Assessing the exceptionality of network motifs

F. Picard[⋆,∗], J-J. Daudin[†], M. Koskas[†], S. Schbath [‡], S. Robin[†].

[⋆] UMR CNRS-8071/INRA-1152, Statistique et Génome, Évry,

[∗] UMR CNRS-5558, Laboratoire de Biométrie et Biologie Evolutive,

[‡] Mathématique, Informatique et Génome, Jouy-en-Josas,

[†] UMR INAPG/ENGREF/INRA MIA 518, Paris.

Statistics for Systems Biology (SSB) group

INRA-MIG	E. Roquain, S. Schbath,
Stat. et Génome-Évry	E. Birmelé, C. Matias, V. Miele.

- **Breaking-down complex networks into functional modules**:

  $\longrightarrow$ patterns of interconnection,

  $\longrightarrow$ **network motifs**.

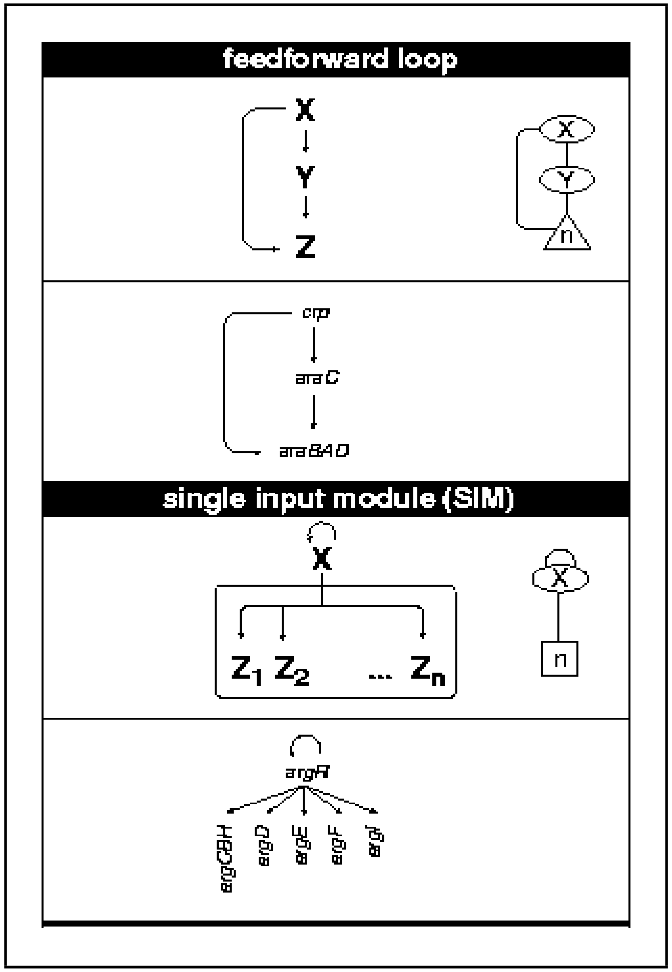
- **Application in Biology**:

  $\longrightarrow$ transcriptional regulatory modules

  $\longrightarrow$ Example: feed-forward loop.

- **Exceptionality of a motif?**

  $\longrightarrow$ when a given motif appears more frequently than **expected**.



From Shen-Orr et al.(2002)

How to assess the exceptionality of a motif

- Count the observed number $N_{\text{obs}}(\mathbf{m})$ of a given motif $\mathbf{m}$ (out of our scope)
- Assess its significance with a $p-$value : need to calculate $\mathbb{P}\{N(\mathbf{m}) \geq N_{\text{obs}}(\mathbf{m})\}$
- Current strategy (Shen-Orr et al.):

  $\longrightarrow$ use simulations to calculate $\mathbb{E}(N)$ and $\mathbb{V}(N)$ under a reference model

  $\longrightarrow$ use a $Z-$score to calculate the $p-$value (implies a Gaussian approximation).

CONTRIBUTION

1. Give an analytic expression of the mean and the variance of the count,
2. Propose another distribution to better approximate the count distribution.

Definition of a motif in random graphs

- A random graph is defined by :

  $\longrightarrow$ $\mathcal{V}$ of fixed vertices with $|\mathcal{V}| = n$.

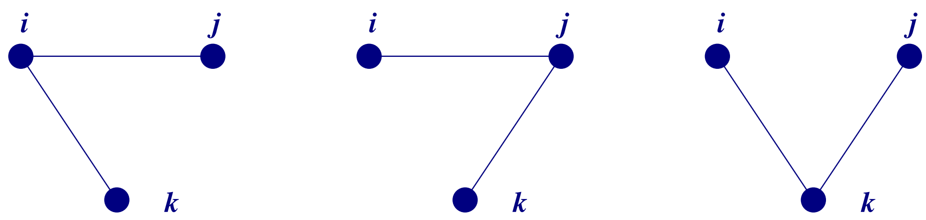
  $\longrightarrow$ $\mathbf{X} = \{X_{ij}, (i,j) \in \mathcal{V}^2\}$ a set of random edges such that X_{ij} equals 1 if nodes i and j are connected, and 0 otherwise.

  $\longrightarrow$ A distribution on X_{ij}. Example: the Erdös-Rényi model: $\mathbb{P}(X_{ij} = 1) = p$.
- **exchangeability hypothesis**: $\mathbb{P}(X_{ij})$ does not depend on (i,j).
- $\mathbf{m}$ stands for a motif of size k: connected subgraph with k vertices,
- It is defined by a fixed topology through its adjacency matrix also denoted by $\mathbf{m}$ such that $\mathbf{m}_{uv} = 1$, if nodes u, v are connected in the motif

Example: the V motif

- 3 versions of the V motif at a **fixed** position $\alpha = (i, j, k)$.

$\mathbf{m}$	$\mathbf{m}'$	$\mathbf{m}''$
$\begin{bmatrix} 0 & 1 & 1 \\ . & 0 & 0 \\ . & . & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 1 & 0 \\ . & 0 & 1 \\ . & . & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 & 1 \\ . & 0 & 1 \\ . & . & 0 \end{bmatrix}$



Position and occurrence of a motif

- Let α be a possible position of $\mathbf{m}$. We consider that α is an ordered $k-$tuple with $i_1 < \ldots < i_k$.
- We introduce the random indicator variable $Y_\alpha(\mathbf{m})$ which equals one if motif $\mathbf{m}$ **occurs at position** α and 0 otherwise :

$$Y_\alpha(\mathbf{m}) = \prod_{1\leq u<v\leq k} \left(X_{i_u i_v}\right)^{m_{uv}}.$$

- Under the exchangeability assumption, the distribution of Y_α does not depend on α. Denoting $\mu(\mathbf{m})$ the **probability of occurrence** of motif $\mathbf{m}$, we have

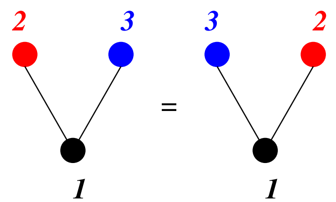
$$Y_\alpha(\mathbf{m}) \sim \mathcal{B}(\mu(\mathbf{m}))$$

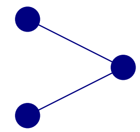
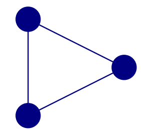
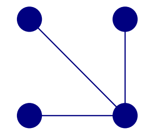
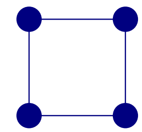
- The number of occurrences of $\mathbf{m}$ is then $N(\mathbf{m}) = \sum_{\alpha\in I_k} \sum_{?} Y_\alpha(\mathbf{m})$.

Redundancy and Motif permutation

- For a given position, permutations of vertices of $\mathbf{m}$ can lead to the same motif

$\mathsf{aut}(\mathsf{V}) = \{\mathsf{Id}, (2,3)\}$



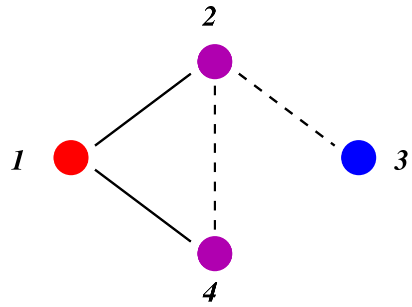
				
$\lvert\mathsf{aut}(\mathbf{m})\rvert$	2	6	6	8
$\rho(\mathbf{m})$	3	1	4	3

- We define $\mathcal{R}(\mathbf{m})$, the set of non redundant permutations of $\mathbf{m}$, $\rho(\mathbf{m}) = |\mathcal{R}(\mathbf{m})|$.
- $\rho(\mathbf{m})$ is calculated with the $k!$ simultaneous permutations of the rows and columns of $\mathbf{m}$.
- The count of motif $\mathbf{m}$ is: $N(\mathbf{m}) = \sum_{\alpha \in I_k} \sum_{\mathbf{m}' \in \mathcal{R}(\mathbf{m})} Y_\alpha(\mathbf{m}')$.
- We aim at calculating the mean of the count

$$\mathbb{E}N(\mathbf{m}) = |I_k| \times \sum_{\mathbf{m}' \in \mathcal{R}(\mathbf{m})} \mathbb{E}Y_\alpha(\mathbf{m}') = \binom{n}{k} \rho(\mathbf{m}) \mu(\mathbf{m}).$$

Calculating the variance is more intricate...

$$N^2(\mathbf{m}) = \sum_{\alpha,\beta \in I_k} \sum_{\mathbf{m}',\mathbf{m}'' \in \mathcal{R}(\mathbf{m})} Y_\alpha(\mathbf{m}') Y_\beta(\mathbf{m}'')$$



- $\mathbf{m}$: V motif
- $\mathbf{m}'$ occurs at $\alpha = (1,2,4)$, $\mathbf{m}''$ occurs at $\beta = (2,3,4)$,
- In this case $\alpha \cap \beta = (2,4)$
- The **super-motif** denoted by $\mathbf{m}' \underset{s}{\Omega} \mathbf{m}''$ is the union of two versions of $\mathbf{m}$

  $\longrightarrow$ In this case, the super-motif is the so-called whisk graph motif
- We need to define:

  $\longrightarrow$ the adjacency matrix of the super-motif $\mathbf{m}' \underset{s}{\Omega} \mathbf{m}''$

  $\longrightarrow$ the non-redundant permutations of $\mathbf{m}' \underset{s}{\Omega} \mathbf{m}''$.

Adjacency matrix for super-motifs

- we break down $\mathbf{m}'$ and $\mathbf{m}''$ such that:

$$\mathbf{m}' = \left(\begin{array}{c|c} \underset{(k-s)\times(k-s)}{\mathbf{m}'_{11}} & \underset{(k-s)\times s}{\mathbf{m}'_{12}} \\ \hline \underset{s\times(k-s)}{\mathbf{m}'_{21}} & \underset{s\times s}{\mathbf{m}'_{22}} \end{array}\right), \qquad \mathbf{m}'' = \left(\begin{array}{c|c} \underset{s\times s}{\mathbf{m}''_{11}} & \underset{s\times(k-s)}{\mathbf{m}''_{12}} \\ \hline \underset{(k-s)\times s}{\mathbf{m}''_{21}} & \underset{(k-s)\times(k-s)}{\mathbf{m}''_{22}} \end{array}\right),$$

where $\mathbf{m}'_{22}$ and $\mathbf{m}''_{11}$ correspond to vertices in $\alpha \cap \beta$,

- We set:

$$\mathbf{m}' \underset{s}{\Omega} \mathbf{m}'' = \left(\begin{array}{c|c|c} \mathbf{m}'_{11} & \mathbf{m}'_{12} & \mathbf{0} \\ \hline \mathbf{m}'_{21} & \max(\mathbf{m}'_{22}, \mathbf{m}''_{11}) & \mathbf{m}''_{12} \\ \hline \mathbf{0} & \mathbf{m}''_{21} & \mathbf{m}''_{22} \end{array}\right).$$

- The $\max$ function in the central term indicates that for the s common vertices of α and β, all edges of $\mathbf{m}'_{22}$ and $\mathbf{m}''_{11}$ must be present. It is equivalent to the logical `OR`.

New formulation for the squared count

- Each term of the sum depends on s, the number of shared vertices between α and β
- If $s = 0$, Y_α and Y_β are independent and $\mathbb{E}\left[Y_\alpha(\mathbf{m})Y_\beta(\mathbf{m})\right] = \mathbb{E}Y_\alpha(\mathbf{m})\mathbb{E}Y_\beta(\mathbf{m})$
- $\forall s \geq 1, \; Y_\alpha(\mathbf{m}')Y_\beta(\mathbf{m}'') \quad = \quad Y_{\alpha\cup\beta}(\mathbf{m}' \underset{s}{\Omega} \mathbf{m}'').$
- The squared count can be rewritten as:

$$N^2(\mathbf{m}) \quad = \quad \sum_{s=0}^{k} \sum_{\substack{\alpha, \beta \in I_k : \\ |\alpha \cap \beta| = s}} \sum_{\mathbf{m}', \mathbf{m}'' \in \mathcal{R}(\mathbf{m})} Y_{\alpha\cup\beta}(\mathbf{m}' \underset{s}{\Omega} \mathbf{m}''),$$

- The expectation of the squared count is:

$$\mathbb{E}N^2(\mathbf{m}) \quad = \quad C_1(n,k)\left[\sum_{\mathbf{m}' \in \mathcal{R}(\mathbf{m})} \mu(\mathbf{m}')\right]^2 + \sum_{s=1}^{k} C_2(n,k,s) \sum_{\mathbf{m}', \mathbf{m}'' \in \mathcal{R}(\mathbf{m})} \mu(\mathbf{m}' \underset{s}{\Omega} \mathbf{m}'').$$

Calculating $\mu(\mathbf{m})$ in ER and ER mixture models

- $\mathbb{E}(N)$ and $\mathbb{V}(N)$ depend on $\mu(\mathbf{m})$ which depend on $\mathbb{P}\{\mathbf{X}\}$ (exchangeable)
- In the Erdös-Rényi model with parameter π, (m_{++} the number of edges in $\mathbf{m}$):

$$\mu_{\mathsf{ER}}(\mathbf{m}) = \pi^{m_{++}}$$

- ERMG is an alternative model. We suppose that nodes are spread among Q hidden classes with proportion $\alpha_1, \dots, \alpha_Q$.

  $\longrightarrow$ We denote by Z_is the independent random variables which equal q if node i belongs to class q, then $X_{ij} | \{Z_i = q, Z_j = \ell\} \sim \mathcal{B}(\pi_{q\ell})$.

  $\longrightarrow$ Under ERMG, we have:

$$\mu_{\mathsf{ERMG}}(\mathbf{m}) = \sum_{c_1=1}^{Q} \dots \sum_{c_k=1}^{Q} \alpha_{c_1} \dots \alpha_{c_k} \prod_{1 \le u < v \le k} \pi_{c_u c_v}^{m_{uv}}.$$

Calculating $\mu(\mathbf{m})$ for Expected-Degree Distribution models

- EDD generates graphes whose degrees follow a given distribution,

$$\mathbb{P}\{X_{ij} = 1 | D_i D_j\} = \gamma D_i D_j$$

- "exchangeable" version of the Fixed Degree Distribution model (FDD),
- $\mu(\mathbf{m})$ can be calculated

$$\mu_{\mathsf{EDD}}(\mathbf{m}) = \gamma^{m_{++}/2} \prod_{u=1}^{k} \mathbb{E}\left(D_{iu}^{m_{u+}}\right).$$

- $\mu_{\mathsf{EDD}}(\mathbf{m})$ only depends on the product of some moments of the expected degree D.

Comparison of theoretical moments on PPI networks

			$\mathbb{E}N(\mathbf{m})$		$\widehat{\mathbb{E}N(\mathbf{m})}$
Ecoli	N_{obs}	ER	EDD	ERMG	FDD
V	248,093	52,744.70	99,126.40	243,846.93	248,093
triangle	11,368	72.47	2,197.38	10,221.17	3,579.49
chain	9,557,956	399,151.00	2,339,200.00	9,555,414.55	5,950,903.40
star	6,425,495	133,050.00	1,537,740.00	5,772,005.15	6,425,495
square	487,408	411.31	38,890.60	417,190.55	76,467.39
whisker	2,154,048	1,645.22	306,789.00	1,929,516.68	547,802.44
halfclique	273,621	3.39	20,117.90	204,093.45	18,422.25
clique	14,882	0.00	867.24	8,904.75	317.27

First remark: the choice of the model has a strong influence on the first two moments.

This influence **depends on the topology of the motif**.

Comparison of theoretical moments on PPI networks

		$\mathbb{E}N(\mathbf{m})$			$\widehat{\mathbb{E}N(\mathbf{m})}$
Ecoli	N_{obs}	ER	EDD	ERMG	FDD
V	**248,093**	52,744.70	99,126.40	243,846.93	248,093
triangle	11,368	72.47	2,197.38	10,221.17	3,579.49
chain	9,557,956	399,151.00	2,339,200.00	9,555,414.55	5,950,903.40
star	**6,425,495**	133,050.00	1,537,740.00	5,772,005.15	6,425,495
square	487,408	411.31	38,890.60	417,190.55	76,467.39
whisker	2,154,048	1,645.22	306,789.00	1,929,516.68	547,802.44
halfclique	273,621	3.39	20,117.90	204,093.45	18,422.25
clique	14,882	0.00	867.24	8,904.75	317.27

- The expected count of **V** and **star** under ER and EDD are far from the observed count.
- Due to observed nodes with high degree which generate lots of occurrences of those motifs.

Comparison of theoretical moments on PPI networks

		$\mathbb{E}N(\mathbf{m})$			$\widehat{\mathbb{E}N(\mathbf{m})}$
Ecoli	N_{obs}	ER	EDD	ERMG	FDD
V	248,093	52,744.70	99,126.40	243,846.93	248,093
triangle	**11,368**	72.47	2,197.38	10,221.17	3,579.49
chain	9,557,956	399,151.00	2,339,200.00	9,555,414.55	5,950,903.40
star	6,425,495	133,050.00	1,537,740.00	5,772,005.15	6,425,495
square	487,408	411.31	38,890.60	417,190.55	76,467.39
whisker	2,154,048	1,645.22	306,789.00	1,929,516.68	547,802.44
halfclique	**273,621**	3.39	20,117.90	204,093.45	18,422.25
clique	**14,882**	0.00	867.24	8,904.75	317.27

- The expected count under ERMG for **triangle**, **halfclique** and **clique** are close to the **observed** count
- Those motifs are linked to local clustering trends which are well captures by ERMG.

Comparison of theoretical moments on PPI networks

			$\sqrt{\mathbb{V}N(\mathbf{m})}$		$\sqrt{\widehat{\mathbb{V}N(\mathbf{m})}}$
Ecoli	N_{obs}	ER	EDD-E	ERMG	FDD
V	248093	1281.87	20851.70	51676.68	0
triangle	11368	8.90	797.30	3041.98	68.58
chain	9557956	14743.70	774109.00	3019630.93	67739.86
star	6425495	5089.62	484152.00	1672086.51	0
square	487408	29.14	19122.60	170502.21	1117.56
whisker	2154048	214.52	145764.00	739836.65	15593.00
halfclique	273621	2.04	12876.60	94018.80	891.99
clique	14882	0.05	707.94	4660.71	32.96

- When using the Fixed Degree Distribution model, the variance is systematically smaller
- Extreme case for the **V** and **star** motifs for which the degree exactly defines the number of occurrences

Compound Poisson approximation

- Exceptionality is assessed with $\mathbb{P}\{N(\mathbf{m}) \geq N_{\text{obs}}(\mathbf{m})\}$, where $N(\mathbf{m})$ the random number of occurrence of $\mathbf{m}$ under the reference model.
- network motifs tend to overlap : clumps are present in the graph and C stand for the number of clumps (random)
- Denoting by S_i the size of clump i, we have $N(\mathbf{m}) = \sum_{i=1}^{C} S_i(\mathbf{m})$.
- If we make the hypothesis that $C \sim \mathcal{P}(\lambda)$, $N(\mathbf{m})$ is compound Poisson

  $\longrightarrow$ We use the **Geometric-Poisson** distribution : we suppose that $S_i(\mathbf{m}) \approx \mathcal{G}(1-a)$.

  $\longrightarrow$ Then we approximate the distribution of $N(\mathbf{m}) \approx \mathcal{CP}(\lambda, a)$.

  $\longrightarrow$ Parameters (λ, a) can be calculated according to $\mathbb{E}N(\mathbf{m})$ and $\mathbb{V}N(\mathbf{m})$

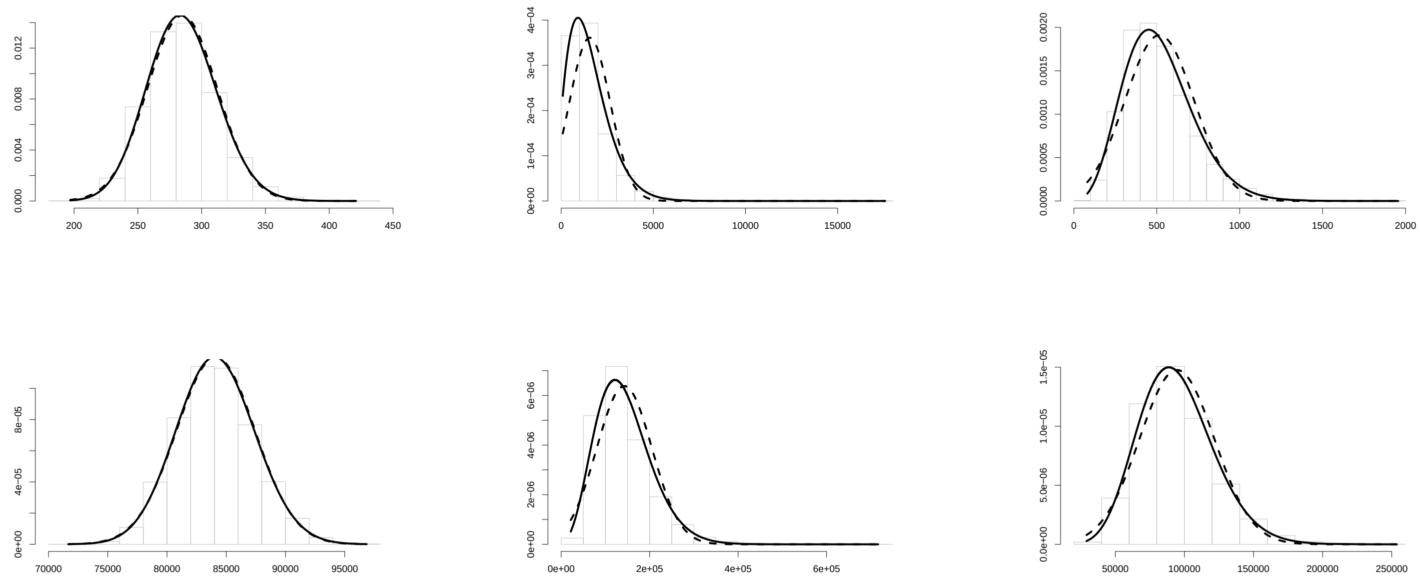
$$a = [\mathbb{E}N(\mathbf{m}) - \mathbb{V}N(\mathbf{m})]/[\mathbb{E}N(\mathbf{m}) + \mathbb{V}N(\mathbf{m})], \quad \lambda = (1-a)\mathbb{E}N(\mathbf{m}).$$

Simulated count distributions

	FDD	EDD	ERMG

square

chain



- The **shape** of the distribution highly depends on the model (whatever the motif)
- FDD generates symetrical distributions (reflect the constraint of the model)
- EDD generates highly skewed distributions (diversity of visited configurations)

Conclusions for the simulation study

- Criteria used to assess the goodness of fit:

  $\longrightarrow$ The Kolmogorov-Smirnoff distance between theoretical and empirical distributions

  $\longrightarrow$ Empirical probabilities of exceeding the 0.999 quantile. It should be close to 0.001.

- The Geometric-Poisson approximation outperforms the Gaussian approximations for both criteria in all cases.
- The 0.999 quantile is underestimated by the Gaussian approximation:

  $\longrightarrow$ the Gaussian approximation can lead to false positive results

- The KS distance is high for both approximations in some cases, especially for frequent and highly self overlapping motifs.
- However the clumps size distribution is not geometric...

Exceptional motifs in PPI networks

Hpylo	N_{obs}	FDD-Pv$_{\mathcal{PA}}$	EDD-Pv$_{\mathcal{PA}}$	ERMG-Pv$_{\mathcal{PA}}$
V	14113	-	4.13e−01	4.06e−01
triangle	75	4.36e−03	9.06e−01	3.31e−01
chain	98697	1.22e−05	7.42e−01	4.12e−01
star	112490	-	3.65e−01	2.34e−01
square	1058	1.80e−52	6.15e−01	1.33e−02
whisker	3535	1.11e−02	8.58e−01	2.63e−01
halfclique	79	2.54e−05	7.51e−01	3.11e−02
clique	0	1.00e−00	1.00e−00	8.50e−01

- Using the FDD model leads to very drastic results (constant accross examples)

- When everything is exceptional the model should be questioned !

- For ERMG, 2 motifs are exceptional in the PPI network of H. Pylori

Exceptional motifs in PPI networks

Ecoli	N_{obs}	FDD-Pv$_{\mathcal{PA}}$	EDD-Pv$_{\mathcal{PA}}$	ERMG-Pv$_{\mathcal{PA}}$
V	248093	-	1.24e−08	4.46e−01
triangle	11368	0.00e+00	7.02e−13	3.30e−01
chain	9557956	0.00e+00	2.33e−10	4.68e−01
star	6425495	-	1.14e−11	3.26e−01
square	487408	0.00e+00	3.48e−23	3.10e−01
whisker	2154048	1.03e−265	1.15e−12	3.49e−01
halfclique	273621	1.24e−115	1.09e−17	2.14e−01
clique	14882	2.61e−41	3.30e−15	1.09e−01

- The behavior of the EDD model is not satisfactory
- Motifs are either all exceptional or all non exceptional
- May be linked to a variable quality of fit of the model to the data.

Exceptional motifs in PPI networks

Scere	N_{obs}	FDD-Pv$_{\mathcal{PA}}$	EDD-Pv$_{\mathcal{PA}}$	ERMG-Pv$_{\mathcal{PA}}$
V	436131	-	6.21e−33	1.44e−01
triangle	10567	1.31e−128	1.13e−22	1.21e−06
chain	7530597	8.44e−99	8.61e−27	1.38e−01
star	12227236	-	3.11e−22	9.54e−03
square	165085	1.09e−322	3.19e−22	2.73e−02
whisker	993733	1.64e−65	9.33e−22	8.90e−04
halfclique	116667	1.71e−33	1.28e−18	7.22e−04
clique	8601	1.54e−10	3.19e−16	5.25e−06

- ERMG could be an alternative : Pvalues are moderate
- $\mu_{\text{ERMG}}(\mathbf{m})$ depends on the number of groups
- Model averaging to stabilize the procedure

Conclusions & future directions

- We propose a method to assess the exceptionality of network motifs.
- The method to calculate the moments of the count is general and can be applied to any random graph model with exchangeable distribution
- The Geometric-Poisson approximation for the count distribution works well on simulated data.
- Directions: how to assess the distribution of the clump size. Is there a general method or does it depend on each motif ?