Network Science

Motifs

Albert-László Barabási
with
Emma K. Towlson, Sebastian Ruf,
Michael Danziger, and Louis Shekhtman

www.BarabasiLab.com

Network Motifs

Inspired by:

Jànos Kertész

Luay Nakhleh

Alain Karma/Marc Santollini

My adventures in brain networks

Structure and function

How do complex systems function?
Complex networks should reflect the function of the systems they represent.
How is the topology related to the function?

Important task: identifying units which are topologically closely related - they are expected to have a functional role.

Scales:
Microscopic
Mesoscopic
Macroscopic

Structure and function

Structures:
Microscopic
Mesoscopic
Macroscopic

Microscopic: node properties + interactions (dyad)
Very system specific

Macroscopic: the network as a whole. Global characterisation, qualitative universality, robustness etc.

Mesoscopic: structures on intermediate scales.
System specific + universal features

Structure and function

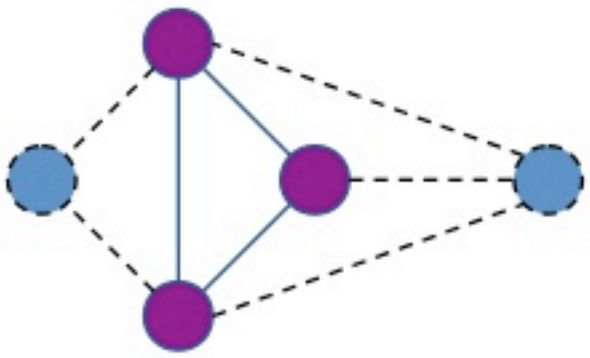
Mesoscopic structures: subgraphs of the original (usually large) graph.

For a graph: $G = \{V, E\}$

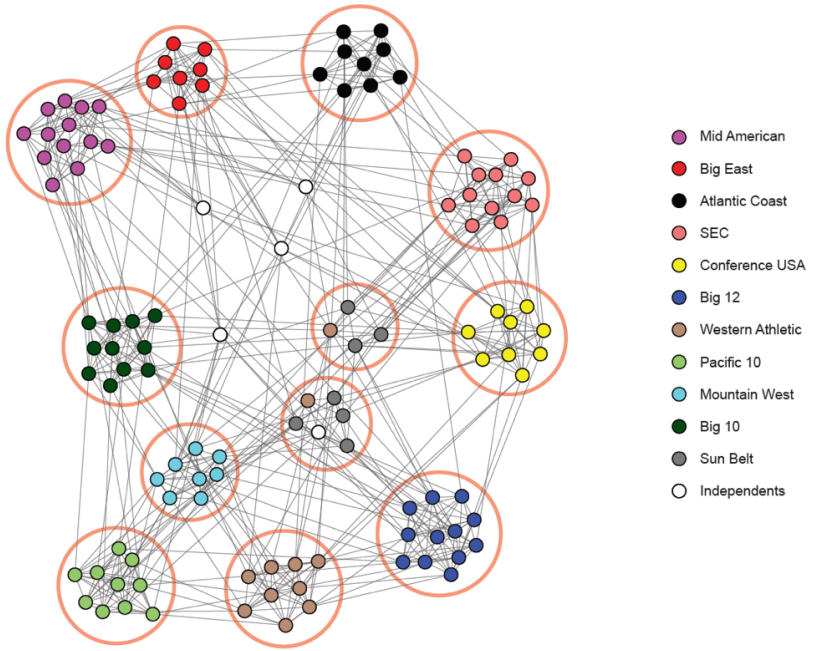
Define a subgraph: $G' = \{V', E'\}$ with $V' \subseteq V; E' \subset E$

such that $i, j \in V' \; \forall e_{ij} \in E'$

Assume: $N' \ll N$



Structure and function



Two approaches:

i) Consider the particular subgraphs after identifying them: egocentric networks, communities...

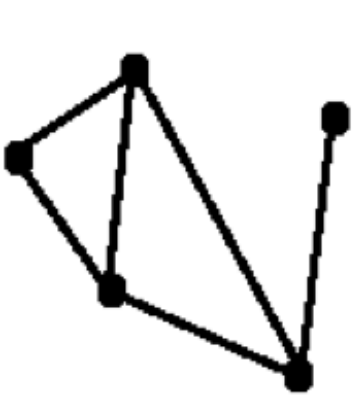
ii) Define a type of subgraph, identify topologically equivalent occurrences and check how significant this class is: motifs.

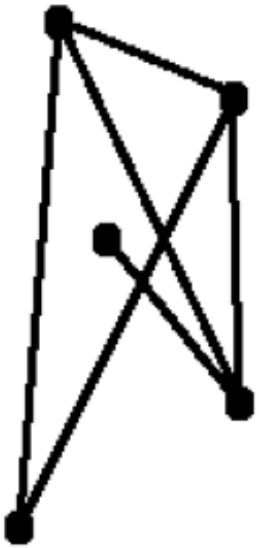
Idea: If a type of subgraph (e.g. a triangle) occurs significantly often in a large network, we can expect that such subgraphs have an important role in its function.

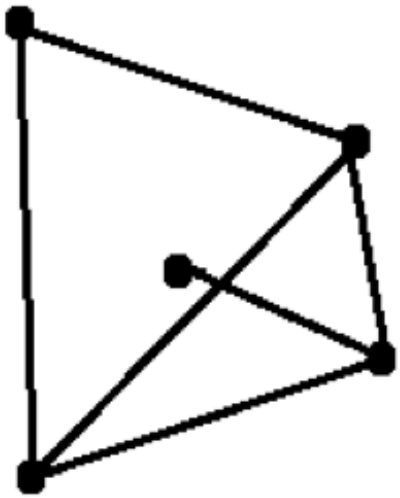
Motifs

Two graphs are topologically equivalent (or isomorphic) if there is an appropriate numbering of nodes such that the adjacency matrices become identical.

Condition that number of nodes and their degrees are the same is necessary but not sufficient.







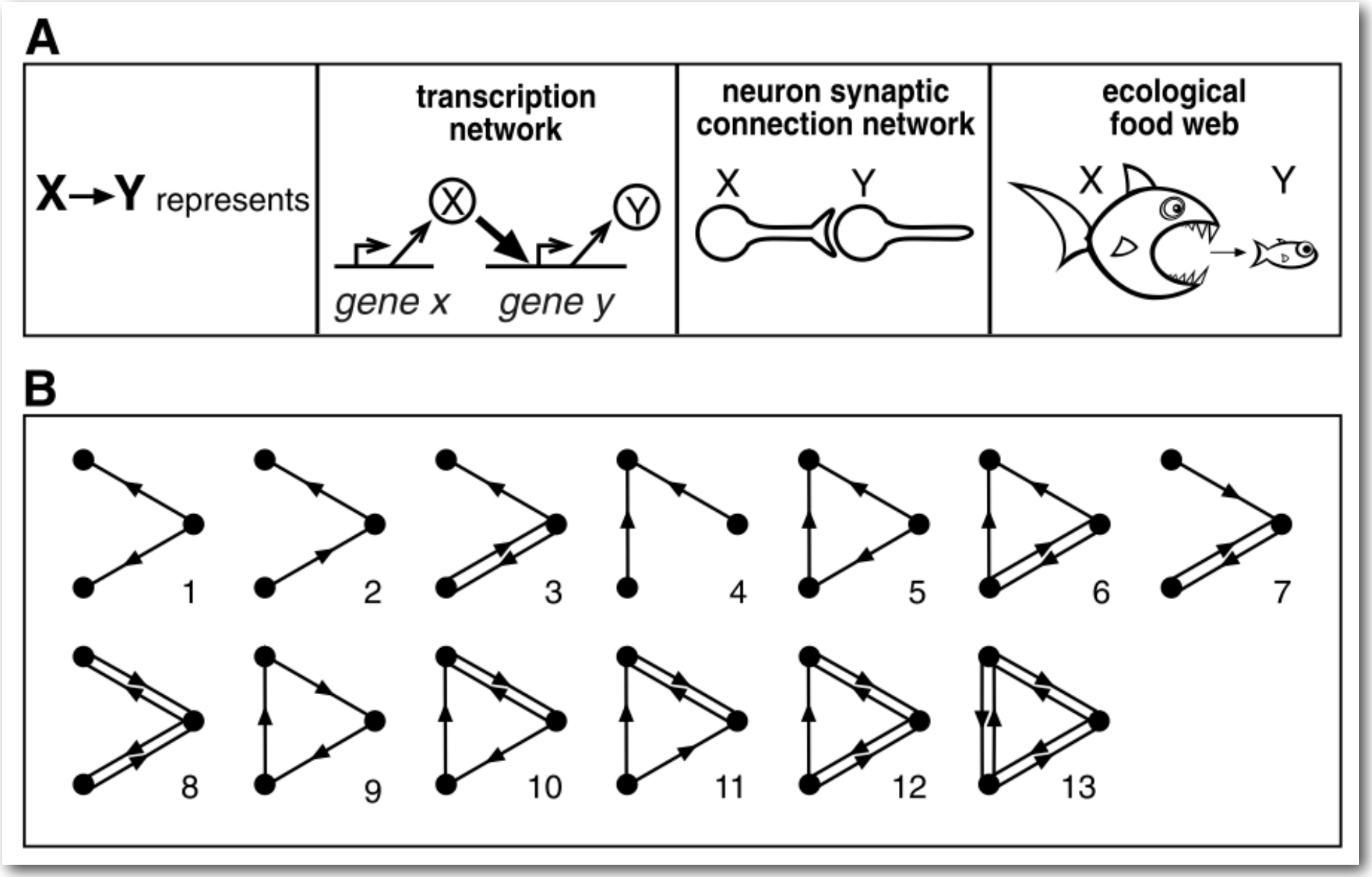
Motifs

Motif: set of topologically equivalent subgraphs in a network

Cardinality of a set: number of elements in the set

If the cardinality of a motif is significantly high, we expect that the motif has an important role in the function of the complex system the network is mapped from.

Motifs



Milo et al, Science 2002

Motifs

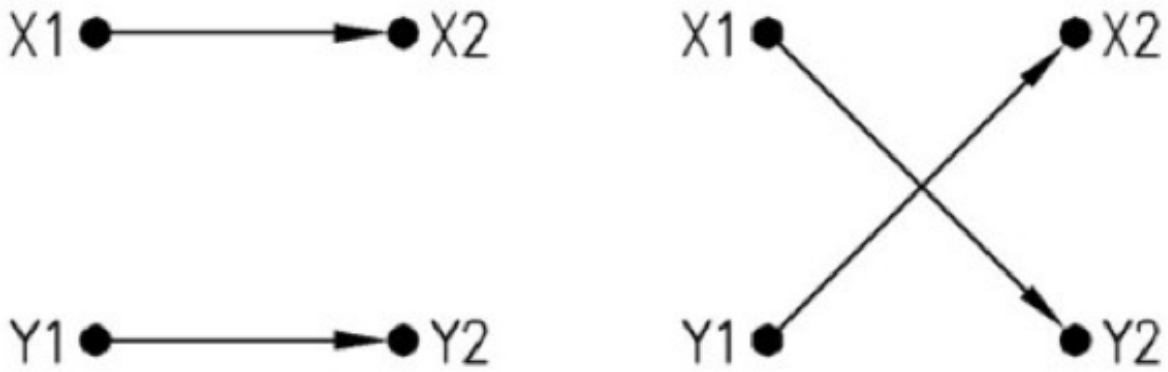
What do we mean by "significantly high"?
We must compare to a reference system.

"Significantly high" means that we have a null hypothesis that a given cardinality of a motif stems from a reference system. If we can exclude this hypothesis then there is an additional origin for the effect - possibly the function of the system.

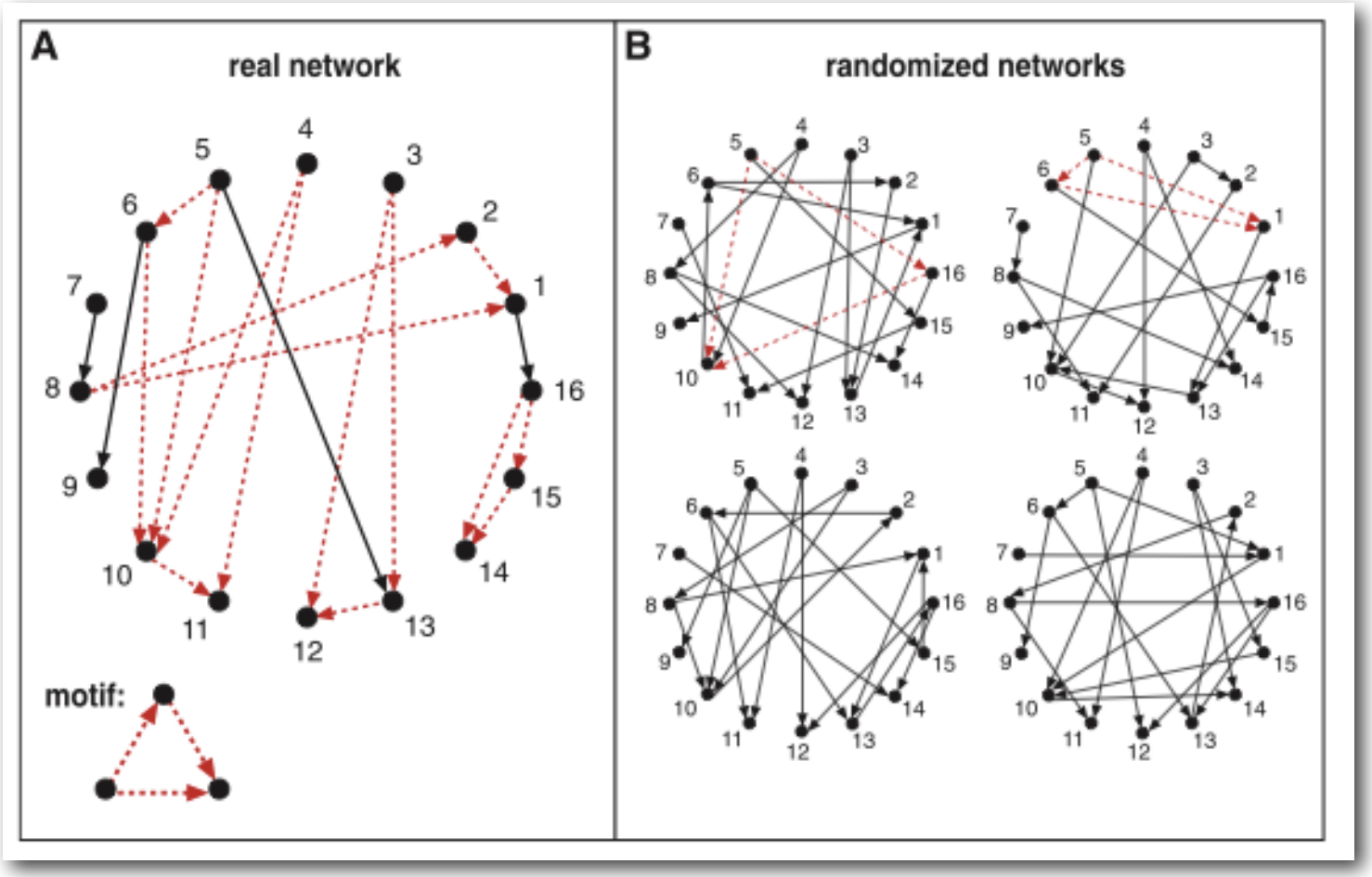
Motifs

Usually we take a random network as a reference system assuming that correlations are caused by the function.

Simplest: ER(N,L). Too simple and far from the global, universal observations of many real networks (broad degree distribution etc). Thus the common reference system is the configuration model. The configuration model can be considered as a result of a degree-preserving randomisation process: link swapping.



Motifs



Milo et al, Science 2002

Motifs

Now we need a measure for the importance of a motif.

The null model is an ensemble, thus with a mean and a variance. We compare the empirical cardinality to the mean cardinality of the ensemble, and judge the significance of the deviations by comparing them to the standard deviation.

z-score:

$$z_m = \frac{N_m(emp) - N_m(rnd)}{\sigma_m}$$

Motifs

Alternative to the z-score: p-value.

As before, create an ensemble of M networks by performing link swaps. For each motif i, count the number of networks which contain more occurrences of the motif than the empirical network.

$$p = \frac{1}{M} \sum_{i=1}^{M} \theta(N_i - N(emp))$$

$$\theta(x) = \begin{cases} 1 \text{ if } x > 0 \\ 0 \text{ otherwise} \end{cases}$$

Motifs

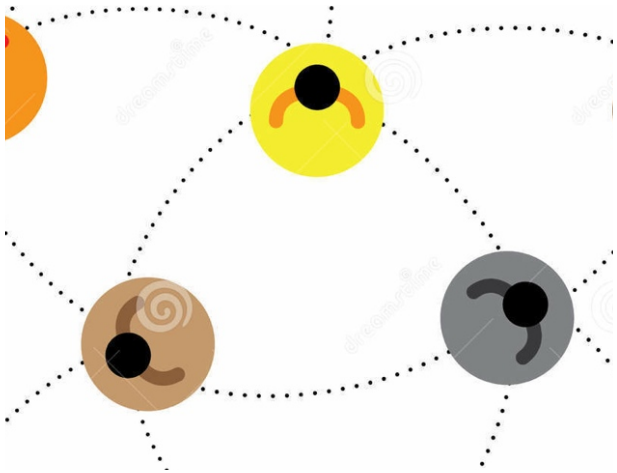
Important to consider the choice of reference system carefully.

E.g. in a biological system, one may conclude using this methodology that significant overrepresentation of a motif reflects an evolutionary advantage. BUT: genetic networks are embedded in space and neighbouring nodes interact more easily than further away nodes. This aspect is entirely ignored if the reference system is the configuration model. Random geometric graph could be a better choice.

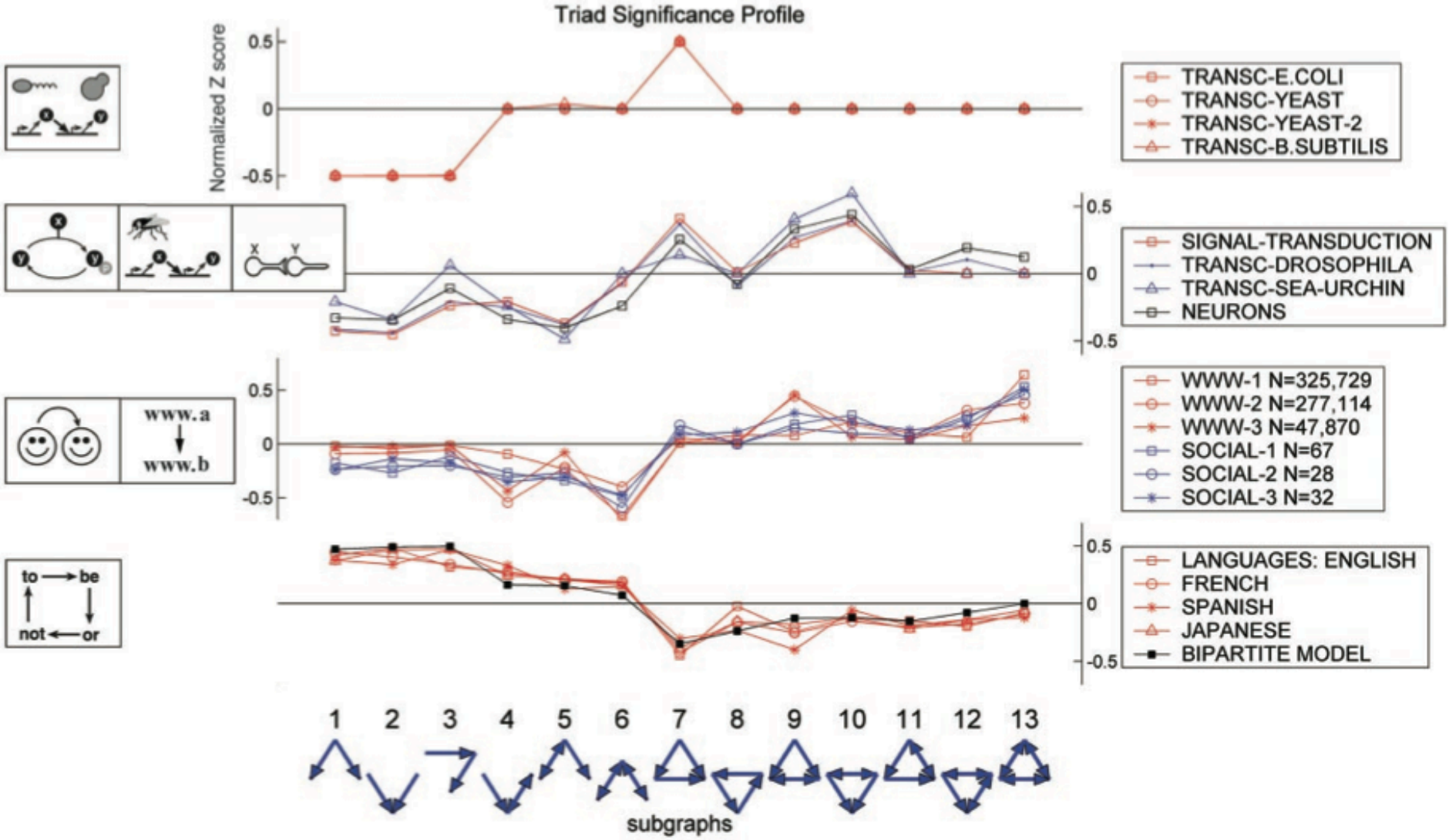
Motifs

Usually directed networks are considered and N≈3-5.

This motif approach has been extensively applied in biology, but not so much in social sciences.
The observation that social networks have high clustering is about a specific motif.



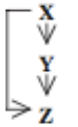
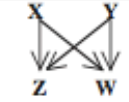
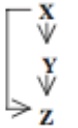
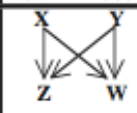
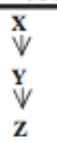
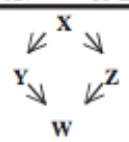
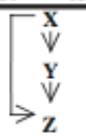
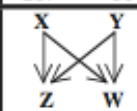
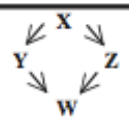
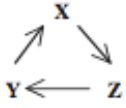
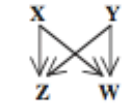
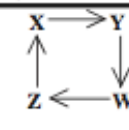
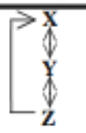
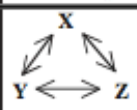
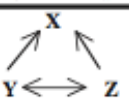
Motifs



Milo et al, Science 2004

Motifs

Different networks have different motif profiles

Network	Nodes	Edges	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score
Gene regulation (transcription)			 **Feed-forward loop**			 **Bi-fan**					
E. coli	424	519	40	7 ± 3	10	203	47 ± 12	13			
*S. cerevisiae**	685	1,052	70	11 ± 4	14	1812	300 ± 40	41			
Neurons			 **Feed-forward loop**			 **Bi-fan**			 **Bi-parallel**		
C. elegans†	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			
Electronic circuits (forward logic chips)			 **Feed-forward loop**			 **Bi-fan**			 **Bi-parallel**		
s15850	10,383	14,240	424	2 ± 2	285	1040	1 ± 1	1200	480	2 ± 1	335
s38584	20,717	34,204	413	10 ± 3	120	1739	6 ± 2	800	711	9 ± 2	320
s38417	23,843	33,661	612	3 ± 2	400	2404	1 ± 1	2550	531	2 ± 2	340
s9234	5,844	8,197	211	2 ± 1	140	754	1 ± 1	1050	209	1 ± 1	200
s13207	8,651	11,831	403	2 ± 1	225	4445	1 ± 1	4950	264	2 ± 1	200
Electronic circuits (digital fractional multipliers)			 **Three-node feedback loop**			 **Bi-fan**			 **Four-node feedback loop**		
s208	122	189	10	1 ± 1	9	4	1 ± 1	3.8	5	1 ± 1	5
s420	252	399	20	1 ± 1	18	10	1 ± 1	10	11	1 ± 1	11
s838‡	512	819	40	1 ± 1	38	22	1 ± 1	20	23	1 ± 1	25
World Wide Web			 **Feedback with two mutual dyads**			 **Fully connected triad**			 **Uplinked mutual dyad**		
nd.edu§	325,729	1.46e6	1.1e5	2e3 ± 1e2	800	6.8e6	5e4±4e2	15,000	1.2e6	1e4 ± 2e2	5000

Milo et al, Science 2002

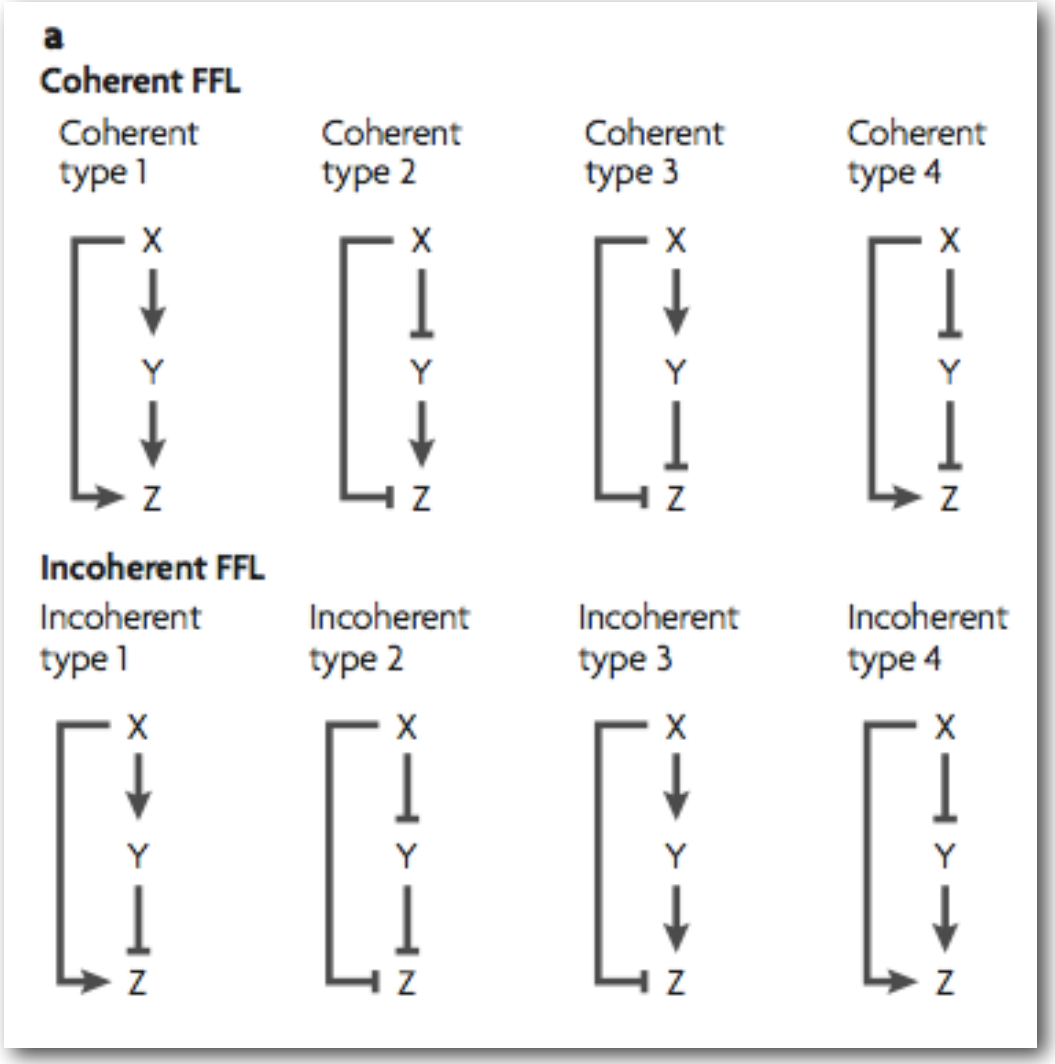
Motifs in transcription regulation networks

1) Feed-forward loop (FFL)

Coherent: if the direct effect of X on Z has the same indirect effect of X on Z through Y

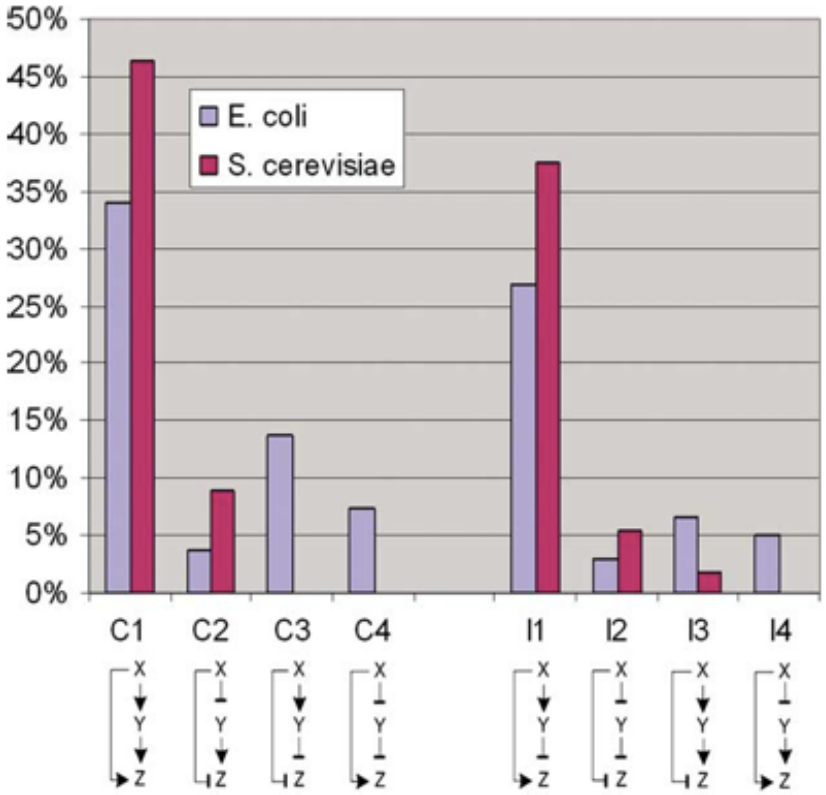
Incoherent: otherwise

8 types of FFL.

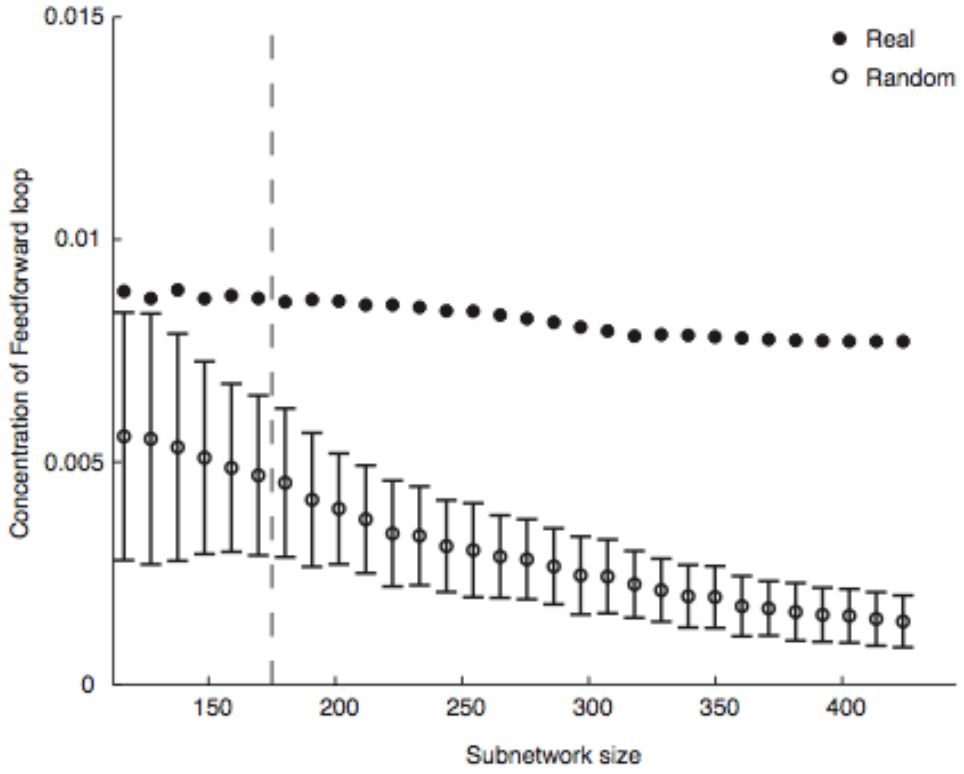


Alon, Nat. Rev. Gen. 2007

Motifs in transcription regulation networks



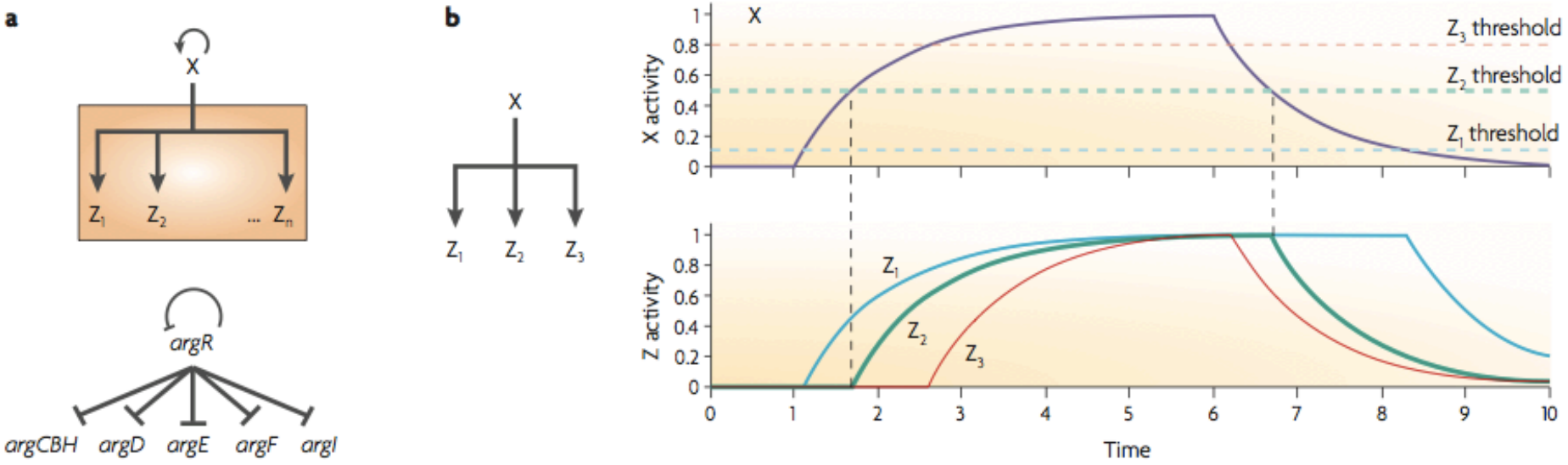
Yeast:



Mangan et al, J. Mol. Bol. 2006

Milo et al, Science 2002

Motifs in transcription regulation networks



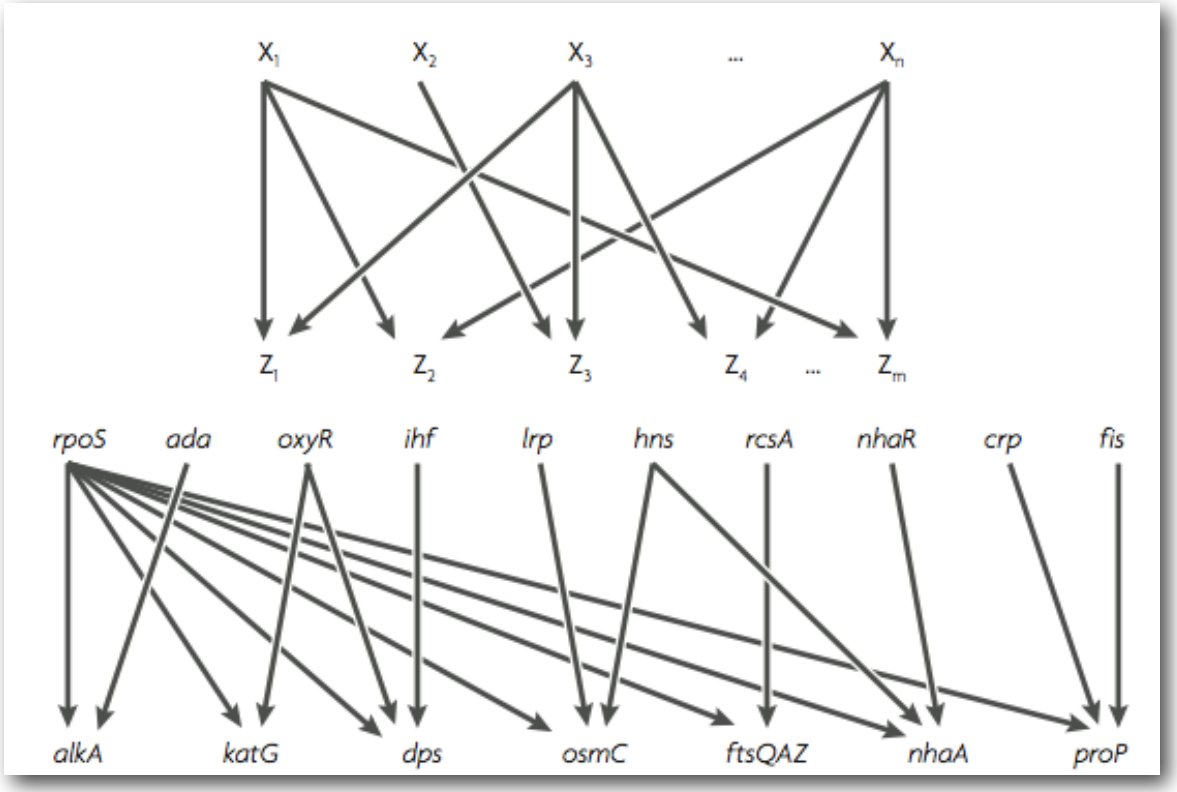
2) Single input module (SIM)

- All operons $Z_1,\ldots,Z_n$ are regulated with the same sign
- None is regulated by a transcription factor other than X
- X is usually autoregulatory

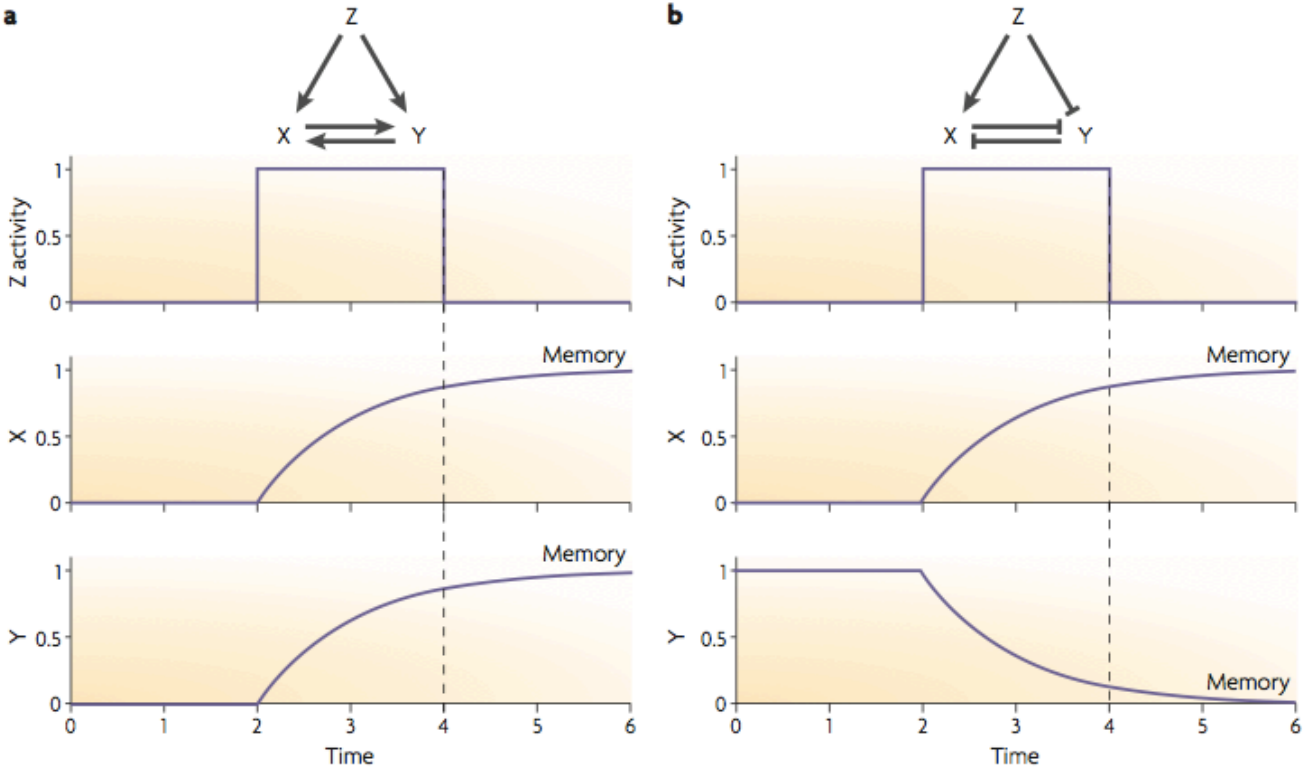
Alon, Nat. Rev. Gen. 2007

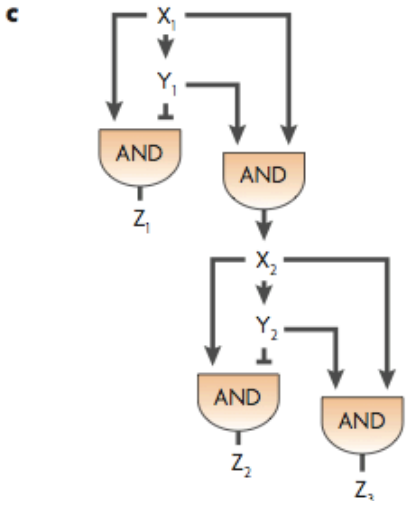
Motifs in transcription regulation networks

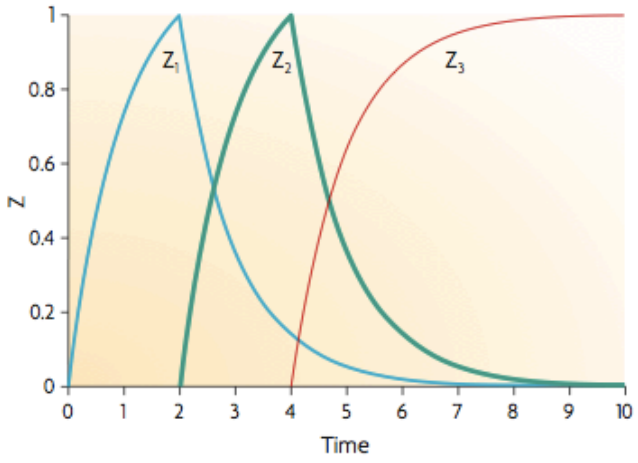
3) Dense overlapping regulon (DOR)



Motifs in transcription regulation networks

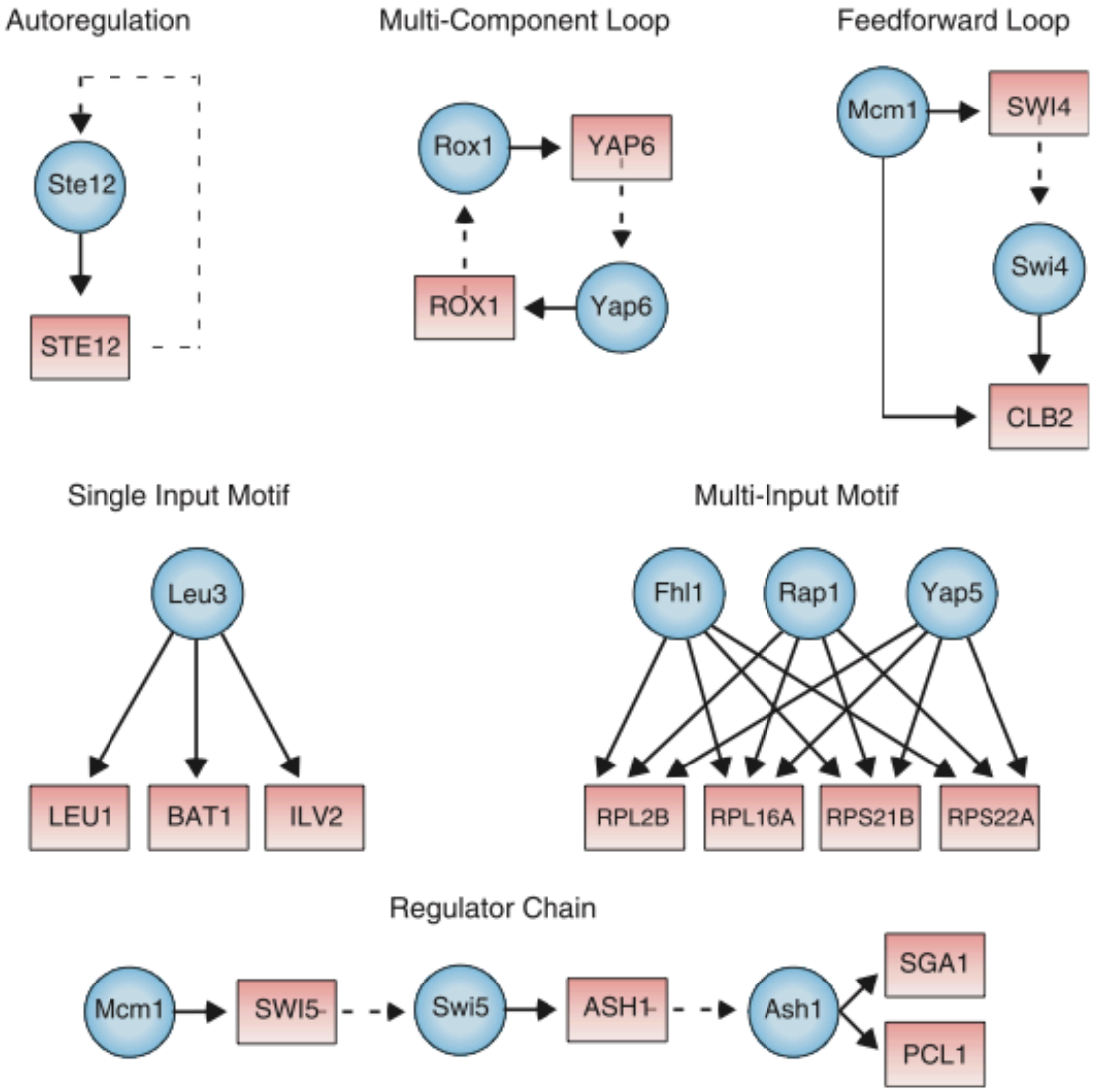






Alon, Nat. Rev. Gen. 2007

Motifs in transcription regulation networks



Lee et al, Science 2002

Motif conservation and convergent evolution

Wuchty et al. showed that in S. cerevisiae, proteins organized in cohesive patterns of interactions are conserved to a substantially higher degree than those that do not participate in such motifs.

They found that the conservation of proteins in distinct topological motifs correlates with the interconnectedness and function of that motif and also depends on the structure of the overall interactome topology.

These findings indicate that motifs may represent evolutionary conserved topological units of cellular networks molded in accordance with the specific biological function in which they participate.

Wuchty et al, Nat. Genetics 2003

Motif conservation and convergent evolution

Experimental set up:
Test the correlation between a protein's evolutionary rate and the structure of the motif it is embedded in

Hypothesis: if there is evolutionary pressure to maintain specific motifs, their components should be evolutionarily conserved and have identifiable orthologs in other organisms

They studied the conservation of 678 S. cerevisiae proteins with an ortholog in each of five higher eukaryotes (Arabidopsis thaliana, C. elegans, D. melanogaster, Mus musculus, and Homo sapiens) from the InParanoid database

Wuchty et al, Nat. Genetics 2003

Motif conservation and convergent evolution

Table 1 Evolutionary conservation of motif constituents

#	Motifs	Number of yeast motifs	Natural conservation rate	Random conservation rate	Conservation ratio
1		9,266	13.67%	4.63%	2.94
2		167,304	4.99%	0.81%	6.15
3		3,846	20.51%	1.01%	20.28
4		3,649,591	0.73%	0.12%	5.87
5		1,763,891	2.64%	0.18%	14.67
6		9,646	6.71%	0.17%	40.44
7		164,075	7.67%	0.17%	45.56
8		12,423	18.68%	0.12%	157.89
9		2,339	32.53%	0.08%	422.78
10		25,749	14.77%	0.05%	279.71
11		1,433	47.24%	0.02%	2,256.67

Table 2 Overrepresentation of human orthologous motifs in various functional classes of yeast proteins

Functional class	Overrepresented motifs
Transport facilitation	(10)
Subcellular localization	(21) (21) (26) (15) (27) (23) (29) (20) (63) (45)
Regulation	(10)
Protein fate	(14) (16) (13) (33) (27) (20) (26) (24) (16) (60) (41)
Cell cycle	(11) (14) (13) (11) (14)
Cellular transport	(11) (12)
Transcription	(12) (16) (17) (13) (16) (19) (17) (15) (14) (21) (23)
Protein synthesis	(12) (11) (17) (11) (24)

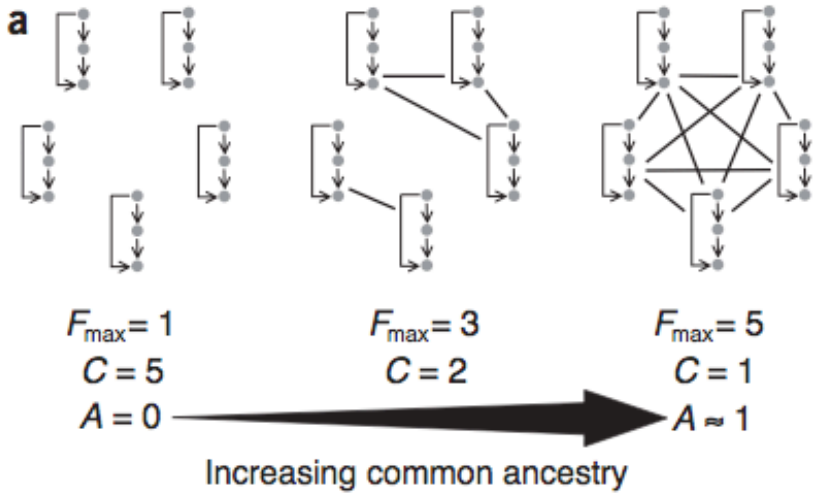
Wuchty et al, Nat. Genetics 2003

Motif conservation and convergent evolution

Convergent evolution is a potent indicator of optimal design
Conant and Wagner recently showed that multiple types of transcriptional regulation circuitry in E. coli and S. cerevisiae have evolved independently and not by duplication of one or a few ancestral circuits

Conant and Wagner, Nat. Genetics 2003

Motif conservation and convergent evolution



b

	Circuit type	Number of circuits	Number of families (*C*)	Index of common ancestry (*A*)	Largest circuit family (F_{max})
Yeast	Feed-forward	48	44 (46.8 ± 1.9; *P* = 0.08)	0.082 (0.023 ± 0.035; *P* = 0.08)	5 (1.9 ± 1.4; *P* = 0.05)
	Bi-fan	542	435 (469.0 ± 37.7; *P* = 0.18)	0.197 (0.135 ± 0.070; *P* = 0.18)	49 (41.0 ± 31.1; *P* = 0.33)
	MIM-2	176	168 (164.5 ± 8.8; *P* = 0.60)	0.045 (0.065 ± 0.050; *P* =0.60)	5 (7.4 ± 6.2; *P* = 0.59)
	Reg. chain (3)	33	33	0	1
E. coli	Feed-forward	11	11	0	1
	Bi-fan	27	27	0	1

Table 1 Gene families are not over-represented in circuit types

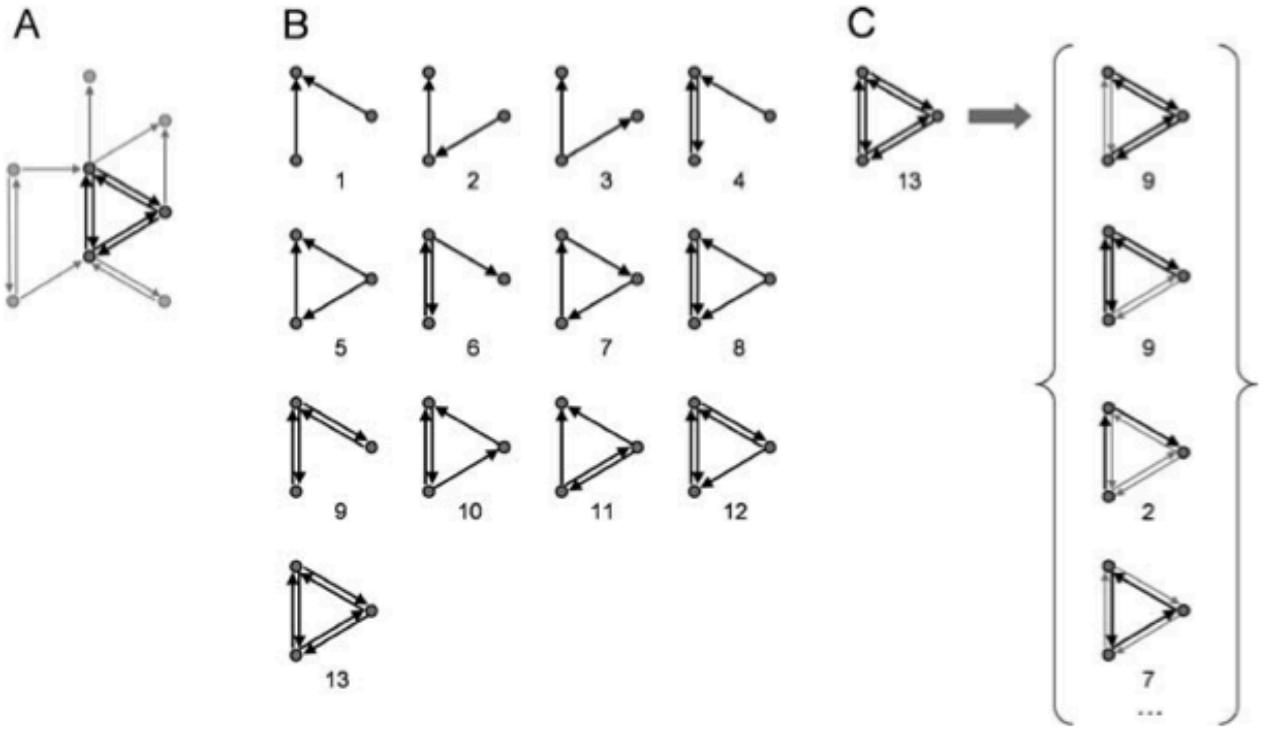
Organism	Circuit type	P_{motif}[a]	$P_{motif\|duplicate}$[b]	*P*[c]
S. cerevisiae	Bi-fan	0.82	0.80	NA
	Feed-forward	0.38	0.42	0.21
	Multi-input motif	0.77	0.76	NA
	Regulator chains	0.64	0.67	0.30
E. coli	Bi-fan	0.50	0.67	0.11
	Feed-forward	0.82	0.67	NA

[a]Probability of a randomly chosen regulatory gene occurring in a given circuit type. [b]Probability of a regulatory gene occurring in a circuit type given that one of its duplicates occurs in that circuit type (see **Supplementary Methods** online). [c]*P* value for one-sided exact binomial test of the null hypothesis $P_{motif} = P_{motif\ |\ duplicate}$. NA indicates that a test has not been carried because $P_{motif} > P_{motif\ |\ duplicate}$. The number of transcriptional regulators was $n = 112$ and $n = 22$ for the yeast and *E. coli* analyses, respectively.

Conant and Wagner, Nat. Genetics 2003

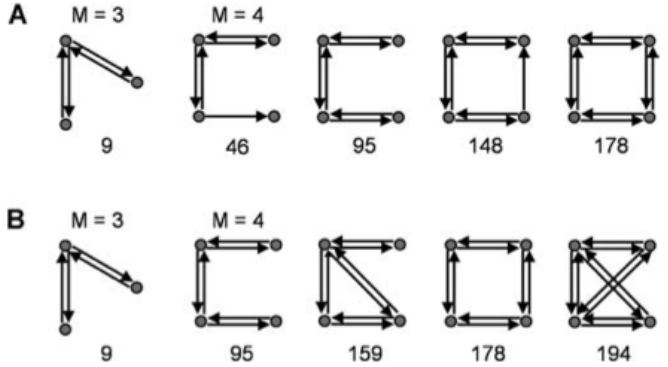
Motifs in brain networks

Structural vs functional motifs: functional motifs are subsets of structural motifs (options for function that are facilitated by the structure).



Sporns and Kötter, PLoS Bio. 2004

Motifs in brain networks



Low number of structural motifs, high number of functional motifs.

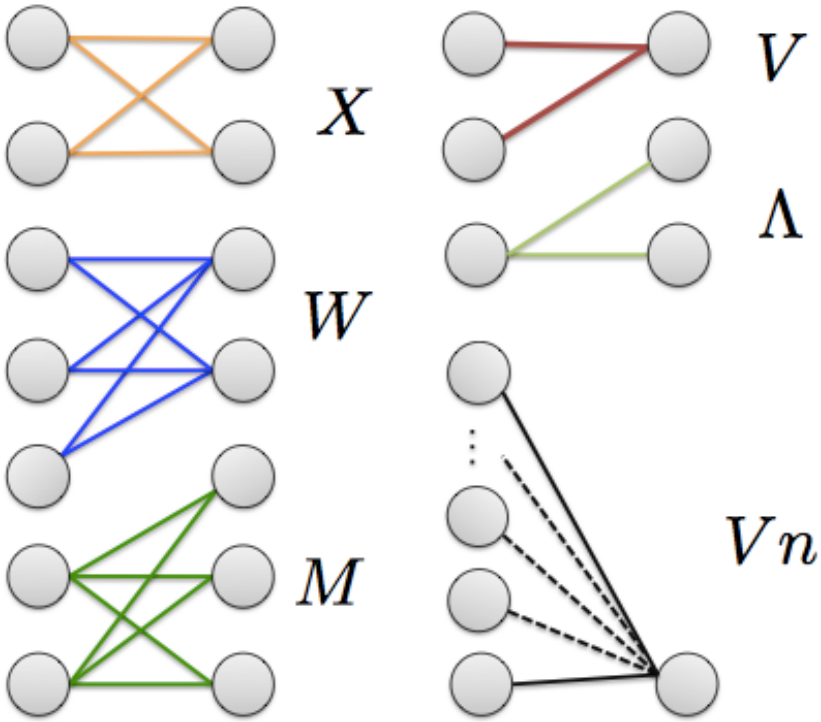
Table 1. Structural and Functional Motif Number for Cortical Connection Matrices and Corresponding Random and Lattice Matrices

Brain Network	*M*	Structural Motifs			Functional Motifs		
		Real	Random	Lattice	Real	Random	Lattice
Macaque Visual	2	190	243 (4)	191 (2)	432	380 (4)	431 (2)
Cortex	3	1,486	2,353 (51)	1,344 (40)	19,769	14,358 (325)	21,120 (308)
	4	10,487	18,076 (391)	8,688 (414)	1,843,308	1,013,131 (55,187)	2,259,970 (90,404)
	5	62,940	105,926 (2,059)	50,278 (2,863)	334,279,477	121,572,738 (13,874,054)	513,004,042 (50,992,845)
Macaque Cortex	2	438	654 (7)	471 (7)	1,054	838 (7)	1,021 (7)
	3	4,584	10,786 (227)	4,439 (143)	53,601	30,449 (648)	56,043 (871)
	4	51,129	173,235 (4,635)	39,345 (2,346)	5,306,188	1,850,355 (87,743)	6,617,493 (272,110)
Cat Cortex	2	519	656 (7)	510 (5)	1,054	838 (7)	1,021 (7)
	3	6,986	10,898 (160)	6,021 (122)	53,601	30,449 (648)	56,043 (871)
	4	87,673	149,791 (2,250)	65,527 (2,150)	5,306,188	1,850,355 (87,743)	6,617,493 (272,110)
C. elegans	2	1,718	1,922 (6)	1,700 (40)	2,230	2,026 (6)	2,248 (40)
	3	31,070	41,707 (279)	23,376 (1,494)	70,911	55,054 (363)	84,245 (4,200)
	4	674,125	1,081,682 (11,105)	316,228 (36,200)	3,430,885	2,160,611 (34,800)	5,326,201 (578,900)

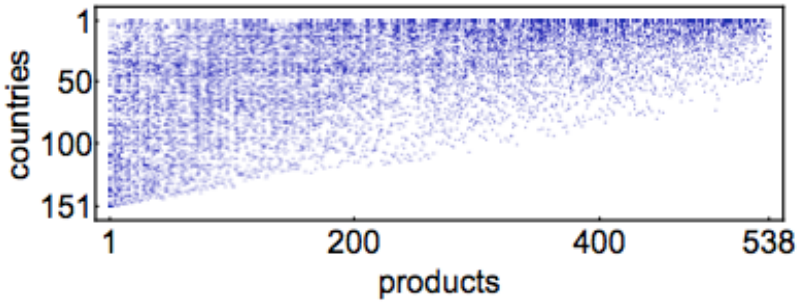
Sporns and Kötter, PLoS Bio. 2004

Bipartite networks

The simplest motifs in bipartite networks:



Example: World Trade Network.
Set A - countries
Set B - products

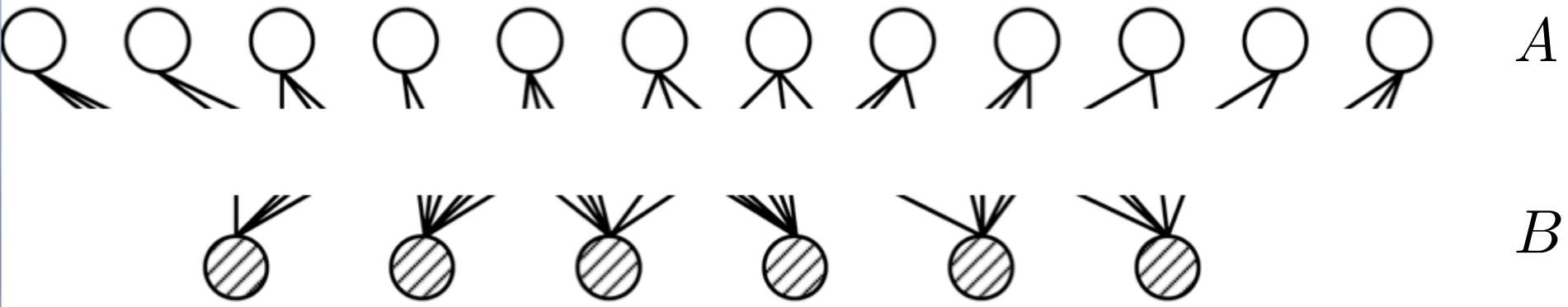


Saracco et al, submitted to New Journal of Physics 2015

Bipartite networks

What is the correct null model here?

Bi-configuration model is obtained by randomly matching stubs from A and B:



$$\sum_{i \in A} k_i = \sum_{j \in B} k_j$$

Bipartite networks

Advanced:

The Shannon entropy S of an ensemble of networks is defined as:

$$S = -\sum_{M} P(M) ln P(M)$$

...where M is an adjacency matrix. The most random configurations define an ensemble for which S is maximal, where restrictions or constraints C(M) should be considered. E.g. C(M) is the degree sequences in the sets A and B. We arrive at the distribution containing as many parameters as constraints (Lagrange multipliers). Their values are calculated by the maximum likelihood method. See the original paper for details.

Saracco et al, submitted to New Journal of Physics 2015

Bipartite networks

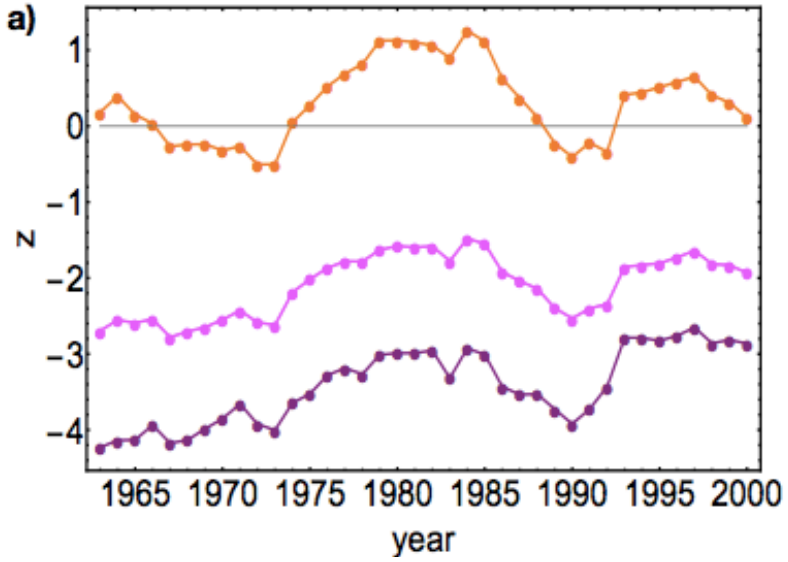
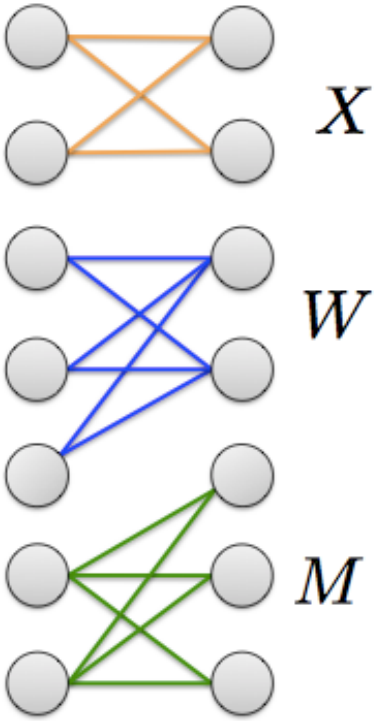
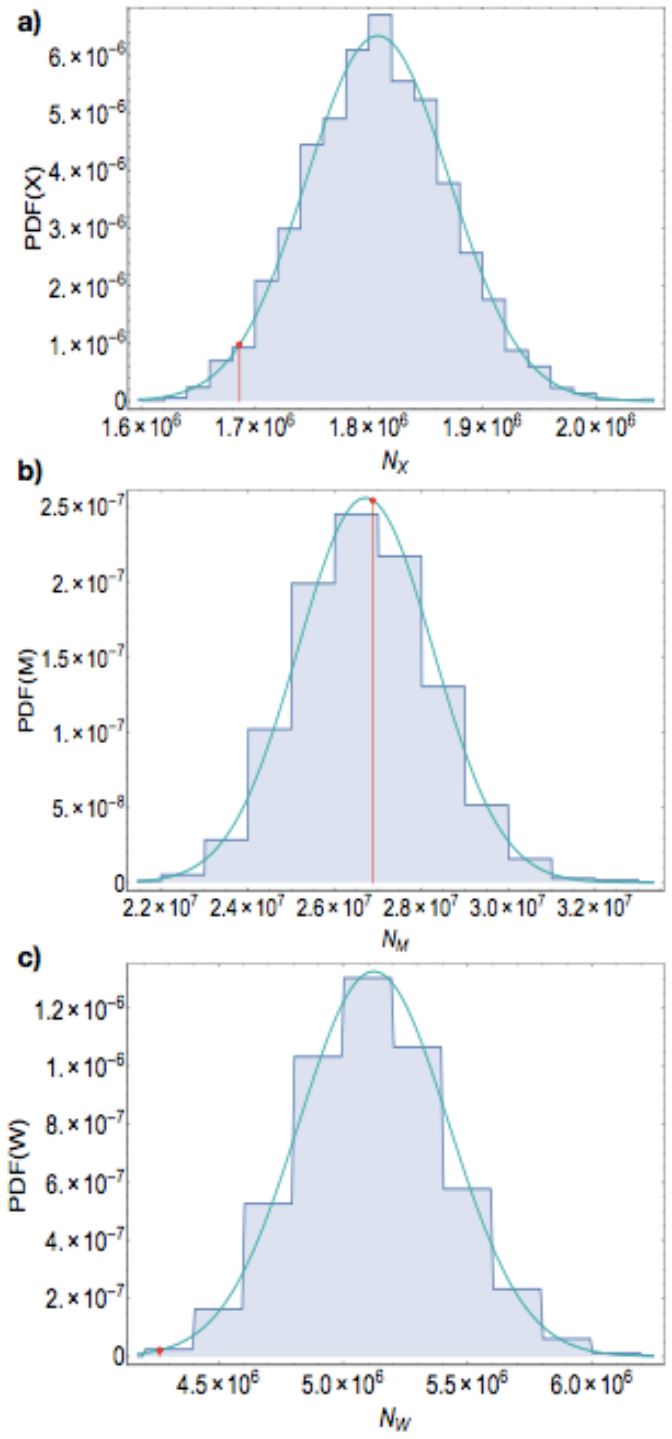


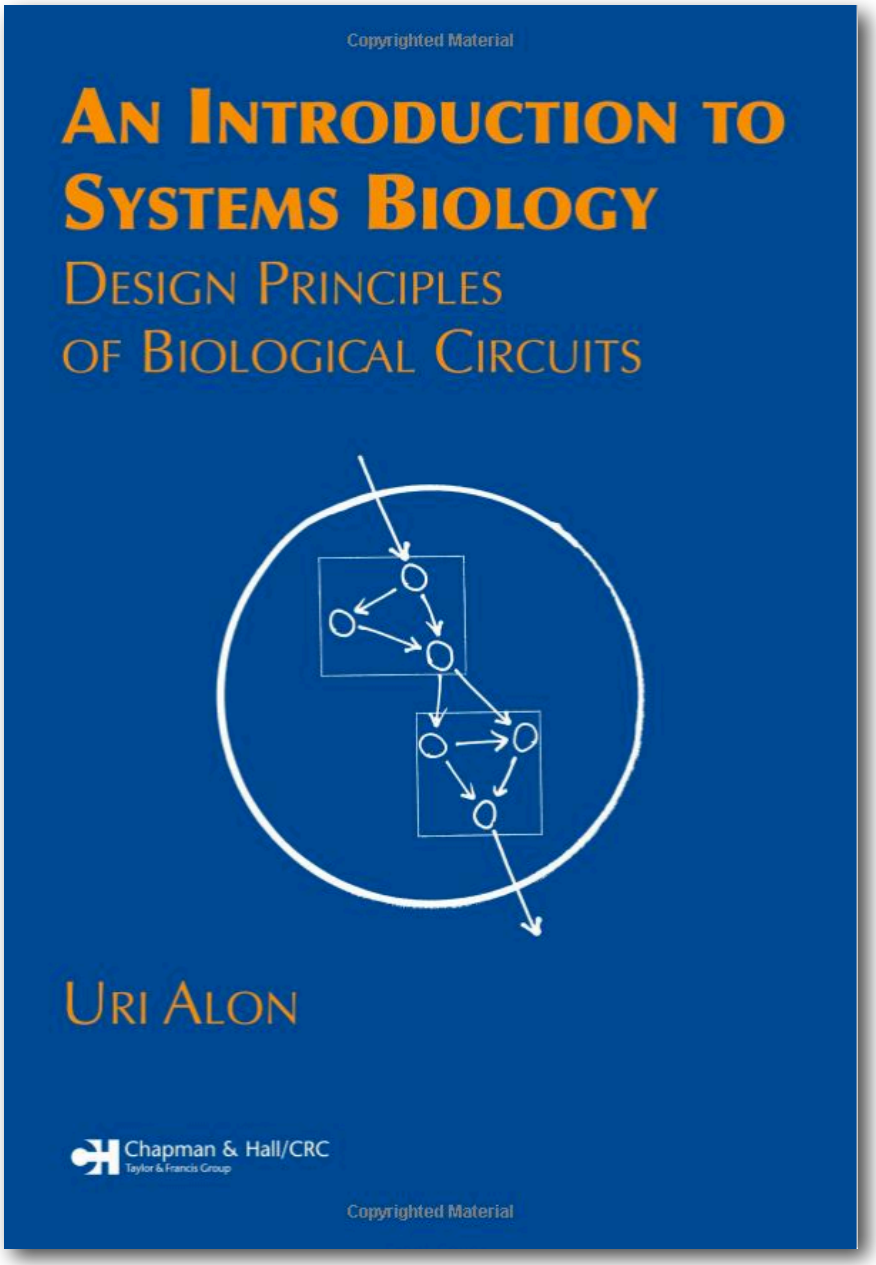
FIG. 13. Analysis of motifs. From top to bottom: z-scores and similarity evolution across our database years of N_M (•), N_X (•), N_W (•).





Saracco et al, submitted to New Journal of Physics 2015

Want to learn (a lot) more about motifs?



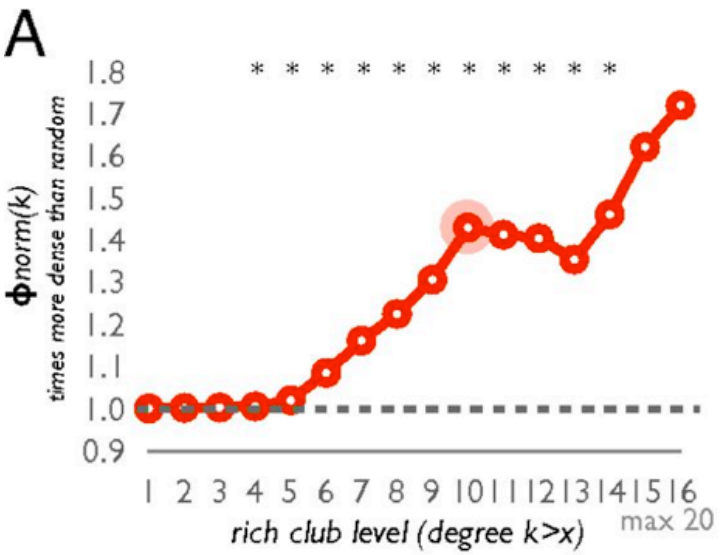
Motifs: an alternative definition

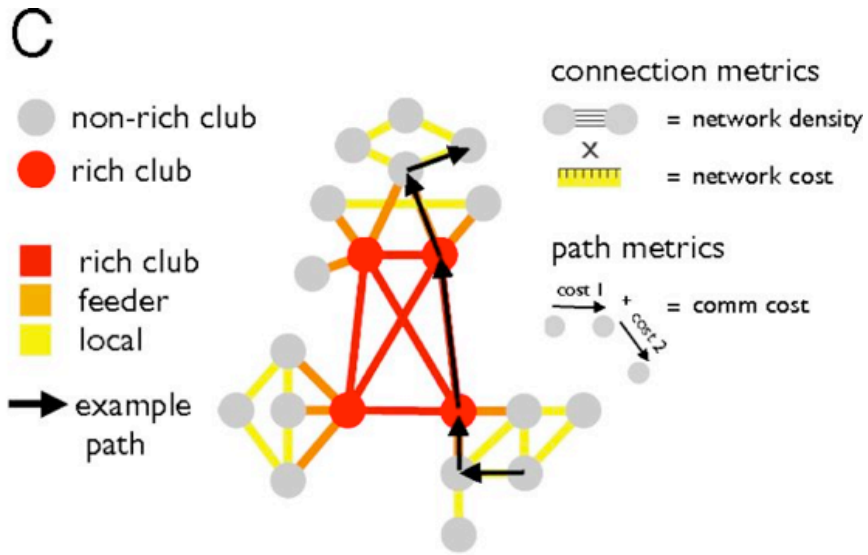
Motifs are usually defined as above - subgraphs with a given number of nodes M.

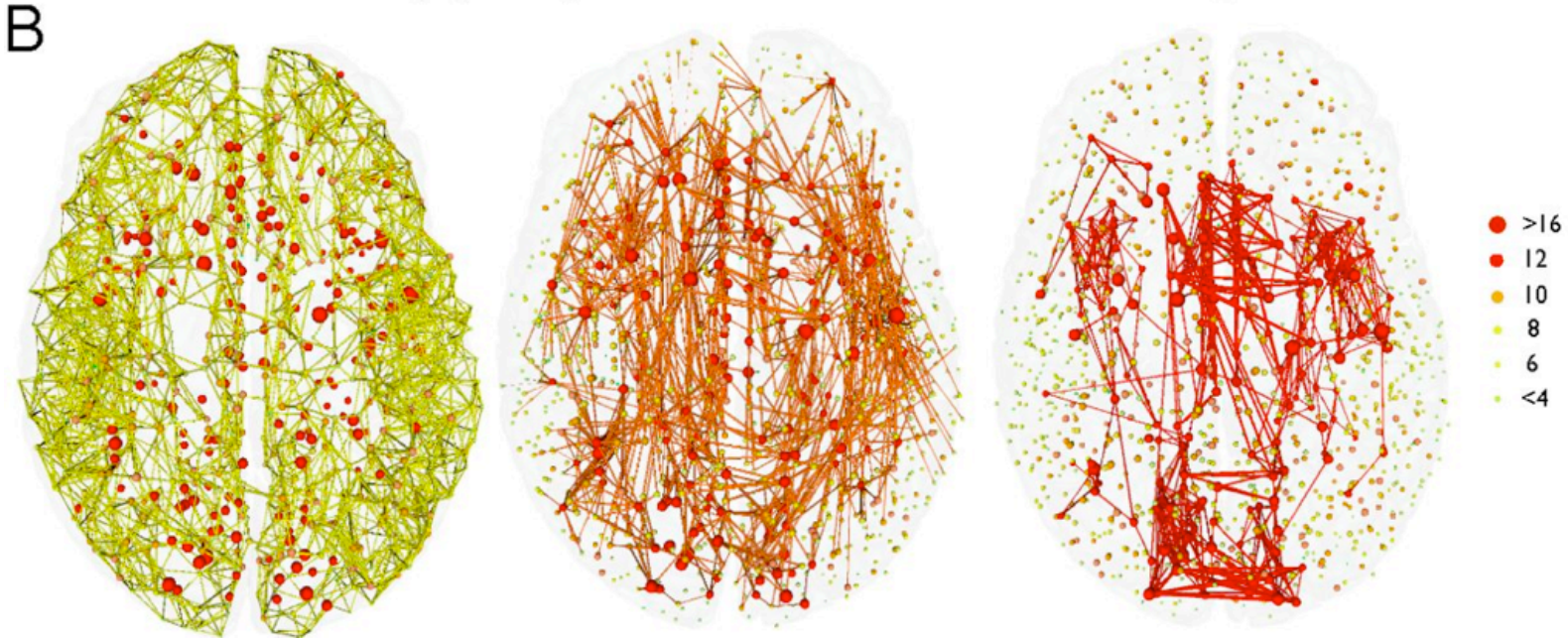
However, one may consider paths as motifs by labeling each path segment and observing the pattern as a motif.

(Sidenote: the concept of 'labeling' here is a general one. One may overlay node or edge attributes in the previous analyses to create motif repertoires.)

Motifs: an alternative definition

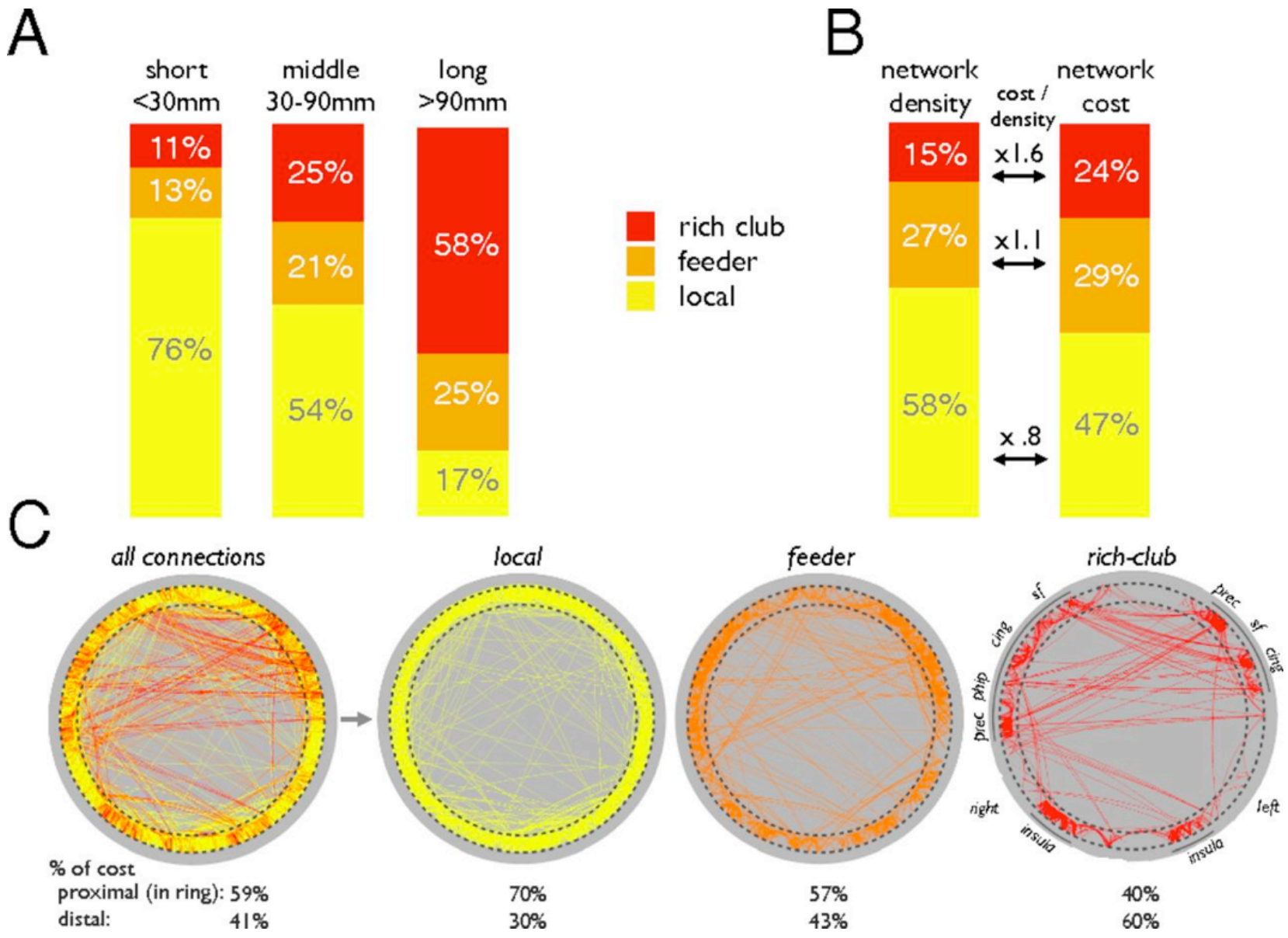






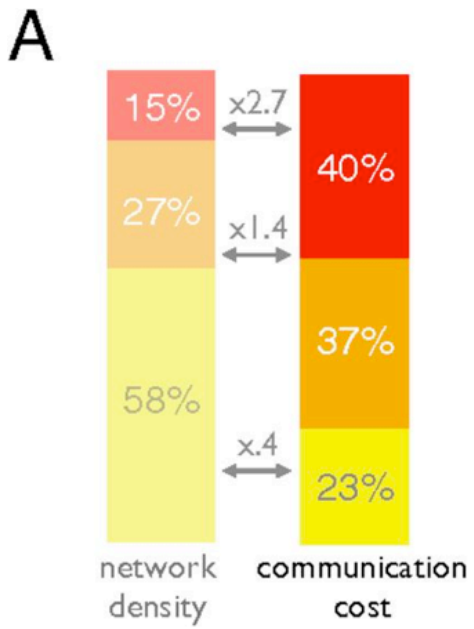
van den Heuvel et al, PNAS 2012

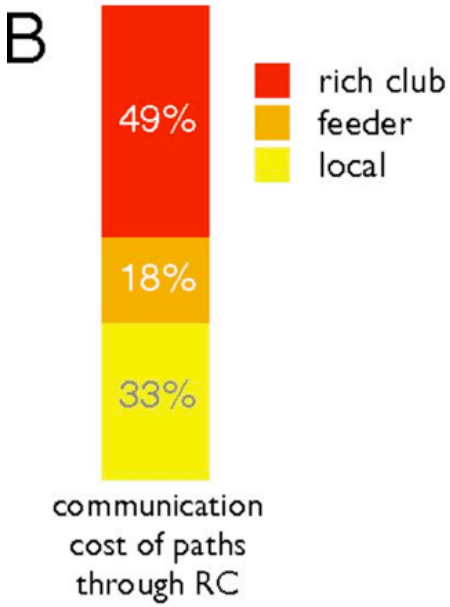
Motifs: an alternative definition

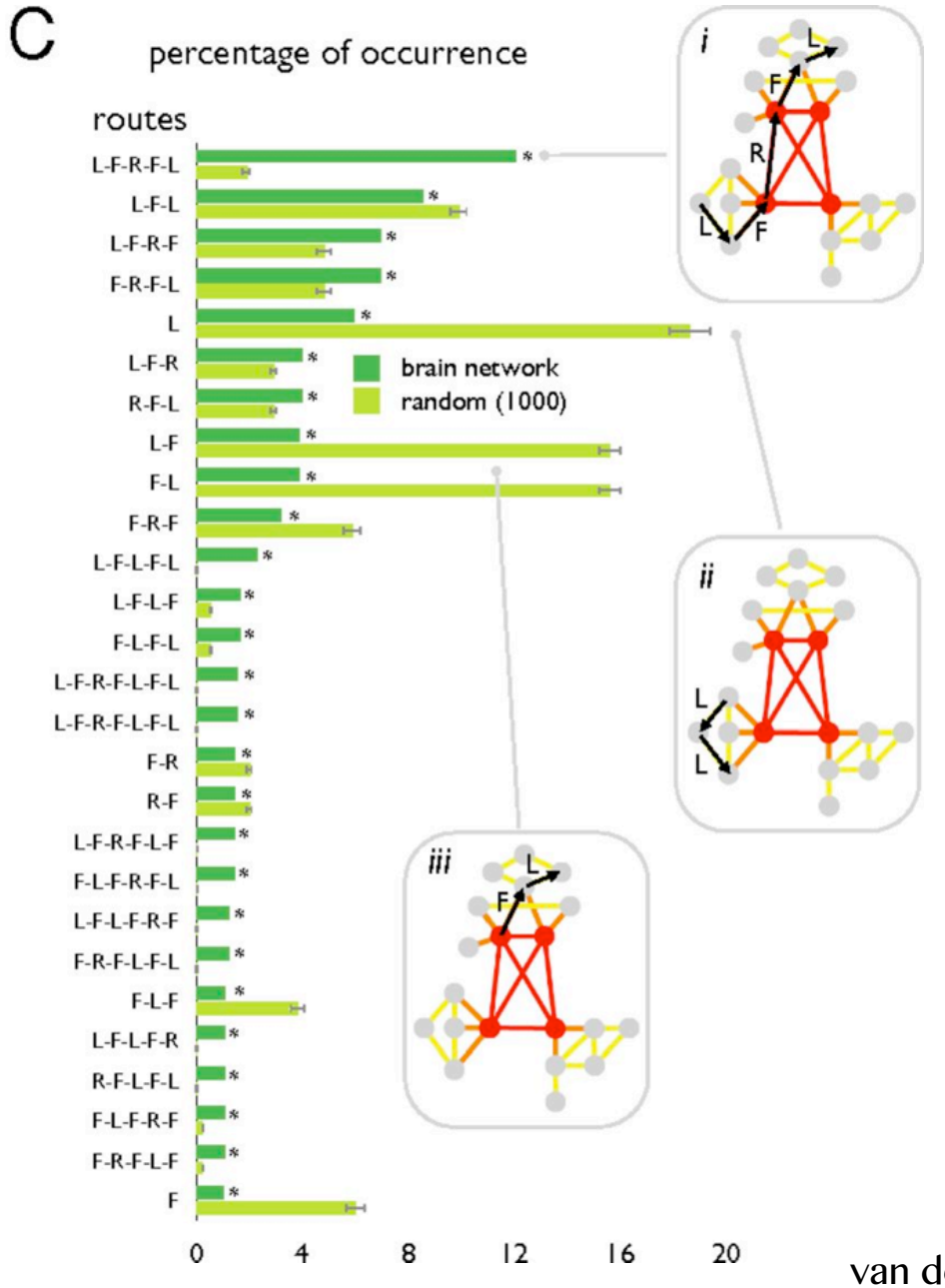


van den Heuvel et al, PNAS 2012

Motifs: an alternative definition

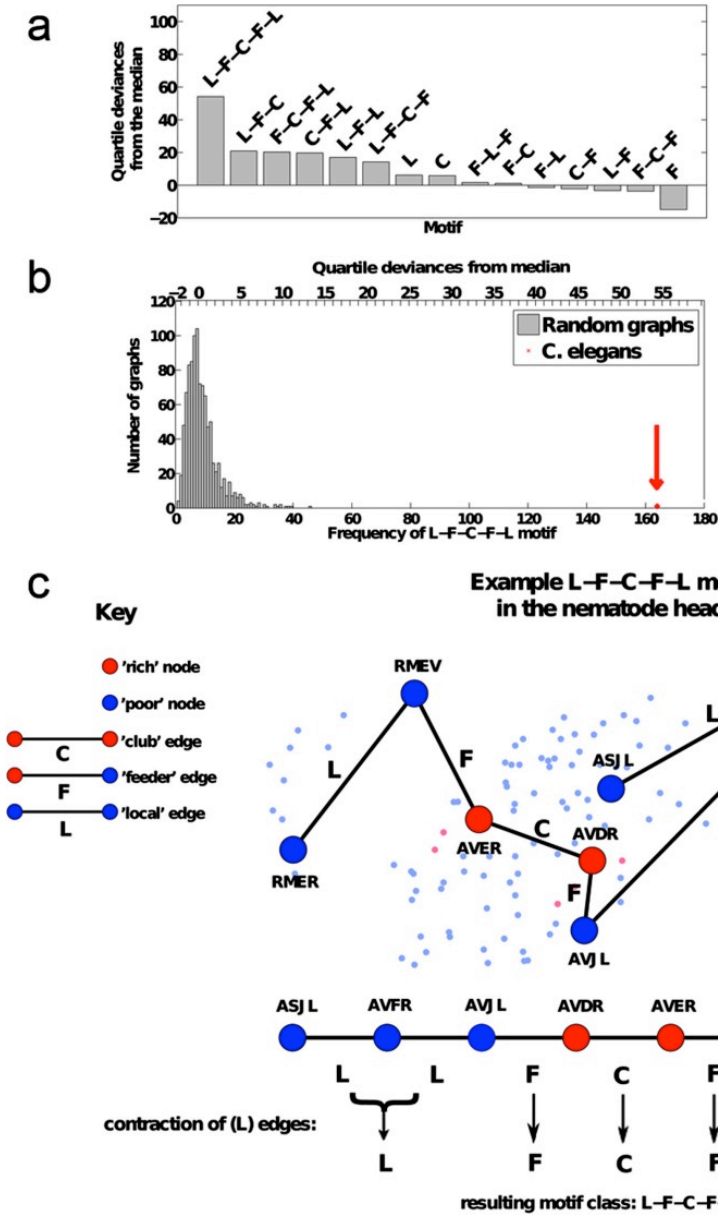




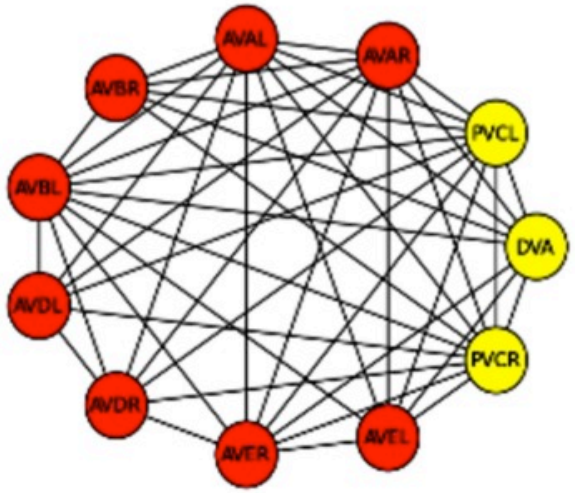


van den Heuvel et al, PNAS 2012

Motifs: an alternative definition



Observe the same pattern in C. elegans as human (and, it turns out, macaque, cat, mouse...)



Towlson et al, J. of Neuroscience 2013

The computational challenges of finding motifs

To perform a motif analysis, one needs to:

i) Calculate the frequency of a given subgraph. The number of possible subgraphs increases exponentially with network or subgraph size.

ii) Determine if graphs are isomorphic: not known if NP complete or solvable in polynomial time. Best known algorithm scales as $2^{\sqrt{nlogn}}$

iii) Calculate statistical significance - repeat the problem many times.

Wong et al, Briefings in Bioinformatics 2011

Software

מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE
Department of Molecular Cell Biology

Uri AlonLab

Design Principles in Biology

Network Motif Software
Pareto Front Software
Downloadable data
Pareto Task Inference (ParTI) method
Collection of complex networks
Dual Kinect Software

NETWORK MOTIF SOFTWARE

Collection of complex networks:

mfinder

Network motifs detection tool

- Manual
- Software mfinder1.21.zip
- Source code for Windows mfinder1.21_source.zip
- Source code for Linux mfinder1.21_unix.tar

  Linux version: run 'tar -xvf mfinder1.21_unix.tar', 'cd mfinder1.21', and then 'make'
- Motif dictionary
- Supplementary materials - Milo et al. Science 2002
- Acknowledgments: We thank Jacobien Carstens for bringing to our attention the bug in the switching method of undirected networks and for suggesting a fix. This fix is now implemented in mfinder1.21 (May 2015). We thank Paul Brodersen for spotting and fixing a bug in the unix makefile that prevented mfinder to compile in some cases.

Software

Brain Connectivity Toolbox

Search this site

Navigation

- **Home**
  - Getting started
  - Latest releases
  - All functions
  - All help headers
- **Network construction**
- **Network measures**
  - List of measures
- **Network models**
- **Network comparison**
  - Network Based Statistic Toolbox
- **Network visualization**
- **Connectivity network data sets**

Home

Download the Toolbox

Getting started

The Brain Connectivity Toolbox (brain-connectivity-toolbox.net) is a MATLAB toolbox for complex-network analysis of structural and functional brain-connectivity data sets.

Reference and citation

Complex network measures of brain connectivity: Uses and interpretations.
Rubinov M, Sporns O (2010) NeuroImage 52:1059-69.

Brain Connectivity Toolbox in other projects

The Brain Connectivity Toolbox codebase is widely used by brain-imaging researchers, and has been wholly or partially ported to, or included in, the following projects:

bctpy: Brain Connectivity Toolbox for Python.
bct-cpp: Brain Connectivity Toolbox in C++.
Human Connectome Project: A consortium for mapping human brain connectivity.

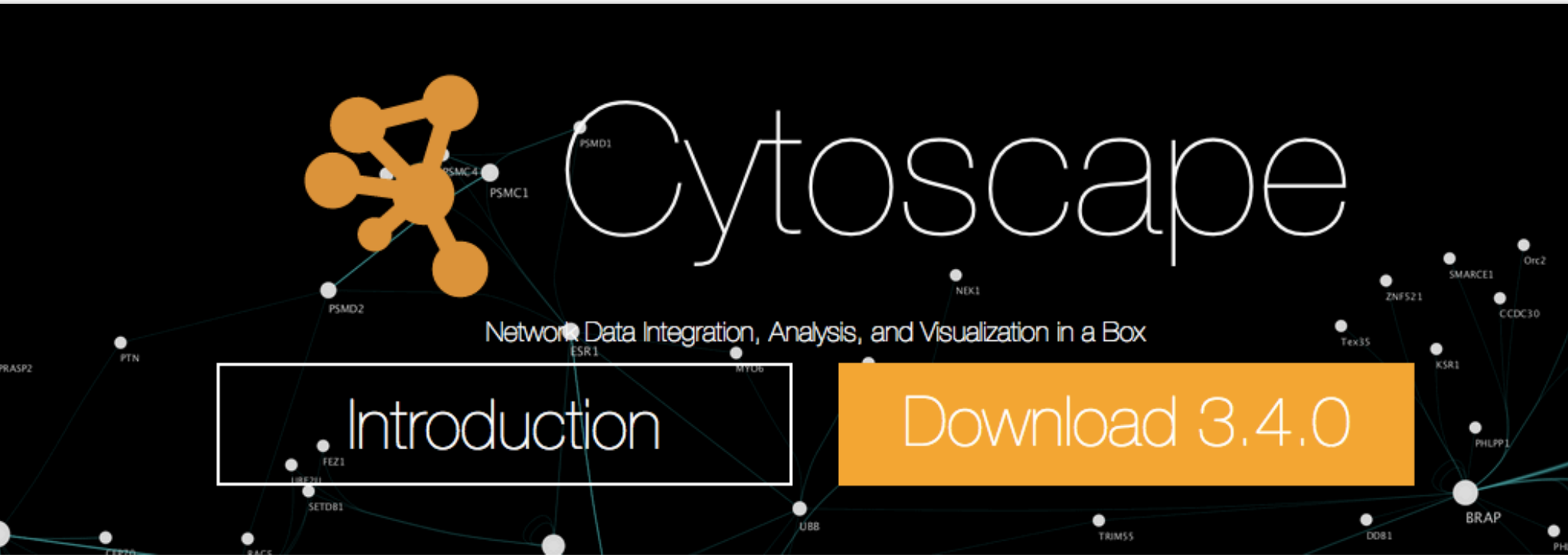
Software

In the interest of using a piece of software that:

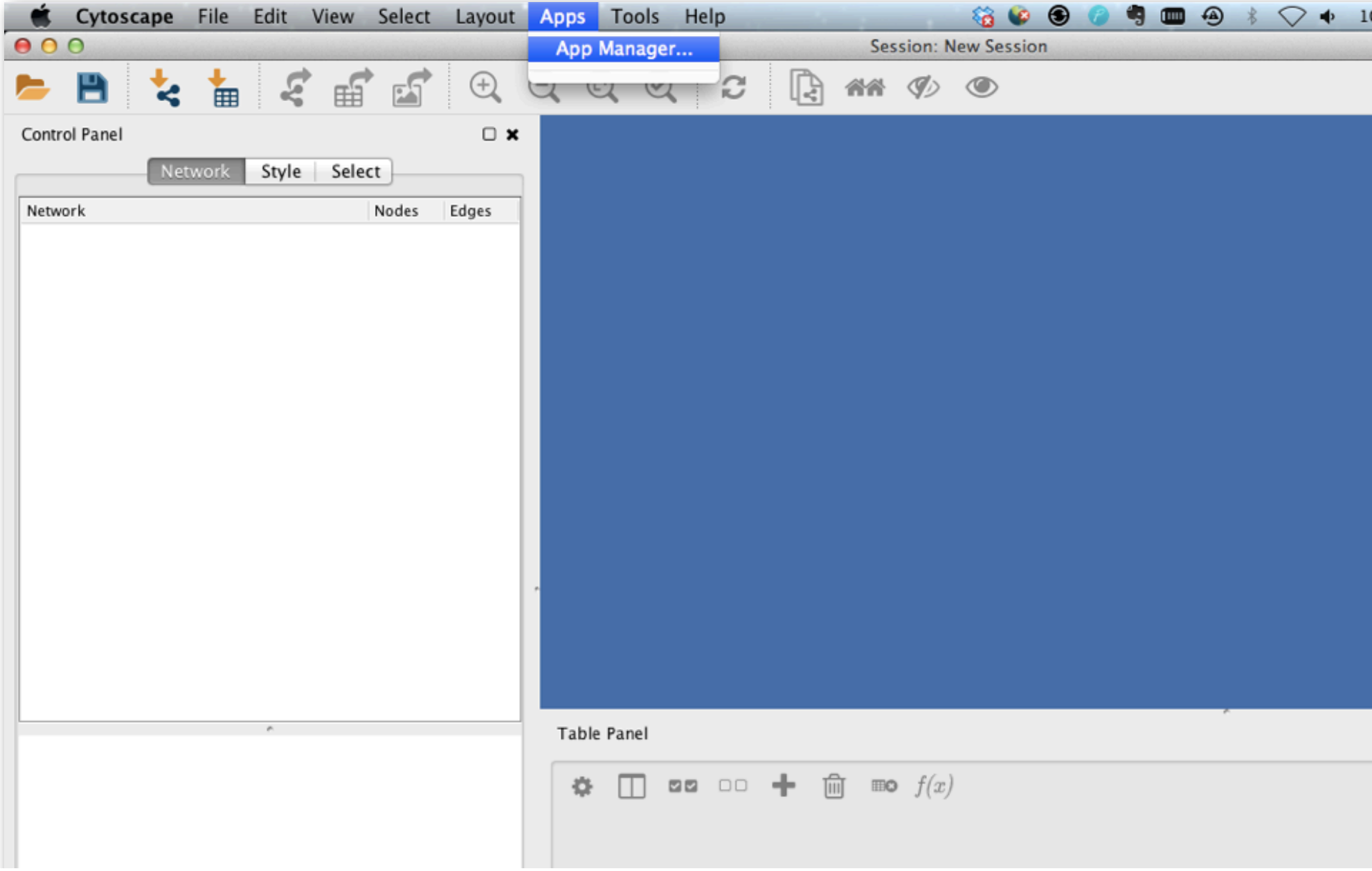
(a) has been updated in the last decade

(b) is open source

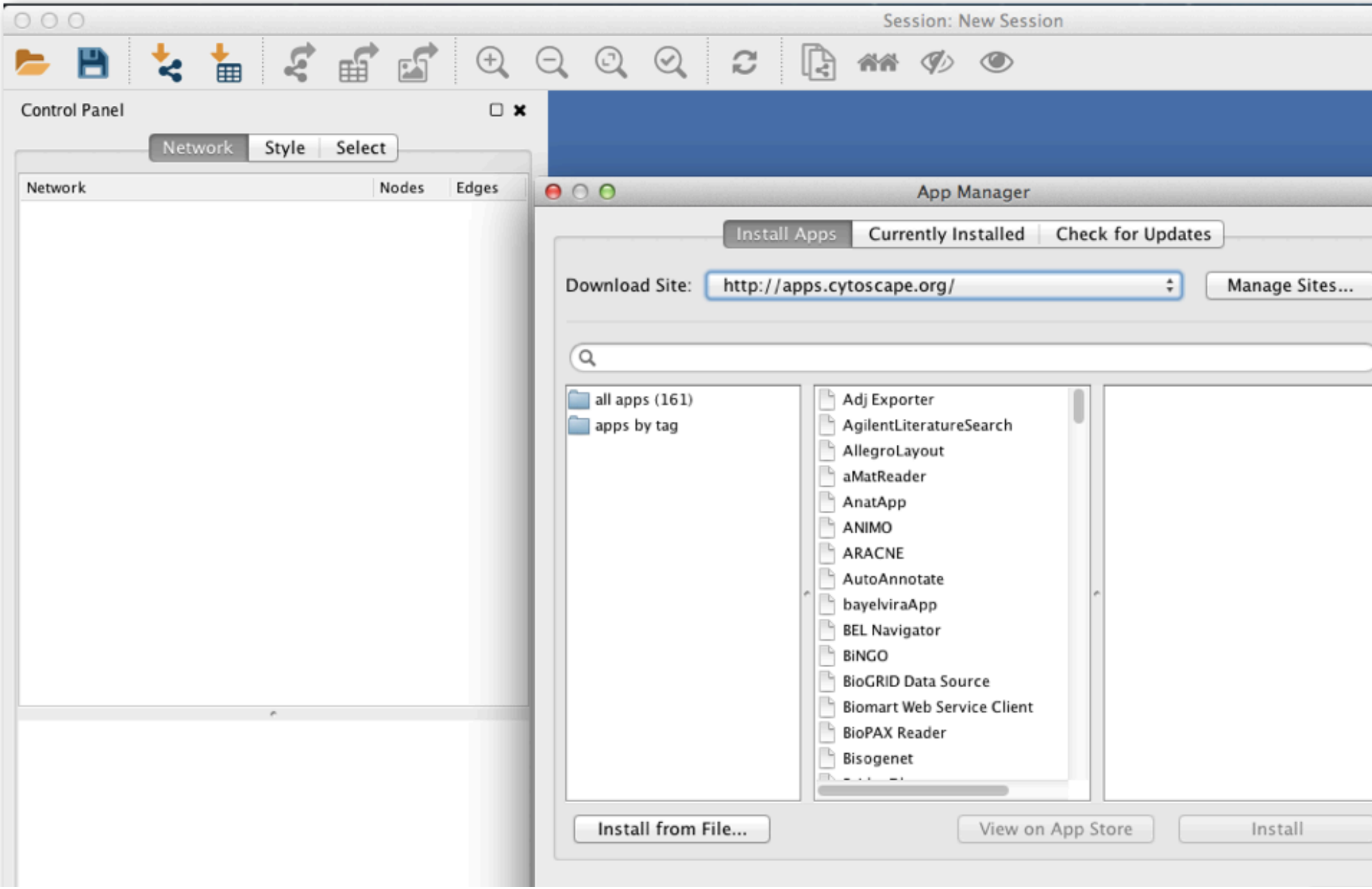
(c) is available for Mac, Linux & Windows



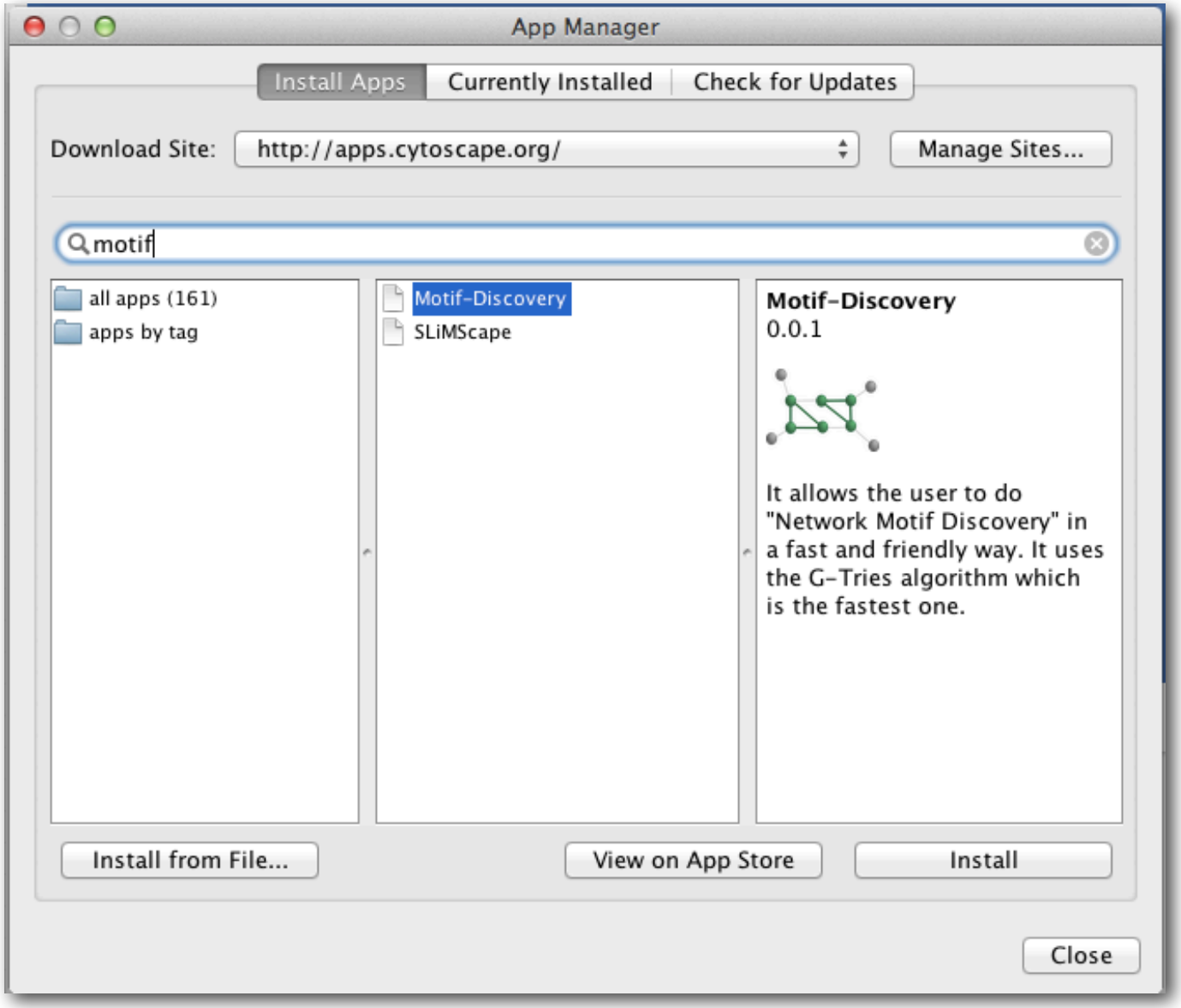
Load the "Motif Discovery" app



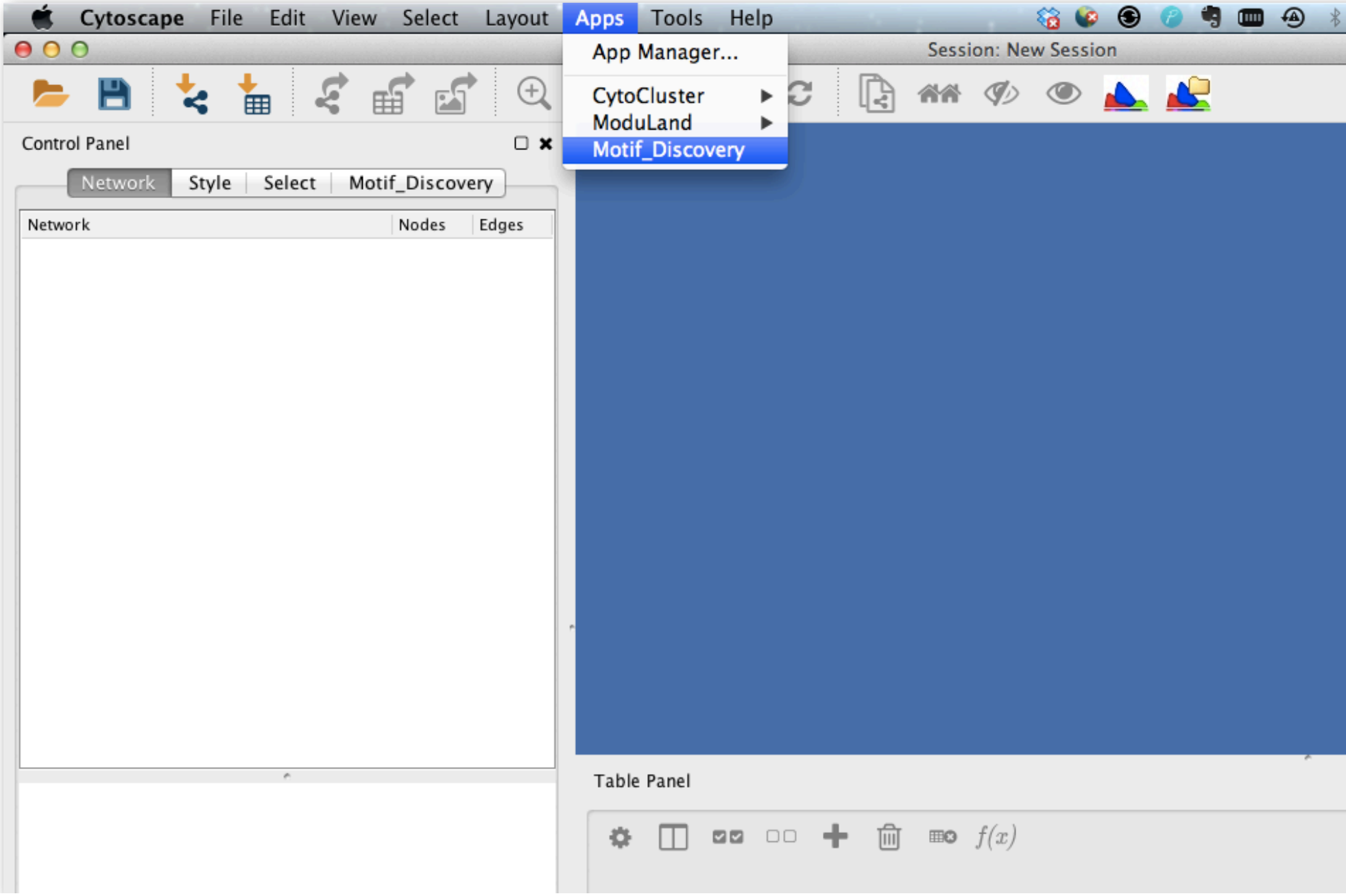
Load the "Motif Discovery" app



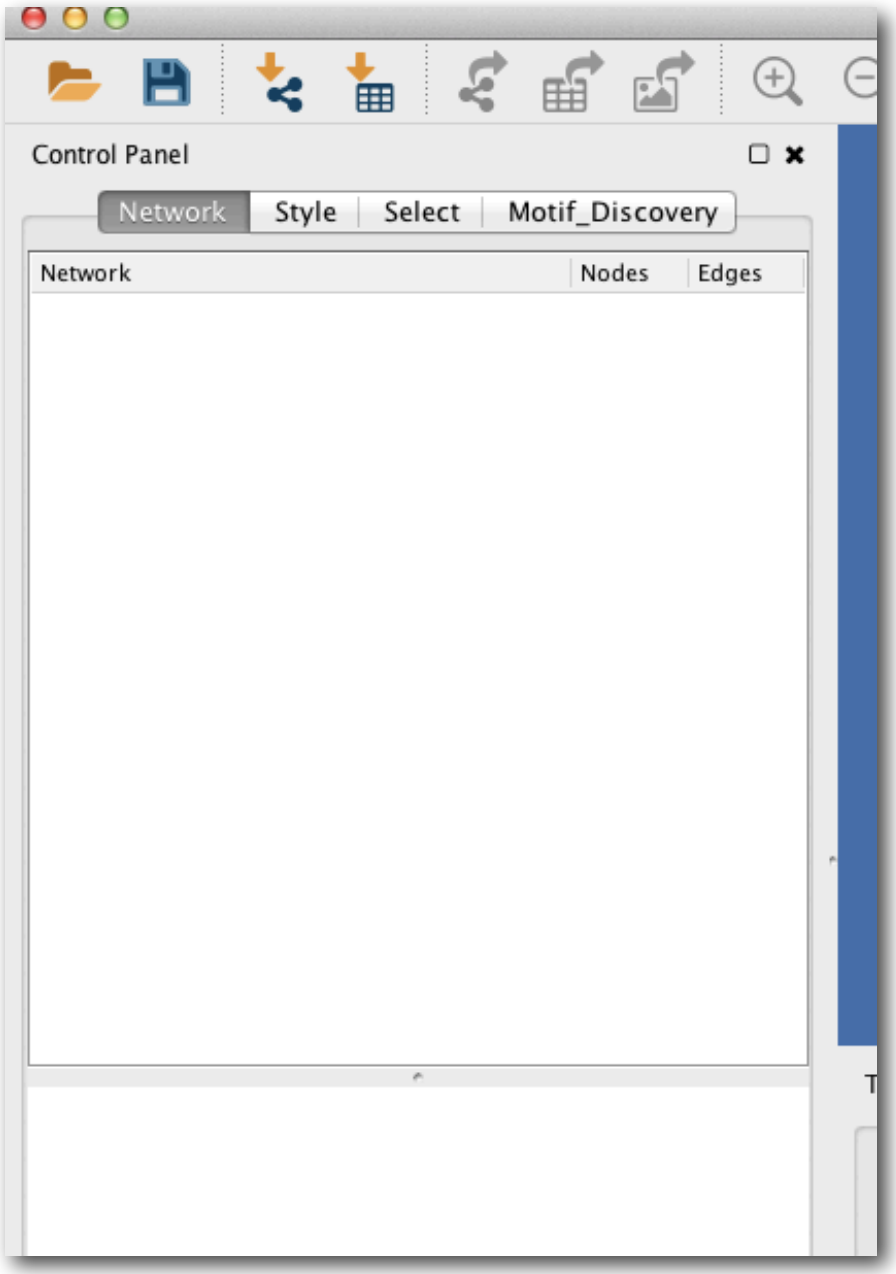
Load the "Motif Discovery" app

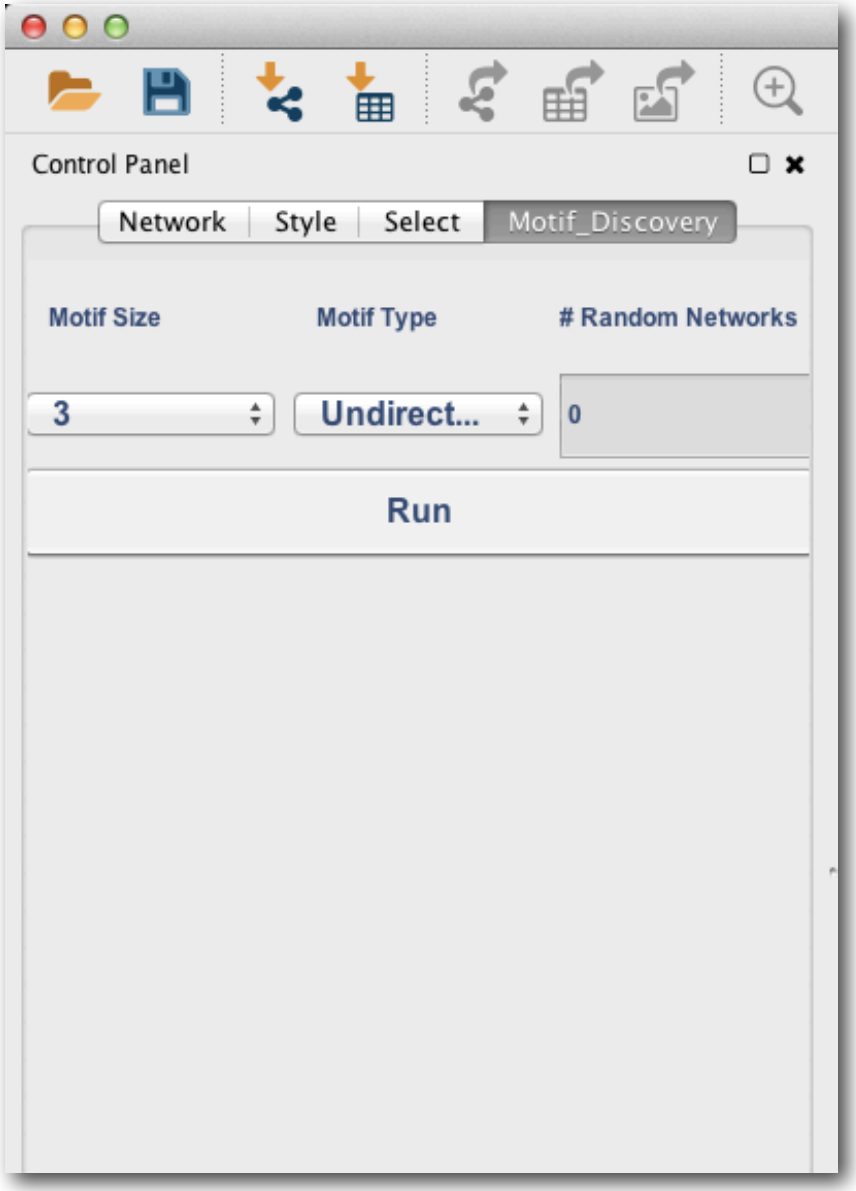


Load the "Motif Discovery" app



Load the "Motif Discovery" app





Your turn!

Open up your networks and look for motifs with a statistically surprisingly high frequency.

- You may work in pairs
- If your network is large (larger than ~1000 nodes), this will be slow - recommend you either work with a subgraph if you have one to hand, or with a partner
- Start with the smaller motifs - 3 nodes and going upwards
- Normally you would want at least 1000 random graphs. For now if this is slow, 100 will suffice.

You may be asked to:

- Describe motifs which appear with high statistical frequency
- Explain your interpretation of the functional roles these motifs might play in your network