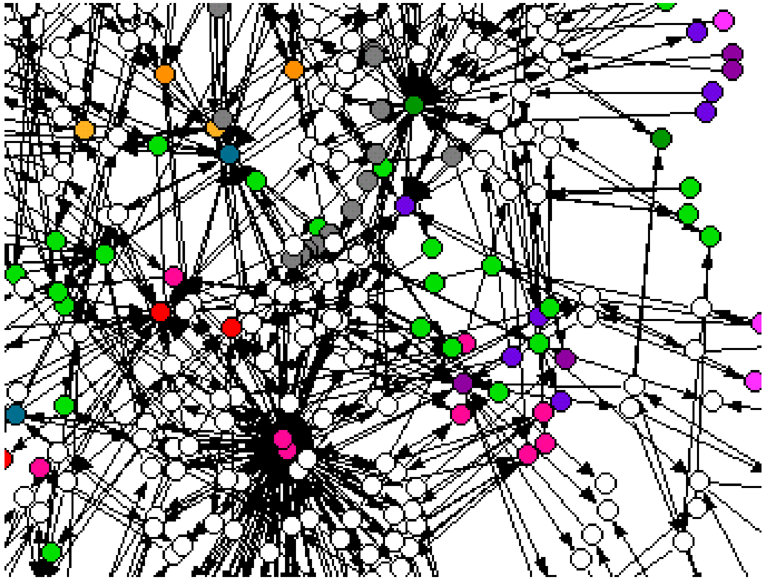
Discovery and analysis of biochemical subnetwork hierarchies

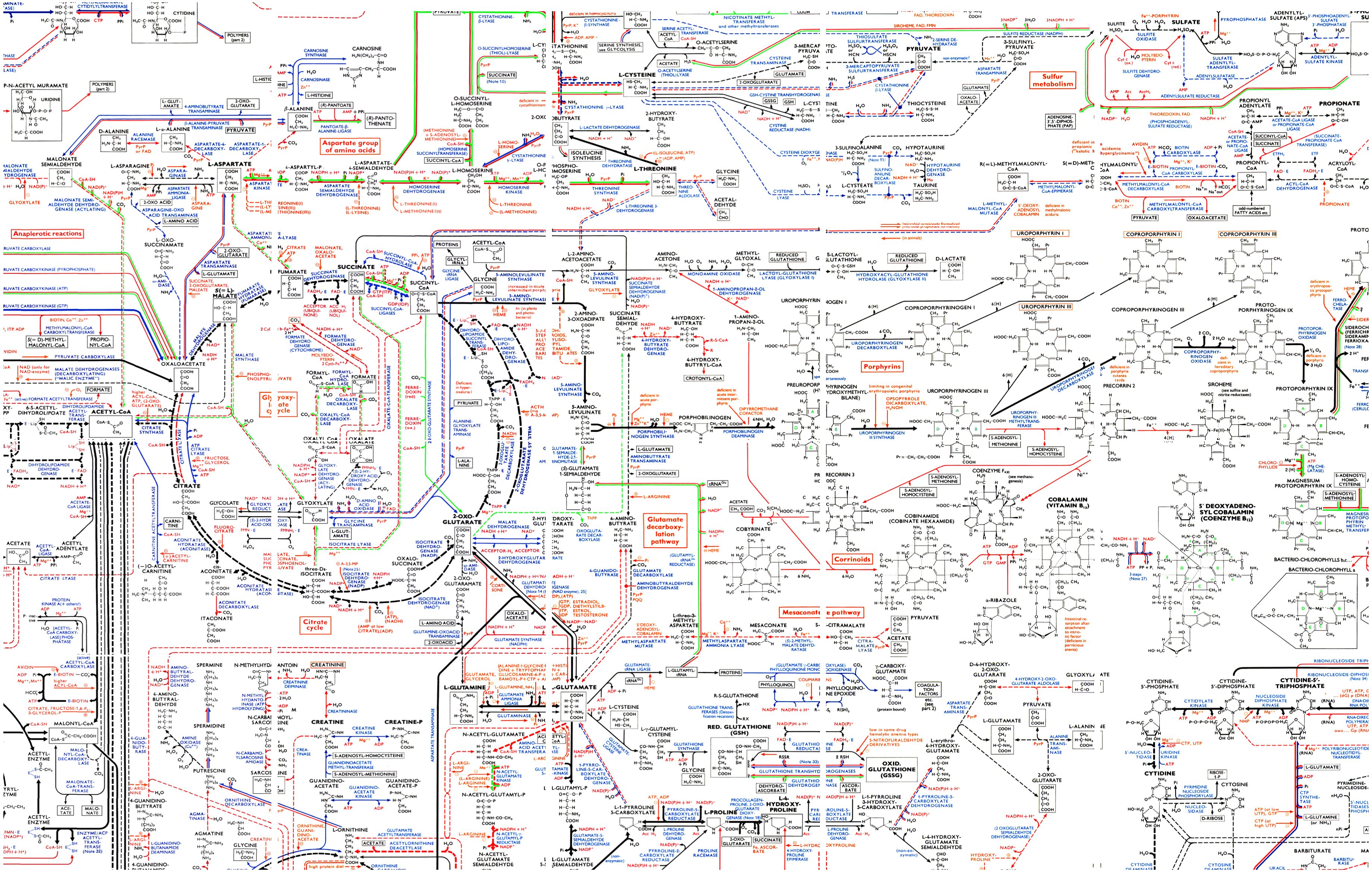
October 6, 2003

Petter Holme
Department of Physics, Umeå
University, Umeå, Sweden
NORDITA, Copenhagen, Denmark

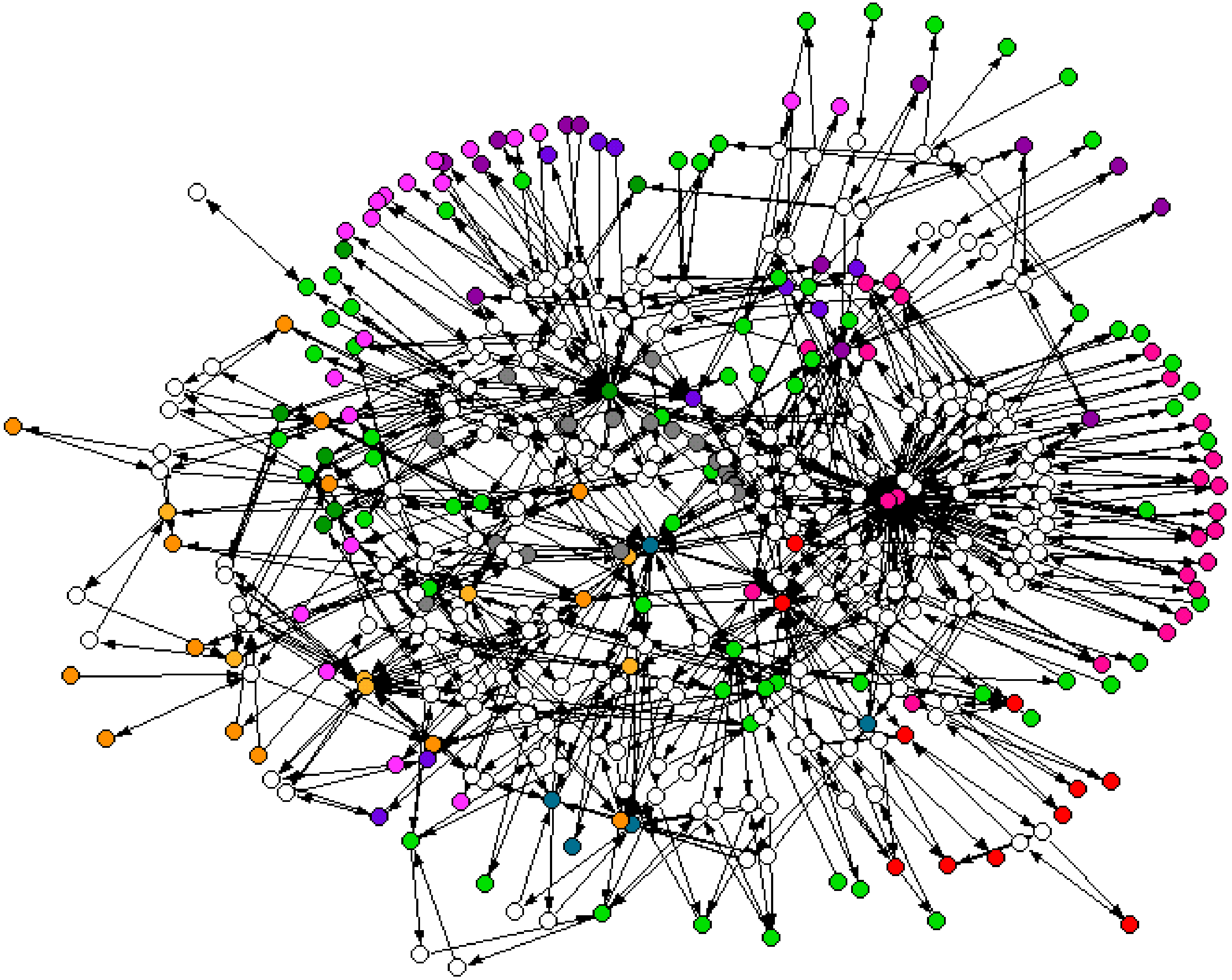
Mikael Huss
SANS, NADA, Royal Institute of
Technology, Stockholm, Sweden



Biochemical networks: The classical view



BIOCHEMICAL PATHWAYS AS GRAPHS



metabolic pathways of *Borrelia burgdorferi* (a bacterium)

Motivation

Complexity

Even *E. coli* has a metabolism involving over 850 substances and 1500 reactions ⇒

- The coarsest level of description—the graph representation—is needed, at least as a complement.
- One would like to decompose the graph into functional subunits. Both for conceptual and analytical purposes.

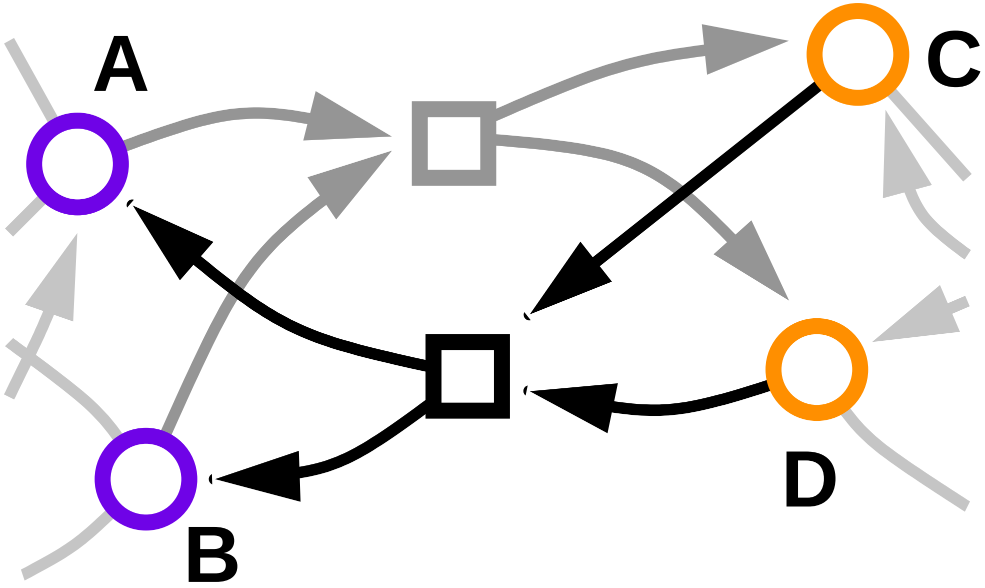
Our work

- Earlier algorithms have been based on local algorithms that may miss some large-scale features.
- Not much known about how the large scale subnetwork ordering looks like. Can the network easily be decomposed into autonomous subnetworks? How independent are the modules? Is it useful to talk about modules at all?

The basic assumption

If we find a subnetwork that is well-connected within, and sparsely connected to the outside, then it is likely to be a relatively autonomously functioning subnetwork.

BIOCHEMICAL NETWORKS AS DIRECTED BIPARTITE NETWORKS



The reaction A + B ↔ C + D in a directed bipartite representation:

- Two types of vertices, representing substrates and chemical reactions.
- Arcs (arrows) between different types of vertices

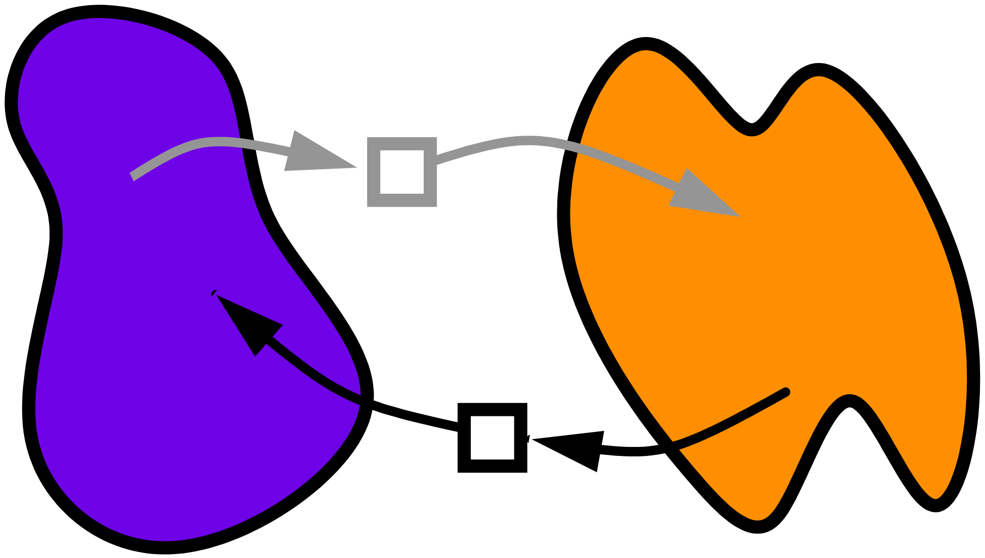
We denote the set of chemical substances by S and the set of reaction vertices by R.

The cluster detection algorithm

Based on: M. Girvan & M. Newman, PNAS **99** (2002), pp. 7821-7826.

Presented in: P. Holme, M. Huss, and H. Jeong, Bioinformatics **19** (2003), pp. 532-538.

The idea Recursively delete reactions situated between densely connected regions.



1. Calculate the effective betweenness $c_B(r)$ for all reaction vertices.
2. Remove the reaction vertex with highest effective betweenness and all its in- and out-going links.
3. Save information about the current state of the network.

The cluster detection algorithm *(continued)*

Let C_B be the *betweenness* of r with respect to the substance-vertices.

$$C_B(r) = \sum_{s\in S} \sum_{s'\in S\setminus\{s\}} \frac{\sigma_{ss'}(r)}{\sigma_{ss'}}\ , \tag{1}$$

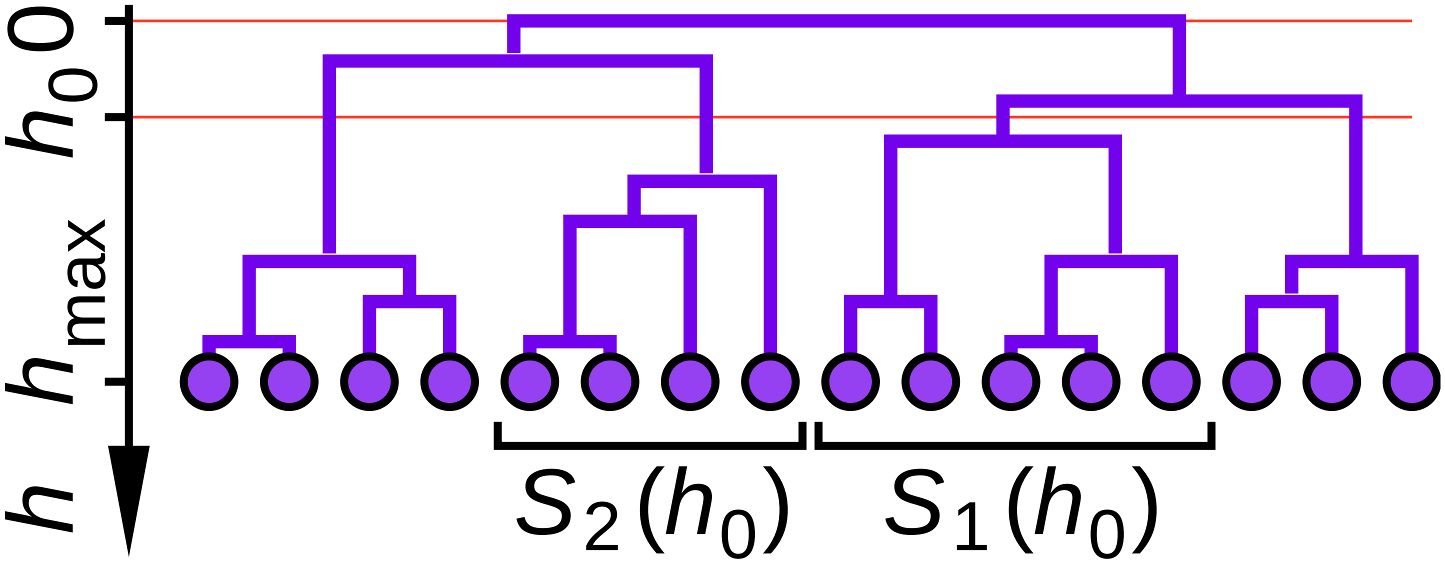
where $\sigma_{ss'}(r)$ is the number of shortest paths between s and s' that passes through r, and $\sigma_{ss'}$ is the total number of shortest paths between s and s'.

The reactions we delete recursively are the one having the highest *effective betweenness*:

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

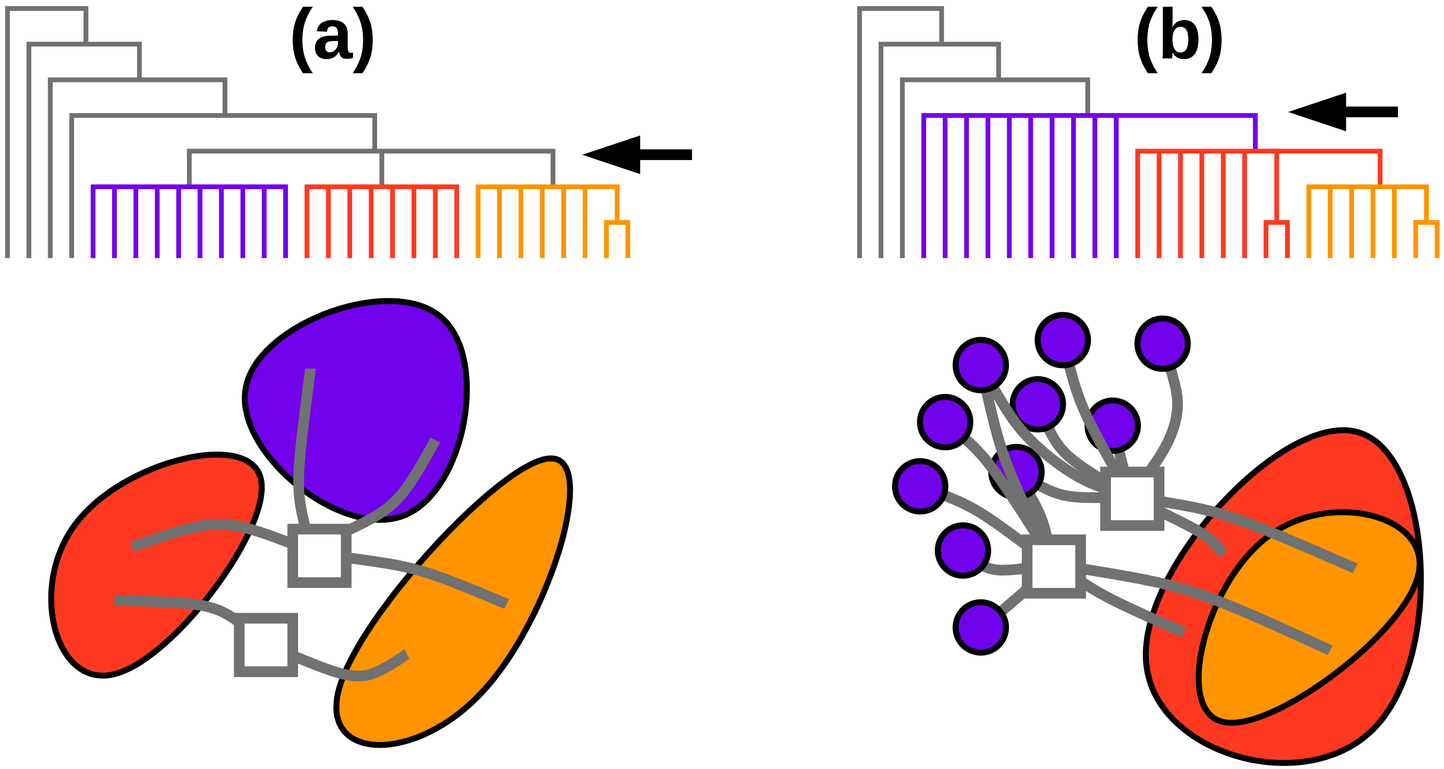
where $k_{\text{in}}(r)$ is the in-degree (# of substrates) of the reaction r. This rescaling is sensible since all substrates needs to be present for a reaction to occur.

SUBNETWORK HIERARCHIES



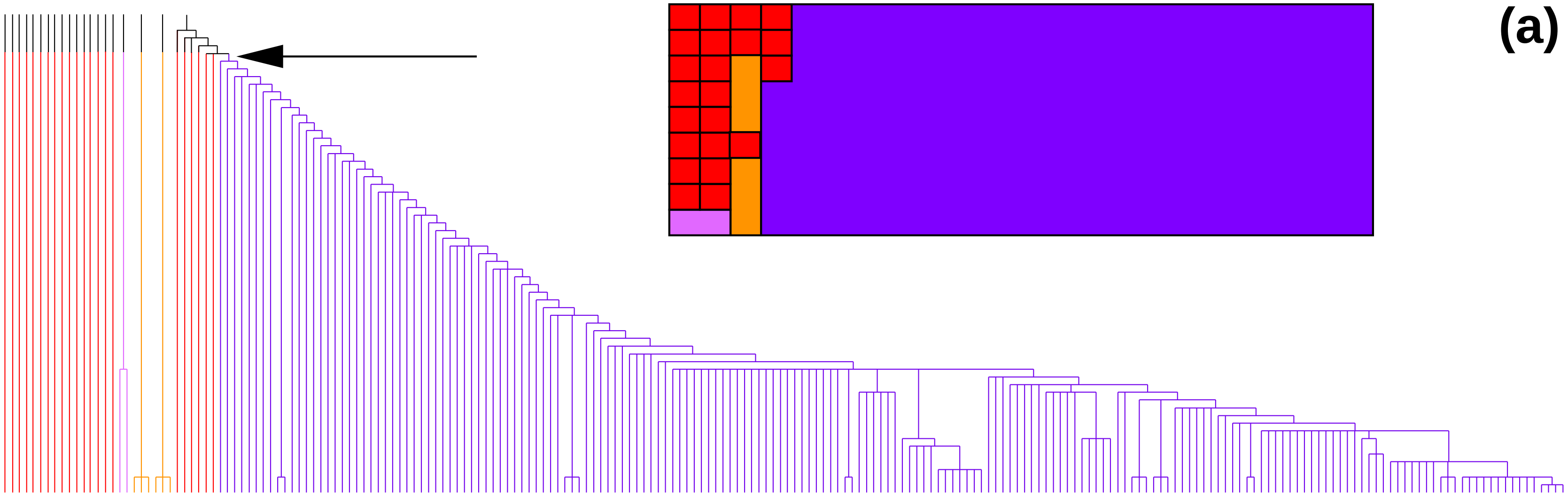
- The substrates are at the base of the tree.
- If a horizontal line is drawn across the tree, the vertices below are connected at that particular level of the hierarchy.
- Clusters that are isolated high in the hierarchy (close to the bottom of the tree) are more entangled in other pathways.

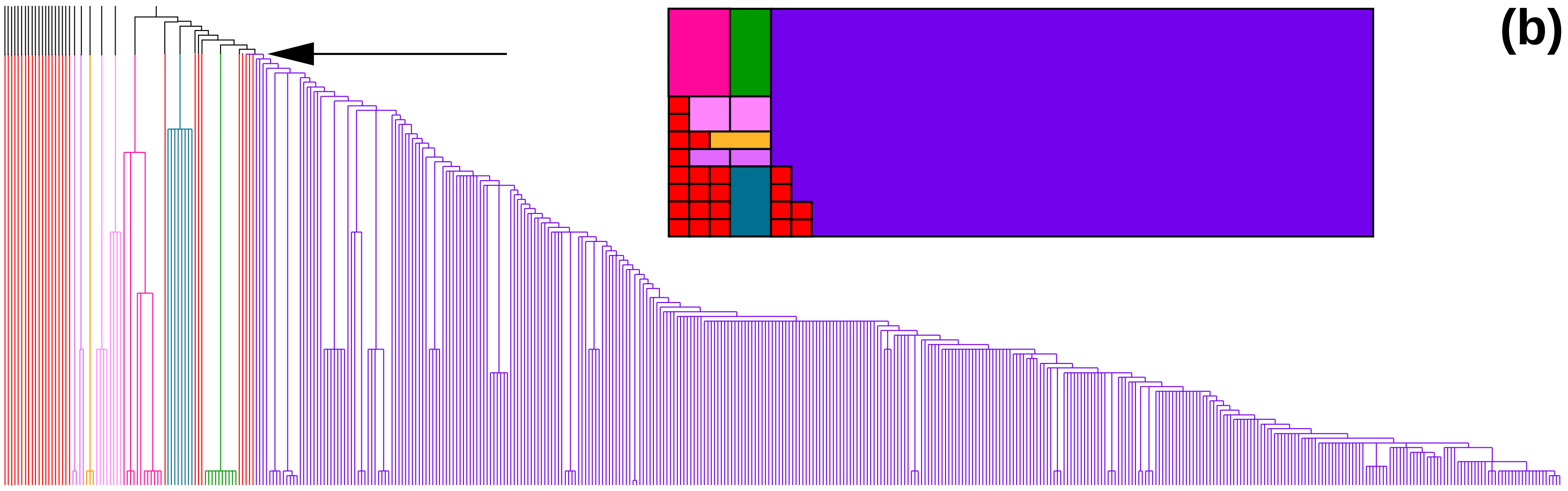
Subnetwork hierarchies



- **(a)** Clusters that get isolated at the same level are more highly wired within, than to its surrounding (and therefore a candidate to a functional module).
- **(b)** Vertices that becomes isolated at the same level forms an outer shell of the cluster in question.

EXAMPLES: *Treponema pallidum*



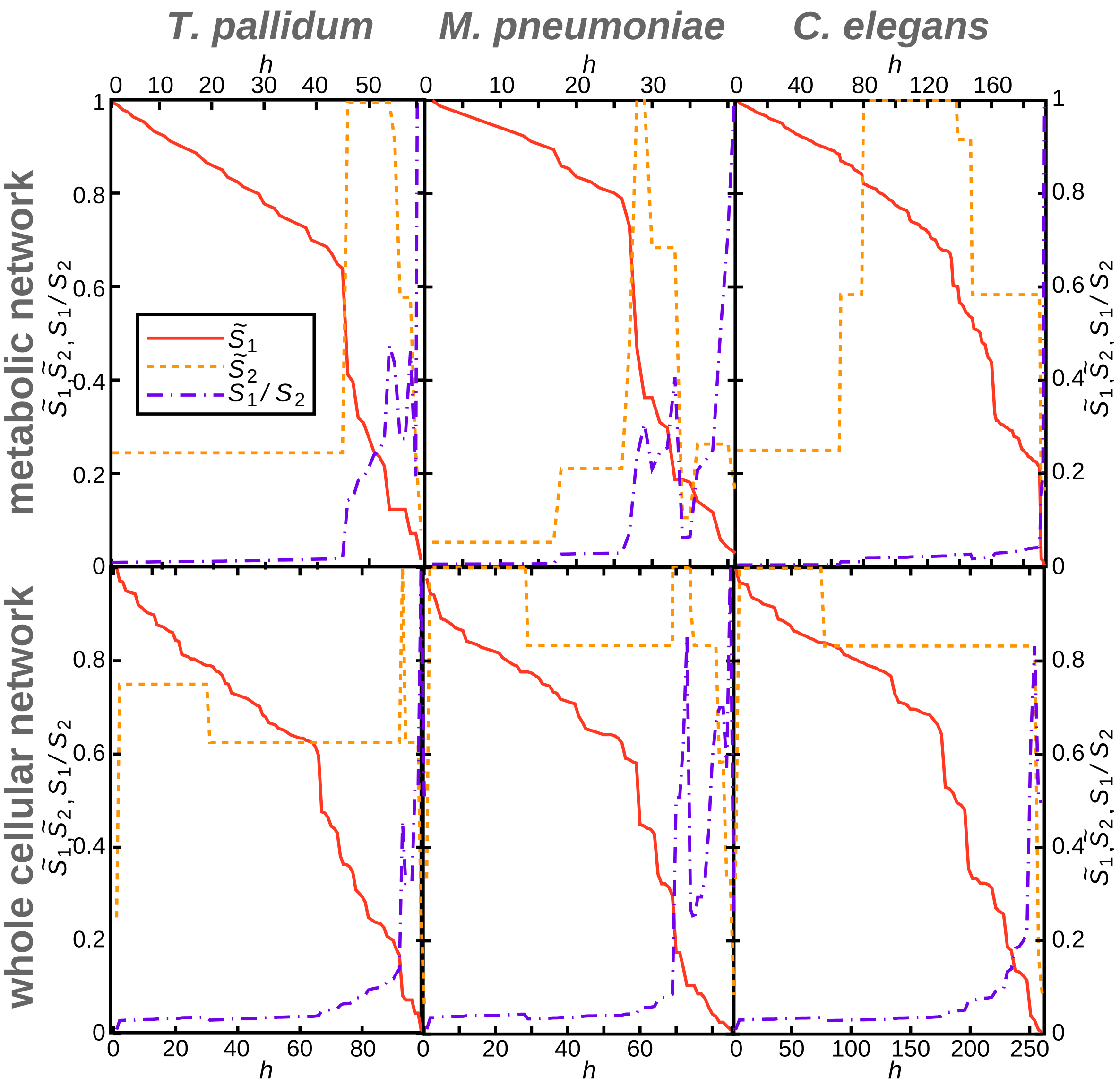


LARGE SCALE SHAPE OF THE TREES

We test 43 organisms of the WIT database.

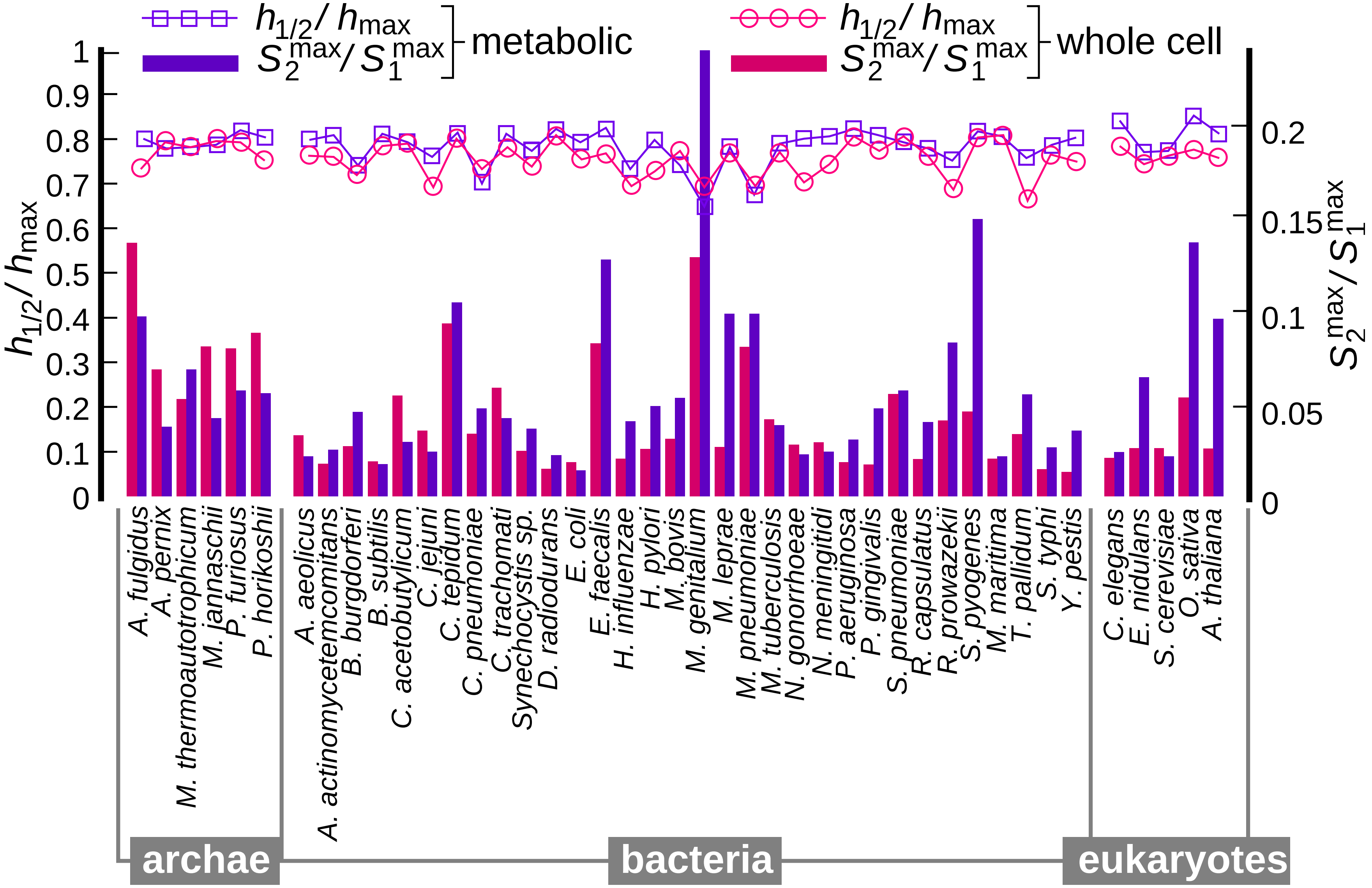
S_1 size of the biggest cluster.

S_2 size of the second biggest cluster.



Large scale shape of the trees

(continued)



Criteria for identifying subnetworks

F. Radicchi *et al.*, preprint 2003 (http://arxiv.org/abs/cond-mat/0309488/):

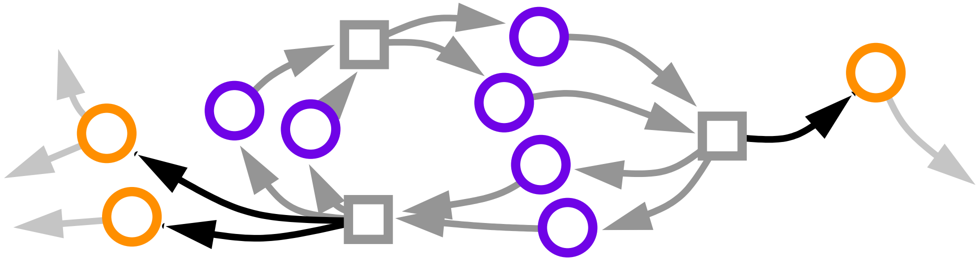
If, during the iterations of the GN algorithm, an isolated vertex set $S' \subset S$ fulfills the following criterion it is said to be a *weak community* if:

$$\sum_{s \in S'} K_{\text{in}}(s) > \sum_{s \in S'} K_{\text{out}}(s) \; , \tag{3}$$

and a *strong community* if:

$$K_{\text{in}}(s) > K_{\text{out}}(s) \text{ for all } s \in S' \; , \tag{4}$$

where $K_{\text{in}}(s)$ is the number of $s \in S$ that are products of a reaction involving a substrate $s \in S$, and $K_{\text{out}}(s)$ is the number of $s \in S \setminus S'$ that are products of a reaction involving a substrate $s \in S$.



Criteria for identifying subnetworks

- These criteria works well for social networks and electronic circuits, but gives *only trivial* clusters for biochemical networks.

Modified criteria

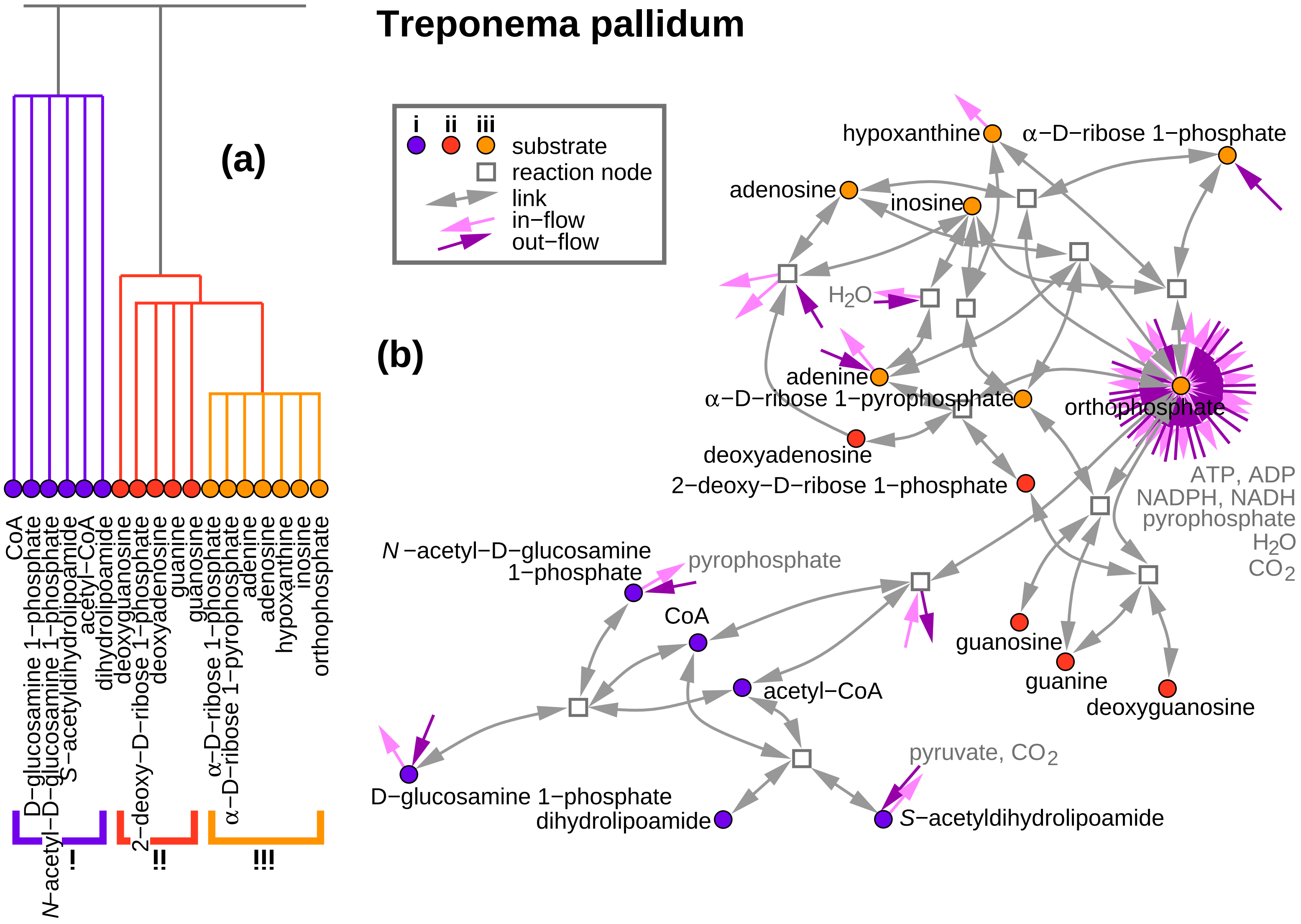
Idea: Networks with some degree of autonomy have loops. To implement this idea, consider the subnetworks with substrate vertex set S' that fulfills:

$$L(S') \leqslant \Lambda |S'| \; , \tag{5}$$

where $L(S')$ is the number of vertices in S' that lies on an elementary cycle (a closed non-self-intersecting path) of only vertices in S' and length larger than three, $|S'|$ is the number of vertices in S', and the parameter $\Lambda \in [0,1]$ is the required fraction of loop vertices.

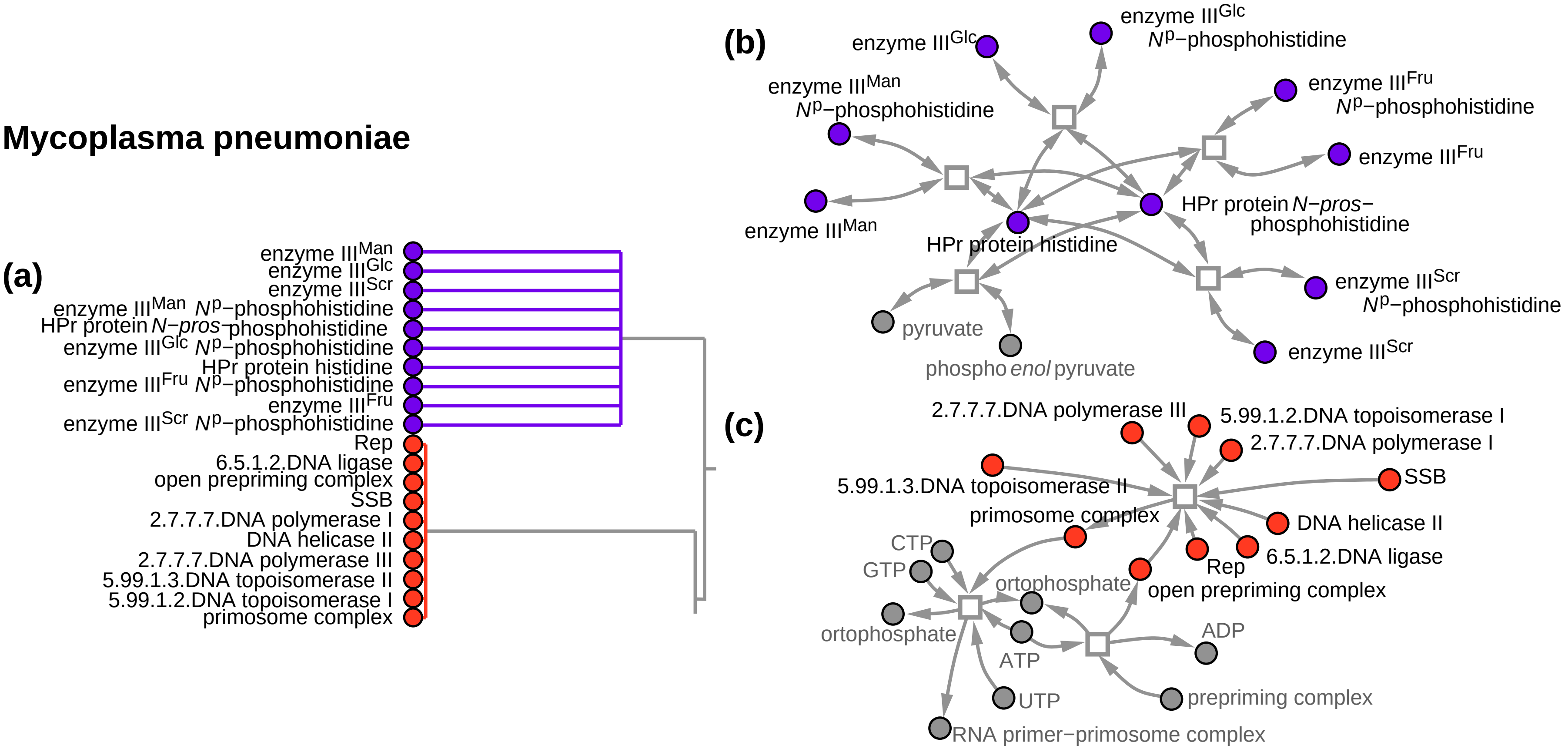
$0.5 < \Lambda \leqslant 1$ gives sensible subnetworks.

Some detected subnetworks



Some detected subnetworks *(continued)*

Mycoplasma pneumoniae



(a)
enzyme III$^{\mathrm{Man}}$
enzyme III$^{\mathrm{Glc}}$
enzyme III$^{\mathrm{Scr}}$
enzyme III$^{\mathrm{Man}}$ N^{p}−phosphohistidine
HPr protein *N−pros−*phosphohistidine
enzyme III$^{\mathrm{Glc}}$ N^{p}−phosphohistidine
HPr protein histidine
enzyme III$^{\mathrm{Fru}}$ N^{p}−phosphohistidine
enzyme III$^{\mathrm{Fru}}$
enzyme III$^{\mathrm{Scr}}$ N^{p}−phosphohistidine
Rep
6.5.1.2.DNA ligase
open prepriming complex
SSB
2.7.7.7.DNA polymerase I
DNA helicase II
2.7.7.7.DNA polymerase III
5.99.1.3.DNA topoisomerase II
5.99.1.2.DNA topoisomerase I
primosome complex
(b)
enzyme III$^{\mathrm{Glc}}$
enzyme III$^{\mathrm{Glc}}$
N^{p}−phosphohistidine
enzyme III$^{\mathrm{Man}}$
N^{p}−phosphohistidine
enzyme III$^{\mathrm{Fru}}$
N^{p}−phosphohistidine
enzyme III$^{\mathrm{Fru}}$
enzyme III$^{\mathrm{Man}}$
HPr protein *N−pros−*
phosphohistidine
HPr protein histidine
enzyme III$^{\mathrm{Scr}}$
N^{p}−phosphohistidine
pyruvate
enzyme III$^{\mathrm{Scr}}$
phospho*enol*pyruvate
(c)
2.7.7.7.DNA polymerase III
5.99.1.2.DNA topoisomerase I
2.7.7.7.DNA polymerase I
SSB
5.99.1.3.DNA topoisomerase II
primosome complex
DNA helicase II
CTP
GTP
Rep
6.5.1.2.DNA ligase
ortophosphate
open prepriming complex
ortophosphate
ADP
ATP
UTP
prepriming complex
RNA primer−primosome complex

Summary & conclusions

Advantages with graph theoretical studies of biochemical networks:

- Detection of autonomous subnetworks important for both conceptual and analytical purposes.
- The large-scale structure of biochemical networks can be described.

Our method:

- We deconstruct biochemical networks using a modified version of Girvan & Newman's algorithm.
- We emphasize the use of hierarchy-trees.
- Objective criteria based on presence of loops can be established.

We find:

- that biochemical networks are dominated by its closely connected core surrounded by increasingly loosely connected substances.
- some interesting subnetworks can be detected.