Detecting network communities: an application to phylogenetic analysis

FESC

Roberto F. S. Andrade
Leonardo B. L. Santos
Charles Santana
Suani T. R. de Pinho
Ivan Rocha

Instituto de Física
Universidade Federal da Bahia

Marcelo V. C. Diniz
Aristóteles Góes-Neto

Departamento de Ciências Biológicas, Universidade Estadual de Feira de Santana

Charbel N. El-Hani

Instituto de Biologia,
Universidade Federal da Bahia

Thierry P. Lobão

Instituto de Matematica,
Universidade Federal da Bahia

Complex networks and phylogenetic analysis

FESC

Complex networks and biological physics

FESC

Other contributions

- Comparative protein analysis of the chitin metabolic pathway in extant organisms: A complex network approach (Biosystems 101, 59 (2010))
- Modularity map of the network of human cell differentiation (PNAS 107, 5750 (2010))
- Detecting network communities: an application to phylogenetic analysis (PLoS Comp. Biol (2011))
- The fragility of protein-protein interaction networks (preprint, arXiv:1010.3531)

Other colaborators

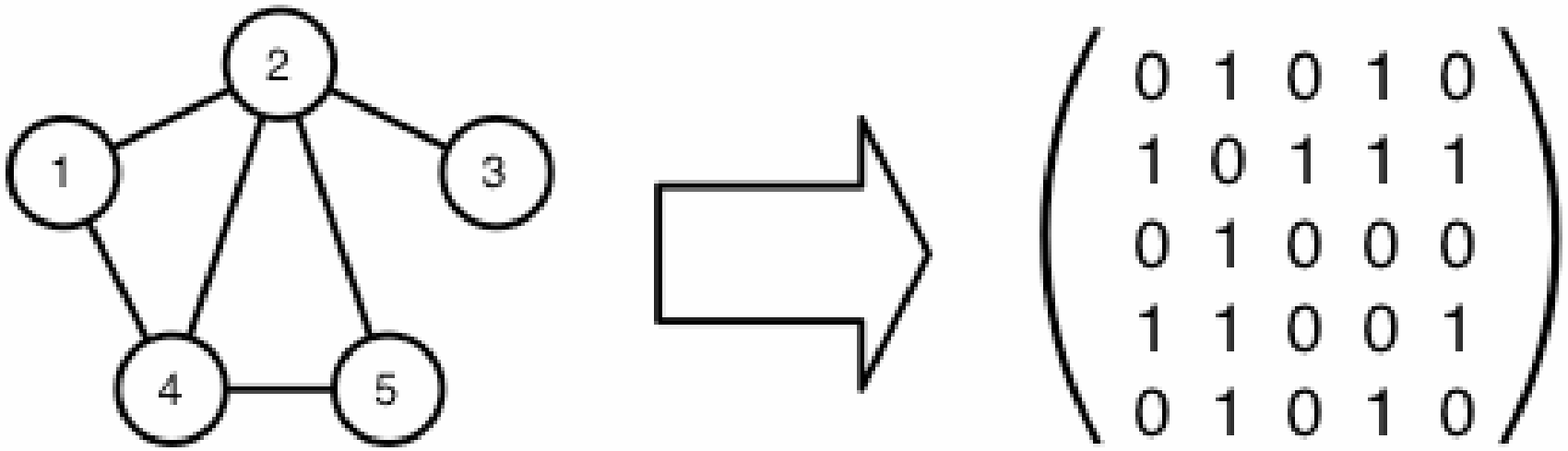
- Ernesto P. Borges
- José G.V. Miranda
- Viviane Galvão
- José S. Andrade Jr.
- Lazaros K. Gallos
- Hernán A. Makse,
- C.M. Schneider
- T. Shinbrot
- H.J. Herrmann

Outline

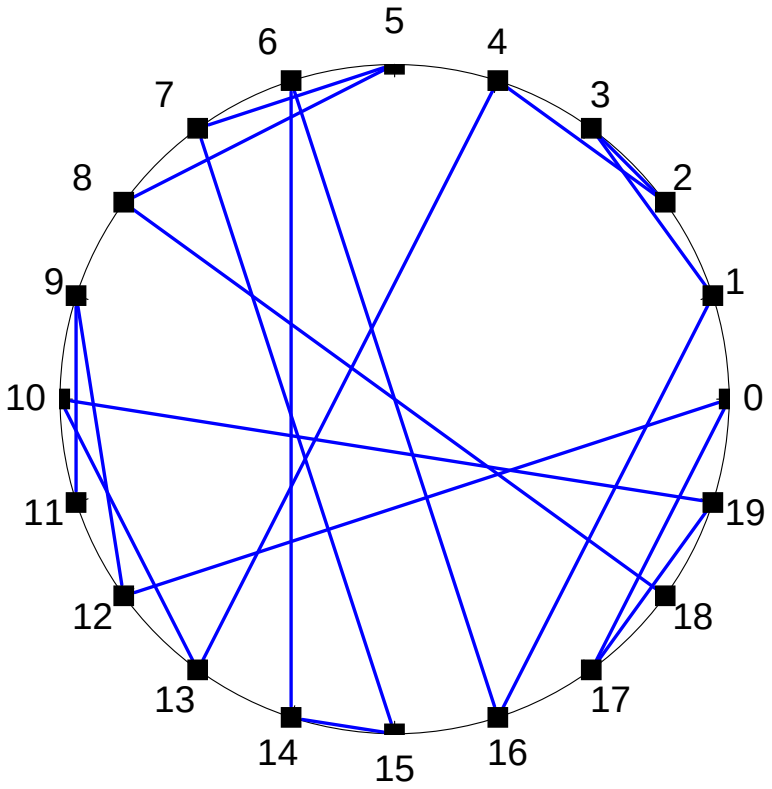
- Comparing network
- Modularity
- Phylogenetic classification
- Protein networks
- Results
- Network robustness
- Fragility of protein networks
- Results
- Conclusions and perspectives

Comparing networks

- Adjacency matrix

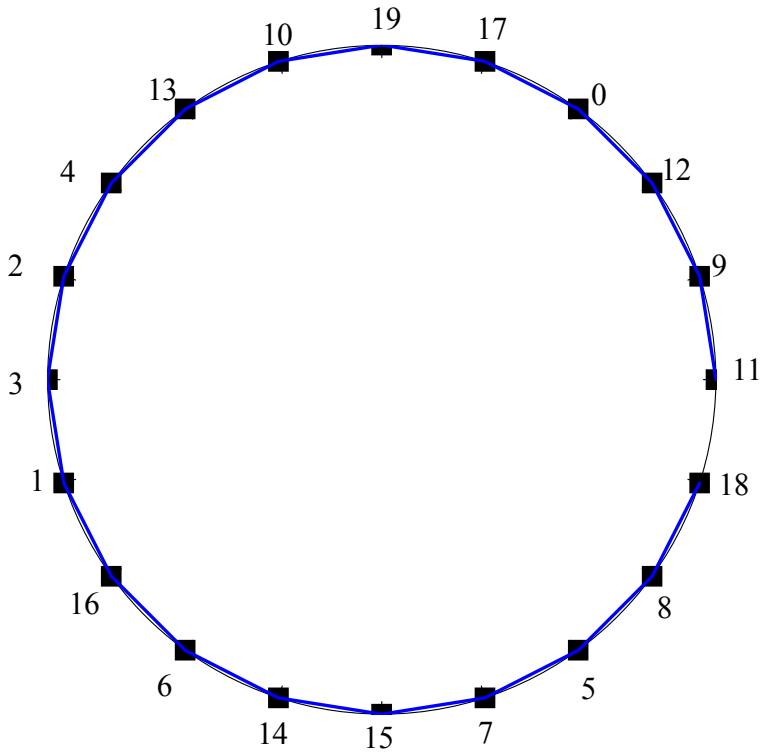


Comparing networks



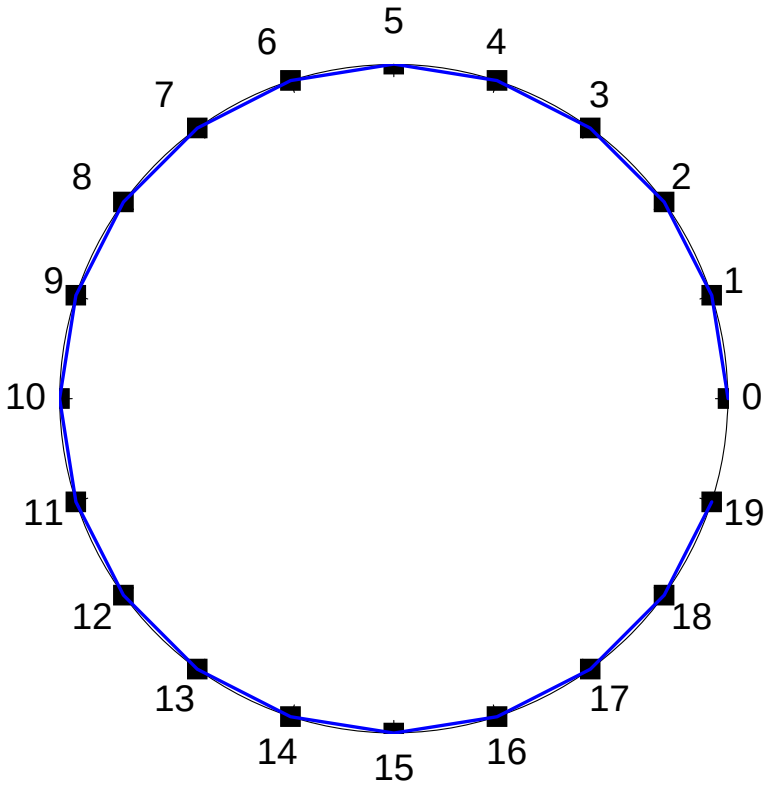

```
00000000000010000100
00010000000000001000
00011000000000000000
01100000000000000000
00100000000001000000
00000001100000000000
00000000000000101000
00000100000000010000
00000100000000000010
00000000000110000000
00000000000001000001
00000000010000000000
10000000010000000000
00001000001000000000
00000010000000010000
00000001000000100000
01000010000000000000
10000000000000000001
00000000100000000000
00000000001000000100
```

Comparing networks




```
00000000000010000100
00010000000000001000
00011000000000000000
01100000000000000000
00100000000001000000
00000001100000000000
00000000000000101000
00000100000000010000
00000100000000000010
00000000000110000000
00000000000001000001
00000000010000000000
10000000010000000000
00001000001000000000
00000010000000010000
00000001000000100000
01000010000000000000
10000000000000000001
00000000100000000000
00000000001000000100
```

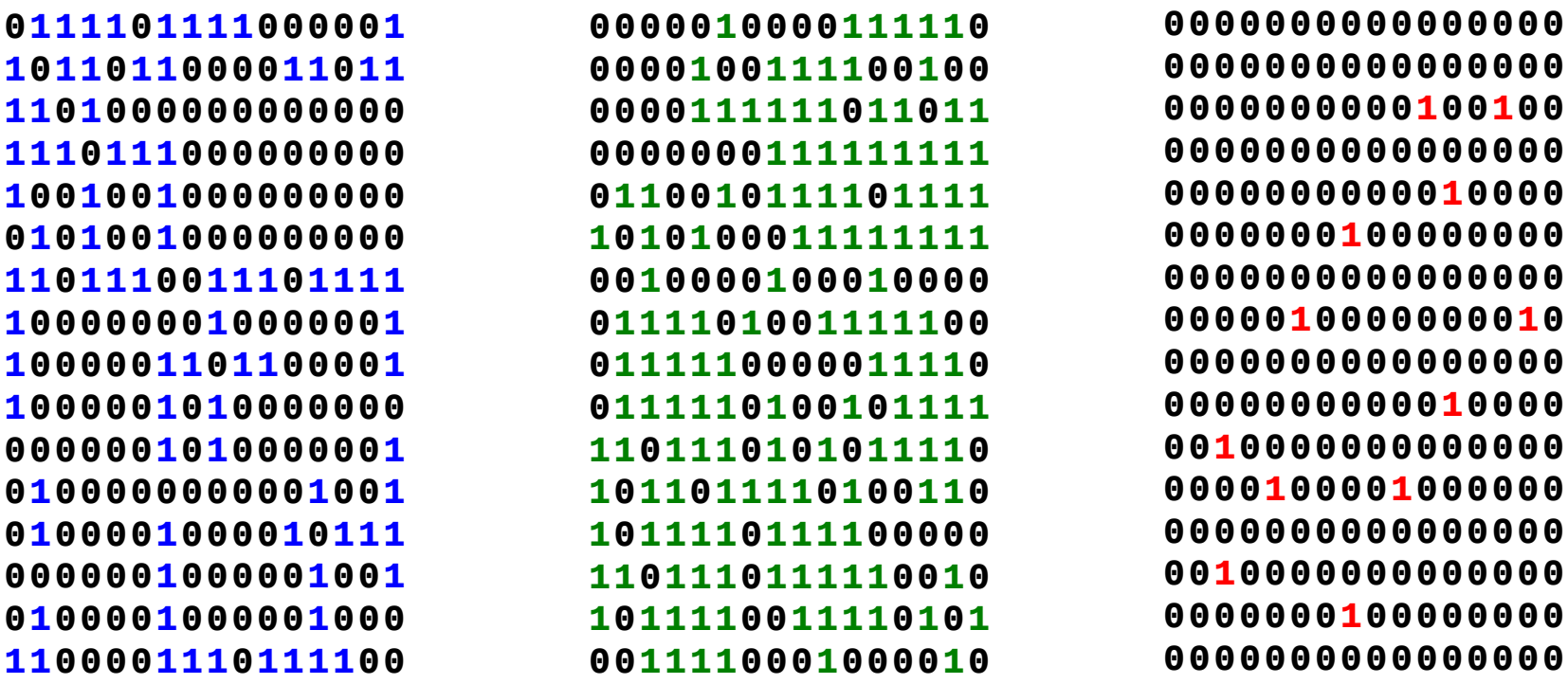
Basic concepts in network theory




```
01000000000000000000
10100000000000000000
01010000000000000000
00101000000000000000
00010100000000000000
00001010000000000000
00000101000000000000
00000010100000000000
00000001010000000000
00000000101000000000
00000000010100000000
00000000001010000000
00000000000101000000
00000000000010100000
00000000000001010000
00000000000000101000
00000000000000010100
00000000000000001010
00000000000000000101
00000000000000000010
```

Comparing networks

- Higher order ℓ neighborhood evaluation
- Description by matrices $M(\ell)$
- $\ell = 1$ $\ell = 2$ $\ell = 3$

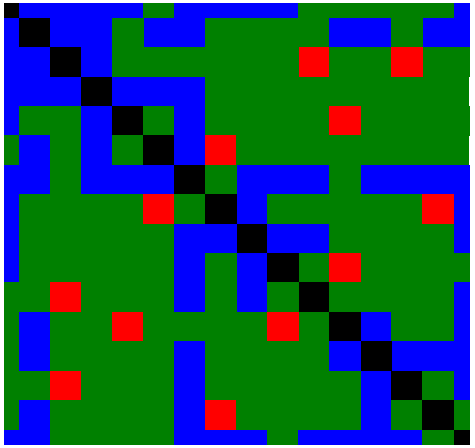


Comparing networks

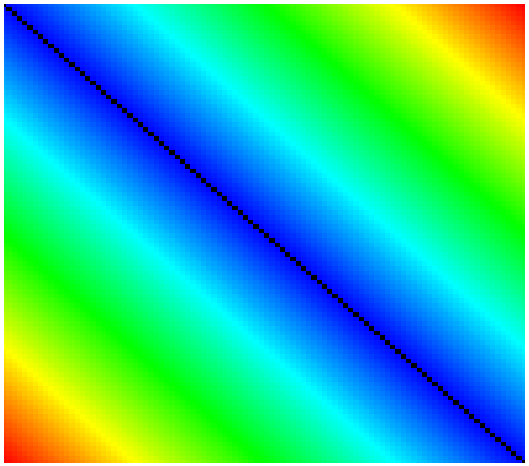
```
0111101111000001
1011011000011011
1101000000000000
1110111000000000
1001001000000000
0101001000000000
1101110011101111
1000000010000001
1000001101100001
1000001010000000
0000001010000001
0100000000001001
0100001000010111
0000001000001001
0100001000001000
1100001110111100
```

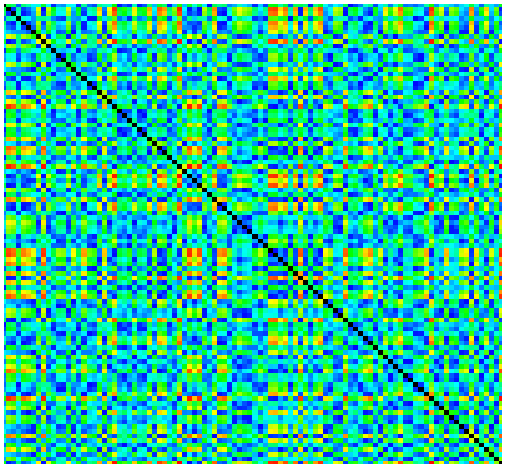
```
0000020000222220
0000200222200200
0000222222022022
0000000222222222
0220020222202222
2020200022222222
0020000200020000
0222202002222200
0222220000022220
0222220200202222
2202222020220220
2022022220200220
2022220222200000
2202222022222020
2022220022220202
0022222000200020
```

```
0000000000000000
0000000000000000
0000000000300300
0000000000000000
0000000000030000
0000000300000000
0000000000000000
0000300000000030
0000000000000000
0000000000030000
0030000000000000
0000300003000000
0000000000000000
0030000000000000
0000000300000000
0000000000000000
```



Comparing networks





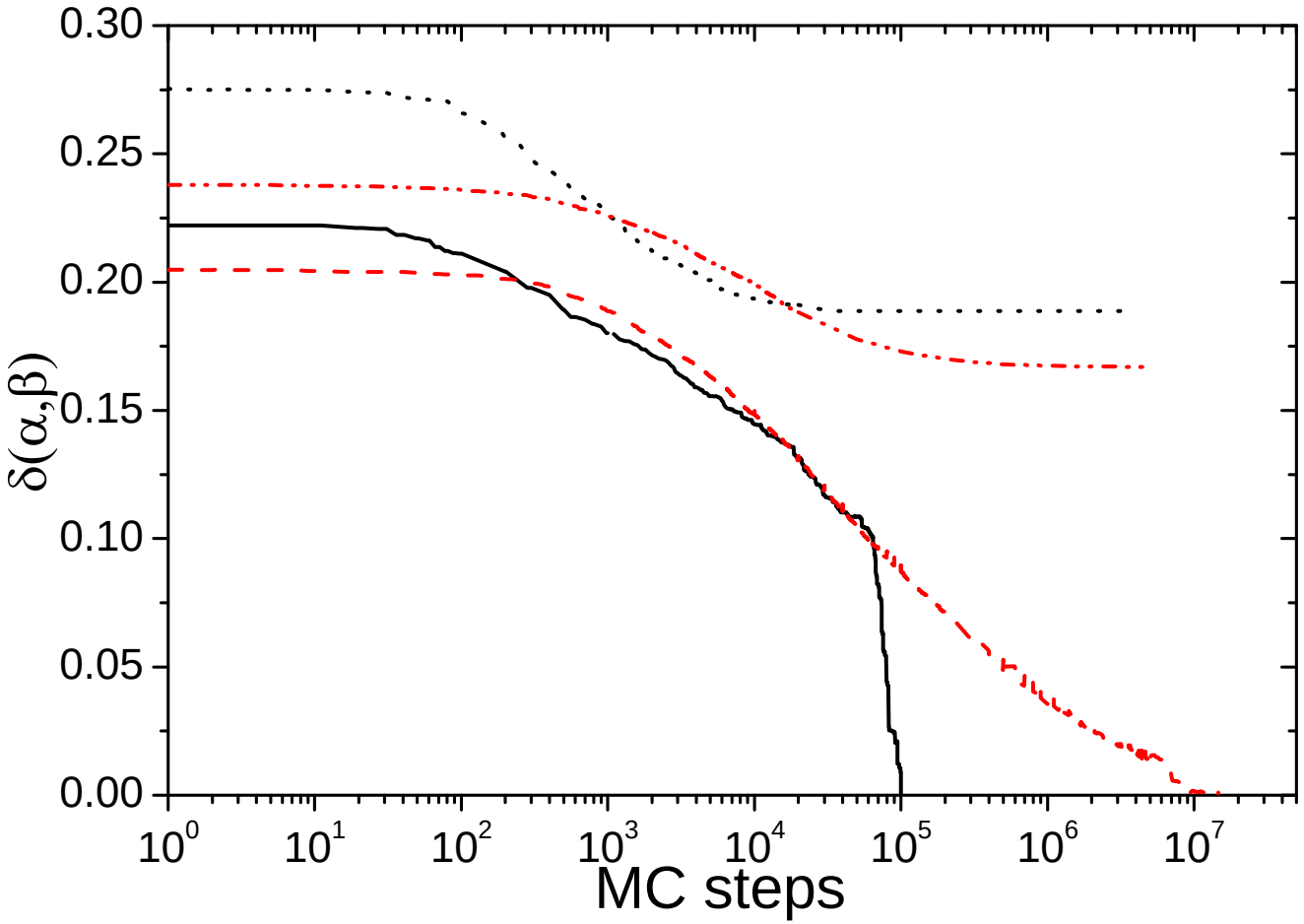
Comparing networks

- Network distance
- Define a neighborhood based distance δ

$$\delta^2(\alpha,\beta) = \frac{1}{N(N-1)} \sum_{i,j=1}^{N} \left[\frac{(\hat{M}_\alpha)_{i,j}}{D_\alpha} - \frac{(\hat{M}_\beta)_{i,j}}{D_\beta} \right]^2$$

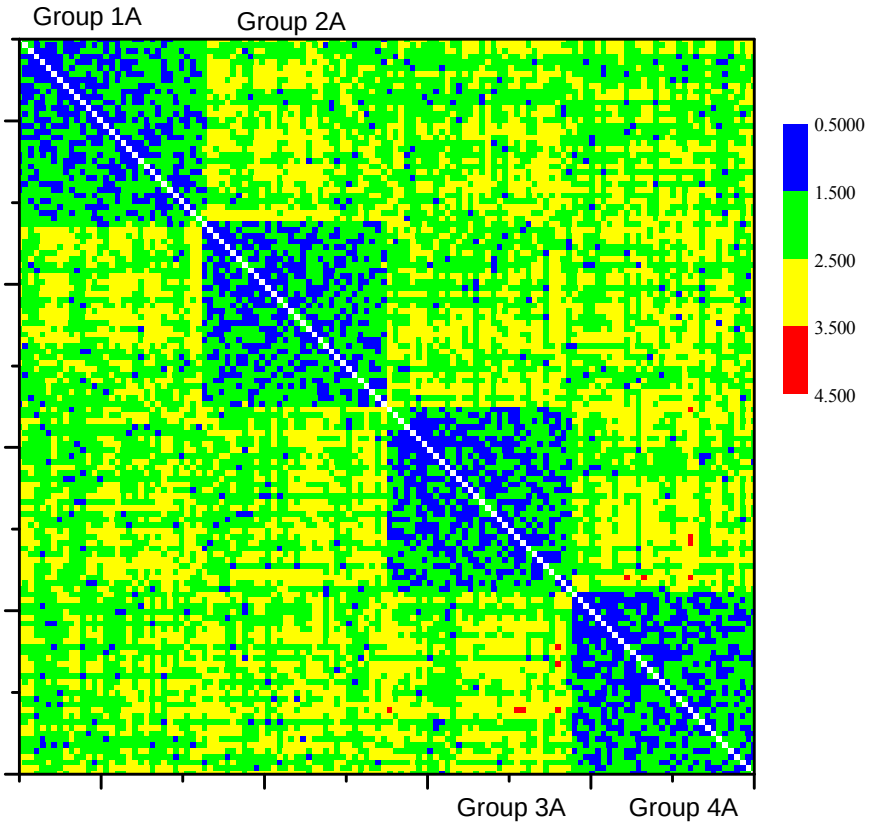
- Minimize δ by Monte-Carlo procedure

Comparing networks



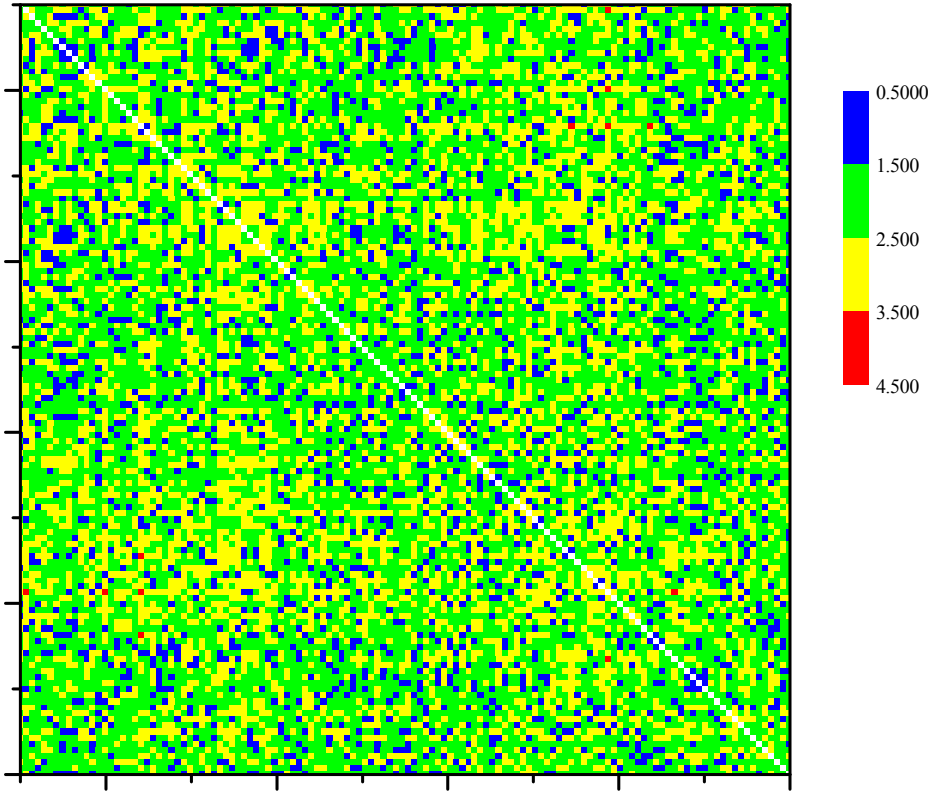
Modularity

- Modularity: number of links among groups of nodes (modules) within a network is much larger than among nodes



Modularity

- Modularity: number of links among groups of nodes (modules) within a network is much larger than among nodes



Modularity

- Finding the modules of a network: difficult task with a large number of proposed algorithms
- One important condition: modules must be there!!!
- Setting links in a network representing actual system requires information on the interaction about the entities the nodes correspond to.
- Reliability of knowledge about node interaction or strength of interaction are key steps.

Modularity

- Finding modules for a given network: efficient algorithm + network own features.

- Interpret such conditions in terms of weighted networks
- Adjacency matrix $\rightarrow$ weight matrix (WM)

$$M_{ij} = 0,1 \rightarrow W_{ij} \in [0,1]$$

- Use W to define a set of networks $M(w)$

Modularity

- Use W to define a set of networks $M(w)$

$$M_{ij}(w) = 0 \text{ if } W_{ij} \leq w$$

$$M_{ij}(w) = 1 \text{ if } W_{ij} > w$$

- Tune w to find $M(w)$ with best modular properties
- Our proposal: use network distance between neighboring networks and look for values of w that cause large peaks in the distance $\leftrightarrow$ important changes in network structure

Modularity

- Take $\alpha = w$ and $\beta = w+\Delta w$

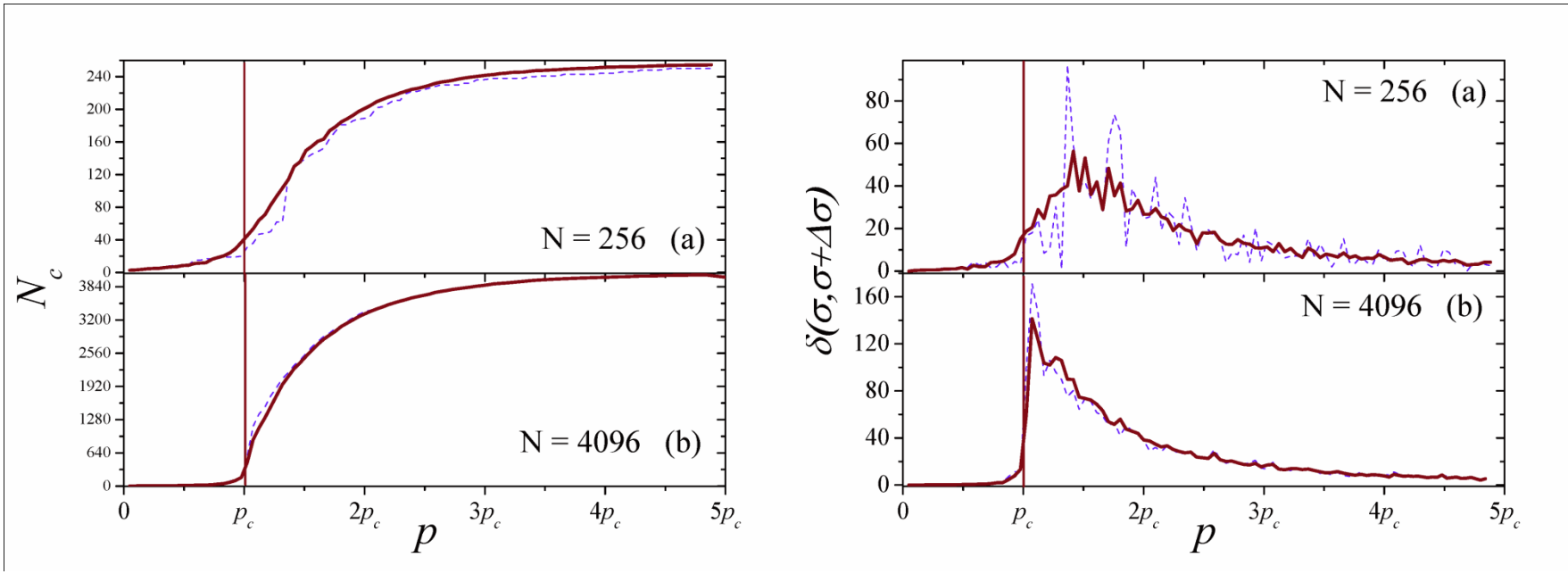
- Evaluate the neighborhood based distance δ

$$\delta^2(w, w+\Delta w) = \frac{1}{N(N-1)} \sum_{i,j=1}^{N} [(\hat{M}_w)_{i,j} - (\hat{M}_{w+\Delta w})_{i,j}]^2$$

- Identify w^*, maxima of $\delta(w,w+\Delta w)$
- Apply community finding algorithms to $M(w^*)$

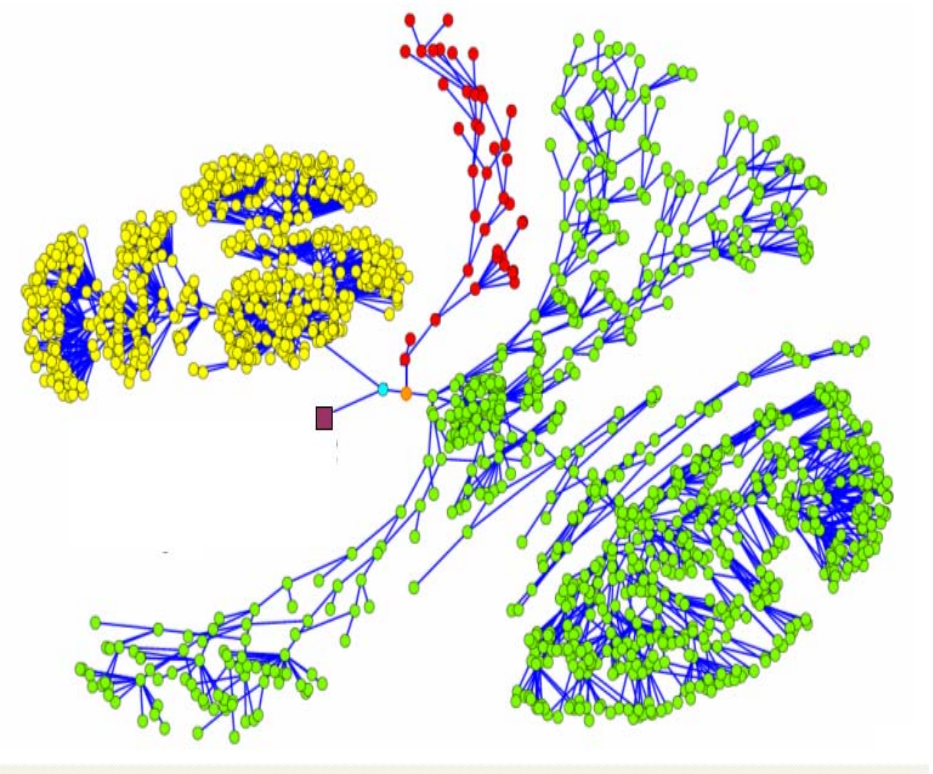
Modularity

- Example for ER network
- Emergence of a giant cluster at the transition point $p_c = 1/N$



Phylogenetic classification

- Phylogenetic trees as "periodic tables" of biologic diversity
- Usual classification: species, genus, family, order class, phylum, kingdom
- Recently introduced domains (archea, bacteria, eukarya) as basic roots of biologic evolution



Phylogenetic classification

- Classical methods of phylogenetic classification (grouping analysis): bayesian, distance, likelihood, parsimony.

- Heavily relies on qualitative biologic features as input to substitution matrices.

- This work: provides phylogenetic classification based networks constructed from protein data from completely sequenced genomes.

Phylogenetic classification

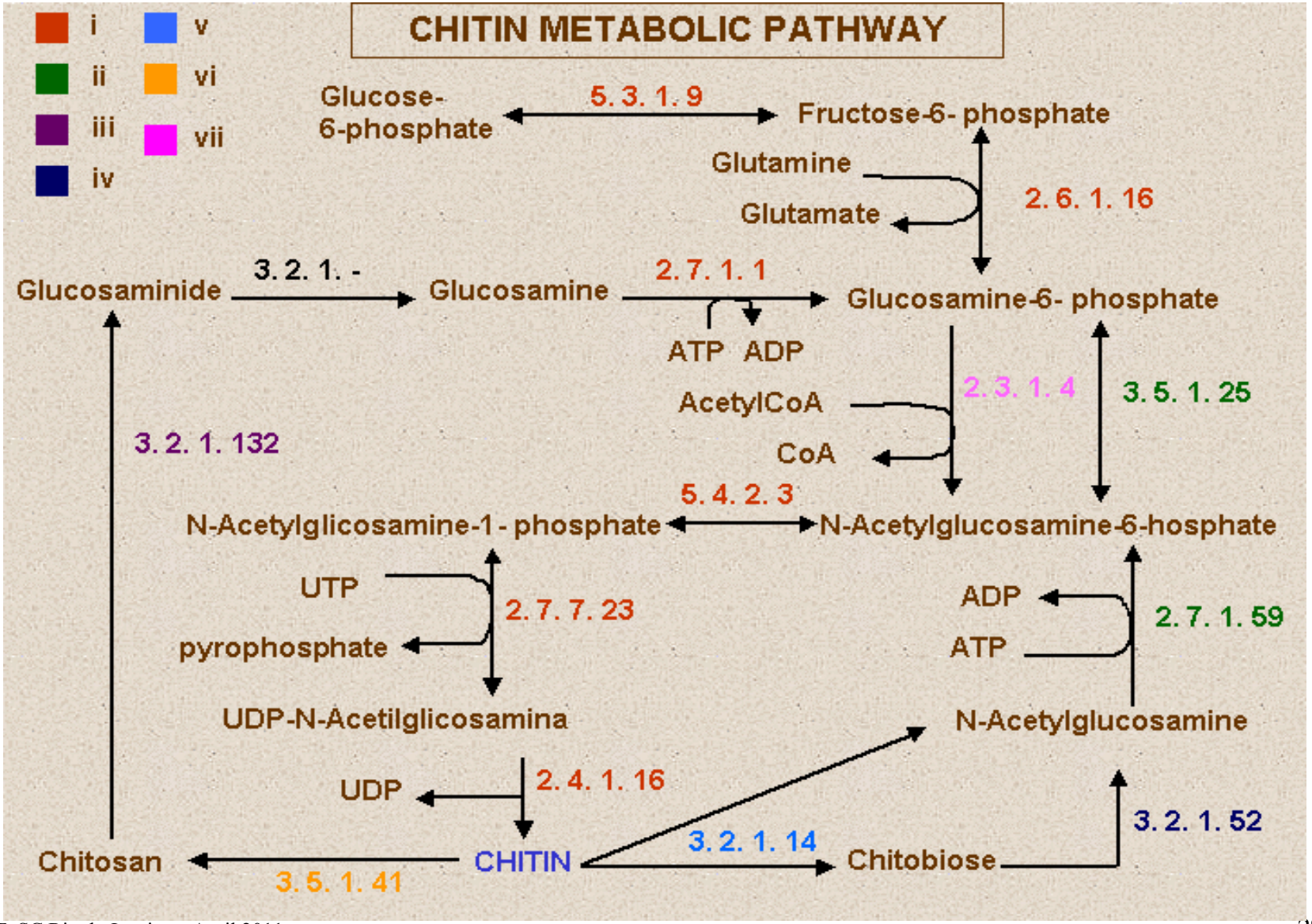
- Biological basis of the method:

- Bio-molecules required for basic reactions present in large number of organisms
- Synthesis of such molecules requires the presence of several enzymes
- Distinct organisms use own enzyme sets (pathways) to obtain the "same" molecule
- Organisms can be classified according to similarity of enzyme sets

Protein and molecular synthesis

- This work: data for chitin synthesis
- Chitin:
  - Structural endogenous carbohydrate, major component of fungal cell walls and arthropod exoskeletons.
  - Second most abundant polysaccharide in nature after cellulose
- Method can use any other molecular synthesis

Protein and molecular synthesis



Protein networks

- Database: Protein sequences from NCBI (19/05/2007)
- 1695 protein sequences for 13 enzymes within chitin metabolic pathway, e.g.
  - UDP-acetylglucosamine pyrophosphorylase
  - Acetylglucosamine phosphate deacetylase
  - Hexosaminidase
  - Phosphoglucoisomerase
  - Glucosaminephosphateisomerase
- Choose one of them along with the subset of organisms that include this or similar enzymes in the pathway

Protein networks

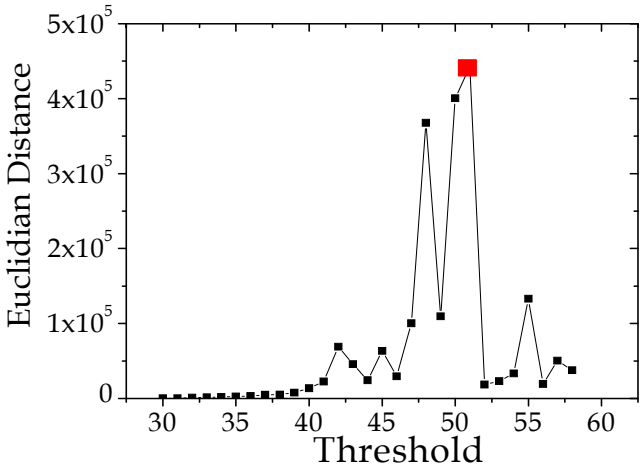
- Network node *i* represents a protein of a sequenced organism
- Network weight: comparison of protein sequences performed by BLAST (v. 2.2.15) $\Rightarrow$ similarity index (*S*) and similarity matrix *SM**.
- Associate *W(w)* with *SM*, symmetric form of *SM** undirected network adjacency matrix
- Nodes *i,j* are connected in a network if SM_{ij} is above a pre-established threshold S_{th} (=*w**)

Results

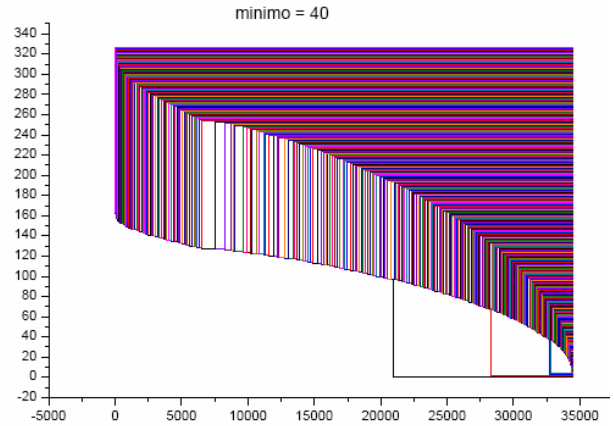
- Network measures for each value S_{th}:
  - Degree distribution $P(k)$
  - Clustering coefficient C
  - Average path-length $\langle d \rangle$
  - Edge betwenness B
  - **Network distance $D_{\alpha\beta}$**
- Networks depend on S_{th}
- Judicious choice of value of S_{th} optimizes reliability of classification scheme, based on Newman-Girvan method

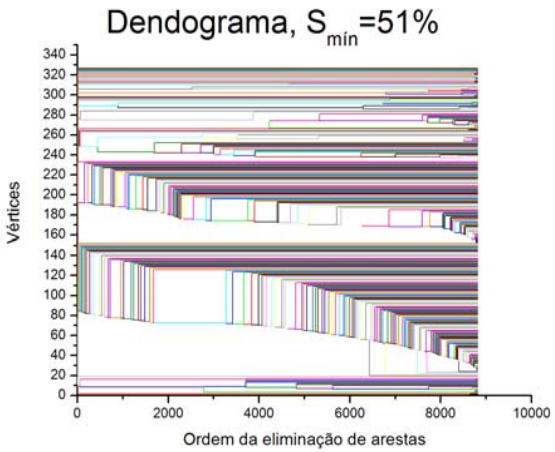
Results

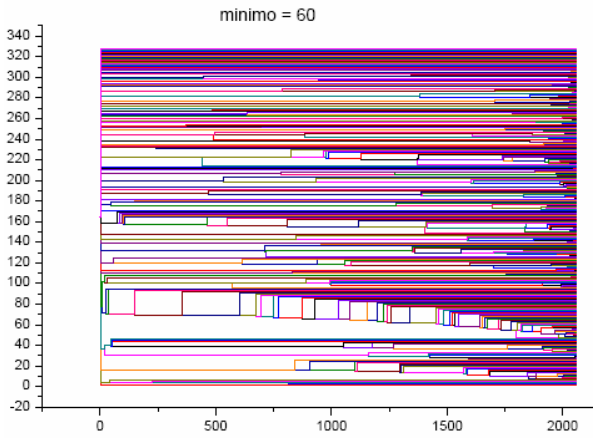
- Enzyme UDP
- $S_{th} \approx 51\%$: sudden transition in network properties
  - Sharp decrease in <*d*>
  - Clustering *C* remains relatively unchanged
  - Sharp change in dendrogram based on *B*
  - **Peak in the distance $D_{\alpha,\alpha+1}$**



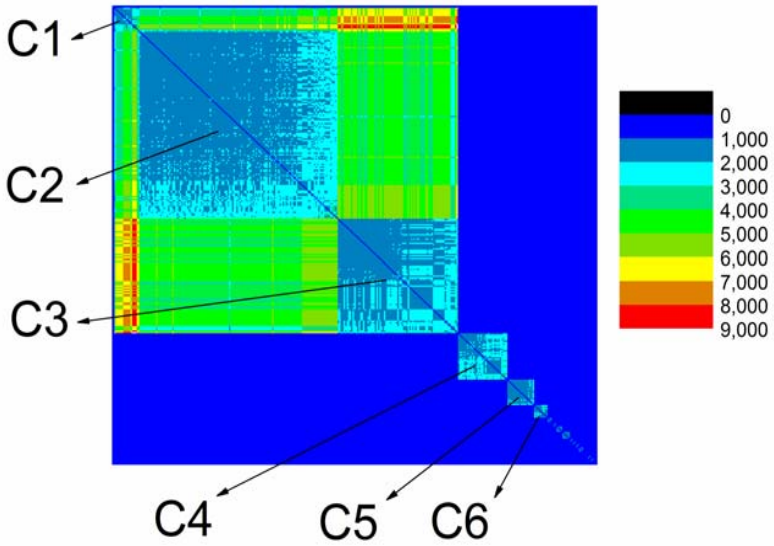
Results





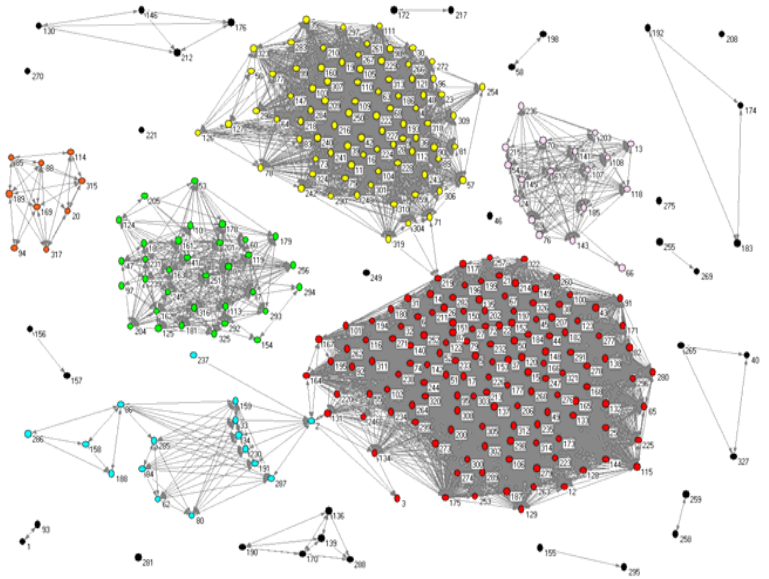


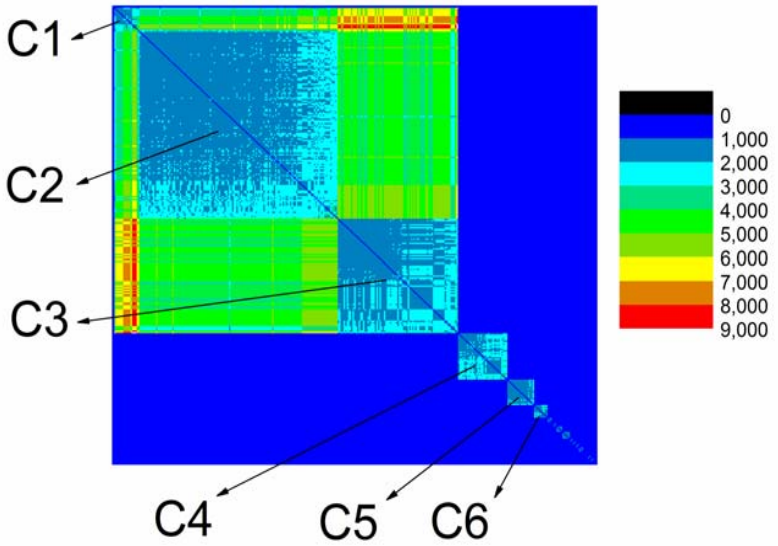
- $D_{\alpha,\alpha+1}$ is reflected in the dendrogram structure
- At S_{th}=51%, main groups identified are reproduced in neighborhood matrix
- Moduli C1-C6 with precise biologic meaning.



Results

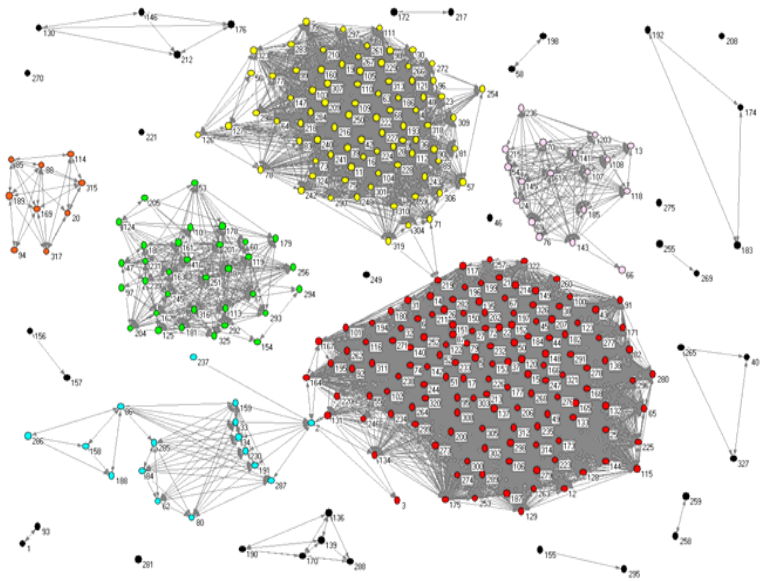
- C1 – Cyanobacteria C2 – Firmicutes
  C3 – β and γ Proteobacteria C4 – α-Proteobacteria
  C5 – Actinobacteria C6 – ε-Proteobacteria





Results

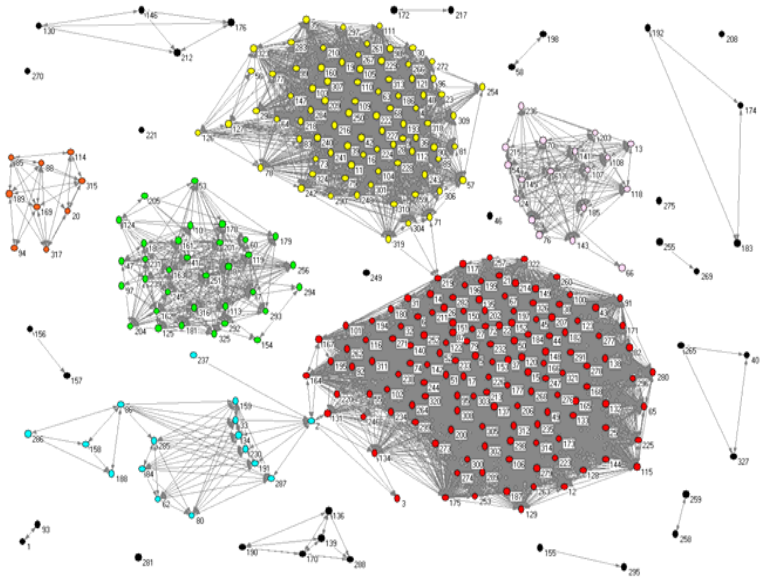
- C1 – Cyanobacteria C2 – Firmicutes
  C3 – β and γ Proteobacteria C4 – α-Proteobacteria
  C5 – Actinobacteria C6 – ε-Proteobacteria



- Identification of these modules in the network.
- Crossing results from our approach with taxonomic and phylogenetic data: the modules correspond in clear and rather precise way to bacterial phyla and/or classes

Results

- C1 – Blue – Cyanobacteria
  C2 – Yellow – Firmicutes
  C3 – Red – Beta and Gamma Proteobacteria
  C4 – Green – Alpha Proteobacteria
  C5 – Pink – Actinobacteria
  C6 – Orange – Epsilon Proteobacteria



Results

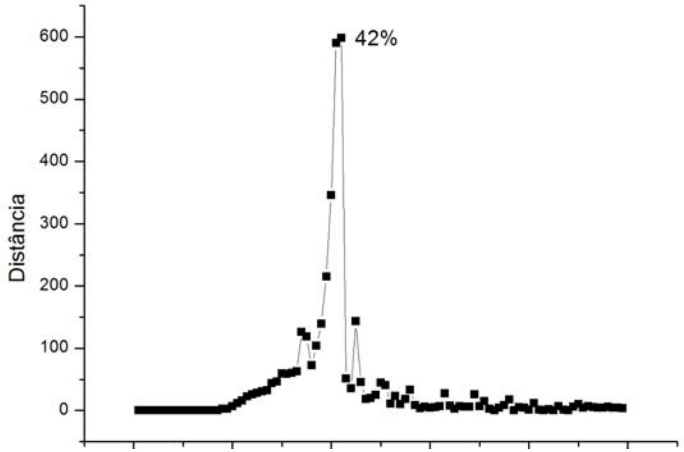
- Same method was applied to other networks (with no. of vertices ≥ 100) $\Rightarrow$ accurately defined grouping suggests robustness of the method.

Enzime	<SIM>	σ	S_t	# Diferents sequences	# Diferents phylum
UDP-acetylglucosamine pyrophosphorylase	39	15.91	51	327	14
Acetylglucosamine phosphate deacetylase	34	11.21	42	176	12
Glucosaminephosphate isomerase	37	15.16	40	313	20
Hexosaminidase	22	21.40	36	328	13
Phosphoglucoisomerase	27	23.45	36	501	20

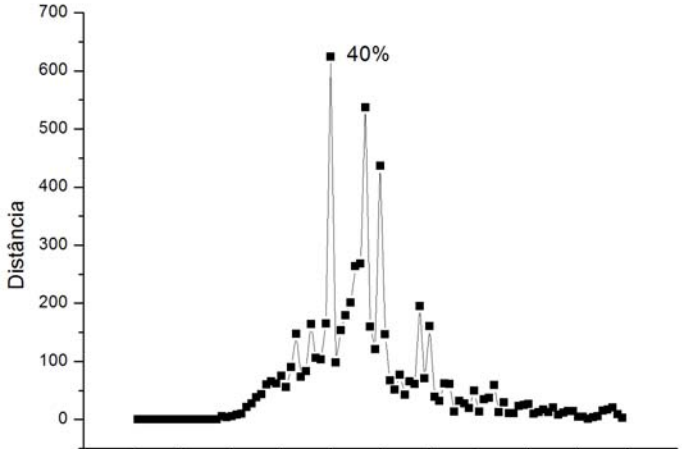
Results

- Network distance $D_{\alpha\beta}$ x threshold S_{th}

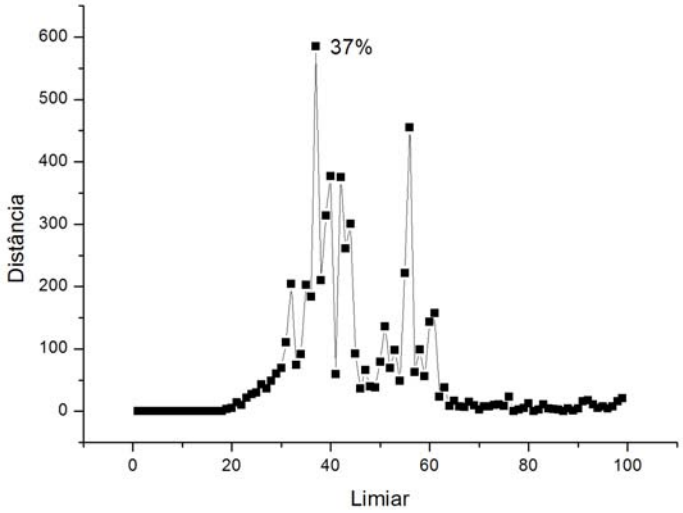
Acetyl



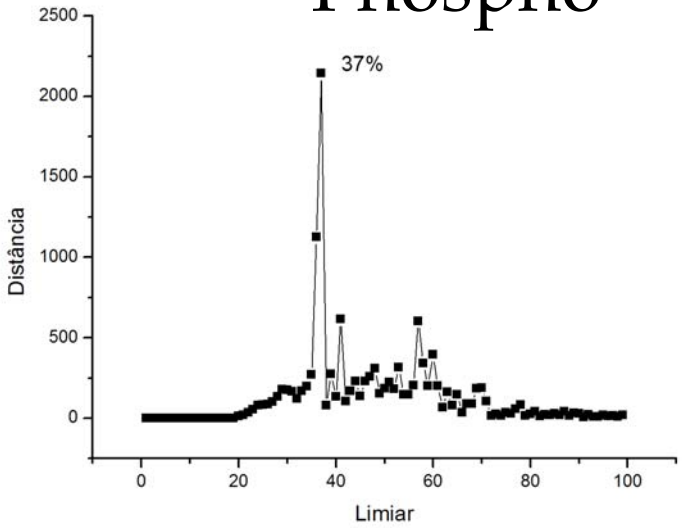
Gluco



Hexo



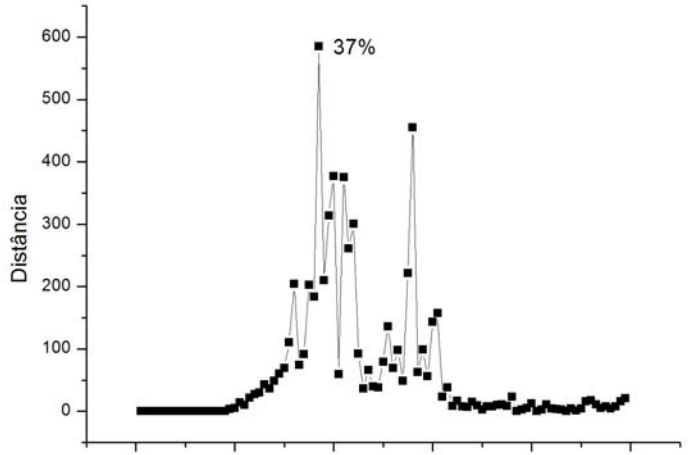
Phospho



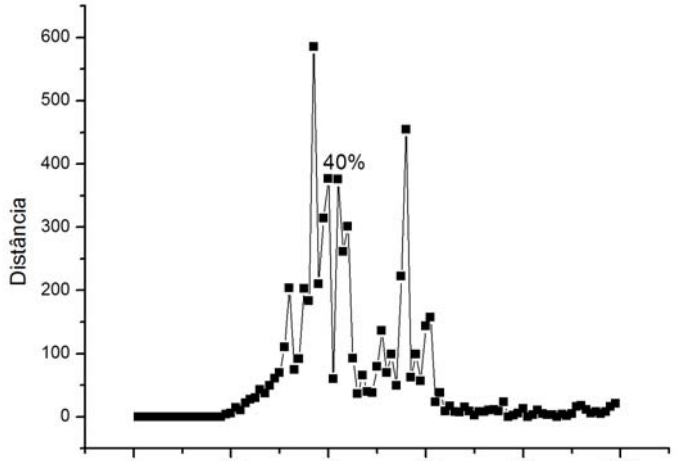
Results

- Hexo: Dependence of network on S_{th}

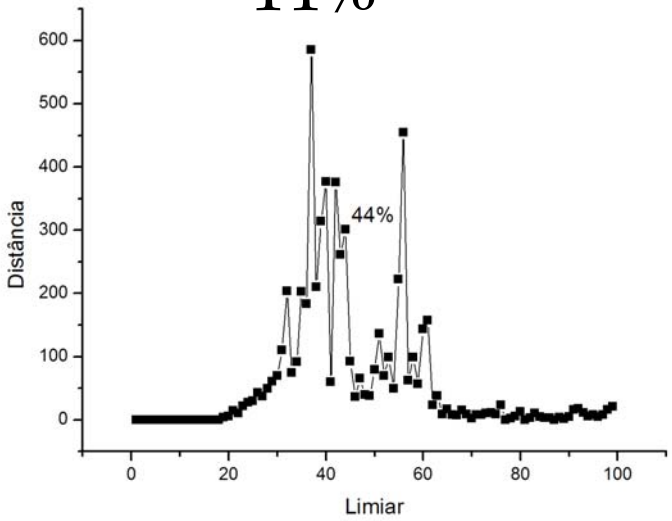
37%



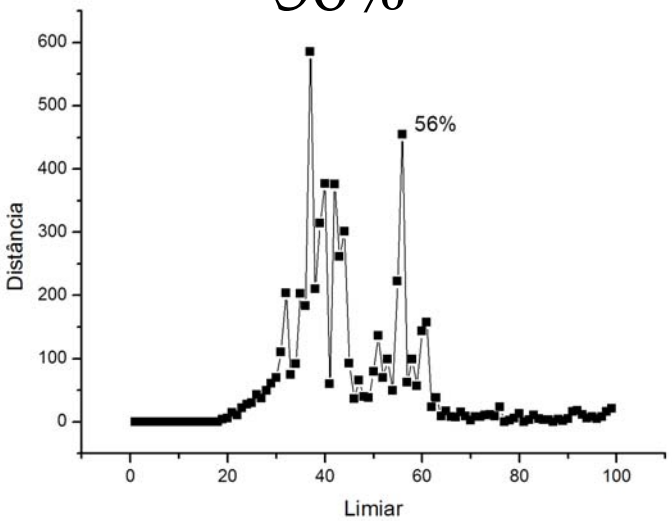
40%



44%



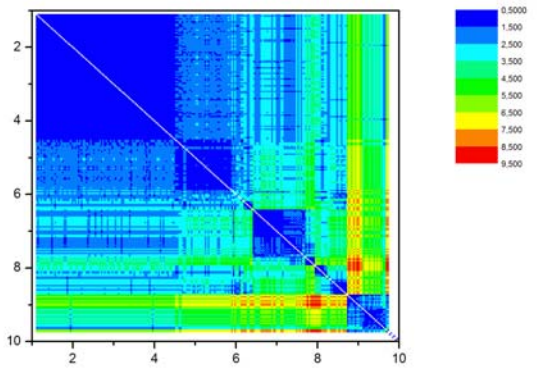
56%



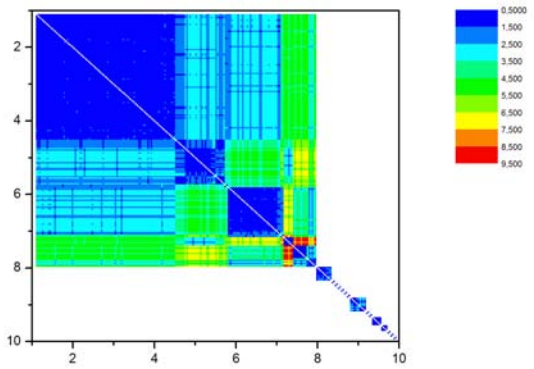
Results

- Hexo: Dependence of network on S_{th}

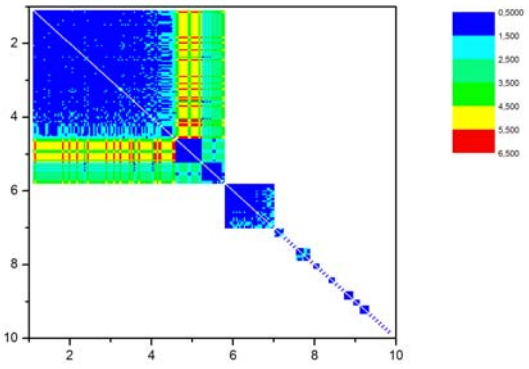
37%



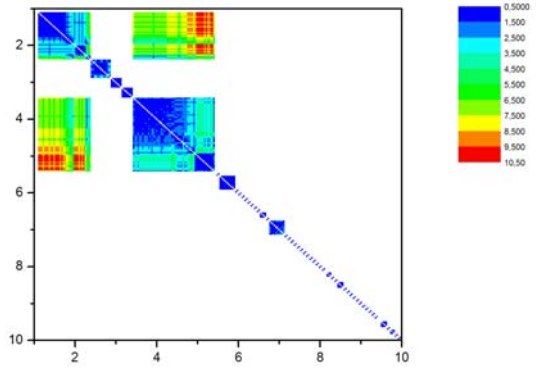
40%



44%



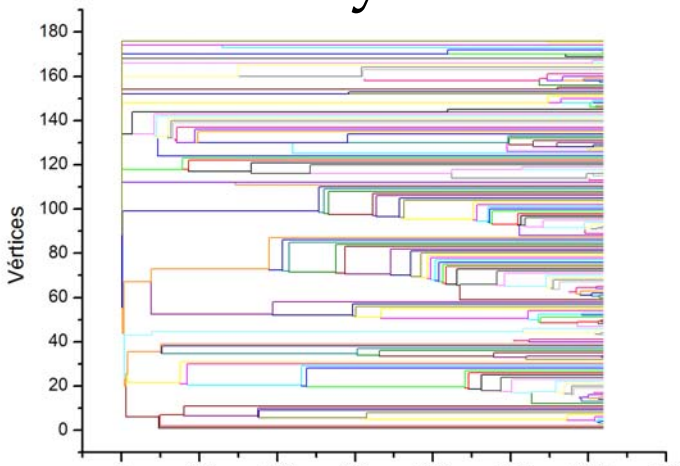
56%



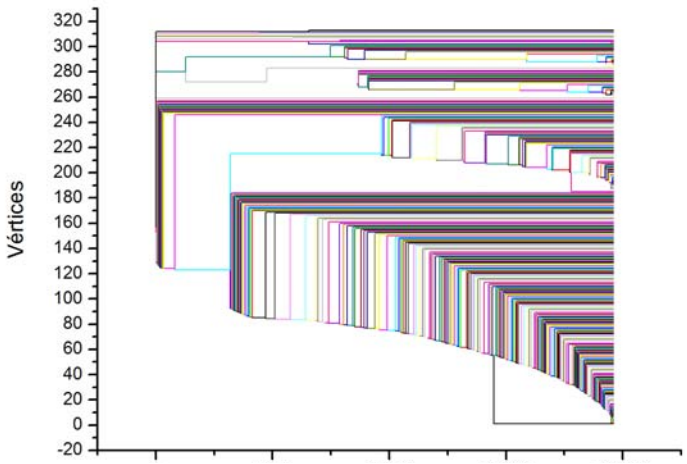
Results

- Dendrograms at first threshold S_{th}

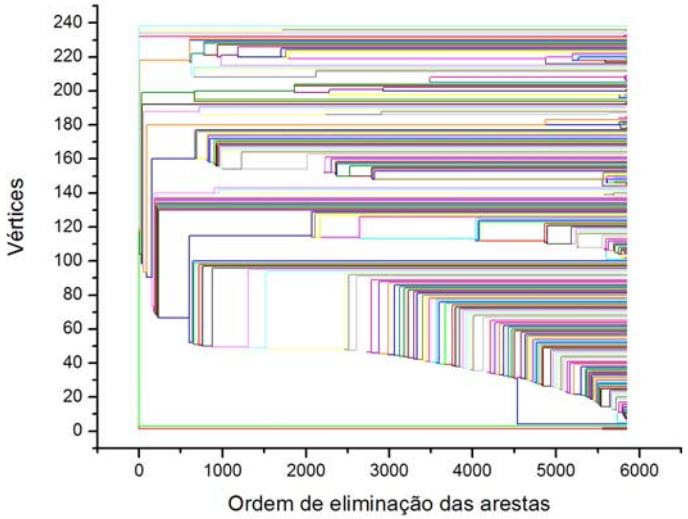
Acetyl



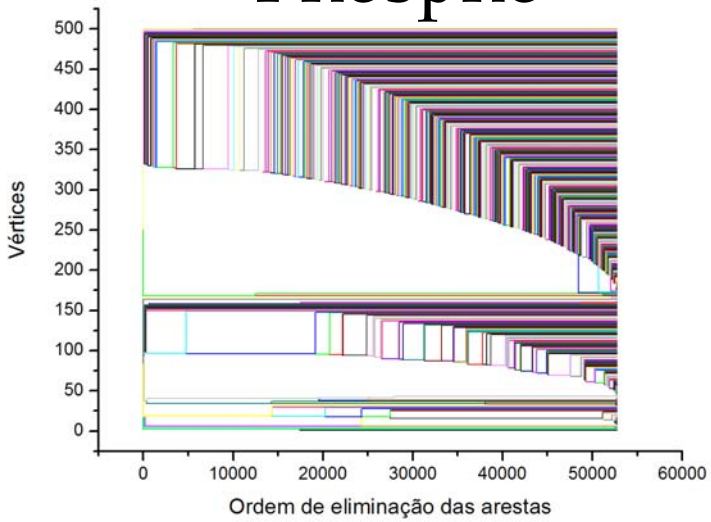
Gluco



Hexo



Phospho



Results

- Same method was applied to other networks (with no. of vertices ≥ 100) $\Rightarrow$ accurately defined grouping suggests robustness of the method.

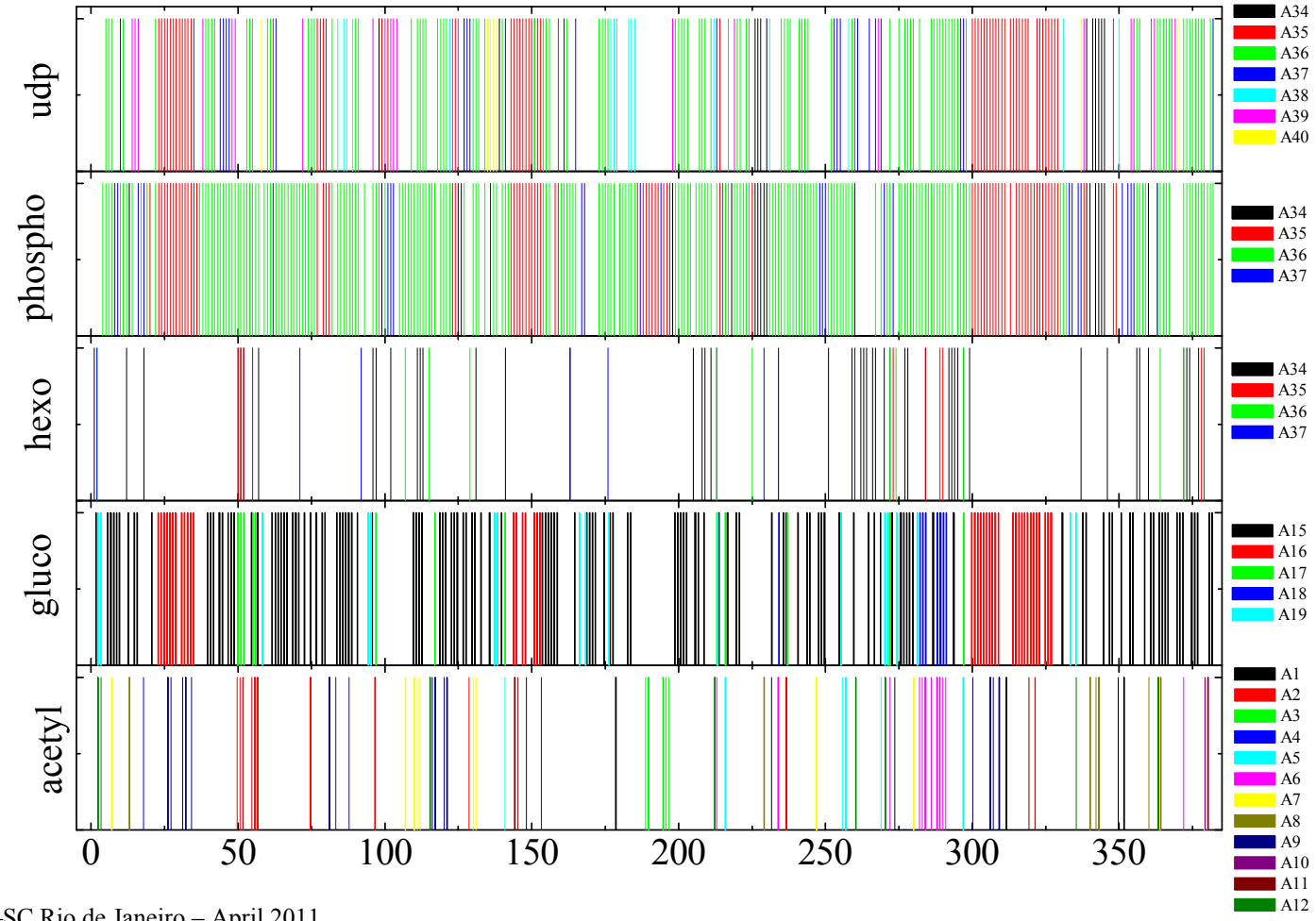
Enzime	<SIM>	σ	S_t	# Diferents sequences	# Diferents phylum
UDP-acetylglucosamine pyrophosphorylase	39	15.91	51	327	14
Acetylglucosamine phosphate deacetylase	34	11.21	42	176	12
Glucosaminephosphate isomerase	37	15.16	40	313	20
Hexosaminidase	22	21.40	36	328	13
Phosphoglucoisomerase	27	23.45	36	501	20

Results

- Number of distinct sequences in different networks totalize 1645 (out of 1695 in data set)
- Each sequence belongs to only one network
- Identification of 382 distinct organisms
- More than one sequence can be present in the same organism
- Congruence of classification by distinct networks

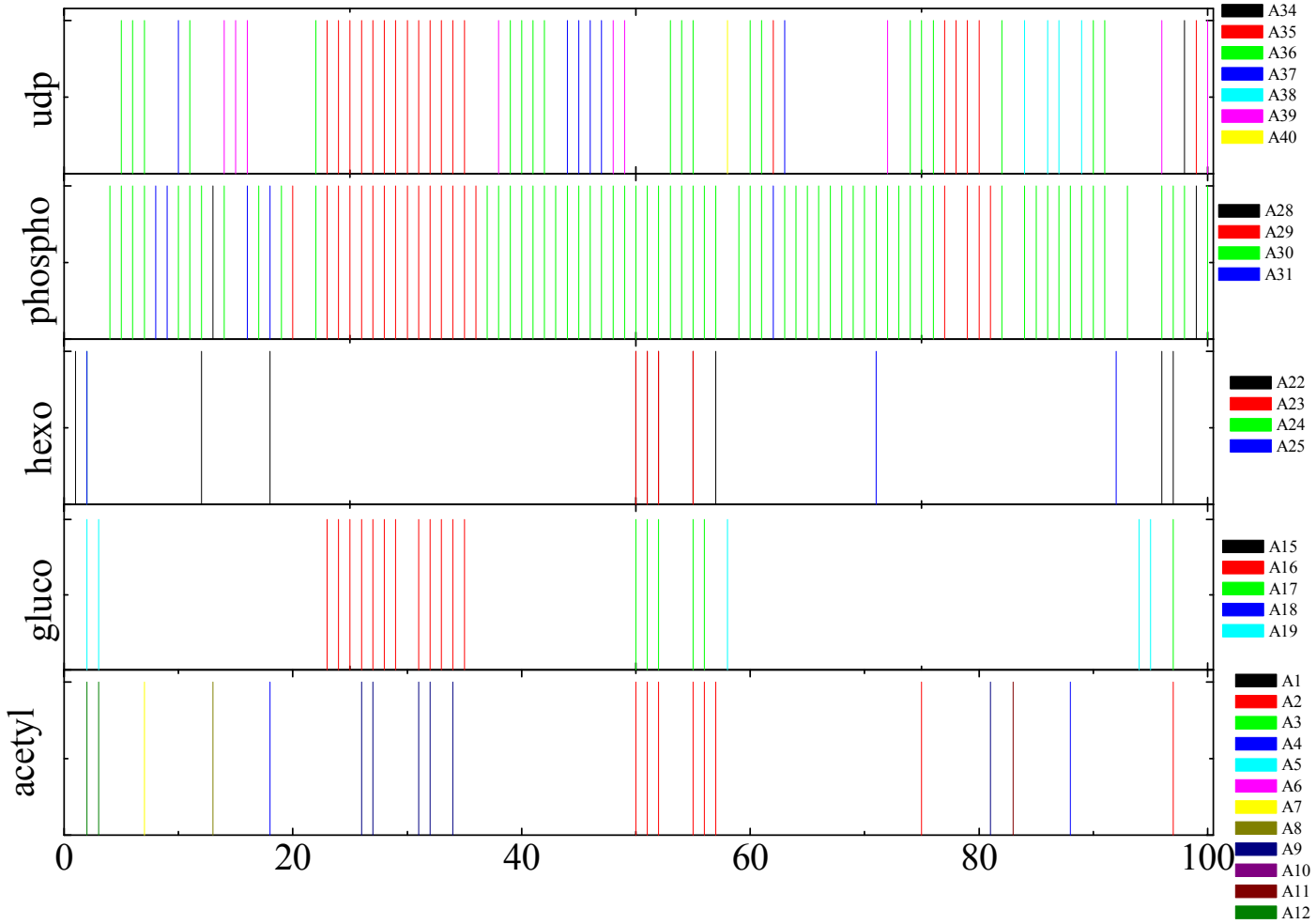
Results

- Congruence of classification by distinct networks
  - Networks with different sizes and communities



Results

- Congruence of classification by distinct networks
  - Networks with different sizes and communities



Results

Protein	σ_{max}	# nodes	# communities
Acetyl	42	176	12
Gluco	40	313	5
Hexo	37	238	10
Phospho	37	501	6
UDP-	51	327	7

Results

- Values of congruence obtained after pair-wise comparison of the phylogenetic analysis provided by two different networks. The average value of the entries in the table is 84%.

	A	G	H	P	U
A		0.79	0.73	0.93	0.91
G	0.79		0.69	0.83	0.87
H	0.73	0.69		0.90	0.79
P	0.93	0.83	0.90		0.95
U	0.91	0.87	0.79	0.95	

Results

- Further results for chitin synthase, another protein of the chitin metabolic pathway.
- For this data set, we have found that the phylogenetic classification obtained through the complex network with other methods agrees with those based
  - Bayesian – 0.56,
  - Distance – 0.53
  - Likelihood – 0.58
  - Parsimony – 0.64
- Scores are similar when methods are compared among themselves

Network robustness

- Robustness: How long a network can stand (~ giant cluster) if successive attacks eliminate nodes or connections
- Usual robustness measure: percolation threshold q_c
- New proposal (JSTAT P01027 (2011) takes into account the size of all largest clusters after the removal of each node:

$$R = \frac{1}{N+1}\sum_{Q=0}^{N} s(Q)$$

Network robustness

- Two different kinds of attack: malicious (targeted at highly connected nodes) and random
- Depending on the topology, networks can be more resistant to one or other type of attack
- Networks with onion like topology (core is occupied by highly connected nodes) is more resistant to targeted attacks
- How do biological networks resist to both kinds of attacks?

Fragility of protein network

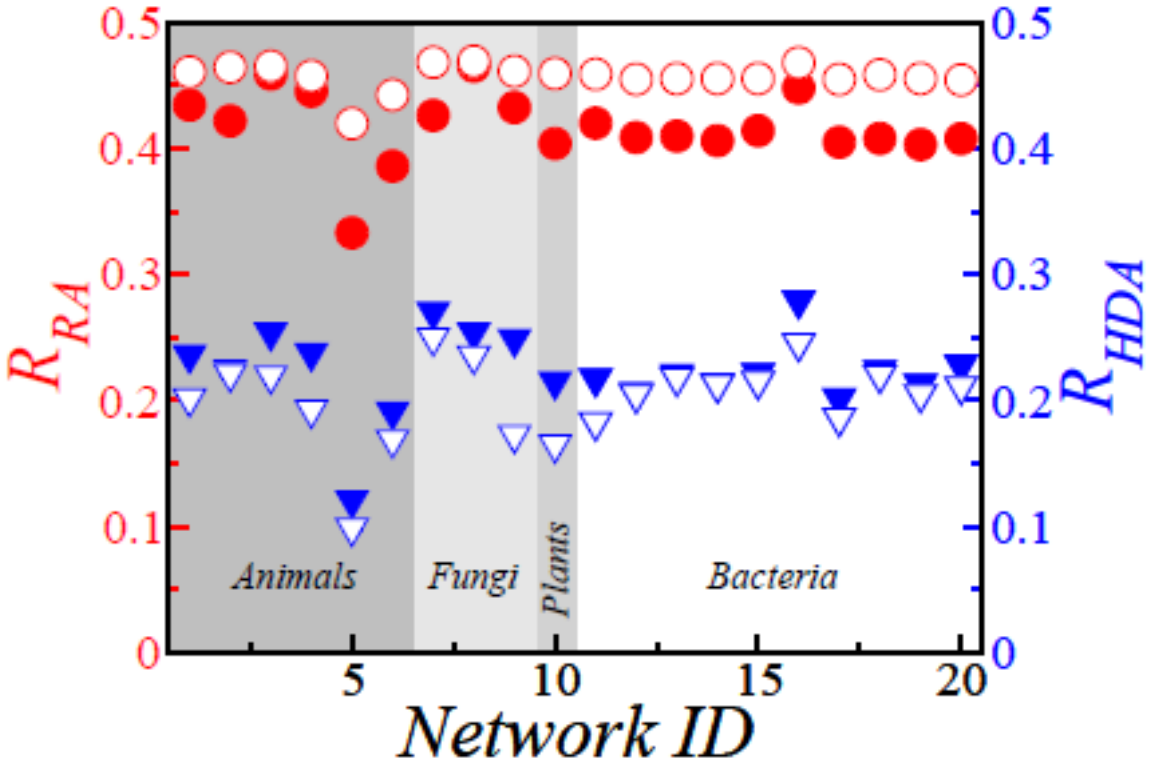
- Protein Interaction Networks (PIN): main source of information for cellular processes.
- If two proteins are present in a same reaction within the organism they maybe linked in a network representation.
- PPI of 20 different organisms in the bacteria and eukarya domains
- Submit each network to a series of malicious and random attacks.
- *R* measure the network robustness.

Fragility of protein network

- STRING 8.2 database
- Combined Score (*CS*) as a measure of the likelihood that two proteins interact in a given network.
- Threshold value $CS_{th} = 70\%$.
- Smaller values produce dramatic growth in numbers of edges, masking relevant information with extraneous information, while larger CS_{th} may exclude known protein interaction
- Compare results with ER surrogates with same number of nodes and edges

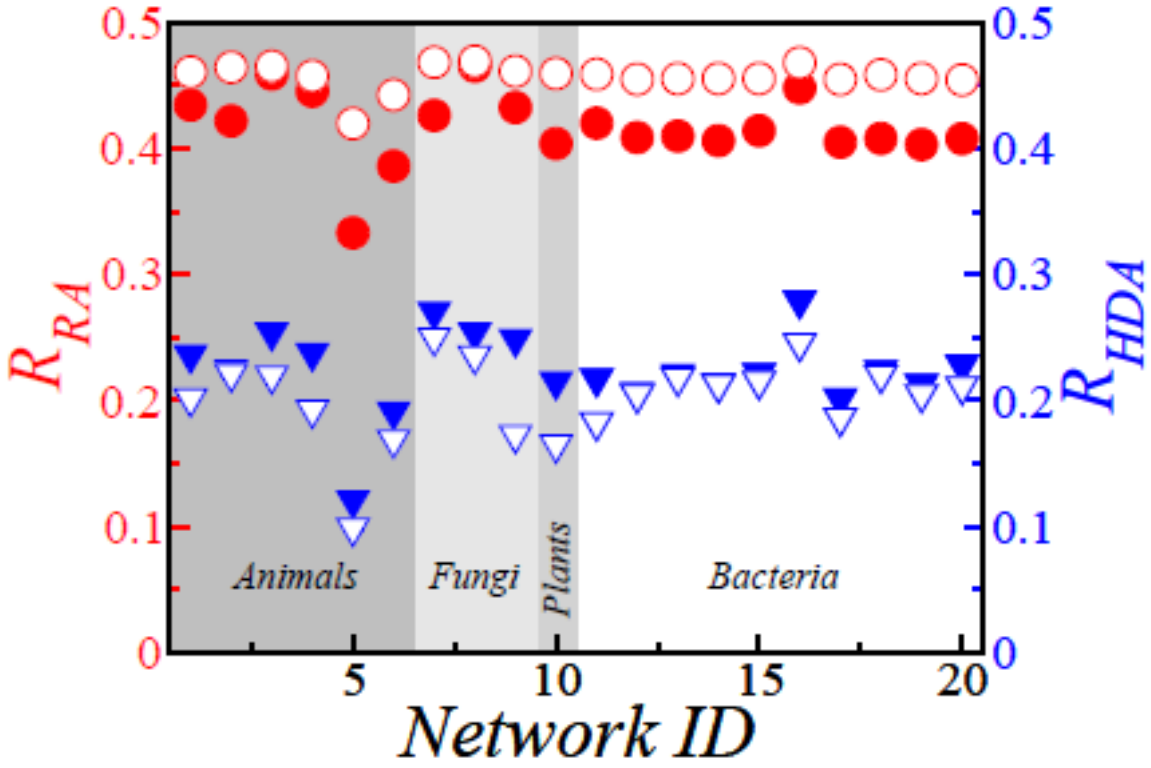
Results

- Robustness of PPI against random and malicious attacks.
- Solid and open symbols correspond to biological data and surrogates.



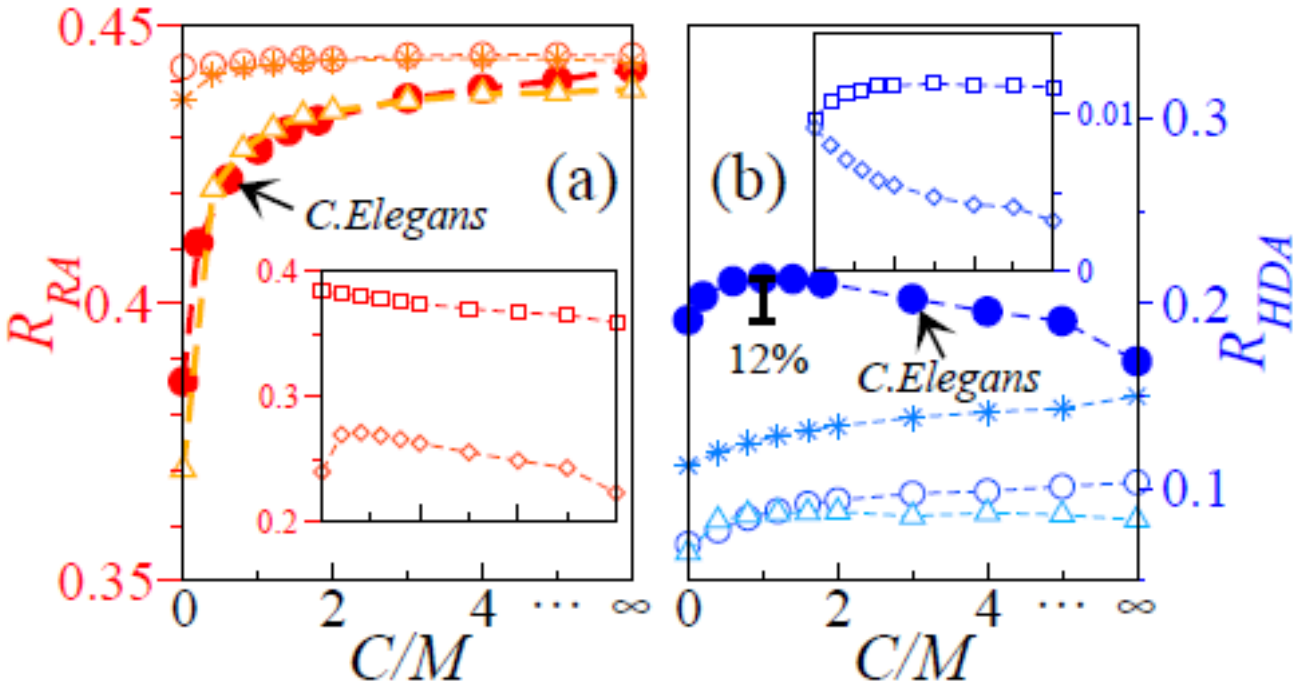
Results

- Robustness against random attacks R_{RA} is smaller than surrogates with identical degree distributions,
- Robustness against malicious attacks R_{HDA} is larger than surrogates.



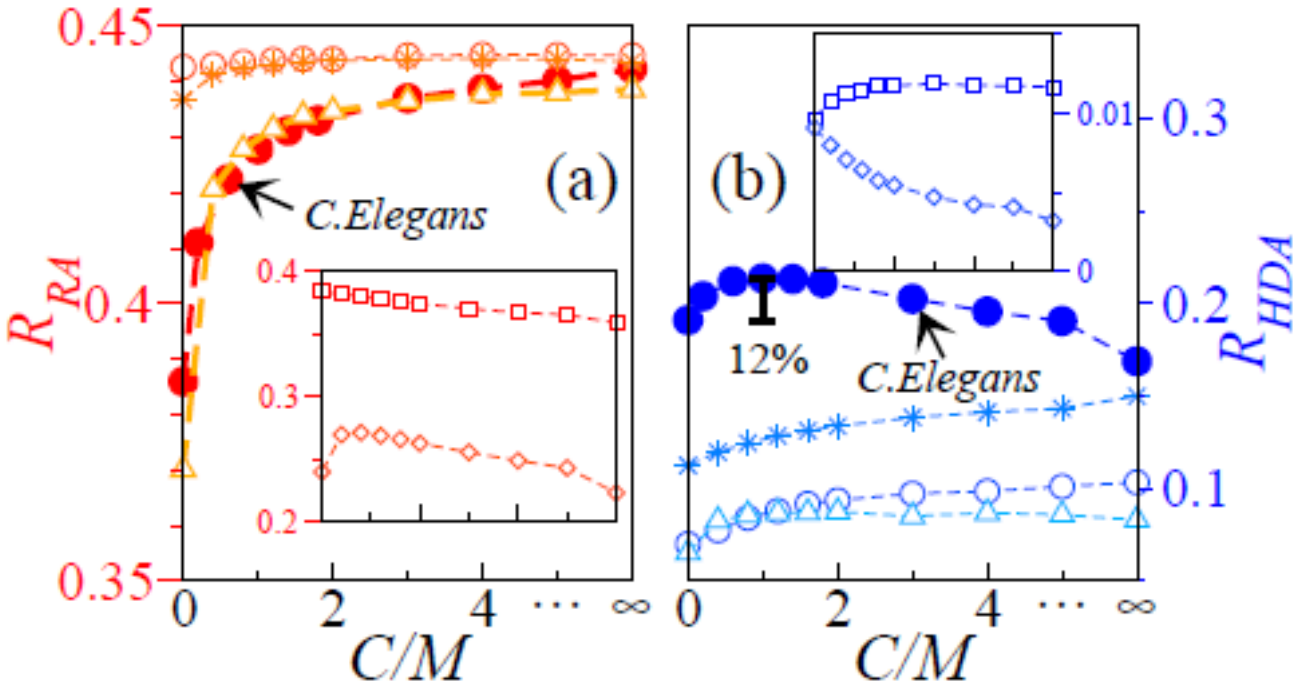
Results

- Paradoxical behavior can be analyzed by evolution of the behavior of *R* for original and randomized networks (measured by *C/M*)
- Highlights the different behavior of fragile and robust networks



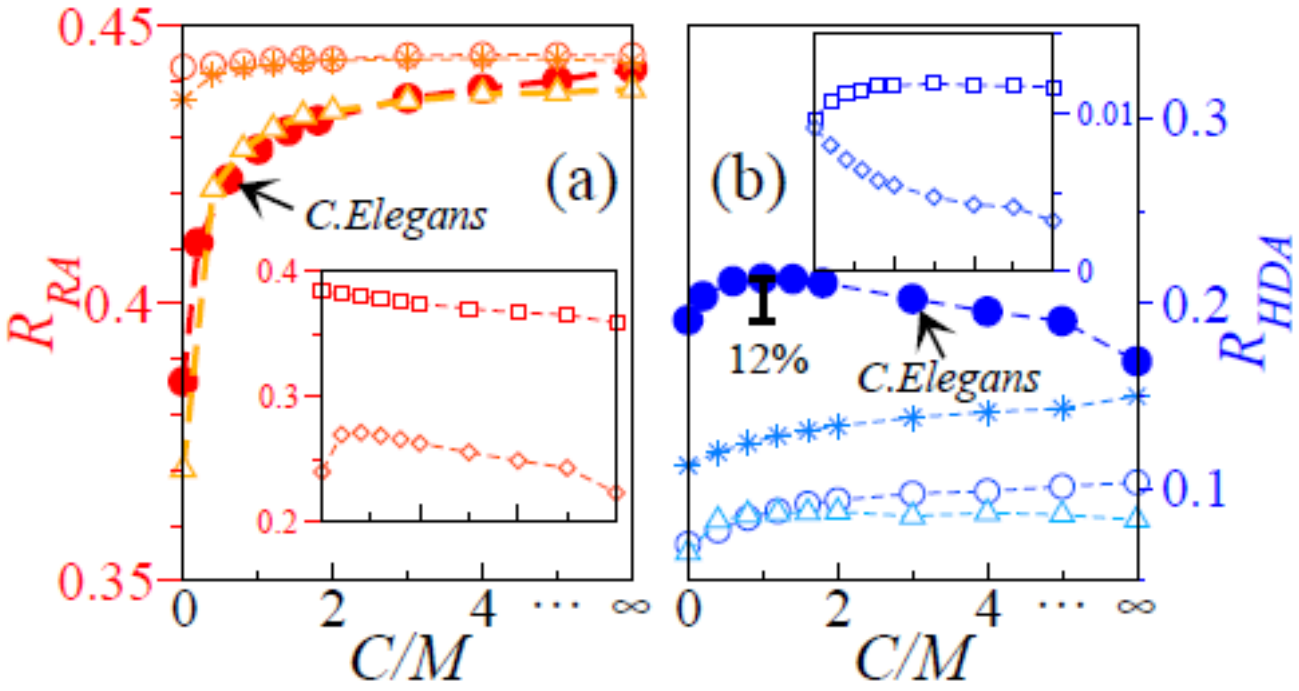
Results

- R_{RA} increases with *C/M* for C. Elegans (●), air-line (Δ), citation (∗), and PoP networks (○).
- R_{RA} decreases with *C/M* for the Internet (□) and corporate ownership network (◊).



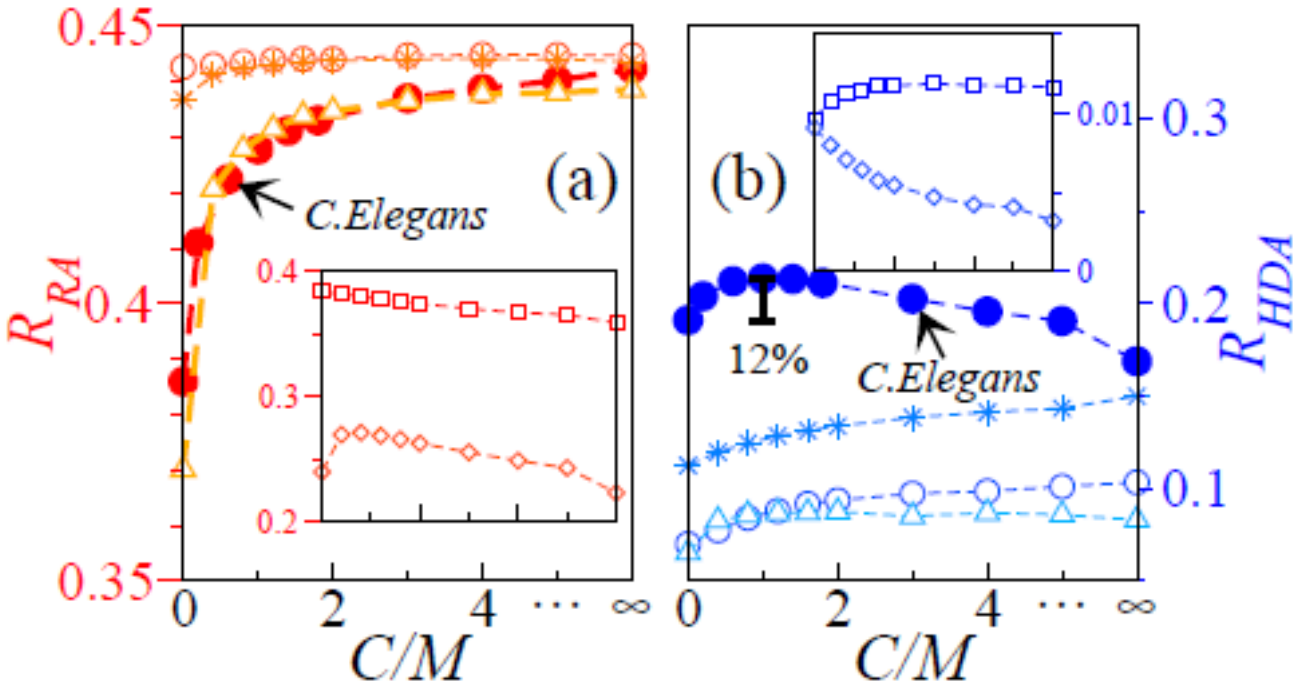
Results

- Improvement R_{RA} significantly larger for C. Elegans and airline networks (modular), in opposition to Internet



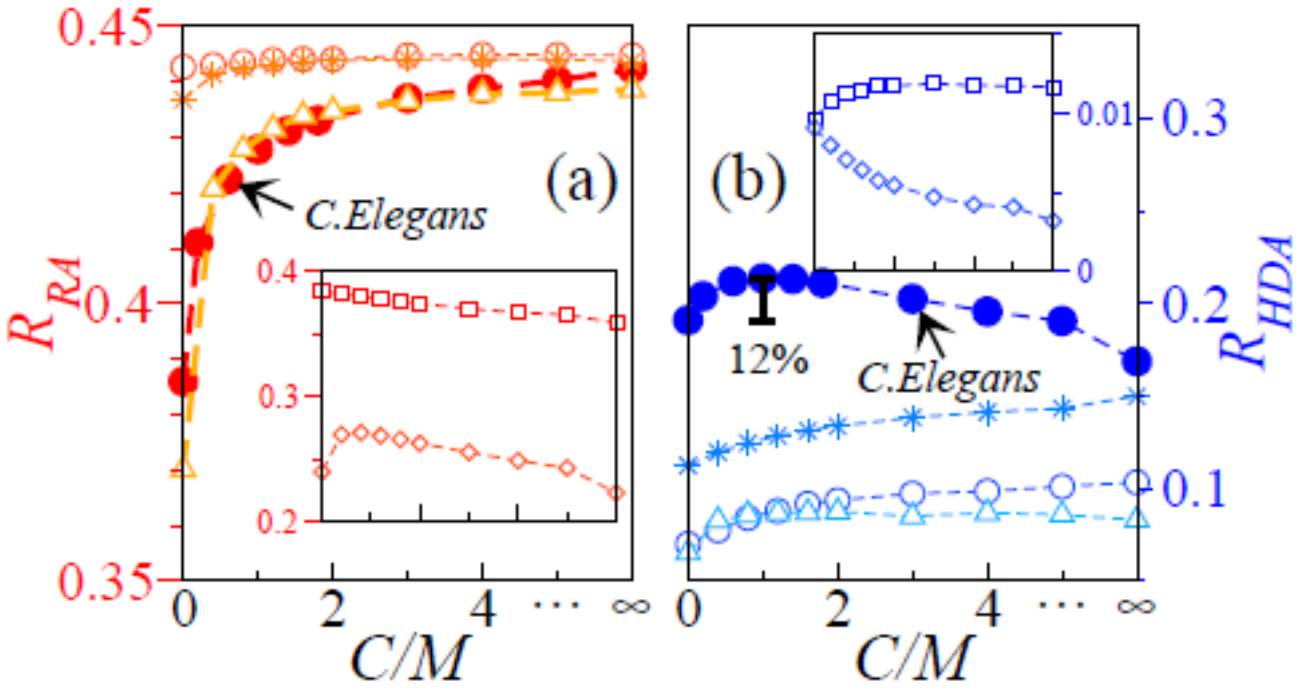
Results

- R_{HDA} also differs between biological and other networks.
- For biological networks, it increases with *C/M* up to *12%* until *C/M~1*, but then decreases



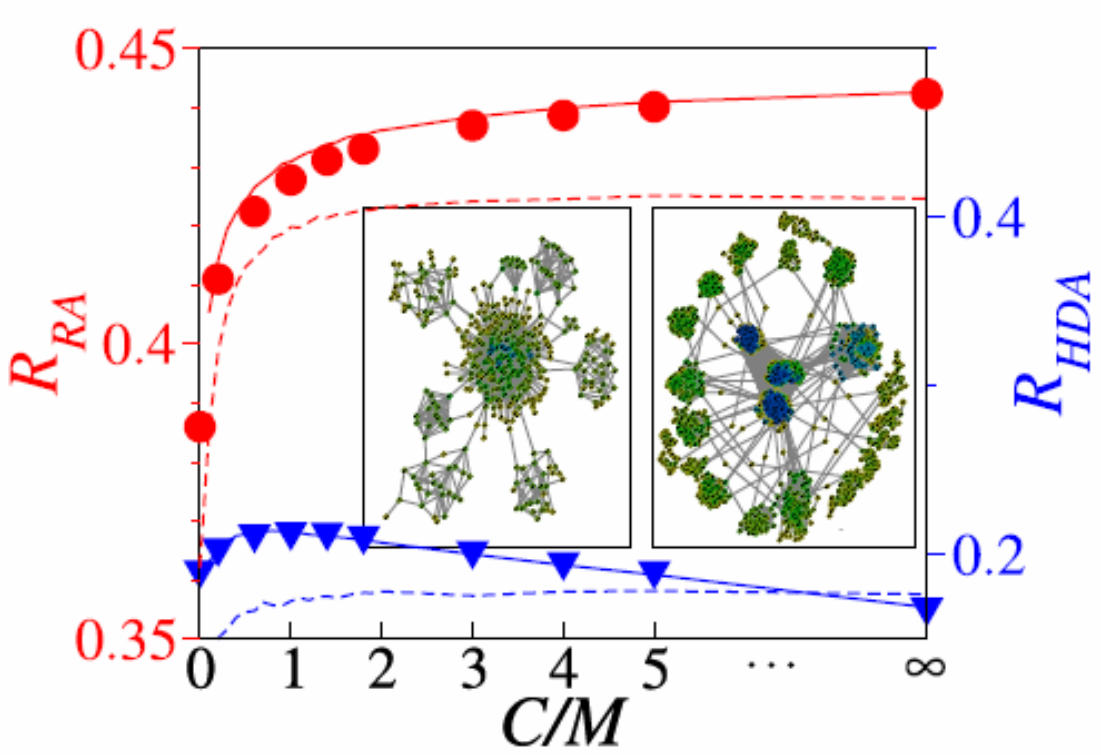
Results

- For all other networks, R_{HDA} monotonically increases with *C/M*.
- Exception only for the ownership network.



Results

- Effects of modularity on robustness in model (curves) and a sample biological networks (C. Elegans – data points).
- Simple network (left figure, dashed curves).
- More complex network (right figure, solid).



Conclusions

- The application of a complex network approach to the comparative analysis of protein sequences of chitin metabolic pathway resulted in the identification of modularity (communities) in a critical region of similarity threshold
- Communities (modules) were automatically revealed by calculating edge betweenness, and a highly significant and remarkably agreement between modules and phylogeny of organisms was retrieved.

Conclusions

- Robustness and fragility of PPI and other biological networks may help understand evolutionary processes and strategies.