Detection of network motifs by local concentration

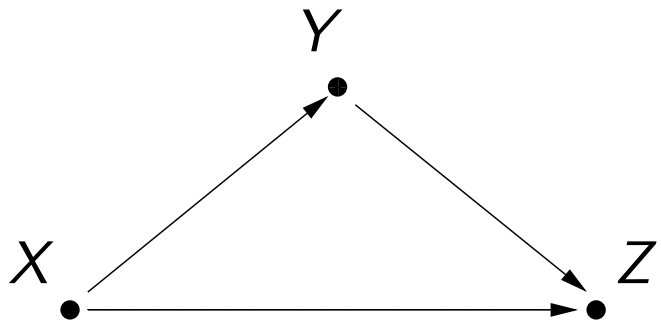
Etienne Birmelé

Laboratoire *Statistique et Génome*, Université d'Evry
Groupe SSB - ANR NeMo

Network motifs

A *motif* is a small graph which is over-represented in a network: it's a candidate to be studied for a potential biological meaning.

Example: the feed-forward loop



Network motif detection

All previous methods look for an overall over-representation:

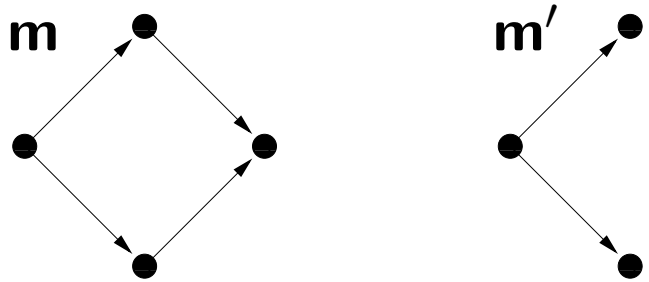
- U. Alon's group (since 2002): simulations for size 3 and 4, Z-score
- J. Berg and M. Lässig (2004): probabilistic motifs by an alignment heuristic
- F. Picard et al (2008): mixture model for the network and *Polya-Aeppli* distribution.

Leading ideas

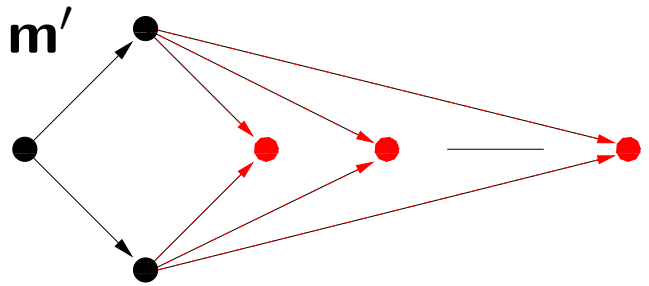
- A small graph $\mathbf{m}$ may be over-represented because one of its subgraphs $\mathbf{m}'$ is over-represented. In that case, $\mathbf{m}'$ is the relevant motif.
- Motifs in regulatory networks are known to be concentrated on some places of the networks (Dobrin & al 04).

Local motifs

Consider a paatern $\mathbf{m}$ and a subgraph $\mathbf{m}'$ of $\mathbf{m}$ obtained by the deletion of a vertex in $\mathbf{m}$.



$\mathbf{m}$ is a *local motif* with respect to $\mathbf{m}'$ if there exist a *theme* (Zhang et al.) of significantly higher order than expected.



Random graph model

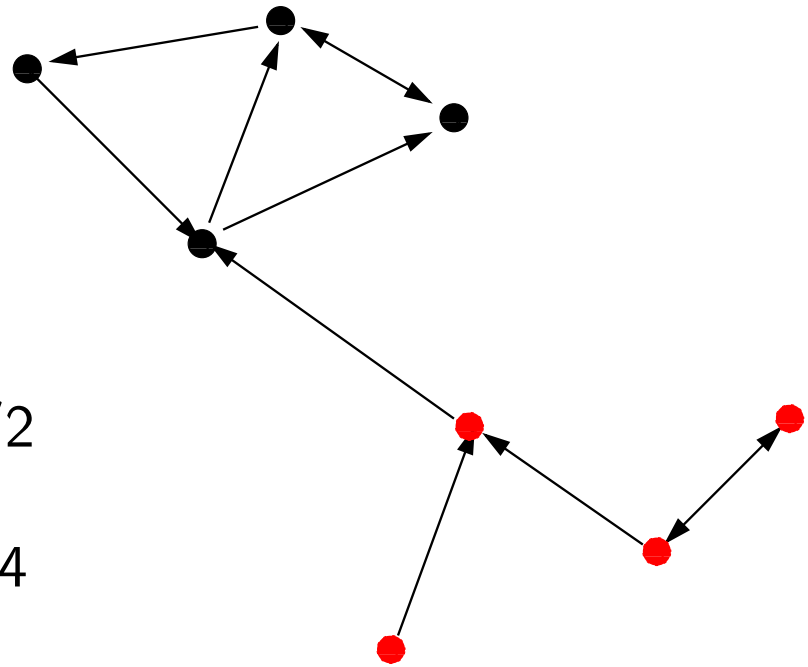
We fix the number n of nodes and the underlying random graph model is defined by a $n \times n$ matrix C: the edge indicators $(X_{ij})_{1 \leq i,j \leq n}$ are independent Bernoulli variables and

$$P(X_{ij} = 1) = c_{ij}$$

In particular, our theory is valid for:

- Edge probability proportional to $d_i d_j$.
- Mixture models on graphs with fixed classes.

Random graph model



$\mathbb{P}(NN) = 1/2$

$\mathbb{P}(RR) = 1/4$

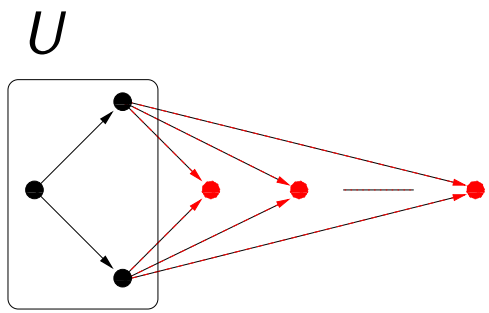
$\mathbb{P}(NR) = 0$

$\mathbb{P}(RN) = 1/16$

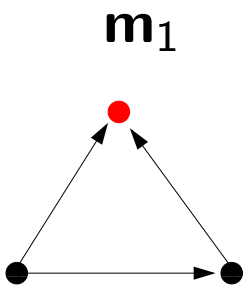
Notations

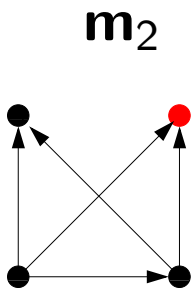
Let $\mathbf{m}$ be a pattern on k vertices and $\mathbf{m}'$ a subpattern. Let U be a set of $k-1$ vertices corresponding to an occurence of $\mathbf{m}'$.

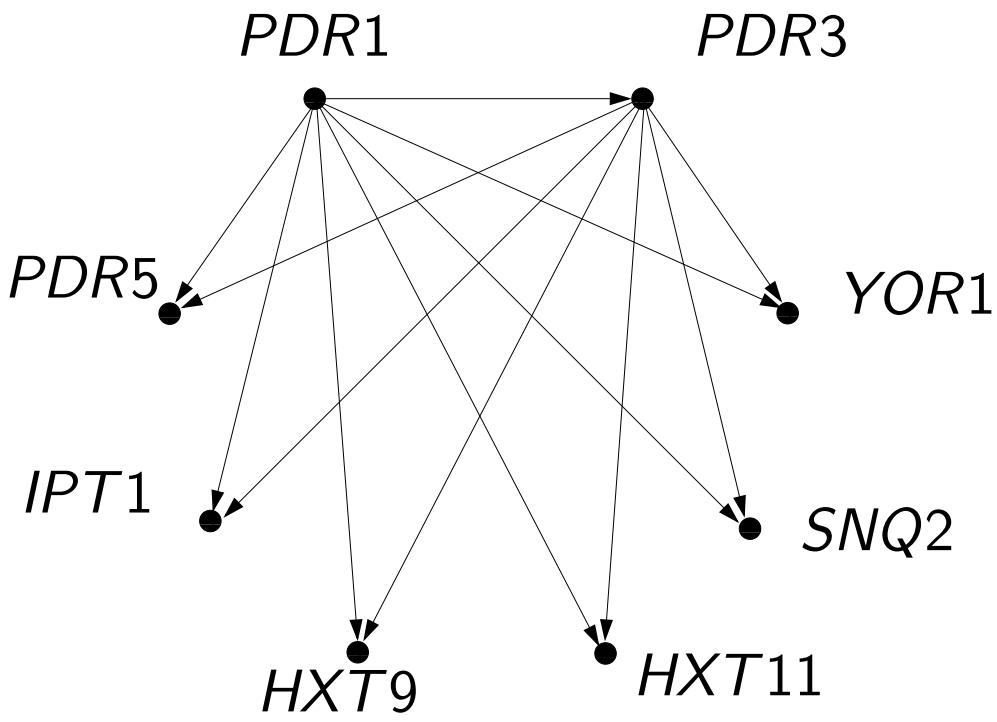
- For all $v \notin U$, let I_U^v be the indicator random variable which is equal to 1 if adding v yields an occurence of $\mathbf{m}$;
- $p_U^v = \mathbb{E}(I_U^v)$
- $N_U(\mathbf{m}) = \sum_{v \notin U} I_U^v$ is the observed order of the theme;
- $\lambda_U = \sum_{v \notin U} p_U^v$ is the expected order of the theme.
- $\Delta_U = \frac{N_U(\mathbf{m}) - \lambda_U}{\lambda_U}$ a normalized quantity



Example







Subnetwork of the Yeast regulation network which is a $(\mathbf{m}'_1, \mathbf{m}_1)$-theme of order 6 and a $(\mathbf{m}'_2, \mathbf{m}_2)$-theme of order 5.

Poisson approximation

$\sum_{v \notin U} I_U^v$ is a sum of independant Bernoulli r.v.'s and can therefore be approximated in total variation distance by a Poisson law of mean λ_U:

$$d_{TV}\left(\mathcal{L}(N_U), \mathcal{L}(Po(\lambda_U))\right) \leq min(1, \lambda_U^{-1}) \sum_{v \notin U} (p_U^v)^2$$

A local statistic

The upper bound approximation is even better for tail probabilities: $\forall K > 2\lambda_U$,

$$\mathbb{P}(N_U \geq K | G[U] \sim \mathbf{m}') \leq \frac{K - \lambda_U}{K - 2\lambda_U} Po(\lambda_U)([K, +\infty))$$

which implies, $\forall t > 1$,

$$\mathbb{P}(\Delta_U \geq t) \leq \mathbb{P}(G[U] \sim \mathbf{m}') \frac{\sqrt{t+1}}{\sqrt{2\pi\lambda_U}(t-1)} e^{-\lambda_U((1+t)\ln(1+t)-t)}$$

Variable change

Problem of multiple testing: order of n^{k-1} positions.

We define

$$g(\lambda, t) = \lambda\big((1+t)\log(1+t) - t\big) - \log\left(\frac{\sqrt{t+1}}{\sqrt{2\pi\lambda_U}(t-1)}\right)$$

$g(\lambda, .)$ is a one-to-one increasing function and goes to infinity like $\lambda t \log(t)$.

$$\begin{aligned} \text{The pattern is a motif} &\Leftrightarrow N_U(\mathbf{m}) >> \lambda_U \\ &\Leftrightarrow \Delta_U >> 1 \\ &\Leftrightarrow g(\lambda_U, \Delta_U) >> 1 \end{aligned}$$

A global statistic

Applying the local theorem to y such that $g(\lambda_U, y) = t$ gives:

$$\mathbb{P}\big(g(\lambda_U, \Delta_U) > t\big) \leq \mathbb{P}(G(U] \sim \mathbf{m}')e^{-t}$$

which yields

Theorem

Let $N(\mathbf{m}')$ the random variable counting the number of occurences of $\mathbf{m}'$. Then, for every $t > 0$,

$$\mathbb{P}\big(\max_U(g(\lambda_U, \Delta_U)) > t\big) \leq \mathbb{E}N(\mathbf{m}')e^{-t} \qquad (1)$$

Tightness of the local bound

Proposition

Let U be a position and define $\lambda_{2,U} = \sum_{v \notin U} (p_U^v)^2$. Let t be such that $\mathbb{P}(N_U(\mathbf{m}) \geq K) = \mathbb{P}(\Delta_U \geq t)$. Denote by $B_U(t)$ the local upper bound.
If $2\lambda_U < K < \frac{\lambda_U}{8\sqrt{\lambda_{2,U}}}$,

$$\frac{\mathbb{P}(N_U(\mathbf{m}) \geq K)}{B_U(t)} \geq \left(1 - 52\frac{K\lambda_{2,U}}{\lambda_U^2}\right)\left(1 - \frac{2\lambda_U}{K}\right)\left(1 - \frac{1}{10K}\right)$$

Tightness of the global bound

Proposition

Let $\mathbf{m}$ *be a pattern admitting a vertex* s *such that* $\mathbf{m} \setminus \{s\}$ *is connected and* s *is a neighbour of every other vertex of* $\mathbf{m}$. *Let* $\rho = \max_{i,j} \pi_{i,j}$ *and suppose that* $\rho = \mathcal{O}(n^{-\frac{1}{2}-\epsilon})$, *with* $\epsilon > \frac{1}{2k}$. *Let* $\delta = min(\epsilon, 2k\epsilon - 1) > 0$. *Then*

$$\mathbb{P}(\max_U(g(\lambda_U, \Delta_U)) > t) = (1-\eta) \sum_U \mathbb{P}(g(\lambda_U, \Delta_U) > t), \qquad \eta =$$

Corollary: For a pattern and a deleted vertex like in the Proposition in an Erdős-Renyi model with linear size growth, the upper bound is asymptotically tight.

Pattern count

ESU Algorithm (Wernicke 05) gives the list of all patterns of a given size in a graph:

- Order the vertices;
- Breadth-first search algorithm based on a set of current vertices and a set of extension vertices;
- Every extension vertex v is added to the currrent set to give a new node. The extension set is updated by keeping only the node sof higher order than v and adding the neighbors of v which see no other current vertex.

That algorithm counts every pattern of size k exactly once and lists their positions.

Motif selection criterion

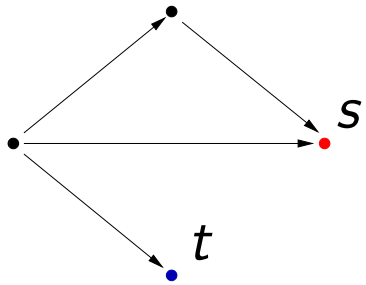
Fix a threshold α.
For every pattern $\mathbf{m}$ of size $\leq k$, every non-disconnecting vertex s of $\mathbf{m}$, do the following steps:

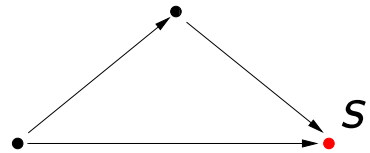
First step Determine if $\mathbf{m}$ is over-represented with respect to $\mathbf{m} \setminus \{s\}$.

Second step If the answer is positive, determine for every non-disconnecting t not linked to s if $\mathbf{m} \setminus \{t\}$ is over-represented with respect to $\mathbf{m} \setminus \{s, t\}$.
$\mathbf{m}$ is a motif with respect to $\mathbf{m} \setminus \{s\}$ if the answer to the the second question is negative for all t.

Example

Example in the Yeast interaction network.





$2.1e-10$ $3.2e-12$

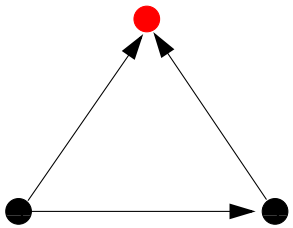
The feed-forward loop beeing over-represented with respect to s, the first pattern is not a local motif.

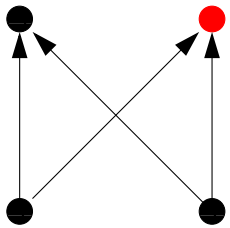
Simulated exemple

Graph sample:

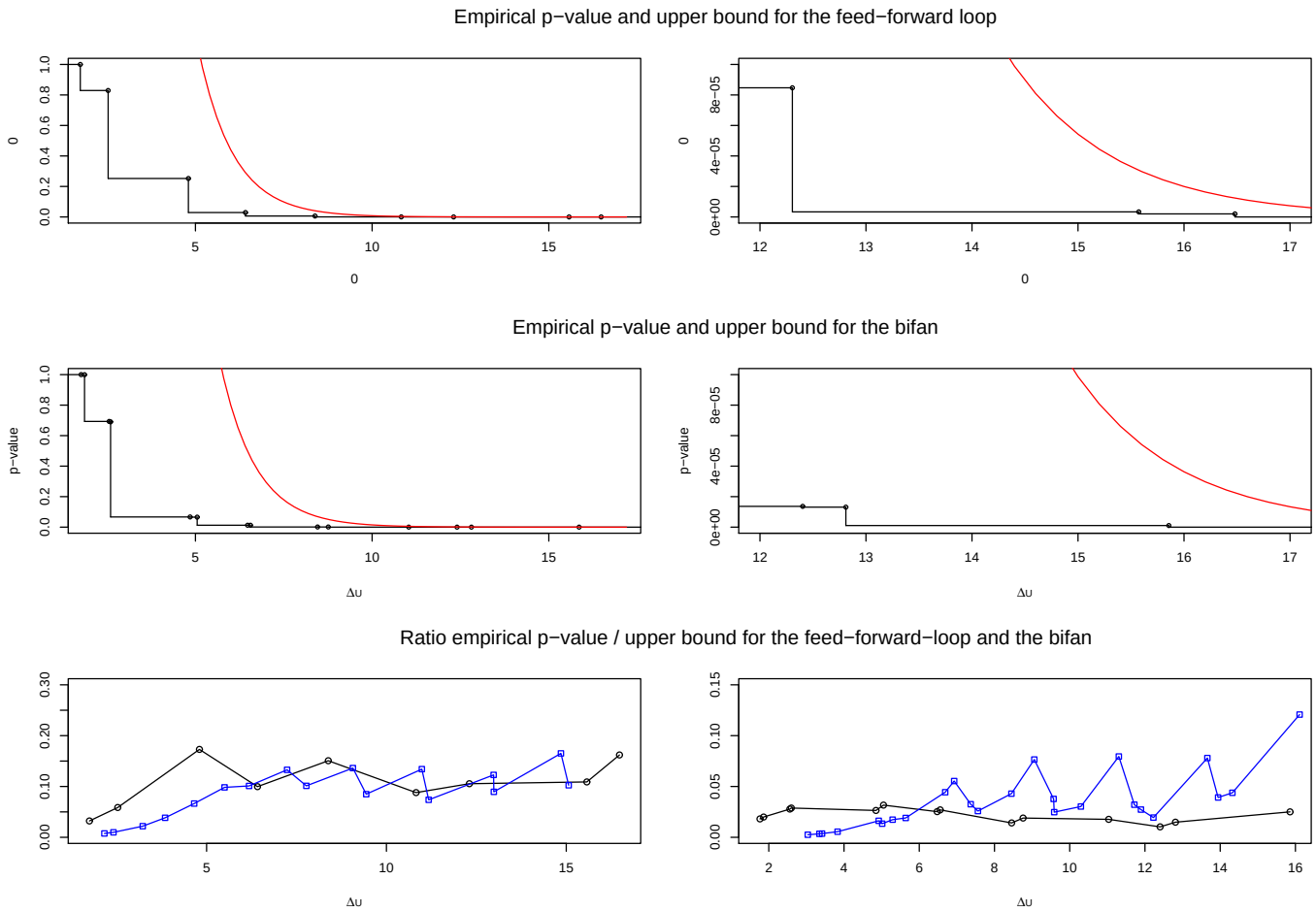
- 2 millions graphs generated using three vertex classes of 30 vertices each and respective probabilities of connection .05 and .01 depending on whether the vertices belong to the same class or not.
- 30000 more dense graphs with connection probabilities .5 and .1.

Patterns: Feed-forward loop and bifan:





Simulated example



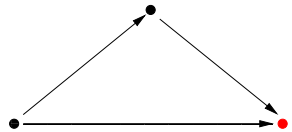
Yeast data

Transcriptional regulatory network available at U. Alon's lab webpage: 690 genes and 1078 interactions.

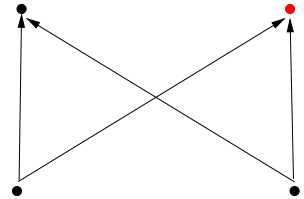
Bayesian estimation of the parameters for a mixture model (Latouche et al., 2008).

Motifs of size 3, 4, 5

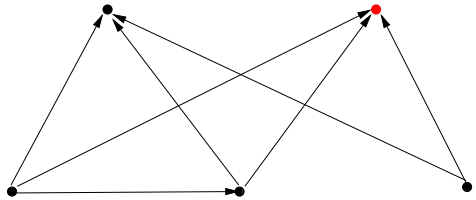
Motifs found by our method with $\alpha = 1e-5$.



$3.2e-12$



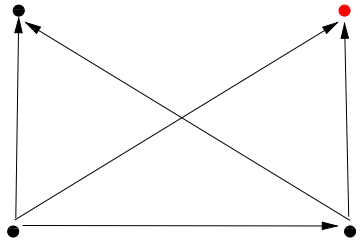
$3.0e-26$

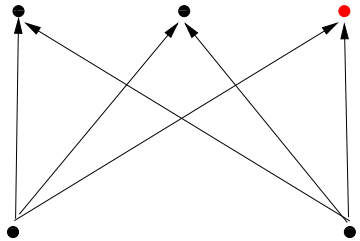


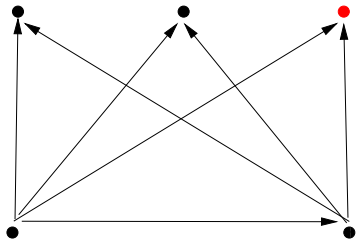
$6.5e-13$

Motifs of size 3, 4, 5

Motifs found by Berg et al. and Alon et al. but not selected by our method.







Conclusion

- New definition of a *motif*: a motif is over-represented with respect to a submotif.
- No false positives but is the procedure to stringent?
- The known relevant motifs in the *Yeast* regulation network are found.

Perspectives

- Deeper biological applications,
- Are bigger motifs meaningfull objects?,
- Network comparisons using the local score lists?