Analyse statistique de réseaux biologiques

Sophie S. Schbath

Analyse statistique de réseaux biologiques

Sophie Schbath

Unité **M**athématique, **I**nformatique et **G**énome
INRA - Jouy-en-Josas

Séminaire Proba-Stat, Orsay, 23 juin 2011

Part 1
Introduction

The network revolution

- **Nature of the data** :
  – n individuals (n large),
  – but also n^2 couples.
- **Many scientific fields** :
  sociology, physics, "internet", biology.
- **Biological networks** :
  protein-protein interaction networks, regulatory networks, metabolic networks.

Biological networks (1/2)

Gene regulatory networks

- nodes = genes
- edges : regulations (directed)

Protein-protein interaction networks

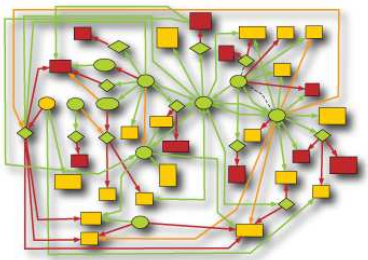
- nodes = proteins
- edges : physical interaction

Metabolic networks

- nodes = chemical compounds
- edges : chemical reactions or enzyme (directed, hyper-edges)

Reaction networks

- nodes = enzymes
- edges : consecutiveness in the metabolic network



Biological networks (1/2)

Gene regulatory networks

- nodes = genes
- edges : regulations (directed)

Protein-protein interaction networks

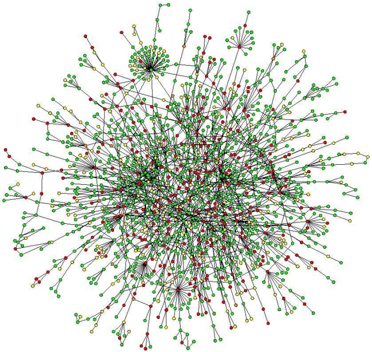
- nodes = proteins
- edges : physical interaction

Metabolic networks

- nodes = chemical compounds
- edges : chemical reactions or enzyme (directed, hyper-edges)

Reaction networks

- nodes = enzymes
- edges : consecutiveness in the metabolic network



Biological networks (1/2)

Gene regulatory networks

- nodes = genes
- edges : regulations (directed)

Protein-protein interaction networks

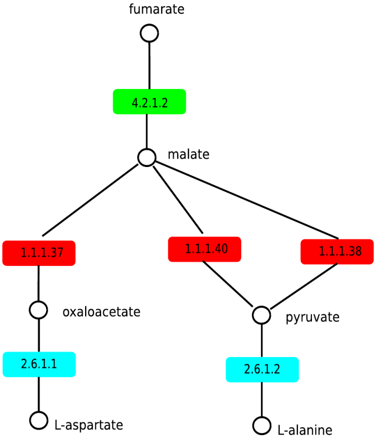
- nodes = proteins
- edges : physical interaction

Metabolic networks

- nodes = chemical compounds
- edges : chemical reactions or enzyme (directed, hyper-edges)

Reaction networks

- nodes = enzymes
- edges : consecutiveness in the metabolic network



Biological networks (1/2)

Gene regulatory networks

- nodes = genes
- edges : regulations (directed)

Protein-protein interaction networks

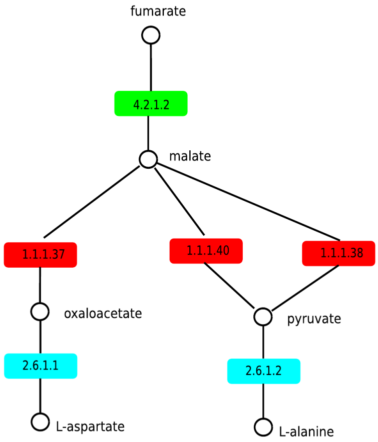
- nodes = proteins
- edges : physical interaction

Metabolic networks

- nodes = chemical compounds
- edges : chemical reactions or enzyme (directed, hyper-edges)

Reaction networks

- nodes = enzymes
- edges : consecutiveness in the metabolic network



Biological networks (2/2)

Main characteristics :

- several thousands of nodes (n)
- sparsity (nb of edges $= O(n)$)
- heterogeneous connexions
- nodes may be coloured (biological function, class of reaction, cellular localization etc.)

The network revolution (fol.)

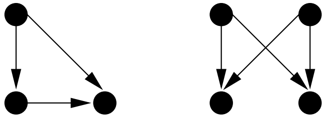
- **Nature of the data** :
  – n individuals (n large),
  – but also n^2 couples.
- **Many scientific fields** :
  sociology, physics, "internet", biology.
- **Biological networks** :
  protein-protein interaction networks, regulatory networks, metabolic networks.
- **Statistical aspects** :
  – network inference,
  – statistical properties of given networks (degrees, diameter, clustering coefficient, modules, motifs etc.),
  – random graph models.

Looking for local structures

- Breaking-down complex networks into functional modules or **basic building blocks** : [*Shen-Orr et al. (02)*]
  $\rightarrow$ network motifs : topological motifs and/or coloured motifs.
- **Focus on exceptional motifs** $=$ motifs appearing more frequently than **expected**.
  [*Milo et al. (02), Shen-Orr et al. (02), Zhang et al. (05), Lacroix et al. (06), Lee et al. (07), Taylor et al. (07)*]

Network motifs (1/2)

- **Topological motif** : **connected pattern of interconnection**
  $\rightarrow$ an occurrence in the network is an isomorphic subgraph

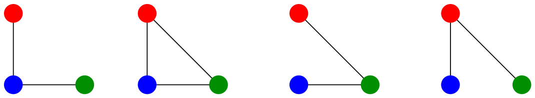


Ex : particular regulatory units like feed-forward loop or bi-fan motifs.

Interpretation of over-represented topological motifs : they are thought to reflect functional units which combine to regulate the cellular behavior as a whole.

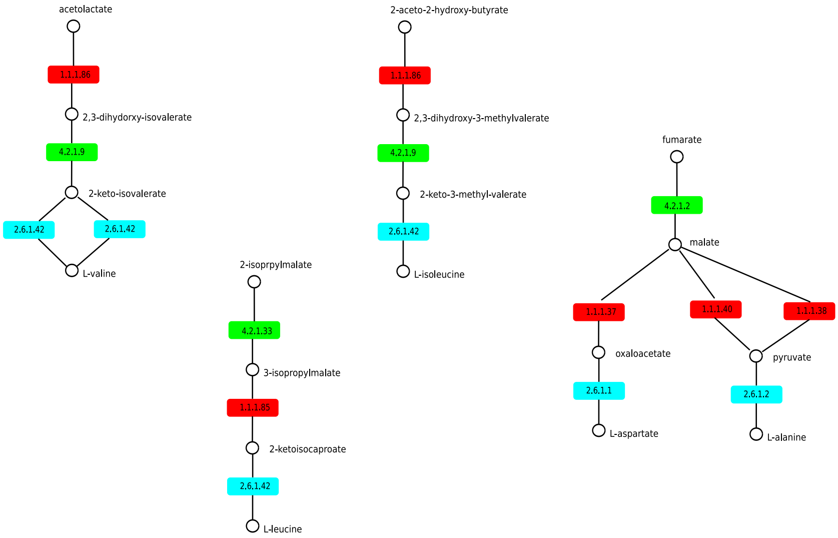
Network motifs (2/2)

- **Coloured motif** : **multiset of node colours**, e.g. {•, •, •}
  → an occurrence in the network is a connected subgraph with the appropriate node colours



Interpretation of over-represented coloured motifs :
they are thought to reflect groups of cooperative enzymes (reaction networks).

Coloured motifs : example



How to assess the exceptionality of a motif ?

Step 1 **To count** the observed number of occurrences $N_{\text{obs}}(\mathbf{m})$ of a given motif **m** (out of my scope)

Its significance is assessed with the *p*-value $\mathbb{P}\{N(\mathbf{m}) \geq N_{\text{obs}}(\mathbf{m})\}$ (*the probability to get as much occurrences at random*)

Step 2 To choose an appropriate **random graph model**

Step 3 To get the **distribution** of the count $N(\mathbf{m})$ under this model

State of the art (1/2)

Analytical approaches :

- The most popular random graph model is the **Erdös-Rényi model** (nodes are connected independently with proba p)
- Some theoretical works exist on Poisson and Gaussian approximations of topological motif count distribution (see [*Janson et al. (00)*] for an overview)

  BUT

  - only for particular motifs (“balanced” property),
  - the Erdös-Rényi model is not a good model for biological networks (e.g. it does not fit the degrees).

State of the art (2/2)

Simulated approaches :

- Random networks are generated by edge swapping, (degrees are preserved)
- Empirical distributions for motif counts are obtained leading either to p-values or to z-scores

  BUT

  - huge number of simulations required to estimate tiny p-values,
  - z-scores are compared to $\mathcal{N}(0, 1)$ which is not always appropriate,
  - edge swapping does not define a probabilistic random graph model.

SSB contributions (1/2)

- To propose probabilistic random graph models
  - adapted for biological networks,
  - allowing probabilistic calculations,
  - with efficient estimation procedures.

[Daudin, Picard, Robin (08)]. A mixture model for random graphs. *Statis. Comput.*
[Birmelé (07)]. A scale-free graph model based on bipartite graphs. *Disc. Appl. Math.*
[Mariadassou, Robin, Vacher (10)]. Uncovering structure in valued graphs : a variational approach. *Ann. Appl. Statist.*
[Latouche, Birmele, Ambroise (10)] Overlapping Stochastic Block Models with Application to the French Political Blogosphere. *Annals of Applied Statistics*
[Daudin, Pierre, Vacher (10).] Model for Heterogeneous Random Networks Using Continuous Latent Variables and an Application to a Tree–Fungus Network. *Biometrics*
[Latouche, Birmele, Ambroise (11)] Variational Bayesian Inference and Complexity Control for Stochastic Block Models. *Statistical Modelling*
[Gazal, Daudin, Robin (11)]. Accuracy of variational estimates for random graph mixture models. *J. Comput. Comput. Simul.*

SSB contributions (2/2)

- To provide general analytical results on motif count distribution :
  - mean and variance of the count in a wide class of random graph models,
  - relevant distribution to approximate the count distribution.

[*Matias, Schbath, Birmelé, Daudin and Robin (06)*] Network motifs : mean and variance for the count, *REVSTAT*. 4 31–51.
[*Picard, Daudin, Schbath and Robin (08)*] Assessing the exceptionality of network motifs, *J. Comput. Biol.*
[*Schbath, Lacroix and Sagot (09)*] Assessing the exceptionality of coloured motifs in networks, *EURASIP*

Part 2

Mixture model for random graphs

(Stochastic Block model)

Random graphs

- A random graph G is defined by :
  - a set $\mathcal{V}$ of fixed vertices with $|\mathcal{V}| = n$,
  - a set of random edges $\mathbf{X} = \{X_{ij}, (i,j) \in \mathcal{V}^2\}$ such that

$$X_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are connected,} \\ 0 & \text{otherwise} \end{cases}$$

  - and a distribution on X_{ij}.
- Examples :
  - the Erdös-Rényi model,
  - the Stochastic Block Model (=mixture of ER models),
  - the Expected Degree Distribution model.

Erdös-Rényi model

- Edges X_{ij}'s are independent ...
- ... and identically distributed according to $\mathcal{B}(p)$

$$\mathbb{P}(X_{ij} = 1) = p$$

- Degrees are Poisson distributed

$$K_i := \sum_{j \neq i} X_{ij} \sim \mathcal{B}(n-1, p) \approx \mathcal{P}((n-1)p)$$

- Bad fit of Erdös-Rényi model on biological networks due to heterogeneous connection probabilities along the network.

Stochastic Block Model (or “Mixnet”)

- Vertices are spread into Q groups.
- Conditionally to the group of vertices, edges are independent and

$$X_{ij} \mid \{i \in q, j \in \ell\} \sim \mathcal{B}(\pi_{q,\ell})$$

  $\pi_{q,\ell}$ is the connection probability between groups q and ℓ.
- Degrees are distributed according to a Poisson mixture

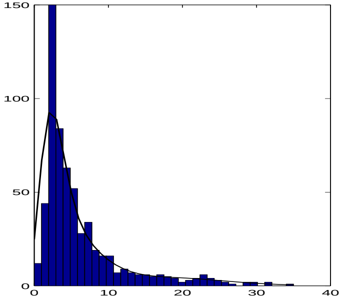
$$K_i \sim \sum_q \alpha_q \mathcal{B}(n-1, \overline{\pi}_q) \text{ with } \overline{\pi}_q = \sum_\ell \alpha_\ell \pi_{q,\ell}$$

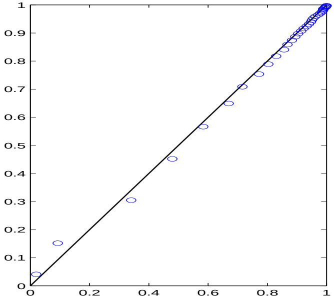
- Introduced by [*Nowicki and Snijers (2001)*]

Mixnet fit

- *E. coli* reaction network : 605 vertices, 1782 edges. (data curated by V. Lacroix and M.-F. Sagot).
- Degrees : Poisson mixture versus empirical distribution

PP-plot

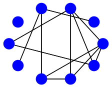
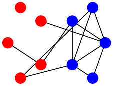
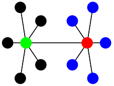
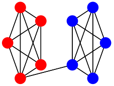




- Clustering coefficient :

Empirical	Mixnet ($Q = 21$)	ER ($Q = 1$)
0.626	0.544	0.0098

Mixnet flexibility

Examples	Network	Q	π
Erdös-Rényi		1	p
Independent model		2	$\begin{pmatrix} a^2 & ab \\ ab & b^2 \end{pmatrix}$
Stars		4	$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$
Clusters (affiliation network)		2	$\begin{pmatrix} 1 & \varepsilon \\ \varepsilon & 1 \end{pmatrix}$

Mixnet : estimation procedure

[Daudin, Picard and Robin (Stat. Comput. 08)]

Classical maximum likelihood procedures fail

- log-likelihood $\mathcal{L}(\mathbf{X})$ not calculable because of hidden groups ($\mathbf{Z}$, Z_i is the group of node i).
- EM algorithm, classical to fit mixture models, cannot be used because $\mathbb{P}(\mathbf{Z} \mid \mathbf{X})$ is not computable (all vertices are potentially connected, no local dependence).

Variational approach (iterative procedure)

- maximization of $\mathcal{L}(\mathbf{X}) - KL(\mathbb{P}(\mathbf{Z} \mid \mathbf{X}), Q_R(\mathbf{Z}))$ where Q_R is the best approximation of $\mathbb{P}(\mathbf{Z} \mid \mathbf{X})$ within a class of 'nice' distributions.
  $\Rightarrow$ estimator of $\mathbb{P}(Z_i = q|\mathbf{X})$.
- analytical expressions for $\widehat{\alpha}_q$ and $\widehat{\pi}_{q,\ell}$

Choice of Q : heuristic penalized likelihood criterion inspired from BIC (ICL)

Part 3

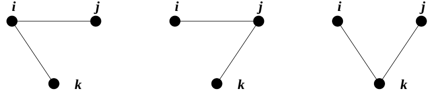
Assessing the significance of topological motif frequencies

Topological motifs

Let **m** be a motif of size k (connected graph with k vertices, $k << n$).

- **m** is defined by its adjacency matrix (also denoted by **m**) : $\mathbf{m}_{uv} = 1$ iff nodes $u \leftrightarrow v$ ($\mathbf{m}_{uv} = 0$ otherwise).
- Let $\mathcal{R}(\mathbf{m})$ be the set of non redundant permutations of **m** (so-called "versions").
- Ex : 3 versions of the V motif at a **fixed** position (i, j, k).

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



Occurrences of a motif

- Let $\alpha = (i_1, \ldots, i_k) \in I_k$ be a possible position of $\mathbf{m}$ in G. G_α denotes the subgraph $(V_{i_1}, \ldots, V_{i_k})$.
- Non strict occurrences :

$$\mathbf{m} \text{ occurs at position } \alpha \Leftrightarrow \mathbf{m} \subseteq G_\alpha$$

- Random indicator of occurrence : $Y_\alpha(\mathbf{m})$

$$Y_\alpha(\mathbf{m}) = \mathbb{1}\{\mathbf{m} \text{ occurs at position } \alpha\} = \prod_{1 \le u,v \le k} X_{i_u i_v}^{m_{uv}}.$$

- The total count $N(\mathbf{m})$ of motif $\mathbf{m}$ is then :

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

- Warning : $N(\mathbf{m}) \neq$ number of induced subgraphs ("$\mathbf{m} = G_\alpha$").

Expected count and variance

Under assumptions (H1) and (H2) on the random graph model :

- (H1) Stationary assumption : $\mathcal{D}(X_{i_1,j_1},\ldots,X_{i_\ell,j_\ell}) = \mathcal{D}(X_{i'_1,j'_1},\ldots,X_{i'_\ell,j'_\ell})$
- (H2) Independence of disjoint occurrences

we have [*Picard, Daudin, Koskas, Schbath, Robin (08)*]

$$\mathbb{E}N(\mathbf{m}) = \binom{n}{k}|\mathcal{R}(\mathbf{m})|\mu(\mathbf{m}).$$

where $\mu(\mathbf{m}) := \mathbb{E}Y_\alpha(\mathbf{m}) = \mathbb{P}(\mathbf{m} \text{ occurs at } \alpha)$ and

$$\mathrm{Var}N(\mathbf{m}) = \sum_{s=0}^{k} C(n,k,s) \sum_{\mathbf{m}' \underset{s}{\Omega} \mathbf{m}''} \mu(\mathbf{m}' \underset{s}{\Omega} \mathbf{m}'') - [\mathbb{E}N(\mathbf{m})]^2.$$

where $\mathbf{m}' \underset{s}{\Omega} \mathbf{m}''$ is a **super-motif** composed of the union of two overlapping occurrences of $\mathbf{m}'$ and $\mathbf{m}''$ sharing s common vertices.

Variance

By definition $\mathbb{V}\mathrm{ar}N(\mathbf{m}) = \mathbb{E}N^2(\mathbf{m}) - [\mathbb{E}N(\mathbf{m})]^2$. We then calculate

$$\begin{aligned}
\mathbb{E}N^2(\mathbf{m}) &= \mathbb{E}\left(\sum_{\alpha,\beta\in I_k}\ \sum_{\mathbf{m}',\mathbf{m}''\in\mathcal{R}(\mathbf{m})} Y_\alpha(\mathbf{m}')Y_\beta(\mathbf{m}'')\right), \\
&= \mathbb{E}\left(\sum_{s=0}^{k}\ \sum_{|\alpha\cap\beta|=s}\ \sum_{\mathbf{m}',\mathbf{m}''\in\mathcal{R}(\mathbf{m})} Y_{\alpha\cup\beta}(\mathbf{m}'\underset{s}{\Omega}\mathbf{m}'')\right) \\
&= \sum_{s=0}^{k} C(n,k,s) \sum_{\mathbf{m}'\underset{s}{\Omega}\mathbf{m}''} \mu(\mathbf{m}'\underset{s}{\Omega}\mathbf{m}''),
\end{aligned}$$

where $\mathbf{m}'\underset{s}{\Omega}\mathbf{m}''$ is a **super-motif** composed of the union of two overlapping occurrences of $\mathbf{m}'$ and $\mathbf{m}''$ sharing s common vertices.

Candidate random graph models

- Erdös-Rényi model (ER) : Edges X_{ij}'s are i.i.d. $\sim \mathcal{B}(p)$

$$\mu(\mathbf{m}) = p^{e(\mathbf{m})}$$

Candidate random graph models

- Erdös-Rényi model (ER) : Edges X_{ij}'s are i.i.d. $\sim \mathcal{B}(p)$

$$\mu(\mathbf{m}) = p^{e(\mathbf{m})}$$

- Mixture of ER model (Mixnet/SBM) with Q groups, proportions $\alpha_1, \ldots, \alpha_Q$ and connection probabilities $\pi_{q,\ell}$

$$\mu(\mathbf{m}) = \sum_{c_1=1}^{Q} \ldots \sum_{c_k=1}^{Q} \alpha_{c_1} \ldots \alpha_{c_k} \prod_{1 \leq u < v \leq k} \pi_{c_u,c_v}^{m_{uv}}.$$

Motif count distribution

- Exact distribution unknown.
- Several approximations exist in the literature under specific conditions (motif and model) :
  - Poisson distribution *[Bollobas (81), Barbour (82), Karónski and Ruciński (83)]*
  - Gaussian distribution *[Barbour et al. (87)]*
  - Compound Poisson distribution *[Stark (01)]*

Compound Poisson distribution

- Distribution of $\sum_{i=1}^{Z} T_i$ when $Z \sim \mathcal{P}(\lambda)$ and T_i's iid.
- Particularly adapted for the count of clumping events : Z is the number of clumps and T_i is the size of the i-th clump.

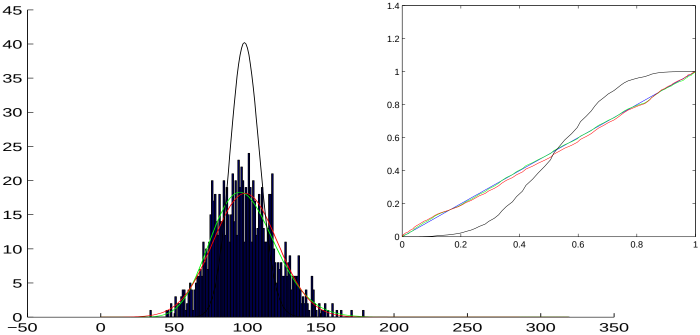
- All network motifs are overlapping : they occur in clumps.
- We proposed to use a Geometric-Poisson(λ, a) distribution, i.e. when $T_i \approx \mathcal{G}(1-a)$
  - analogy with sequence motifs [S. (95)],
  - (λ, a) can be calculated according to $\mathbb{E}N(\mathbf{m})$ and $\mathbb{V}\mathrm{ar}N(\mathbf{m})$:

$$a = \frac{\mathbb{E}N(\mathbf{m}) - \mathbb{V}N(\mathbf{m})}{\mathbb{E}N(\mathbf{m}) + \mathbb{V}N(\mathbf{m})} \quad \text{and} \quad \lambda = (1-a)\mathbb{E}N(\mathbf{m}).$$

Simulation study

Model = mixnet with 2 groups, $n = 200$, etc.

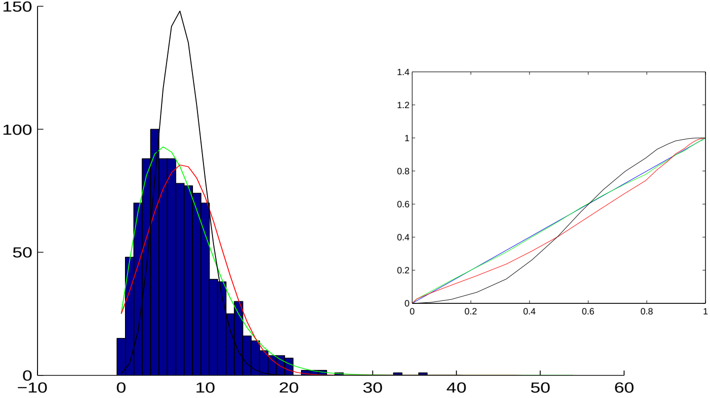
For expectedly frequent motifs :



Gaussian (–), Poisson (–) and Geometric-Poisson (–)

Simulation study (fol.)

For expectedly rare motifs :



Gaussian (–), Poisson (–) and Geometric-Poisson (–)

Application to the *H. pylori* PPI network

- PPI network : 706 proteins and 1420 interactions (edges).
- Mixnet was fitted to the network $\rightarrow$ 4 groups of connectivity.

Motif	N_{obs}	$\mathbb{E}_{\text{mixnet}} N$	$\sigma_{\text{mixnet}}(N)$	$\mathbb{P}(\mathcal{GP} \leq N_{\text{obs}})$	$\mathbb{P}(\mathcal{GP} \geq N_{\text{obs}})$
	14113	13602	2659		$4.06\,10^{-1}$
	75	66.9	20.4		$3.31\,10^{-1}$
	98697	94578	27039		$4.12\,10^{-1}$
	112490	93741	27257		$2.34\,10^{-1}$
	1058	516.6	208.7		$1.33\,10^{-2}$
	3535	2897	1120		$2.63\,10^{-1}$
	79	34.8	20.0		$3.11\,10^{-2}$
	0	0.17	0.45	$8.5\ 10^{-1}$	

Part 3

Assessing the significance of coloured motif frequencies

Coloured motifs

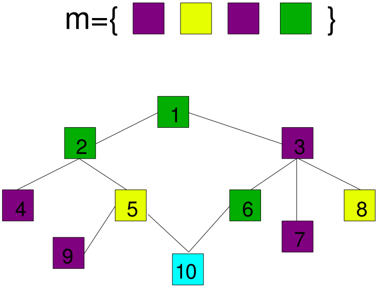
Graph : n nodes coloured with colours in $\mathcal{C}$

Coloured motif m of size k is a multiset of k colours $\{m_1, \ldots, m_k\} \in \mathcal{C}^k$.

Multiplicity of colour c in **m** : $s_{\mathbf{m}}(c) = s(c)$.

Indicator of occurrence at position α : $Y_\alpha(\mathbf{m})$

Number of occurrences : $N(\mathbf{m}) = \sum_{\alpha \in I_k} Y_\alpha(\mathbf{m})$.



Model for coloured graph

- **Topology** : Erdös-Rényi model with probability p
- **Colours** : Let f be a probability measure on $\mathcal{C}$; Nodes are coloured independently in color $c \in \mathcal{C}$ with probability $f(c)$.

This model allows to derive analytical formulas for mean and (co)variance of motif counts *[Schbath, Lacroix, Sagot (09)]*

The motif count distribution is then approximated by a Geometric-Poisson distribution.
$\longrightarrow$ approximate p-value $\mathbb{P}(N(\mathbf{m}) \geq N^{\text{obs}}(\mathbf{m}))$.

Coloured motifs : Expected count

$$\begin{aligned}\mathbb{E}N(\mathbf{m}) &= \sum_{\alpha\in I_k}\mathbb{E}Y_\alpha(\mathbf{m}) = \binom{n}{k}\,\mathbb{P}(\mathbf{m}\text{ occurs at }\alpha)\\ &= \binom{n}{k}\,g(k,p)\times\underbrace{\frac{k!}{\prod_{c\in\mathcal{C}} s(c)!}\prod_{i=1}^{k} f(m_i)}_{:=\gamma(\mathbf{m})}\end{aligned}$$

where $g(k,p)$ is the probability for an ER(p) graph of size k to be connected *[Gilbert, 59]* :

$$g(k,p) = 1 - \sum_{i=1}^{k-1}\binom{k-1}{i-1} g(i,p)(1-p)^{i(k-i)}.$$

$(g(1,p)=1)$.

Coloured motifs : Variance of the count (1/2)

Let us just compute $\mathbb{E}N^2(\mathbf{m})$.

$$\begin{aligned}\mathbb{E}N^2(\mathbf{m}) &= \sum_{\alpha\in I_k}\sum_{\beta\in I_k}\mathbb{E}[Y_\alpha(\mathbf{m})Y_\beta(\mathbf{m})].\\ &= \sum_{\ell=0}^{k}\sum_{|\alpha\cap\beta|=\ell}\underbrace{\mathbb{P}(\mathbf{m}\text{ occurs at }\alpha\text{ and }\beta)}_{=K(\alpha,\beta)\times Q_{\mathbf{m}}(\alpha,\beta)}.\end{aligned}$$

where

$$\begin{aligned}K(\alpha,\beta) &= \mathbb{P}(G(\alpha)\text{ and }G(\beta)\text{ are connected})\\ Q_{\mathbf{m}}(\alpha,\beta) &= \mathbb{P}(C(\alpha)=C(\beta)=\{m_1,\dots,m_k\}).\end{aligned}$$

Coloured motifs : variance of the count (2/2)

color term :

$$\begin{aligned} Q_{\mathbf{m}}(\alpha, \beta) &= \mathbb{P}(C(\alpha) = C(\beta) = \{m_1, \ldots, m_k\}). \\ &= \sum_{\mathbf{m}^* \subset \mathbf{m}} \frac{\gamma(\mathbf{m}^*)[\gamma(\mathbf{m}^-)]^2}{s(\mathbf{m}^*)} \end{aligned}$$

Coloured motifs : variance of the count (2/2)

color term :

$$\begin{aligned} Q_{\mathbf{m}}(\alpha,\beta) &= \mathbb{P}(C(\alpha) = C(\beta) = \{m_1,\dots,m_k\}). \\ &= \sum_{\mathbf{m}^*\subset\mathbf{m}} \frac{\gamma(\mathbf{m}^*)[\gamma(\mathbf{m}^-)]^2}{s(\mathbf{m}^*)} \end{aligned}$$

connectedness term :

$$\begin{aligned} K(\alpha,\beta) &= \mathbb{P}(G(\alpha) \text{ and } G(\beta) \text{ are connected}) \\ &= \begin{cases} g(k,p), & \text{if } \ell = k \\ g^2(k,p), & \text{if } \ell = 0 \text{ or } 1. \\ & \\ & \end{cases} \end{aligned}$$

Coloured motifs : variance of the count (2/2)

color term :

$$\begin{aligned} Q_{\mathbf{m}}(\alpha,\beta) &= \mathbb{P}(C(\alpha) = C(\beta) = \{m_1,\ldots,m_k\}). \\ &= \sum_{\mathbf{m}^*\subset\mathbf{m}} \frac{\gamma(\mathbf{m}^*)[\gamma(\mathbf{m}^-)]^2}{s(\mathbf{m}^*)} \end{aligned}$$

connectedness term :

$$\begin{aligned} K(\alpha,\beta) &= \mathbb{P}(G(\alpha) \text{ and } G(\beta) \text{ are connected}) \\ &= \begin{cases} g(k,p), & \text{if } \ell = k \\ g^2(k,p), & \text{if } \ell = 0 \text{ or } 1. \\ \text{ad-hoc polynoms} & \text{otherwise} \end{cases} \end{aligned}$$

Coloured motifs : variance of the count (2/2)

color term :

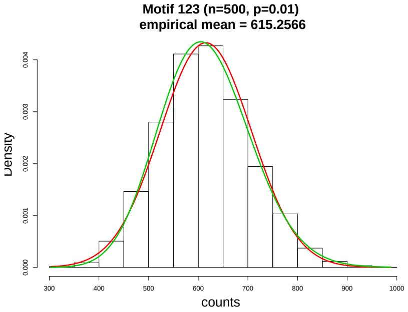
$$\begin{aligned} Q_{\mathbf{m}}(\alpha,\beta) &= \mathbb{P}(C(\alpha) = C(\beta) = \{m_1,\dots,m_k\}). \\ &= \sum_{\mathbf{m}^* \subset \mathbf{m}} \frac{\gamma(\mathbf{m}^*)[\gamma(\mathbf{m}^-)]^2}{s(\mathbf{m}^*)} \end{aligned}$$

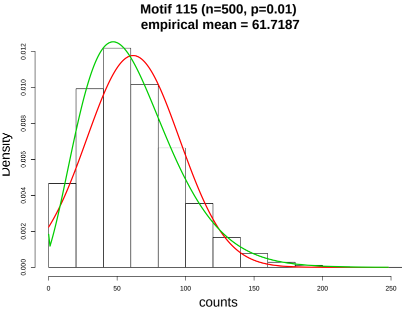
connectedness term :

$$\begin{aligned} K(\alpha,\beta) &= \mathbb{P}(G(\alpha) \text{ and } G(\beta) \text{ are connected}) \\ &= \begin{cases} g(k,p), & \text{if } \ell = k \\ g^2(k,p), & \text{if } \ell = 0 \text{ or } 1. \\ \text{ad-hoc polynoms} & \text{otherwise} \\ 4p^3 - 3p^4 & \text{if } \ell = 3 \text{ and } k = 3 \\ & \text{etc.} \end{cases} \end{aligned}$$

Geometric-Poisson approximation (1/2)

Both parameters of the GP distribution can be derived from the first 2 moments of the count.

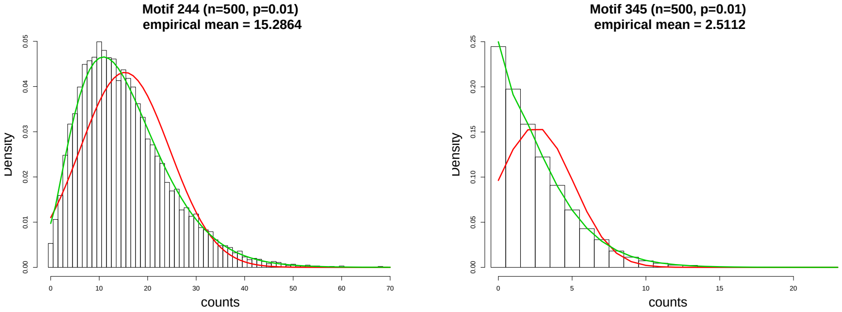




Gaussian (red curve), Geometric-Poisson (green curve)

Geometric-Poisson approximation (2/2)

Both parameters of the PA distribution can be derived from the first 2 moments of the count.



Gaussian (red curve), Geometric-Poisson (green curve)

Part 4

Comparison of reaction networks

(ongoing work with S. Robin and L. Benaroya)

Aim

Let G_1 and G_2 be two coloured graphs of size n_1 and n_2 (typically reaction networks from 2 different species).

Each graph is caracterized by the count vector of M given motifs of size k : $\mathbf{N}_g = (N_{g1}, N_{g2}, \ldots, N_{gM})$, for $g = 1, 2$.

Questions :

- Do they share common exceptional motifs ?
- Have both graphs similar k-motif compositions ?
- Whose motifs are the most discriminant ?

Idea : to define a motif-based distance taking care of

- the deviations from the models,
- the dependence between motif counts.

Motif-based distance

Normalization by the size of the graphs :
Since $\mathbb{E}N(\mathbf{m}) = \binom{n}{k}\,\mathbb{P}(\mathbf{m} \text{ occurs at } \alpha)$, we define :

$$\widetilde{N}_{gm} = \binom{n_g}{k}^{-1} N_{gm}, \quad g = 1, 2$$

Box-Cox Transformation to make the counts “more” Gaussian :

$$N^*_{gm} = 2(\sqrt{\widetilde{N}_{gm}} - 1), \quad g = 1, 2$$

Euclidian distance on z-scores :

$$d^2(\mathbf{N}^*_1, \mathbf{N}^*_2) = ||(\Sigma^*_1)^{-1/2}(\mathbf{N}^*_1 - \mathbb{E}\mathbf{N}^*_1) - (\Sigma^*_2)^{-1/2}(\mathbf{N}^*_2 - \mathbb{E}\mathbf{N}^*_2)||^2_2$$

where $\mathbb{E}\mathbf{N}^*_g$ and the covariance matrix $\Sigma^*_g = (\mathrm{Cov}(N^*_g(\mathbf{m}_i), N^*_g(\mathbf{m}_j)))_{i,j}$ can be calculated from $\mathbb{E}\mathbf{N}_g$ and the covariance matrix $(\mathrm{Cov}(N_g(\mathbf{m}_i), N_g(\mathbf{m}_j)))_{i,j}$ (previous part).

Sequential distance

1. Consider all single motif sets ($\dim(\mathbf{N}_1^*)=\dim(\mathbf{N}_2^*)=1$), and take

$$\widehat{\mathbf{m}}^1 = \mathrm{argmax}_{\mathbf{m}_1,\dots,\mathbf{m}_M} d^2(\mathbf{N}_1^*, \mathbf{N}_2^*)$$

2. Consider all motif pairs $(\widehat{\mathbf{m}}^1, \mathbf{m}_j)$ with $\mathbf{m}_j \neq \widehat{\mathbf{m}}^1$, and take

$$\widehat{\mathbf{m}}^2 = \mathrm{argmax}_{\mathbf{m}_j \neq \widehat{\mathbf{m}}^1} d^2(\mathbf{N}_1^*, \mathbf{N}_2^*)$$

3. and so on

Exemple (1/2)

Reaction networks with threshold 3 on the EC numbers :

	Escherichia coli	*Buchnera aphidicola*
number of nodes	886	248
number of edges	4630	473
number of colors	107	62
motifs of size 3	6402	597

Exemple (2/2)

Rank	Motif	Cumulative distance
1	{ 2.7.1 2.7.4 6.3.4 }	300.90
2	{ 1.1.1 1.3.1 1.14.14 }	577.15
3	{ 1.1.1 1.1.1 1.6.1 }	835.41
4	{ 1.1.1 1.6.1 2.5.1 }	1029.83
5	{ 1.1.1 2.3.1 2.3.1 }	1177.35
6	{ 2.3.1 2.3.1 2.5.1 }	1324.18
7	{ 2.7.1 2.7.4 2.7.9 }	1467.98
8	{ 2.7.1 2.7.2 2.7.4 }	1606.38
9	{ 2.7.4 2.7.4 6.3.1 }	1747.93
10	{ 2.7.4 2.7.4 2.7.10 }	1876.69
11	{ 1.1.1 1.2.1 1.6.1 }	2003.00
12	{ 2.7.4 2.7.4 3.6.4 }	2127.12
13	{ 2.3.1 3.1.2 6.2.1 }	2250.14
14	{ 2.7.1 2.7.4 6.3.3 }	2372.24
15	{ 1.1.1 1.6.1 3.5.1 }	2489.55

Another approach

- To model the vector $\mathbf{N} = (N_1, N_2, \dots, N_M)$
- Need for a “multidimensional (compound) Poisson distribution with given covariance matrix”
- Our choice = the multivariate Poisson-log normal distribution from [*Aitchison and Ho, 89*] :

$$\begin{aligned} N_m &\sim \mathcal{P}(e^{\lambda_m}) \\ \Lambda &\sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \end{aligned}$$

- $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ can be explicitly derived from the expectation and covariance matrix of **N**.
- Distance = euclidian distance between Λ_1 and Λ_2
- Λ is estimated by $\mathbb{E}(\Lambda \mid \mathbf{N})$

Limitations :

- $\boldsymbol{\Sigma}$ may be not positive
- No analytical expression for $\mathbb{E}(\Lambda \mid \mathbf{N})$ (Monte Carlo)

Acknowledgement

SSB net subgroup

AgroParisTech
Jean-Jacques Daudin
Michel Koskas
Stéphane Robin

MIG, Jouy
Mahendra Mariadassou

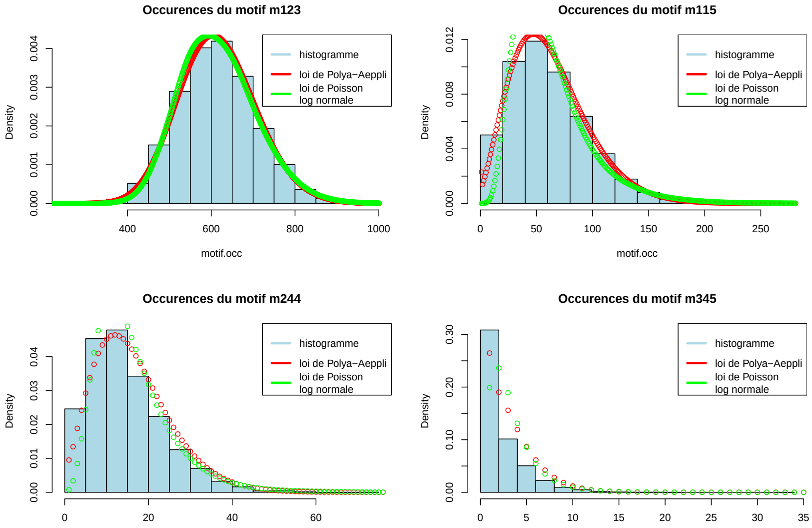
Stat&Génome, Evry
Christophe Ambroise
Etienne Birmelé
Camille Charbonnier
Julien Chiquet
Gilles Grasseau
Pierre Latouche
Catherine Matias

LBBE, Lyon
Vincent Lacroix
Vincent Miele
Franck Picard
Marie-France Sagot

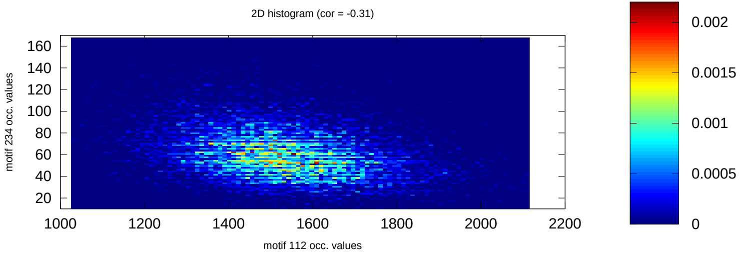
R package (Mixer) and C++ program (Mixnet) on `www.ssbgroup.fr`
An R package nemo soon available for network motif analysis.

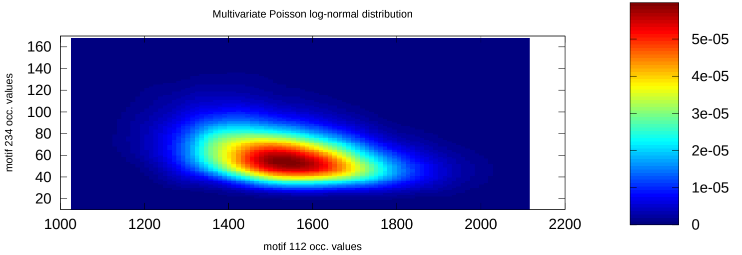
NeMo project supported by the French ANR

Multivariate Poisson-log normal distribution



Multivariate Poisson-log normal distribution





Multivariate Poisson-log normal distribution