ASSESSING THE EXCEPTIONALITY OF NETWORK MOTIFS

Picard F.⋆ [1],Daudin J-J.†, Schbath S.‡, Robin S.†.

⋆ UMR CNRS 8071 / INRA 1152 / Univ. d'Évry, Évry, France.
† UMR ENGREF / INA PG / INRA 518, Paris, France.
‡ INRA, Unité Mathématique, Informatique et Génome, Jouy-en-Josas, France.

Abstract

Getting and analyzing biological interaction networks is at the core of systems biology. To help understanding these complex networks, many recent works have suggested to focus on motifs which occur more frequently than expected in random. To identify such exceptional motifs in a given network, we propose a statistical and analytical method which does not require any simulation. For this, we first provide an analytical expression of the mean and variance of the count under any stationary random graph model. Then we approximate the motif count distribution by a compound Poisson distribution whose parameters are derived from the mean and variance of the count. Thanks to simulations, we show that the quality of our compound Poisson approximation is very good and highly better than a Gaussian or a Poisson one. The compound Poisson distribution can then be used to get an approximate p-value and to decide if an observed count is significantly high or not.

1 Introduction

The important progress of high-throughput biology allows us now to consider the cell as a whole system under study. This complex system is mainly represented by various networks of interacting components (e.g. transcriptional regulatory networks, protein-protein interaction networks, metabolic networks). To help understanding the organization and dynamics of cell functions, one usually tries to break down these complex networks into functional modules [5] or into basic building blocks [12], which are also called network motifs. For transcriptional regulatory networks, some motifs such as the three-node feed-forward loop or the four-node bi-fan, may perform specific regulatory functions [16]. Many recent works have suggested to focus on motifs which occur more frequently than expected in random [12, 16]. Such motifs seem indeed to reflect functional or computational units which combine to regulate the cellular behavior as a whole. Their possible function can be provided by common themes of the system in which they appear.

The common method that has been used for now to detect significantly over-represented motifs is based on simulations. Random graphs are first generated such that they preserve some characteristics of the biological network like the numbers of vertices and edges or the degree sequence (numbers of edges per vertex) [12, 11]. Then, either a z-score is calculated thanks to the empirical mean and variance of the count [12, 10], or an estimation of the empirical p-value is derived from the empirical distribution of the count [16, 12]. Such methods are not totally satisfactory from a probabilistic point of view. Indeed, using a z-score means to assume that the motif count follows a Gaussian distribution which is only true asymptotically under some restrictive conditions. Moreover, to evaluate a p-value close to zero, a huge number of simulations have to be performed, which is usually not the case in previous studies because of high computational times. Getting theoretical properties on the motif count distribution would then be very valuable to identify exceptional motifs.

Several approximations have been proposed under the so-called Erdös model [8]. This basic model assumes that edges are independent and distributed according to a Bernoulli distribution with

[1] picard@genopole.cnrs.fr

same parameter p. It means in particular that the probability to connect two nodes does not depend on the nodes. Under these assumptions, Poisson and compound Poisson approximations have been first proposed for rare motifs satisfying some conditions on their number of vertices and edges ([3, 17]). The asymptotic normality of the motif count has been also extensively studied and bounds on the approximation error have been derived (e.g. [2] and references therein). However, except for the mean count which is simple to derive under the Erdös model, no explicit formula of the parameters of these limiting distributions has never been provided. In particular, no general expression exists for the variance of the count.

However, the Erdös random graph model does not fit biological networks essentially because it does not take heterogeneities into account. Indeed, some nodes are very connected to others. A very general model assuming that edges are still independent but depend on both connected vertices has been recently studied and a method has been proposed to get the exact formulas for the mean and variance of the motif count [4]. However, their computations appear very time consuming. The first question we address in this paper is how to calculate in a unified way the exact mean and variance of a motif count under any stationary random graph model. These two quantities are indeed crucial to identify unexpected motifs. Provided that the occurrence probability of a given motif does not depend on the occurrence position (stationary assumption), we derive the expression of the first two moments of the count. In particular we treat the case of the ERMG model in detail. This model, introduced in [6], is a mixture of a finite number of Erdös models, nodes being spread into classes. It encompasses the Erdös model (a unique class of nodes) and the heterogeneous model proposed in [4] (one class per node). Moreover, the ERMG model is rich and flexible enough to capture relevant information on the network topology.

The second question we focus on is which approximation of the motif count distribution to use in order to get accurate p-values. Note that no result exists now on the exact distribution of this count. As regard to existing theoretical results under the Erdös model, we compare the approximation quality of the three following distributions: the Gaussian distribution, the Poisson distribution and the Pólya-Aeppli distribution. The later is a special compound Poisson distribution with only two parameters. This comparison will be done thanks to simulations under the ERMG model. Parameters of these approximate distributions will be set from the exact mean and variance of the count we provide.

For the sake of simplicity, we consider undirected graphs and motifs. However, our methodology can be easily generalized to a directed framework as it is discussed in the conclusion.

2 Definitions and notations

Let us define a random graph G, where $\mathcal{V}$ denotes the set of fixed vertices with $|\mathcal{V}| = n$. Random edges are described by a set of random variables $\mathbf{X} = \{X_{ij}, (i,j) \in \mathcal{V}^2\}$ such that X_{ij} equals 1 if nodes i and j are connected, and 0 otherwise. In the following, we consider random graphs with stationary probability distributions, for which $\mathbb{P}(\mathbf{X})$ does not depend on nodes. Moreover, we consider the case of non-directed graphs, meaning that $X_{ij} = X_{ji}$.

We denote by $\mathbf{m}$ a network motif of size k, which is a connected subgraph with k vertices. It is defined by a fixed topology through its adjacency matrix also denoted by $\mathbf{m}$, with general term $m_{uv} = 1$ if nodes u and v are connected, and 0 otherwise. A typical example is the V motif, which can be defined by three adjacency matrices depending on the position of the central edge, as shown in Table 1.

To define an occurrence of motif $\mathbf{m}$ we introduce notation I_k which is the set of all k-tuples of $\mathcal{V}$, namely $I_k = \left\{\{i_1, \ldots, i_k\} \subset \{1, \ldots, N\}^k \mid i_j \neq i_\ell, \forall j \neq \ell\right\}$. We consider $\alpha \in I_k$, a potential position of $\mathbf{m}$ in G. The number of such positions is $\binom{n}{k}$. In order to match a position with an

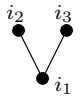
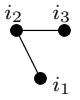
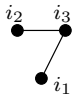
$\mathbf{m}$	$\mathbf{m}'$	$\mathbf{m}''$
$\begin{bmatrix} 0 & 1 & 1 \\ . & 0 & 0 \\ . & . & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 1 & 0 \\ . & 0 & 1 \\ . & . & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 & 1 \\ . & 0 & 1 \\ . & . & 0 \end{bmatrix}$
		

Table 1: Non redundant permutations of the V motif at position $\alpha = (i_1, i_2, i_3)$.

adjacency matrix, we consider a specific element of $\{i_1, \ldots, i_k\}$ which is $\alpha = (i_1, \ldots, i_k)$ with $i_1 < \ldots < i_k$. Then we introduce the random indicator variable $Y_\alpha(\mathbf{m})$ which equals one if motif $\mathbf{m}$ occurs at position α and 0 otherwise : $Y_\alpha(\mathbf{m}) = \prod_{1 \leq u < v \leq k} X_{i_u i_v}^{m_{uv}}$. Since the distribution of $\mathbf{X}$ is stationary, the distribution of Y_α does not depend on α, and Y_α is distributed according to a Bernoulli distribution $\mathcal{B}(\mu(\mathbf{m}))$, where $\mu(\mathbf{m})$ is the probability of occurrence of motif $\mathbf{m}$ at any position.

Considering the occurrence of the V motif at position $\alpha = (i_1, i_2, i_3)$ (Table 1), one can see that V occurs at α with a given permutation on indices. This is why we need to define $\mathcal{R}(\mathbf{m})$, the set of non redundant permutations of $\mathbf{m}$, and we denote $\rho(\mathbf{m}) = |\mathcal{R}(\mathbf{m})|$, which equals 3 in the case of the V motif, and 1 for the triangle. Note that $\rho(\mathbf{m}) = k!/|\text{aut}(\mathbf{m})|$, where $\text{aut}(\mathbf{m})$ is the set of automorphisms of motif $\mathbf{m}$: $\text{aut}(\mathbf{m}) = \{\sigma \in \mathfrak{S},\ \sigma(\mathbf{m}) = \mathbf{m}\}$, with $\mathfrak{S}$ the set of permutations on the vertices of $\mathbf{m}$. We consider permutations of the motif rather than permutations of positions.
From a practical point of view, we propose to avoid the calculation of $|\text{aut}(\mathbf{m})|$, and to focus on $\rho(\mathbf{m})$. This calculation can be done by considering the $k!$ simultaneous permutations of the rows and columns of $\mathbf{m}$, each new element being compared with the previous ones to check for redundancy. The complexity of this method is then in $\mathcal{O}(k!^2)$ and does not depend on the size of the complete graph. Moreover, since we are searching for small-size motifs ($k = 3, 4$ typically), the computational time of this procedure is moderate.

Finally we define $N(\mathbf{m})$ the count of motif $\mathbf{m}$ such that: $N(\mathbf{m}) = \sum_{\alpha \in I_k} \sum_{\mathbf{m}' \in \mathcal{R}(\mathbf{m})} Y_\alpha(\mathbf{m}')$.

3 Calculating moments under a stationary model

In this section, we aim at providing an automatic method to calculate the first and second moments of $N(\mathbf{m})$. This method requires the knowledge of $\mu(\mathbf{m})$, the probability of occurrence of motif $\mathbf{m}$. In a first step, we develop our method with $\mu(\mathbf{m})$ as a general term. This probability depends on the distribution of $\mathbf{X}$ and its derivation under different models will be given in the next section.

The calculation of the mean can be done directly since the distribution of Y_α does not depend on α. Indeed, the stationarity assumption implies that permutations of motif $\mathbf{m}$ have the same probability of occurrence ($\mu(\mathbf{m}) = \mu(\mathbf{m}')\ \forall \mathbf{m}' \in \mathcal{R}(\mathbf{m})$). It follows that:

$$\mathbb{E}N(\mathbf{m}) \quad = \quad |I_k| \times |\mathcal{R}(\mathbf{m})| \sum_{\mathbf{m}' \in \mathcal{R}(\mathbf{m})} \mathbb{E}Y_\alpha(\mathbf{m}') \quad = \quad \binom{n}{k} \rho(\mathbf{m}) \mu(\mathbf{m}). \tag{1}$$

The calculation of the variance is based on the expectation of the squared count:

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

and each term of this sum depends on the cardinality of the intersection $\alpha \cap \beta$ denoted by s. When $s = 0$, variables Y_α and Y_β are independent and $\mathbb{E}\left[Y_\alpha(\mathbf{m}) Y_\beta(\mathbf{m})\right] = \mathbb{E}Y_\alpha(\mathbf{m})\mathbb{E}Y_\beta(\mathbf{m})$. For $s \geq 1$, $\mathbf{m}'$ at α and $\mathbf{m}''$ at β share s vertices. Then we consider all possible overlaps between the two versions of $\mathbf{m}$ occurring at each position. We define the overlapping operation with s common vertices (denoted by $\underset{s}{\Omega}$) between motifs $\mathbf{m}'$ and $\mathbf{m}''$. Consequently,

$$\forall s \geq 1,\ Y_\alpha(\mathbf{m}') Y_\beta(\mathbf{m}'') \quad = \quad Y_{\alpha \cup \beta}(\mathbf{m}' \underset{s}{\Omega} \mathbf{m}''),$$

where $\mathbf{m}' \underset{s}{\Omega} \mathbf{m}''$ represents what we call a "super-motif", which is a motif with $(2k - s)$ edges made of two overlapping occurrences of $\mathbf{m}'$ and $\mathbf{m}''$, two versions of $\mathbf{m}$.

To define the adjacency matrix of the super-motif, 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 $\alpha \cap \beta$, and 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 $\mathfrak{OR}$. Note that the operation $\underset{s}{\Omega}$ is not symmetric. Note that we also have to consider the number of possible overlaps of $\mathbf{m}' \underset{s}{\Omega} \mathbf{m}''$ which is $|\mathcal{R}(\mathbf{m})|^2$. The complexity of this enumeration is therefore smaller than $\mathcal{O}(k!^2)$.

The squared count can be rewritten as $N^2(\mathbf{m}) = \sum_{s=0}^{k} \sum_{\substack{\alpha,\beta \in I_k : \\ |\alpha \cap \beta| = s}} \sum_{\mathbf{m}',\mathbf{m}'' \in \mathcal{R}(\mathbf{m})} Y_{\alpha \cup \beta}(\mathbf{m}' \underset{s}{\Omega} \mathbf{m}'')$,
and its expectation is:

$$\mathbb{E}N^2(\mathbf{m}) = \binom{n}{n-2k,k,k} \left[\sum_{\mathbf{m}' \in \mathcal{R}(\mathbf{m})} \mu(\mathbf{m}') \right]^2 + \sum_{s=1}^{k} \binom{n}{k-s,s,k-s,n-2k-s} \sum_{\mathbf{m}',\mathbf{m}'' \in \mathcal{R}(\mathbf{m})} \mu(\mathbf{m}' \underset{s}{\Omega} \mathbf{m}''). \tag{3}$$

Coefficients $\binom{a}{b,c,d}$ stand for multinomial coefficients. Put together, we can derive the formula for the variance of the count since $\mathbb{V}N(\mathbf{m}) = \mathbb{E}N^2(\mathbf{m}) - \mathbb{E}^2N(\mathbf{m})$.

4 Two stationary random graph models

Once the method to calculate the first and second moments of the count has been settled, we need to choose an appropriate model for the distribution of $\mathbf{X}$ in order to calculate $\mu(\mathbf{m})$, the probability of occurrence of motif $\mathbf{m}$. One basic model is the Erdös model [7] in which the probability of connection

between two vertices is constant and equals p. In this framework, $\mu(\mathbf{m})$ resumes to the probability for vertices of motif $\mathbf{m}$ to be connected. Denoting by $v(\mathbf{m})$ the number of vertices of $\mathbf{m}$ we have $\mu(\mathbf{m}) = p^{v(\mathbf{m})}$. Note that even if the expectation of the count is very simple in this case, calculating the variance still requires to calculate the occurrence probabilities of all super-motifs for which no simplification exists.

Despite a simple formulation and important theoretical results, it is now well accepted that the Erdös model fits the data poorly [1]. Alternative models have been proposed to describe real networks [1, 13]. Nevertheless, they are mainly based on summary statistics such as the degree distribution, which hampers the exact calculation of $\mu(\mathbf{m})$ which requires a probabilistic distribution on $\mathbf{X}$. In a recent publication, the calculus of the expectation and the variance of the count has been proposed under a heterogeneous model with $\Pr\{X_{ij} = 1\} \propto k_i k_j$, where k_i stands for the degree of node i [4]. Nevertheless, the fact that the distribution of $\mathbf{X}$ is not stationary leads to a procedure which is difficult to use in practice.

In this work, we propose to use ERMG (Erdös-Rényi Mixture for Graphs) as an alternative to the Erdös model [6]. This model has been developed to fit the connection heterogeneity which is observed in real networks. Its core hypothesis is that nodes are spread among Q hidden classes with proportion $\alpha_1, \ldots, \alpha_Q$. Denoting by Z_is the independent random variables which equal q if node i belongs to class q, the conditional distribution of X_{ij} is such that:

$$X_{ij} | \{Z_i = q, Z_j = \ell\} \sim \mathcal{B}(\pi_{q\ell}).$$

Consequently, the marginal distribution of $\mathbf{X}$ is a mixture of Bernoulli distributions. Under the ERMG model, the probability of occurrence of motif $\mathbf{m}$ is: $\mu(\mathbf{m}) = \sum_{c_1=1}^{Q} \ldots \sum_{c_k=1}^{Q} \alpha_{c_1} \ldots \alpha_{c_k} \prod_{1 \leq u < v \leq k} \pi_{c_u c_v}^{m_{uv}}$.

5 Compound Poisson approximation

The knowledge of the moments of the count under some null model is not sufficient to assess its significance. To decide whether a motif $\mathbf{m}$ is over-represented in a given network, one typically needs to calculate the probability $\Pr\{N(\mathbf{m}) \geq N_{\text{obs}}(\mathbf{m})\}$, where $N_{\text{obs}}(\mathbf{m})$ is the observed number of occurrences of $\mathbf{m}$ and $N(\mathbf{m})$ the random number of occurrences under the reference model. To do so, we need to specify the distribution of $N(\mathbf{m})$ under the reference model. Unfortunately, even in the Erdös model, the exact distribution seems very difficult to derive, so only an approximate distribution can be proposed at this time.

One particularity of network motifs is that their occurrences naturally tend to overlap. Two occurrences of a motif $\mathbf{m}$ overlap if they share at least one vertex; A motif is overlapping if two of its occurrences may overlap. Consequently all network motifs are actually overlapping. Thus, the approximate distribution must account for the existence of clumps, i.e. sets of overlapping occurrences. Clumps result in numerous occurrences with a reduced number of vertices. For example, an occurrence of the four-branch star motif accounts for 4 overlapping occurrences of the three-branch star motif, i.e. for a clump of size 4 involving only 5 vertices.

Compound Poisson distributions are particularly relevant to describe how the count of an event occurring in clumps may vary. The number of clumps is supposed to have a Poisson distribution with mean λ, and the clump sizes are supposed to be independent with common distribution. The Pólya-Aeppli (denoted by $\mathcal{PA}$) distribution (or geometric Poisson, [9]) is obtained when the clump size has a geometric distribution $\mathcal{G}(1-a)$, so the mean size of a clump is $(1-a)^{-1}$. In this case, the number

of observed events W has distribution $\mathcal{PA}(\lambda, a)$:

$$W \sim \mathcal{PA}(\lambda, a) \quad \Leftrightarrow \quad \Pr\{W = w\} = \begin{cases} e^{-\lambda} a^w \sum_{c=1..w} \frac{1}{c!} \binom{w-1}{c-1} \left[\frac{\lambda(1-a)}{a}\right]^c & \text{if } w > 0, \\ e^{-\lambda} & \text{if } w = 0. \end{cases}$$

We propose to use the Pólya-Aeppli distribution as an approximation of the distribution of the count $N(\mathbf{m})$ for several reasons. (i) This distribution is an excellent approximation (from both a theoretical and a practical point of view) for word counts in random sequences [15]) (ii) The Pólya-Aeppli distribution only involves two parameters that can be easily computed when the first two moments are known. (iii) Finally, simulations show that it fits the observed distribution well for the ERMG model (see next section). Note that argument (i) is not sufficient because the topology of a network motif is quite different from the topology of a sequence motif. Still, parameter a can be interpreted as the overlapping probability of the motif $\mathbf{m}$, i.e. the probability that an occurrence of $\mathbf{m}$ overlaps another one.

The first two moments of the $\mathcal{PA}(\lambda, a)$ distribution are $\lambda/(1-a)$ and $\lambda(1+a)/(1-a)^2$. Given these moments, parameters can be calculated as:

$$a = [\mathbb{E}N(\mathbf{m}) - \mathbb{V}N(\mathbf{m})]/[\mathbb{E}N(\mathbf{m}) + \mathbb{V}N(\mathbf{m})], \quad \lambda = (1-a)\mathbb{E}N(\mathbf{m}). \tag{4}$$

p-values can be calculated in a quadratic time using the algorithm given in [14].

6 Simulation study

The objective of the simulations is to compare the Gaussian, Poisson and Pólya-Aeppli approximations of the distribution of the motif counts. We simulate networks following an ERMG model with two groups. Connectivity parameters are $\pi_{11} = \pi_{22} = \eta\gamma$ and $\pi_{12} = \pi_{21} = \eta(1-\gamma)$, where η is a scale parameter and γ characterizes the between and within group connectivities. The proportions of the groups are α and $1-\alpha$. The mean connectivity is $\overline{\pi} = \alpha^2\pi_{11} + 2\alpha(1-\alpha)\pi_{12} + (1-\alpha)^2\pi_{22}$. Parameters α (0.1 and 0.5), γ (0.1, 0.5 and 0.9), n (20 and 200), and $\overline{\pi}$ ($1/n$ and $2/n$), have been chosen to cover a large range for $\mathbb{E}(N(\mathbf{m}))$ (from 0.07 to 1075.5). This design would lead to $3 \times 2^3 = 24$ cases. However the ERMG with $\gamma = 0.5$ is the Erdös model with $p = \overline{\pi}$ for any α, a feature which reduces the number of cases to 20. The numbers of simulations are 10 000 for $n = 20$ and 1 000 for $n = 200$. The motifs studied, defined in Table 2, are of size 3 or 4 and have a high or low self-overlapping structure.

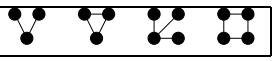
Motif	
Name	V ▽ ⩘ □

Table 2: Motifs used in the simulation

There is a good agreement between the exact and the empirical means and variances for all the simulated cases (results not shown, see supplementary material). We can see in Table 3 that, excepted for rare motifs, the variance is much higher than the mean, so that the Poisson approximation is not valid. This is illustrated by Figure 1 (second and third rows).

Parameters of the approximate Gaussian distribution are $\mathbb{E}(N(\mathbf{m}))$ and $\mathbb{V}(N(\mathbf{m}))$. Parameters ($\lambda$, a) of the approximate Pólya-Aeppli distribution are computed using formula (4). Parameter values are given in Table 3 for motifs V and □. Motifs V and ⩘ have a high overlapping probability a, which

means that they are more clumped than the two other motifs. For example, parameter a for motif $\nwarrow$ is equal to 0.982 in the case $n = 200$, $\overline{\pi} = 0.01$, $\alpha = 0.1$, $\gamma = 0.1$. According to the Pólya-Aeppli paradigm, the picture is the following: motifs $\nwarrow$ occur in clumps with mean size $1/(1-a) = 55.6$, the mean number of clumps being equal to $\lambda = 19.28$. As the Pólya-Aeppli distribution is only an approximation of the true distribution, we do not claim that the latter description is exact. It only gives an interesting indication about the clumping of the motifs.

Histograms and PP-plots (Figure 1), for the motifs V and □ are given in three cases ($\mathbb{E}(N(\mathbf{m})) \simeq 1$, $\mathbb{E}(N(\mathbf{m})) \simeq 10$ and $\mathbb{E}(N(\mathbf{m})) \simeq 100$) in Figure 1. Table 3 contains the results about the quality of approximation of respectively the Gaussian and the Pólya-Aeppli distributions based on the exact first two moments for the motifs V and □ (corresponding tables for the motifs $\bigtriangledown$ and $\nwarrow$ are presented in the supplementary material). The quality of the Poisson approximation is not given, because it is clearly not valid for frequent motifs, and is not different from the Pólya-Aeppli approximation for rare motifs. The comparison is based on two criteria: (1) the total variation distance $D = \frac{1}{2}\sum_i |o_i - t_i|$ between the theoretical and empirical distributions, where o_i and t_i are respectively the observed and theoretical frequency for the count i. This criterion allows a comparison along the whole range of the random variable. In practice we are often more concerned by the tails of the distribution for computing the p-values of the counts. (2) $\hat{F}(Q_G)$ and $\hat{F}(Q_P)$ are the empirical probability of exceeding the 0.99 Gauss (Q_G) and Pólya-Aeppli (Q_P) quantiles respectively. This criterion should be close to 0.01. Similar results are obtained for the lower tail of the distribution (not shown).

Three conclusions appear: (i)The Pólya-Aeppli approximation outperforms the Gaussian approximation for both criteria in all cases.(ii) The 0.99 quantile is underestimated by the Gaussian approximation. This implies that using z-scores leads to many false positives. On the opposite this quantile is well estimated by the Pólya-Aeppli approximation. However the tail values greater than 0.99, such as 0.999 or 0.9999, have not been explored because a too small number of simulations, so that this conclusion is limited to the studied range of moderate tail values. (iii)The total variation distance between approximate and empirical distributions is high for both approximations in some cases, especially for frequent and highly self-overlapping motifs. This is explained by the odd distribution of the simulation counts: it is not smooth, has many modes and present periodic patterns (see Figure 1). This could be due by the clump size distribution which is not geometric and seems to have several modes. However, even in these cases, the Pólya-Aeppli approximation of the 0.99 quantile is good.

7 Conclusion

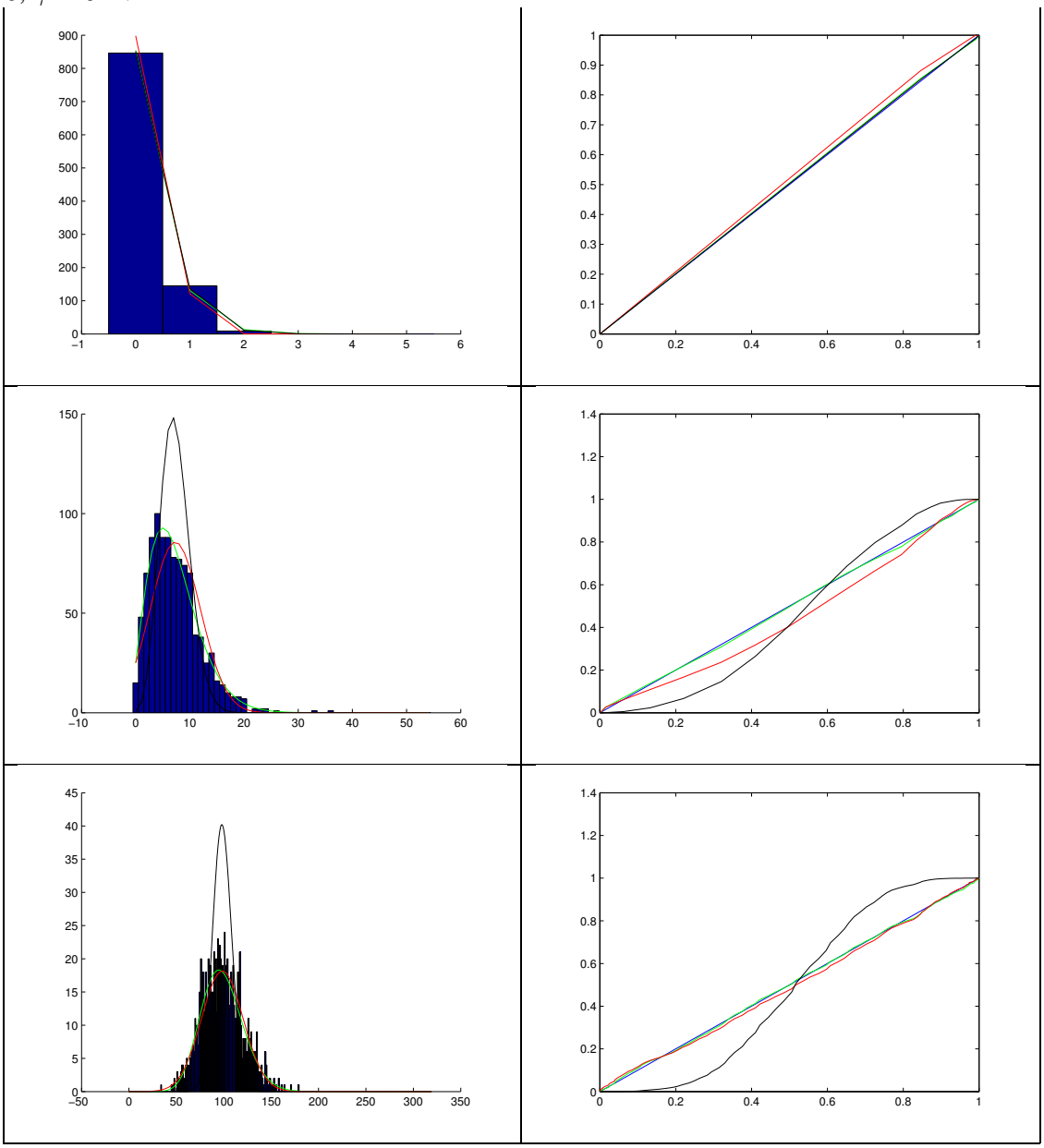
We provide an exact method to calculate the mean and variance of the count of any network motif, whatever its topology. These formula hold for any stationary random graph model. The Pólya-Aeppli approximation is accurate in a large range of situations, and we demonstrated that the Gaussian and Poisson approximations are not satisfactory. Consequently, strategies based on z-scores are not reliable. In addition, p-values can be easily computed thanks to the Pólya-Aeppli approximation we propose. This direct computation avoids simulations which should be very numerous to be accurate in the case of small p-values. Typically, a p-value of about 10^3 would require at least 10^5 simulations.

The generalization of our method to the oriented case is straightforward. In this case adjacency matrices $\mathbf{X}$ and $\mathbf{m}$ are not symmetric anymore, and formula to calculate the mean and variance still hold. The ERMG model can also be used for directed graphs.

Table 3: Mean and variance of the number of occurrences of motifs V and □ in an ERMG network defined by n, α, γ and $\overline{\pi}$, and quality of the Gaussian and Pólya-Aeppli approximations. $\mathbb{E}$ and $\mathbb{V}$: exact mean and variance of the motif count. λ, a: corresponding parameters of the Pólya-Aeppli distribution. DG and DP: total variation distances between the empirical distribution and the Gaussian and the Pólya-Aeppli distributions respectively. $\hat{F}_G$ and $\hat{F}_P$: empirical probabilities of exceeding the 0.99 Gaussian and Pólya-Aeppli quantiles. $\alpha, \gamma, \overline{\pi}, DG, DP, \hat{F}_G$ and $\hat{F}_P$ are expressed in percentage.

simulation parameters				motif V								motif □							
n	$\overline{\pi}$	α	γ	$\mathbb{E}$	$\mathbb{V}$	λ	a	DG	DP	$\hat{F}_G$	$\hat{F}_P$	$\mathbb{E}$	$\mathbb{V}$	λ	a	DG	DP	$\hat{F}_G$	$\hat{F}_P$
20	5	10	10	13.84	142.8	2.45	0.82	18.2	6.8	3.1	1.1	0.34	0.96	0.18	0.47	36.0	5.2	3.7	0.9
20	5	10	90	9.1	44.6	3.08	0.66	15.6	5.1	3.2	1.2	0.12	0.17	0.1	0.17	2.8	1.4	10.3	0.3
20	5	50	10	8.55	35.9	3.29	0.62	14.8	4.4	2.9	1.3	0.13	0.18	0.11	0.17	2.2	1.0	10	0.5
20	5	50	50	8.55	37.0	3.21	0.62	15.9	5.3	3.2	1.1	0.09	0.12	0.08	0.13	12.1	0.6	7.8	0.9
20	5	50	90	8.55	38.3	3.12	0.63	15.6	5.7	3.4	1.4	0.13	0.19	0.1	0.2	3.4	1.6	10.5	0.5
20	10	10	10	55.38	1406.3	4.2	0.92	11	9.2	2.1	0.5	5.47	75.69	0.74	0.87	34.8	18.6	3.8	1
20	10	10	90	36.41	330.9	7.22	0.8	12.3	4.4	3	1.5	1.93	6.40	0.89	0.54	26.8	11.0	4.4	1.5
20	10	50	10	34.2	236.2	8.65	0.75	10.3	4.3	2.5	1.1	2.05	6.55	0.98	0.52	25.1	11.2	4.4	1.1
20	10	50	50	34.2	249.3	8.25	0.76	10.7	4.1	2.6	1.3	1.45	3.84	0.8	0.45	25.7	8.5	4.6	1.3
20	10	50	90	34.2	267.0	7.76	0.77	12.5	5.9	2.8	1.4	2.05	7.84	0.85	0.59	26.9	14.8	3.4	1.1
200	0.5	10	10	159.5	2034.0	23.1	0.85	20.4	19.7	2.5	1.6	0.46	0.58	0.41	0.11	21.8	1.8	7.9	0.7
200	0.5	10	90	104.9	590.5	31.6	0.7	15.2	14	1.9	1.2	0.16	0.17	0.16	0.03	4.2	1.4	15.4	0.9
200	0.5	50	10	98.5	484.0	33.3	0.66	13.1	12.6	1.1	0.7	0.17	0.18	0.17	0.03	4.0	2.5	17.8	1.9
200	0.5	50	50	98.5	484.0	33.2	0.66	14.3	13.2	1.6	1.1	0.12	0.13	0.12	0.02	12.1	0.5	11.7	0.9
200	0.5	50	90	98.5	488.4	33.1	0.66	14.5	14.8	2.5	0.9	0.17	0.18	0.17	0.03	3.1	1.6	16.9	1.9
200	1	10	10	638	21345.2	37.1	0.94	32.8	32.7	1.5	1.2	7.31	21.72	3.68	0.5	11.8	5.4	3.2	0.9
200	1	10	90	419.4	4637.6	69.7	0.83	23.1	22.9	2	1.5	2.57	3.42	2.21	0.14	9.3	2.7	3.6	0.5
200	1	50	10	394	3457.4	80.6	0.8	21.4	21	2.9	0.8	2.74	3.69	2.33	0.15	12.3	3.6	4.7	1.2
200	1	50	50	394	3457.4	80.2	0.8	20.8	20.6	1.3	0.8	1.94	2.40	1.74	0.1	11.3	2.0	3.2	1.6
200	1	50	90	394	3492.8	79.8	0.8	19.8	19.7	1.4	1	2.74	3.72	2.32	0.15	10.8	4.5	3.7	0.7

Figure 1: Empirical, Gaussian (red), Poisson (black) and Pólya-Aeppli (green) distributions. Left column: histograms, right column: PP-plots. First row: rare motif, number of motif □ in an ERMG with $n = 200$,$\overline{\pi} = 0.005$, $\alpha = 0.1$, $\gamma = 0.9$, second row: medium motif, number of motif □ in an ERMG with $n = 200$,$\overline{\pi} = 0.01$, $\alpha = 0.1$, $\gamma = 0.1$, third row: frequent motif, number of motif V in an ERMG with $n = 200$,$\overline{\pi} = 0.005$, $\alpha = 0.5$, $\gamma = 0.1$.



References

[1] A. L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[2] A.D. Barbour, M. Karoński, and A. Ruciński. A central limit theorem for decomposable random variables with applications to random graphs. *J. Combinat. Theory, series B*, 457:125–145, 1987.

[3] B. Bollobas. Random graphs. In *Combinatorics*, London Math. Soc. lecture Note Ser. 52. Cambridge University Press, 1981.

[4] Matias C., Schbath S., Birmelé E., Daudin J-J, and Robin S. Network motifs: mean and variance for the count. *REVSTAT*, 4(1):1–20, 2006.

[5] J. Chen and B. Yuan. Detecting functional modules in the yeast protein-protein intereaction network. *Bioinformatics*, 22:2283–2290, 2006.

[6] J-J. Daudin, F. Picard, and S. Robin. A mixture model for random graphs. Research Report RR-5840, INRIA, France, February 2006.

[7] P. Erdös and A. Rényi. On random graphs. *Publicationes Mathematicae*, 6:290–297, 1959.

[8] S. Janson, A. Rucinski, and T. Luczak. *Random graphs*. Wiley, 2000.

[9] N. L. Johnson, S. Kotz, and A. W. Kemp. *Univariate Discrete Distributions*. Wiley: New-York, 1992.

[10] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon. Superfamilies of evolved and designed networks. *Science*, 303:1538–1542, 2004.

[11] R. Milo, N. Kashtan, S. Itzkovitz, M.E.J. Newman, and U. Alon. On the uniform generation of random graphs with prescibed degree sequences. *cond-mat*, 0312028, 2004.

[12] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Networks motifs: simple building blocks of complex networks. *Science*, 298:824–827, 2002.

[13] M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E*, 64(2):026118, 2001.

[14] S. Robin. A compound Poisson model for words occurrences in DNA sequences. *J. R. Statist. Soc. C*, 51:437–451, 2002.

[15] S. Schbath. Compound Poisson approximation of word counts in DNA sequences. *ESAIM: Probability and Statistics*, 1:1–16. (http://www.emath.fr/ps/), 1995.

[16] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Networks motifs in the transcriptional regulation network of *escherichia coli*. *Nature Genetics*, 31:64–68, 2002.

[17] D. Stark. Compound poisson approximation of subgraph counts in random graphs. *Random Structures and Algorithms*, 18:39–60, 2001.