A complex network approach can reveal phylogenetic relationships of extant organisms

FESC

Roberto F. S. Andrade
Leonardo B. L. Santos
Charles Santana
Suani T. R. de Pinho
José G. V. Miranda
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

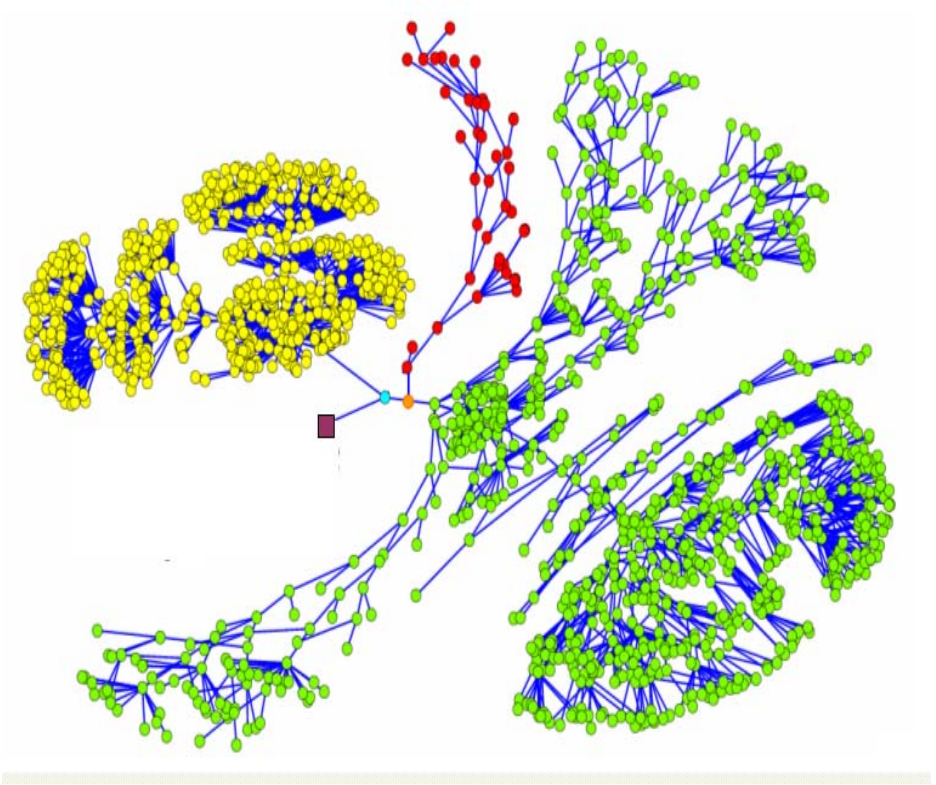
Instituto de Matematica,
Universidade Federal da Bahia

Outline

- Phylogenetic classification
- Protein and molecular synthesis
- Protein networks
- Results
- Conclusions and perspectives

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)
- Heavily relies on qualitative biologic features as input to substitution matrices
- Large computing facilities

Protein and molecular synthesis

- Several organic bio-molecules required for basic purposes are 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 pathways to some basic molecular synthesis

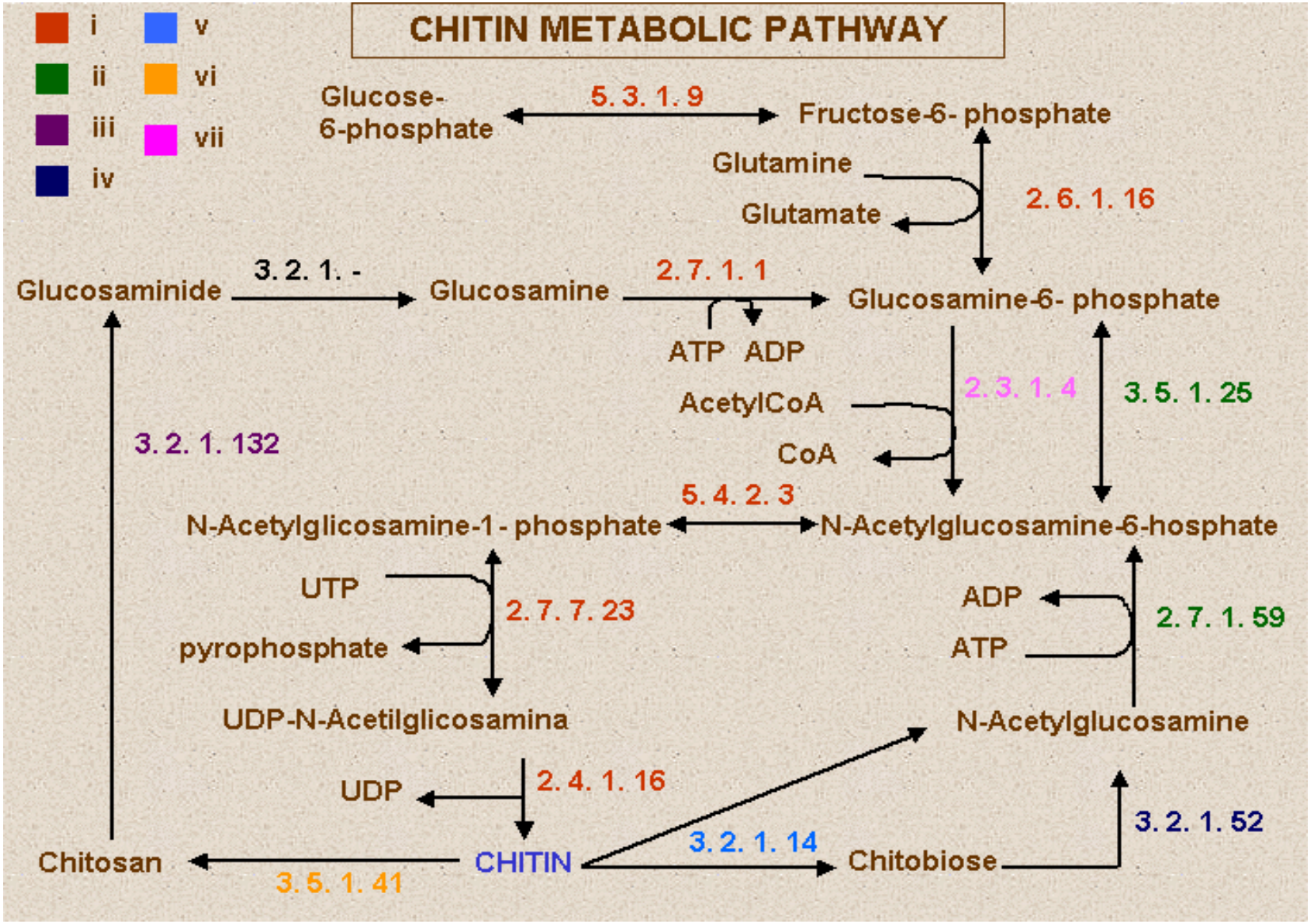
Protein and molecular synthesis

- This work: similarity based on protein data from completely sequenced genomes of extant organisms is used for network construction
- Classification: distance between networks

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 Genebank (May 19, 2007)
- Extract 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

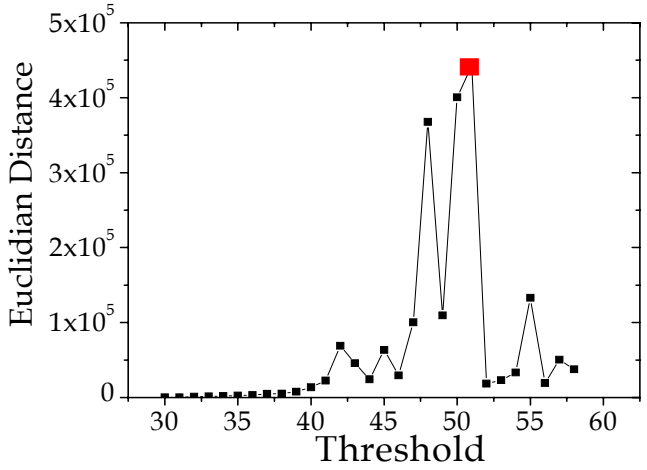
- Comparison of protein sequences for organism sequences based on similarity index (S) BLAST (v. 2.2.15) $\Rightarrow$ similarity matrix (SM)
- Symmetrization of SM
- Symmetrized SM leads to undirected network adjacency matrix AM
- Network nodes i represent sequenced organisms
- Nodes i,j are connected if similarity index S_{ij} is above a pre-established threshold S_{th}

Results

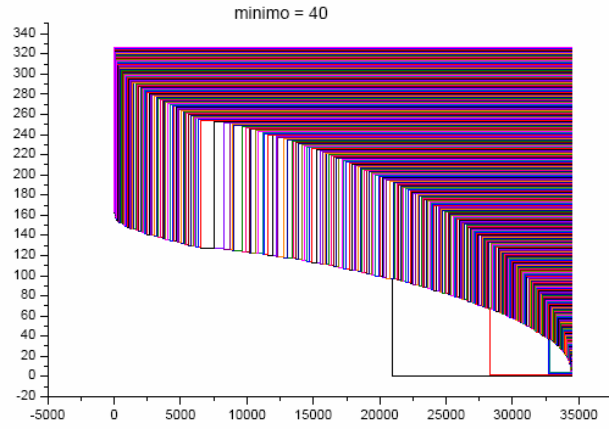
- Network measures:
  - Degree distribution $P(k)$
  - Clustering coefficient C
  - Average path-length $\langle d \rangle$
  - Edge betwenness B
  - Network distance $D_{\alpha\beta}$
- Networks depend on S_{th}
- Judicious choice of value of S_{th} optimizes reliability of classification scheme

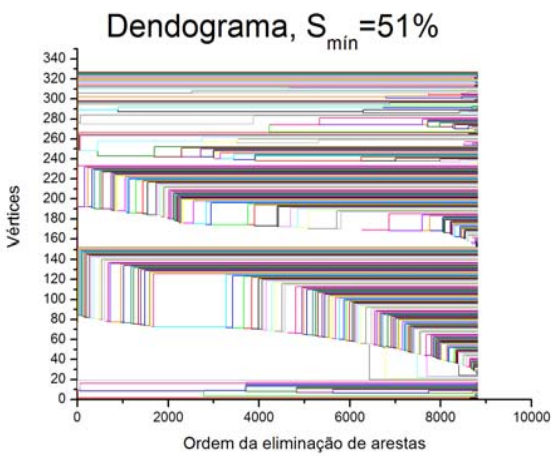
Results

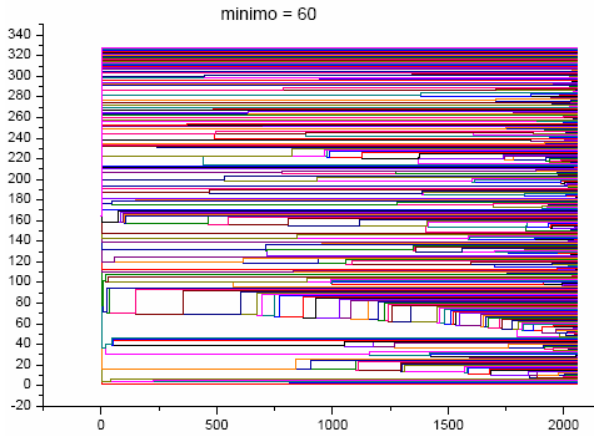
- Enzyme UDP
- $S_{th} \approx 51\%$: sudden transition in network properties
  - Sharp decrease in <d>
  - Clustering C remains relatively unchanged
  - Sharp change in dendrogram based on B
  - Peak in the distance $D_{\alpha,\alpha+1}$



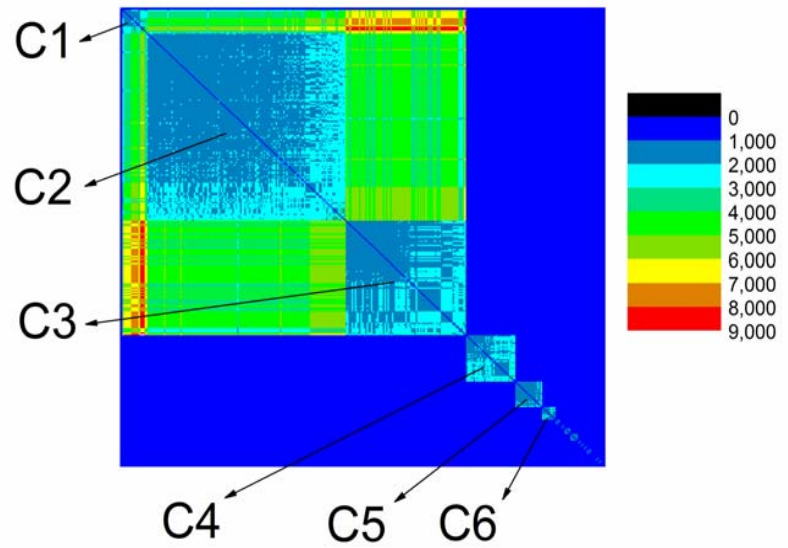
Results





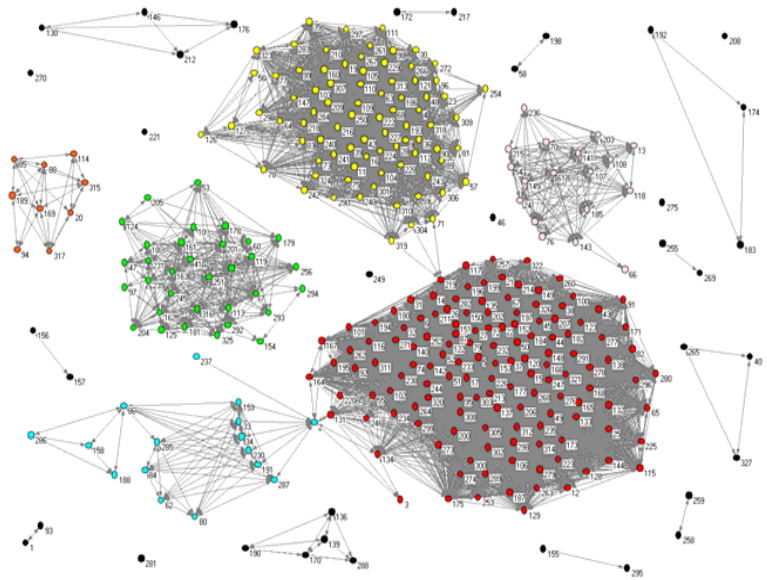


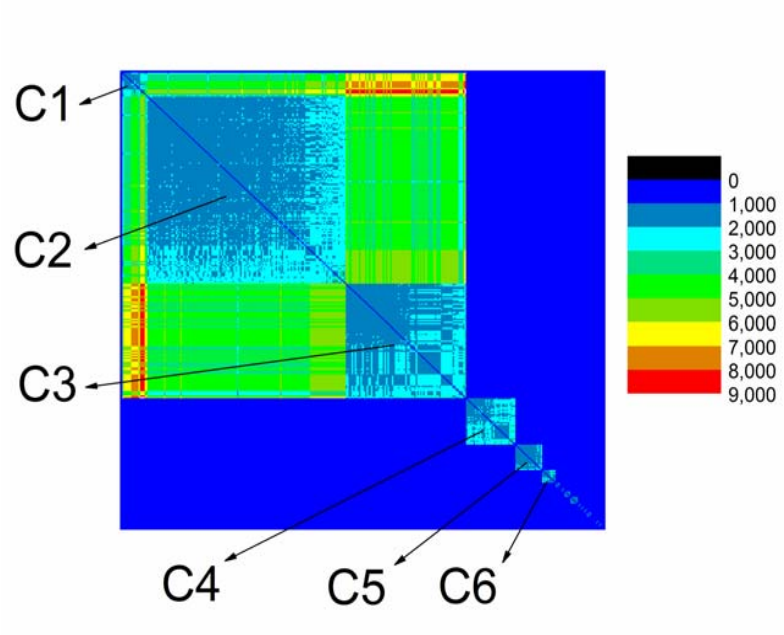
- $D_{\alpha,\alpha+1}$ is reflected in the dendrogram structure
- At S_{th}=51%, main groups identified are reproduced in neighborhood matrix
- Moduli C1-C6 with precise biologic meaning.



Results

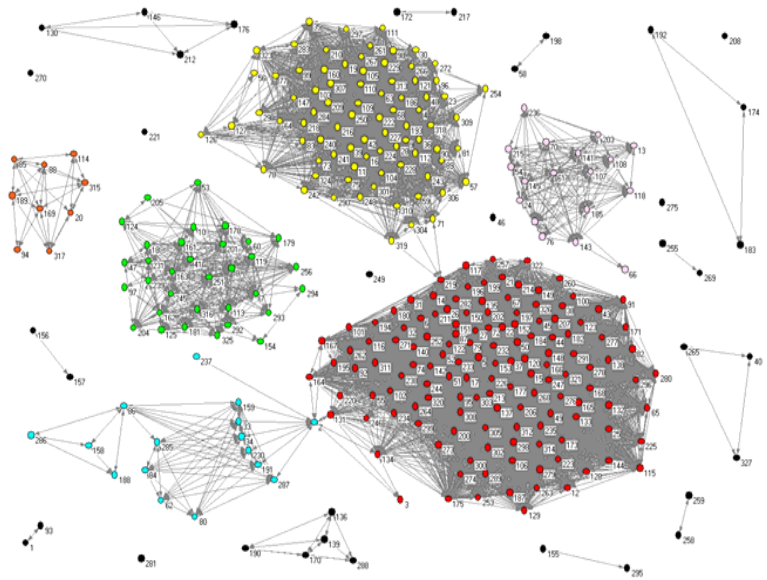
- C1 – Cyanobacteria C2 – Firmicutes
  C3 – β and γ Proteobacteria C4 – α-Proteobacteria
  C5 – Actinobacteria C6 – ε-Proteobacteria





Results

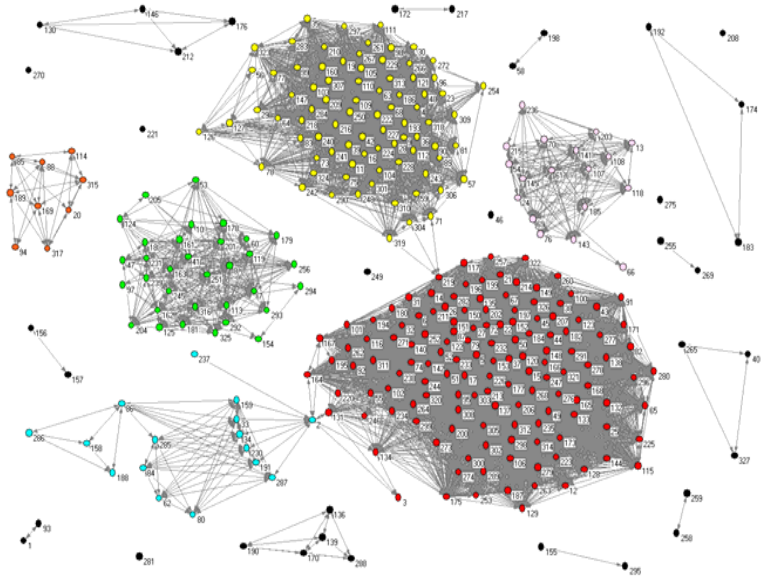
- C1 – Cyanobacteria C2 – Firmicutes
  C3 – β and γ Proteobacteria C4 – α-Proteobacteria
  C5 – Actinobacteria C6 – ε-Proteobacteria



- 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

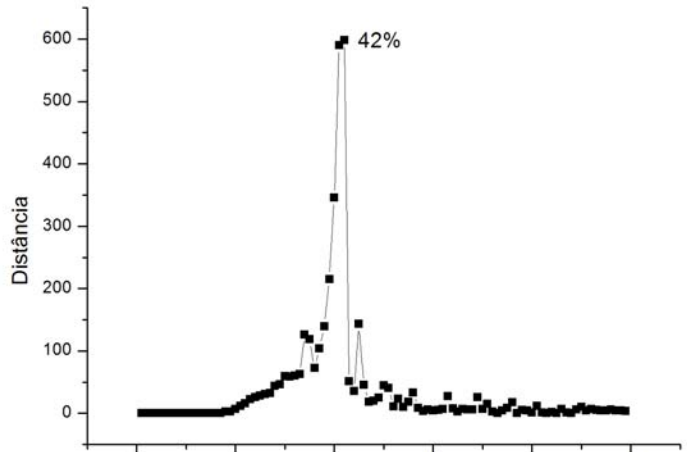
- Same method was applied to other networks (with no. of vertices ≥ 100) $\Rightarrow$ accurately defined grouping suggests robustness of the method.

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

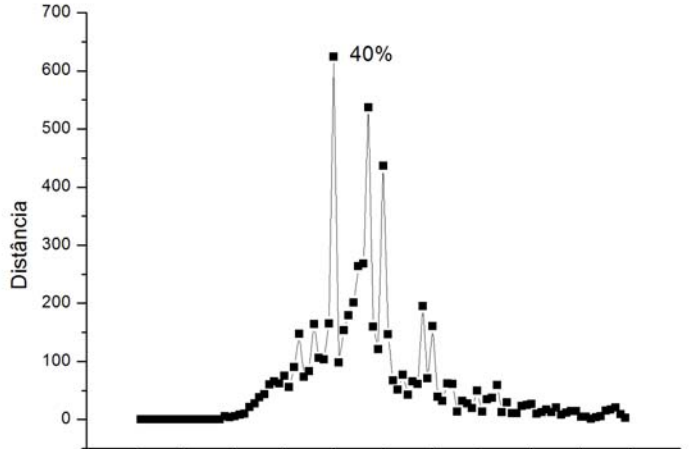
Results

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

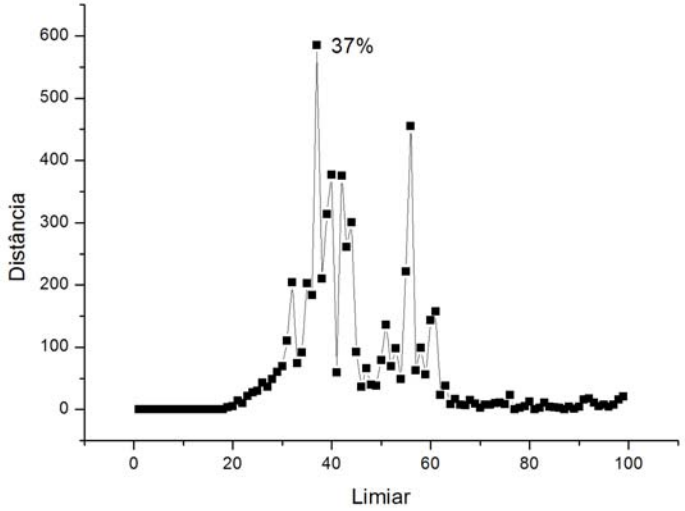
Acetyl



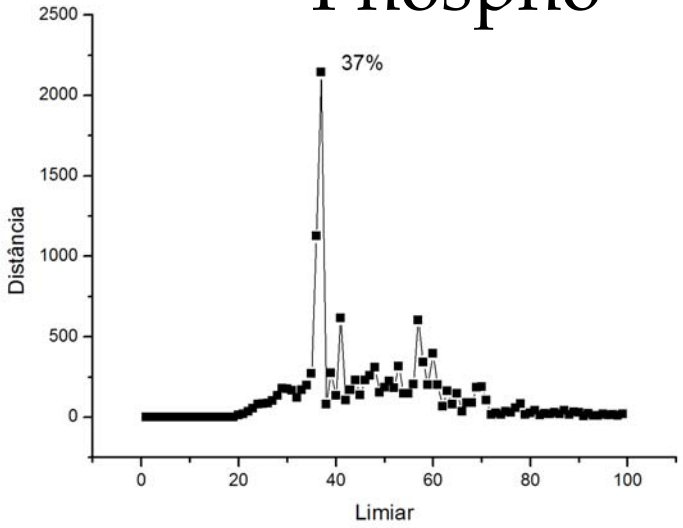
Gluco



Hexo



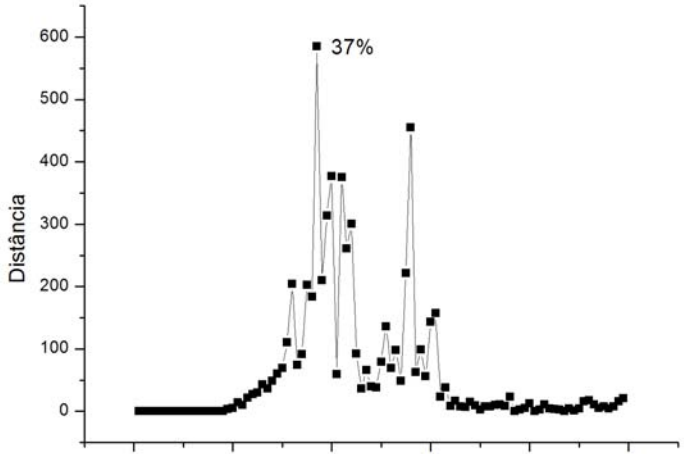
Phospho



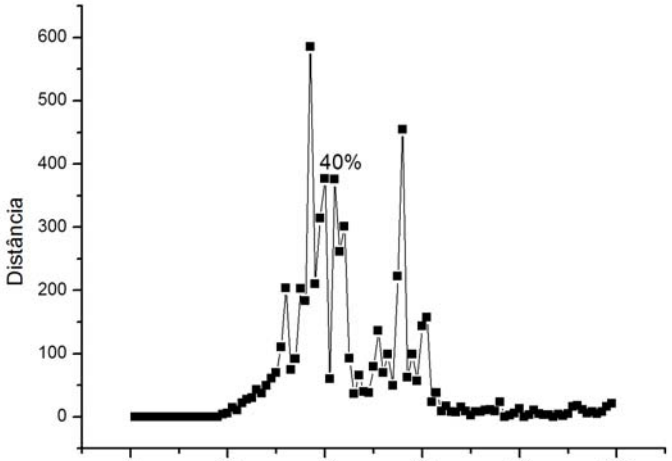
Results

- Hexo: Dependence of network on S_{th}

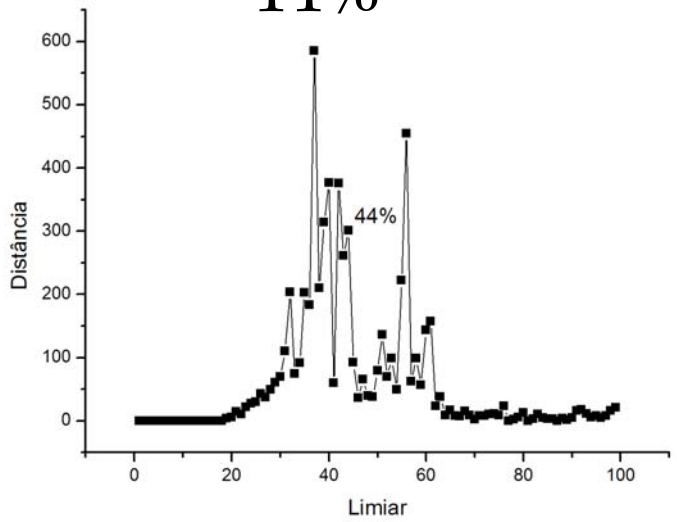
37%



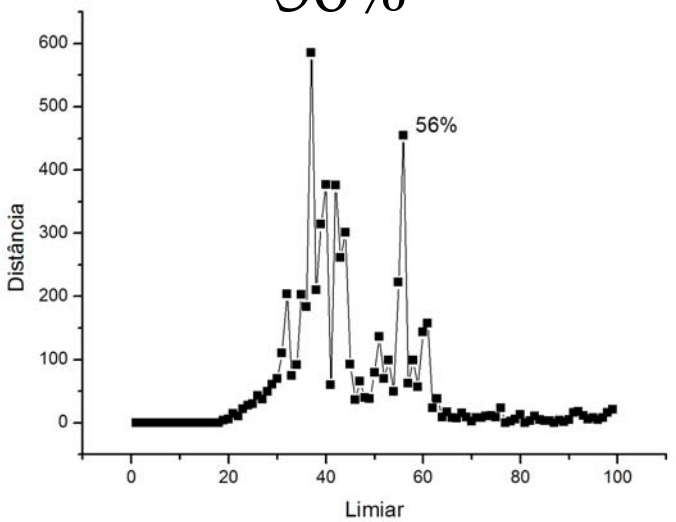
40%



44%



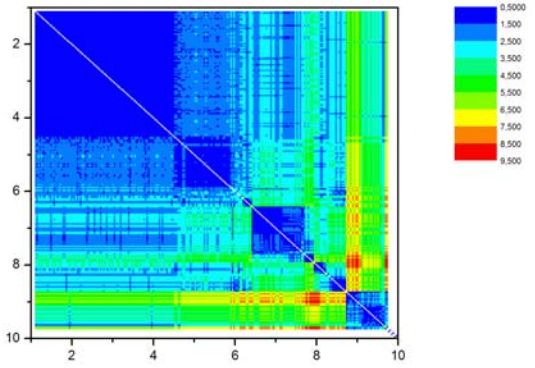
56%



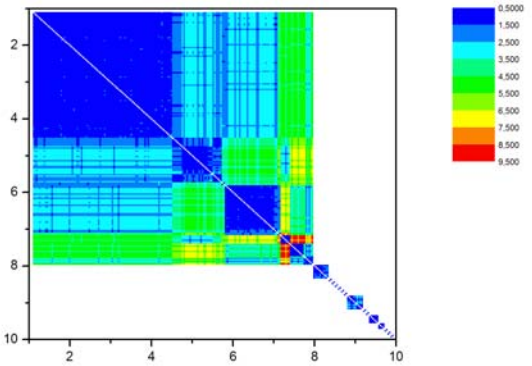
Results

- Hexo: Dependence of network on S_{th}

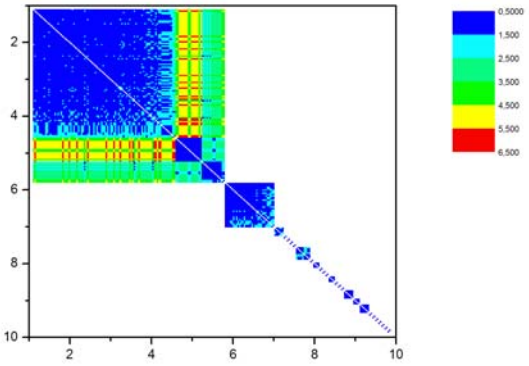
37%



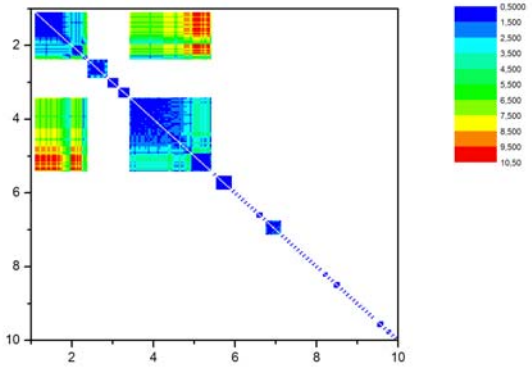
40%



44%



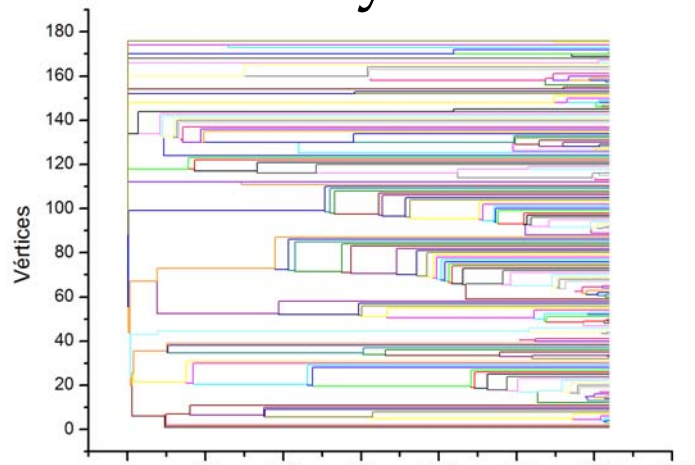
56%



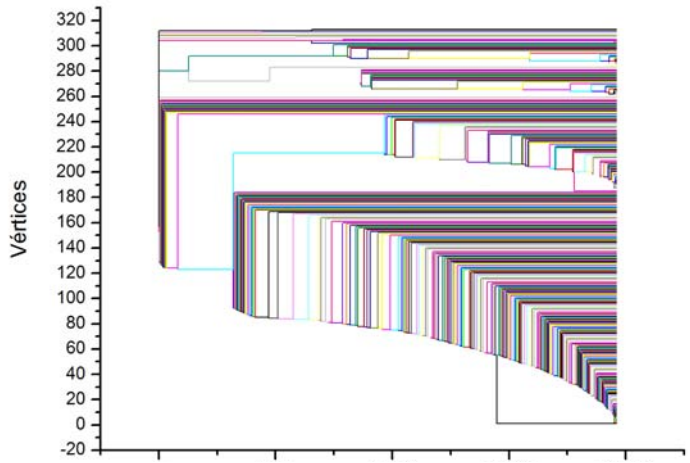
Results

- Dendrograms at first threshold S_{th}

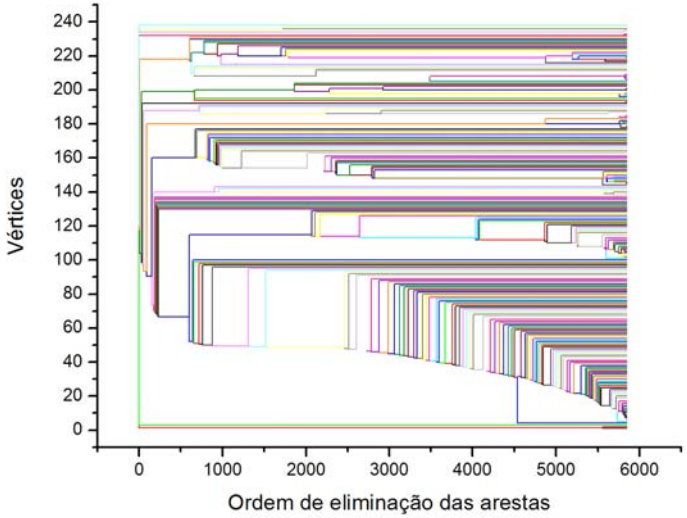
Acetyl



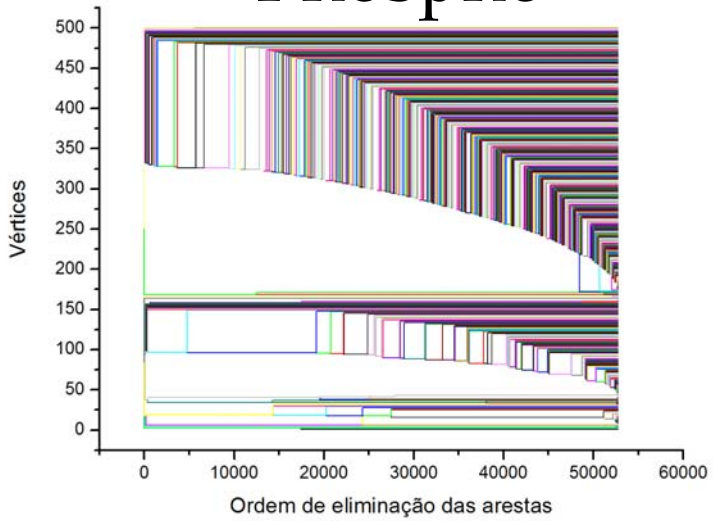
Gluco



Hexo



Phospho



Results

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

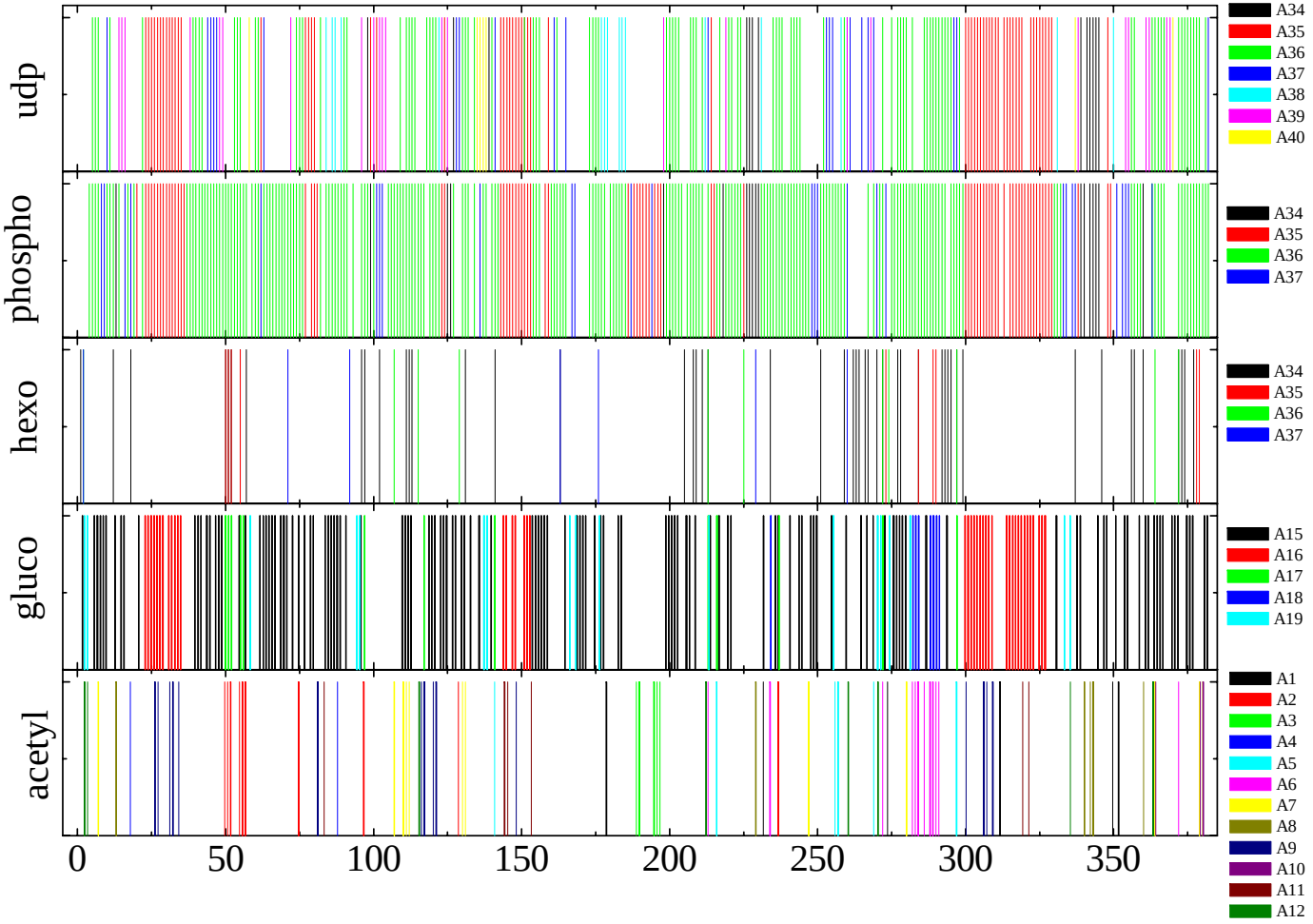
Enzime	<SIM>	σ	S_t	# Diferents sequences	# Diferents phylum
UDP-acetylglucosamine pyrophosphorylase	39	15.91	51	327	14
Acetylglucosamine phosphate deacetylase	34	11.21	42	176	12
Glucosaminephosphate isomerase	37	15.16	40	313	20
Hexosaminidase	22	21.40	36	328	13
Phosphoglucoisomerase	27	23.45	36	501	20

Results

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

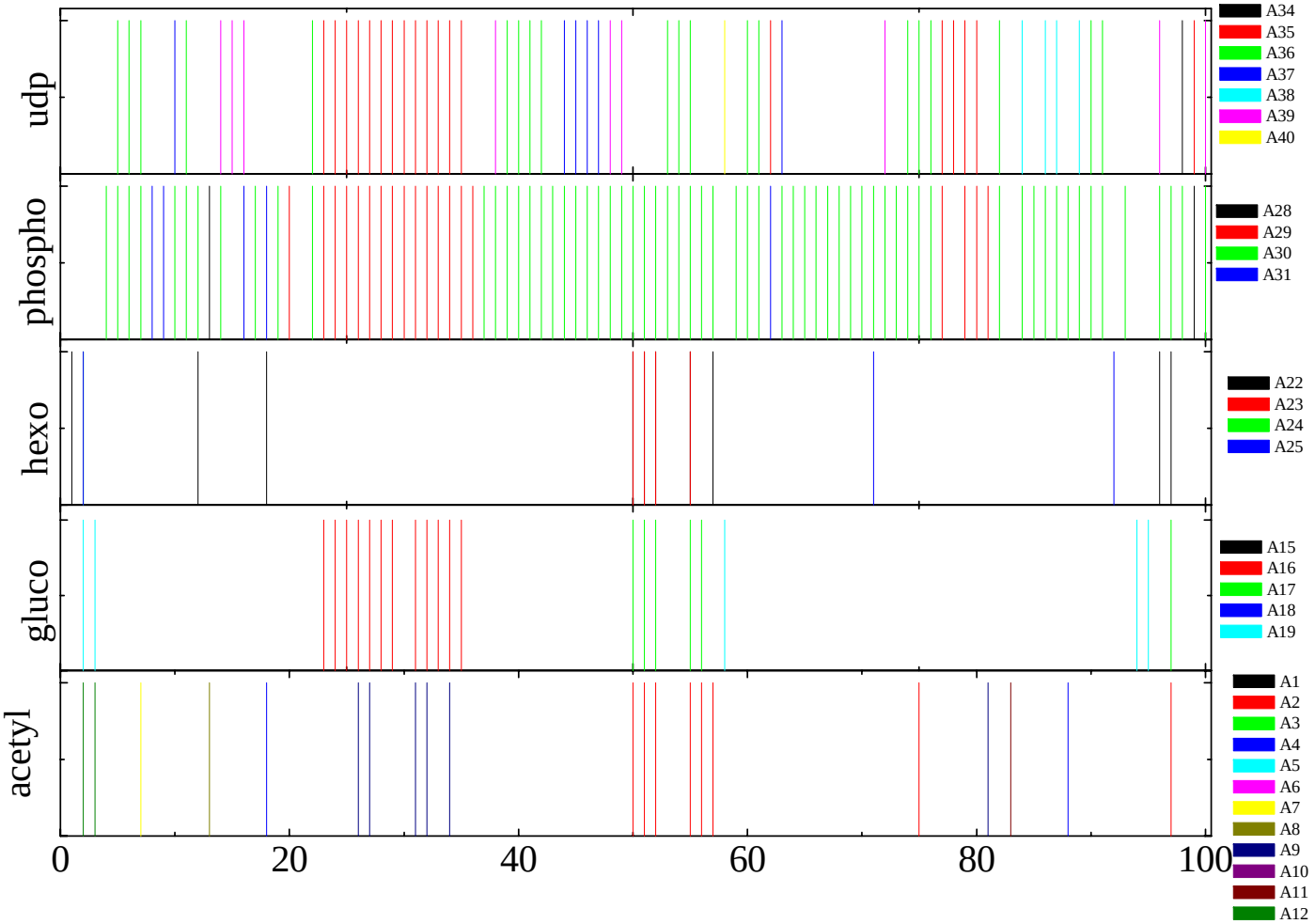
Results

- Congruence of classification by distinct networks



Results

- Congruence of classification by distinct networks



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.