Detection of network motifs by local concentration

Etienne Birmelé

Laboratoire Statistique et Génome, Université d'Evry
Groupe SSB - ANR NeMo
Detection of network motifs by local concentration

Etienne Birmelé

1. Context

2. Local and global statistics

3. Motif detection procedure

4. Simulation and application to Yeast

5. 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

\[ \begin{array}{c}
& & Y \\
X & \rightarrow & \rightarrow \\
& & Z \\
\end{array} \]
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
Leading ideas

- A small graph $m$ may be over-represented because one of its subgraphs $m'$ is over-represented. In that case, $m'$ 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 pattern $m$ and a subgraph $m'$ of $m$ obtained by the deletion of a vertex in $m$.

$m$ is a local motif with respect to $m'$ 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.
Detection of network motifs by local concentration

Etienne Birmélé

Context
Local and global statistics

Motif detection procedure

Simulation and application to Yeast

Conclusion

Random graph model

\[ P(\text{NN}) = 1/2 \]

\[ P(\text{RR}) = 1/4 \]

\[ P(\text{NR}) = 0 \]

\[ P(\text{RN}) = 1/16 \]
Detection of network motifs by local concentration

Etienne Birmelé

1. Context

2. Local and global statistics

3. Motif detection procedure

4. Simulation and application to Yeast

5. Conclusion
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 occurrence of $\mathbf{m}'$.

- For all $v \notin U$, let $I^V_U$ be the indicator random variable which is equal to 1 if adding $v$ yields an occurrence of $\mathbf{m}$;
- $p^V_U = \mathbb{E}(I^V_U)$
- $N_U(\mathbf{m}) = \sum_{v \notin U} I^V_U$ is the observed order of the theme;
- $\lambda_U = \sum_{v \notin U} p^V_U$ is the expected order of the theme.
- $\Delta_U = \frac{N_U(\mathbf{m}) - \lambda_U}{\lambda_U}$ a normalized quantity
Subnetwork of the Yeast regulation network which is a $(m'_1, m_1)$-theme of order 6 and a $(m'_2, m_2)$-theme of order 5.
Poisson approximation

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

$$d_{TV}(\mathcal{L}(N_U), \mathcal{L}(Po(\lambda_U))) \leq \min(1, \lambda_U^{-1}) \sum_{v \notin U} (p_U^v)^2$$
The upper bound approximation is even better for tail probabilities: \( \forall K > 2\lambda_U, \)

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

which implies, \( \forall t > 1, \)

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

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

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

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

The pattern is a motif $\iff N_U(m) \gg \lambda_U$
$\iff \Delta_U \gg 1$
$\iff g(\lambda_U, \Delta_U) \gg 1$
A global statistic

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

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

which yields

**Theorem**

Let $N(m')$ the random variable counting the number of occurrences of $m'$. Then, for every $t > 0$,

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

(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 $\mathbb{P}(N_U(m) \geq K) = \mathbb{P}(\Delta_U \geq t)$. Denote by $B_U(t)$ the local upper bound.

If $2\lambda_U < K < \frac{\lambda_U}{8\sqrt{\lambda_{2,U}}}$,

$$\frac{\mathbb{P}(N_U(m) \geq K)}{B_U(t)} \geq (1 - 52\frac{K\lambda_{2,U}}{\lambda_U^2})(1 - \frac{2\lambda_U}{K})(1 - \frac{1}{10K})$$
Tightness of the global bound

Proposition

Let \( m \) be a pattern admitting a vertex \( s \) such that \( m \setminus \{s\} \) is connected and \( s \) is a neighbour of every other vertex of \( m \). 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

\[
P\left( \max_{U} \left( g(\lambda_U, \Delta_U) > t \right) \right) = (1 - \eta) \sum_{U} P\left( g(\lambda_U, \Delta_U) > t \right), \quad \eta = \text{correction factor}.
\]

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

Local and global statistics

Motif detection procedure

Simulation and application to Yeast

Conclusion
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 current set to give a new node. The extension set is updated by keeping only the node of 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 $\alpha$.

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

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

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

$m$ is a motif with respect to $m \setminus \{s\}$ if the answer to the the second question is negative for all $t$. 
Example

Example in the Yeast interaction network.

![Diagram](image)

2.1e \(-10\)  
3.2e \(-12\)

The feed-forward loop being over-represented with respect to s, the first pattern is not a local motif.
Detection of network motifs by local concentration

Etienne Birmelé

1. Context
2. Local and global statistics
3. Motif detection procedure
4. Simulation and application to Yeast
5. Conclusion
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

Detection of network motifs by local concentration

Etienne Birmelé

Context
Local and global statistics
Motif detection procedure
Simulation and application to Yeast
Conclusion

Empirical p-value and upper bound for the feed-forward loop

Empirical p-value and upper bound for the bifan

Ratio empirical p-value / upper bound for the feed-forward-loop and the bifan
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 $\alpha = 1e - 5$.

- $3.2e - 12$ (smaller)
- $3.0e - 26$ (larger)
- $6.5e - 13$ (medium)
Motifs of size 3, 4, 5

Motifs found by Berg et al. and Alon et al. but not selected by our method.
Detection of network motifs by local concentration

Etienne Birmelé

1. Context
2. Local and global statistics
3. Motif detection procedure
4. Simulation and application to Yeast
5. Conclusion
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 meaningful objects?,
- Network comparisons using the local score lists?