FESC

A Complex Network Approach for Phylogenetic Analysis

Suani Pinho
(suani@ufba.br)

The Team

- **Physicists and Mathematicians**
  - Suani Pinho (UFBa)
  - Leonardo Santos (UFBa / INPE)
  - Roberto Andrade (UFBa)
  - Thierry Petit Lobão (UFBa)
  - José Miranda (UFBa)
  - Ernesto Borges (UFBa)
- **Computer Scientists**
  - Marcelo Diniz (UEFS / IFBa))
  - Ivan do Carmo Neto (UFBa)
  - Charles Santana (UFBa / IMEDEA – Sapin)
- **Biologists**
  - Aristóteles Goés-Neto (UEFS)
  - Charbel El-Hani (UFBa)

Mathematical and Computer Modelling of Biosystems

- Nonlinear differential equations (population-based models)
  - population dynamics models (Malthus, Verhulst)
  - ecological models (Lotka-Volterra)
  - epidemics models (Kermack and Mc. Kendrick)
- Cellular automata (individual-based models) – Von Neumann and Ulam
- Complex networks (based on graph theory) – Barabasi-Albert, Watts-Strogatz models
- Data mining (genome project) – Watson and Crick

Interdisciplinary Approach

- Biomathematics
- Biostatistics
- Physical Biology
- Computational Biology

System Biology

Outline

- Introduction – complex networks
- Fundamental concepts
  - Neighborhood matrices
  - Distance between networks
  - Identifying community structure
- Application in Biology
  - Protein similarity networks
  - The computational methodology
- Concluding remarks

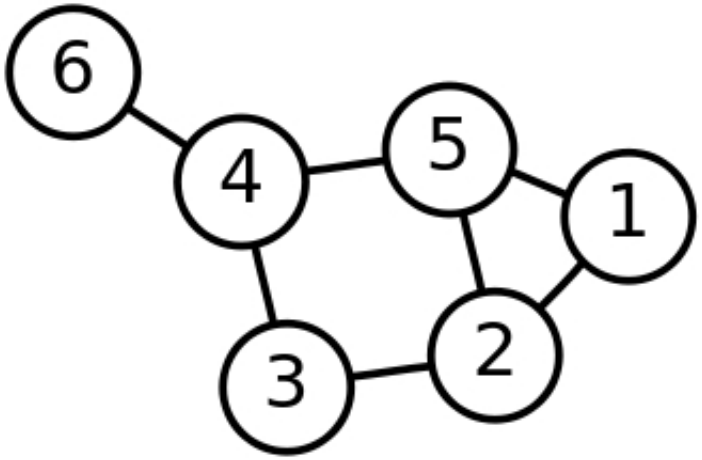
Complex Networks

- The regular and random networks are not suitable to describe some biological and social networks. In the last 20 years, complex networks, neither regular nor random, were proposed and applied to real systems.
- The concepts of Graph Theory and, more recently, Statistical Physics are used to study complex networks, comparing with regular and random networks.
- Identification of community structure is a hot topic in the context of complex networks. It is particularly relevant for biological networks.
- In this talk, we present and analyse a neighborhood matrix representation of network and its computational powerful to its characterization. We apply this concept to phylogentic analysis based on protein similarity networks.

Basic Concepts

Neighborhood Matrix

Graph



$M_1 = A$

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

$\hat{M}$

$$b)\begin{pmatrix} 0 & 1 & 2 & 2 & 1 & 3 \\ 1 & 0 & 1 & 2 & 1 & 3 \\ 2 & 1 & 0 & 1 & 2 & 2 \\ 2 & 2 & 1 & 0 & 1 & 1 \\ 1 & 1 & 2 & 1 & 0 & 2 \\ 3 & 3 & 2 & 1 & 2 & 0 \end{pmatrix}$$

$$\hat{M} = 1\begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} + 2\begin{pmatrix} 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 \end{pmatrix} + 3\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

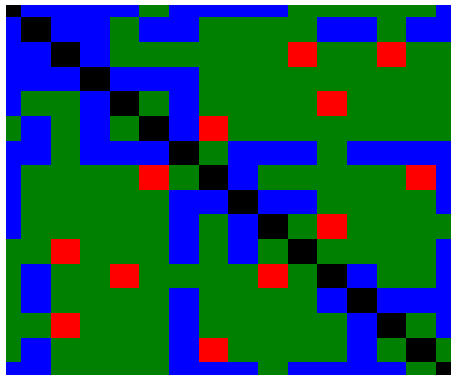
Higher order neighborhood

- Define neighborhood of order ℓ
- ℓ steps to walk from node to another in the ℓ -neighborhood
- Boolean matrix M_ℓ describes ℓ-neighborhood
- $M_0 = I$: each node is in *0*-neighborhood of itself
- $M_1 = A$
- $M_2 = [(M_0 \oplus M_1) \otimes M_1] - (M_0 \oplus M_1) = [(I+A) \otimes A] - (I \oplus A)$
- *In general:*

$$M_\ell = \left(\oplus_{g=0}^{\ell-1} M_g\right) \otimes M_1 - \left(\oplus_{g=0}^{\ell-1} M_g\right)$$

Color Representation of Neighborhood Matrix

```
0111101111000001     00000200002222O     0000000000000000
1011011000011011     0000200222002OO     0000000000000000
1101000000000000     0000222222022O2     0000000000300300
1110111000000000     0000000222222222     0000000000000000
1001001000000000     022002022202222     0000000000030000
0101001000000000     202020002222222     0000000300000000
1101110011101111     002000020002O000     0000000000000000
1000000010000001  +  022220200222220O  +  0000030000000030
1000001101100001     022222000022220     0000000000000000
1000001010000000     022222020020222     0000000000030000
0000001010000001     220222020202222O     0030000000000000
0100000000001001     202202222020022O     0000300003000000
0100001000010111     202222022220000O     0000000000000000
0000001000001001     220222022222002O     0030000000000000
0100001000001000     202222002222020 2     0000000300000000
1100001110111100     002222000200002O     0000000000000000
```



$$\hat{M} = \sum_{\ell=1}^{D} \ell \, M_{\ell}$$

Andrade, Miranda, Petit Lobao, PRE 73, 046101 (2006).

Using neighborhood matrix to characterize the network

- ✔ It carries out the same information of the adjacency matrix but this information is processed.
- ✔ Diameter ⇒ the maximum value of its elements
- ✔ Average shortest path $\langle d \rangle$ ⇒ mean value of its elements
- ✔ Edge betweenness ⇒ a simple algorithm to calculate the sum of fractions of shortest paths between the nodes that pass through an edge
- ✔ Color code plots to visualize the neighborhood structure of network

Andrade, Miranda, Pinho, Petit Lobao, Eur. Phys. J. B 61, 2470-256 (2008)

Distance between Networks

- Define distance (or "numbering energy") **E** between networks **P** and **Q** :

$$E = \sum_{i,j} [(\hat{M}_P)_{i,j} - (\hat{M}_Q)_{i,j}]^2$$

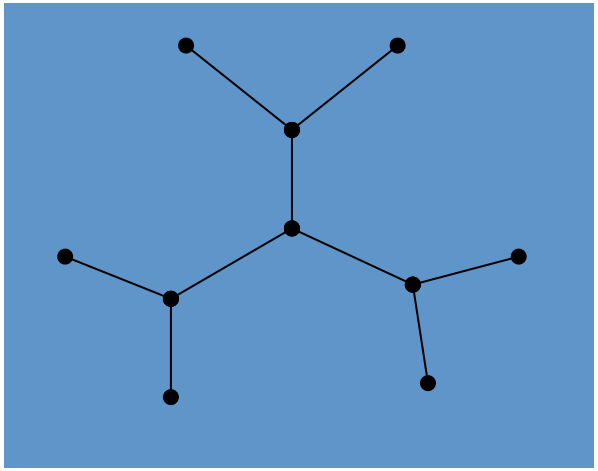
- Project **P** over **Q**, i.e., minimize **E** by Monte-Carlo procedure
- Changing randomly lines and columns of $\hat{M}_P$, the minimization of **E** leads to $\hat{M}_P$ with properties of $\hat{M}_Q$

Andrade, Miranda, Pinho, Petit Lobao; Phys. Let. A 372, 5265-5269 (2008).

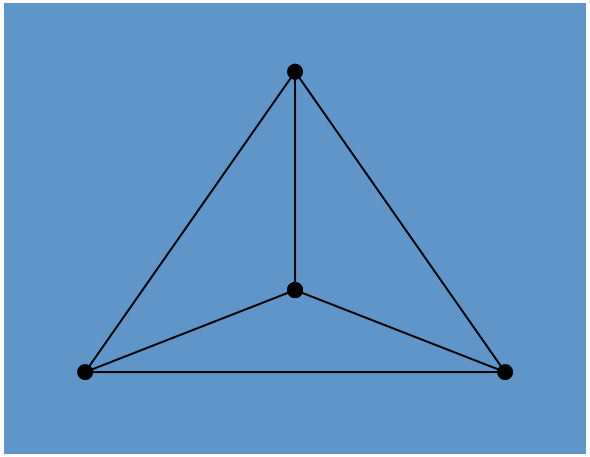
Deterministic networks

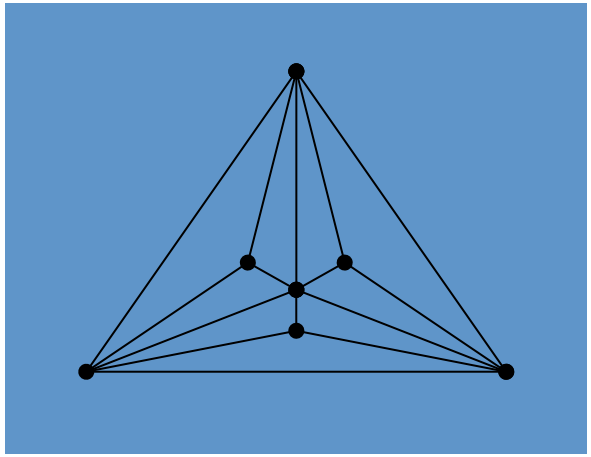
Bethe lattice

(Cayley tree)

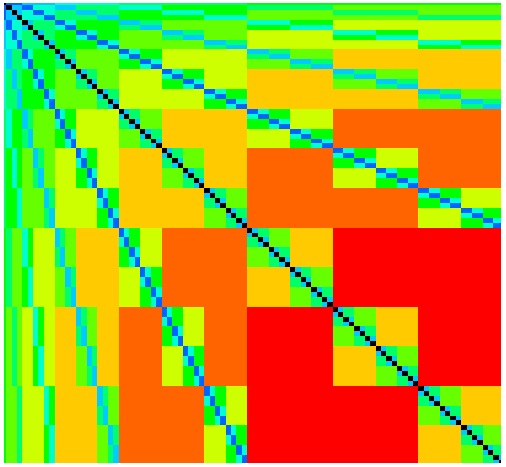


Appollonian network

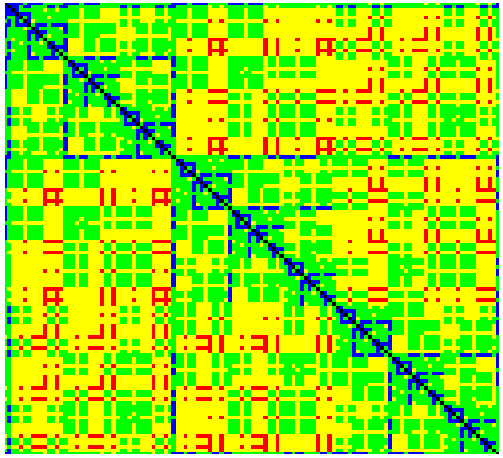




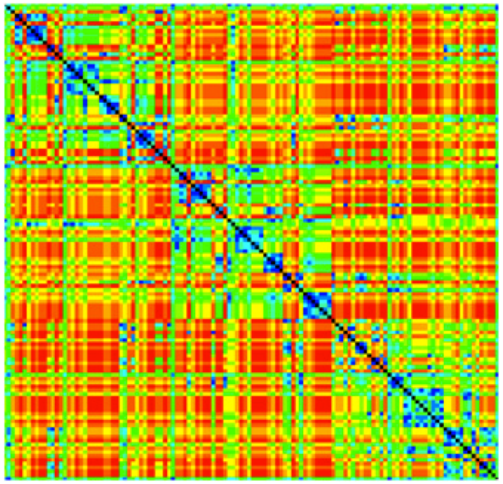
Projection based on the distance



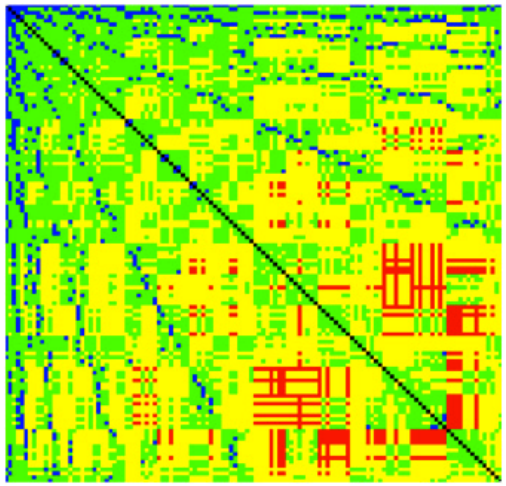
Cayley (C)



Apollonian (A)



C➔A



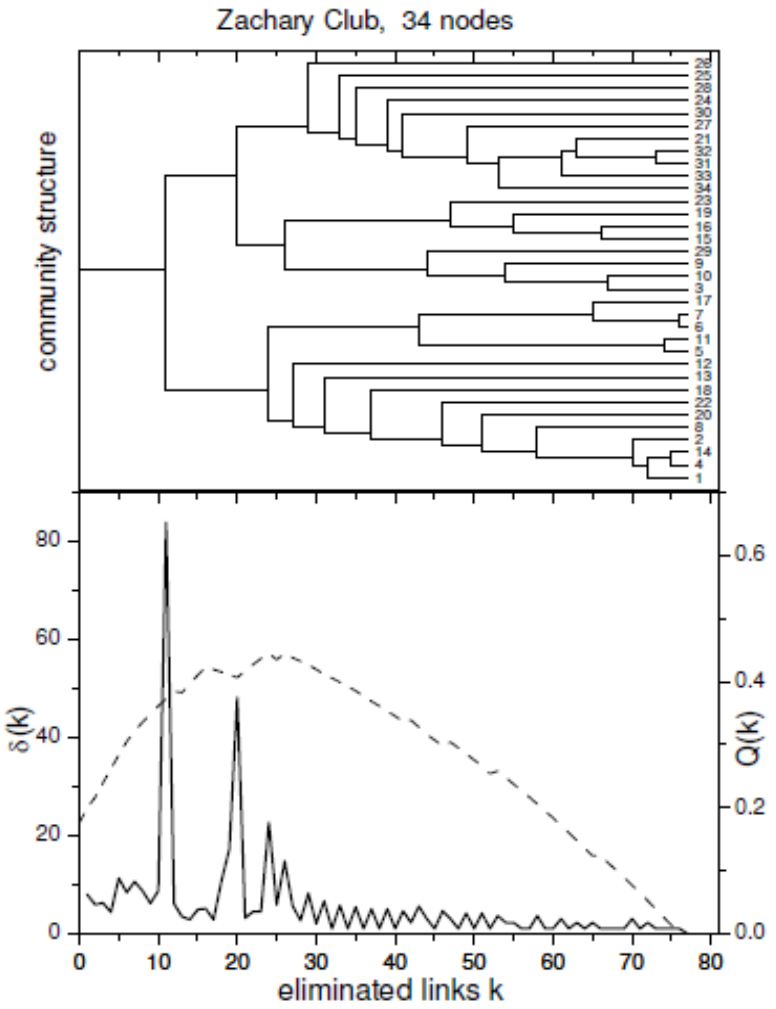
A➔C

Modular structure of a network

✔Newman-Girvan procedure – elimination of links with maximum edge betweenness using the modularity to indicate branching in dendrogram

✔Distance between successive eliminations of links in NG procedure is more precise than modularity.

Andrade, Pinho, Petit Lobao, Int. J. Bif. Chaos App. Sci. Eng. 19, 2677-2685 (2009)



2: a) Dendrogram produced by the sequence

Protein similarity networks

Phylogenetic analysis using networks

Methodology:

Database and network construction

- ✔ Select the protein sequences corresponding to the enzymes of a metabolic pathway of genomes of organisms. Select also the relevant information to set up the similarity level between the sequences (Genebank - NCBI – www.ncbi.nlm.nih.gov).

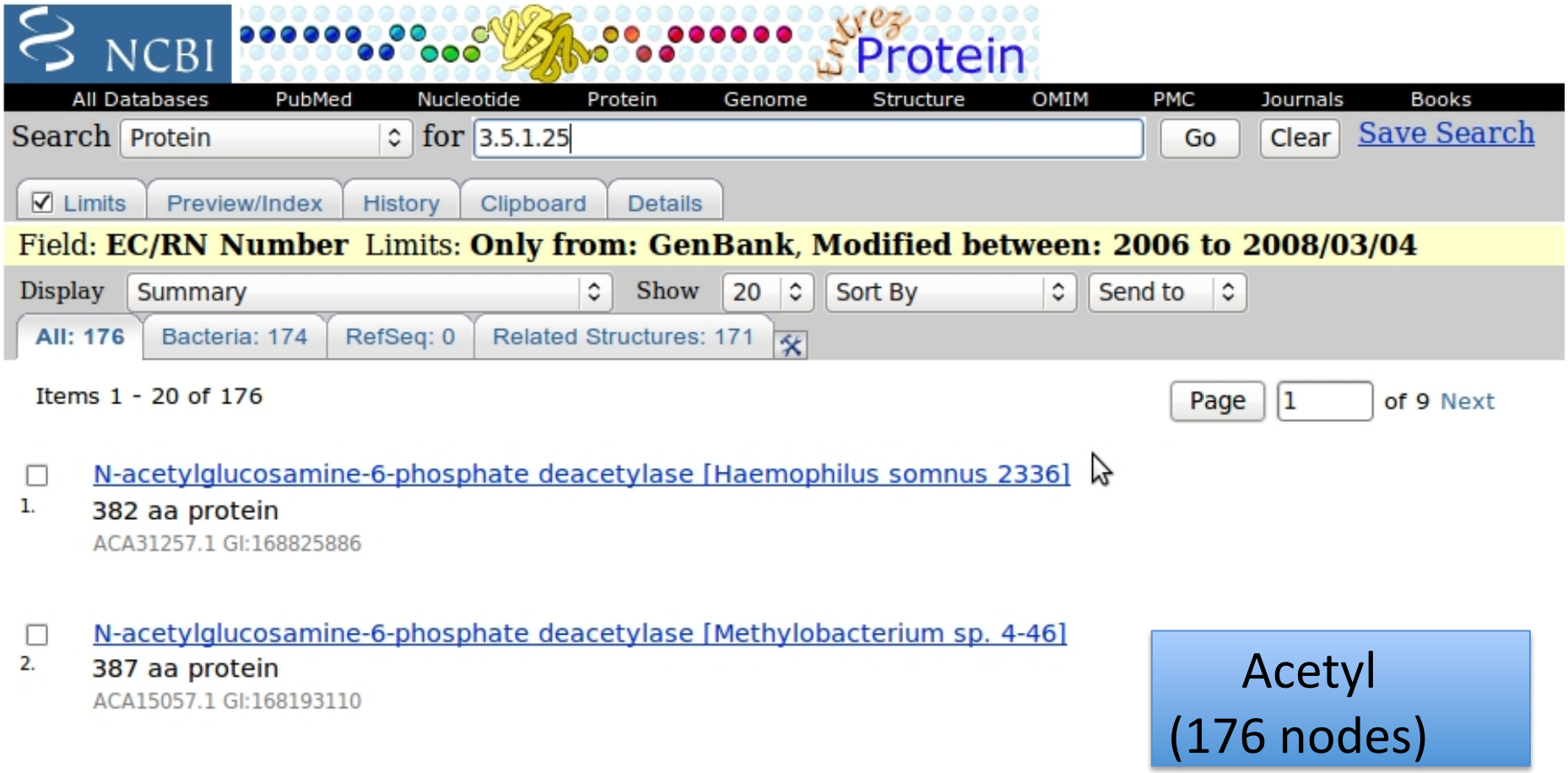
- ✔ The comparison between the sequences is performed by the program BLAST . As a result a similarity matrix N x N is set up, associated with each enzyme of the metabolic pathway, with N as the number of protein sequences.

- ✔ Generate 101 networks associated with the similarity threshold (σ) from 0 to 100 : the nodes are the protein sequences and there is a link betwen nodes if the similarity is greater or equal to σ.

Metabolic pathway: chitin

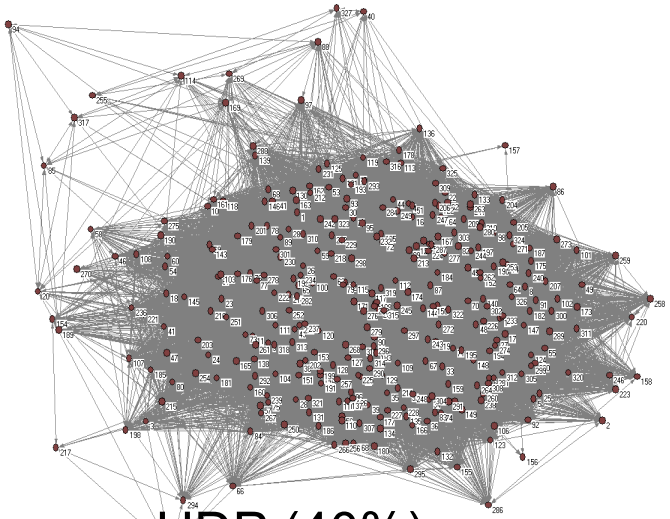
There are 1695 protein sequences corresponding to 13 enzymes

Enzyme	Enzymatic classification	Domain (#)
UDP-acetylglucosamine pyrophosphorylase	2.7.7.23	E(2), B(324), A(2)
Acetylglucosamine phosphate deacetylase	3.5.1.25	B(170), A(6)
Glucosaminephosphate isomerase	2.6.1.16	E(23), B(285), A(5)
Hexosaminidase	3.2.1.52	E(3), B(235)
Phosphoglucoisomerase	5.3.1.9	E(16), B(472), A(12)

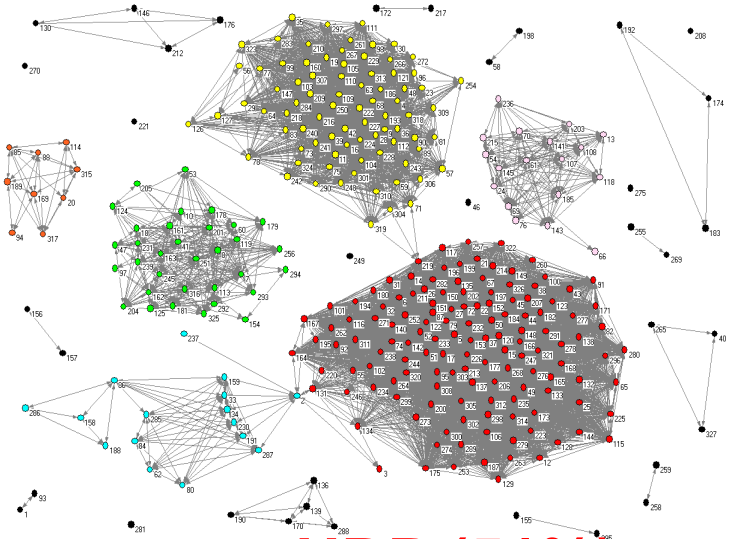

NCBI
Entrez Protein
All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books
Search Protein for 3.5.1.25 Go Clear Save Search
Limits Preview/Index History Clipboard Details
Field: EC/RN Number Limits: Only from: GenBank, Modified between: 2006 to 2008/03/04
Display Summary Show 20 Sort By Send to
All: 176 Bacteria: 174 RefSeq: 0 Related Structures: 171
Items 1 - 20 of 176
Page 1 of 9 Next
1. N-acetylglucosamine-6-phosphate deacetylase [Haemophilus somnus 2336]
382 aa protein
ACA31257.1 GI:168825886
2. N-acetylglucosamine-6-phosphate deacetylase [Methylobacterium sp. 4-46]
387 aa protein
ACA15057.1 GI:168193110
3. N-acetylglucosamine-6-phosphate deacetylase [Xylella fastidiosa M12]
385 aa protein
ACA11796.1 GI:167964786
4. N-acetylglucosamine-6-phosphate deacetylase [Haemophilus parasuis 29755]
383 aa protein
EDS25027.1 GI:167853786
Acetyl
(176 nodes)

Characterizing the networks

- To set up the neighborhood matrix from the adjacency matrix.
- To calculate the size of the largest cluster for different values of σ.
- For low values of σ, there is a large cluster; for high values of σ, there are many sub-networks. For intermediate values of σ, the modular structure is revealed.

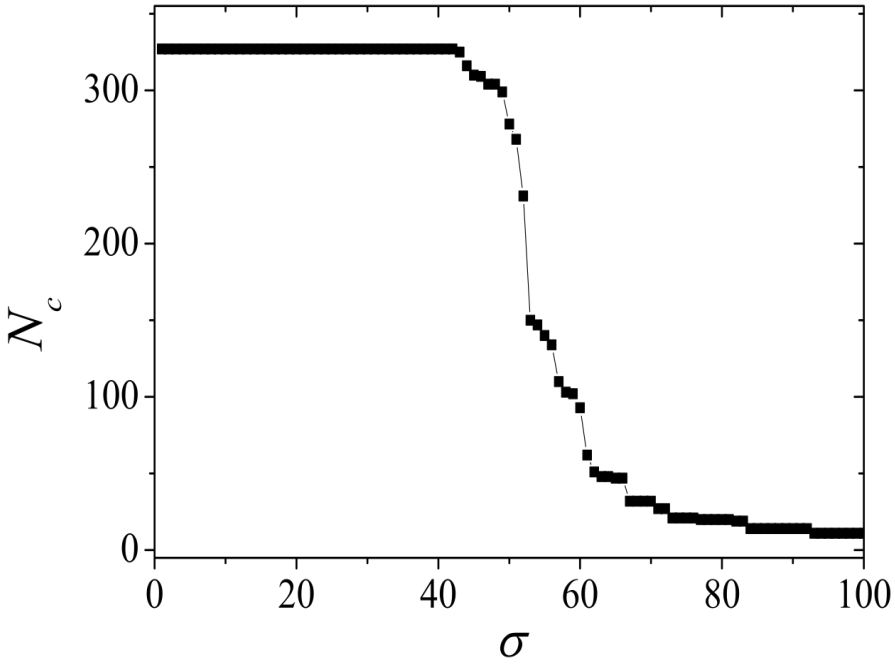


UDP (40%)

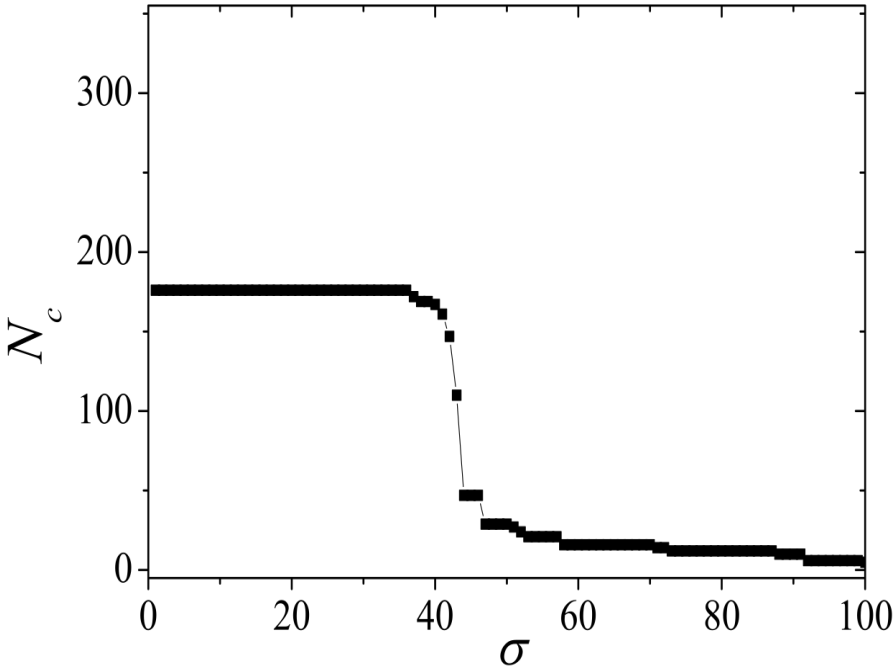


UDP (51%)

Size of the largest cluster (UDP e Acetyl)



UDP

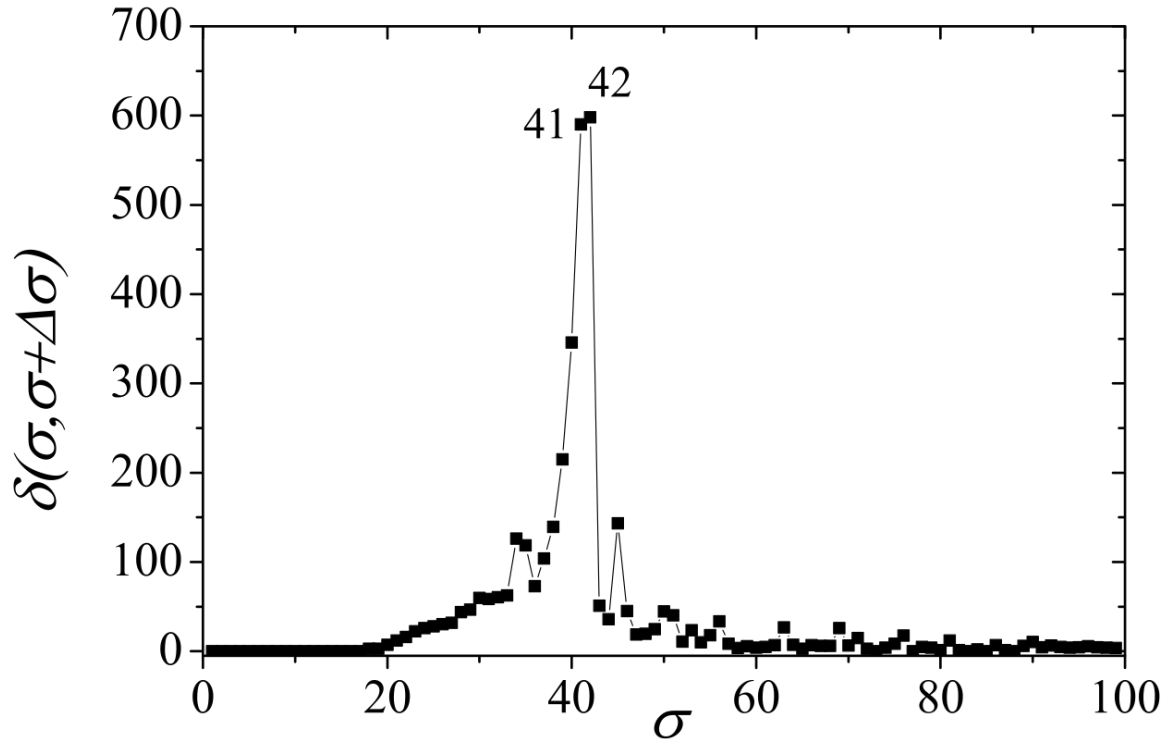


Acetyl

Goes-Neto, Diniz, Santos, Pinho, Miranda, Andrade, Petit Lobao, Borges, El-Hani. Biosystems 101: 59-66 (2010).

Critical network

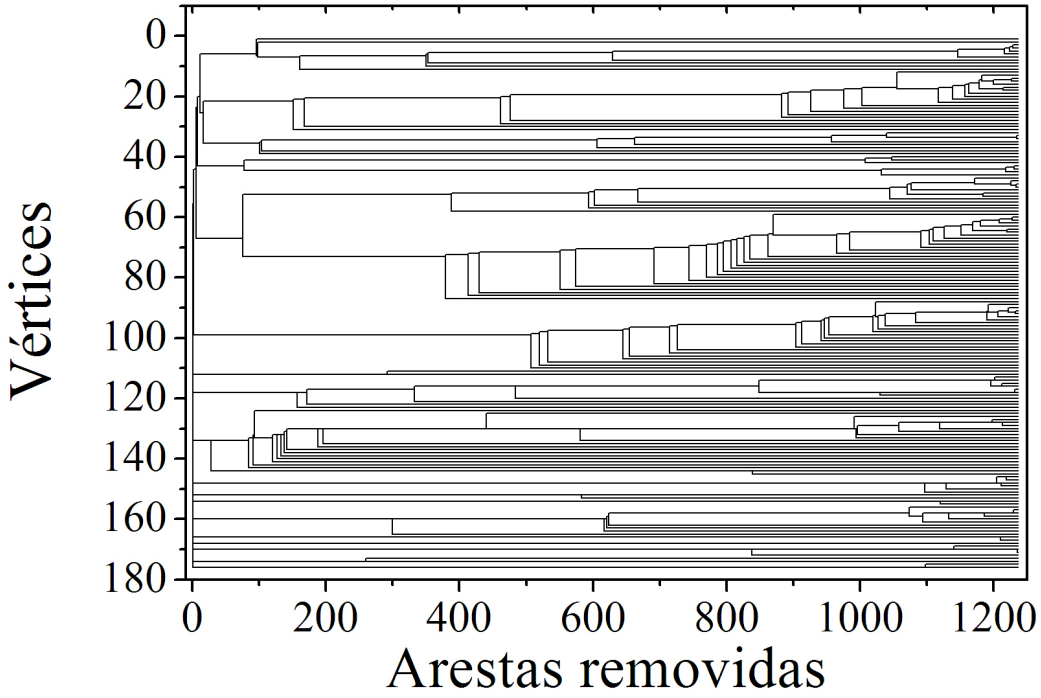
- Distance between the networks (based on the neighborhood matrices) before and after the removing of the links: the maximum value corresponds to the critical network (acetyl: 42%).



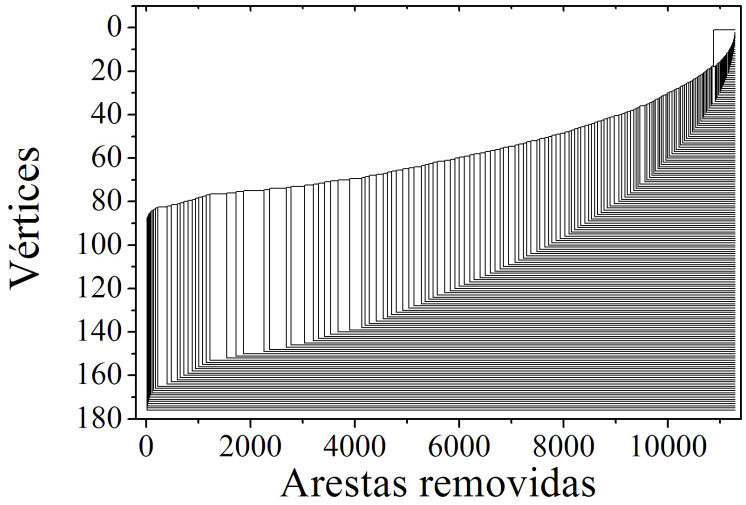
Edge Betweenness

- Edge Betweenness: sum of fractions of shortest paths connecting the pairs of nodes through a certain edge.
- Assuming the NG procedure (Newman and Girvan, PRE **69**, 026113 (2004):
  - Remove the edge with the maximal value of edge betweenness.
  - Repete the process until there is no link.
- In order to set up the modular structure, we set up the dendrogram for the critical network as well as the color representation of the neighborhood matrix.

Dendrograms (acetyl)

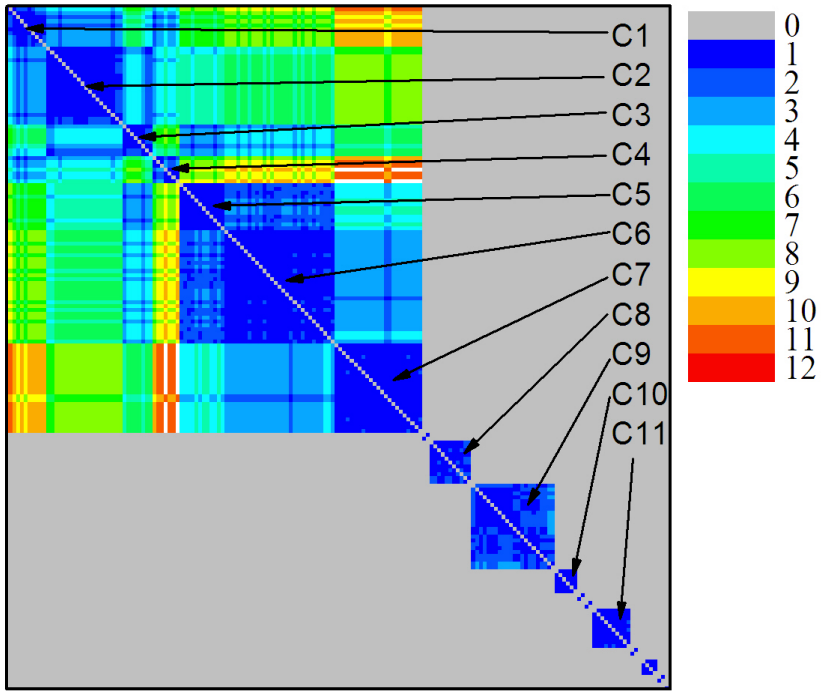


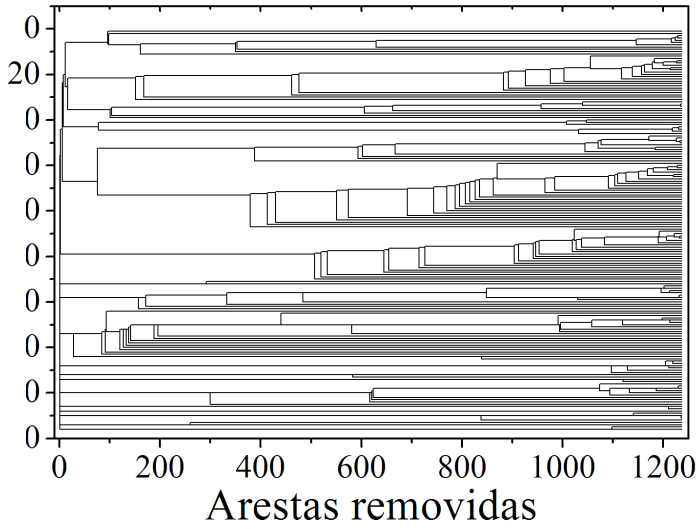
Dendrogram of critical netwrok (42%)

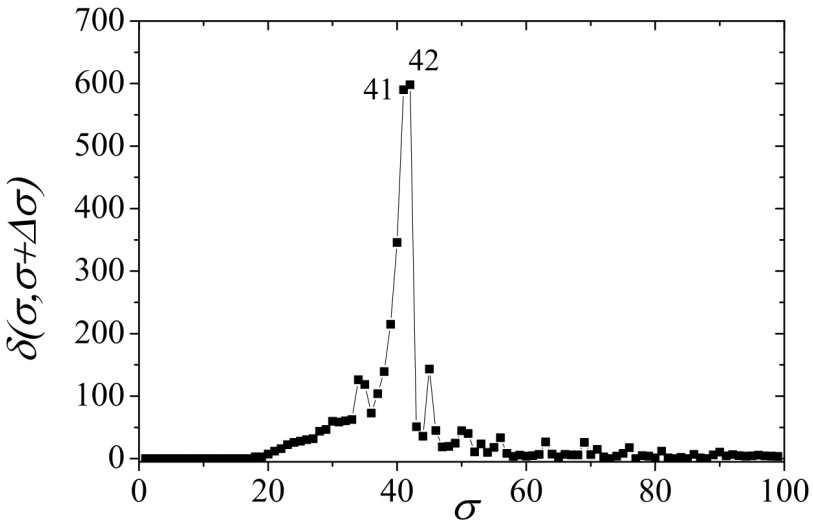


Dendrogram (30%)

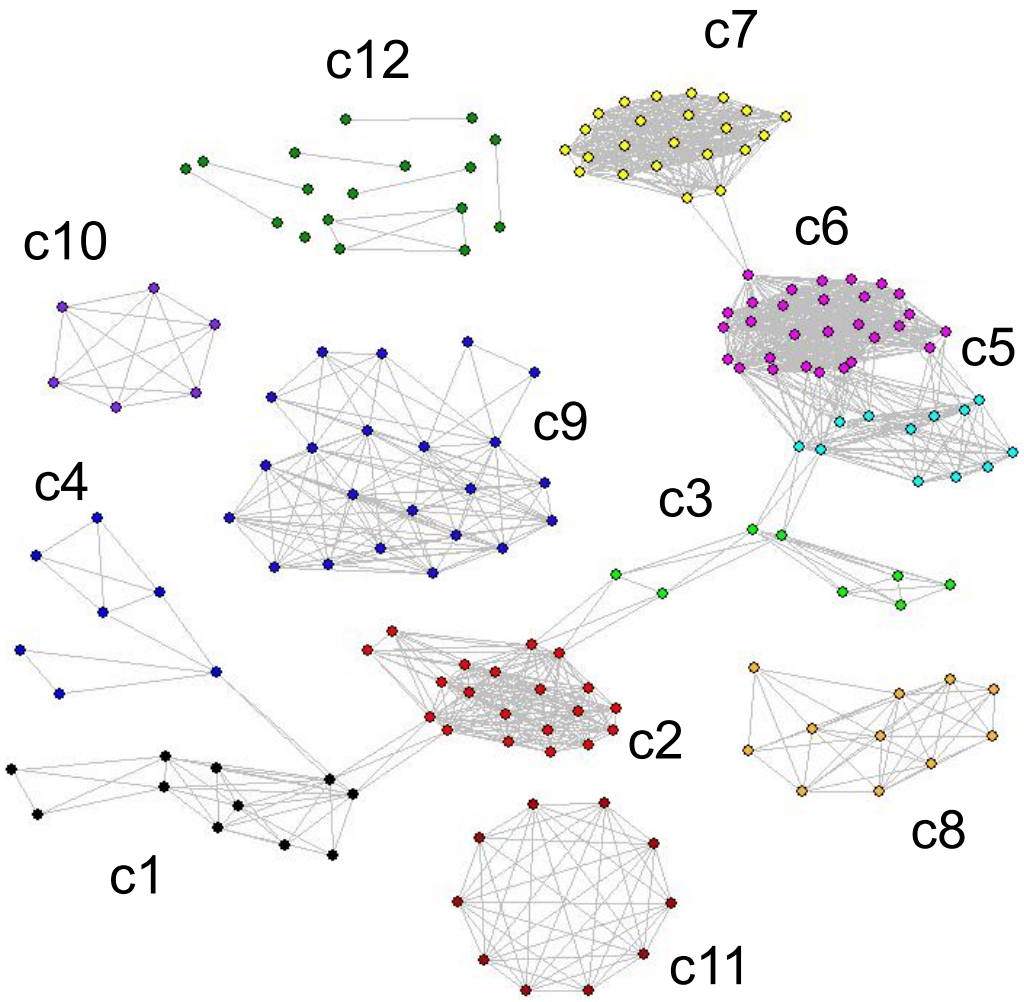
Identifying the modular structure







Modular network of Acetyl



Com_1.dat: 7 Actinobacteria; 4 Crenarchaeota;
Com_2.dat: 14 Betaproteobacteria; 2 Gammaproteobacteria; 2 Deinococcus-Thermus; 2 Alphaproteobacteria;
Com_3.dat: 8 Firmicutes;
Com_4.dat: 7 Actinobacteria;
Com_5.dat: 12 Alphaproteobacteria;
Com_6.dat: 27 Gammaproteobacteria; 2 Alphaproteobacteria;
Com_7.dat: 23 Gammaproteobacteria;
Com_8.dat: 12 Cyanobacteria;
Com_9.dat: 16 Firmicutes; 6 Gammaproteobacteria;
Com_10.dat: 6 Gammaproteobacteria;
Com_11.dat: 9 Firmicutes;
Outros.dat: 6 Actinobacteria; 4 Planctomycetes; 2 Crenarchaeota; 2 Bacteroidetes; 2 Acidobacteria; 1 Betaproteobacteria;

Summary of Results

Enzyme	$S_{th}^{(crit)}$	# nodes	# communities
UDP-acetylglucosamine pyrophosphorylase	51	327	6
Acetylglucosamine phosphate deacetylase	42	176	11
Glucosaminephosphate isomerase	40	313	4
Hexosaminidase	37	238	9
Phosphoglucoisomerase	37	501	5

Concluding Remarks

✔ The interdisciplinary character of this research project reveals the importance of getting together the knowledge of biologists with physicists, mathematicians and computer scientists. However the team was concerned about the computational features of it.

✔The neighborhhod matrix present the processed information about the network. It leads to the concept of distance between networks.

✔ The distance is more efficient in revealing the critical network in which the modular structure of the network, if it is the case, is revealed.

✔The above concepts are applied in protein similarity proteins revealing the classification of organisms that present these protein sequences.

✔The methodology may be applied to other networks that presents a modular character.

Thank you!

Congruence

Let *p* and *q* the critical networks associated with two enzymes; N_{pq} the number of common organisms; the number of matching organisms M_{pq}, i.e., number of organisms that are placed in the same communities in the two networks. If the number of communities in networks *p* and *q* are different, it is necessary to make a correspondence of two or more communities of network *p* to the same community in network *q*. The value G_{pq} is just the ratio M_{pq}/N_{pq}.

Ex: UDP (Norg=245) e Acetyl (Norg=88) : N_{pq}=44 ; M_{pq}=40; G_{pq} = 91%

UDP (327)

Acetyl (176)

	u1	u2	u3	u4	u5	u6
a1	0	0	0	0	2	0
a2	0	0	3	1	0	0
a3	0	0	0	0	0	0
a4	0	0	0	0	0	0
a5	0	0	0	3	0	0
a6	0	0	8	1	0	0
a7	0	0	8	0	0	0
a8	2	0	0	0	0	0
a9	0	10	2	0	0	0
a10	0	0	0	0	0	0
a11	0	4	0	0	0	0

Interseção	44

Associação	40
u1a8	2
u2a9a11	14
u3a2a6a7	19
u4a5	3
u5a1	2

Não-congruentes	4
u3a9	2
u4a2	1
u4a6	1

Congruência	40/44
Congruência	90,9%

References

[1] Andrade, R. F. S.; Miranda, J. G. V.; Lobão, T. P.; *Neighborhood concepts in complex networks.* Phys. Rev. E **73**, 046101 (2006).

[2] Andrade, R. F. S. ; Miranda, J. G. V. ; Pinho, S. T. R. ; Lobao, T. C. P.; *Characterization of complex networks by higher order neighborhood properties.* Eur. Phys. J. B **61**, 247-256 (2008).

[3] Andrade, R. F. S. ; Miranda, J. G. V. ; Pinho, S. T. R. ; Lobao, T. P.; *Measuring distances between complex networks*, Phys. Lett. A **372**, 5265–5269 (2008).

[4] Andrade, R.F.S.; Pinho, S.T.R.; Petit Lobão, T.C.; *Identification of community structure in networks using higher order neighborhood concepts.* Int. J. Bifurc. Chaos. 19, 2677-2685 (2009).

[5] Góes-Neto, A. ; Diniz, M. V. C. ; Santos, L. B. L. ; Pinho, S. T. R. ; Miranda, J. G. V. ; Andrade, R. F. ; Petit Lobao, T.; Borges, E.P.; El-Hani; *Comparative protein analysis of the chtin metabolic pathway in extant organisms: a complex network approach.* Biosystems 101: 59-66 (2010).