



Recovering evolutionary  
history by complex network  
modularity analysis



Roberto F. S. Andrade



Roberto F. S. Andrade, Suani T. R. de Pinho,  
Aristóteles Góes-Neto, Charbel N. El-Hani, Thierry P.  
Lobão, Gilberto C. Bomfim, Marcelo V. C. Diniz,  
Leonardo B. L. Santos, Charles Santana, Ivan Rocha,  
Daniel S. Carvalho, Arthur M. Y. R. Sousa, André P.  
Vieira, Carmen P. C. Prado



Roberto F. S. Andrade, Suani T. R. de Pinho,  
Aristóteles Góes-Neto, Charbel N. El-Hani, **Thierry P.  
Lobão**, Gilberto C. Bomfim, Marcelo V. C. Diniz,  
Leonardo B. L. Santos, **Charles Santana**, **Ivan Rocha**,  
**Daniel S. Carvalho**, Arthur M. Y. R. Sousa, André P.  
Vieira, Carmen P. C. Prado



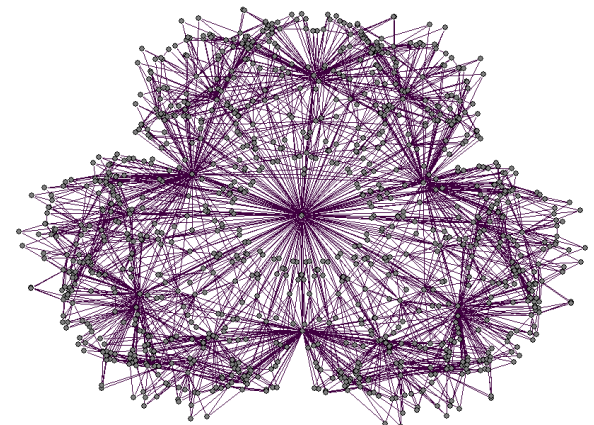
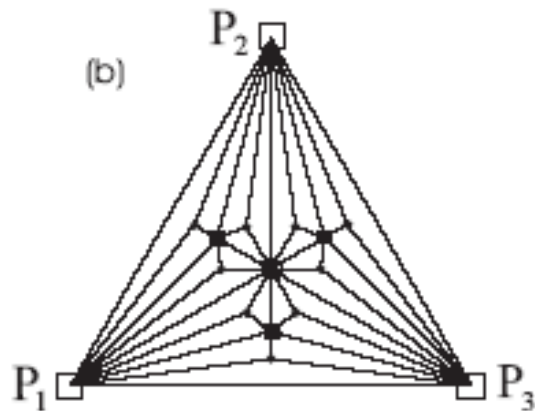
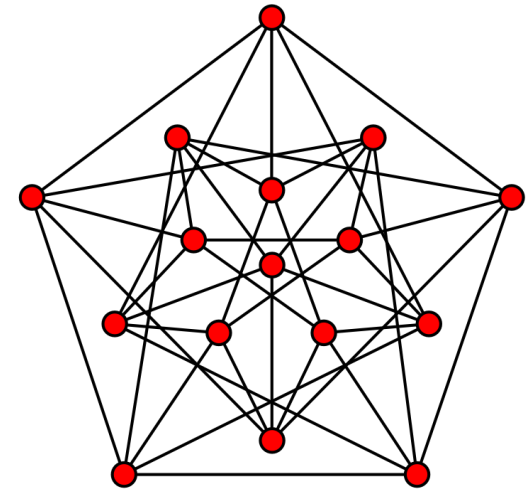
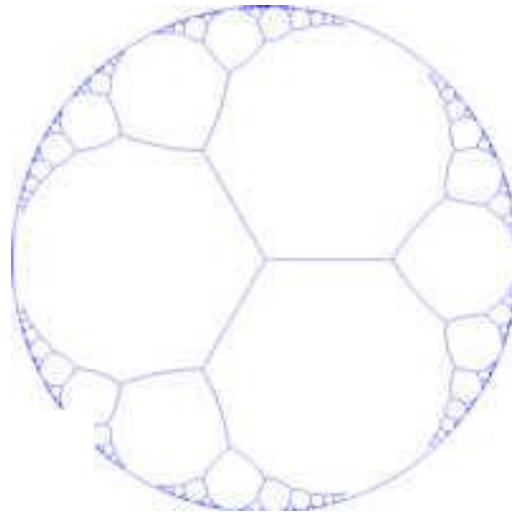
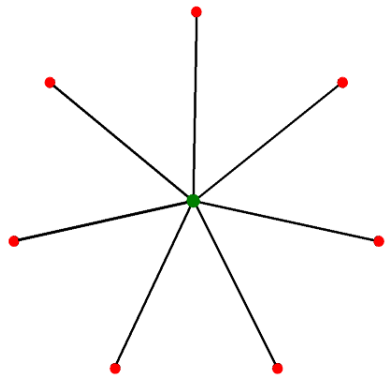
- ❑ What are the Evolutionary Origins of Mitochondria? A Complex Network Approach - PLOS ONE 0134988, 2015
- ❑ Detecting Network Communities: ...Phylogenetic Analysis - PLOS Computational Biology 7, e1001131, 2011
- ❑ Comparative protein analysis of the... : A complex network approach - BioSystems 101, 59–66, 2010
- ❑ Measuring distances between complex networks - Physics Letters A 372, 5265-5269, 2008
- ❑ Neighborhood properties of complex networks - Physical Review E 73, 046101 2006



# Outline

- Complex network representation
- Adjacency matrix, neighborhood matrix
- Distance between networks
- Distance between weighted networks
- Phylogeny and evolution
- Results
  - Chitin synthesis enzymes (pathways)
  - Chitin pathways in fungi: comparison to other methods
  - Evolutionary origins of mitochondria
  - Evolutionary history recovery
- Conclusions

# Complex network representation





# Complex network representation

## ■ Network measures

- Degree  $k_i$  and degree distribution  $P(k)$
- Node clustering coefficient  $c_i = 2n_i / k_i(k_i - 1)$  and  $C = \langle c_i \rangle$
- Average path-length  $\langle d \rangle$
- Diameter  $D$
- Edge betweenness centrality  $B$
- Degree assortativity
- Modularity
- Fractal dimension



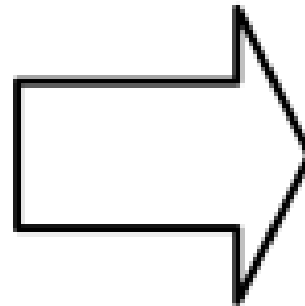
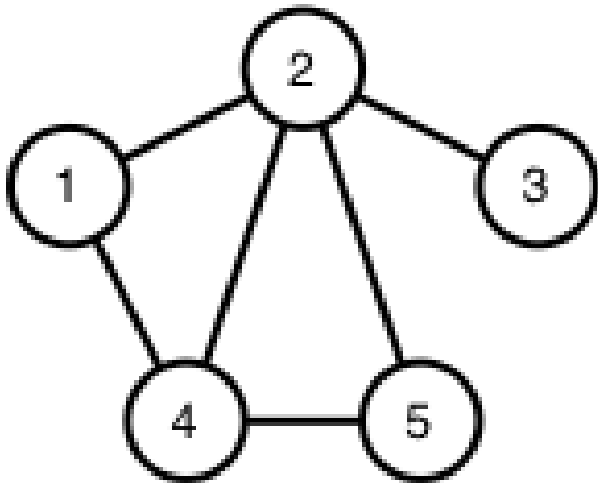
# Complex network representation

## ■ Network measures

- Degree  $k_i$  and degree distribution  $P(k)$
- Node clustering coefficient  $c_i = 2n_i / k_i(k_i - 1)$  and  $C = \langle c_i \rangle$
- Average path-length  $\langle d \rangle$
- Diameter  $D$
- Edge betweenness centrality  $B$
- Degree assortativity
- Modularity
- Fractal dimension
- Network distance  $\delta(\alpha, \beta)$
- Weighted network distance  $\delta(\sigma, \sigma + \Delta\sigma)$

# Adjacency matrix, neighborhood matrix

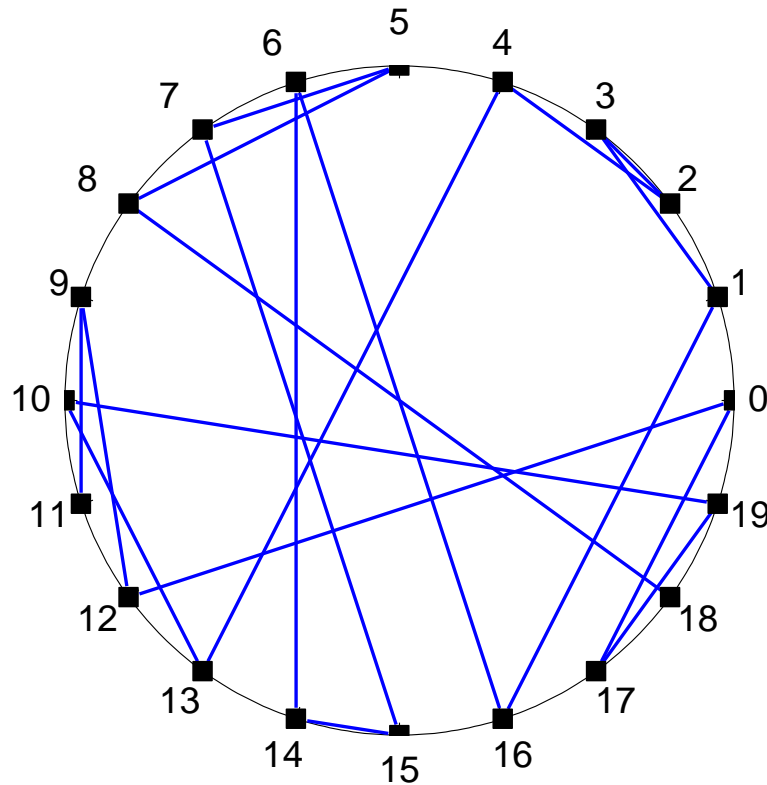
- Adjacency matrix:



$$\begin{pmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \end{pmatrix}$$

# Adjacency matrix, neighborhood matrix

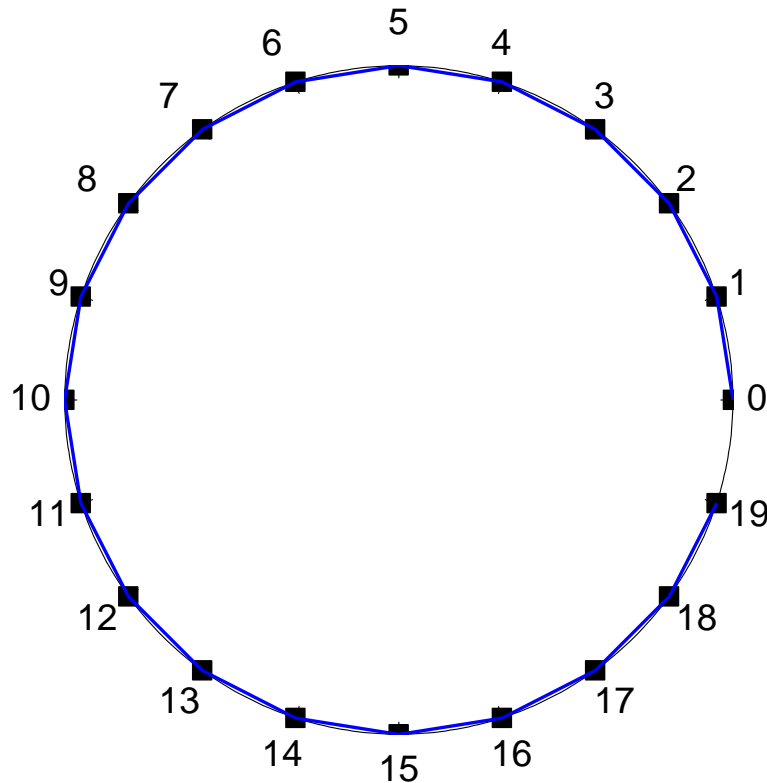
Numbering dependent



```
000000000000010000100
000100000000000001000
000110000000000000000
011000000000000000000
001000000000001000000
000000011000000000000
000000000000001010000
000001000000000100000
000001000000000000010
000000000000110000000
000000000000010000010
000000000100000000000
100000000100000000000
000010000010000000000
000000100000000100000
000000010000001000000
010000100000000000000
100000000000000000001
000000001000000000000
000000000100000001000
```

# Adjacency matrix, neighborhood matrix

Numbering dependent



```

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

# Adjacency matrix, neighborhood matrix

- Higher order  $\ell$  neighborhood evaluation
- Description by matrices  $M(\ell)$

- $\ell = 1$                        $\ell = 2$                        $\ell = 3$

```

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

```

0000010000111110
0000100111100100
0000111111011011
0000000111111111
0110010111101111
1010100011111111
0010000100010000
0111101001111100
0111110000011110
0111110100101111
1101110101011110
1011011110100110
1011110111100000
1101110111110010
1011110011110101
0011110001000010
    
```

```

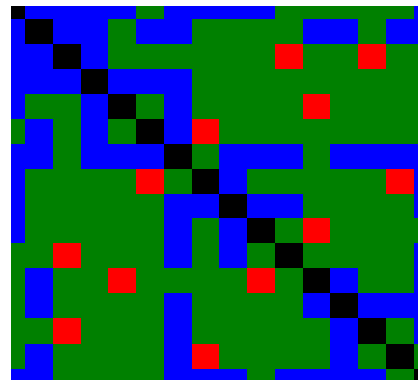
0000000000000000
0000000000000000
0000000000100100
0000000000000000
0000000000001000
0000000100000000
0000000000000000
0000010000000010
0000000000000000
0000000000010000
0010000000000000
0000100001000000
0000000000000000
0010000000000000
0000000100000000
0000000000000000
    
```

# Adjacency matrix, neighborhood matrix

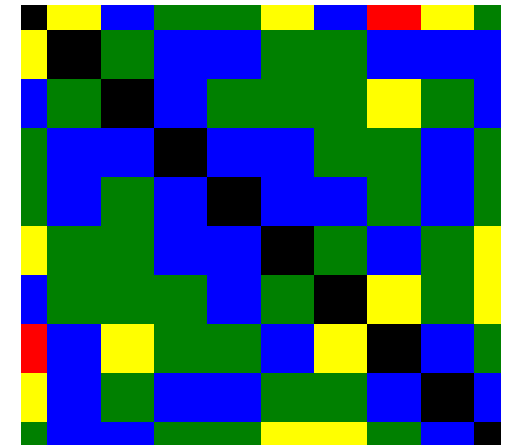
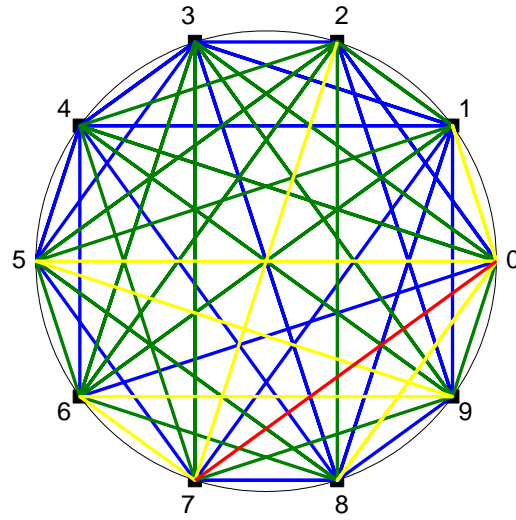
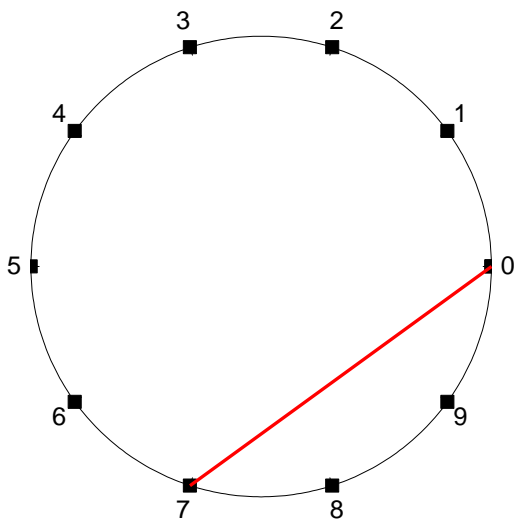
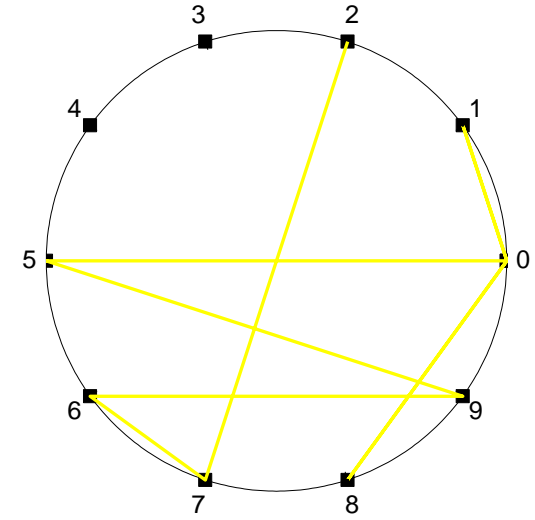
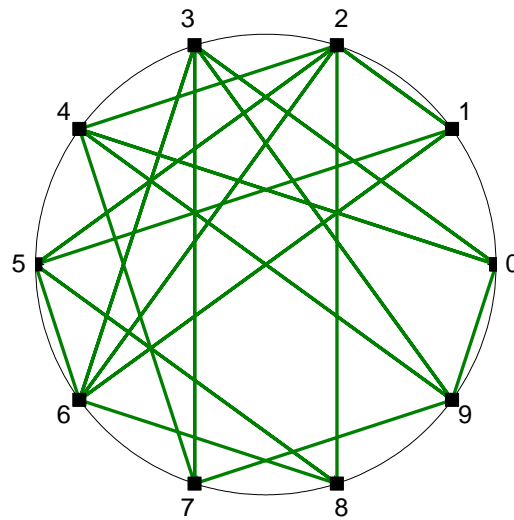
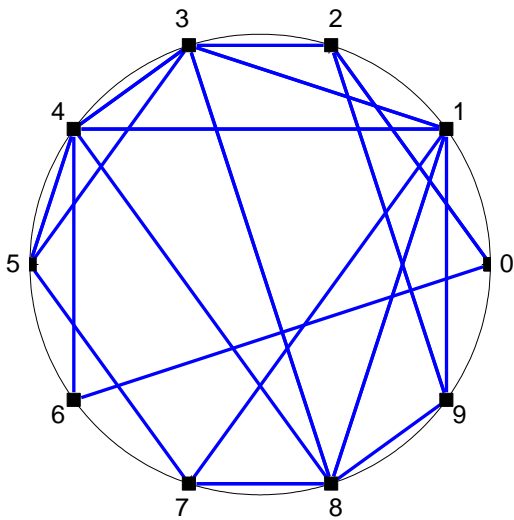
```
0111101111000001
1011011000011011
1101000000000000
1110111000000000
1001001000000000
0101001000000000
1101110011101111
1000000010000001
1000001101100001
1000001010000000
0000001010000001
0100000000001001
0100001000010111
0000001000001001
0100001000001000
1100001110111100
```

```
0000020000222220
0000200222200200
0000222222022022
0000000222222222
0220020222202222
2020200022222222
0020000200020000
0222202002222200
0222220000022220
0222220200202222
2202220202022220
20222022220200220
2022220222200000
2202220222220020
2022220022220202
0022220002000020
```

```
0000000000000000
0000000000000000
0000000000300300
0000000000000000
0000000000030000
0000000300000000
0000000000000000
0000030000000030
0000000000000000
0000000000030000
0030000000000000
0000300003000000
0000000000000000
0030000000000000
0000000300000000
0000000000000000
```

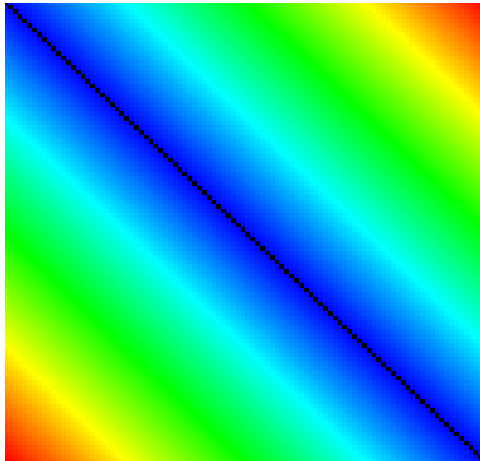


# Adjacency matrix, neighborhood matrix

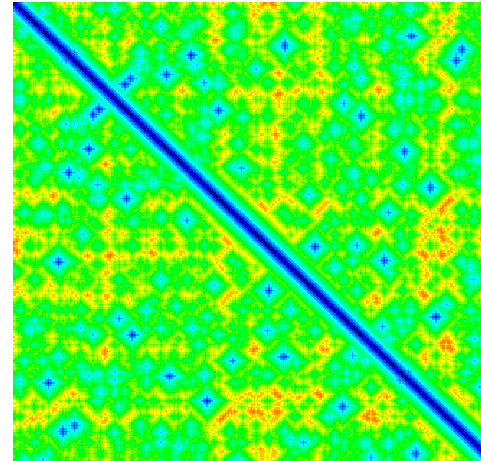


# Complex network representation

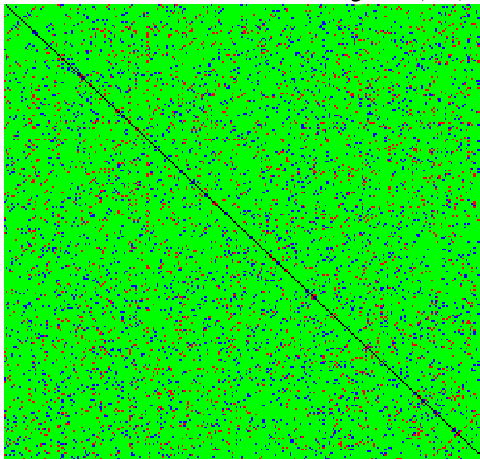
Ordered (100)



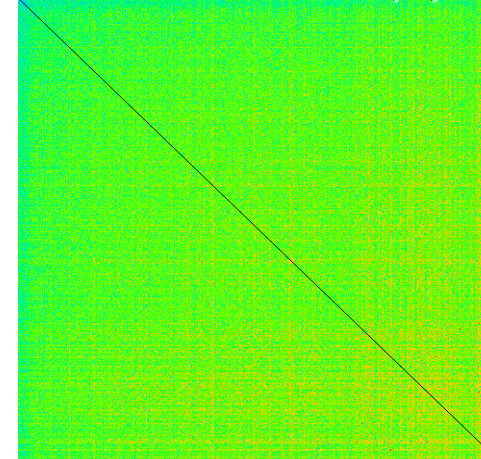
Smallworld(15)



Erdős-Rényi(3)



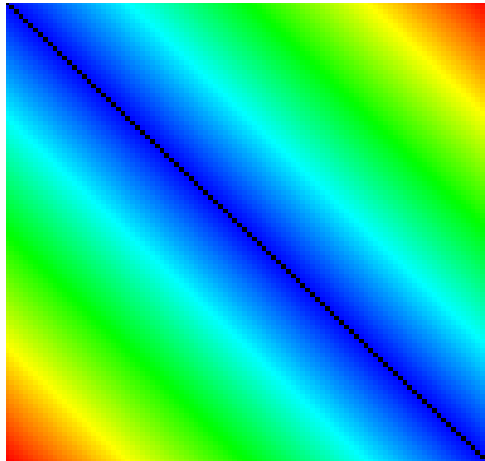
Scale-free (5)



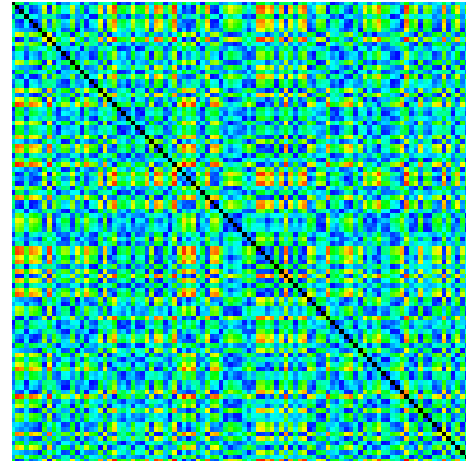


# Complex network representation

Ordered (100)

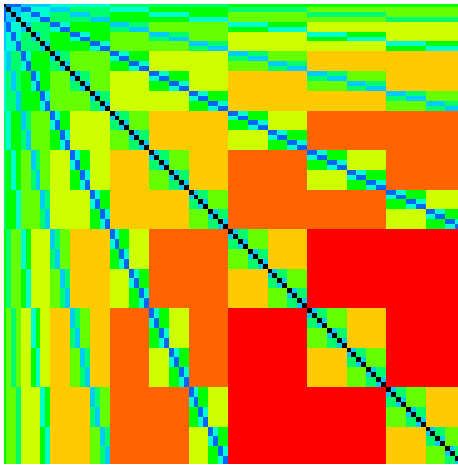


Ordered + Shuffled(100)

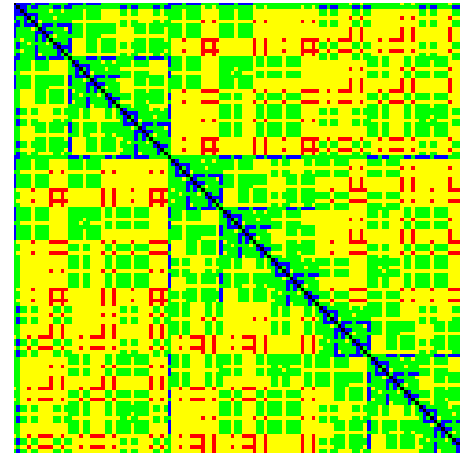


# Complex network representation

Cayley(10)



Apollonian(4)



# Distance between networks

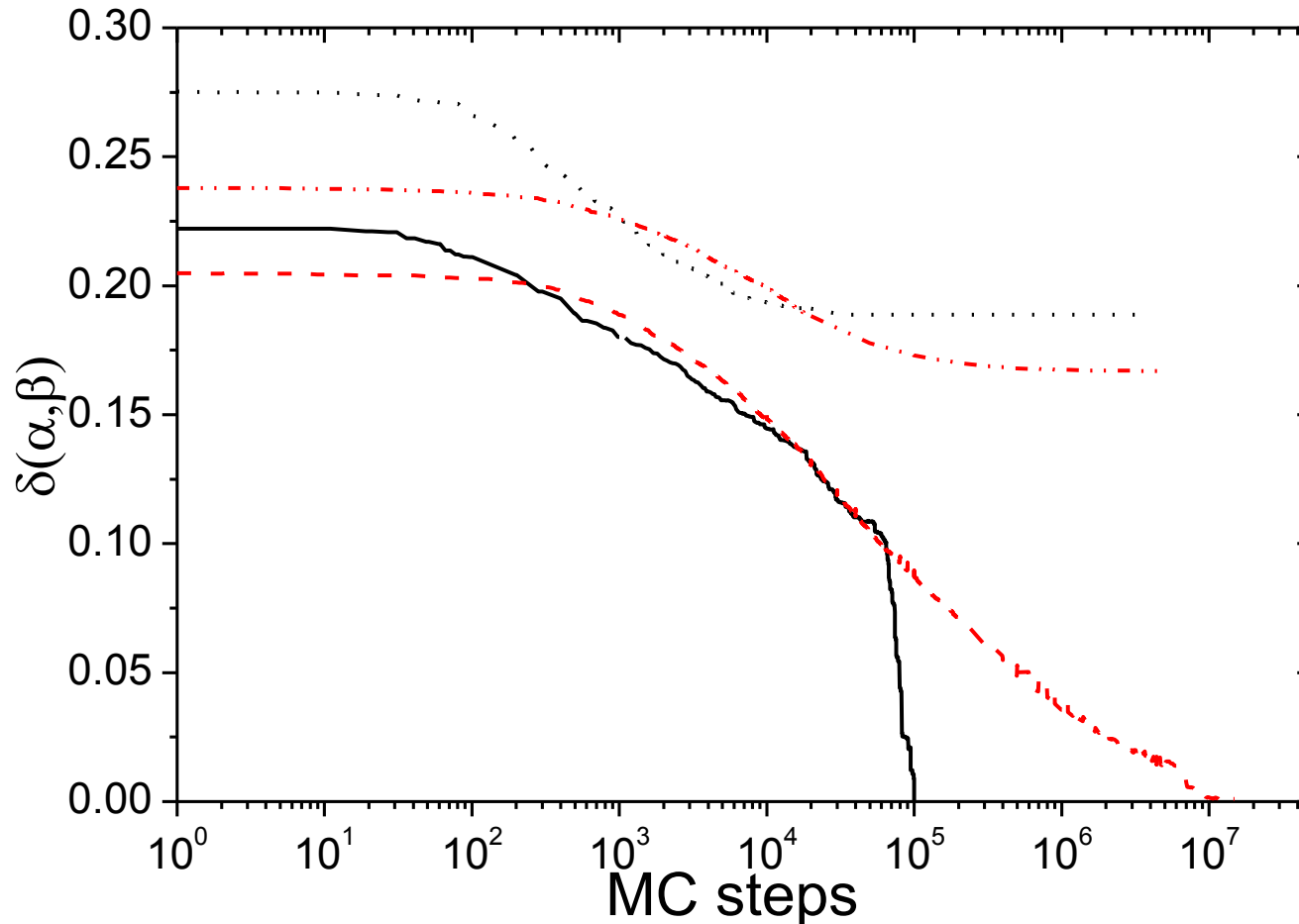
- Define a neighborhood based distance  $\delta(\alpha, \beta)$

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

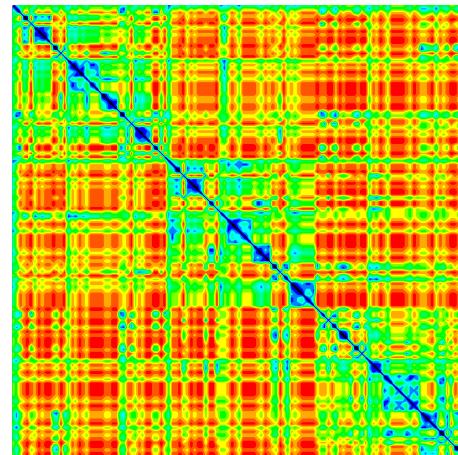
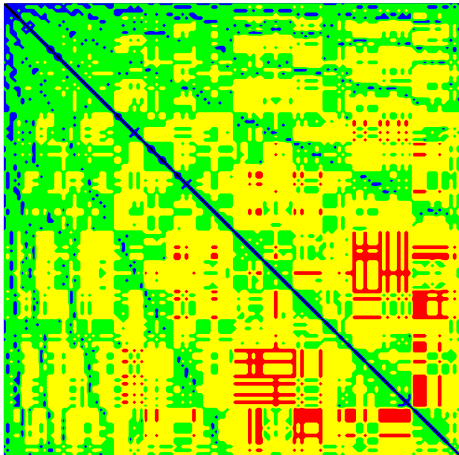
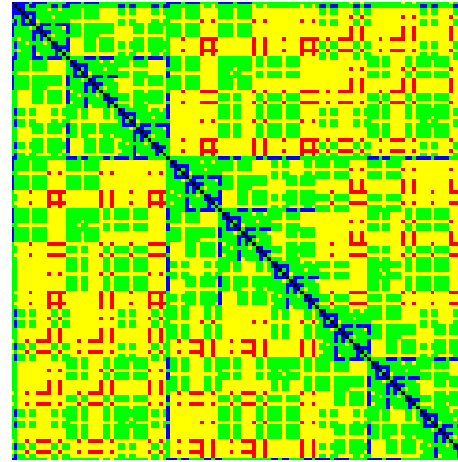
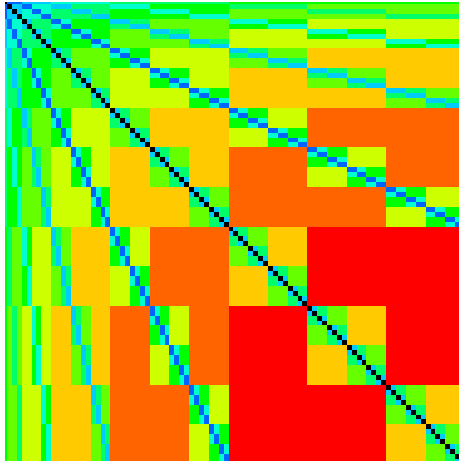
- Minimize  $\delta(\alpha, \beta)$  by Monte-Carlo procedure

# Distance between networks

- Monte-Carlo time evolution of  $\delta(\alpha, \beta)$



# Distance between networks



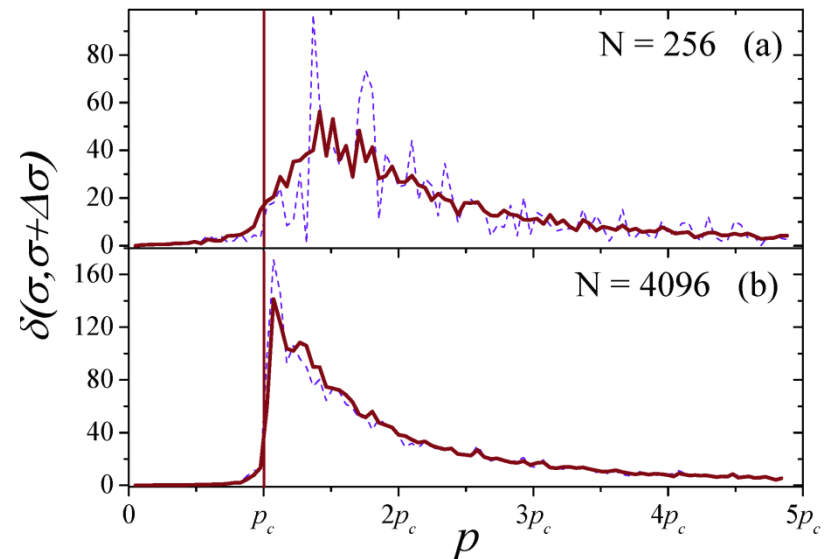
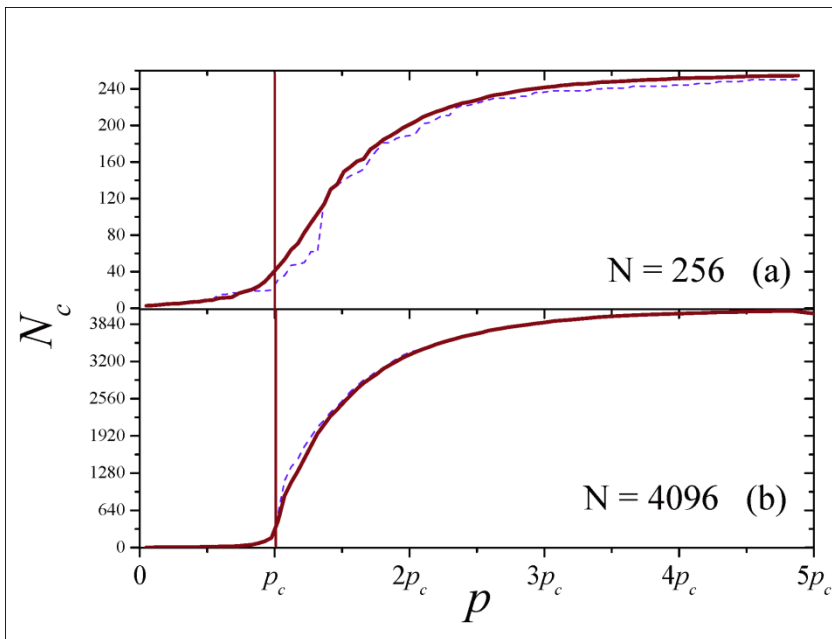
# Distance between weighted networks

- ER network with  $N$  nodes.
- Choose random weight matrix  $0 < W_{i,j} < 1$ .
- For given threshold  $p = W_{th}$  define adjacency matrix:
  - $M_{i,j}(p) = 1$ , if  $W_{i,j} < p$ .
  - $M_{i,j}(p) = 0$ , if  $W_{i,j} > p$ .
- Giant percolating cluster at  $p = p_c$ .
- Extend framework to any regular or random lattice.
- Distance between neighboring weighted networks  $\delta(p, p + \Delta p)$  detects relevant changes in network structure as function of  $p$ .

# Distance between weighted networks

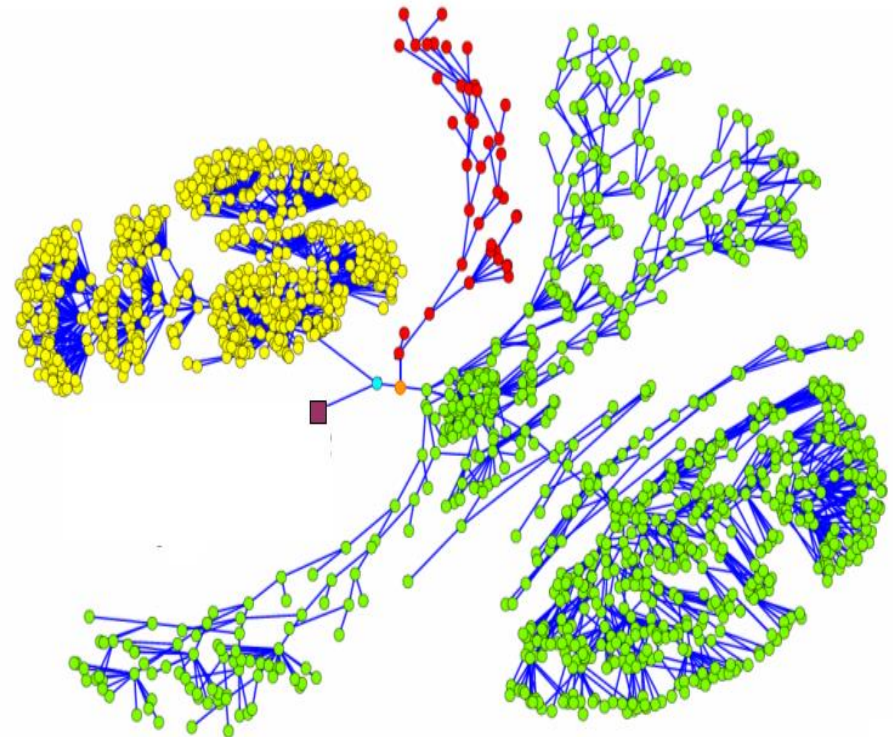
$$\delta(p, p + \Delta p) = \frac{1}{N(N-1)} \sum_{i,j=1}^N \left[ \frac{\hat{M}(p)_{i,j}}{D_p} - \frac{\hat{M}(p + \Delta p)_{i,j}}{D_{p+\Delta p}} \right]^2$$

## ■ Square lattice



# Phylogeny and evolution

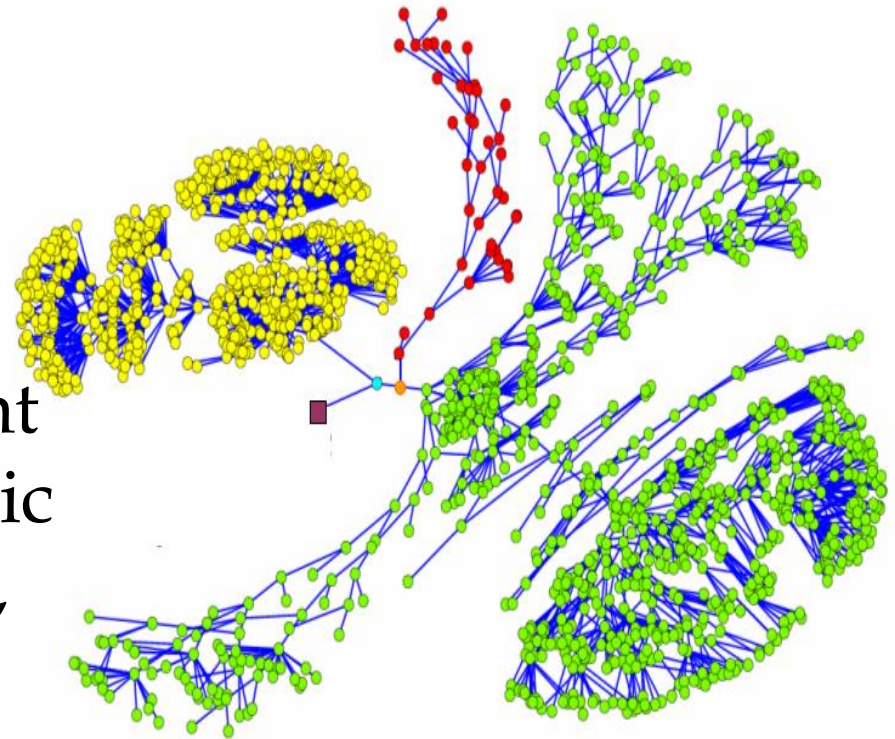
- Phylogenetic trees as “periodic tables” of biologic diversity.
- Usual classification: species, genus, family, order, class, phylum, kingdom.
- Recently introduced domains (archaea, bacteria eukarya) as basic roots of biologic evolution.





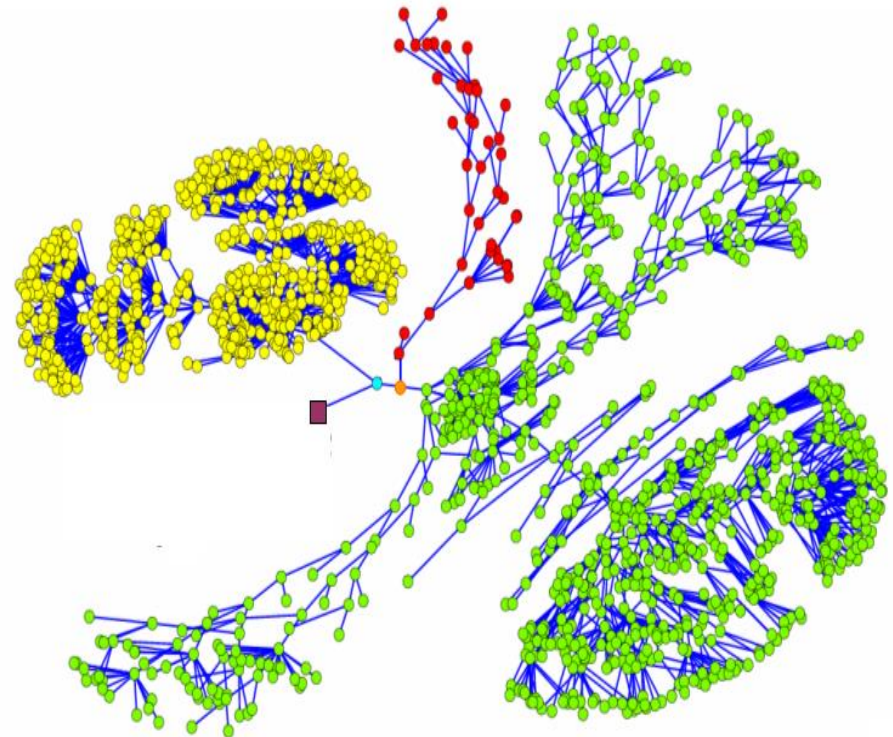
# Phylogeny and evolution

- Classical methods of phylogeny and evolution based on grouping analysis criteria:
  - Bayesian analysis;
  - Maximum likelihood;
  - Neighbor joining distance;
  - Parsimony.
- Take into account different features, from morphologic to molecular composition, structure and interactions
- Comparatively large computational efforts



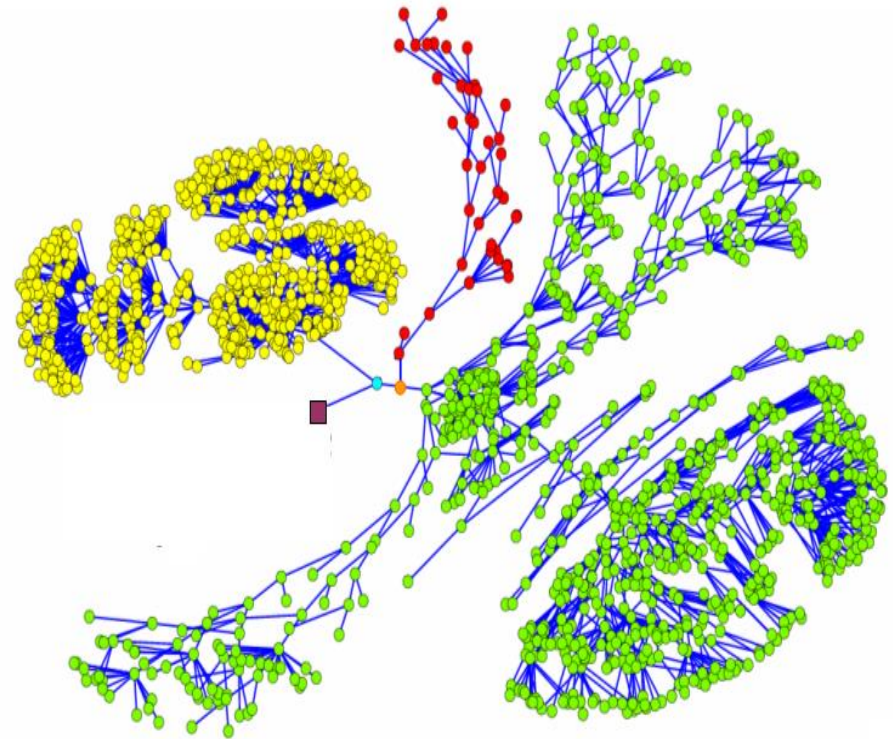
# Phylogeny and evolution

- Molecular features intensively used in phylogeny and evolution of mitochondria and chloroplasts.
- Proteins encoded by nuclear or mitochondrial DNA
- Conserved sequences suitable for ancestry investigations.
- Phylogenetic studies: important also for understanding evolutionary relationships



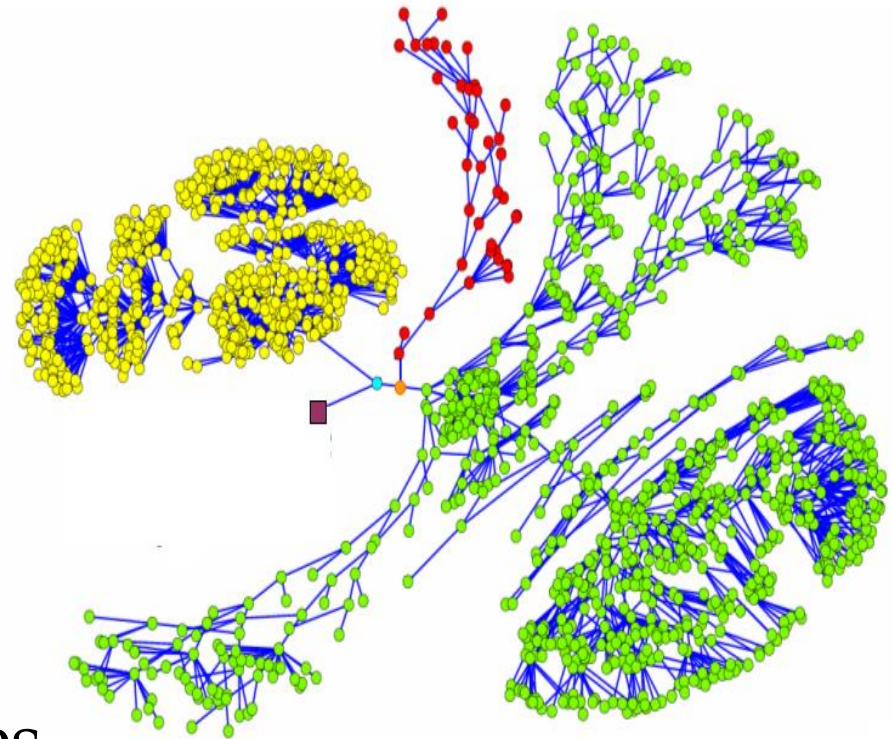
# Phylogeny and evolution

- Our proposal: classify organisms based on similarity of components involved in basic molecular syntheses
- Several basic purpose bio-molecules present in large number of organisms
- Their synthesis require presence of enzymes
- Organisms use own sets of enzyme (pathways) to obtain “same” molecule



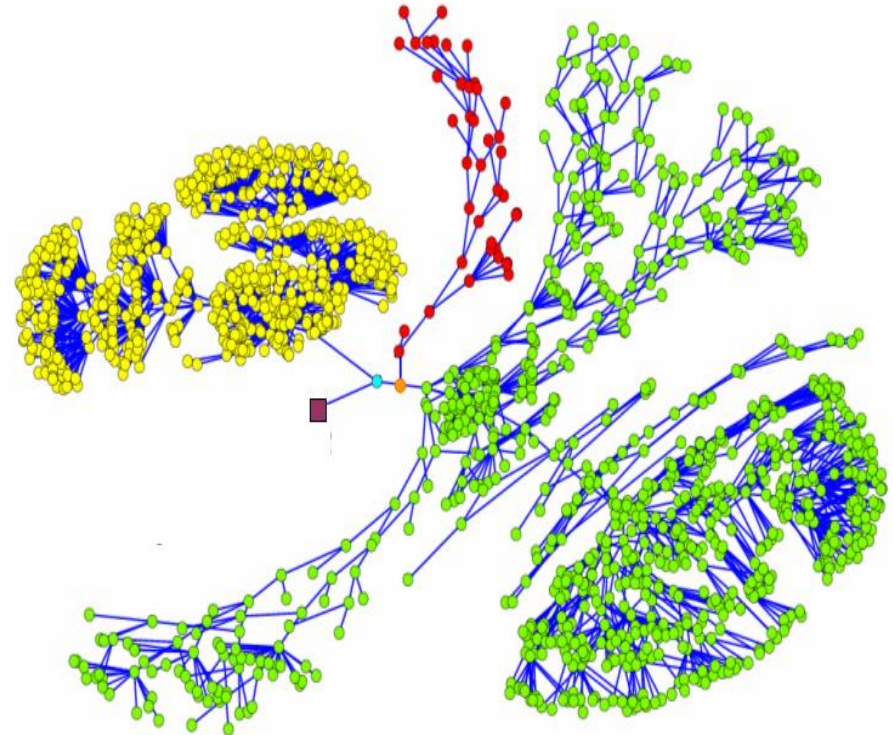
# Phylogeny and evolution

- **Basic steps:** protein structure from NCBI data base
- Weighted network  $\Leftrightarrow$  BLAST similarity score  $S_{ij}$
- Complete sequences of extant organisms only
- Modularity properties  
Newman-Girvan or ...
- Network modules  $\Leftrightarrow$  species, genus, family...
- Network  $\delta$ -distance locates best modularity expression



# Phylogeny and evolution

- Results obtained so far:
- Chitin synthesis enzymes (pathways)
- Chitin synthase in Fungi: comparison to other methods
- Evolutionary origins of mitochondria
- Evolutionary history recovery





# Results: Chitin synthesis pathway

- First system: Devise, implement and test method
- Classification source: enzymes in 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



# Results: Chitin synthesis pathway

- Database: Protein sequences from NCBI
- 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

# Results: Chitin synthesis pathway

- 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  $\sigma = S_{th}$



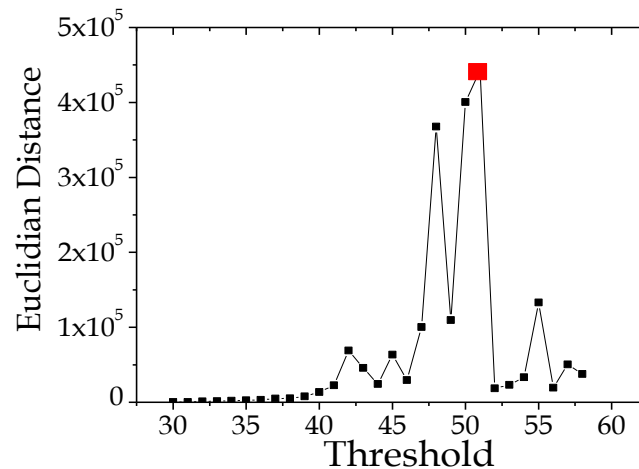


# Results: Chitin synthesis pathway

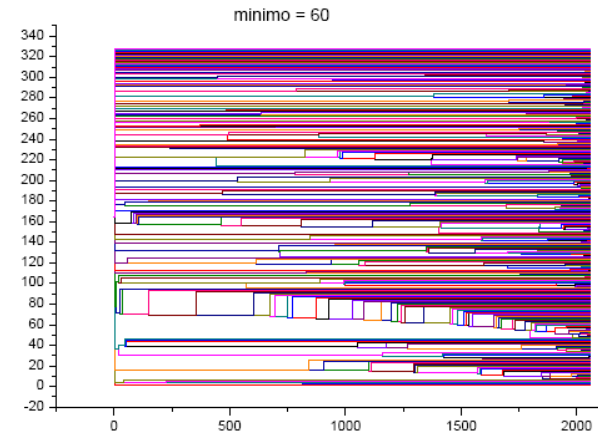
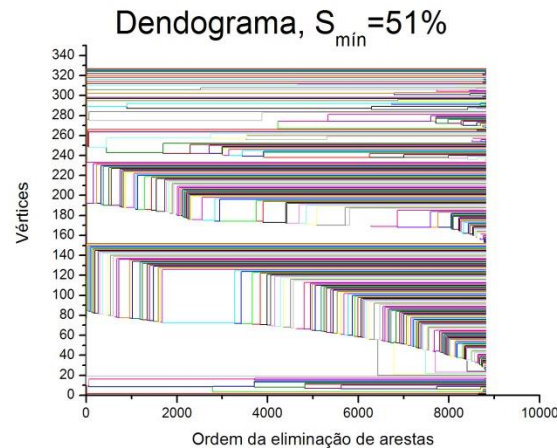
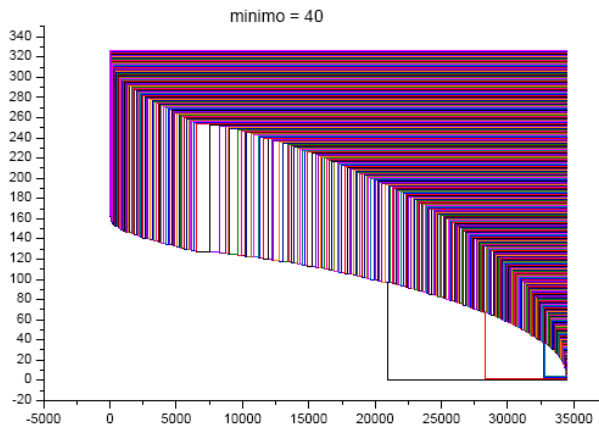
- Network measures:
  - Degree distribution  $P(k)$
  - Clustering coefficient  $C$
  - Average path-length  $\langle d \rangle$
  - Edge betweenness  $B$
  - **Network distance**  $\delta(\alpha, \beta)$
- Networks depend on  $\sigma$
- Judicious choice of value of  $\sigma$  optimizes reliability of classification scheme
- NG community finding method

# Results: Chitin synthesis pathway

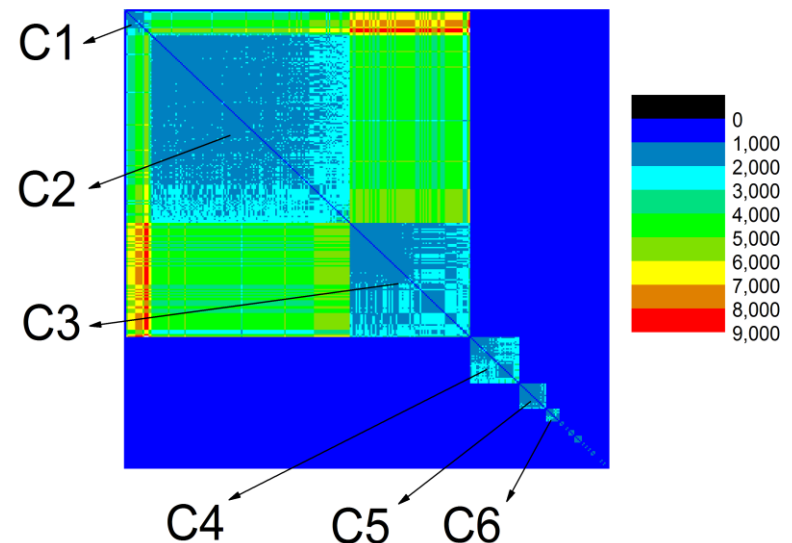
- Enzyme UDP
- $S_{th} = \sigma \approx 51\%$ : sudden transition in network properties
  - Sharp decrease in  $\langle d \rangle$
  - Clustering  $C$  remains relatively unchanged
  - Sharp change in dendrogram based on  $B$
  - Peak in the distance  $\delta(\sigma, \sigma + \Delta\sigma)$



# Results: Chitin synthesis pathway

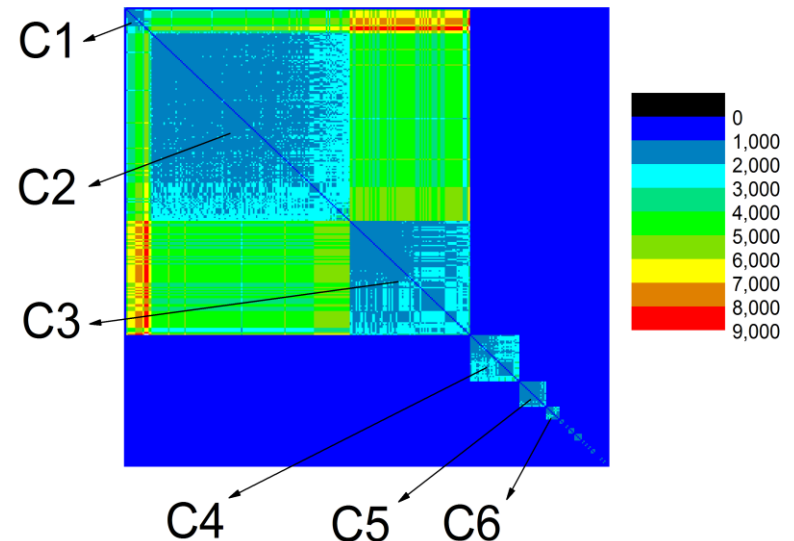
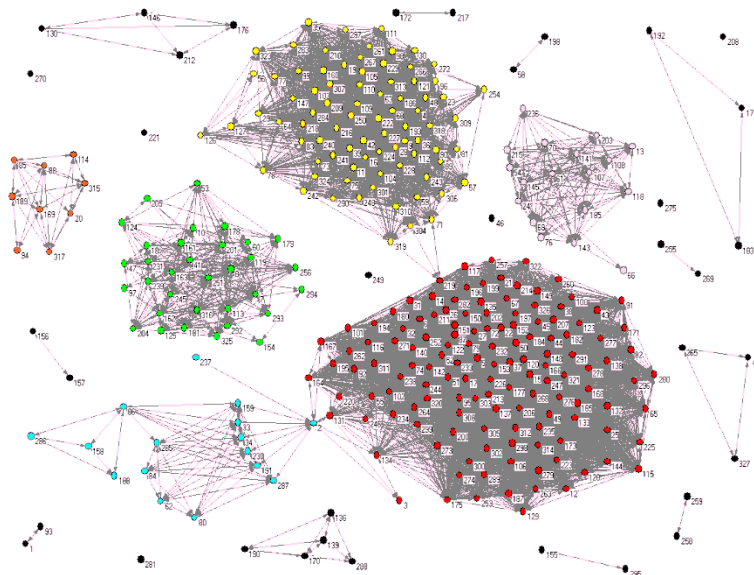


- $D_{\alpha, \alpha+1}$  is reflected in the dendrogram structure
- At  $S_{th}=51\%$ , main groups identified are reproduced in neighborhood matrix
- Moduli C1-C6 with precise biologic meaning.



# Results: Chitin synthesis pathway

- C1 - Cyanobacteria
- C2 - Firmicutes
- C3 -  $\beta$  and  $\gamma$  Proteobacteria
- C4 -  $\alpha$ -Proteobacteria
- C5 - Actinobacteria
- C6 -  $\epsilon$ -Proteobacteria

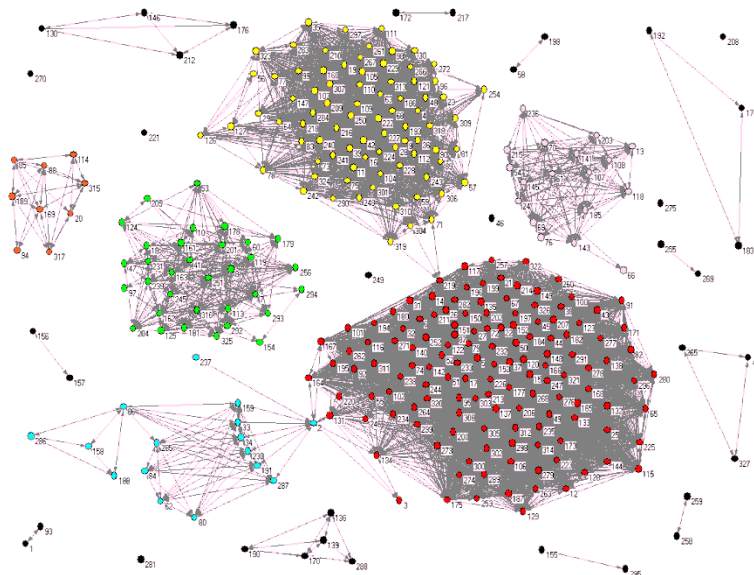


# Results: Chitin synthesis pathway

- C1 – Cyanobacteria
- C2 – Firmicutes
- C3 –  $\beta$  and  $\gamma$  Proteobacteria
- C4 –  $\alpha$ -Proteobacteria
- C5 – Actinobacteria
- C6 –  $\epsilon$ -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



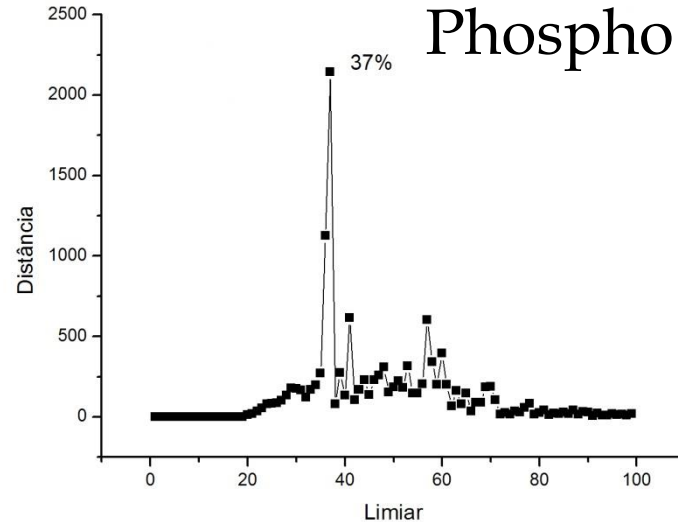
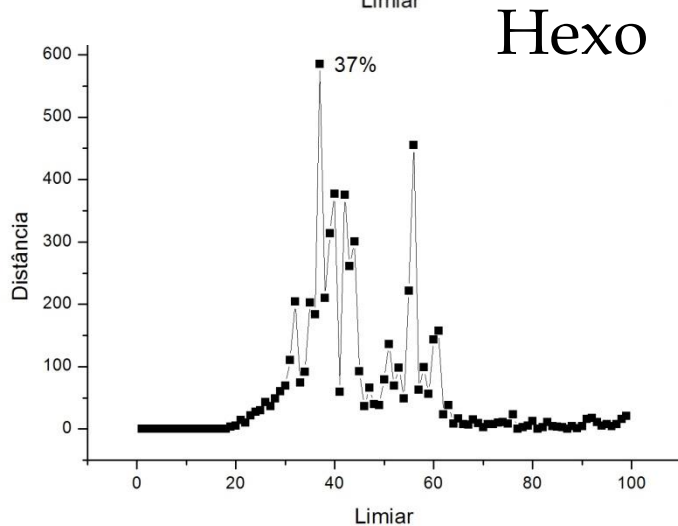
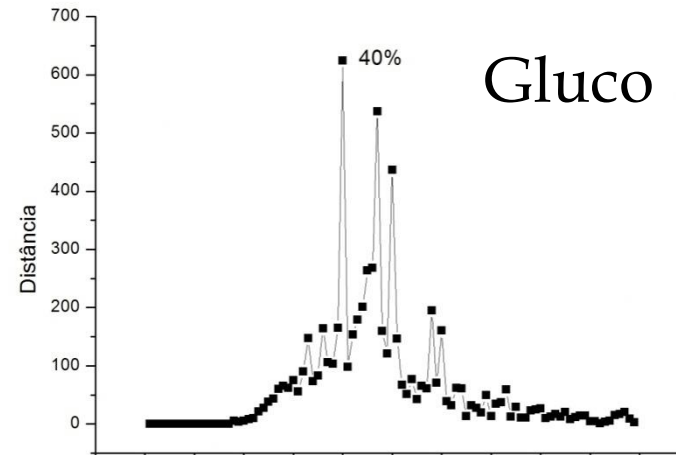
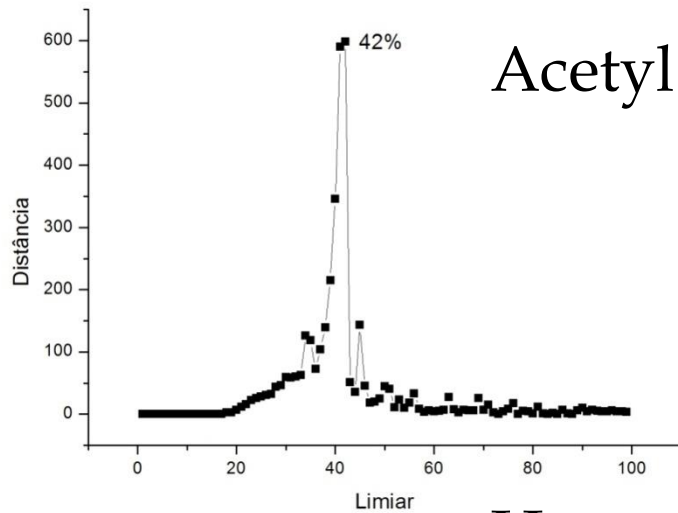
# Results: Chitin synthesis pathway

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

Enzyme	<SIM>	$\sigma$	$S_t$	# 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: Chitin synthesis pathway

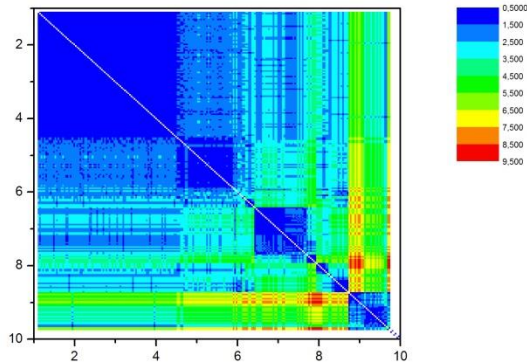
- Network distance  $D_{\alpha\beta}$  x threshold  $S_{th}$



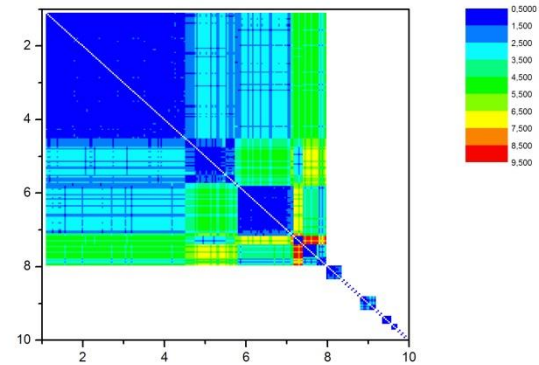
# Results: Chitin synthesis pathway

- Hexo: Dependence of network on  $S_{th}$

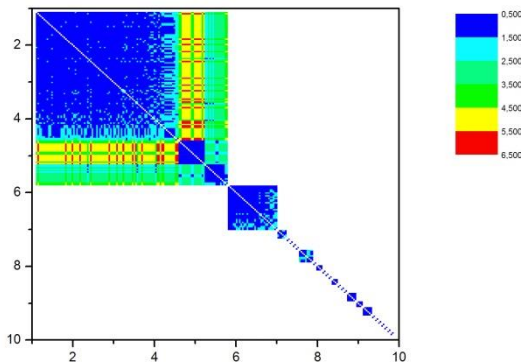
37%



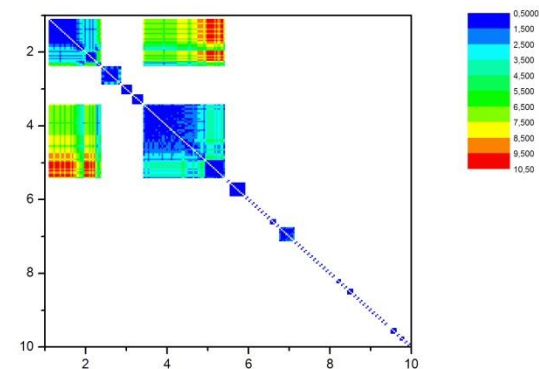
40%



44%



56%





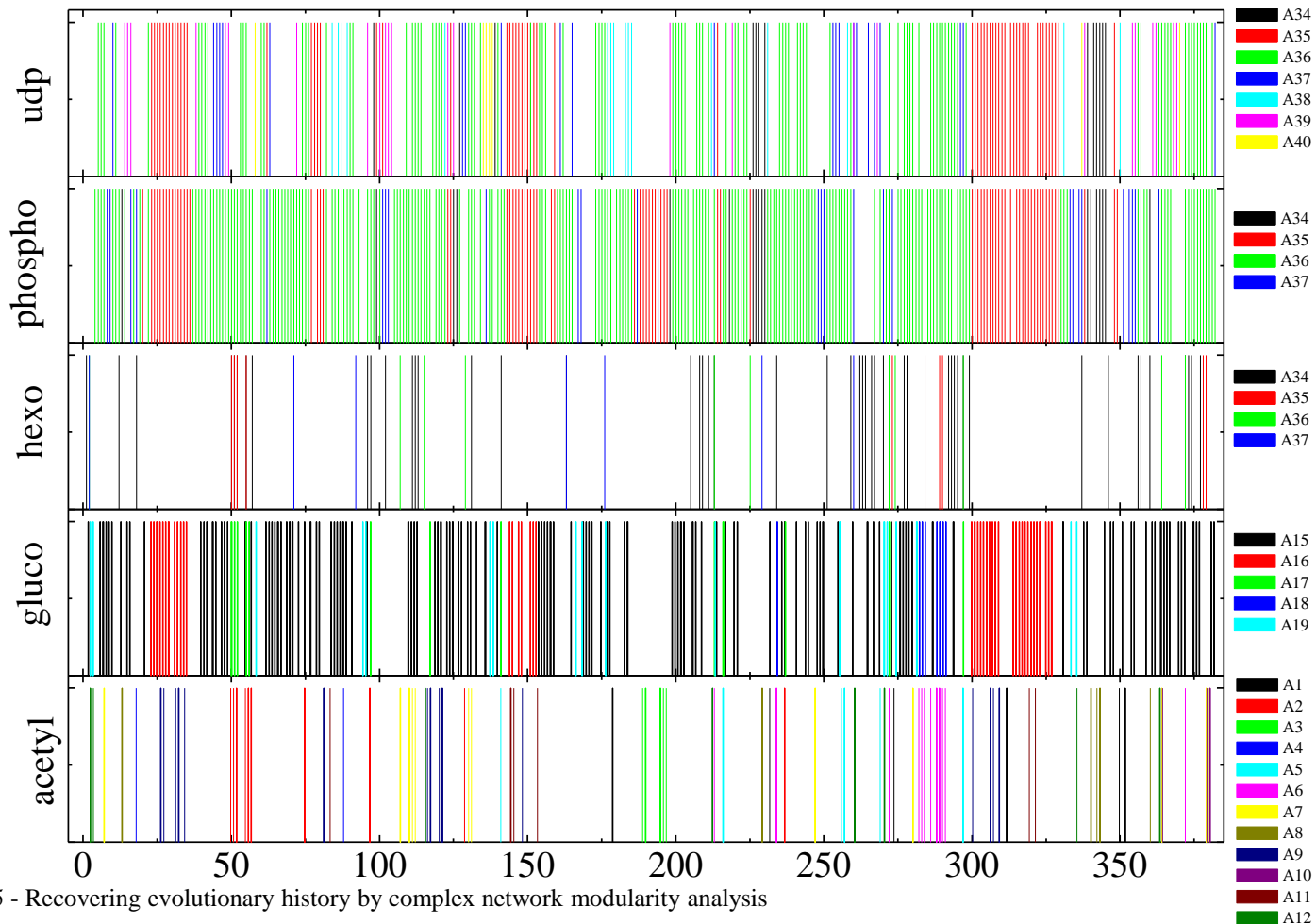


## Results: Chitin synthesis pathway

- 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

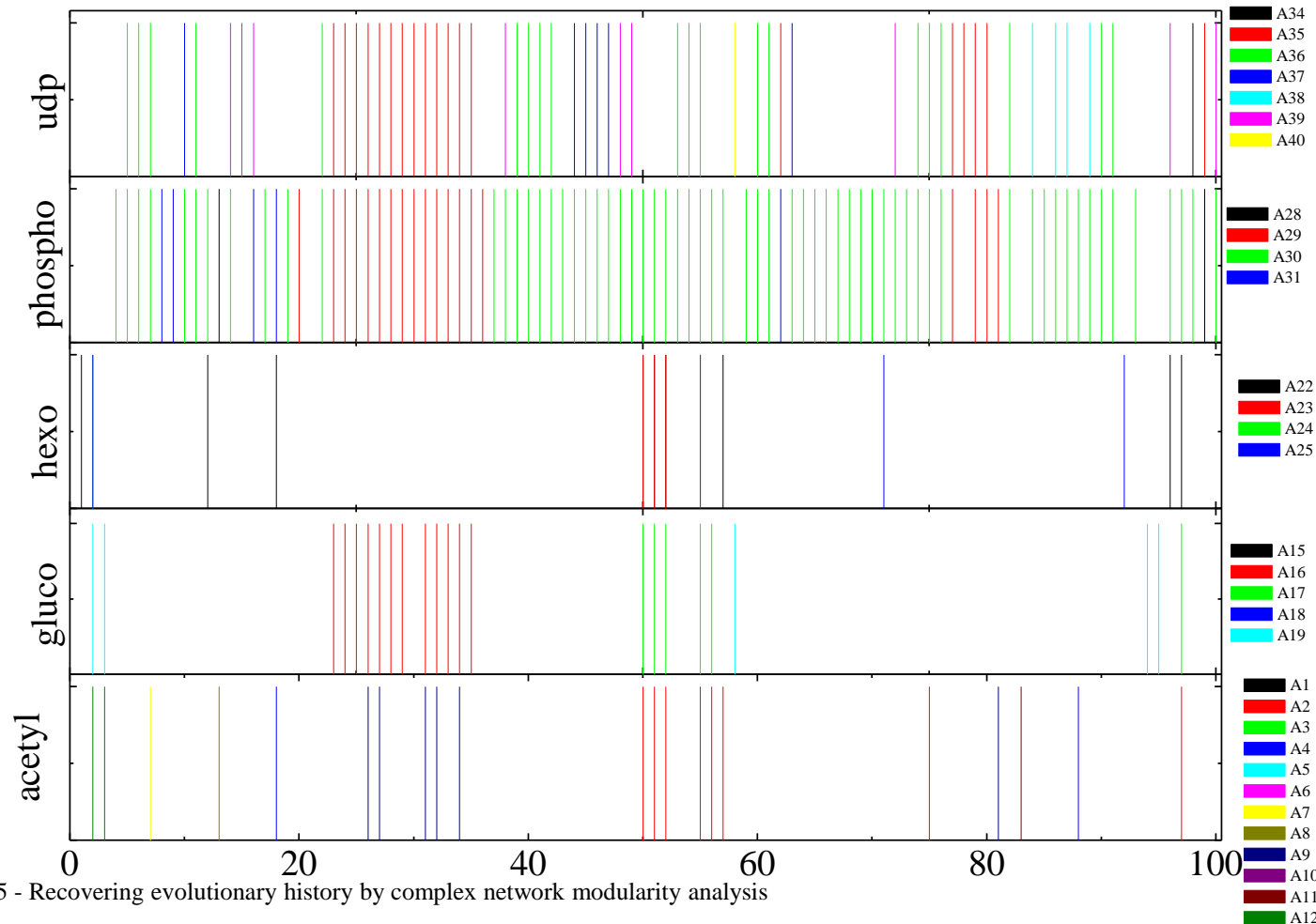
# Results: Chitin synthesis pathway

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



# Results: Chitin synthesis pathway

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



# Results: Chitin synthesis pathway

- Congruence scores for pair-wise phylogeny comparison provided by two different networks.
- Average table score: 0.84.

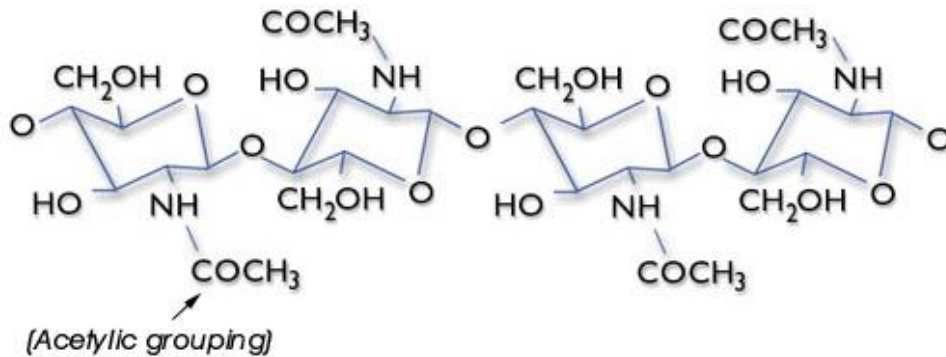
	A	G	H	P	U
A		0.79	0.73	0.93	0.91
G	0.79		0.69	0.83	0.87
H	0.73	0.69		0.90	0.79
P	0.93	0.83	0.90		0.95
U	0.91	0.87	0.79	0.95	



## Results: Fungi chitin synthase

- Second system – Compare with other phylogenetic methods
- Fungal cell wall controls interaction of fungi with surroundings, protects cell environmental stresses.
- Directly involved in important biological processes: morphogenesis, antigenic expression, adhesion, ...
- Chitin is crucial to the architecture and integrity of the fungal cell wall.
- Disruption of chitin synthesis leads to malformed, osmotically unstable cells, resulting in cell death

# Results: Chitin synthesis pathway



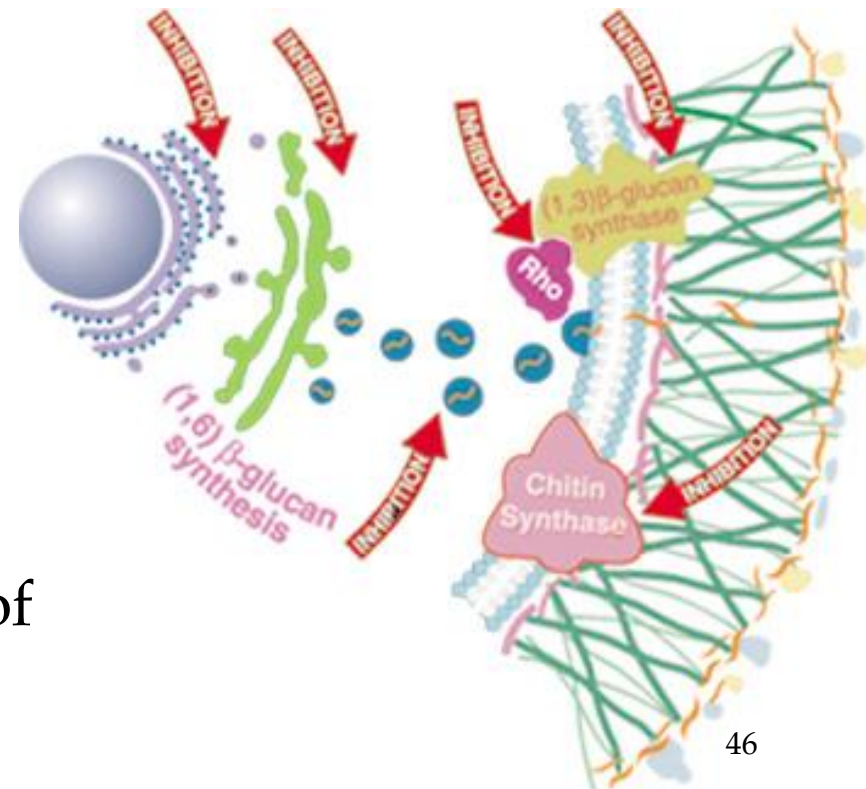
Chitin:  $\beta$ -1,4-linked linear homopolymer of N-acetylglucosamine

Chitin Synthases (CHS)  
EC 2.4.1.16  
Glucosyltransferase

Converts UDP-N-acetyl-D-glucosamine into chitin and UDP (irreversible)

Multiple isoforms

Different expression levels (stage of the life cycle and cellular location)



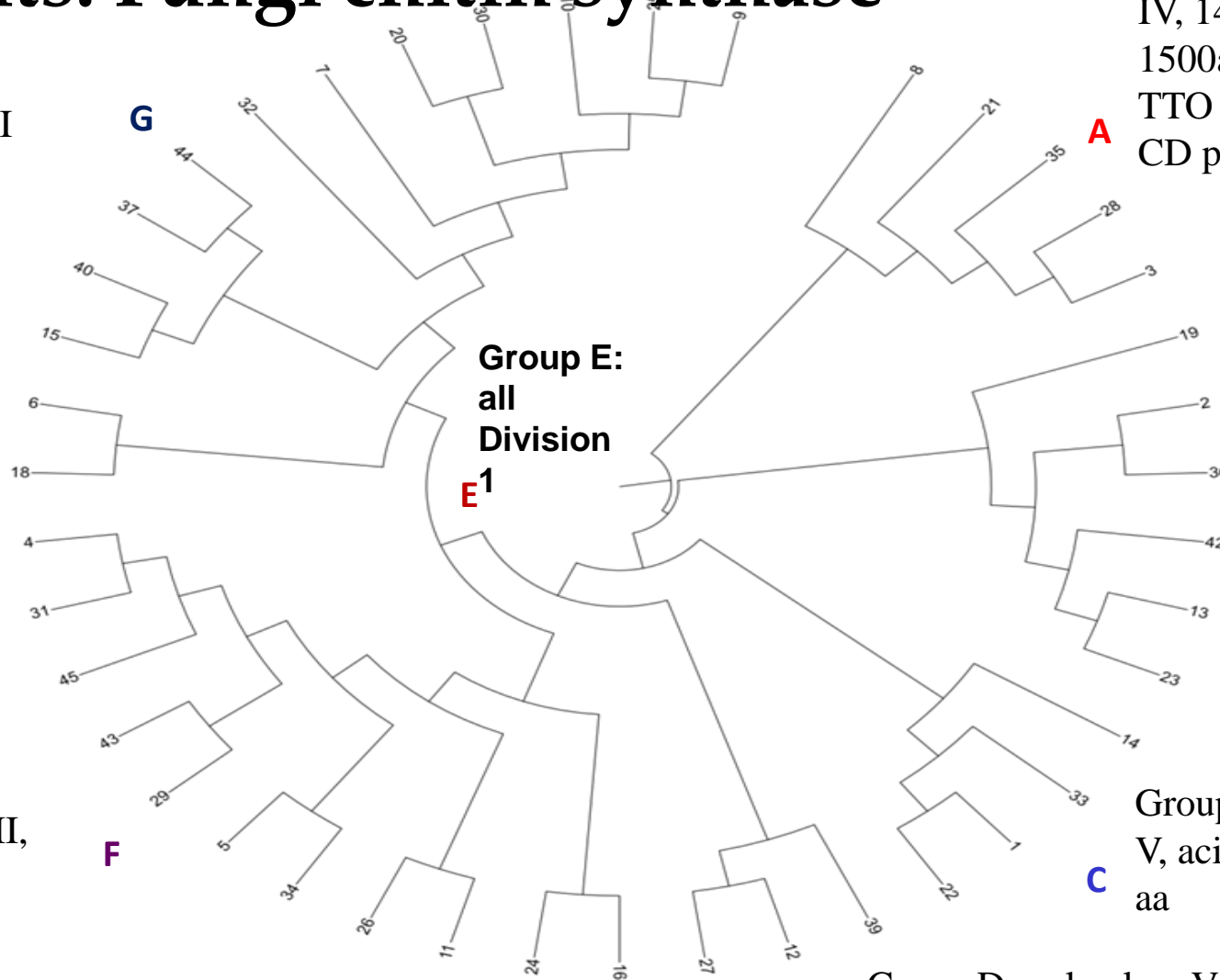


# Results: Fungi chitin synthase

- Previous studies addressed to Ascomycota CHS.
- CHSBasidio database: data mining from NCBI on 10/01/2011, with 347 sequences, 72 distinct species.
- Relational database of all CHSBasidio protein and nucleotide sequences constructed and validated
- 62 complete protein sequences (18% entries)
- Excluding redundancies  $\Rightarrow$  42 unique sequences.
- Sequences with 864-1271 aminoacids.
- Chemically basic, Division I (class II and III) enzymes, with six to seven transmembrane helices
- Conserved domains PF01644 and PF08407.

# Results: Fungi chitin synthase

Group G: class II



Group A: class IV, 1400-1500aa, basic TTO profile 1, CD profile 1

Group B: class IV, 1000-1250 aa, basic, TTO profile 1, CD profile 1

Group F: class III, 755-977 aa

Group C: only class V, acid, 1150-2070 aa

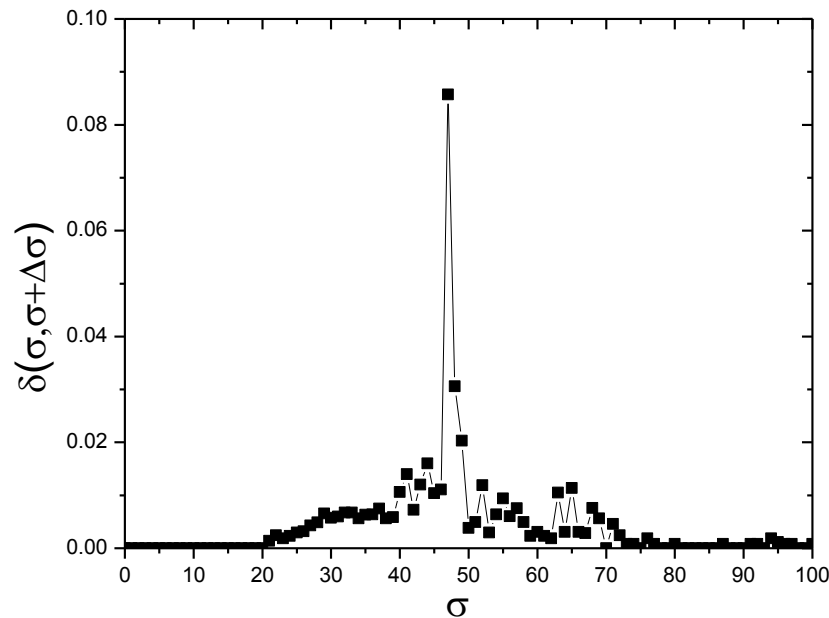
Group D: only class V, acid, TTO profile 3, CD profile 3

Dendrogram of sequence similarity analysis

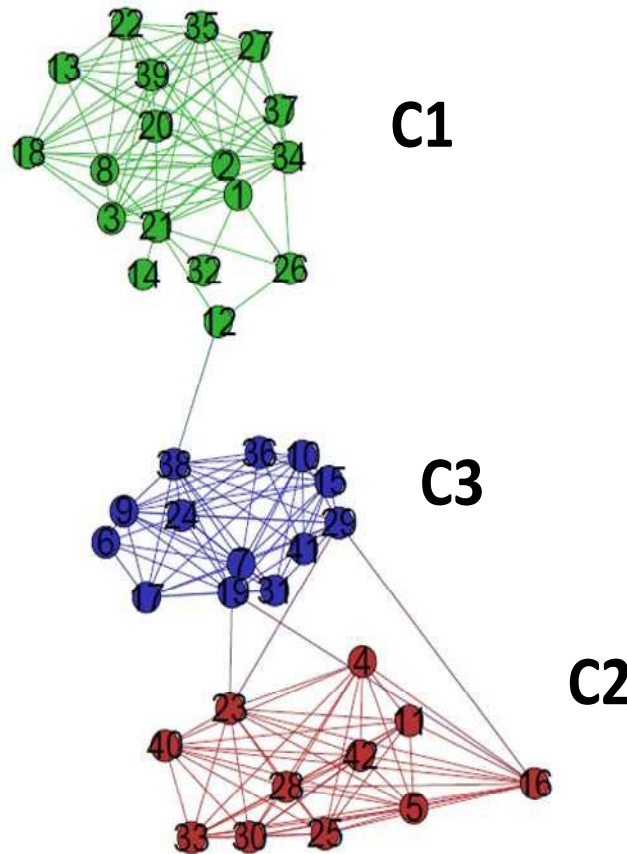


# Results: Fungi chitin synthase

- $\delta$ -distance network approach



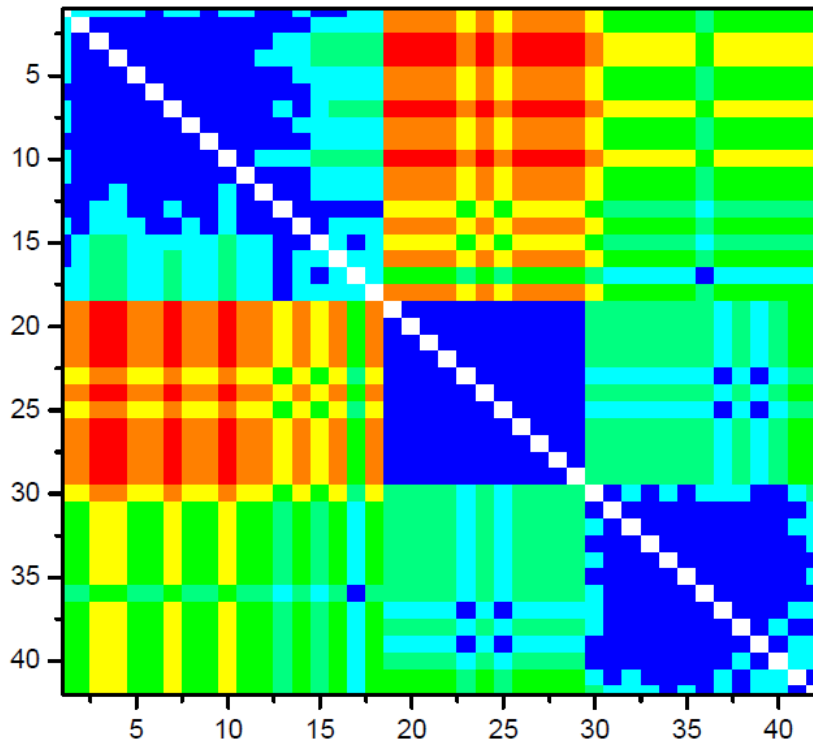
# Results: Fungi chitin synthase



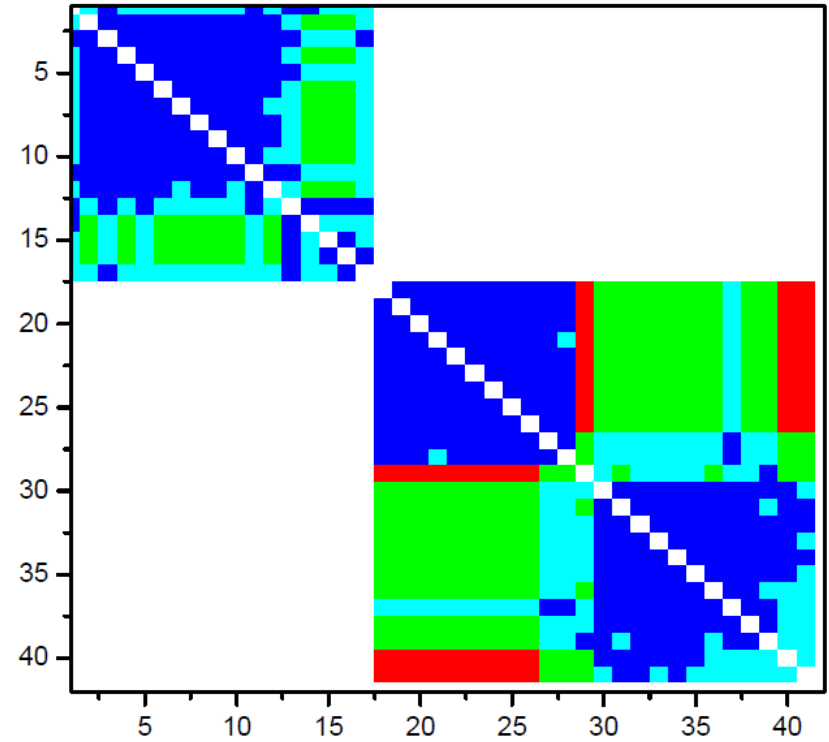
# Results: Fungi chitin synthase

## ■ Neighborhood matrix

$$\sigma = 0.46$$

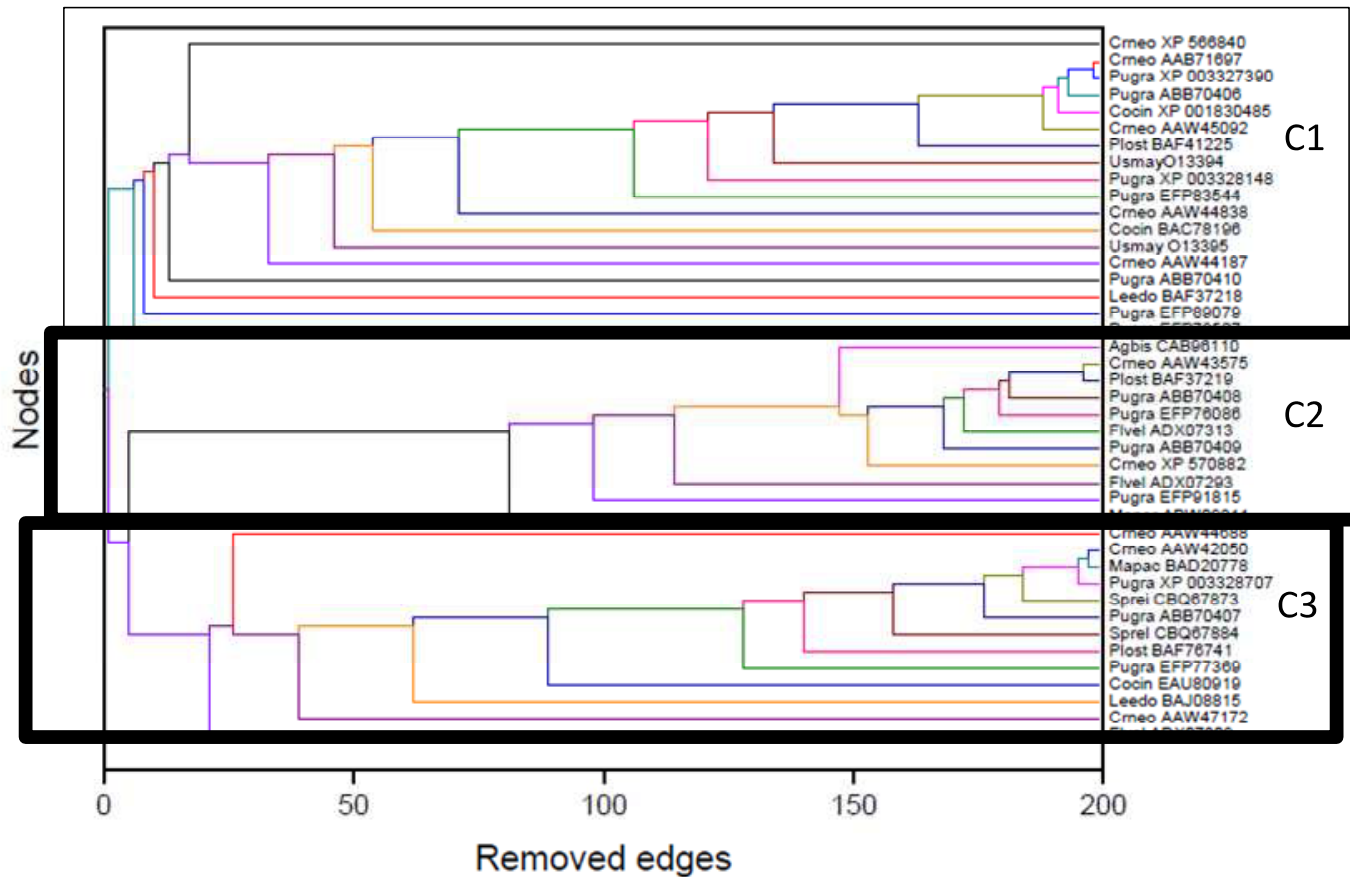


$$\sigma = 0.47$$



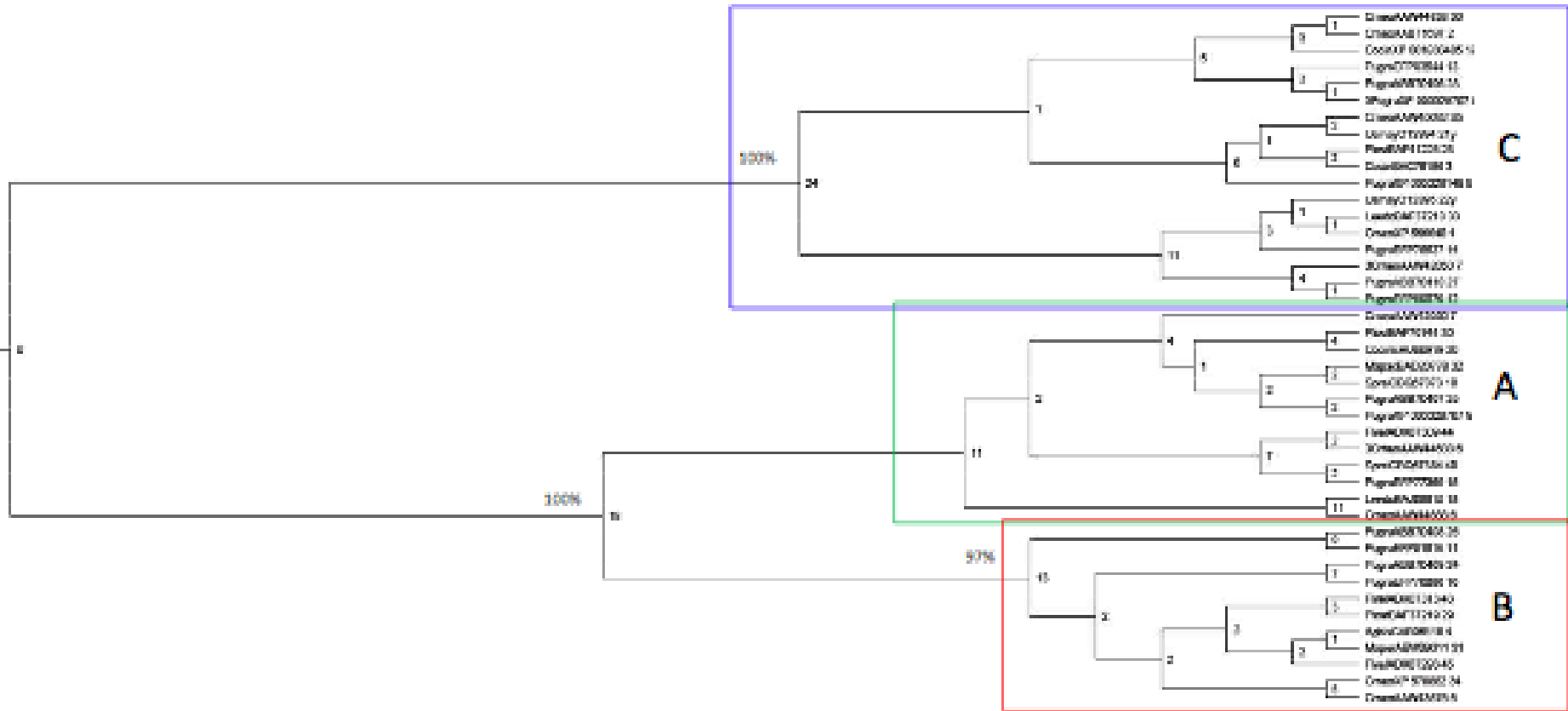
# Results: Fungi chitin synthase

- NG community finding method



# Results: Fungi chitin synthase

- Phylogenetic tree Bayesian tree - Posterior Probabilities values above 50% are exhibited above the branches for the main groups.



# Results: Fungi chitin synthase

- Congruence obtained after pairwise comparison of the phylogenetic analysis based on chitin synthase sequences provided by five different methods

Groups	MP	D	ML	B	CN
MP	1	1	1	1	1
D	1	1	1	1	1
ML	1	1	1	1	1
B	1	1	1	1	1
CN	1	1	1	1	1



# Results: Origins of mitochondria

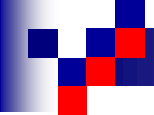
- Endosymbiotic theory: all organelles in eukaryotic cells have their origin by the inclusion of pre-existent organisms, together with their complete genomes.
- Mitochondria and chloroplasts, with own genome, have undergone gene transfer to the eukaryotic host cell nucleus
- Horizontal or lateral gene transfer phenomenon
- Horizontal transfer favored by endosymbiotic origin of these organelles.
- Conserved sequences suitable for ancestry studies.



# Results: Origins of mitochondria

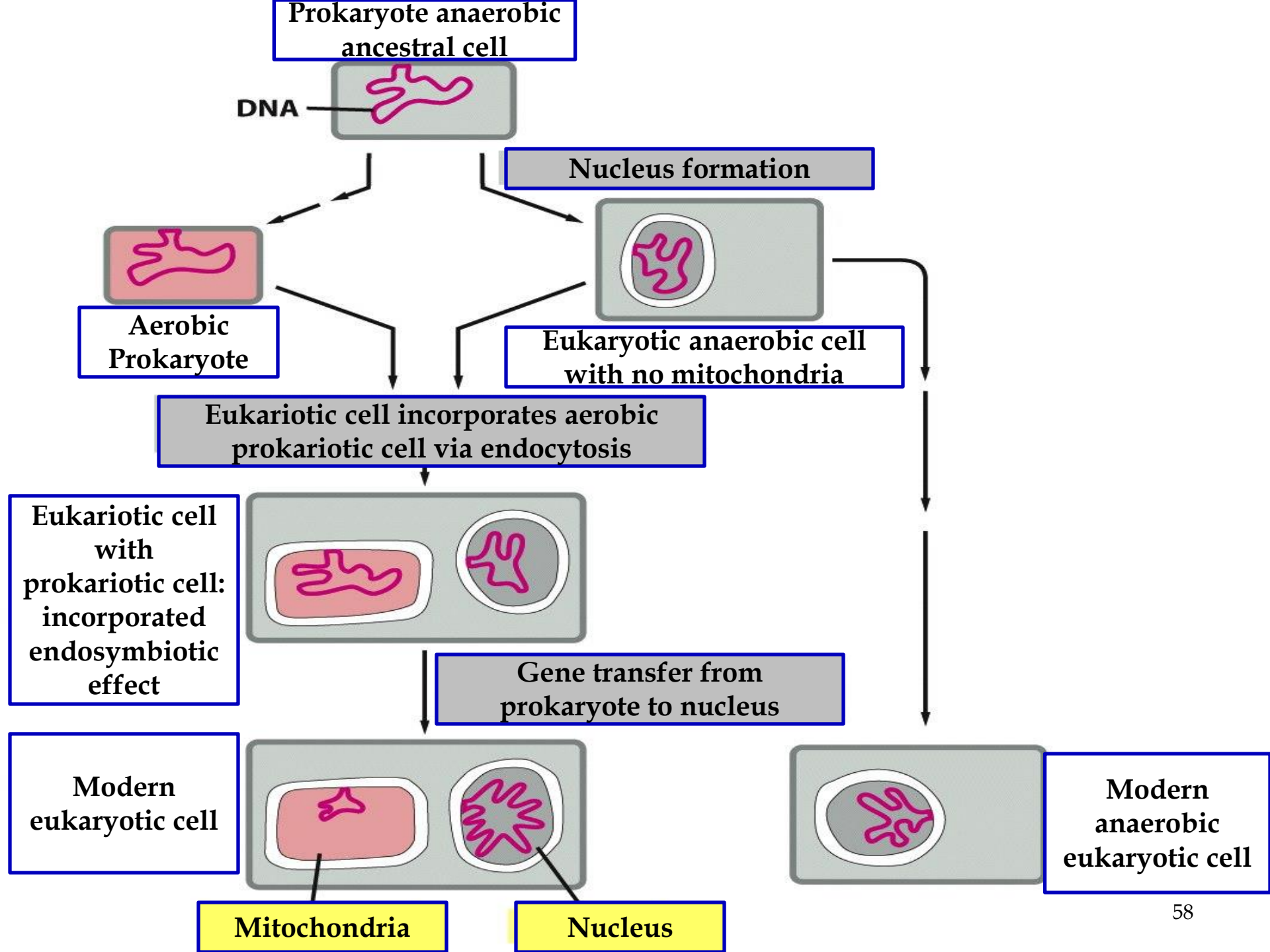
- Phylogenetic relationship between mitochondria and bacteria.
- Investigation related to evolutionary aspects.
- From prokaryotic anaerobic cell to eukaryotic aerobic cell and multi-cellular organisms.
- Find evidences of the common ancestor of bacteria and mitochondria.
- Which class, order, family of extant organisms are closer to common ancestor.

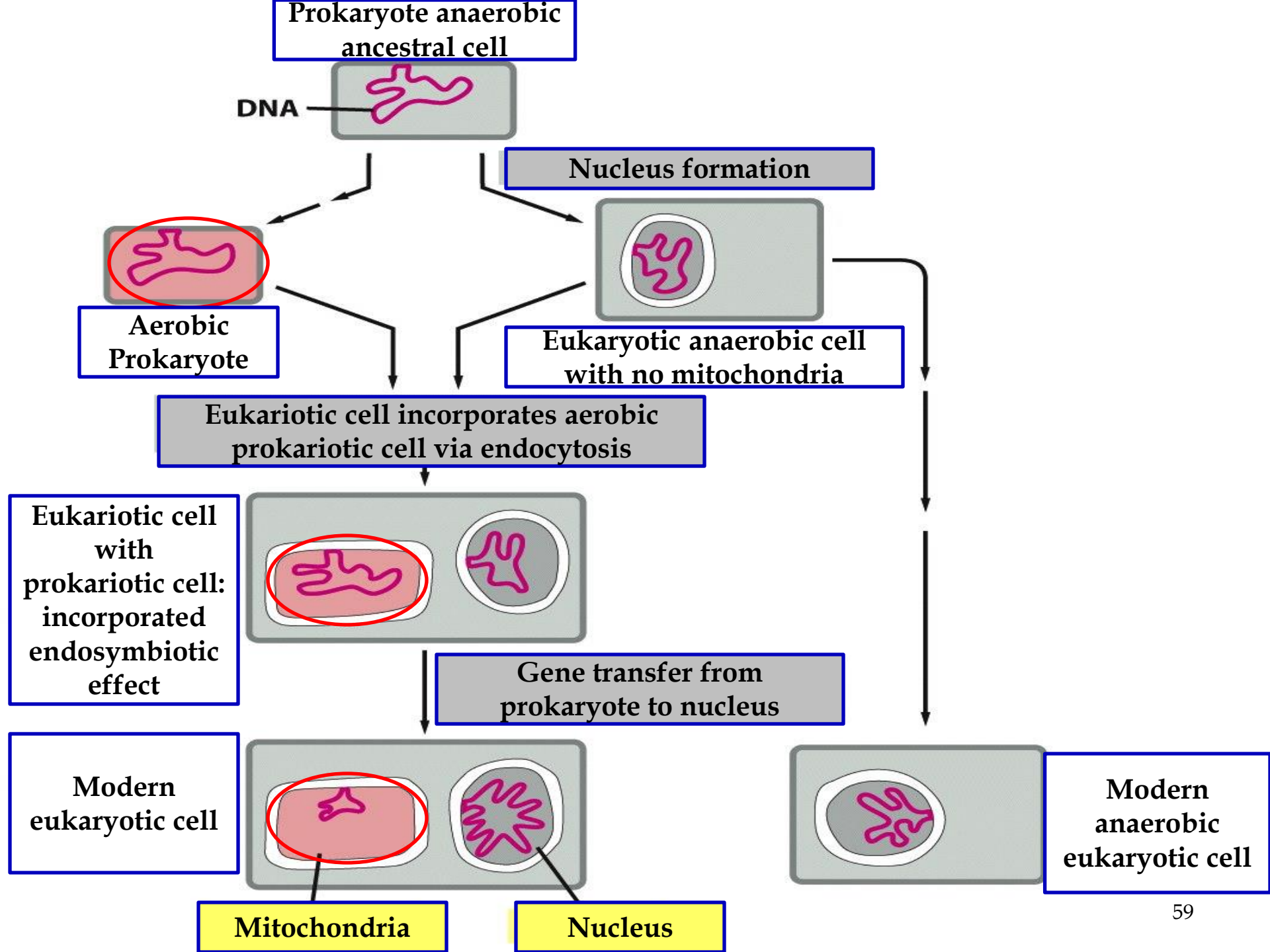


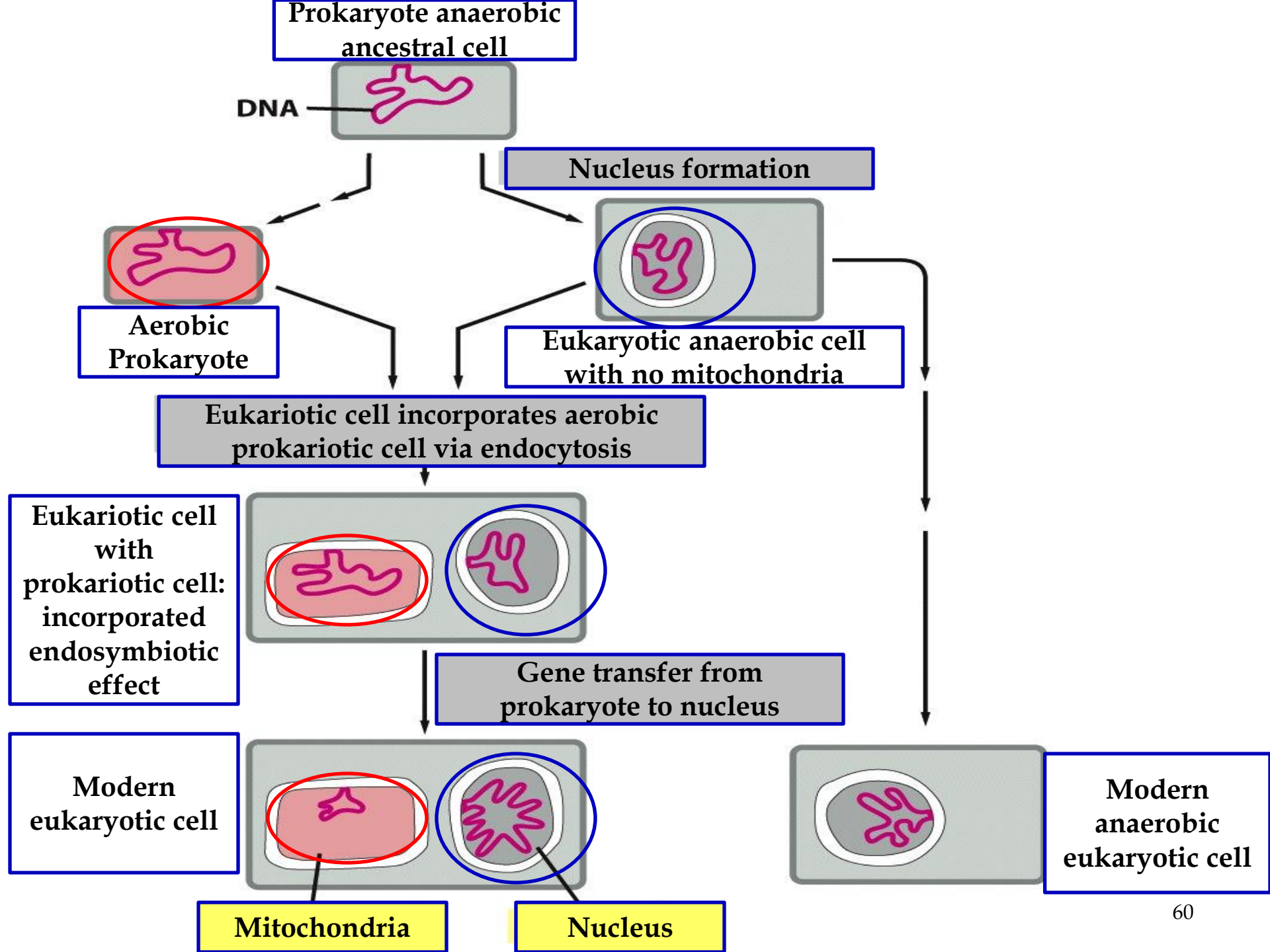


# Results: Origins of mitochondria

- Basic steps of endosymbiotic dynamics

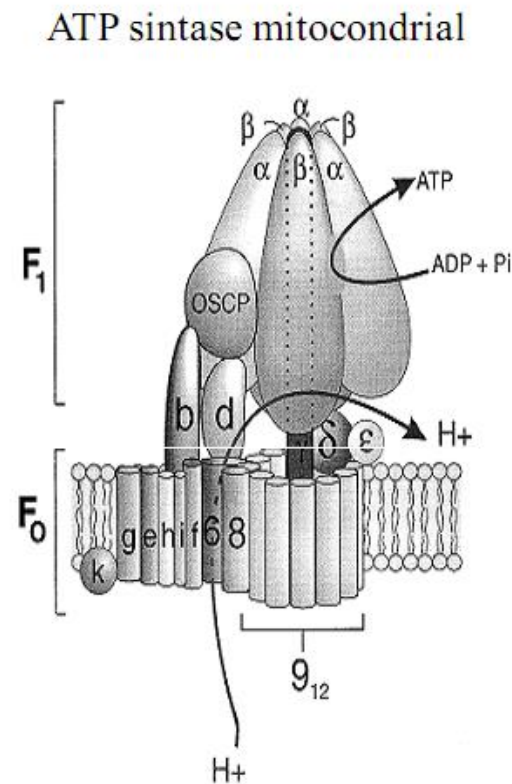




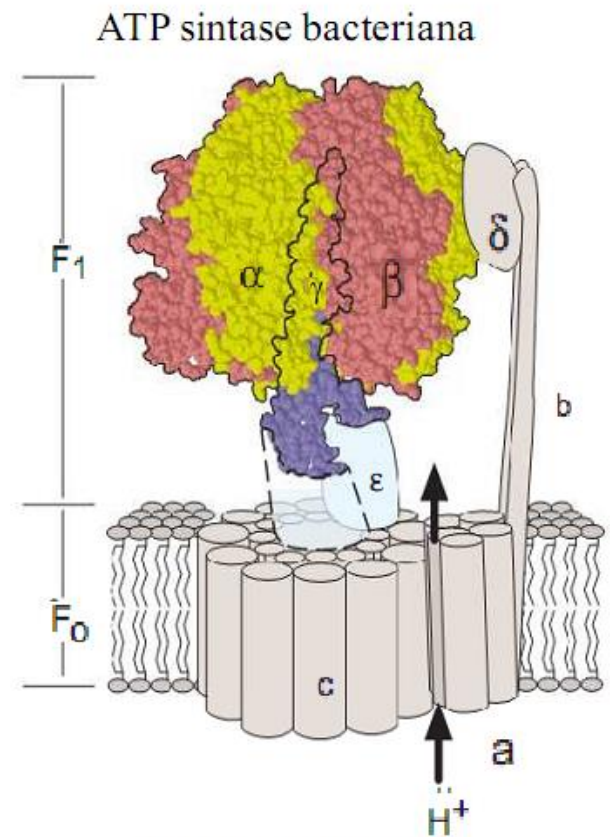


# Results: Origins of mitochondria

- Study based on ATP synthase complex



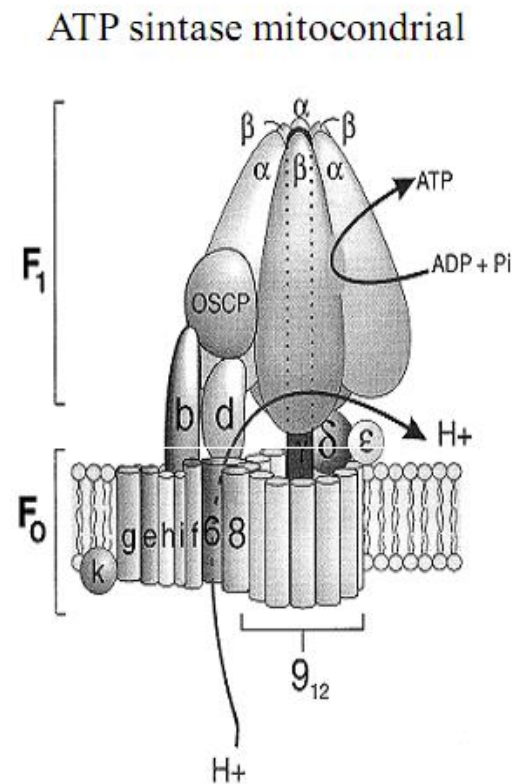
DEVENISH *et al*, 2000



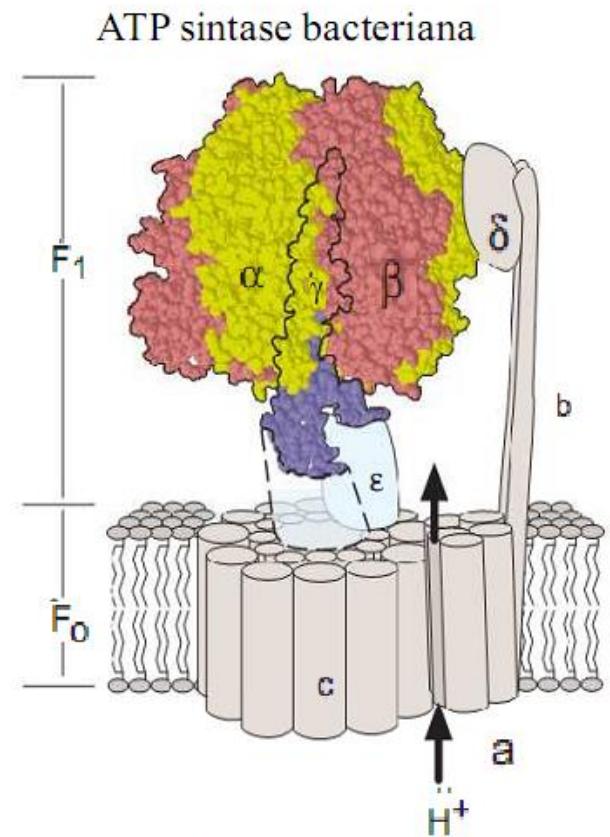
WANG & OSTER, 1998

# Results: Origins of mitochondria

- Study based on ATP synthase complex: responsible for ATP synthesis in mitochondria, chloroplasts and bacteria.



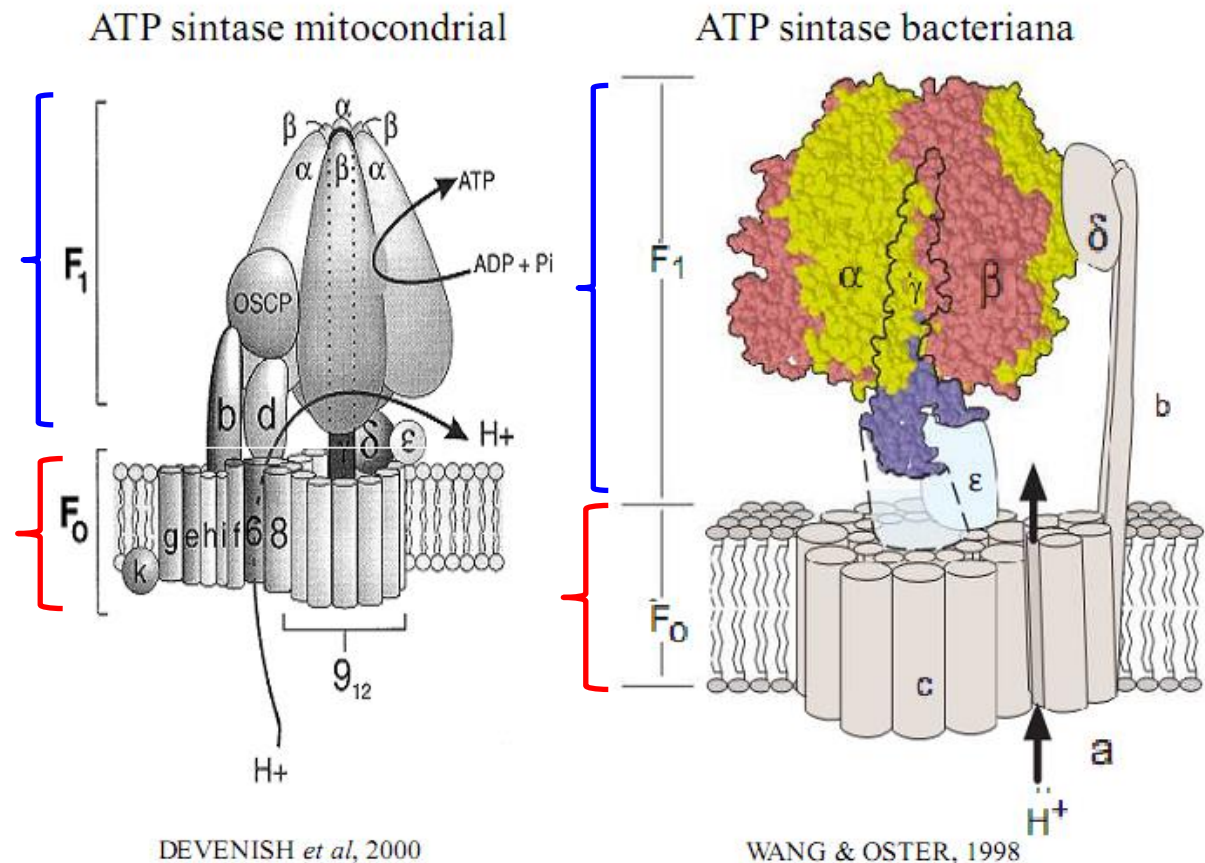
DEVENISH *et al*, 2000



WANG & OSTER, 1998

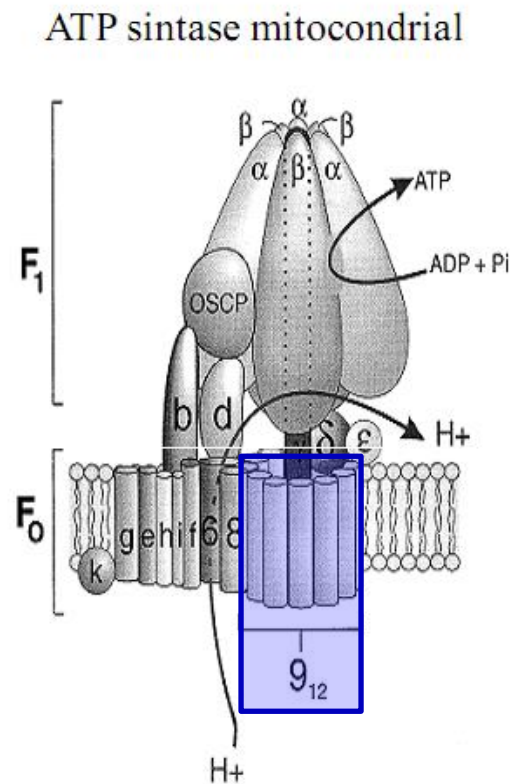
# Results: Origins of mitochondria

- ATP synthase complex:  
formed by two modules  $F_0$  and  $F_1$

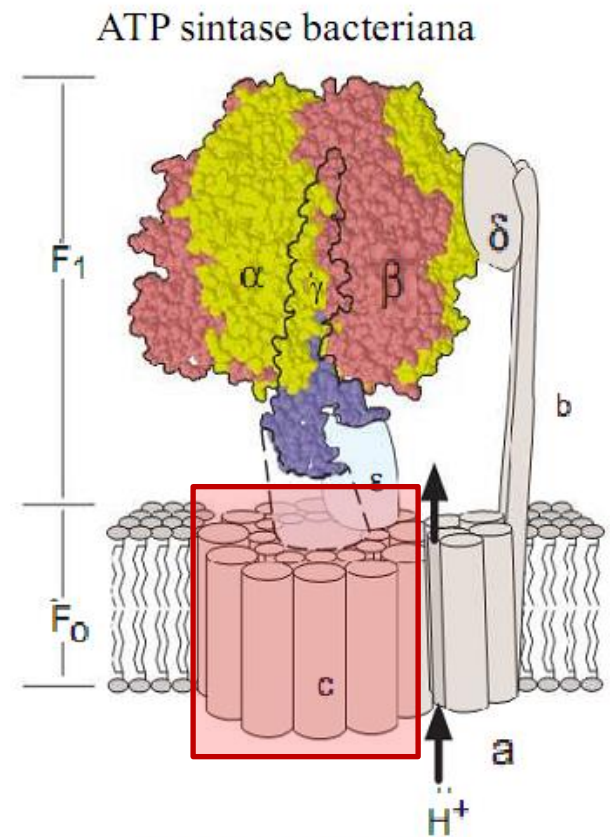


# Results: Origins of mitochondria

- ATP synthase complex: subunits 9 and *c* act in the proton channel mechanism in mitochondria and prokaryotes



DEVENISH *et al*, 2000

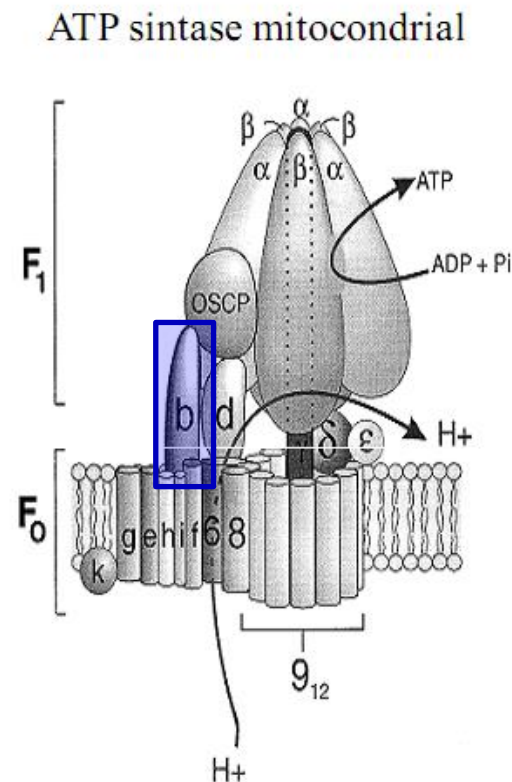


WANG & OSTER, 1998

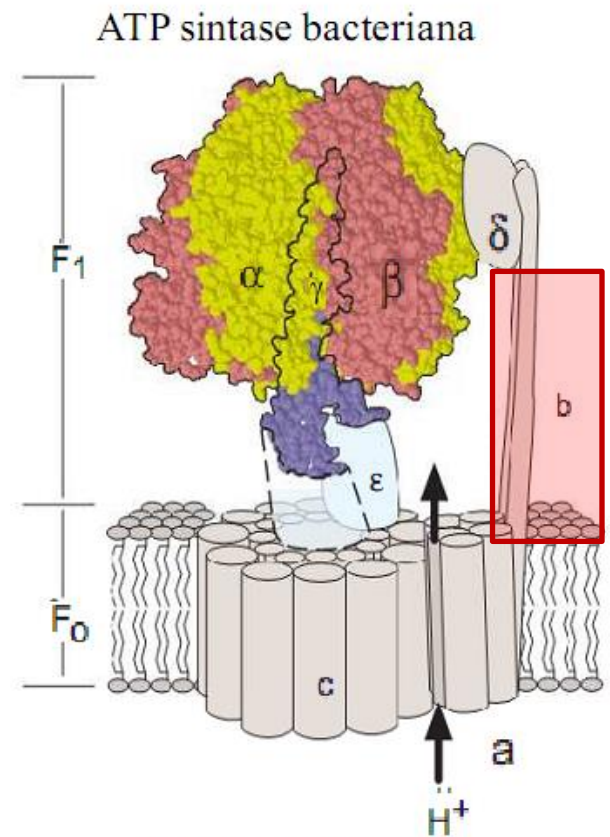


# Results: Origins of mitochondria

- ATP synthase complex: subunits 4 and *b* act as connecting rod between  $F_0$  and  $F_1$



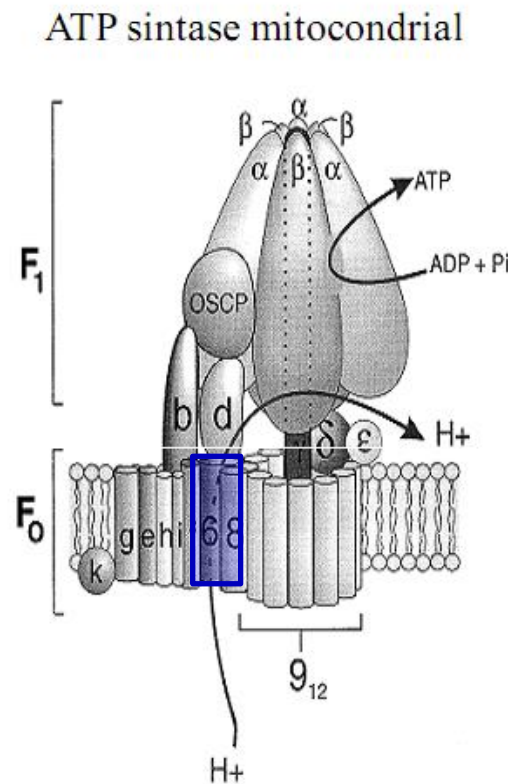
DEVENISH *et al*, 2000



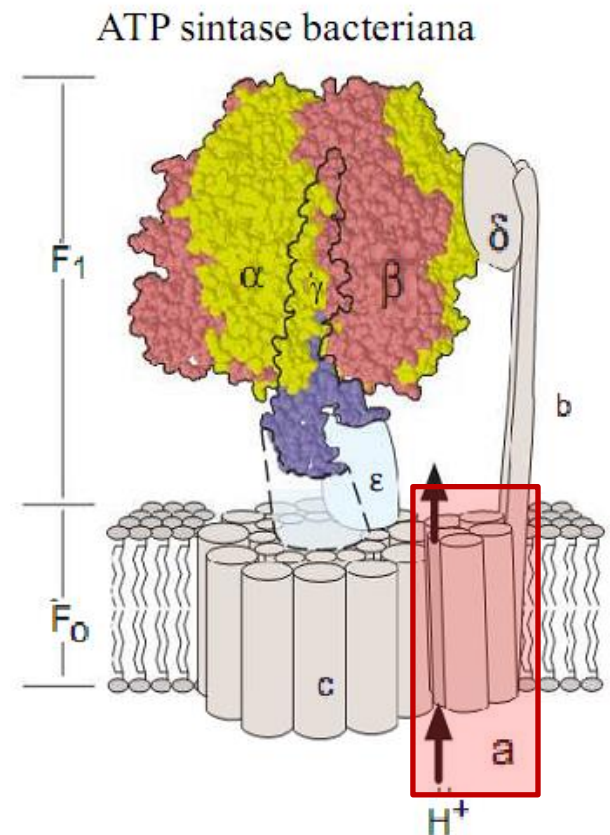
WANG & OSTER, 1998

# Results: Origins of mitochondria

- ATP synthase complex: subunits  $\delta$  and  $a$  constitute the proton channel in mitochondria and prokaryotes



DEVENISH *et al*, 2000



WANG & OSTER, 1998



# Results: Origins of mitochondria

- Mitochondria originated endosymbiotically from an Alphaproteobacteria-like ancestor.
- Relationship between mitochondria within eukarya and bacteria.
- Protein sequences of ATP synthase: subunits 4, 6, and 9 (eukarya), and *b*, *a*, and *c* (bacteria).
- BLAST similarity scores between sequences;
- $\delta$ -distance locates best phylogenetic relationship;
- NG community finding method
- Distribution of mitochondrial and bacteria sequences in different network communities



# Results: Origins of mitochondria

- Complete data base: downloaded from NCBI on 08/04/2011.
- Filtered bank
  - sub-unities 4 and *b*: 597 sequences ⇒ nuclear encoded
  - sub-unities 6 and *a*: 2945 sequences ⇒ mitochondrially encoded
  - sub-unities 9 and *c*: 890 sequences ⇒ mitochondrially encoded
- Similarity evaluation by BLAST 2.2.21 StandAlone



# Results: Origins of mitochondria

- Extant group of Alphaproteobacteria phylogenetically closer to the mitochondrial ancestor is uncertain
- Proposed groups:
  - the order **Rickettsiales**
  - the family *Rhodospirillaceae*
  - the genus *Rickettsia*.
- For all subunits, results do not support hypothesis that **Rickettsiales** are closely related to the mitochondrial ancestor.



## Results: Origins of mitochondria - details

- Mitochondrial ATP synthase far less related to the bacterial homologs in **Rickettsiales** × Rhizobiales, Rhodobacterales, Rhodospirillales, Sphingomonadales, and clusters SAR11 and SAR116.
- Agree better with proposal that Rhodospirillaceae includes closest extant relatives to mitochondria.
- Agree with results that Alphaproteobacteria other than in the **Rickettsiales** show closer relationships to mitochondria
- Confirm that Rhodobacterales, Rhizobiales and Rhodospirillales are sister groups of mitochondria.



## Results: Origins of mitochondria - details

- Other results show that trees for alphaproteobacterial orders may differ if different methods or models are used to infer phylogenetic placement.
- **Rickettsiales** appears as an order diverging earlier than the other orders in most of the phylogenies for the Alphaproteobacteria.
- Combining results provided herein with other phylogenetic analyses support hypothesis that mitochondria share a common ancestor with alphaproteobacterial orders other than **Rickettsiales**.



## Results: Evolutionary history recovery

- Work to measure the reliability of used procedure to treat proteomic data of actual extant organisms
- Simple evolutionary model where parameter set controls mutation probability and population of different species.
- Phylogeny produced by framework accurately reproduces classification resulting from actual evolutionary history for parameter values where species originated by the evolutionary dynamics have clear community structure.
- Corroborate previous reliability tests of framework.





# Results: Evolutionary history recovery

- Microscopic evolutionary dynamics of organism set characterized by a genetic strand of binary bases.
- Organisms differentiate from a single common ancestor through a cumulative process of random changes in the strand.
- Neutral model no explicit selection acts on species.
- Phylogenetic classification based on same framework used for actual data.



# Results: Evolutionary history recovery

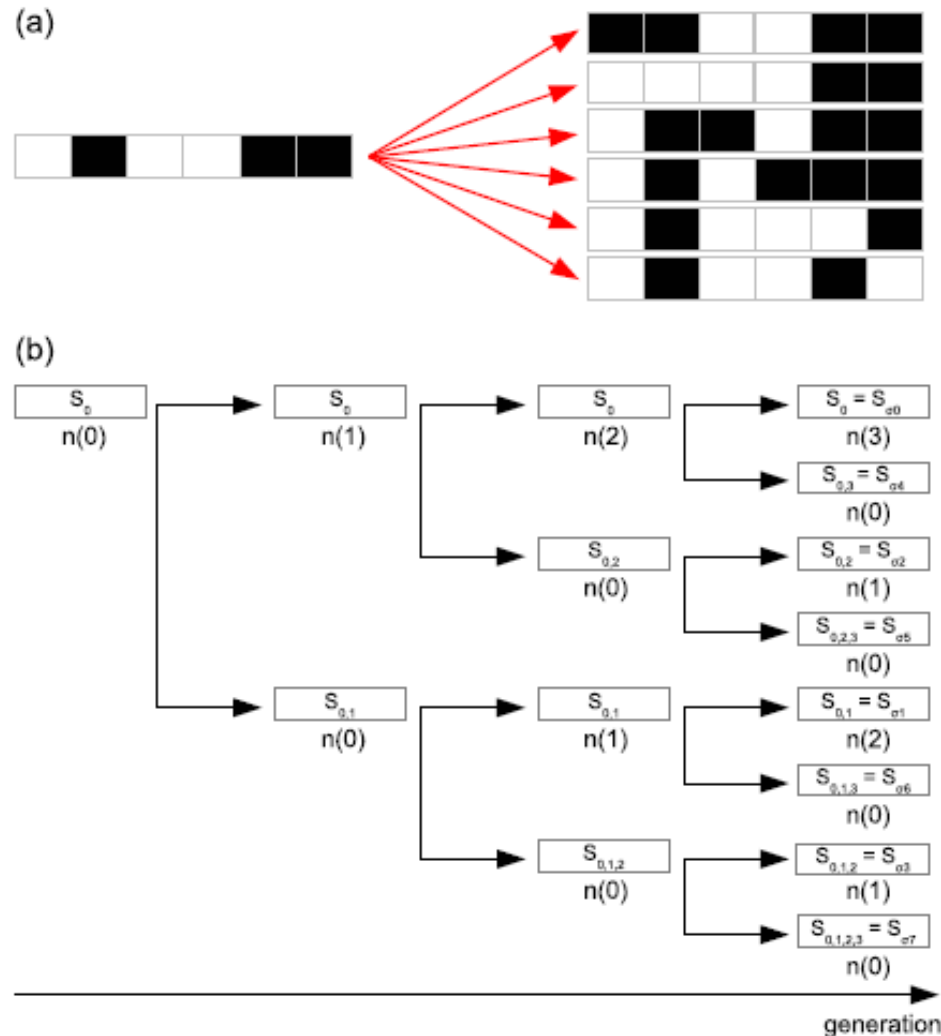
- Model initialized with one species ( $S_0$ ) with a single individual.
- At each time step  $T$ , update number of individuals of an existing species  $i$  according to a function  $n_i(t-T_i)$ , with  $T_i$  the generation at which the species  $S_i$  appeared
- Each living individual can make a transition to a random neighbor species with probability  $X$ .
- Evolutionary link between two species registered in the evolutionary history.
- Model stops at generation  $T_f$ .



## Results: Evolutionary history recovery

- Model outputs: set of species, their evolutionary history, similarity between each pair of species.
- Focus on genomes and the precise genealogy of the various species.
- Differs to works based on fixed-population Wright-Fisher model, where statistics of distances between species or time to the most recent common ancestor are main features.

# Results: Evolutionary history recovery





# Results: Evolutionary history recovery

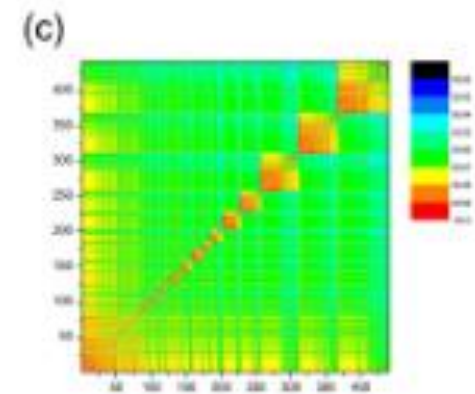
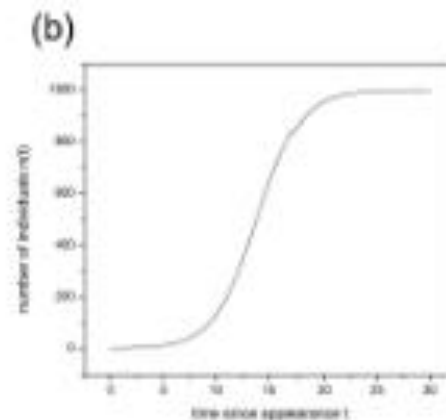
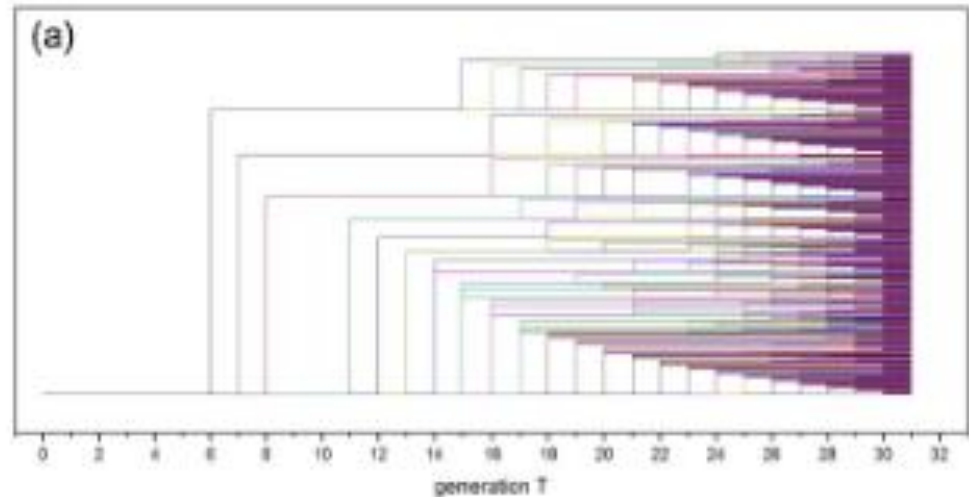
- (a)
- A species is represented by a strand of  $N$  genes (here 6), on active (black) or inactive (white) state.
- In each step, this species can mutate with probability  $X$  to a neighbor species.
- (b)
- Set of all species  $S$  divided into sub-sets  $S_\sigma$ , with  $\sigma$  a numeric sequence identifying each subset.
- Numbers in the sequence indicate all generations at which species belonging to that subset suffered a mutation.

# Results: Evolutionary history recovery

- $T=0$ ; single species in the subset identified as  $\sigma_0=0$ .
- For  $T>1$ , there are  $2^{T-1}$  new subsets, labelled by  $2^{T-1}$  new sequences  $\sigma_k$ ,  $k=2^{T-1}, 2^{T-1}+1, \dots, 2^T-1$ .
- At  $T=1$ , there is the  $\sigma_0$  subset, and  $\sigma_1=0,1$ .
- $N_\sigma$  denotes the number of different species in  $S_\sigma$  with  $N_{\sigma_0}=1$ .
- $N_\sigma$  depends on the random introduction of changes in the genome of the species ancestor.
- If no change occurs,  $N_\sigma=0$  for that particular sequence and for all sequences resulting by adding new numbers to this sequence.

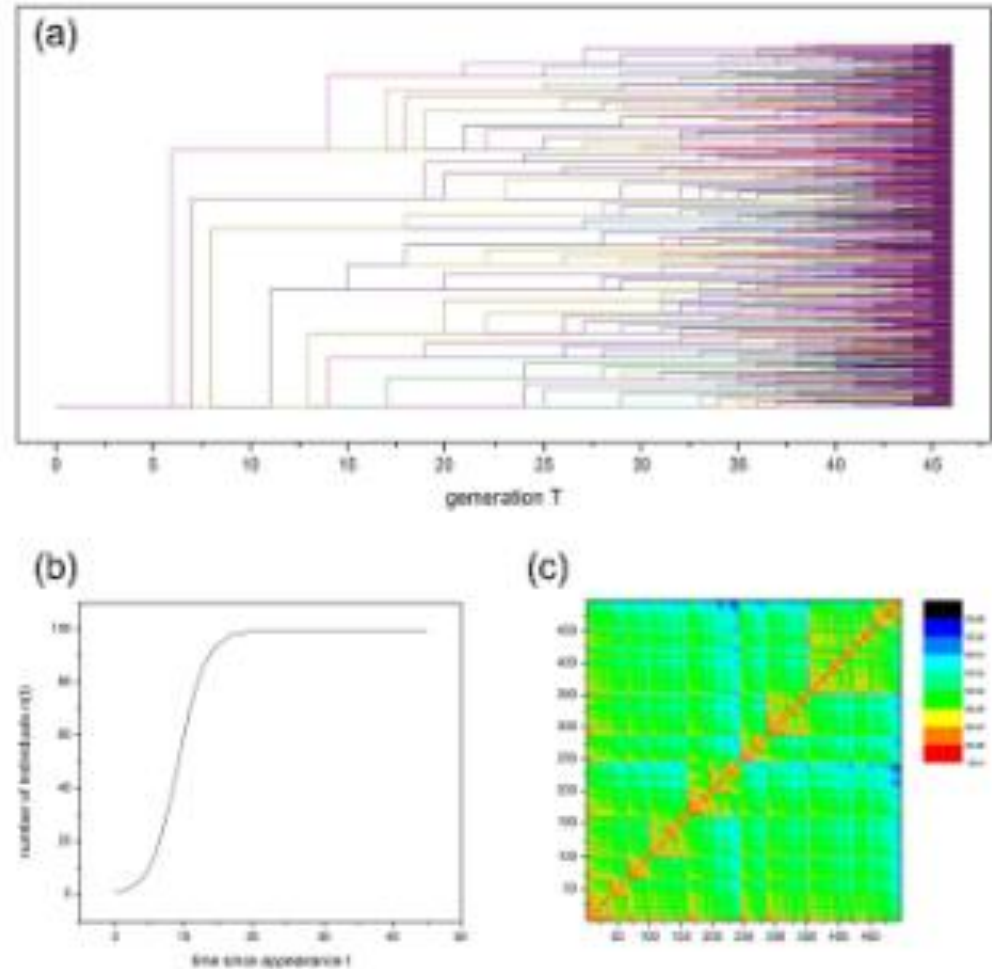
# Results: Evolutionary history recovery

- Dendrogram, growth  $n(t)$ , and similarity matrix using dendrogram numbering.
- $N=10000$ ,  $X = 0.005$ ,  $T_f=30$ ,  $r=0.5$ ,  $k=1000$ .
- Well-defined modular structures.



# Results: Evolutionary history recovery

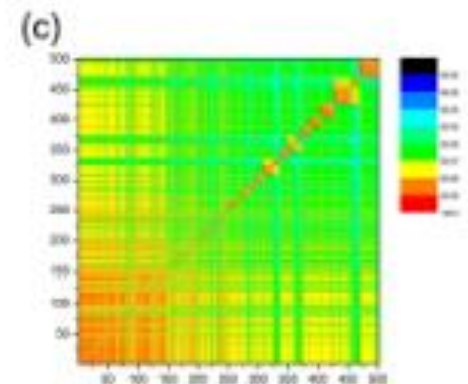
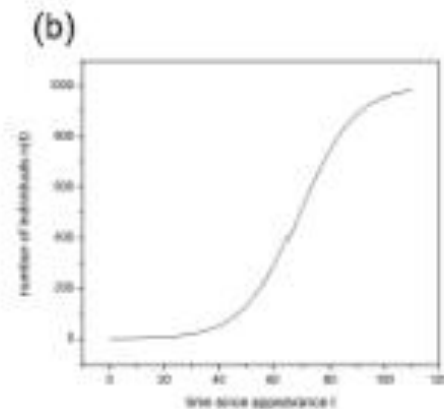
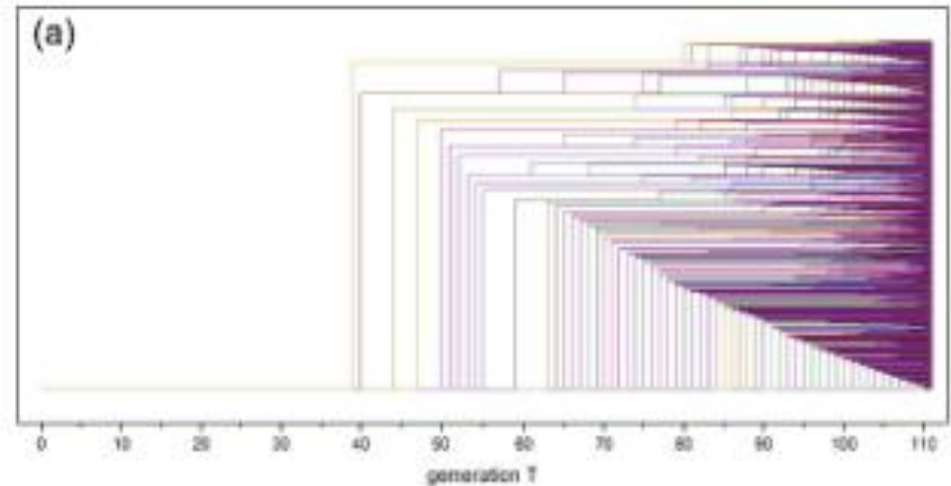
- Dendrogram, growth  $n(t)$ , and similarity matrix using dendrogram numbering.
- $N=10000$ ,  $X = 0.005$ ,  $T_f=45$ ,  $r=0.5$ ,  $k=100$ .
- Modular structure is not well developed.



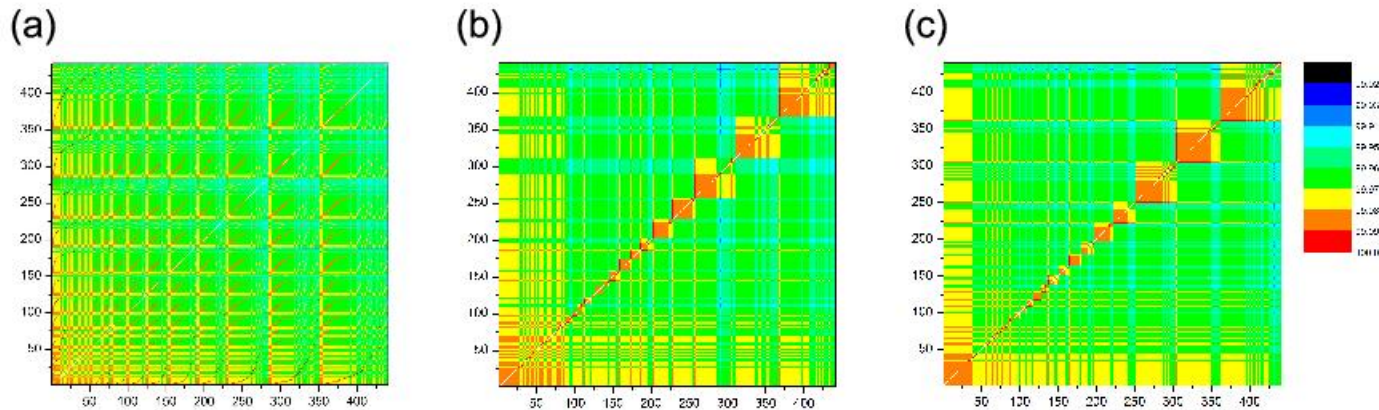


# Results: Evolutionary history recovery

- Dendrogram, growth  $n(t)$ , and similarity matrix using dendrogram numbering.
- $N=10000$ ,  $X = 0.005$ ,  $T_f=110$ ,  $r=0.1$ ,  $k=1000$ .
- Large community comprising neighbors of original species.



# Results: Evolutionary history recovery



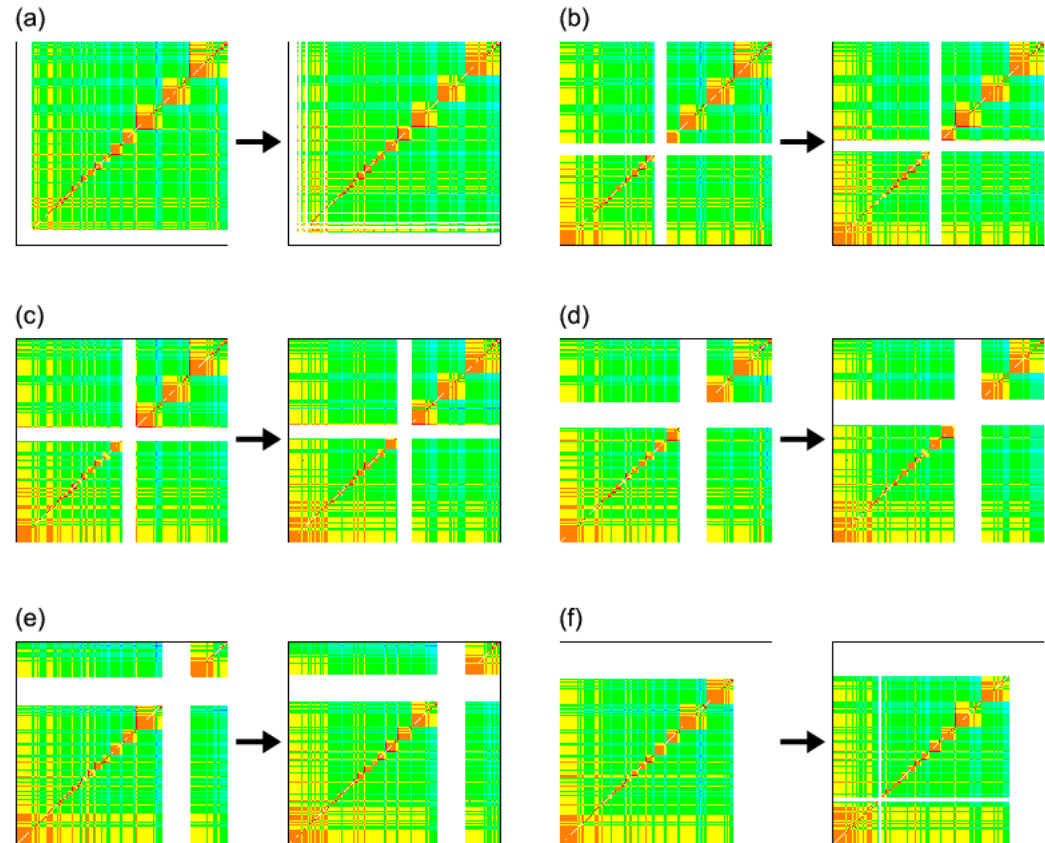
## ■ Similarity matrix

- (a) original numbering.
- (b) NG numbering.
- (c) dendrogram numbering.

# Results: Evolutionary history recovery

- (a) from 1 to 34
- (b) from 197 to 221;
- (c) from 222 to 250;
- (d) from 251 to 304;
- (e) from 305 to 361;
- (f) from 362 to 441.

■ Community comprising species in the NG and dendrogram similarity matrix;





# Conclusions

- Networks based on similarity of protein structure successfully used for phylogenetic analysis
- Identification of similarity threshold where modularity is likely to be expressed
- Community identification at a set of optimal threshold values
- Results for different sets of issues
  - Chitin synthesis enzymes (pathways)
  - Chitin pathways in fungi: comparison to other methods
  - Evolutionary origins of mitochondria
  - Evolutionary history recovery



# Conclusions

- Phylogenetic groups revealed by community finding algorithms (NG)
- Significant and remarkably agreement between communities and validated phylogenetic methods (ML, parsimony, Bayesian, distance).
- Able to address relevant biological issues as mitochondria evolutionary history.
- Recovers evolutionary dynamics of simple models.



**Thank you for your attention**