

Detection of network motifs by local concentration

Etienne Birmelé

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

1 Context

2 Local Statistics

3 A global statistic

4 Motif detection procedure

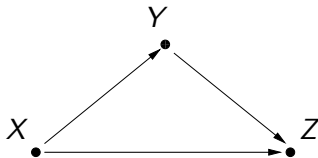
5 Application to Yeast

6 Conclusion

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.

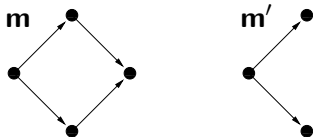
Leading ideas

- 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).
- $Z = f(X_1, \dots, X_n)$ is highly concentrated around its mean when the X_i 's are independent and changing the value of one of them does affect Z by less than a constant.

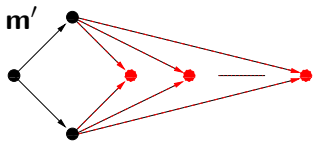
Changing the definition of a motif

Etienne Birmelé

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



\mathbf{m} is a *motif* with respect to \mathbf{m}' if there exist an occurrence of \mathbf{m}' in the network which has a surprisingly high number of extensions to occurrences of \mathbf{m} .



Context

Local Statistics

A global statistic

Motif detection procedure

Application to Yeast

Conclusion

Random graph model

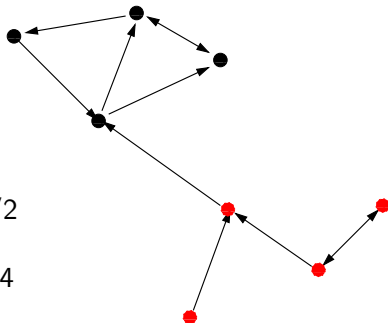
We fix the number n of nodes and the underlying random graph model is defined by a $n \times n$ matrix C : the edge indicators $(X_{ij})_{1 \leq i, j \leq n}$ are independent Bernoulli variables and

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

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



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

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

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

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

Context

Local Statistics

A global statistic

Motif detection procedure

Application to Yeast

Conclusion

① Context

② Local Statistics

③ A global statistic

④ Motif detection procedure

⑤ Application to Yeast

⑥ Conclusion

Notations

Let \mathbf{m} be a small graph on k vertices (r_1, \dots, r_{k-1}, s) and \mathbf{m}' the subgraph obtained by deleting s .

Let $U = (u_1, \dots, u_{k-1})$ be an ordered set of $k - 1$ vertices.

We define:

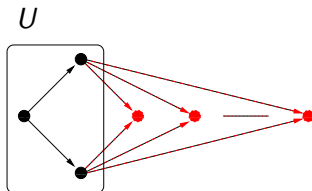
- $N_U(\mathbf{m})$ the number of occurrences of \mathbf{m} which restriction to U is isomorphic to \mathbf{m}' ;
- $Y_U(\mathbf{m}') = \mathbb{I}_{G[U] \sim \mathbf{m}'}$
- $ext_U^v(\mathbf{m}', \mathbf{m}) = 1 \Leftrightarrow \forall i, X_{u_i v} = e_{r_i s}$
 $ext_U^v = 1$ if adding the vertex v yields an occurrence of \mathbf{m} .
- $\lambda_U = \mathbb{E}(\sum_{v \notin U} ext_U^v)$ the mean number of valid extensions.

Notations

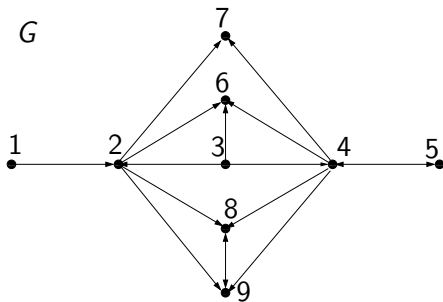
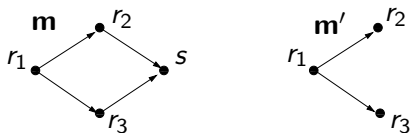
Then

$$N_U(\mathbf{m}) = Y_U(\mathbf{m}') \sum_{v \notin U} \text{ext}_U^v(\mathbf{m}', \mathbf{m})$$

and Y_U and ext_U^v are independent.



Example



For $U = (3, 2, 4)$, $Y_U(m') = 1$ and $N_U(m) = 3$.

Poisson approximation

$\sum_{v \notin U} ext_U^v$ 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 :

$$\forall A \subset \mathbb{Z}^+,$$

$$|\mathbb{P}(N_U(\mathbf{m}) \in A | Y_U(\mathbf{m}')) - Po(\lambda_U)(A)| \leq \min(1, 1/\lambda_U) \sum_v p_v^2$$

with $p_v = \mathbb{P}(ext_U^v = 1)$.

In practice, p_v 's are small and that bound is quite sharp (between $1.8e - 9$ and $5.0e - 3$ for the different positions of the feed-forward loop in the Yeast regulatory network)

A local statistic

The upper bound approximation is even better for tail probabilities:

If $t = \frac{m - \lambda_U}{\lambda_U} > 1$,

$$\begin{aligned}\mathbb{P}(N_U(\mathbf{m}) \geq m | Y_U(\mathbf{m}')) &\leq \frac{t}{t-1} \text{Po}(\lambda_U)([m, +\infty)) \\ &\leq \frac{t+1}{t-1} \text{Po}(\lambda_U)(m)\end{aligned}$$

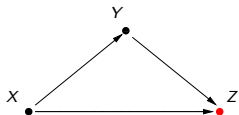
which implies

$$\mathbb{P}\left(\frac{N_U(\mathbf{m}) - \lambda_U}{\lambda_U} > t\right) \leq \mathbb{P}(Y_U(\mathbf{m}') = 1) \frac{t+1}{\sqrt{2\pi}(t-1)} e^{-((1+t)\ln(1+t))}$$

Simulated exemple

$1e7$ graphs were generated using three vertex classes of 100 vertices each and respective probabilities of connection .25 and .05 depending on whether the vertices belonged to the same class or not.

The pattern \mathbf{m} is the feed-forward loop and the vertex deleted to obtain \mathbf{m}' is Z .

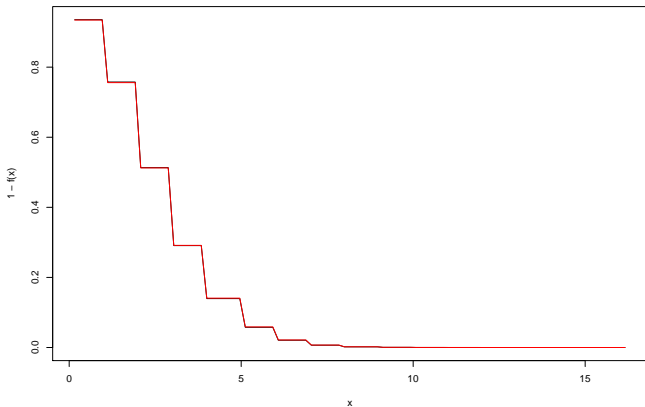


The position U contains one vertex of class 1 and one vertex of class 2. The mean number of extensions is $\lambda_U = 2.725$

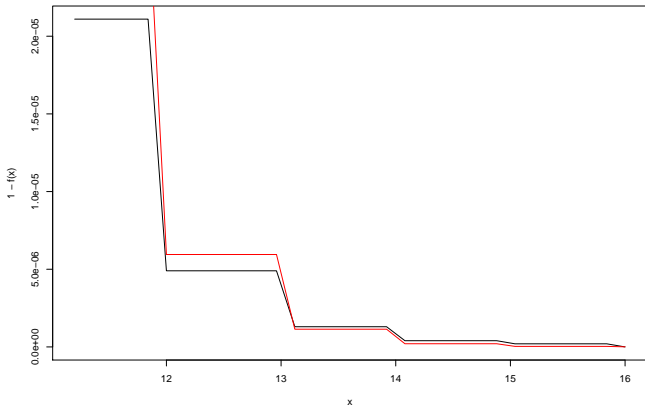
Simulated exemple

Black: empirical p-values

Red: $Po(\lambda)$ p-values



Zoom on large deviations



Refining the bounds for moderate deviations

Theorem

Let $\eta = \sum_v p_v^2$ and $\phi = \frac{\eta}{\lambda_U} + \eta(1+t)^2$.

If $\lambda_U + \sqrt{\lambda_U} \leq m \leq \lambda_U^2/\eta$,

$$1 - 15\phi \leq \frac{\mathbb{P}(N_U(\mathbf{m}) \geq m)}{\text{Po}(\lambda)([m, +\infty))} \leq 1 + 15\phi$$

Example: ER model with $p = \frac{\epsilon}{n}$. For $1 \leq m \leq \sqrt{15\alpha n}$,

$$(1 - \alpha)c_m \leq \frac{\mathbb{P}(N_U(\mathbf{m}) \geq m)}{\text{Po}(\lambda)([m, +\infty))} \leq 1 + \alpha$$

with $\lim_{m \rightarrow \infty} c_m = 1$.

1 Context

2 Local Statistics

3 A global statistic

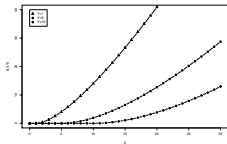
4 Motif detection procedure

5 Application to Yeast

6 Conclusion

A nice function

$$h(X, Y) = \begin{cases} 0 & \text{if } X \leq Y \\ X \ln\left(\frac{X}{eY}\right) + Y & \text{else} \end{cases}$$



At fixed Y , $h_Y : X \rightarrow h(X, Y)$ is increasing, asymptotically equivalent to $X \ln(X)$.

A global statistic

The local inequality can be rewritten as:

$$\forall t > 0, \mathbb{P}(h(N_U(\mathbf{m}), \lambda_U) > t) \leq \mathbb{P}(G[U] \sim \mathbf{m}')e^{-t}$$

As

$$\mathbb{P}\left(\max_U(h(N_U(\mathbf{m}), \lambda_U)) > t\right) \leq \sum_U \mathbb{P}(h(N_U(\mathbf{m}), \lambda_U) > t),$$

Theorem

$$\mathbb{P}\left(\max_U(h(N_U(\mathbf{m}), \lambda_U)) > t\right) \leq aut(\mathbf{m}')\mathbb{E}N_U(\mathbf{m}')e^{-t}$$

An user-friendly corollary

Corollary

$$\mathbb{P}(\exists U/N_U(\mathbf{m}) > e^2 \lambda_U + t) \leq \text{aut}(\mathbf{m}') \mathbb{E} N_U(\mathbf{m}') e^{-t}$$

That is:

*The probability that there exist **any** occurrence of \mathbf{m}' in the network which has a surprisingly high number of extensions to \mathbf{m} decreases exponentially.*

- 1 Context
- 2 Local Statistics
- 3 A global statistic
- 4 Motif detection procedure**
- 5 Application to Yeast
- 6 Conclusion

Motif selection criterion

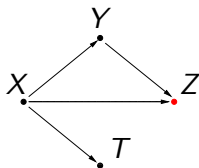
Fix a threshold α .

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

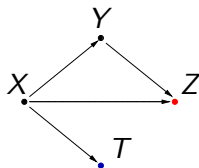
First step Determine if \mathbf{m} is over-represented with respect to $\mathbf{m} \setminus \{s\}$.

Second step If the answer is positive, determine for every non-disconnecting t distinct from 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



$2.1e - 10$



2.4

The feed-forward loop being over-represented with respect to Z, the pattern is not a motif.

- 1 Context
- 2 Local Statistics
- 3 A global statistic
- 4 Motif detection procedure
- 5 Application to Yeast**
- 6 Conclusion

Data

Transcriptional regulatory network available at U. Alon's lab webpage: 680 genes and 1078 interactions.

Bayesian estimation of the parameters for a mixture model for graphs gives 7 vertex classes, of respective sizes 11,13,29,78,87,88 and 384.

Motifs of size 3, 4, 5

Context

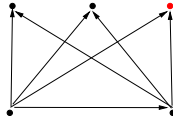
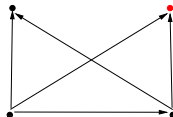
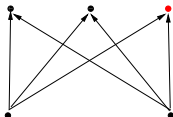
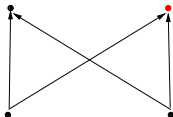
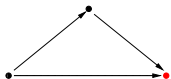
Local Statistics

A global statistic

Motif detection procedure

Application to Yeast

Conclusion



- 1 Context
- 2 Local Statistics
- 3 A global statistic
- 4 Motif detection procedure
- 5 Application to Yeast
- 6 Conclusion

Conclusion

- New definition of a *motif*: a motif is over-represented with respect to a submotif.
- Fast algorithm.
- The known relevant motifs in the *Yeast* regulation network are found.

Perspectives

- Lower bounds,
- Deeper biological applications,
- Network comparisons using the local score lists.