

Ch4 Network topology

Part2 Network modules and motifs

Zhuo Wang

zhuowang@sjtu.edu.cn

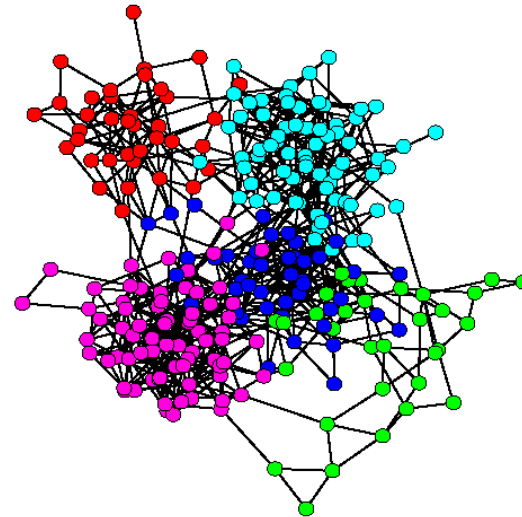
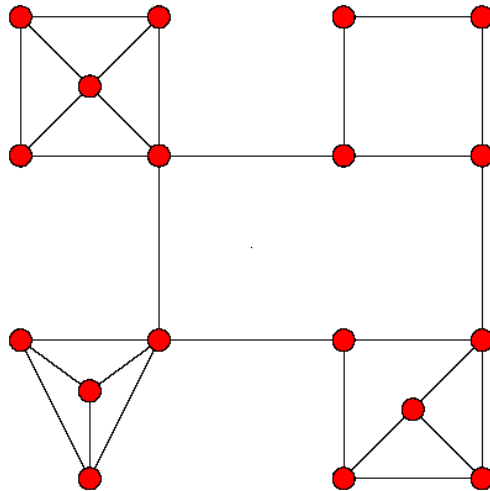
Modularity in Cellular Networks

➤ Hypothesis:

Biological functions are carried by discrete functional modules.

❖ Hartwell, L.-H., Hopfield, J. J., Leibler, S., & Murray, A. W., *Nature*, 1999.

➤ Traditional view of modularity:



➤ Question: Is modularity a myth, or a structural property of biological networks?
(are biological networks fundamentally modular?)

From molecular to modular cell biology

Leland H. Hartwell, John J. Hopfield, Stanislas Leibler and Andrew W. Murray

Cellular functions, such as signal transmission, are carried out by 'modules' made up of many species of interacting molecules. Understanding how modules work has depended on combining phenomenological analysis with molecular studies. General principles that govern the structure and behaviour of modules may be discovered with help from synthetic sciences such as engineering and computer science, from stronger interactions between experiment and theory in cell biology, and from an appreciation of evolutionary constraints.

Although living systems obey the laws of physics and chemistry, the notion of function or purpose differentiates biology from other natural sciences. Organisms exist to reproduce, whereas, outside religious belief, rocks and stars have no purpose. Selection for function has produced the living cell, with a unique set of properties that distinguish it from inanimate systems of interacting molecules. Cells exist far from thermal equilibrium by harvesting energy

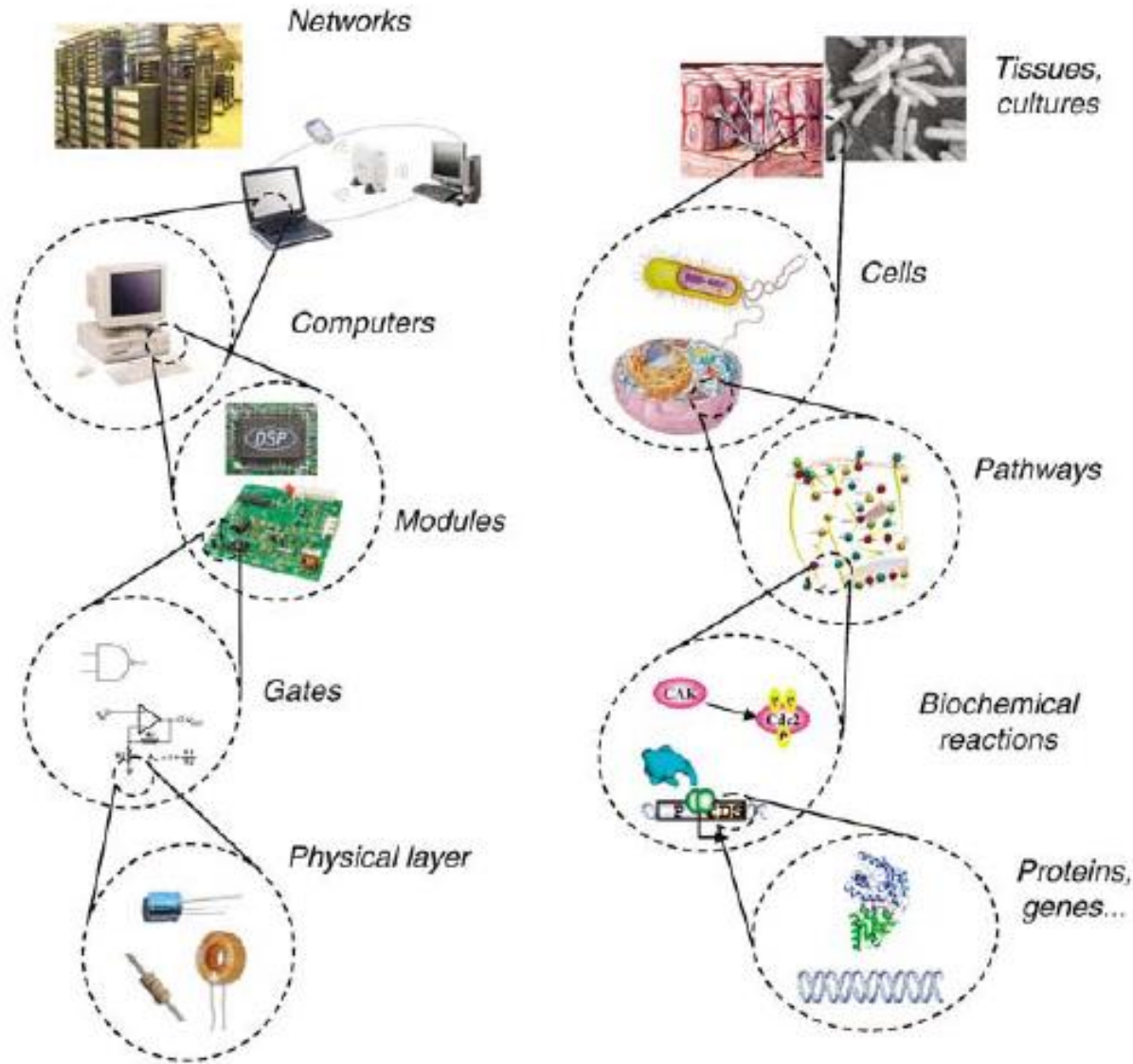
from many components. For example, in the signal transduction system in yeast that converts the detection of a pheromone into the act of mating, there is no single protein responsible for amplifying the input signal provided by the pheromone molecule.

To describe biological functions, we need a vocabulary that contains concepts such as amplification, adaptation, robustness, insulation, error correction and coincidence detection. For example, to decipher how the

Having described such concepts, we need to explain how they arise from interactions among components in the cell.

We argue here for the recognition of functional 'modules' as a critical level of biological organization. Modules are composed of many types of molecule. They have discrete functions that arise from interactions among their components (proteins, DNA, RNA and small molecules), but these functions cannot easily be predicted by studying

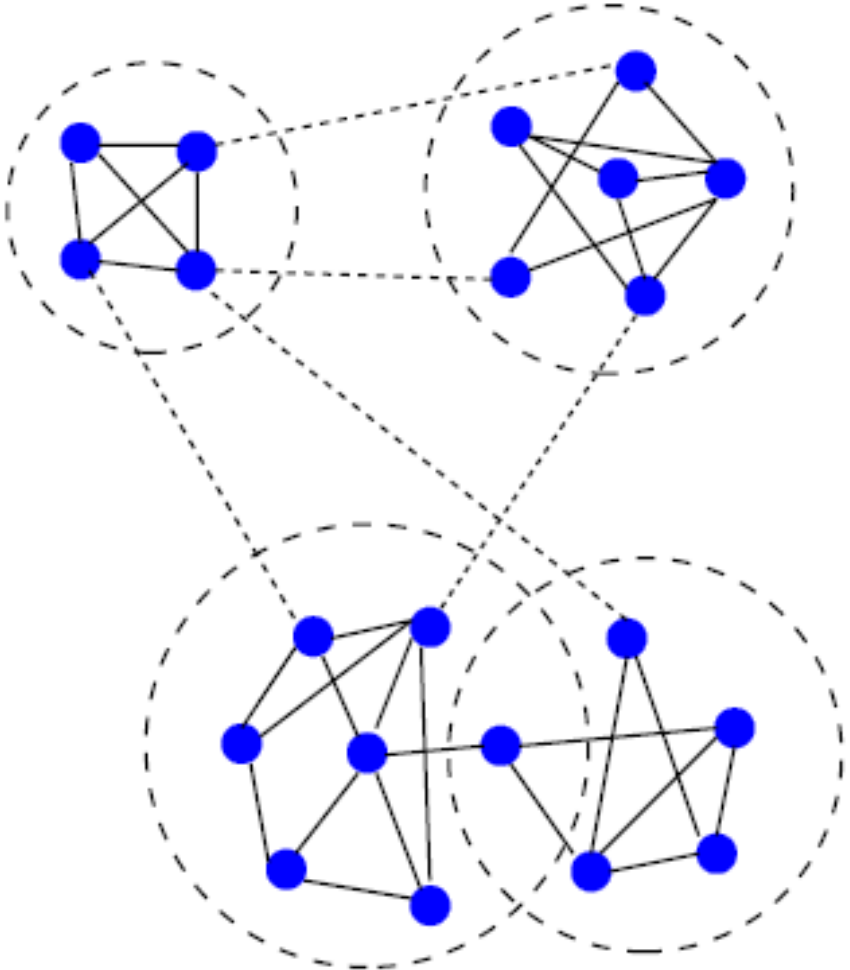
Modularity in cell biology



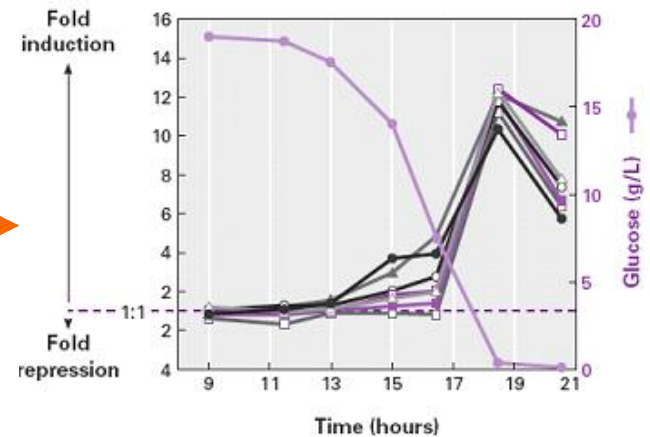
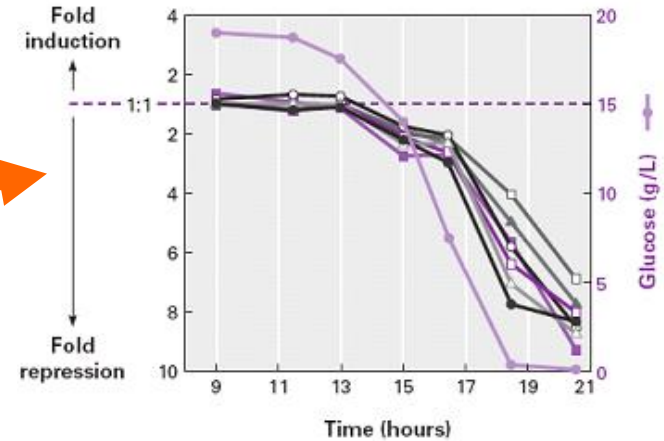
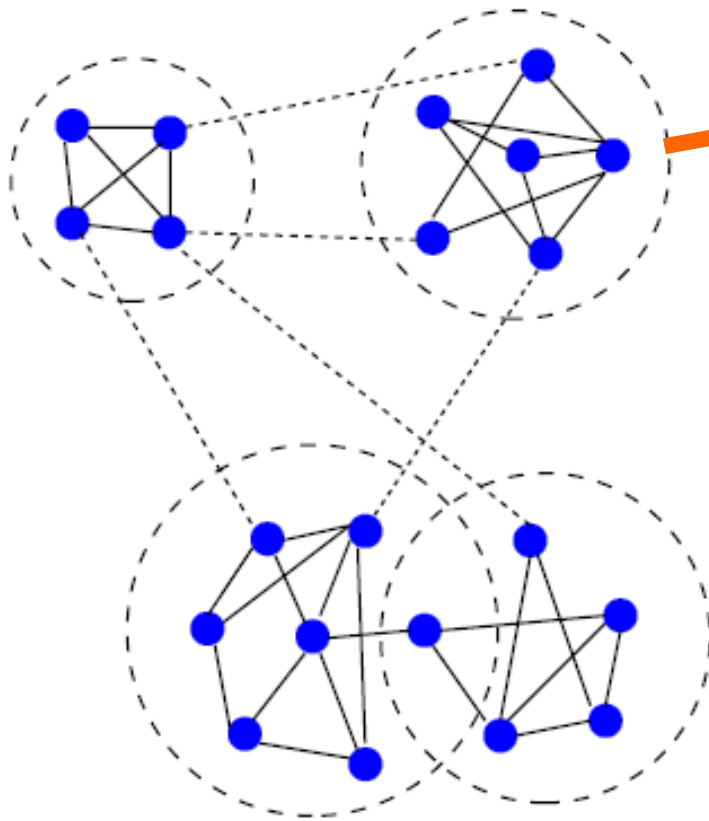
Definition of a module

- Loosely linked island of densely connected nodes
- Groups of co-expressed genes

Concept of modules in a network



Concept of modules in a network



Computational analysis of modular structures

Data clustering approach

Concept of data clustering analysis

- Partitioning a data set into groups so that points in one group are similar to each other and are as different as possible from the points in other groups.
- The validity of a clustering is often in the eye of beholder.

Concept of data clustering analysis

- In order to describe two data points are similar or not, we need to define a **similarity measure**.
- We also need a **score function** for our objectives.
- A clustering algorithm can be used to partition the data set with **optimized score function**.

Types of clustering algorithms

- Partition-based clustering algorithms
- Hierarchical clustering algorithms
- Probabilistic model-based clustering algorithms

Similarity measure for network clustering

- Correlation
- Shortest path length
- Edge betweenness

Score function for network clustering

- To maximize the intra group connections as many as possible and to minimize the inter group connection as few as possible.

Quantitative measurement of network modularity

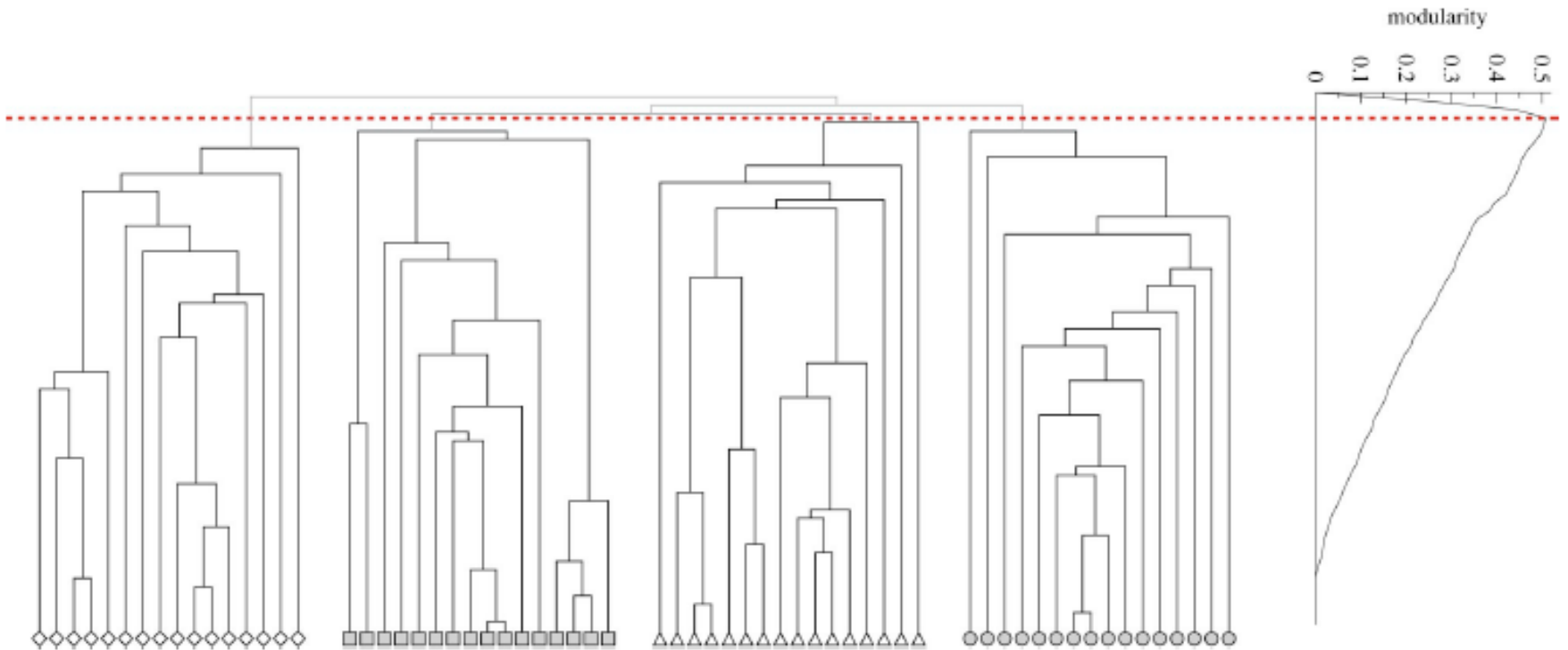
Modularity Q

$$Q = \sum_i e_{ii} - a_i^2$$

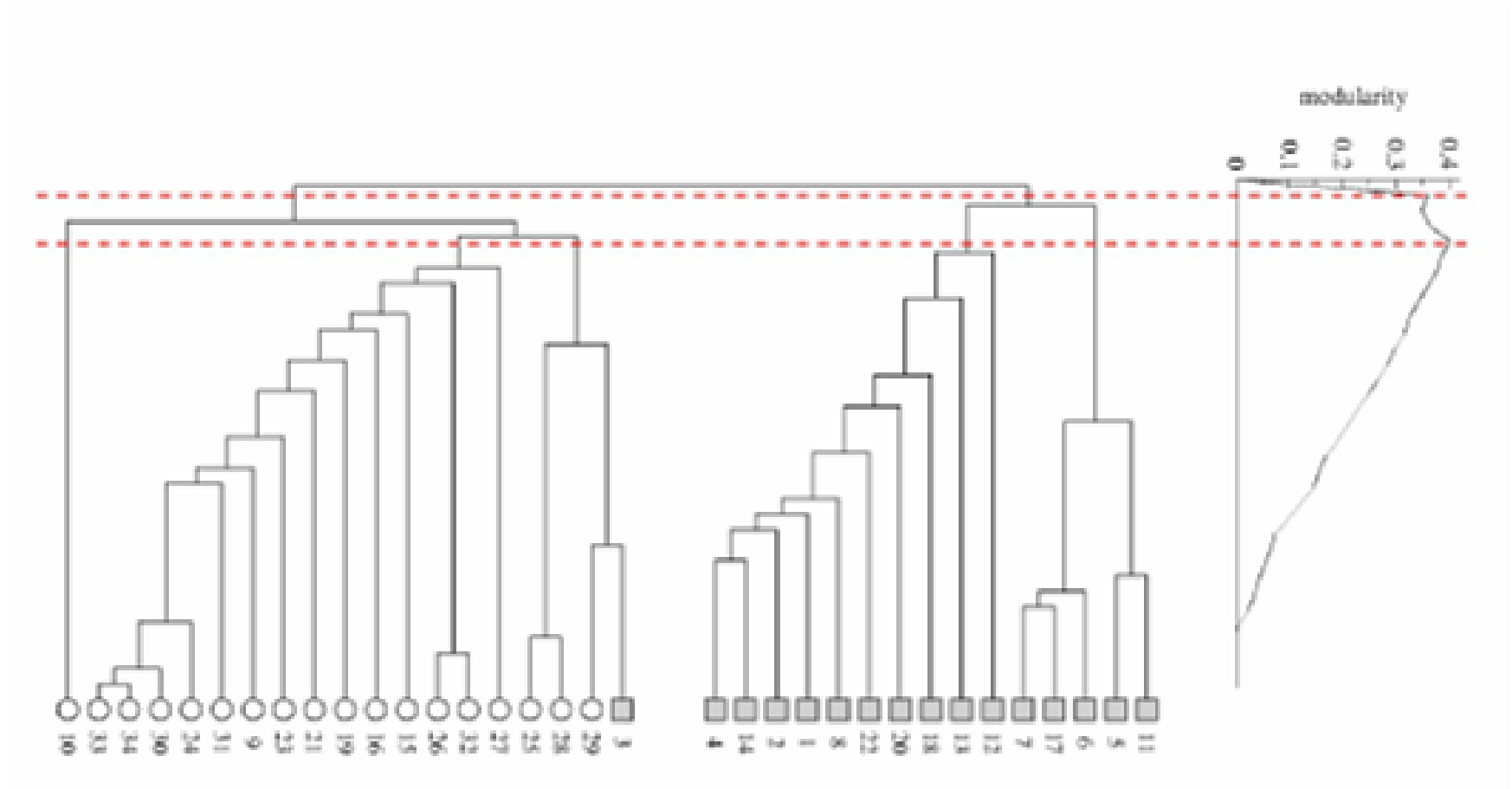
$$a_i = \sum_j e_{ij}$$

e_{ij} is the fraction of edges in network connecting module i and j

Threshold selection



club network

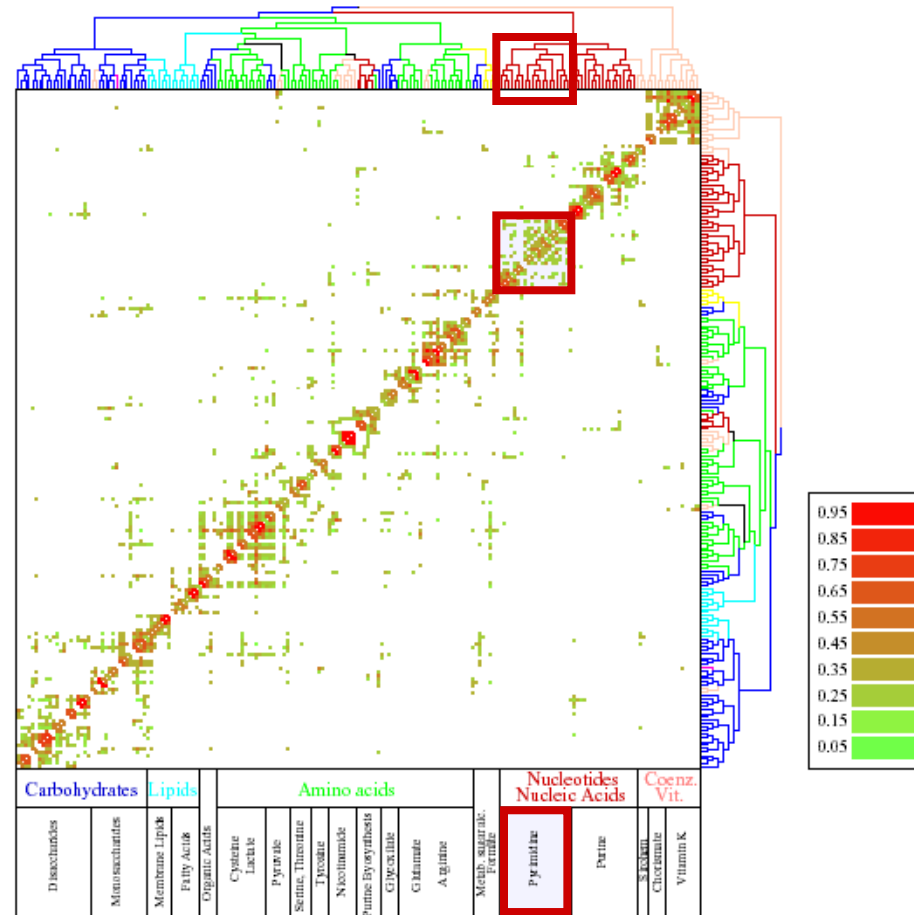
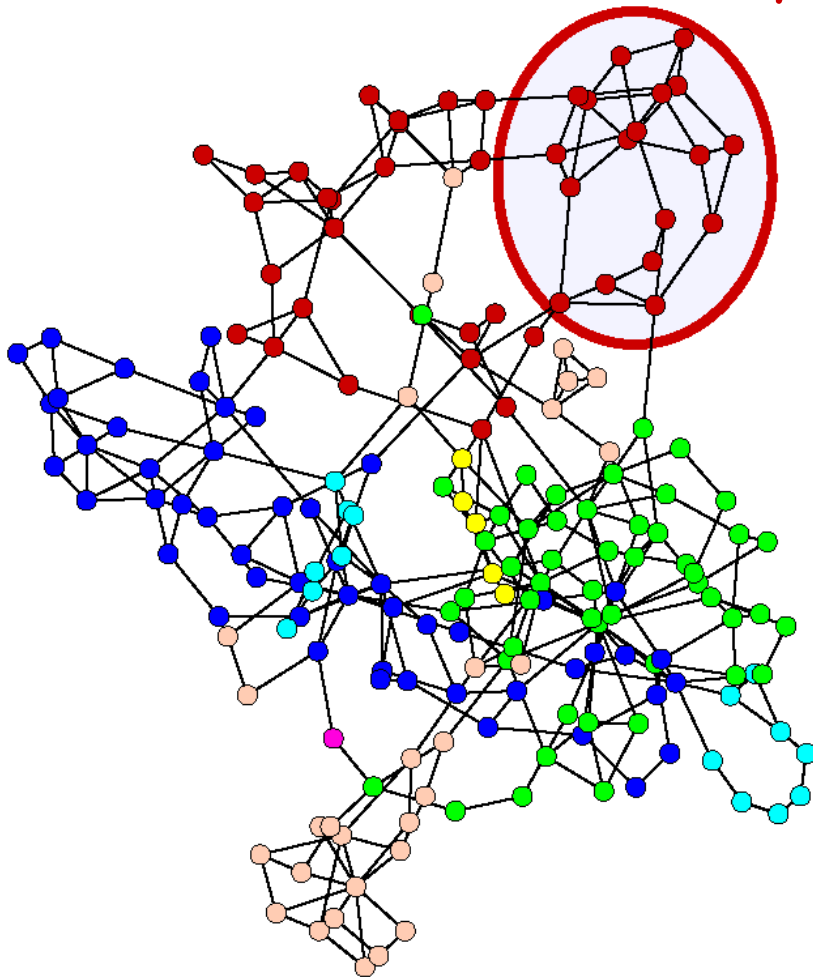


Examples of agglomerative hierarchical clustering

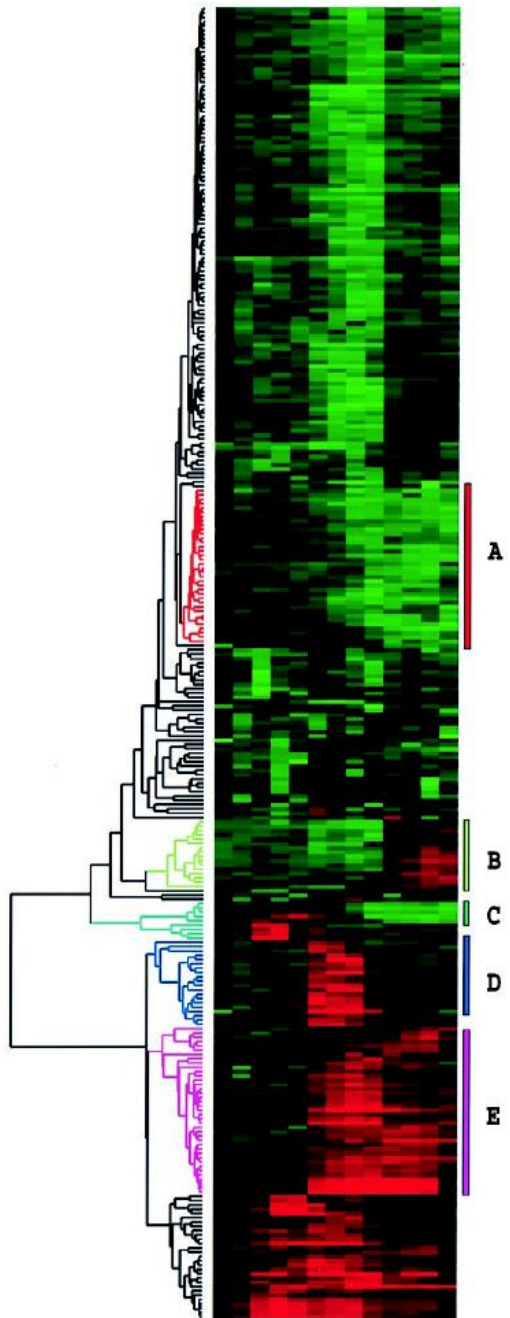
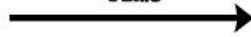
Modules in the *E. coli* metabolism

E. Ravasz et al., *Science*, 2002

Pyrimidine metabolism



Time



Spotted microarray for *Saccharomyces cerevisiae*

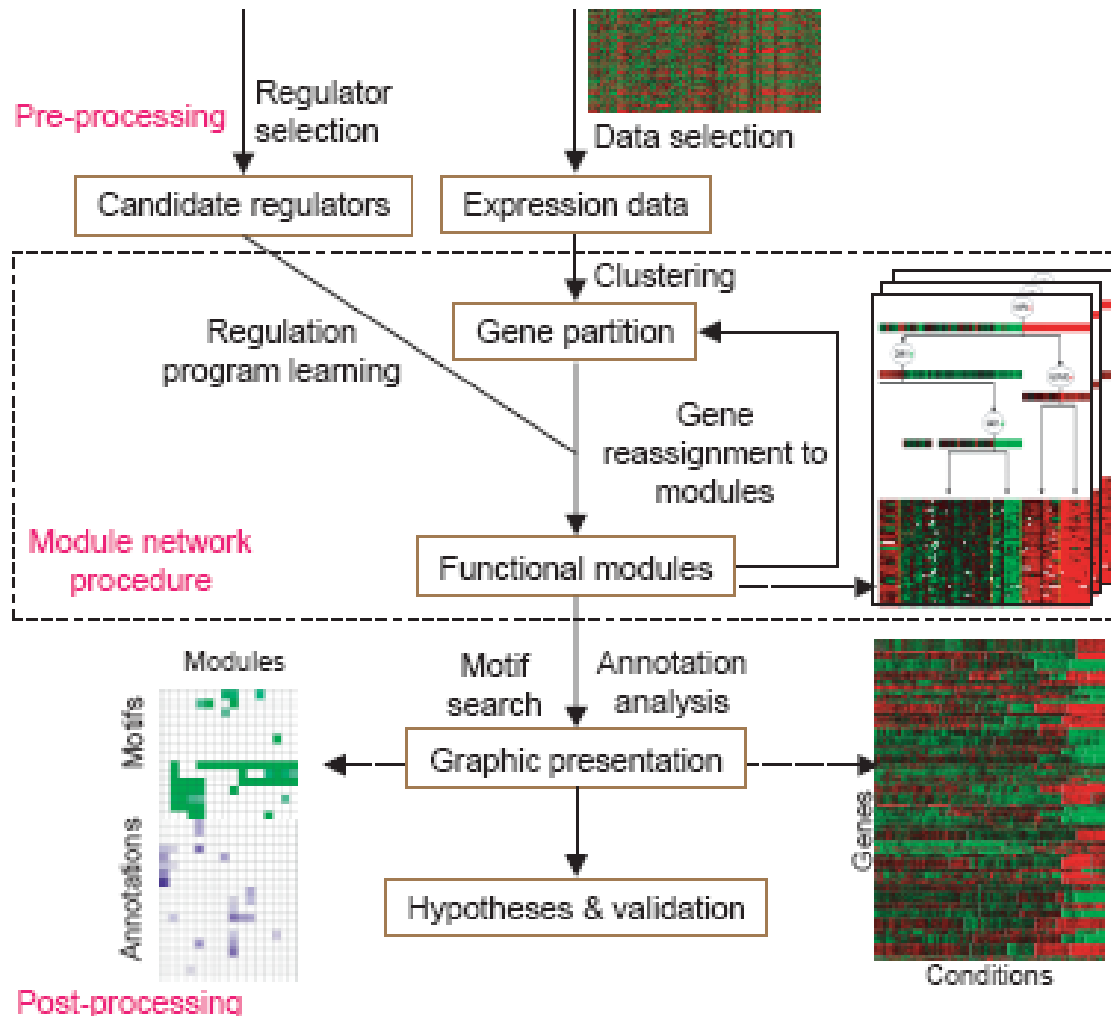
Similarity measure

$$S(X, Y) = \frac{1}{N} \sum_{i=1, N} \left(\frac{X_i - X_{offset}}{\Phi_X} \right) \left(\frac{Y_i - Y_{offset}}{\Phi_Y} \right)$$

where

$$\Phi_G = \sqrt{\sum_{i=1, N} \frac{(G_i - G_{offset})^2}{N}}$$

Regulatory module network



Method

High-resolution analysis of condition-specific regulatory modules in *Saccharomyces cerevisiae*

Hun-Goo Lee^{✉**†}, Hyo-Soo Lee^{✉*}, Sang-Hoon Jeon^{*}, Tae-Hoon Chung[‡],
Young-Sung Lim^{*} and Won-Ki Huh[†]

Addresses: ^{*}Department of Bioinformatics, Dong-a Seetech Research Institute, Seoul 135-010, Republic of Korea. [†]School of Biological Sciences and Research Center for Functional Cellulomics, Institute of Microbiology, Seoul National University, Seoul 151-747, Republic of Korea. [‡]Computational Biology Division, TGEN, N 5th St, Phoenix, Arizona 85004, USA.

✉ These authors contributed equally to this work.

Correspondence: Won-Ki Huh. Email: wkh@snu.ac.kr

Published: 3 January 2008

Genome Biology 2008, **9**:R2 (doi:10.1186/gb-2008-9-1-r2)

The electronic version of this article is the complete one and can be found online at <http://genomebiology.com/2008/9/1/R2>

Received: 18 July 2007

Revised: 15 October 2007

Accepted: 3 January 2008

© 2008 Lee et al.; licensee BioMed Central Ltd.

This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

We present an approach for identifying condition-specific regulatory modules by using separate units of gene expression profiles along with ChIP-chip and motif data from *Saccharomyces cerevisiae*. By investigating the unique and common features of the obtained condition-specific modules, we detected several important properties of transcriptional network reorganization. Our approach reveals the functionally distinct coregulated submodules embedded in a coexpressed gene module and provides an effective method for identifying various condition-specific regulatory events at high resolution.

Genome Biology, 9, R2, (2008).

Limitations of Hierarchical Clustering

- Once a decision is made to combine two clusters, it cannot be undone
- No objective function is directly minimized
- Different schemes have problems with one or more of the following:
 - Sensitivity to noise and outliers
 - Difficulty handling different sized clusters and convex shapes
 - Breaking large clusters

Simulated Annealing (SA)

- Simulated annealing is a powerful technique to provide high quality solutions to some difficult combinatorial problems
- It keeps a variable Temperature (T) which determines the behavior of the annealing process. This variable T is initialized to a very large value at the beginning, and will be gradually decreased (cooled down).

Simulated Annealing

Algorithm:

Initialize T and a feasible solution f

While ($T \geq$ a threshold)

- Make a slight modification to f to get g
- Check if g is better than f , i.e., $\text{cost}(g) \leq \text{cost}(f)$?
- If yes, accept g , i.e., $f \leftarrow g$; else, compute p as $e^{-k(\text{cost}(g)-\text{cost}(f))/T}$ where k is a positive constant, and then, accept g with probability p
- Update T

Basic Ingredients of Simulated Annealing

- Solution Space
- Neighboring Structure
- Cost Function
- Annealing Schedule
 - Moves are selected randomly, and the probability that a move is accepted is proportional to system's current temp

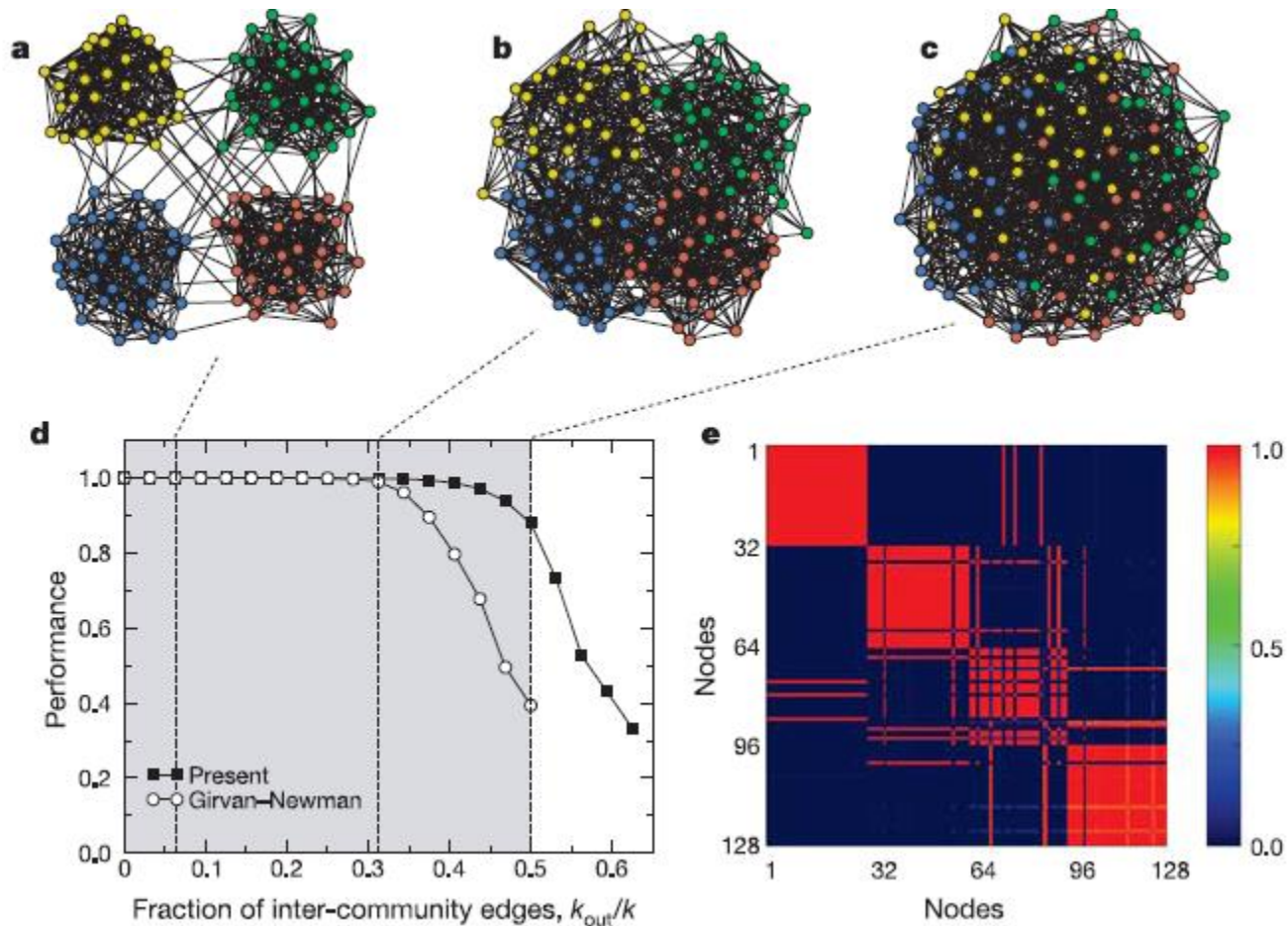
模拟退火算法是模拟固体的退火过程，对 **Metropolis** 算法进行迭代的组合优化算法。设组合优化问题的一个解 i 和目标函数 $f(i)$ 分别与固体的一个微观状态 i 和能量状态 E_i 等价，并用控制参数 t 担当固体退火过程中温度 T 的角色，则对于控制参数 t 的每一取值，算法持续进行“产生新解—判断—接受/舍弃”的迭代过程，控制参数 t 随算法进程递减其值，使得整个迭代过程与固体在某一恒定温度下趋于热平衡的过程相对应。

Functional cartography of complex metabolic networks

Roger Guimerà & Luís A. Nunes Amaral

NICO and Department of Chemical and Biological Engineering, Northwestern University, Evanston, Illinois 60208, USA

Guimerà and Amaral, Nature 2005.



利用模拟退火法（SA）进行模块划分

使网络模块性(**Modularity**)最大

$$M \equiv \sum_{s=1}^{N_M} \left[\frac{l_s}{L} - \left(\frac{d_s}{2L} \right)^2 \right]$$

其中

N_M 模块的数目，

L 是整个网络内所有链接数目的总和，

l_s 是模块 s 内部节点间链接的数目，

d_s 是模块 s 内部节点的连接度的总和。

- 能量函数 $C = -M$
- 接受准则 Metropolis

$$p = \begin{cases} 1 & \text{if } C_f \leq C_i \\ \exp\left(-\frac{C_f - C_i}{T}\right) & \text{if } C_f > C_i \end{cases}$$

其中 C_f 是更新后的能量函数， C_i 是更新前的能量函数

在每个温度 T ，通过两类随机移动改变网络模块结构：

- 1) 某一个节点从一个模块随机移动到另一个模块，做 n_i 次，定义 $n_i=f\omega^2$ ；
- 2) 随机合并两个模块或者将某一个模块随机地划分成为两个模块，做 n_c 次，定义 $n_c=f\omega$ 。

其中， ω 为整个网络的节点数目； f 为迭代因子；温度 T 按照冷却因子 Δ 进行改变， $T'=\Delta T$ ，一般 $\Delta\in[0.990, 0.999]$ 。当温度连续改变25次模块性 M 都不变时，收敛，得到的收敛状态即为最终的模块化结果

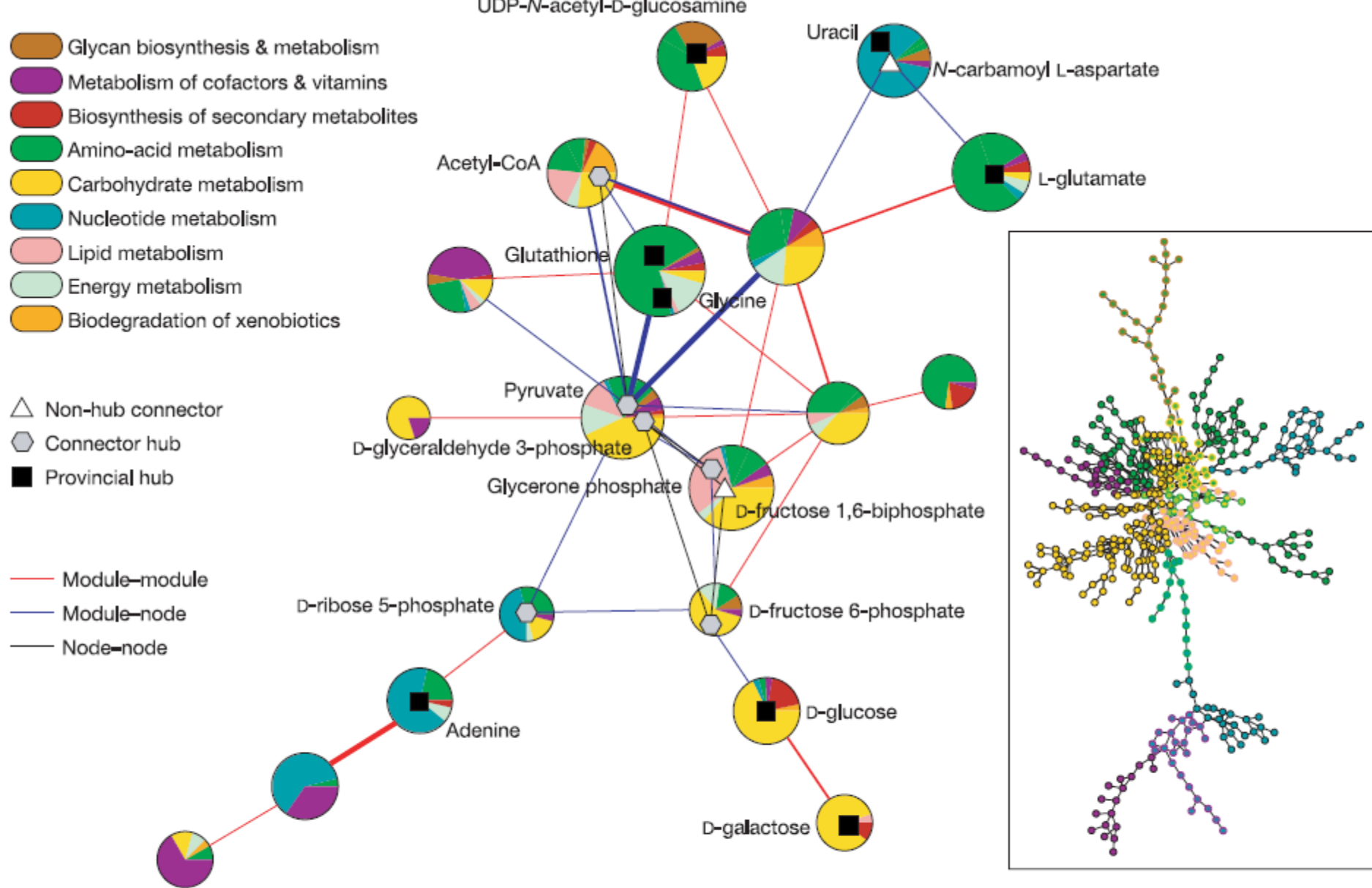


Figure 3 Cartographic representation of the metabolic network of *E. coli*. Each circle represents a module and is coloured according to the KEGG pathway classification of the metabolites it contains. Certain important nodes are depicted as triangles (non-hub connectors), hexagons (connector hubs) and squares (provincial hubs). Interactions between modules and nodes are depicted using lines, with thickness proportional to the

number of actual links. Inset: metabolic network of *E. coli*, which contains 473 metabolites and 574 links. This representation was obtained using the program Pajek. Each node is coloured according to the 'main' colour of its module, as obtained from the cartographic representation.

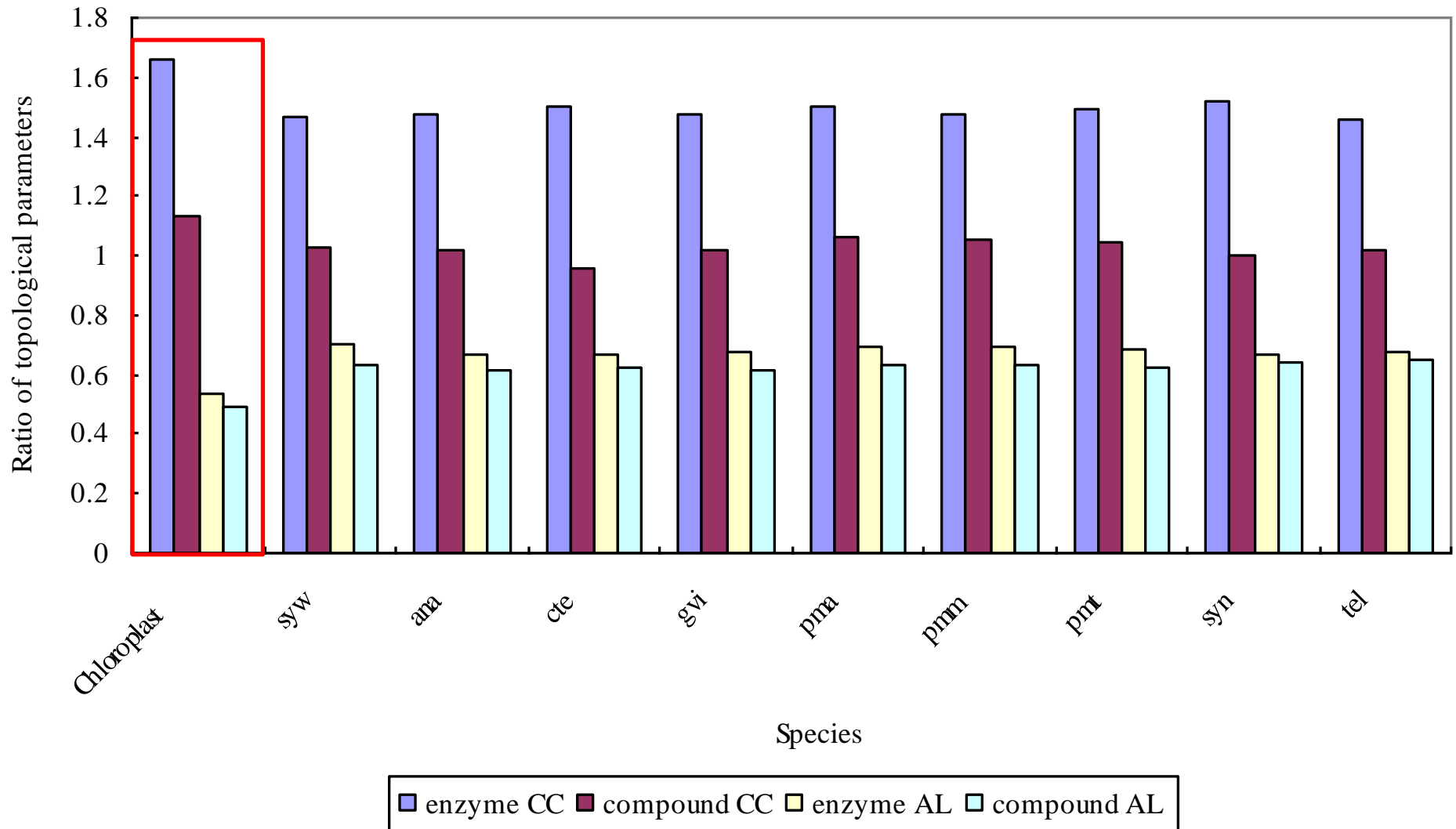
Case study

Comparative analysis of metabolic networks between chloroplast and cyanobacteria.

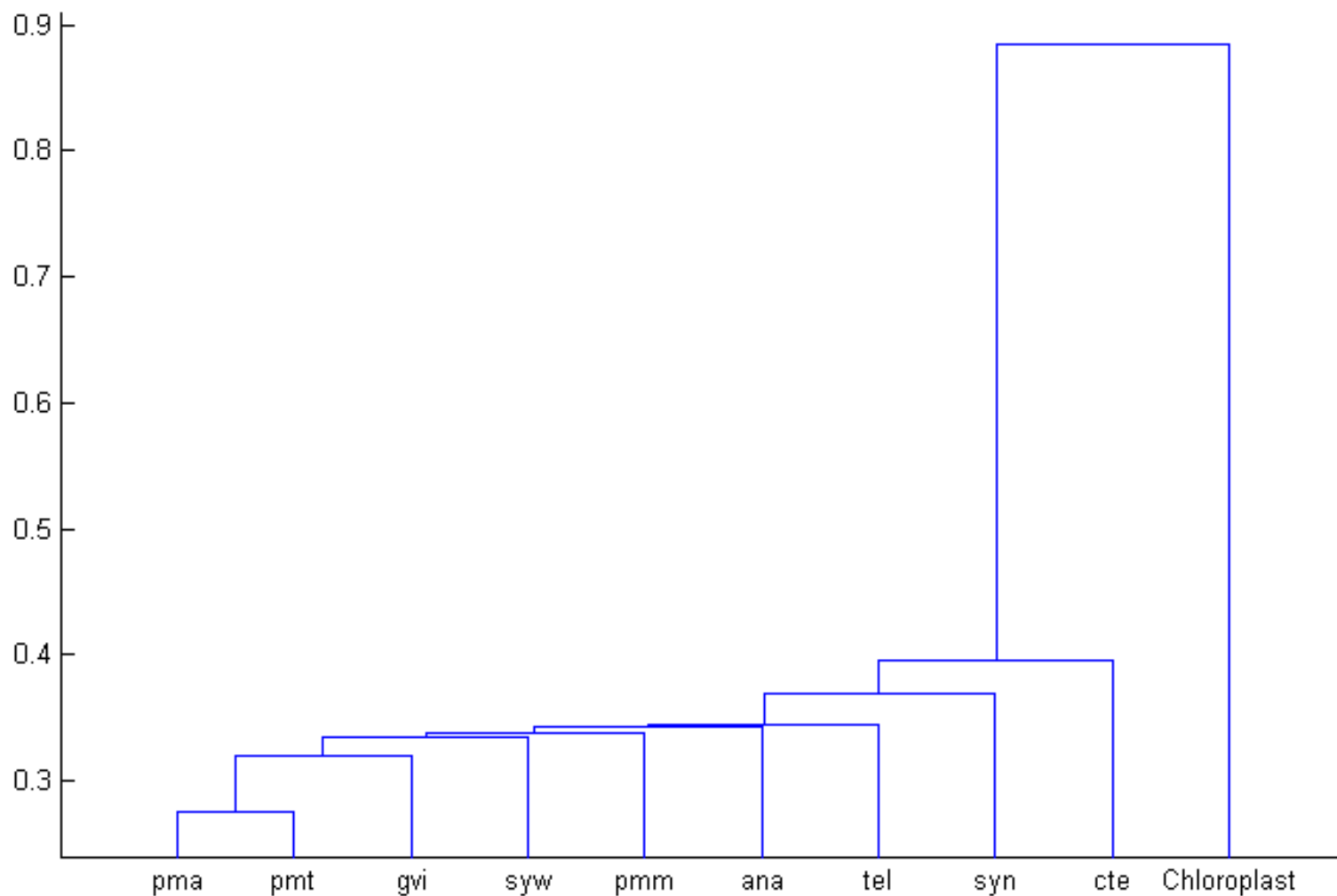
Global Topology of metabolic networks of chloroplast and different cyanobacterias

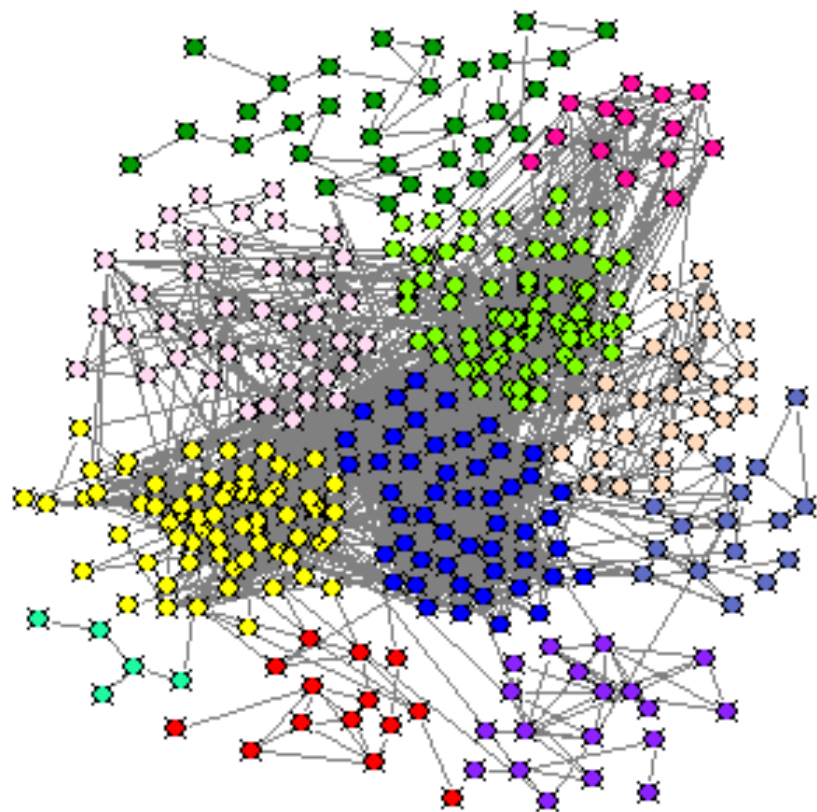
	enzyme cluster coefficient	compound cluster coefficient	average enzyme path length	enzyme diameter	average compound path length	compound diameter
Chloroplast	0.534371	0.431872	5.07847	19	4.83902	19
syw	0.59365	0.503954	4.07523	11	3.972854	12
ana	0.590467	0.513945	4.15901	11	3.95608	12
cte	0.577056	0.506881	4.12231	12	3.94473	12
gvi	0.594211	0.518726	4.15974	12	3.95251	12
pma	0.577878	0.487342	4.09658	12	3.92037	12
pmm	0.590459	0.48967	4.06937	10	3.92196	11
pmt	0.581159	0.495484	4.09455	12	3.98362	12
syn	0.590339	0.501971	4.1349	12	3.91225	12
tel	0.593009	0.488283	4.11994	11	3.87589	12

Ratio of topological parameters between calvin cycle –centered subnetwork and whole network

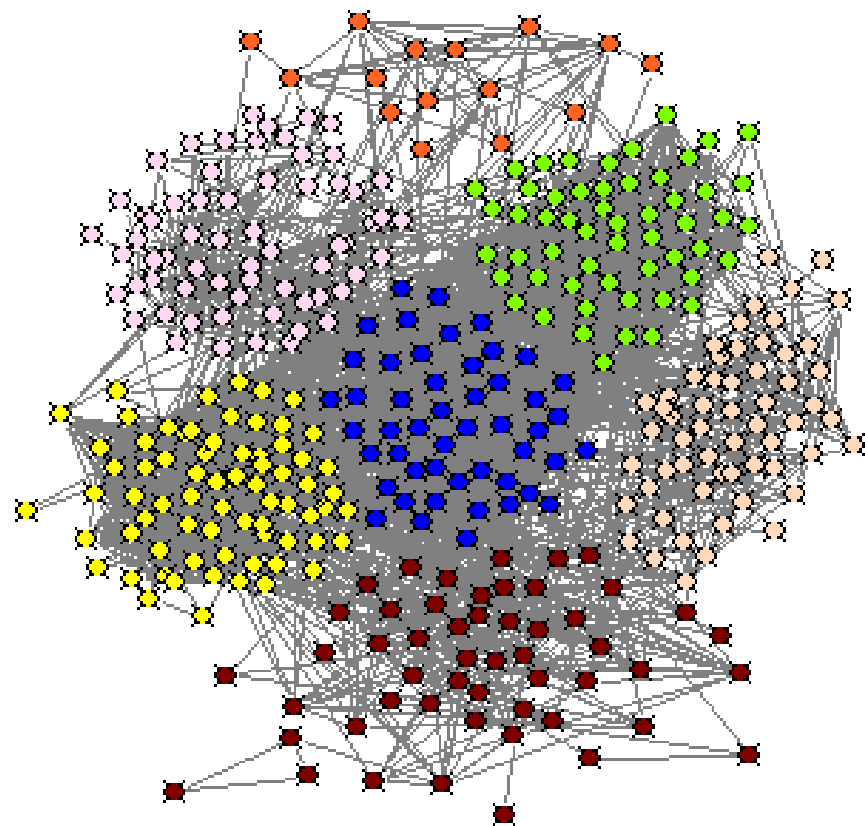


叶绿体和各种蓝藻模块化结构的相似性





叶绿体



蓝藻syw

Network Motifs

We have looked at some global features of real complex networks:

- Short distance between nodes (small-world)
- Low number of hubs (scale-free)
- High local clustering (modular)

Network Motifs

Now we look at “patterns” in complex networks

Network motif = small subgraphs that are significantly over-represented

Example of a 3-node motif: 

Do you expect this motif to be over-represented?

First focus on directed networks and look at 3- and 4-node motifs

What is a 2-node motif?

How many 3-node motifs are there?

How many 4-node motifs are there?

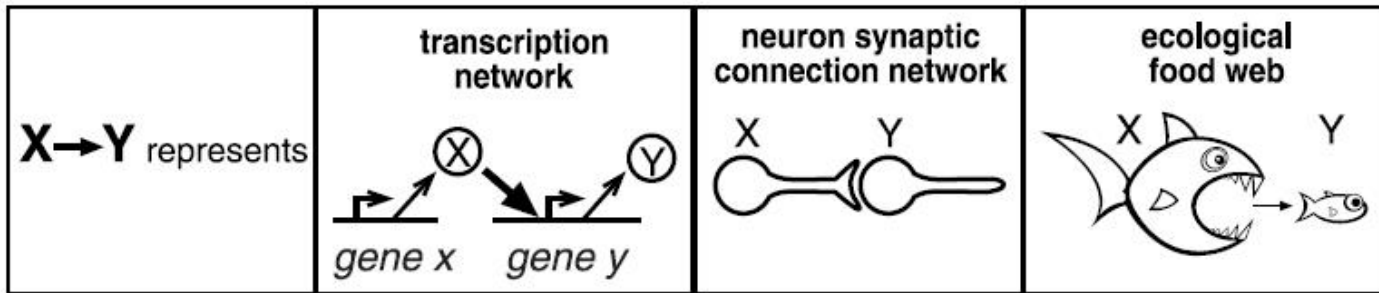
Beware of overcounting due to isomorphisms(同构)!

Detection of important network motifs

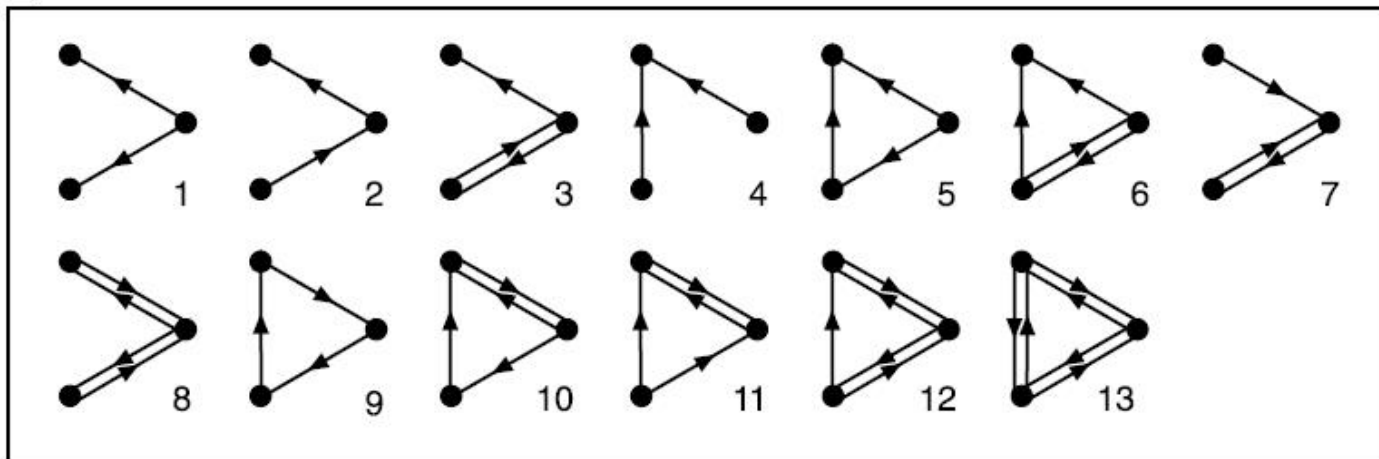
- Technique:
 - construct many random graphs with the same number of nodes and degree distribution
 - count the number of motifs in those graphs
 - calculate the Z score: the probability that the same or larger number of motifs in the real world network could have occurred in a random one
- Software available: mfinder
 - <http://www.weizmann.ac.il/mcb/UriAlon/>

Enumeration of directed 3-node motifs

A



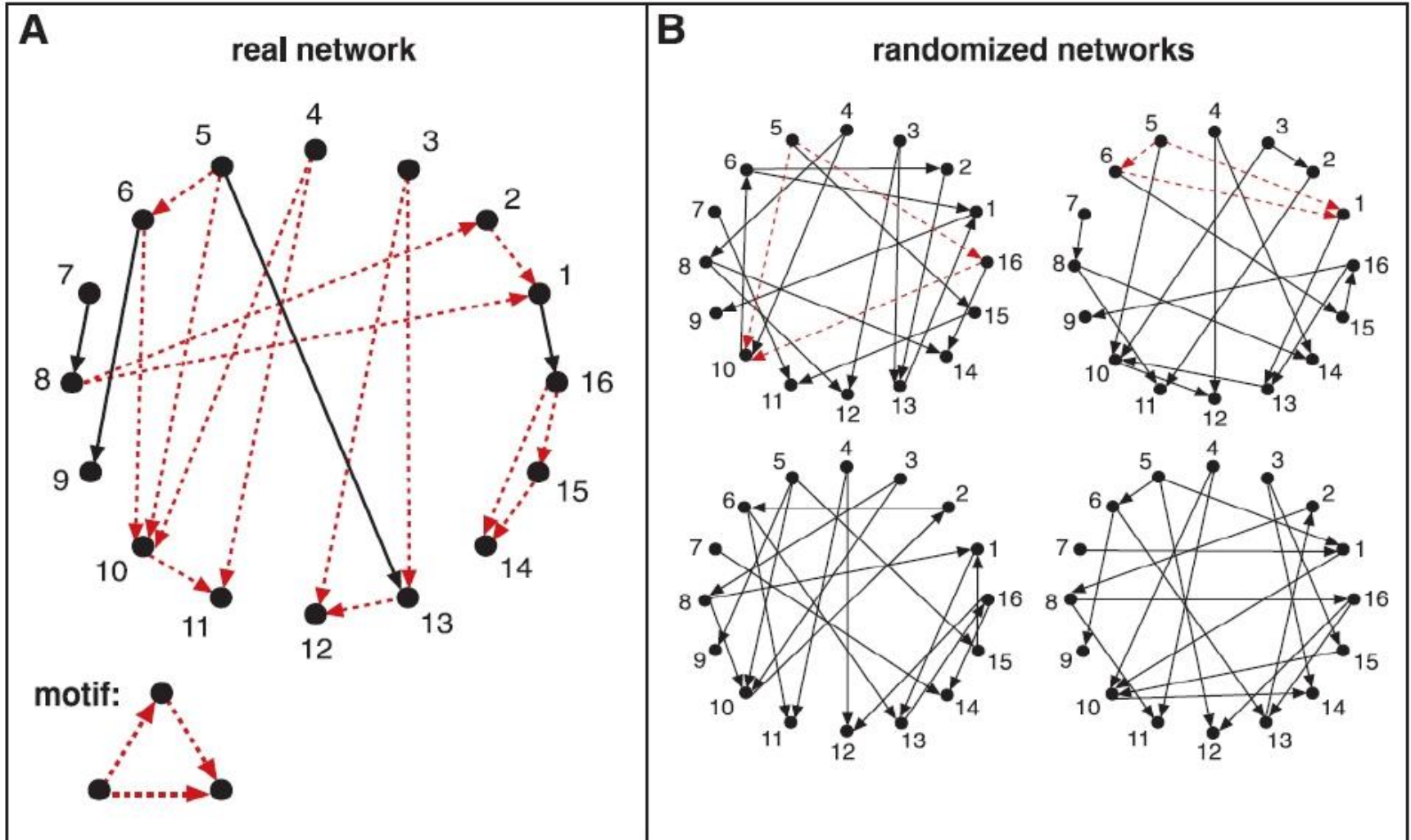
B



Again, interpretations (what those formalisms actually mean)!
Does $X \rightleftarrows Y$ make sense in the food web context?

Exercise: How many undirected 3-node motifs are there?

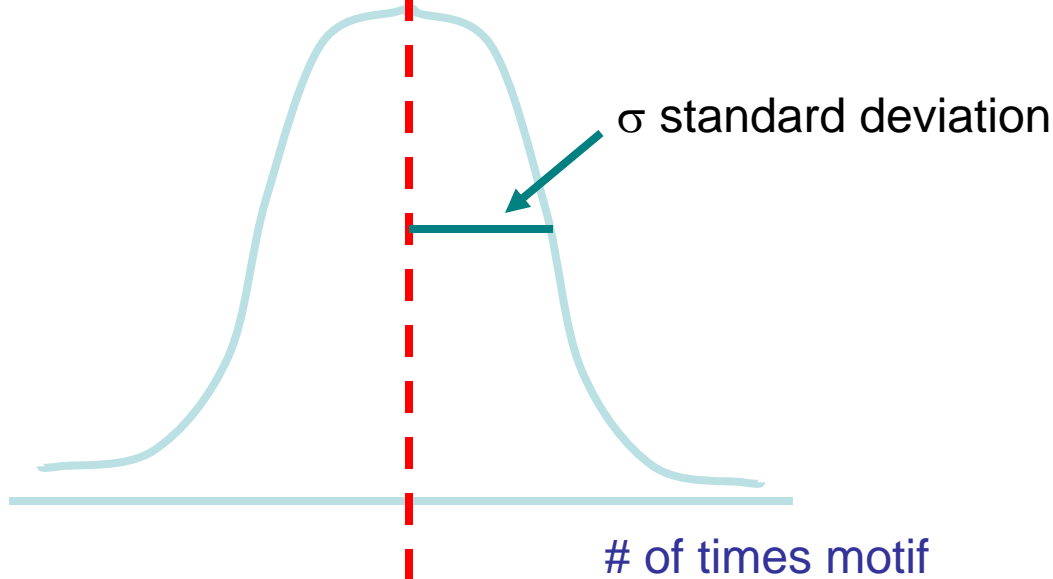
Example: Feed-forward loop



- › Count how many times it appears in the real network
- › Count how many times it appears in “comparable” random networks (through edge-swapping)
- › Compute empirical p-value or z-score.

What the Z score means

μ = mean number of times the motif appeared in the random graph



In the context of motifs:
 $Z > 0$, motif occurs more often than for random graphs
 $Z < 0$, motif occurs less often than in random graphs

$$Z_x = \frac{X - \mu_x}{\sigma_x}$$

$|Z| > 1.65$, only a 5% chance of random occurrence

Examples of network motifs (3 nodes)

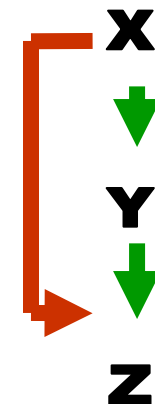
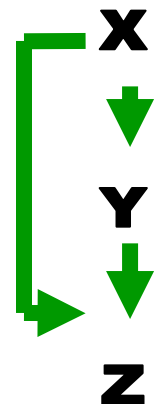
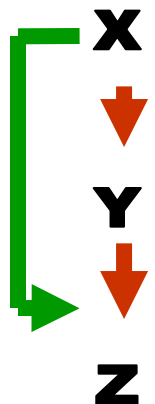
- Feed forward loop
 - Found in many transcriptional regulatory networks



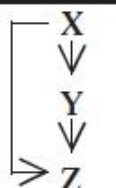
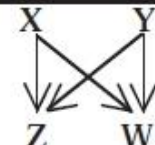
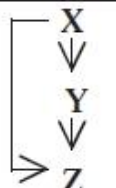
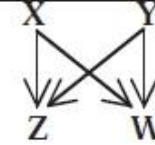
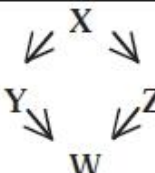

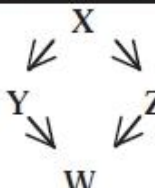
Structure	Appearances in real network	Appearances in randomized network (mean \pm s.d.)	<i>P</i> value
Coherent feedforward loop	34	4.4 \pm 3	<i>P</i> < 0.001
Incoherent feedforward loop	6	2.5 \pm 2	<i>P</i> ~ 0.03

coherent

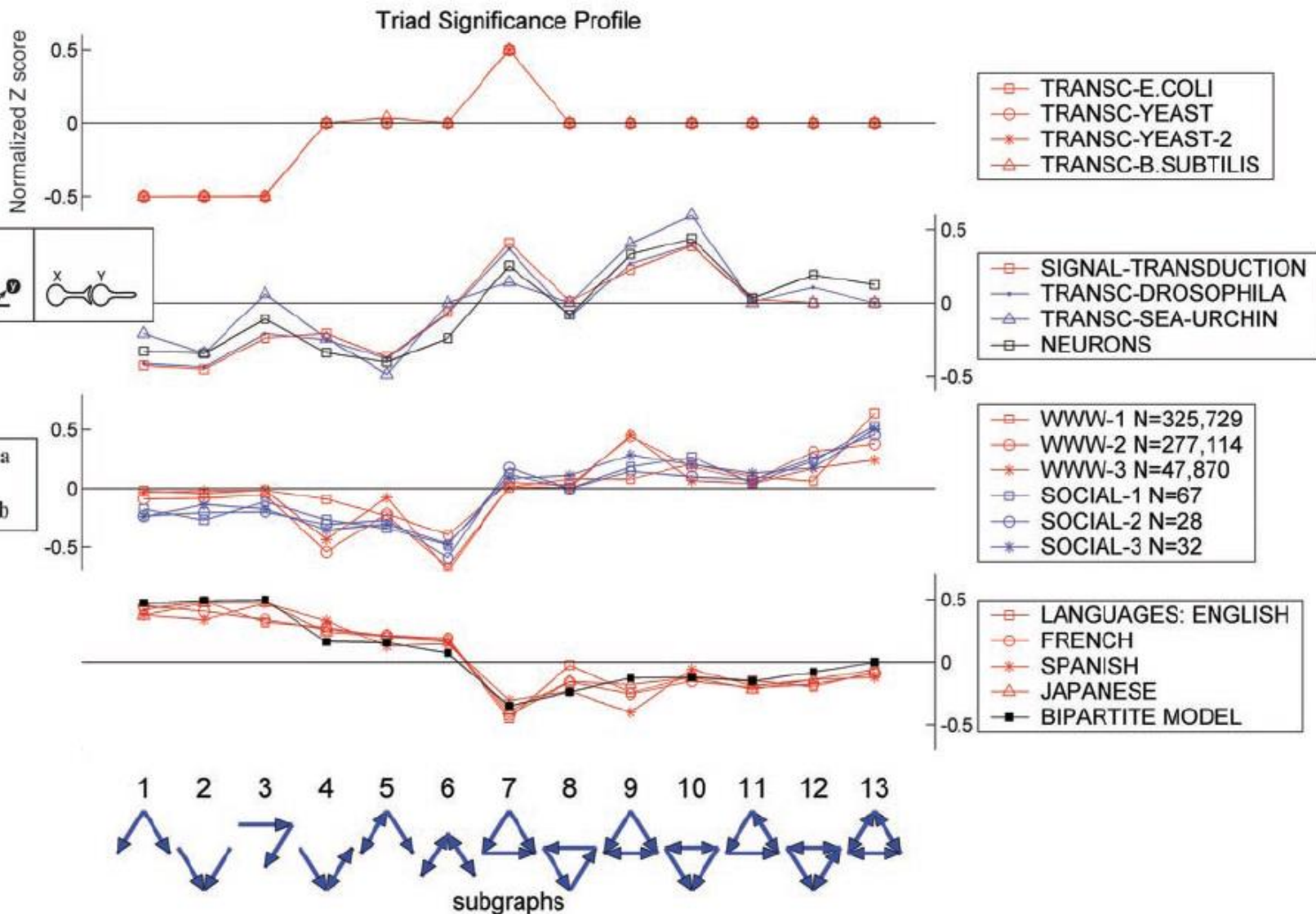
incoherent



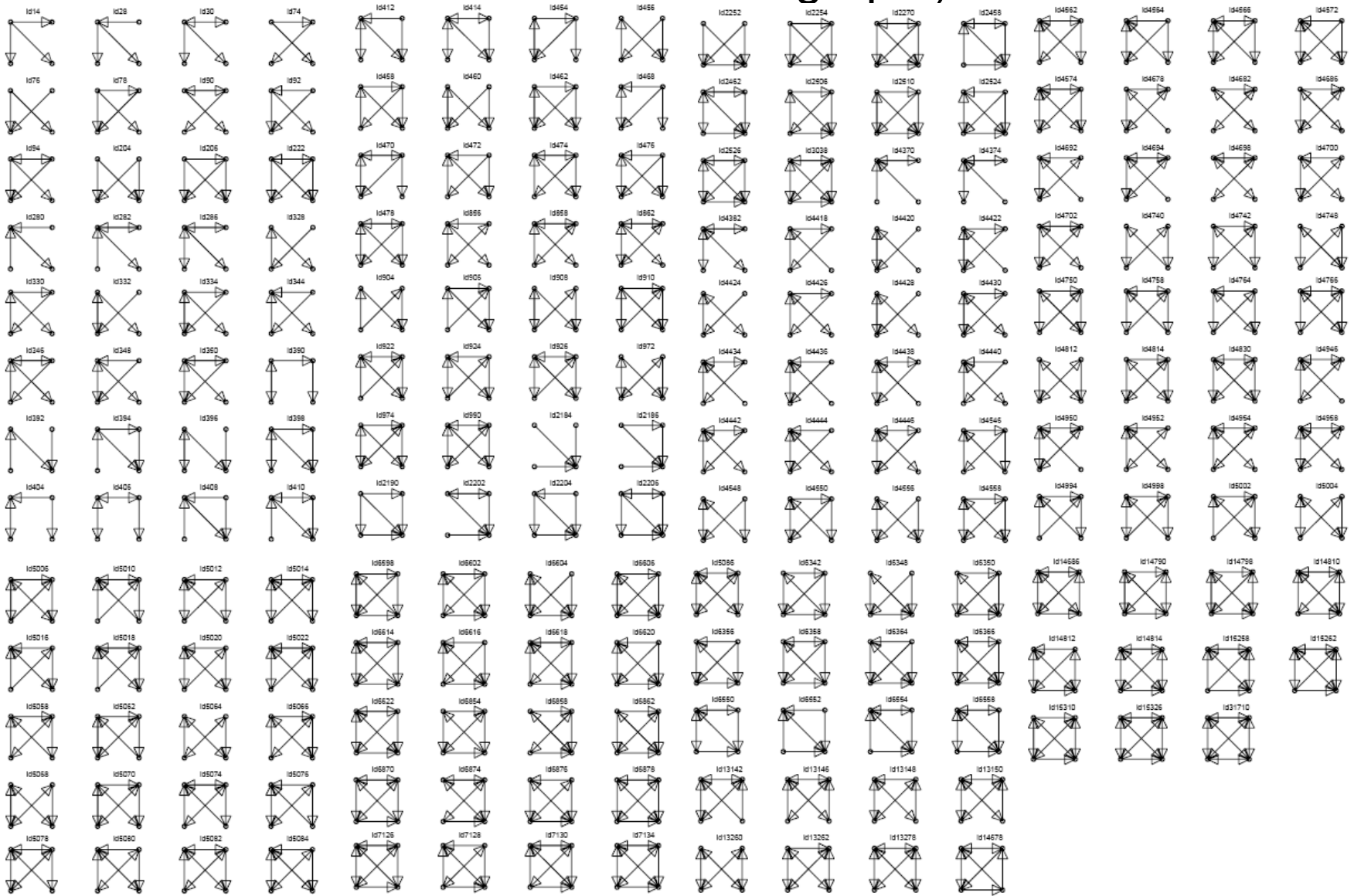
Different classes of networks prefer different network motifs

Network	Nodes	Edges	N_{real}	$N_{\text{rand}} \pm \text{SD}$	Z score	N_{real}	$N_{\text{rand}} \pm \text{SD}$	Z score	N_{real}	$N_{\text{rand}} \pm \text{SD}$	Z score
Gene regulation (transcription)				Feed-forward loop			Bi-fan				
<i>E. coli</i>	424	519	40	7 ± 3	10	203	47 ± 12	13			
<i>S. cerevisiae</i> *	685	1,052	70	11 ± 4	14	1812	300 ± 40	41			
Neurons				Feed-forward loop			Bi-fan			Bi-parallel	
<i>C. elegans</i> †	252	509	125	90 ± 10	3.7	127	55 ± 13	5.3	227	35 ± 10	20
Food webs				Three chain			Bi-parallel				
Little Rock	92	984	3219	3120 ± 50	2.1	7295	2220 ± 210	25			
Ythan	83	391	1182	1020 ± 20	7.2	1357	230 ± 50	23			
St. Martin	42	205	469	450 ± 10	NS	382	130 ± 20	12			
Chesapeake	31	67	80	82 ± 4	NS	26	5 ± 2	8			
Coachella	29	243	279	235 ± 12	3.6	181	80 ± 20	5			
Skipwith	25	189	184	150 ± 7	5.5	397	80 ± 25	13			
B. Brook	25	104	181	130 ± 7	7.4	267	30 ± 7	32			

Finding classes on graphs based on their motif “profiles”



All 4 node subgraphs (computational expense increases with the size of the graph!)



Thanks !

