) symbolizes the second moment. Simply speaking, a vertex of the other type contributes to \( k \) by the square of its degree, because the probability of an edge leading to a vertex of degree \( K \) is proportional to \( K \), and if an edge leads from \( i \) to a neighbor of degree \( K \), all the \( K - 1 \) other vertices will be attached to \( i \) in the projected network. Now, since \( K \geq 1 \) for both types of vertices, \( \mu_2(K) \) is the leading term in Eq. 1. Since the second moment weights high degree vertices (of the opposite type), higher, we understand that the average degree would be much higher for the reaction networks.
Appendix 2:
Currency metabolites

Substance graphs are optimal simple-graph representations of metabolism

Petter Holme and Mikael Huss

In this file, we provide a figure illustrating the change of relative modularity $\Delta$ as vertices are deleted in the algorithm defining currency metabolites. The darker bars indicate the currency metabolites. The figure represents a substance-graph representation of the human network from the KEGG database. After that we list the sets of currency metabolites for all organisms and network types studied. The order (from top to bottom) represents the order in which they were deleted (like moving from left to right in the figure below).
### Optimal simple-graph representations of metabolism

| substr.–prod. | substr.–substr. | substa. | reac. |
|---------------|----------------|---------|-------|
| H₂O          | H₂O            | H₂O     |       |
| H⁺           | H⁺             | H⁺      |       |
| Oxygen       | Oxygen         | ATP     |       |
| NADPH        | NADP⁺          | NADPH   |       |
| NADP⁺        | ATP            | ATP     |       |
| ATP          | NAD⁺           | NAD⁺    |       |
| NADH         | NADH           | NADH    |       |
| Orthophosphate | Orthophosphate | ADP     |       |
|              |                | Pyrophosphate | Coenzyme A |

**Human (KEGG)**

| substr.–prod. | substr.–substr. | substa. | reac. |
|---------------|----------------|---------|-------|
| H⁺           | H⁺             | H⁺      |       |
| H₂O          | H₂O            | H₂O     |       |
| Oxygen       | ATP            | ATP     |       |
| ATP          | Coenzyme A     | Oxygen  | NADP⁺ |
| Orthophosphate | NADP⁺         | NADP⁺   |       |
| NADPH        | Oxygen         | NADPH   |       |
| NAD⁺         | Orthophosphate | Coenzyme A | NAD⁺ |
|              | ADP            | ADP     |       |
|              | UDP            | UDP     |       |
|              | NADH           | NADH    |       |
|              | Pyrophosphate  | Pyrophosphate | CO₂ |

**Human (BiGG)**
### S. cerevisiae KEGG

| substr.–prod. | substr.–substr. | substa. | reac. |
|---------------|----------------|---------|-------|
| H<sub>2</sub>O | H<sub>2</sub>O | H<sub>2</sub>O | H<sub>2</sub>O |
| H<sup>+</sup>  | H<sup>+</sup>  | ATP     | ATP   |
| NADPH         | NADPH         | H<sup>+</sup> | H<sup>+</sup> |
| NADP<sup>+</sup> | NADP<sup>+</sup> | Coenzyme A | Coenzyme A |
| ATP           | ATP           | CO<sub>2</sub> | CO<sub>2</sub> |
| NAD<sup>+</sup> | NAD<sup>+</sup> | Oxygen | Oxygen |
| 2–Oxoglutarate | 2–Oxoglutarate | NADPH | NADPH |
| ADP           | ADP           | NADP<sup>+</sup> | NADP<sup>+</sup> |
| NADPH         | NADPH         | AMP     | AMP   |
| Orthophosphate | Orthophosphate | CO<sub>2</sub> | CO<sub>2</sub> |
| NADP<sup>+</sup> | NADP<sup>+</sup> | NAD<sup>+</sup> | NAD<sup>+</sup> |
| AMP           | AMP           | NADH    | NADH  |
| CO<sub>2</sub> | CO<sub>2</sub> | CoA     | CoA   |
| NH<sub>4</sub>| NH<sub>4</sub>| Oxygen | Oxygen |

### S. cerevisiae BiGG

| substr.–prod. | substr.–substr. | substa. | reac. |
|---------------|----------------|---------|-------|
| H<sub>2</sub>O | H<sub>2</sub>O | H<sub>2</sub>O | H<sub>2</sub>O |
| H<sup>+</sup>  | H<sup>+</sup>  | ATP     | ATP   |
| NADPH         | NADPH         | H<sup>+</sup> | H<sup>+</sup> |
| NADP<sup>+</sup> | NADP<sup>+</sup> | Coenzyme A | Coenzyme A |
| ATP           | ATP           | CO<sub>2</sub> | CO<sub>2</sub> |
| NAD<sup>+</sup> | NAD<sup>+</sup> | Oxygen | Oxygen |
| 2–Oxoglutarate | 2–Oxoglutarate | NADPH | NADPH |
| ADP           | ADP           | NADP<sup>+</sup> | NADP<sup>+</sup> |
| NADPH         | NADPH         | AMP     | AMP   |
| Orthophosphate | Orthophosphate | CO<sub>2</sub> | CO<sub>2</sub> |
| NADP<sup>+</sup> | NADP<sup>+</sup> | NAD<sup>+</sup> | NAD<sup>+</sup> |
| AMP           | AMP           | NADH    | NADH  |
| CO<sub>2</sub> | CO<sub>2</sub> | CoA     | CoA   |
| NH<sub>4</sub>| NH<sub>4</sub>| Oxygen | Oxygen |
Optimal simple-graph representations of metabolism

| substr.–prod. | substr.–substr. | substa. | reac. |
|---------------|----------------|---------|-------|
| H₂O           | H₂O            | H₂O     |       |
| H⁺            | H⁺             | H⁺      |       |
| Oxygen        | ATP            |         |       |
| NADPH         | UDP            |         |       |
| NADP⁺         | CO₂            |         |       |
|               | Coenzyme A    |         |       |

**C. elegans**

| substr.–prod. | substr.–substr. | substa. | reac. |
|---------------|----------------|---------|-------|
| H₂O           | H₂O            | H₂O     |       |
| H⁺            | H⁺             | H⁺      |       |
| Oxygen        | ATP            |         |       |
| NADPH         | Coenzyme A    |         |       |
| NADP⁺         | Oxygen         |         |       |
| ATP           | UDP            |         |       |
| NADH          | CO₂            |         |       |
| NAD⁺          | Orthophosphate |         |       |
|               | ADP             |         |       |
|               | Pyrophosphate  |         |       |
|               | Coenzyme A     |         |       |
|               | CO₂            |         |       |
|               | AMP            |         |       |

**M. musculus**

| substr.–prod. | substr.–substr. | substa. | reac. |
|---------------|----------------|---------|-------|
| H₂O           | H₂O            | H₂O     |       |
| H⁺            | H⁺             | H⁺      |       |
| Oxygen        | ATP            |         |       |
| NADP⁺         | Oxygen         |         |       |
| NADPH         | Coenzyme A    |         |       |
| NAD⁺          | UDP            |         |       |
| NADH          | CO₂            |         |       |
| ATP           | Orthophosphate |         |       |

**R. norvegicus**
### Optimal simple-graph representations of metabolism

| substr.–prod. | substr.–substr. | subst. | reac. |
|---------------|-----------------|--------|-------|
| H₂O           | H₂O             | H₂O    | H₂O   |
| H⁺            | H⁺              | H⁺     | H⁺    |
| Oxygen        | ATP             | ATP    | ATP   |
| NADPH         | Coenzyme A      | Coenzyme A | Coenzyme A |
| NADP⁺         | CO₂             | CO₂    | CO₂   |
| NAD⁺          |                 |        |       |
| NADH          |                 |        |       |
| ATP           |                 |        |       |
| Orthophosphate|                 |        |       |
| ADP           |                 |        |       |

**D. melanogaster**

| substr.–prod. | substr.–substr. | subst. | reac. |
|---------------|-----------------|--------|-------|
| Pyrophosphate | ATP             | ATP    | ATP   |
| ATP           | H₂O             | H₂O    | H₂O   |
| H₂O           | Pyrophosphate   | Pyrophosphate | Pyrophosphate |
|               | ADP             | ADP    | ADP   |

**M. genitalium**

| substr.–prod. | substr.–substr. | subst. | reac. |
|---------------|-----------------|--------|-------|
| H₂O           | H₂O             | H₂O    | H₂O   |
| H⁺            | H⁺              | H⁺     | H⁺    |
| ATP           | ATP             | ATP    | ATP   |
| NAD⁺          | CO₂             | CO₂    | CO₂   |
| NADH          |                 |        |       |
| Orthophosphate|                 |        |       |

**E. coli**
Appendix 3: Cluster–Function Maps

Substance graphs are optimal simple-graph representations of metabolism

Petter Holme and Mikael Huss

This file contains cluster–function maps for the rest of the organisms (corresponding to Fig. 4 of the paper). All figures are for substance graph representations of the various organisms.
The figure for *M. genitalium* looks a bit different from the picture outlined in the text. Note however that this network does not have any patch with a strong coupling to currency metabolites. It is also an outlier in size, \( \Delta \) and other quantities.
Appendix 1:
Degree distributions

Substance graphs are optimal simple-graph representations of metabolism

Petter Holme and Mikael Huss

This file contains degree distributions of the studied organisms and the various graph representations. The data is binned logarithmically. Diamonds represent the original degree distribution, including the currency metabolites. Triangles represent the functionally more relevant network, with the currency metabolites deleted.

Human (KEGG)
Optimal simple-graph representations of metabolism

Human (BiGG)

S. cerevisiae (BiGG)
Optimal simple-graph representations of metabolism

S. cerevisiae (BiGG)

C. elegans
Optimal simple-graph representations of metabolism

$M.\ musculus$

$R.\ norvegicus$
Optimal simple-graph representations of metabolism

---

**D. melanogaster**

---

**M. genitalium**