



## Subnetwork hierarchies of biochemical pathways

Petter Holme<sup>1,\*</sup>, Mikael Huss<sup>2</sup> and Hawoong Jeong<sup>3</sup>

<sup>1</sup>Department of Theoretical Physics, Umeå, 901 87 Umeå, Sweden, <sup>2</sup>SANS, NADA, Royal Institute of Technology, 100 44 Stockholm, Sweden and <sup>3</sup>Department of Physics, Korea Advanced Institute of Science and Technology, Taejeon, 305-701, Korea

Received on May 21, 2002; revised on September 13, 2002; accepted on November 1, 2002

### ABSTRACT

**Motivation:** The vastness and complexity of the biochemical networks that have been mapped out by modern genomics calls for decomposition into subnetworks. Such networks can have inherent non-local features that require the global structure to be taken into account in the decomposition procedure. Furthermore, basic questions such as to what extent the network (graph theoretically) can be said to be built by distinct subnetworks are little studied.

**Results:** We present a method to decompose biochemical networks into subnetworks based on the global geometry of the network. This method enables us to analyze the full hierarchical organization of biochemical networks and is applied to 43 organisms from the WIT database. Two types of biochemical networks are considered: metabolic networks and whole-cellular networks (also including for example information processes). Conceptual and quantitative ways of describing the hierarchical ordering are discussed. The general picture of the metabolic networks arising from our study is that of a few core-clusters centred around the most highly connected substances enclosed by other substances in outer shells, and a few other well-defined subnetworks.

**Availability:** An implementation of our algorithm and other programs for analyzing the data is available from <http://www.tp.umu.se/forskning/networks/meta/>

**Supplementary information:** Supplementary material is available at <http://www.tp.umu.se/forskning/networks/meta/>

**Contact:** holme@tp.umu.se; hussm@kth.se; hjeong@nd.edu.

### INTRODUCTION

In the last few years, several studies have addressed graph theoretical aspects of biochemical networks (see for example, Schuster and Hilgetag, 1994; Schuster *et al.*, 1999, 2002; Jeong *et al.*, 2000, 2001; Fell and

Wagner, 2000; Wagner and Fell, 2001). This, the coarsest level of describing cellular biochemistry, is a valuable complement to more detailed studies in that it can shed light on the global organization of biochemical networks (cf. Wagner and Fell, 2001). Besides the findings of universal graph-structural properties, such methods have been used to identify arguably biologically significant subnetworks (Schuster *et al.*, 2002). The desire for finding subnetworks arises from the vastness of the biochemical networks; even a prokaryote such as *E.coli* has a metabolism involving over 850 substances and 1500 reactions. Attempts to elucidate for example a bacterium's metabolic repertoire, thus face the problem of combinatorial explosion. A fundamental question is therefore what the hierarchical organization of subnetworks looks like. Can relevant subnetworks be found at arbitrary sizes? Is it at all relevant to talk of subnetworks, or must the whole network always be taken into account? This paper aims to answer these questions by proposing a general method for partitioning a biochemical network into subnetworks by successively removing reactions of high betweenness centrality—reactions situated between areas of many interior pathways (i.e. well-defined subnetworks). Besides finding explicit subnetworks of arbitrary sizes, this method also enables us to investigate the full hierarchical organization of a cellular network—how a subnetwork can be divided into sub-subnetworks and so on.

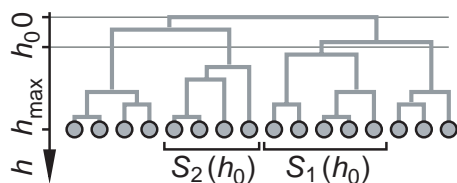
### TRACING THE HIERARCHICAL SUBNETWORK STRUCTURE

#### Networks

We represent the metabolic network as a directed bipartite graph  $G = (S, R, L)$  where  $S$  is the set of nodes representing substrates,  $R$  is the set of nodes representing chemical reactions, and  $L$  is the set of directed links—ordered pairs of one node in  $S$  and one node in  $R$ .  $s_1, \dots, s_n \in S$  is involved in a reaction  $r \in R$  with products  $s'_1, \dots, s'_{n'}$  if and only if  $(s_1, r), \dots, (s_n, r) \in L$  and  $(r, s'_1), \dots, (r, s'_{n'}) \in L$ .

The networks we use (the same data set as in Jeong

\*To whom correspondence should be addressed.



**Fig. 1.** A simple hierarchical clustering tree. A horizontal cut gives the tighter connected subgraphs below, and looser connections above.  $S_i(h_0)$  is the size of the  $i$ th largest connected subgraph at height  $h_0$ . Note that the root is at the top and  $h$  grows downwards.

*et al.*, 2001) were constructed from the WIT database<sup>†</sup> (Overbeek *et al.*, 2000) consisting of 43 organisms from all domains of life—6 archae, 32 bacteria, and 5 eukaryotes. We distinguish between metabolic networks and whole-cellular networks—networks representing the full set of cellular pathways. WIT divides the latter into the following subcategories: intermediate metabolism and bioenergetics, information pathway, electron transport, transmembrane transport, signal transduction and structure and function of cell, of which intermediate metabolism and bioenergetics constitute the metabolic networks. Due to the current pace of database development our data is somewhat aged. Furthermore the signal transduction and transport part of our data is limited, and genetic regulation data missing (so the term ‘whole-cell network’ is a rather crude overstatement). Nevertheless, the big picture should be fairly insensitive to discrepancies in the database (see also discussions in the works cited in the **Introduction**), which adds confidence to our analysis of the global organization of the networks (which along with the method itself is the main theme of the paper).

### Decomposition algorithms

The standard set of metabolic pathways described in biological literature is sometimes too rigid to capture the essence of what is happening in an organism; the sub-systems identified in this way will often to a large extent overlap and intertwine. As a complement to the traditional approach there is a need for unbiased analytic methods, such as the one proposed by Schuster *et al.* (2002), a method for decomposing a biochemical network into sub-networks based on the degree of the metabolites.<sup>‡</sup> The idea of Schuster *et al.* is to label metabolites with degree  $k$  larger than some threshold value  $k_{\max}$  as ‘external’—as ei-

ther a source or sink, and then consider connected components of ‘internal’ metabolites as subnetworks, the motivation being that the system can be regarded as buffered with respect to the substrates participating in the largest number of reactions (Fell and Wagner, 2000). If the relabelling of an internal node as external is interpreted as deleting the node from the network of internal nodes, this method is equivalent to the attack vulnerability study of Albert *et al.* (2000), where networks response to the removal of nodes in order of degree was discussed.

The heuristic motivations and *a posteriori* success (for subnetwork detection) of the method of Schuster *et al.* makes it an important contribution to biochemical pathway analysis. A potential drawback of this method is that networks might have inherent non-local features not possible to be detected by local quantities such as degree. (Non-local effects are known to be an important factor in for example social networks; see Granovetter, 1973.) To be specific, consider a node  $m$  with degree  $k_m > k_{\max}$  neighbours, all except one having  $k = 1$ . This is a local centre but globally (for a large enough network) a peripheral node. Then all these  $k = 1$ -nodes would be classified as belonging to individual one-node clusters, while a more informative categorization would be to have the  $k = 1$  and  $k = k_m$  nodes in the same cluster. Indeed, this kind of configuration actually exists, which will be shown later. The method for identifying subnetworks (presented in the following two sections) is similar to that of Schuster *et al.* only that our algorithm removes reaction nodes rather than substrates, and the removal is based on a global centrality measure (betweenness) rather than a local (degree).

### Constructing hierarchies of subnetworks

Structural studies of networks have a long history in sociology, and many methods and concepts can be brought over to biological network studies as well. The traditional way of detecting hierarchies of subnetworks (or the ‘community structure’) in social networks has been by hierarchical clustering methods (Johnson, 1976), where one reconstructs the network by adding links in order of some measure of strength. In this way one can construct a hierarchy tree, where the tightest connected subgraphs are joined by links close to the root of the tree, and the most long-range inter-community links are close to the root (see Fig. 1). However, these hierarchical clustering methods have some inherent flaws. For example, just as by Schuster’s method above, nodes with one neighbour often become classified as belonging to a one-node cluster. In remedy, Girvan and Newman (2002) proposed an elegant method where one deconstructs the network by successively deleting links carrying many shortest paths—which are likely to lie between tightly connected subnetworks.

<sup>†</sup> Similar information can be obtained from the KEGG (Kanehisa and Goto, 2000), EcoCyc (Karp *et al.*, 2000) and EMP (Selkov *et al.*, 1996) databases.

<sup>‡</sup> Degree is the number of neighbours to a node of a graph. (Sometimes degree is called connectivity.)

Another potentially interesting cluster identification algorithm applied to backbone clusters of residues in proteins is presented in Patra and Vishveshwara (2000).

## Betweenness centrality

For an undirected graph the betweenness centrality  $C_B$  (Freeman, 1977)—or for short, betweenness—of a node  $v$  is the number of shortest paths between pairs of nodes that pass  $v$  (if more than one shortest path exist between  $u$  and  $u'$  passes  $v$ , the fraction of shortest paths through  $v$  contributes to its betweenness). For the purposes of this work we are interested in reaction nodes that are central for paths between metabolites or other molecules; thus we redefine the betweenness centrality of reaction nodes as follows—for  $r \in R$ :

$$C_B(r) = \sum_{m \in M} \sum_{m' \in M \setminus \{m\}} \frac{\sigma_{mm'}(r)}{\sigma_{mm'}}, \quad (1)$$

where  $\sigma_{mm'}(r)$  is the number of shortest paths between  $m$  and  $m'$  that passes through  $r$ , and  $\sigma_{mm'}$  is the total number of shortest paths between  $m$  and  $m'$ . For calculating betweenness we use the fast algorithm of Brandes (2001).

In a biochemical context, a node of high betweenness will thus represent a bottleneck in the flow between highly connected (more independently functioning) regions. Since the betweenness does not contain reaction kinetic information, a high betweenness does not necessarily mean that the reaction is frequent (and it should not—then high betweenness would not correspond to being situated between highly connected areas).

## Our algorithm

As we represent the biochemical network as a bipartite graph (so that all substances—metabolites, macromolecules, complexes, etc.—are separated by reaction nodes and an even number of links), we modify the algorithm of Girvan and Newman (2002) and successively delete reaction nodes with high betweenness with respect to substrates/products and enzymes. A reaction corresponds to a passage through all inward links to a reaction node. To take this into account we regard the effectiveness of a link in the betweenness definition as proportional to  $1/k_{\text{in}}(r)$  ( $k_{\text{in}}(r)$  being the in-degree, or number of substrates to a reaction). The rescaled, effective betweenness thus becomes:

$$c_B(r) = C_B(r)/k_{\text{in}}(r) \quad (2)$$

With these definitions the algorithm comes closer to that of Girvan and Newman (2002), where link betweenness is the operative centrality measure— $c_B(r)$  could also be interpreted as the average link betweenness of  $r$ 's inward links.

With the above modifications, the algorithm consists of the following steps repeated until no reaction nodes remain:

- (1) calculate the effective betweenness  $c_B(r)$  for all reaction nodes;

- (2) remove the reaction node with highest effective betweenness and all its in- and out-going links;
- (3) save information about the current state of the network (such as how many clusters there are, and what nodes that belongs to a specific cluster).

If many reaction nodes have the same betweenness in step , we remove all of them at once.<sup>§</sup> The information obtained in step is used to construct the hierarchy trees and statistics about the ordering. The worst case running time of the algorithm is  $O(M N_S N_R)$ , where  $N_S$  is the number of substrates and  $N_R$  is the number of reaction nodes (cf. Girvan and Newman, 2002).

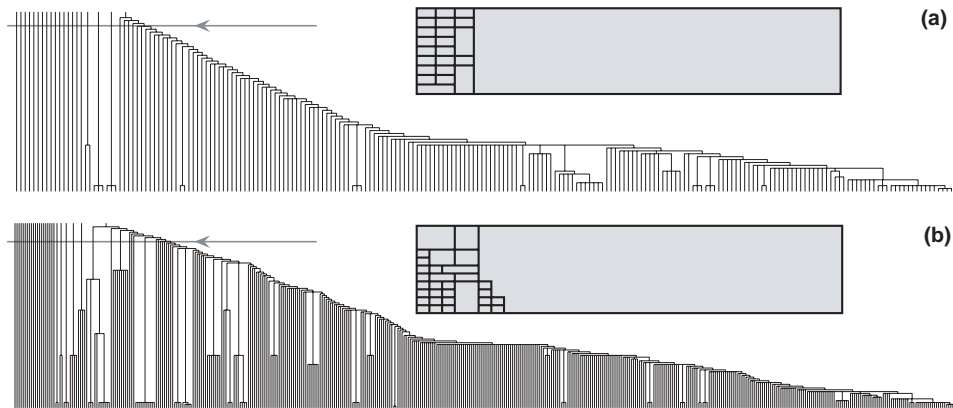
## RESULTS

### General shape of the hierarchy trees

As an example of hierarchical clustering trees (Fig. 2), of metabolic and whole-cellular networks we consider those of the bacterium *Treponema pallidum* (*T.pallidum*).<sup>¶</sup> (*T.pallidum* is the pathological agent of syphilis. A recent review of its functions in a genomic perspective is given in Norris *et al.* 2001.) Most constituents are connected into a giant component (a cluster whose size scales linearly with the total number of nodes, see for example, Janson *et al.*, 2000). Close to the root (Fig. 2, top) the giant component is still existent, but at height  $h \approx 0.8 h_{\text{max}}$  of the hierarchy tree the giant component starts to break into well-defined clusters. When a cluster breaks into subclusters of similar sizes, we say the hierarchy tree has ‘community-type ordering’ at the hierarchy level in question (Fig. 3a). Biochemically this means that the mass flow within the subclusters are more complexly and densely routed than between them, and the function of the original cluster can be understood as composed by relatively distinct modules (i.e. the subclusters). When, on the other hand, the cluster breaks into one large subcluster and many isolated nodes the level has a ‘shell-type ordering’ (Fig. 3b) where the outer nodes often act as in/out-flow or buffers to the core-cluster, rather than being involved in complex interconversion processes. These concepts designed to signify the extreme cases of a most or least symmetric splitting of the clusters, intermediate cases where the cluster splits into subclusters of various sizes can of course also occur. In the real hierarchy trees we study, shell-type ordering is frequent in the whole tree, and dominates the levels closest to the root (with small  $h$ ). Community-type ordering, on the other hand, is only frequent at high  $h$ . This lack of a community-type

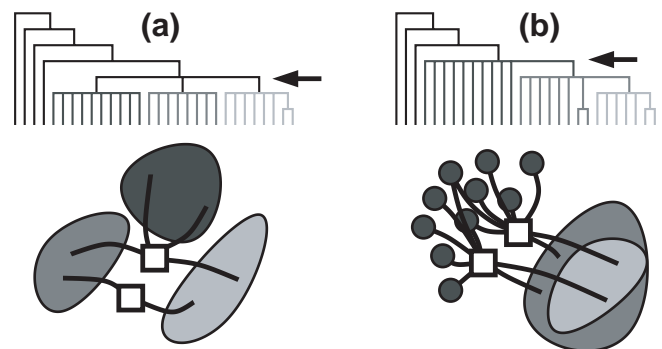
<sup>§</sup> This is needed to make the algorithm consistent and deterministic, and the A C-implementation of the algorithm, along with a test data set can be obtained from <http://www.tp.umu.se/forskning/networks/meta/>.

<sup>¶</sup> Hierarchical clustering trees of the other 43 organisms of the WIT database (Overbeek *et al.*, 2000) used in this study can be seen on <http://www.tp.umu.se/forskning/networks/meta/>.



**Fig. 2.** The hierarchical clustering trees of *T. pallidum*. (a) Shows the tree for the metabolic network, (b) shows the whole-cellular network. The squares represent the subnetwork configuration at  $h = 0.1 h_{\max}$  (the height indicated by the arrow). Sizes of the squares are proportional to the size of the clusters they represent.

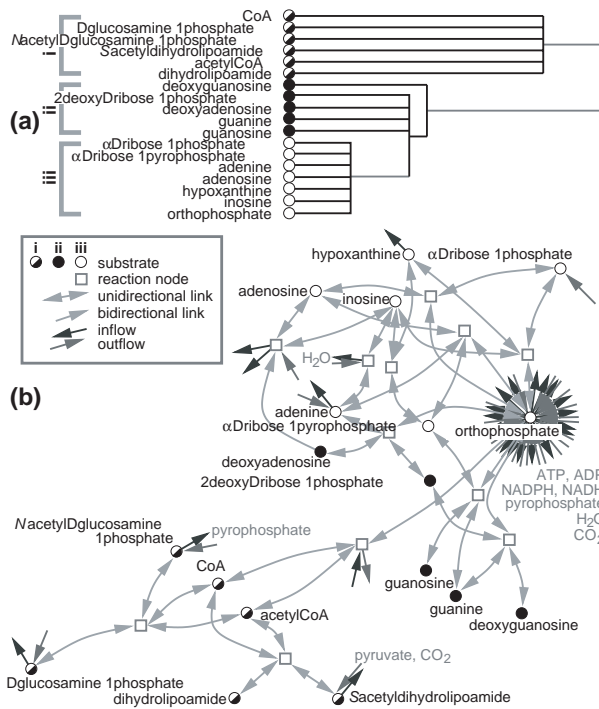
ordering close to the root of the tree (such as seen in for example social networks and ecological food-webs; Girvan and Newman 2002) is related to the highly heterogeneous centrality distribution (Jeong *et al.*, 2001) of cellular biochemical networks: The giant component is tightly connected by the many paths involving the most connected substances, ATP, NADH,  $H_2O$ , and so on. The core of the metabolism is centred around these most connected substances, hence most well-defined subnetworks must contain these, but this is precisely to say that the subnetwork containing these is sure to dominate most levels of the organization. Community-type ordering occurs when either none of these most connected substances is central in some subnetwork, or when these substances fall into different subnetworks. Both of these cases occur in for example *T. pallidum* (Fig. 2; a functional description of these subnetworks will be given later): *N*-acetyl-D-glucosamine 1-phosphate, D-glucosamine 1-phosphate, dihydrolipoamide, *S*-acetyldihydrolipoamide, CoA, and acetyl-CoA define small subnetworks not including any of the most connected substances (Fig. 4). The highly connected orthophosphate and the substances most tightly connected to it ( $\alpha$ -D-ribose 1-phosphate,  $\alpha$ -D-ribose 1-phosphate adenine, adenosine, hypoxanthine, and inosine) define another subnetwork that at a higher level of organization (lower  $h$ ) is joined by other substances (2-deoxy-D-ribose 1-phosphate, deoxyadenosine, guanine, and guanosine) to a more loosely connected subnetwork. The general picture that arises from the study of hierarchy trees is thus that the cellular biochemical networks consist of outer shells encapsulating a core of the most connected substances, with a few well-defined subnetworks at an intermediate level of the organization. This picture is (more or less) the same for all the 43 organisms examined.



**Fig. 3.** Schematic picture of the two different orderings in hierarchy trees. (a) Community-type ordering—same level core-clusters connected by outer parts of the network. (b) Shell-type—a sequence of core-clusters contained in each other. The squares symbolizes the reaction nodes that are deleted at the height marked by the arrow. In (a) three subnetworks of similar sizes gets disconnected when the reaction nodes are removed. In (b) many individual metabolite nodes (circles) get isolated.

### Statistics of the hierarchical ordering

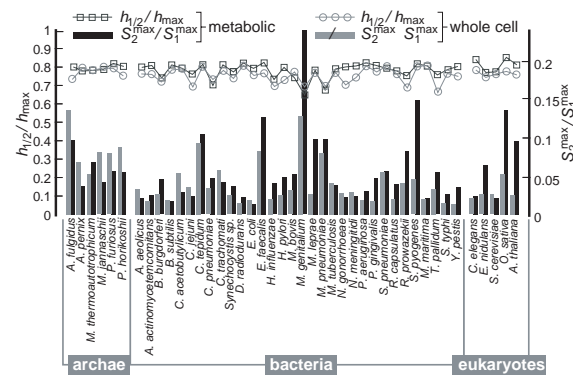
An immediate impression from looking at the hierarchy trees of the 43 WIT organisms is that they are similar in the large scale and more diverse locally. Furthermore, the shell-type ordering dominates much of the small- $h$  region of metabolic networks, whereas in the whole cell networks, community-type ordering also occurs close to the root. This section aims to quantify these observations. A third observation is that, as  $h$  increases, the tree often splits in a non-uniform way, so that only individual nodes or smaller clusters are removed from the largest connected component at a time.



**Fig. 4.** Subnetworks from the metabolic networks of *T.pallidum*. (a) Shows the part of the hierarchy tree that (b) corresponds to. Grey substance names show where shortest paths to the hubs (most connected substances of the network) enter.

The large-scale shape of the tree can be measured in many ways. One simple and informative quantity is the half-height of the largest cluster  $h_{1/2}$ , i.e. the height  $h$  where the size of the largest cluster  $S_1$  has decreased to half of its original value (For notations see Fig. 1). If uniform ordering, where clusters break into clusters of similar sizes, would dominate all levels of organization (such as the examples from sociology and ecology in Girvan and Newman, 2002) we expect a very small relative half-height  $h_{1/2}/h_{max}$  ( $h_{1/2} \propto \log h_{max}$ ). In Figure 5 the relative height of the tree where  $S_1$  has decreased to half of its original value,  $h_{1/2}/h_{max}$  is displayed. Averaged over all 43 organisms this happens at  $h = 0.79(4) h_{max}$  for the metabolic networks and  $h = 0.76(4) h_{max}$  for the whole-cellular networks—a very narrow region suggesting an universal behaviour (see Figure 5). Even though the almost constant  $h_{1/2}/h_{max}$  is not trivially related to other universal features, such as the constant average shortest-path length (Jeong et al., 2001), it is consistent with the general picture of a great diversity of organisms having a very similar large-scale organization of the biochemical pathways.

To measure the magnitude of community-type ordering we study the  $h$ -evolution of the size of the second largest

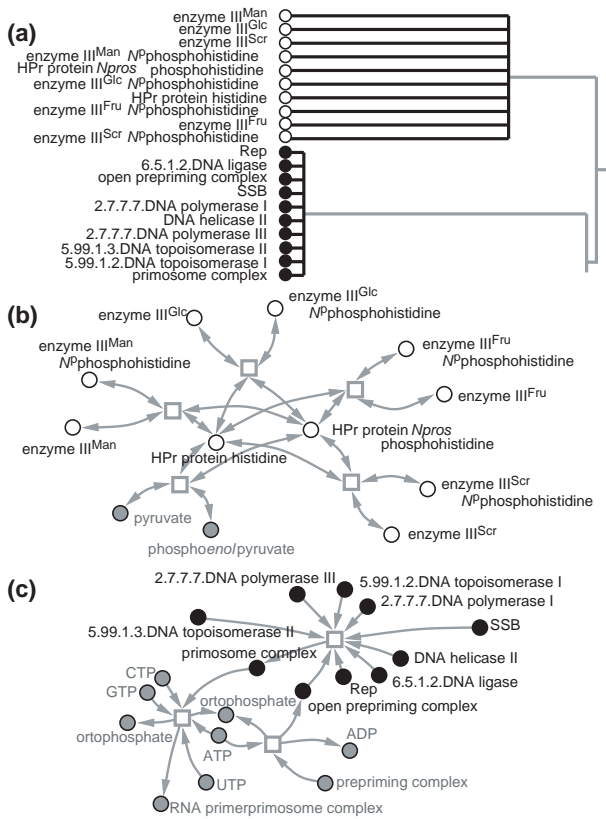


**Fig. 5.** The relative size of the network  $N/N_{max}$ ; the ratio between the largest values of the second largest and largest connected subgraphs  $S_2^{max}/S_1^{max}$ ; and the relative half-height  $h_{1/2}/h_{max}$  for the 43 studied organisms.

cluster  $S_2$ . In Figure 5  $S_2^{max}/S_1^{max}$  is displayed for all 43 organisms. A large value of this quantity means that the network at some time has at least two subnetworks of a large and similar size and thus a pronounced community-type order. Networks with a high  $S_2^{max}/S_1^{max}$  in fact also have a high  $S_3^{max}/S_1^{max}$  and so on, so this quantity works well as a measure of the degree of community-type order. Figure 5 shows that there is a large variance in  $S_2^{max}/S_1^{max}$  (with  $S_2^{max}/S_1^{max} = 0.06(5)$  for the metabolic networks and  $S_2^{max}/S_1^{max} = 0.05(3)$  for the whole-cellular networks). Although larger databases would be needed to obtain statistical certainty,  $S_2^{max}/S_1^{max}$  orders the organisms as archae > bacteria > eukaryotes. It is interesting to note that eukaryotes have the lowest value. A more uniform (small  $S_2$ ) organization is more robust, which suggests that eukaryote biochemical networks are more robust than those in bacteria and archae.

### Detected subnetworks

To give an explicit example how community- and shell-ordering are manifested in a metabolic network, we consider two small subnetworks at  $h = 40$  of the hierarchy tree of *T.pallidum*'s metabolic network (Fig. 2) shown in Figure 4. These subnetworks contains reactions associated with purine metabolism and pyruvate/acetyl-CoA conversion. The pyruvate-acetyl-CoA part ((i) in the hierarchy tree, Roman numbers refer to Figure 4a) is a tightly interconnected, fairly independent subnetwork, while the purine metabolism part consists of an outer shell (ii) encapsulating a smaller core (iii), which is centred around orthophosphate and has to do with interconversions between adenosine and related nucleosides. Deoxyadenosine ends up in the outer shell (ii) while adenosine is located in the core (iii) because there are two reactions involving adenosine and orthophosphate, but only one with de-



**Fig. 6.** An excerpt of the hierarchy tree of (a) *M.pneumoniae* and the corresponding subnetworks associated with sugar import and (b) DNA replication (c). Symbols are the same as in Figure 4, except that nodes adjacent to the two clusters are marked with grey circles. Of the grey nodes' links, only those between the grey node and the cluster are shown. One cluster is three reactions away from the other by, for example, the reaction phosphoenolpyruvate + ADP → pyruvate + ATP.

oxyadenosine and orthophosphate. Overall, however, the subnetworks in this case can be readily seen to represent metabolic processes of increasing homogeneity.

We exemplify whole-cellular networks of the bacterium *Mycoplasma pneumoniae* (*M.pneumoniae*), a bacterium causing respiratory tract infections (see the review by Principi and Esposito, 2001) in Figure 6. In general the whole-cellular networks are, perhaps not surprisingly, even more functionally distinct than the metabolic networks. One of the subnetworks (Fig. 6b) is a part of the bacterial phosphotransferase system, the function of which is to import carbohydrates into the cell (see Saier, 2001, for an overview). Enzyme III is an older collective name for enzyme IIA and enzyme IIB. Each of these enzymes is specific for a certain kind of carbohydrate; in Figure 6b, we see enzymes specific for mannitol, glucose, sucrose and fructose, respectively. The other

network (Fig. 6c) has to do with DNA replication. (see for example, Lewin, 1997, for an overview of DNA replication). Although both of these subnetwork differ from the metabolic networks in that the nodes are not metabolites which are interconverted but rather enzymes that interact with or are part of macromolecular complexes, they both nevertheless represent biologically meaningful groups of substances. Furthermore, the DNA replication subnetwork is centred around a reaction node with high degree (local centrality), but relatively low betweenness (global centrality). Thus local, degree-based, algorithms would have difficulties identifying such a subnetwork. Note that the subnetwork of Figure 6b is ordered higher (is connected at a lower  $h$ ) in the hierarchy than that in Figure 6c since the reversibility of the reactions in Figure 6b increases their betweenness.

## SUMMARY AND DISCUSSION

We propose an algorithm for decomposing biochemical networks into subnetworks based on the global network structure. The algorithm—a development of the algorithms by Girvan and Newman (2002), and Schuster *et al.* (2002)—is purely graph theoretical and uses no biological criteria (cf. Schilling and Palsson, 2000). The data we study are the sets of metabolic and whole-cellular networks of 43 organisms (archae, bacteria and eukaryotes) from the WIT database. We emphasize the study of hierarchy trees to get a general view on the organization of subnetworks. To characterize the hierarchical organization we introduce (to the biochemical network studies) the concepts of community- and shell-ordering and quantitative measures (the relative half-height  $h_{1/2}/h_{\max}$ , and the relative largest size of the second largest cluster  $S_2^{\max}/S_1^{\max}$ ).

The large-scale shape of biochemical network trees is conspicuously uniform among organisms as manifested in a universal relative half-height ( $h_{1/2} = 0.79(4) h_{\max}$  for the metabolic networks and  $h_{1/2} = 0.76(4) h_{\max}$  for the whole-cellular networks). The spread in  $S_2^{\max}/S_1^{\max}$  shows that community-ordering is much more pronounced in some organisms than in others. A small  $S_2^{\max}/S_1^{\max}$  implies a more robust network, which means that it is a quantity of potential interest for evolutionary studies (when the databases reach a size where sufficiently good statistics can be generated).

Well-defined subnetworks occur at different levels of organization (at different heights in the hierarchy tree). This is a strong argument for looking at the whole hierarchy tree rather than the subnetwork configuration at a specific level. For the metabolic networks, the dominating structure at most levels of organization is the largest connected component. For the whole-cellular network, non-metabolic subnetworks, such as those representing

information pathways, signal transduction and the like, are often branched off from the metabolic circuitry close to the root of the hierarchy tree. Still, the largest component is dominant over a large portion of the tree's levels. The general picture of the hierarchical organization that emerges from our study is thus that biochemical networks have individual core-clusters of the most connected substances and their closest related substances; the rest of the substances are then organized as a more and more loosely connected outer shell, with the exception of some well-defined clusters at intermediate levels. The fact that one cluster dominates most of the organization suggests that in some contexts it might be deceptive to generalize properties of subnetworks to the whole network. Thus, there might be global effects that will be overlooked in a purely modular description—so for a comprehensive picture of biochemical networks, one needs to integrate studies at all levels of description.

## ACKNOWLEDGEMENTS

The authors thanks P. Minnhagen for fruitful discussions. P.H. acknowledges partial support from the Swedish Natural Research Council through Contract No. F5102-659/2001. H.J. acknowledges financial support from the Ministry of Information and Communication of Korea through IMT2000-B3-2. Drawing of the figures was aided by the Pajek package (<http://vlado.fmf.uni-lj.si/pub/networks/pajek/>).

## REFERENCES

- Albert,R., Jeong,H. and Barabási,A.-L. (2000) Error and attack tolerance of complex networks. *Nature*, **406**, 378–381.
- Brandes,U. (2001) A faster algorithm for betweenness centrality. *J. Math. Sociol.*, **25**, 163–177.
- Fell,D.A. and Wagner,A. (2000) The small world of metabolism. *Nature Biotechnol.*, **18**, 1121–1122.
- Freeman,M.L. (1977) A set of measures of centrality based upon betweenness. *Sociometry*, **40**, 35–41.
- Granovetter,M.S. (1973) The strength of weak ties. *Am. J. Sociol.*, **78**, 1360–1380.
- Girvan,M. and Newman,M.E.J. (2002) Community structure in social and biological networks. *Proc. Natl Acad. Sci. USA*, **99**, 7821–7826.
- Janson,S., Łuczac,T. and Ruciński,A. (2000) *Random Graphs*. Wiley, New York, pp. 107–126.
- Jeong,H., Tombor,B., Albert,R., Oltvai,Z.N. and Barabási,A.-L. (2000) The large-scale organization of metabolic networks. *Nature*, **407**, 651–654.
- Jeong,H., Mason,S.P., Barabási,A.-L. and Oltvai,Z.N. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.
- Johnson,S.C. (1976) Hierarchical clustering schemes. *Psychometrika*, **32**, 241–253.
- Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
- Karp,P.D. et al. (2000) The EcoCyc and MetaCyc databases. *Nucleic Acids Res.*, **28**, 56–59.
- Lewin,B.M. (1997) *Genes VI*, 6th edn, Oxford University Press, Oxford.
- Norris,S.J., Cox,D.L. and Weinstock,G.M. (2001) Biology of *Treponema pallidum*: correlation of functional activities with genome sequence data. *J. Mol. Microb. Biotechnol.*, **3**, 37–62.
- Overbeek,R. et al. (2000) WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res.*, **28**, 123–125.
- Patra,S.M. and Vishveshwara,S. (2000) Backbone cluster identification in proteins by a graph theoretical method. *Biophys. Chem.*, **84**, 13–25.
- Principi,N. and Esposito,S. (2001) Emerging role of *Mycoplasma pneumoniae* and *Chlamydia pneumoniae* in paediatric respiratory-tract infections. *Lancet Infect. Dis.*, **1**, 334–344.
- Saier,Jr,M.H. (2001) The bacterial phosphotransferase system: structure, function, regulation and evolution. *J. Mol. Microbiol. Biotechnol.*, **3**, 325–327.
- Schilling,C.H. and Palsson,B.Ø. (2000) Assessment of the metabolic capabilities of *Haemophilus Influenzae* Rd through a genome-scale pathway analysis. *J. Theor. Biol.*, **203**, 229–248.
- Schuster,S. and Hilgetag,C. (1994) On elementary flux modes in biochemical reaction systems at steady state. *J. Biol. Syst.*, **2**, 165–182.
- Schuster,S., Dandekar,T. and Fell,D. (1999) Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol.*, **17**, 53–60.
- Schuster,S., Pfeiffer,T., Moldenhauer,F., Koch,I. and Dandekar,T. (2002) Exploring the pathway structure of metabolism: Decomposition into subnetworks and application to *Mycoplasma pneumoniae*. *Bioinformatics*, **18**, 351–361.
- Selkov,E. et al. (1996) The metabolic pathway collection of EMP: the enzymes and metabolic pathway database. *Nucleic Acids Res.*, **24**, 26–28.
- Wagner,A. and Fell,D.A. (2001) The small world inside large metabolic networks. *Proc. R. Soc. Lond. B*, **268**, 1803–1810.