HAL
open science

Compound Poisson Approximation and Testing for Gene Clusters with Multigene Families

Simona Grusea, Etienne Pardoux, Olivier Chabrol, Pierre Pontarotti

To cite this version:

Simona Grusea, Etienne Pardoux, Olivier Chabrol, Pierre Pontarotti. Compound Poisson Approximation and Testing for Gene Clusters with Multigene Families. Journal of Computational Biology, 2011, 18 (4), pp.579-594. 10.1089/cmb.2010.0043 . hal-00636411

HAL Id: hal-00636411

https://hal.science/hal-00636411v1

Submitted on 27 Oct 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 18, Number 4, 2011
© Mary Ann Liebert, Inc.
Pp. 579–594
DOI: 10.1089/cmb.2010.0043

Compound Poisson Approximation and Testing for Gene Clusters with Multigene Families

S. GRUSEA,[1,2] E. PARDOUX,[1] O. CHABROL,[1] and P. PONTAROTTI [1]

ABSTRACT

We present in this article a compound Poisson approximation for computing probabilities involved in significance tests for conserved genomic regions between different species. We consider the case when the conserved genomic regions are found by the reference region approach. An important aspect of our computations is the fact that we are taking into account the existence of multigene families. We obtain convergence results for the error of our approximation by using the Stein-Chen method for compound Poisson approximation.

Key words: compound Poisson approximation, multigene families, reference-region approach, significance test for gene clusters, Stein-Chen method.

1. INTRODUCTION

ORTHOLOGOUS GENES ARE TWO GENES, IN TWO DIFFERENT SPECIES, that descend from the same gene at the ancestor of the two species, as the result of a speciation event. We call *conserved genomic region* or *gene cluster* two chromosomic regions, in two different species, that have in common a certain number of orthologous genes, not necessarily adjacent or in the same order in the two genomes. We do not impose any restriction on the gap length between consecutive orthologs. In the literature, various definitions for gene clusters exist (Bergeron et al., 2002; Danchin et al., 2004; Durand et al., 2003; Hoberman and Durand, 2005, Hoberman et al., 2005; Raghupathy et al., 2005; 2009). We have chosen here a very unrestrictive definition, in order to be able to detect evolutionary signals even between very distant species. The conserved genomic regions can represent signs of evolutionary relatedness between species or of functional selective pressures acting on certain groups of genes. But for this to be the case, the conserved genomic regions have to be *significant*, i.e., very improbable to have appeared by chance.

During the evolutionary time, the gene order in one genome can be affected by various genome rearrangement events such as inversions, translocations, transpositions, chromosomic fissions, and fusions. Therefore, in the absence of certain constraints due to functional selective pressures, the gene order is randomized during evolution. This is one reason why, in general, the null hypothesis taken in the significance tests for gene clusters is the hypothesis of random gene order.

There exist different approaches when searching for gene clusters (Durand et al., 2003). In this article, we focus on the case when the gene clusters are found by the ''reference region'' approach, which consists

[1]LATP–UMR CNRS 6632, Équipe Évolution Biologique et Modélisation, Université de Provence, Marseille, France.
[2]Institut de Mathématiques de Toulouse, INSA de Toulouse, Université de Toulouse, Toulouse, France.

in starting with a fixed genomic region in a certain species A (called the *reference region*) and searching for significant orthologous gene clusters in the genome of another species B.

In general, the orthology relation between the genes of two species is not one-to-one. For a given gene in one species, we may find more than one orthologous gene in another species, as the result of duplication events happened after the separation of the two species. The genes in one species which are orthologous to the same gene in another species are called *co-orthologs* of this gene and form what we call a *multigene family*. The existence of multigene families is an important fact which needs to be considered when testing for gene cluster significance, but very few of the existent statistical tests consider it. Danchin et al. (2004) propose to weigh the orthologs in inverse proportion to the sizes of the multigene families, but their use of a binomial distribution is not adequate in these settings. Raghupathy et al. (2005, 2009) take also into account the existence of multigene families, but their test is suitable only for clusters found by the window-sampling approach, and not by the reference region approach, as in our case.

In this article, we adopt the idea of Danchin et al. (2004) for taking into account the multigene families and propose a compound Poisson approximation for computing the probabilities, under the null hypothesis, of different gene clusters.

The article is organized as follows. In Section 2, we present the mathematical framework. We explain the simplified mathematical model that we use and we describe the way in which we take into account the existence in the genome B of multiple co-orthologs for the genes in the reference region. We give the mathematical formulation of the problem and we start by considering, for technical convenience, the case of a circular genome. In Subsection 2.3, we give a very short presentation of the Stein-Chen method for compound Poisson approximation—the coupling approach. We present in Theorem 1 a result of Roos (1993), which gives a convergence result for the error of the approximation under the existence of a certain coupling. Section 3 is the core of the article, containing the compound Poisson approximations for our probability of interest, together with the convergence results that we have obtained using the Stein-Chen method. Following the approach of Roos (1993a,b), we construct explicitly the coupling needed in Theorem 1, and we estimate the terms appearing in the error bound in Theorem 1. Theorem 2 states the obtained convergence result. In Subsection 3.2, we describe a "Markovian" approximation for computing, in practice, the parameters of the compound Poisson distribution. In Subsection 3.3, we extend the results to the case of a linear genome. In Section 4, we present some numerical results, both in the circular and in the linear case, for a set of selected values for the parameters which are interesting in our biological framework. We also discuss the biological implications of our results. Section 5 presents three applications of our results on real biological data. We analyze three examples: the first one is a comparison between the human genome and the genome of Oryzias-Latipes, the second one is a comparison between the human genome and the genome of Ciona-Intestinalis, and the third one is a comparison between the human genome and the genome of Danio-Rerio.

2. MATHEMATICAL FRAMEWORK

2.1. Mathematical formulation of the problem

We model the genome as an ordered set of genes, the length of a genomic region being measured in number of genes. We ignore the separation into chromosomes and the physical distances between genes.

The data that we dispose of are: the number m of genes in the reference region from the genome A which have at least one ortholog in the genome B; for each of those genes $i = 1, \ldots, m$, the number of orthologs it has in B, which we denote ϕ_i; the positions in B of these orthologs; the total number N of genes in the genome B.

We make the (natural) assumption that there exists a maximal size $\phi_{\max}$ for the multigene families. Based on the fact that we are in the case $m << N$, we make a further approximation and consider the genome B as the continuous interval [0, 1], in which the "new" positions of the orthologs are obtained by dividing by N their real positions in the genome.

We will use a pure significance test, with the null hypothesis H_0 : *random gene order in the genome B*. All the probabilities and distributions appearing throughout the paper are implicitly considered under the null hypothesis H_0.

For $i = 1, \ldots, m$, we let $U_{ij}, j = 1, \ldots, \phi_i$ represent the positions in B of the orthologs of the gene i from the reference region. Under H_0, the r.v.'s $U_{ij}, j = 1, \ldots, \phi_i$, $i = 1, \ldots, m$ are i.i.d. uniformly distributed on [0, 1].

Let $n := \phi_1 + \cdots + \phi_m$ denote the total number of genes in B which are orthologous to genes in the reference region in A. We are interested only in these n genes. We want to test whether they cluster together in a significant way, i.e., in a way which is very improbable by chance, under the null hypothesis.

For taking into account the existence in B of multiple orthologs for the genes in the reference region, we consider the following counting measure:

$$\mu_m := \sum_{i=1}^{m} \frac{1}{\phi_i} \sum_{j=1}^{\phi_i} \delta_{U_{ij}}.$$

For an ortholog belonging to a multigene family of size ϕ_i, we call $\frac{1}{\phi_i}$ its *label*. For an interval $I \subset [0,1]$, we will refer to $\mu_m(I)$ as its *weight*.

For applying a statistical test we need to compute, for a given weight h and a given length r, the probability, under the null hypothesis, of finding somewhere in the genome B an orthologous cluster of weight greater than h and of length smaller than r. We will call such a cluster *of type* $(h : r)$.

In this article, we focus on the computation of this probability. For technical simplifications, we first consider the case when B is a circular genome, hence the circle of length 1 in our model.

2.2. The circular case

Let h be fixed, of the form $h = \sum_{i=1}^{m} \frac{n_i}{\phi_i}$, with $0 \le n_i \le \phi_i, i = 1, \ldots, m$.

Let $r \in (0,1)$ be also fixed.

We denote by $U_{(1)} \le U_{(2)} \le \cdots \le U_{(n)}$ the ordered positions in B of the n orthologs, i.e. the order statistics of n i.i.d. r.v.'s uniformly distributed on the circle of length 1.

Let $W_m = W_m(h, r)$ denote the r.v. representing the number of (possibly overlapping) clusters of type $(h : r)$ in the genome B. Let also

$$A_k = A_k(h, r) := \{\mu_m([U_{(k)}, U_{(k)} + r]) \ge h\}$$

denote the event of having in B a cluster of type $(h : r)$ starting with the k-th ortholog.

We have $W_m = \sum_{k=1}^{n} 1_{A_k}$. We are interested in computing $\mathbb{P}(W_m \ge 1)$, the probability of finding, somewhere in the genome B, at least one cluster of type $(h : r)$.

We will further simplify the parametrization of the problem.

Let $\phi'_1 < \cdots < \phi'_J$ denote all the different values among the multigene families' sizes $\phi_1, \ldots, \phi_m$, and let $g_j = |\{i = 1, \ldots, m : \phi_i = \phi'_j\}|, j = 1, \ldots, J$ denote their multiplicities.

Remark 1. *We can represent the measure μ_m as $\mu_m = \sum_{i=1}^{n} L_i \delta_{U_{(i)}}$, where δ_x denotes the Dirac measure in x and $\boldsymbol{L} = (L_1, \ldots, L_n)$ is a random vector independent of the $U_{(i)}$'s and uniformly distributed over the set of all possible labelings of the n orthologs:*

$$\Lambda = \left\{ \ell = (\ell_1, \ldots, \ell_n) \in \left\{ \frac{1}{\phi'_1}, \ldots, \frac{1}{\phi'_J} \right\}^n : \left| \left\{ i : \ell_i = \frac{1}{\phi'_j} \right\} \right| = g_j \phi'_j, \forall j \right\}.$$

Let $n_{\min} := g_1 = |\{i : \phi_i = \phi'_1\}|$, where $\phi'_1 = \min\{\phi_i : i = 1, \ldots, m\}$. We let also $h_* := \lceil h\phi'_1 \rceil$ and we assume that $n_{\min} \ge h_*$, s.t. h_* is the minimal number of orthologs in a cluster of weight greater than h.

For every labeling $\ell \in \Lambda$ and every $k = 1, \ldots, n$, let

$$h_k(\ell) := \min\{d : \ell_k + \cdots + \ell_{k+d-1} \ge h\}$$

be the minimal number of orthologs in a cluster starting with the k-th ortholog so as to be of weight greater than h. Therefore,

$$A_k \cap \{\boldsymbol{L} = \ell\} = \{U_{(k+h_k(\ell)-1)} - U_{(k)} \le r\} \cap \{\boldsymbol{L} = \ell\}.$$

Let $h^* := \max_{\ell,k}\{h_k(\ell)\}$.

We have $h^* \le \lceil h\phi'_J \rceil$, where $\phi'_J = \max\{\phi_i : i = 1, \ldots, m\}$.

Remark 2. *We place ourselves in the asymptotic settings of $m \to \infty$, or equivalently, $n \to \infty$.*

Note that we are in the case of a sum of indicators which are in a *short-range dependence* and a *long-range (almost) independence*. Because of the strong dependence between the neighboring indicators, the events A_k will tend to occur in clumps. Consequently, it seems reasonable to approach the distribution of W_m by a compound Poisson distribution $CP(\lambda)$, with a "good" choice of the parameter λ.

We will quantify the error using the Kolmogorov distance. We recall that the Kolmogorov distance between two measures μ and ν on $\mathbb{R}_+$ is

$$d_K(\mu,\nu) = \sup_{k\in\mathbb{N}} |\mu([k,\infty) - \nu([k,\infty))|.$$

We will approximate our probability of interest $\mathbb{P}(W_m \geq 1)$ by the corresponding probability $CP(\lambda)([1,\infty)) = 1 - \exp\{-\sum_{i=1}^{\infty} \lambda_i\}$ for the compound Poisson distribution, with an error

$$|\mathbb{P}(W_m \geq 1) - CP(\lambda)([1,\infty))| \leq d_K(\mathcal{L}(W_m), CP(\lambda)).$$

It is therefore sufficient to obtain bounds for the Kolmogorov distance between the two distributions. For bounding the Kolmogorov distance, we will use the Stein-Chen method for compound Poisson approximation.

The Stein-Chen method is a general method to obtain bounds on the distance between two probability distributions with respect to a probability metric. It was originally formulated for normal approximations by Stein (1972), to obtain a bound for the Kolmogorov distance between the distribution of a sum of m-dependent sequence of random variables and a standard normal distribution. The Poisson approximation version was developed by Chen (1975). See also Stein (1986).

The Stein-Chen method for compound Poisson approximation was introduced by Barbour et al. (1992). In this article, we use, more precisely, the *coupling approach* developed by Roos (1993 a,b).

2.3. The Stein-Chen method, the coupling approach

Let $W = \sum_{\alpha\in\Gamma} I_\alpha$ be a sum of indicators. We assume that there exists a local-dependence structure between these indicators (of the type short-range dependence, long-range independence), so that we can, for every $\alpha \in \Gamma$, divide Γ into four disjoint subsets $\{\alpha\}$, Γ_α^{vs}, Γ_α^{vw} and Γ_α^{b}:

$$\begin{aligned}
\Gamma_\alpha^{vs} &:= \{\beta \in \Gamma\backslash\{\alpha\} : I_\beta \text{ ``very strongly'' dependent on } I_\alpha\},\\
\Gamma_\alpha^{vw} &:= \{\beta \in \Gamma\backslash\{\alpha\} : I_\beta \text{ ``very weakly'' dependent on } \{I_\gamma, \gamma \in \{\alpha\} \cup \Gamma_\alpha^{vs}\}\},\\
\Gamma_\alpha^{b} &:= \Gamma\backslash\{\{\alpha\} \cup \Gamma_\alpha^{vs} \cup \Gamma_\alpha^{vw}\}.
\end{aligned}$$

We let $U_\alpha := \sum_{\beta\in\Gamma_\alpha^{vs}} I_\beta, Z_\alpha := I_\alpha + U_\alpha, X_\alpha := \sum_{\beta\in\Gamma_\alpha^{b}} I_\beta$.

The *canonical choice* for the parameter of the approximating compound Poisson distribution is the following: $\lambda = \sum_{i=1}^{G+1} \lambda_i\delta_i$, where $G = \max_{\alpha\in\Gamma}\{|\Gamma_\alpha^{vs}|\}$ and

$$\lambda_i = \frac{1}{i}\sum_{\alpha\in\Gamma} \mathbb{E}(I_\alpha \mathbf{1}_{\{Z_\alpha = i\}}).$$

In practice, it is not always easy to compute the canonical parameters. Sometimes it is useful to keep only a smaller number of parameters: $\hat{\lambda} = \sum_{i=1}^{\ell} \hat{\lambda}_i\delta_i$, with $\ell < G+1$, where $\hat{\lambda}_i = \lambda_i$ for $i = 2, \ldots, \ell$, $\hat{\lambda}_i = 0$ for $i \geq \ell+1$ and $\hat{\lambda}_1 = \mathbb{E}(W) - \sum_{i=2}^{\ell} i\lambda_i$.

For every $\alpha \in \Gamma$, let V_α be a r.v. and $\mathcal{V}_\alpha$ its set of values.

We have the following theorem (see Theorem 4.F. in Roos, 1993a).

Theorem 1. *Assume that for every $\alpha \in \Gamma$ and $v \in \mathcal{V}_\alpha$ we can construct, on the same probability space, the indicators $\{I''_{\beta i v}(\alpha), \beta \in \Gamma, i = 1, \ldots, |\Gamma_\alpha^{vs}| + 1\}$ and $\{I'_{\beta v}(\alpha), \beta \in \Gamma\}$ in such a way that*

$$\mathcal{L}(I''_{\beta i v}(\alpha), \beta \in \Gamma) = \mathcal{L}(I_\beta, \beta \in \Gamma | I_\alpha \mathbf{1}_{\{Z_\alpha = i\}} = 1, V_\alpha = v), \forall i \tag{1}$$

$$\mathcal{L}(I'_{\beta v}(\alpha), \beta \in \Gamma) = \mathcal{L}(I_\beta, \beta \in \Gamma). \tag{2}$$

Then, for all choices of the sets Γ_α^{vs} and Γ_α^{vw} and for all bounded functions $g : g : \mathbb{N} \longrightarrow \mathbb{R}$, we have

$$d_K(\mathcal{L}(W), CP(\hat{\lambda})) \leq c_{\mathcal{K}}(\hat{\lambda})\left\{\sum_{\alpha\in\Gamma}[(\mathbb{E}I_\alpha)^2 + \mathbb{E}I_\alpha\mathbb{E}(U_\alpha + X_\alpha) + \mathbb{E}(I_\alpha X_\alpha)\right.$$
$$\left. + \sum_{i=1}^{|\Gamma_\alpha^{vs}|+1}\sum_{\beta\in\Gamma_\alpha^{vw}}\mathbb{E}(I_\alpha \mathbf{1}_{\{Z_\alpha=i\}}\theta_{\beta,\alpha,i}(V_\alpha))] + \sum_{i=\ell+1}^{G+1} i(i-1)\lambda_i\right\},$$

where $c_{\mathcal{K}}(\hat{\lambda}) = \sup_{A\in\mathcal{K}}\sup_{i\geq 1}|g_{\hat{\lambda},A}(i+1) - g_{\hat{\lambda},A}(i)|$, *with* $\mathcal{K} = \{[k,\infty) : k \in \mathbb{N}\}$ *and* $g_{\hat{\lambda},A}$ *being the solution of the Stein-Chen equation for compound Poisson approximation (Roos, 1993a) and* $\theta_{\beta,\alpha,i}(v) = \mathbb{E}|I''_{\beta iv}(\alpha) - I'_{\beta v}(\alpha)|$.

Barbour et al. (2000) showed that, if the following condition is fulfilled:

$$\lambda_1 \geq 2\lambda_2 \geq 3\lambda_3 \geq \cdots, \tag{3}$$

then

$$c_{\mathcal{K}}(\lambda) \leq \min\left\{\frac{1}{2}, \frac{1}{\lambda_1+1}\right\}. \tag{4}$$

In the next section, we apply this method to our W_m, using Theorem 1.

3. COMPOUND POISSON APPROXIMATION FOR $\mathbb{P}(W_m \geq 1)$

3.1. The circular case

We place ourselves in the asymptotic settings of $n \to \infty$ and $r \to 0$, with $nr \to 0$.

We let $I_k := \mathbf{1}_{A_k}, k = 1, \ldots, n$.

For every $k \in \{1, \ldots, n\}$, we choose the dependence sets as follows:

$$\Gamma_k^{vs} := \{k - h_* + 2, \ldots, k-1, k+1, \ldots, k + h_* - 2\},$$
$$\Gamma_k^{vw} := \{j : |j-k| > 2(h^* - 2)\},$$
$$\Gamma_k^{b} := \Gamma\backslash\{\{\alpha\} \cup \Gamma_\alpha^{vs} \cup \Gamma_\alpha^{vw}\} = \{j : h_* - 2 < |j-k| \leq 2(h_* - 2)\}.$$

Here, $G = \max_{k=1,\ldots,n}|\Gamma_k^{vs}| = 2(h_* - 2)$ and $Z_k = \sum_{j=k-h_*+2}^{k+h_*-2} I_j$.

We will explicitly construct the coupling described in Theorem 1.

Let us define the spacings $S_j := U_{(j+1)} - U_{(j)}$, $j = 1, \ldots, n$, with the circular convention modulo n.

Notation 1. *For a sequence* $(a_j)_{j\geq 1}$ *we will denote* $a_{i,k} := a_i + \cdots + a_{i+k-1}$.

For every $k \in \{1, \ldots, n\}$ and $\ell \in \Lambda$, we have $A_k \cap \{\boldsymbol{L} = \ell\} = \{S_{k,hk(\ell)-1} \leq r\}$.

Let $k \in \{1, \ldots, n\}$ be fixed. The indicators appearing in the expression of Z_k are those from I_{k-h^*+2} to I_{k+h^*-2}. Consequently, if $\boldsymbol{L} = \ell$, the spacings appearing in the expression of Z_k are $S_{k-h_*+2}, \ldots, S_{k+h^*+hk+h^*-2(\ell)-4}$.

Let $V_k := (\boldsymbol{L}, S_{k-h_*+2}, \ldots, S_{k+h_*+h^*-4})$. Note that V_k contains all the spacings which may appear in Z_k, for different values of ℓ.

For every $v = (\ell, z_1, \ldots, z_{2h_*+h^*-5})$, with $\ell \in \Lambda, z_1, \ldots, z_{2h_*+h^*-5} > 0$ and $z_1 + \cdots + z_{2h_*+h^*-5} < 1$, we will construct on the same probability space the indicators $\{I''_{jiv}(k), j = 1, \ldots, n\}$ and $\{I'_j(k), j = 1, \ldots, n\}$ (not depending on v) verifying the relations (1) and (2) in Theorem 1.

Note that the event $\{I_k \mathbf{1}_{\{Zk=i\}} = 1\}$ is V_k - measurable and thus, for the condition (1) to be fulfilled, it suffices to construct the family of indicators $\{I'_j(k), j = 1, \ldots, n\}$ (not depending on i), s.t.

$$\mathcal{L}(I''_{jv}(k), j = 1, \ldots, n) = \mathcal{L}(I_j, j = 1, \ldots, n | V_k = (\ell, z_1, \ldots, z_{2h_*+h^*-5})).$$

Let $U'_1, \ldots, U'_n$ be r.v.'s independent on $\boldsymbol{L}$ and such that $\mathcal{L}(U'_1, \ldots, U'_n) = \mathcal{L}(U_{(1)}, \ldots, U_{(n)})$.

Define the corresponding spacings $S'_j = U'_{j+1} - U'_j, \forall j = 1, \ldots, n$ (with the circular convention $U'_{n+1} = U'_1$). We thus have $\mathcal{L}(S'_1, \ldots, S'_n) = \mathcal{L}(S_1, \ldots, S_n)$.

For $v = (\ell, z_1, \ldots, z_{2h_*+h^*-5})$ with $\ell \in \Lambda, z_1, \ldots, z_{2h_*+h^*-5} > 0$ and $z_1 + \cdots + z_{2h_*+h^*-5} < 1$, we let

$$S''_j = \frac{1 - \sum_{i=1}^{2h_* + h^* - 5} z_i}{1 - \sum_{i=k-h_*+2}^{k+h_*+h^*-4} S'_i} S'_j, j \in \{1, \ldots, n\} \backslash \{k - h_* + 2, \ldots, k + h_* + h^* - 4\},$$

$$S''_{k-h_*+2} = z_1, \ldots, S''_{k+h_*+h^*-4} = z_{2h_*+h^*-5}.$$

Note that

$$\mathcal{L}(S''_1, \ldots, S''_n) = \mathcal{L}(S_1, \ldots, S_n | S_{k-h_*+2} = z_1, \ldots, S_{k+h_*+h^*-4} = z_{2h_*+h^*-5}).$$

Let also $\mu'_m := \sum_{i=1}^n L_i \delta_{U'_i}$.

For every $j \in \{1, \ldots, n\}$ we construct the indicators needed in Theorem 1 as

$$I'_j(k) := \mathbf{1}_{\{\mu'_m([U'_j, U'_j + r]) \geq h\}}, I''_{jv}(k) := \mathbf{1}_{\{S''_j + \cdots + S''_{j+h_j(\ell)-2} \leq r\}}.$$

It is easy to see that the indicators defined above verify the conditions (1) and (2). It remains to compute all the quantities appearing in the error bound in Theorem 1.

The canonical choice for the parameters of the compound Poisson distribution is $\boldsymbol{\lambda} = \sum_{i=1}^{2h_*-3} \lambda_i \delta_i,$, with $\lambda_i = \frac{1}{i} \sum_{k=1}^n \mathbb{E}(I_k \mathbf{1}_{\{Z_k = i\}})$.

In our approximation we will use only half of the parameters, by truncating at $\ell = h_* - 1$. Instead of λ we will use $\hat{\lambda} = \sum_{i=1}^{h_*-1} \hat{\lambda}_i \delta_i$, where $\hat{\lambda}_i = \lambda_i$ for $i = 2, \ldots, h_* - 1$ and $\hat{\lambda}_1 = \mathbb{E}(W_m) - \sum_{i=2}^{h_*-1} \lambda_i = \lambda_1 + \sum_{i=h_*}^{2h_*-3} i\lambda_i$.

We will approximate the probability of interest $\mathbb{P}(W_m \geq 1)$ by

$$p := 1 - \exp\left\{ - \sum_{i=1}^{h_*-1} \hat{\lambda}_i \right\}.$$

Remark 3. *As the indicators $\{I''_{jv}(k), j = 1, \ldots, n\}$ do not depend on i, also the term $\theta_{j,k}(v) = \mathbb{E}\left| I''_{jv}(\alpha) - I'_{jv}(\alpha) \right|$ appearing in Theorem 1 does not depend on i and thus*

$$d_K(\mathcal{L}(W_m), CP(\hat{\lambda})) \leq c_K(\hat{\lambda}) \left\{ \sum_{k=1}^n \{(\mathbb{E}I_k)^2 + \mathbb{E}I_k \mathbb{E}(U_k + X_k) + \mathbb{E}(I_k X_k) \right.$$
$$\left. + \sum_{j \in \Gamma_k^{vw}} \mathbb{E}(I_k \theta_{j,k}(V_k))\} + \sum_{i=h_*}^{2h_*-3} i(i-1)\lambda_i \right\},$$

where $U_k = \sum_{j=k-h_*+2}^{k+h_*-2} I_j - I_k$ and $X_k = \sum_{j=k-2h^*+4}^{k-h_*+1} I_j + \sum_{j=k+h_*-1}^{k+2h^*-1} I_j$.

Using classic results on uniform spacings (Pyke, 1965, 1972), one can easily prove

Lemma 2. *For fixed k, assume that $n \to \infty$, $r \to 0$ s.t. $nr \to 0$. Then, uniformly with respect to $0 < nr < 1$, we have*

$$\mathbb{P}(S_{1,k} \leq r) = \frac{(nr)^k}{k!}(1 + \mathcal{O}\left(\frac{1}{n}\right) + \mathcal{O}(nr))$$

and for fixed i and j,

$$\text{if } i < k : \mathbb{P}(S_{1,i} \leq r, S_{k,j} \leq r) = \frac{(nr)^{i+j}}{i!j!}(1 + \mathcal{O}\left(\frac{1}{n}\right) + \mathcal{O}(nr)),$$

$$\text{if } i \geq k : \mathbb{P}(S_{1,i} \leq r, S_{k,j} \leq r) = \frac{(2k-i+j-2)!}{(k+j-1)!(k-1)!(k-i+j-1)!}(nr)^{k+j-1}$$
$$\times (1 + \mathcal{O}\left(\frac{1}{n}\right) + \mathcal{O}(nr)).$$

For every $k=1, \ldots, n$, we have $\mathbb{E}(I_k)=\frac{1}{|\Lambda|}\sum_{\ell\in\Lambda}\mathbb{P}(S_{k,h_k(\ell)-1}\le r)$.

For every $\ell\in\Lambda$, from Lemma 1 and the exchangeability of the spacings, we have

$$\mathbb{P}(S_{k,h_k(\ell)-1}\le r)=\mathbb{P}(S_{1,h_k(\ell)-1}\le r)=\frac{(nr)^{h_k(\ell)-1}}{(h_k(\ell)-1)!}(1+\mathcal{O}\left(\frac{1}{n}\right)+\mathcal{O}(nr)).$$

We hence obtain, for $0<nr<1$, the following upper bound:

$$\mathbb{E}(I_k)\le\frac{(nr)^{h_*-1}}{(h_*-1)!}(1+\mathcal{O}\left(\frac{1}{n}\right)+\mathcal{O}(nr)). \tag{5}$$

We make the following (biologically realistic) assumption on the data.

Assumption 1. *We assume that we have $n_{\min}\asymp n$, i.e., $n_{\min}=\alpha n(1+\mathcal{O}(\frac{1}{n}))$, with $\alpha\le 1$ fixed.*

Based on Assumption 1, we obtain

$$|\{\ell\in\Lambda : h_1(\ell)=h_*\}|\asymp|\Lambda|. \tag{6}$$

This further implies that $\mathbb{E}(I_k)\asymp(nr)^{h_*-1}$ and $\mathbb{E}(W_m)\asymp n(nr)^{h_*-1}$.

Let $k<j$. We have $\mathbb{E}(I_kI_j)=\frac{1}{|\Lambda|}\sum_{\ell\in\Lambda}\mathbb{P}(S_{k,h_k(\ell)-1}\le r, S_{j,h_j(\ell)-1}\le r)$, where for each $\ell\in\Lambda$, using Lemma 1, we have

$$\begin{aligned}
&\mathbb{P}(S_{k,h_k(\ell)-1}\le r, S_{j,h_j(\ell)-1}\le r)=\mathbb{P}(S_{1,h_k(\ell)-1}\le r, S_{j-k+1,h_j(\ell)-1}\le r)\\
&=\frac{(2(j-k)+h_j(\ell)-h_k(\ell))!}{(j-k)!(j-k+h_j(\ell)-h_k(\ell))!(j-k+h_j(\ell)-1)!}(nr)^{j-k+h_j(\ell)-1}\\
&\quad\times(1+\mathcal{O}(\frac{1}{n})+\mathcal{O}(nr)), \qquad (7)\\
&\quad\text{if } k<j\le k+h_k(\ell)-2 \text{ (we say that } \textit{the two clusters intersect}\text{)}\\
&=\frac{1}{(h_k(\ell)-1)!(h_j(\ell)-1)!}(nr)^{h_k(\ell)+h_j(\ell)-2}(1+\mathcal{O}(\frac{1}{n})+\mathcal{O}(nr)), \qquad (8)\\
&\quad\text{if } j>k+h_k(\ell)-2 \text{ (we say that } \textit{the two clusters do not intersect}\text{)}.
\end{aligned}$$

From Assumption 1, we can also obtain that

$$|\{\ell\in\Lambda : h_k(\ell)=h_*, h_j(\ell)=h_*\}|\asymp|\Lambda|. \tag{9}$$

Next we will estimate the error terms appearing in Theorem 1. We have

Proposition 3. *Assume that $n\to\infty$, $r\to 0$ s.t. $nr\to 0$ and $n_{\min}\asymp n$. Then, uniformly in $\frac{1}{n}\le nr<1$ and $n>2(2h_*+h^*-4)\vee\exp\left\{\frac{4(h_*+h^*-5)}{3(h_*-1)+h^*}\right\}$, we have the following estimates:*

$$(a)\ \sum_{k=1}^{n}(\mathbb{E}I_k)^2\le\frac{n(nr)^{2(h_*-1)}}{[(h_*-1)!]^2}(1+\mathcal{O}(\frac{1}{n})+\mathcal{O}(nr));$$

$$(b)\ \sum_{k=1}^{n}\mathbb{E}(I_k)\mathbb{E}(U_k+X_k)\le 4(h^*-2)\frac{n(nr)^{2(h_*-1)}}{[(h_*-1)!]^2}(1+\mathcal{O}\left(\frac{1}{n}\right)+\mathcal{O}(nr));$$

$$(c)\ \sum_{k=1}^{n}\mathbb{E}(I_kX_k)\le 2(2h^*-h_*-2)\frac{n(nr)^{2(h_*-1)}}{[(h_*-1)!]^2}(1+\mathcal{O}\left(\frac{1}{n}\right)+\mathcal{O}(nr));$$

$$\begin{aligned}(d)\ \sum_{k=1}^{n}\sum_{j\in\Gamma_k^{ww}}\mathbb{E}(I_k\theta_{j,k}(V_k))&\le 2(h_*-1)\{2h_*+h^*-5+2^{h_*-2}(h_*+h^*-4)\}\\
&\quad\times\frac{n(nr)^{2(h_*-1)}}{[(h_*-1)!]^2}(1+\mathcal{O}\left(\frac{1}{n}\right)+\mathcal{O}(nr));\end{aligned}$$

$$(e)\ \sum_{i=h_*}^{2h_*-3}i(i-1)\lambda_i\le(h_*-2)2^{2h_*-5}\frac{n(nr)^{2(h_*-1)}}{[(h_*-1)!]^2}(1+\mathcal{O}\left(\frac{1}{n}\right)+\mathcal{O}(nr)).$$

Proof. The bounds in (*a*) and (*b*) follow easily using (5). ■

Proof of (c). We have

$$\sum_{k=1}^{n} \mathbb{E}(I_k X_k) = \sum_{k=1}^{n} \left\{ \sum_{j=k-2h^*+4}^{k-h_*+1} \mathbb{E}(I_j I_k) + \sum_{j=k+h_*-1}^{k+2h^*-4} \mathbb{E}(I_k I_j) \right\},$$

where $\mathbb{E}(I_k I_j) = \frac{1}{|\Lambda|} \sum_{\ell \in \Lambda} \mathbb{P}(S_{k,h_k(\ell)-1} \le r, S_{j,h_j(\ell)-1} \le r)$.

For $j = k + h_* - 1$, using Lemma 1,

– if $h_j(\ell) = h_*$ and $h_k(\ell) = h_*$, then the two clusters do not intersect and, from (8), we obtain

$$\mathbb{P}(S_{k,h_k(\ell)-1} \le r, S_{j,h_j(\ell)-1} \le r) = \frac{(nr)^{2(h_*-1)}}{[(h_*-1)!]^2}(1 + \mathcal{O}(\frac{1}{n}) + \mathcal{O}(nr));$$

– if $h_j(\ell) = h_*$ and $h_k(\ell) > h_*$, then the two clusters intersect and, using (7), we obtain

$$\mathbb{P}(S_{k,h_k(\ell)-1} \le r, S_{j,h_j(\ell)-1} \le r) \le \frac{1}{2}\frac{(nr)^{2(h_*-1)}}{[(h_*-1)!]^2}(1 + \mathcal{O}(\frac{1}{n}) + \mathcal{O}(nr));$$

– for every other ℓ we have $\mathbb{P}(S_{k,h_k(\ell)-1} \le r, S_{j,h_j(\ell)-1} \le r) = (nr)^{2(h_*-1)}\mathcal{O}(nr)$.

It follows from (9) that $\mathbb{E}(I_k I_{k+h_*-1}) \le \frac{(nr)^{2(h_*-1)}}{[(h_*-1)!]^2}(1 + \mathcal{O}(\frac{1}{n}) + \mathcal{O}(nr))$.

The other cases for j can be treated in a similar manner and the upper bound stated in (*c*) easily follows. ■

Proof of (d). We will condition on the r.v. $V_k = (\boldsymbol{L}, S_{k-h_*+2}, \ldots, S_{k+h_*+h^*-4})$. Given that $V_k = (\ell, z_1, \ldots, z_{2h_*+h^*-5})$, we have $I_k = \mathbf{1}_{\{z_{h_*-1,h_k(\ell)-1} \le r\}}$ (hence deterministic) and thus

$$\sum_{k=1}^{n} \sum_{j \in \Gamma_k^{vw}} \mathbb{E}(I_k \theta_{j,k}(V_k)) = \sum_{k=1}^{n} \sum_{j \in \Gamma_k^{vw}} \frac{1}{|\Lambda|} \sum_{\ell \in \Lambda} d(k,j,\ell),$$

where for each $k = 1, \ldots, n$, $j \in \Gamma_k^{vw}$ and $\ell \in \Lambda$, we let

$$\begin{aligned}
d(k,j,\ell) &:= \mathbb{E}[I_k \theta_{j,k}(V_k) | \boldsymbol{L} = \ell] = d_1(k,j,\ell) + d_2(k,j,\ell),\\
d_1(k,j,\ell) &:= \int \mathbf{1}_{\{z_{h_*-1,h_k(\ell)-1} \le r\}} \mathbb{P}(S''_{j,h_j(\ell)-1} \le r, S'_{j,h_j(\ell)-1} > r)\\
&\qquad dF(z_1, \ldots, z_{2h_*+h^*-5}),\\
d_2(k,j,\ell) &:= \int \mathbf{1}_{\{z_{h_*-1,h_k(\ell)-1} \le r\}} \mathbb{P}(S''_{j,h_j(\ell)-1} > r, S'_{j,h_j(\ell)-1} \le r)\\
&\qquad dF(z_1, \ldots, z_{2h_*+h^*-5}),
\end{aligned}$$

with F being the distribution of $(S_{k-h_*+2}, \ldots, S_{k+h_*+h^*-4})$.

We further decompose $d_1(k,j,\ell) := d'_1(k,j,\ell) + d''_1(k,j,\ell)$, where

$$\begin{aligned}
d'_1(k,j,\ell) &:= \int \mathbf{1}_{\{z_{h_*-1,h_k(\ell)-1} \le r\}} \mathbf{1}_{\{z_{1,2h_*+h^*-5} > ar\}}\\
&\qquad \times \mathbb{P}(S''_{j,h_j(\ell)-1} \le r, S'_{j,h_j(\ell)-1} > r) dF(z_1, \ldots, z_{2h_*+h^*-5}),\\
d''_1(k,j,\ell) &:= \int \mathbf{1}_{\{z_{h_*-1,h_k(\ell)-1} \le r\}} \mathbf{1}_{\{z_{1,2h_*+h^*-5} \le ar\}}\\
&\qquad \times \mathbb{P}(S''_{j,h_j(\ell)-1} \le r, S'_{j,h_j(\ell)-1} > r) dF(z_1, \ldots, z_{2h_*+h^*-5}),
\end{aligned}$$

with $a = a(n)$ to be chosen a little further. ■

We will simplify the notation by writing h_k instead of $h_k(\ell)$. We have:

$$
\begin{aligned}
d_1'(k,j,\ell) &\le \int \mathbf{1}_{\{z_{h_*-1,h_k-1}<r\}}\mathbf{1}_{\{z_{1,2h_*+h^*-5}>ar\}}dF(z_1,\ldots,z_{2h_*+h^*-5})\\
&= \int_0^r \frac{n(nu)^{h_k-2}}{(h_k-2)!}\int_{ar-u}^1 \frac{n(nv)^{2h_*+h^*-h_k-5}}{(2h_*+h^*-h_k-5)!}(1-u-v)^{n-(2h_*+h^*-4)}\\
&\quad\times(1+\mathcal{O}\left(\frac{1}{n}\right))dvdu\\
&\le \int_0^{nr}\frac{x^{h_k-2}}{(h_k-2)!}e^{-x/2}\int_{anr-x}^n \frac{y^{2h_*+h^*-h_k-5}}{(2h_*+h^*-h_k-5)!}e^{-y/2}\\
&\quad\times\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right)dydx \text{ (by a change of variable} + \text{Lemma 3)}\\
&\le 2^{2h_*+h^*-5}\int_0^{nr/2}\frac{z^{h_k-2}}{(h_k-2)!}e^{-z}\int_{anr/2-z}^\infty \frac{t^{2h_*+h^*-h_k-5}}{(2h_*+h^*-h_k-5)!}e^{-t}\\
&\quad\times\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right)dydx\\
&\le \frac{4(nr)^{h_k-1}}{(h_k-1)!}\frac{(anr)^{2h_*+h^*-h_k-5}}{(2h_*+h^*-h_k-5)!}e^{-anr/2}\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right)\text{(by Lemma 2)}\\
&\le \frac{4(nr)^{h_k+h_*-2}}{(h_*-1)!}\frac{1}{n}\left[\frac{n}{(nr)^{h_*-1}}\frac{(anr)^{2h_*+h^*-h_k-5}}{(2h_*+h^*-h_k-5)!}e^{-anr/2}\right]\\
&\quad\times\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right)\\
&\le \frac{1}{n}(nr)^{2(h_*-1)}\mathcal{O}(nr),
\end{aligned}
$$

if $\frac{1}{n}\le nr$ and $(3h_*+h^*-3)\log n\le anr\le\sqrt{n}$, entailing that $n(anr)^{2h_*+h^*-h_k-5}e^{-anr/2}\le(nr)^{h_*}$, and if moreover $a>1$, $nr<1$ and $anr>4(2h_*+h^*-4)$ for applying Lemma 2 and Lemma 3. The last inequality is hence valid for

$$4(2h_*+h^*-4)\vee(3h_*+h^*-3)\log n\le anr\le\sqrt{n}\text{ and }\frac{1}{n}\le nr<1$$

In a similar manner we can bound $d_1''(k,j,\ell)$, then decompose and bound $d_2(k,j,\ell)$.

We finally obtain that, if we take $a:=\frac{(3h_*+h^*-3)\log n}{nr}$ then, uniformly in $\frac{1}{n}\le nr<1$ and $n>4(2h_*+h^*-4)\vee\exp\left\{\frac{4(2h_*+h^*-4)}{3h_*+h^*-3}\right\}$, we have the upper bound stated in (*d*).

Proof of (e). For every $k=1,\ldots,n$ we let $\mathcal{C}_{ik}$ denote the class of all the subsets of size $i-1$ of $\Gamma_k^{vs}=\{k-h_*+2,\ldots,k-1,k+1,\ldots,k+h_*-2\}$. We obtain

$$i\lambda_i=\sum_{k=1}^n\sum_{C\in\mathcal{C}_{ik}}\mathbb{E}\left(I_k\prod_{t\in C}I_t\prod_{t\in\Gamma_k^{vs}\setminus C}(1-I_t)\right)\le\sum_{k=1}^n\sum_{C\in\mathcal{C}_{ik}}\mathbb{E}(I_{\inf C}I_{\sup C}).$$

For every $k=1,\ldots,n$ and $C\in\mathcal{C}_{ik}$ we have

$$\mathbb{E}(I_{\inf C}I_{\sup C})=\frac{1}{|\Lambda|}\sum_{\ell\in\Lambda}\mathbb{P}(S_{\inf C,h_{\inf C}(\ell)-1}\le r,S_{\sup C,h_{\sup C}(\ell)-1}\le r)$$

and $h_*-1\le i-1\le\sup C-\inf C$.

If $h_{\inf C}(\ell)=h_{\sup C}(\ell)=h_*$, then the two clusters do not intersect and we have

$$\mathbb{P}(S_{\inf C,h_{\inf C}(\ell)-1}\le r,S_{\sup C,h_{\sup C}(\ell)-1}\le r)=\frac{(nr)^{2(h_*-1)}}{[(h_*-1)!]^2}\left(1+\mathcal{O}\left(\frac{1}{n}\right)+\mathcal{O}(nr)\right).$$

It follows that $\mathbb{E}(I_{\inf C}I_{\sup C})\le\frac{(nr)^{2(h_*-1)}}{[(h_*-1)!]^2}\left(1+\mathcal{O}\left(\frac{1}{n}\right)+\mathcal{O}(nr)\right)$ and hence

$$i\lambda_i \le \binom{2(h_*-2)}{i-1} \frac{n(nr)^{2(h_*-1)}}{[(h_*-1)!]^2}\left(1+\mathcal{O}\left(\frac{1}{n}\right)+\mathcal{O}(nr)\right).$$

The bound stated in (*e*) then follows from

$$\sum_{i=h_*}^{2h_*-3}(i-1)\binom{2(h_*-2)}{i-1} = (2h_*-4)\sum_{j=h_*-2}^{2h_*-5}\binom{2h_*-5}{j} = (h_*-2)2^{2h_*-5}.$$

■

We have used the following two elementary lemmas that we state without proof. For a proof of Lemma 4, see Grusea (2008), and for a proof of Lemma 5, see Roos (1993a).

Lemma 4. *Let $X_1, \ldots, X_n$ be i.i.d. r.v.'s with distribution* Exp(1) *and let $i, k \ge 1$ s.t. $i+k-1 \le n$. Then, uniformly in $\alpha \ge 1$, $\beta < 1$, $\alpha\beta > 2(n-k-1)$, we have the following inequality:*

$$\mathbb{P}(X_{1,n} > \alpha\beta, X_{i,k} < \beta) \le 2\frac{\beta^k}{k!}\frac{(\alpha\beta)^{n-k-1}}{(n-k-1)!}e^{-\alpha\beta}.$$

Lemma 5. *For $0 \le x \le 1$ and $n \ge 2(m+1)$ we have $(1-x)^{n-(m+1)} \le e^{-nx/2}$.*

In the following lemma, we show that the chosen parameters for the approximating compound Poisson distribution verify the relation (3), and hence we can use the bound (4) of Barbour et al. (2000).

Lemma 6. *If $0 < nr < 1$ and $n_{\min} \asymp n$, then for every $i \in \{1, \ldots, h_*-1\}$ we have $\hat{\lambda}_i \asymp n(nr)^{i+h_*-2}$. If $n_{\min} \asymp n$ and $nr \le \gamma$, where γ is a fixed constant $\gamma < 1$, then $i\hat{\lambda}_i \ge (i+1)\hat{\lambda}_{i+1}, \forall i$.*

Proof. We have $i\lambda_i = \sum_{k=1}^n \mathbb{E}(I_k \mathbf{1}_{\{Z_k=i\}})$. One can easily show that the leading terms in the expression of $i\lambda_i$ are those which are expectations of products of i consecutive indicators. For a term with i consecutive indicators, of the form

$$\mathbb{E}(I_j \cdots I_{j+i-1}) = \frac{1}{|\Lambda|}\sum_{\ell\in\Lambda}\mathbb{E}(I_j \cdots I_{j+i-1}|\boldsymbol{L}=\ell),$$

we have that for each ℓ the extreme clusters intersect (because of the fact that $j+i-1 \le j+h_j(\ell)-2$, as $i < h_* \le h_j(\ell)$, $\forall j$) and hence

$$\begin{aligned}\mathbb{P}(S_j + \cdots + S_{j+i-1+h_{j+i-1}(\ell)-2} \le r) &\le \mathbb{E}(I_j \cdots I_{j+i-1}|\mathbf{L}=\ell)\\ &\le \mathbb{E}(I_j I_{j+i-1}|\mathbf{L}=\ell),\end{aligned}$$

implying that $\mathbb{E}(I_j \cdots I_{j+i-1}|\boldsymbol{L}=\ell \asymp (nr)^{i+h_{j+i-1}(\ell)-2}$..

Using (6) we obtain $\mathbb{E}(I_j \cdots I_{j+i-1}) \asymp (nr)^{i+h_*-2}, \forall j$.

The results in the statement easily follow. ■

For the detailed proofs of Proposition 1 and Lemma 2, see Grusea (2008).

From Proposition 1 and Lemma 4, together with Theorem 1 and relation (4), we obtain the following upper bound on the error of approximating $\mathbb{P}(W_m \ge 1)$ by $p = 1 - \exp\{-\sum_{i=1}^{h_*-1}\hat{\lambda}_i\}$.

Theorem 7. *Suppose that $n \longrightarrow \infty$, $r \longrightarrow 0$ and $n_{\min} \asymp n$. Then, uniformly in $\frac{1}{n} \le nr < 1$ and $n > 2(2h_*+h^*-4) \vee \exp\left\{\frac{4(2h_*+h^*-4)}{3(h_*-1)+h^*}\right\}$, we have:*

$$|\mathbb{P}(W_m \ge 1) - p| \le C\frac{n(nr)^{2(h_*-1)}}{[(h_*-1)!]^2}\left(1+\mathcal{O}\left(\frac{1}{n}\right)+\mathcal{O}(nr)\right),$$

where $C = 4h^* - h_* - 6 + (h_*-1)\{2h_*+h^*-5+2^{h_*-2}(h_*+h^*-4)\} + (h_*-2)2^{2h_*-6}$.

Moreover, if $\mathbb{E}(W_m) = \pi_\infty$ is held constant when $n \longrightarrow \infty$, then

$$|\mathbb{P}(W_m \geq 1) - p| = \mathcal{O}\left(\frac{1}{n}\right).$$

3.2. The computation of the parameters

In practice, based on the fact that the leading terms in the expression of $\hat{\lambda}_i$ are those containing products of i consecutive indicators, and using a "Markovian" approximation, we make the following approximation for the parameters:

$$\hat{\lambda}_i \approx n\pi q^{i-1}(1-q)^2, \text{ for } i = 2, \ldots, h_* - 1,$$
$$\hat{\lambda}_1 = n\pi - \sum_{i=2}^{h_*-1} i\hat{\lambda}_i,$$

where $\pi := \mathbb{P}(A_1)$ and $q := \mathbb{P}(A_2|A_1)$.

For computing π we sum over all labelings ℓ:

$$\pi = \frac{1}{|\Lambda|}\sum_{\ell\in\Lambda} \mathbb{P}(S_{1,h_1(\ell)-1} \leq r),$$

where $\mathbb{P}(S_{1,h_1(\ell)-1} \leq r)$ is given by a Beta distribution function (Glaz et al., 1983; Pyke, 1965).

Note that it suffices to sum only over all different $(\ell_1, \ldots, \ell_{h^*})$ possible.

We compute $q = \mathbb{P}(A_2|A_1) = \mathbb{P}(A_1 \cap A_2)/\pi$ in a similar way. We have

$$\mathbb{P}(A_1 \cap A_2) = \frac{1}{|\Lambda|}\sum_{\ell\in\Lambda} \mathbb{P}(A_1 \cap A_2|\boldsymbol{L} = \ell)$$
$$= \frac{1}{|\Lambda|}\sum_{\ell\in\Lambda} \mathbb{P}(S_{1,h_1(\ell)-1} \leq r, S_{2,h_2(\ell)-1} \leq r).$$

In this case it suffices to sum over all different $(\ell_1, \ldots, \ell_{h^*+1})$.

For calculating $\mathbb{P}(S_{1,h_1(\ell)-1} \leq r)$ and $\mathbb{P}(S_{1,h_1(\ell)-1} \leq r, S_{2,h_2(\ell)-1} \leq r)$ we use classic results on uniform spacings (Glaz et al., 1983).

3.3. The linear case

Next we briefly consider the case of a linear genome. As in the circular case, we see the genome B as the interval [0, 1] and the positions of the n orthologs as i.i.d. r.v.'s uniformly distributed on [0, 1]. The events A_k are defined as before, but in this case we have a smaller number of possible events, precisely $n - h_* + 1$. We also have a boundary effect which consists in the fact that for $k = n - h^* + 2, \ldots, n - h_* + 1$ the events A_k have a smaller probability.

Similarly to the circular case, we approximate the distribution of the number of clusters of type $(h : r)$ in the genome B, $W_m := \sum_{k=1}^{n-h_*+1} 1_{A_k}$, by a compound Poisson distribution of parameter $\hat{\lambda} = \sum_{i=1}^{h_*-1} \hat{\lambda}_i \delta_i$, where

$$\hat{\lambda}_i = \frac{1}{i}\sum_{k=1}^{n-h_*+1} \mathbb{E}(I_k \mathbf{1}_{\{Z_k = i\}}), \; i = 2, \ldots, h_* - 1, \; \hat{\lambda}_1 = \mathbb{E}(W_m) - \sum_{i=1}^{h_*-1} i\hat{\lambda}_i$$

and $Z_k = \sum_{j=k-h_*+2}^{k+h_*-2} \mathbf{1}_{A_j}$. Notice that Theorem 2 is valid in this case, too.

For the computation of the parameters we ignore the boundary effects and use a Markovian approximation as in the circular case. Note that, based on Assumption 1 (see also the relations (6) and (9)), the error introduced in the computation of the parameters by ignoring the boundary effects is negligible.

4. NUMERICAL RESULTS AND DISCUSSION

We denote by $\phi' := (\phi'_1, \ldots, \phi'_J)$ the vector containing all the distinct values $\phi'_1 < \cdots < \phi'_J$ among the sizes of the multigene families in the genome B, and we denote by $\boldsymbol{g} := (g_1, \ldots, g_J)$ the vector containing their multiplicities. We present two sets of numerical results for our compound Poisson approximation, see the two tables below. In Table 1 we give the results for the circular case and in Table 2 for the linear case.

TABLE 1. NUMERICAL RESULTS FOR THE CIRCULAR CASE

$(\phi', \boldsymbol{g}, h, r)$	$\hat{p}_{MC} \pm \varepsilon$	p
(1, 100, 8, 0.01)	0.0053 ± 0.000146	0.0053
(1, 100, 8, 0.012)	0.0150 ± 0.000245	0.0153
((1, 2), (100, 10), 8, 0.01)	0.0061 ± 0.000156	0.0060
((1, 2), (100, 10), 8, 0.012)	0.0174 ± 0.000264	0.0178
((1, 2, 3), (100, 15, 5), 8, 0.01)	0.0070 ± 0.000167	0.0071
((1, 2, 3), (100, 15, 5), 8, 0.02)	0.0208 ± 0.000288	0.0212
((1, 2, 3, 4), (100, 15, 5, 3), 8, 0.01)	0.0072 ± 0.000170	0.0073
((1, 2, 3, 4), (100, 15, 5, 3), 8, 0.012)	0.0219 ± 0.000296	0.0222
((1, 2, 3, 4, 5), (100, 15, 5, 3, 2), 8, 0.01)	0.0071 ± 0.000169	0.0075
((1, 2, 3, 4, 5), (100, 15, 5, 3, 2), 8, 0.012)	0.0219 ± 0.000296	0.0227

The vector ϕ' contains all the distinct sizes of the multigene families in the genome B and the vector $\boldsymbol{g}$ contains their multiplicities. p is our compound Poisson approximation for $\mathbb{P}(W_m(h, r) \geq 1)$ and $\hat{p}_{MC} \pm \varepsilon$ is a 95%–confidence interval for the same probability based on

We have selected values for ϕ', $\boldsymbol{g}$, h, r which are interesting in practice, for our biological purpose of statistically testing the significance of gene clusters found by the reference region approach. In both tables, p is our compound Poisson approximation for the probability of interest $\mathbb{P}(W_m(h, r) \geq 1)$ and $\hat{p}_{MC} \pm \varepsilon$ is a Monte Carlo estimate, based on 10^6 simulations, of the 95%-confidence interval for the same probability. We have estimated ε using the Central Limit Theorem.

Notice that, although Theorem 2 does not apply very well for these selected values and the theoretical bound given by the theorem is poor, the numerical results are very satisfactory.

We present here numerical results only for some selected values for ϕ', $\boldsymbol{g}$, h, r, but the approximation remains very good for a large panel of values for the parameters.

Note that in the case of a one-to-one orthology mapping ($\phi_i = 1$, $\forall i$) the weight of an interval is exactly the number of orthologs it contains. The probability $\mathbb{P}(W_m(h, r) \geq 1)$ can then be expressed in terms of the distribution of a continuous conditional scan statistic, for which a lot of approximations exist (Glaz, 2001) and also an exact (even if quite computationally demanding) expression (Huntington et al., 1975).

However, in the more general case that we treat in this article, trying to find an exact expression for this probability by using the method in Huntington et al. (1975) seems very difficult.

In biological applications, $\boldsymbol{h}$ and r will be the weight and respectively the (normalized) length of a given observed orthologous cluster in the genome B, for which we want to test its significance. The so-called *p-value* of this cluster is exactly the probability $\mathbb{P}(W_m(h, r) \geq 1)$, for which we give here a compound Poisson approximation. The observed cluster is significant, and hence interesting from the biological point of view, provided its p-value is smaller than a fixed threshold (0.01 for example).

TABLE 2. NUMERICAL RESULTS FOR THE LINEAR CASE

(ϕ', g, h, r)	$\hat{p}_{MC} \pm \varepsilon$	p
(1, 100, 8, 0.01)	0.0052 ± 0.000144	0.0052
(1, 100, 8, 0.012)	0.0146 ± 0.000242	0.0151
((1, 2), (100, 10), 8, 0.01)	0.0060 ± 0.000155	0.0060
((1, 2), (100, 10), 8, 0.012)	0.0173 ± 0.000263	0.0176
((1, 2, 3), (100, 15, 5), 8, 0.01)	0.0068 ± 0.000165	0.0070
((1, 2, 3), (100, 15, 5), 8, 0.02)	0.0203 ± 0.000285	0.0211
((1, 2, 3, 4), (100, 15, 5, 3), 8, 0.01)	0.0072 ± 0.000170	0.0073
((1, 2, 3, 4), (100, 15, 5, 3), 8, 0.012)	0.0216 ± 0.000294	0.0220
((1, 2, 3, 4, 5), (100, 15, 5, 3, 2), 8, 0.01)	0.0073 ± 0.000171	0.0074
((1, 2, 3, 4, 5), (100, 15, 5, 3, 2), 8, 0.012)	0.0217 ± 0.000295	0.0226

The vector ϕ' contains all the distinct sizes of the multigene families in the genome B and the vector $\boldsymbol{g}$ contains their multiplicities. p is our compound Poisson approximation for $\mathbb{P}(W_m(h, r) \geq 1)$ and $\hat{p}_{MC} \pm \varepsilon$ is a 95% confidence interval for the same probability based on

A java program for computing our compound Poisson approximation for the p-value of a given gene cluster, in the linear case, is available via the following web address: `www.math.univ-toulouse.fr/~grusea/Program/program.html`.

For a rigorous statistical test, the threshold must be chosen in order to bound the Type 1 error of the test. However, as we do not fix in advance the weight h when we search for conserved orthologous clusters, we deal here with a multiple testing problem and the choice of the threshold becomes quite complicated. We are currently trying to find solutions to this problem.

5. APPLICATIONS TO BIOLOGICAL DATA

This section is devoted to some applications of our results to real biological data. In the three examples given here, the detection of the orthologs and the identification of the conserved genomic regions were performed using the expert system CASSIOPE (Lopez et al., 2009).

5.1. Comparison of Homo-Sapiens and Oryzias-Latipes

In this example, we are interested in finding signs for the conservation of the Major Histocompatibility Complex (MHC) between the human genome and the genome of Oryzias-Latipes (or Japanese killifish, a very small ricefish, popular as an aquarium fish native to Southeast Asia).

The MHC contains genes involved in the immune defense. In the human genome, as the result of two rounds of polyploidization (whole genome duplication), we find four MHC paralogous regions (Abi-Rached et al., 2002).

We choose as reference region for our analysis the MHC paralogous region on the human chromosome 9 (129045207–140191570). The numbers in brackets represent the positions on the chromosome of the starting and, respectively, the ending nucleotide of the region. It has been shown that this region evolves slower than the other three.

This region contains 38 genes which have at least one ortholog in the genome of Oryzias-Latipes. Among those 38 genes, eight have two orthologs in Oryzias-Latipes and 30 have a single ortholog.

Therefore, using the notations from Section 2, the data are the following: $m = 38$, the number of genes in the reference region in the human genome (the species A) which have at least one ortholog in the genome of Oryzias-Latipes (the species B); $\phi' = (1, 2)$, the vector containing all the distinct values for the sizes of the multigene families in the Oryzias-Latipes genome; $\boldsymbol{g} = (30, 8)$, the vector containing the multiplicities of the different sizes in ϕ'; $n = 46$, the total number of genes in Oryzias-Latipes which are orthologous of genes in the human reference region; $N = 19686$, the size of the genome of Oryzias-Latipes (the total number of genes).

After locating the 46 orthologs in the genome of Oryzias-Latipes, nine conserved genomic regions were identified: three regions on the chromosome 9 and six others on the chromosome 12.

For each of these regions we determine its weight h and its normalized length r, and then we calculate its p-value using our compound Poisson approximation. The results are as follows.

The region #1, on the chromosome 9 (899561–1206257), contains three orthologs, of labels $3 \times \frac{1}{2}$; hence, its weight is $h = 1.5$. The total number of genes in the region is 9, thus the normalized length of the region is $r = \frac{9}{19686}$. We obtain a p-value of 0.00956, and hence this region is significant.

The region #2, on the chromosome 9 (28437906–29203467), contains four orthologs, of labels $2 \times 1, 2 \times \frac{1}{2}$; hence, $h = 3$. For this region $r = \frac{22}{19686}$ and its p-value is 0.0148. This region is thus significant at the level $\alpha = 0.05$.

The region #3, on the chromosome 9 (31902437–32170260), contains three orthologs, of labels $1, 2 \times \frac{1}{2}$; hence, $h = 2$. The length of this region is $r = \frac{3}{19686}$ and the p-value is 0.00104. The region is very significant.

The region #4, on the chromosome 12 (993203–5399518), contains four orthologs, of labels $3 \times 1, \frac{1}{2}$; hence, $h = 3.5$. For this region $r = \frac{7}{19686}$ and its p-value is 1.63×10^{-5}. Therefore, this region is highly significant.

The region #5, on the chromosome 12 (6945906–8246163), contains six orthologs, of labels $3 \times 1, 3 \times \frac{1}{2}$, and thus $h = 4.5$. The length of the region is $r = \frac{52}{19686}$ and its p-value is 1.27×10^{-4}. This region is hence very significant.

The region #6, on the chromosome 12 (10049683–10113348), contains three orthologs, of labels 3×1; therefore, $h = 3$. The length of the region is $r = \frac{3}{19686}$ and its p-value is 2.82×10^{-4}. This region is hence very significant.

The region #7, on the chromosome 12 (11625364–11880175), contains four orthologs, of labels 4×1; hence, $h=4$. For this region $r=\frac{16}{19686}$ and the p-value is 5.83×10^{-5}. The region is hence highly significant.

The region #8, on the chromosome 12 (15301431–15697269), contains four orthologs, of labels 3×1, 3×1, $\frac{1}{2}$; hence, $h=3.5$. The region has length $r=\frac{18}{19686}$ and p-value 2.71×10^{-4}. This region is thus very significant.

The region #9, on the chromosome 12 (25421295–26996650), contains six orthologs, of labels 2×1, $2\times 1, 4\times\frac{1}{2}$; hence, $h=4$. The length of the region is $r=\frac{30}{19686}$. We obtain the p-value 3.81×10^{-4}; hence, this region is also very significant.

The results indicate a high conservation between the human MHC region on the chromosome 9 and the nine regions on the chromosomes 9 and 12 of Oryzias-Latipes.

5.2. Comparison of Ciona-Intestinalis and Homo-Sapiens

In this second analysis, we present a comparison between the human genome and the genome of Ciona-Intestinalis, which is a Urochordata (sea squirt) whose genome has been sequenced and which has become, over the past decade, a major experimental model for developmental biologists. For more details about the comparison presented here, see Zucchetti et al. (2009).

We start with a reference genomic region in Ciona, spread over chromosomes 4 and 10 and containing genes of the immunoglobulin superfamily. The concatenated reference region contains 14 genes having at least one ortholog in the human genome.

The data are the following: $m=14$, the number of genes in the reference region in Ciona which have at least one ortholog in the human genome; $\phi'=[1, 2, 3, 4, 7, 8, 16]$, the vector containing all the distinct values for the sizes of the multigene families in the human genome; $\boldsymbol{g}=[5, 3, 2, 1, 1, 1, 1]$, the vector containing the multiplicities of the different sizes in ϕ'; $n=52$, the total number of genes in the human genome which are orthologous of genes in the reference region in Ciona; $N=36396$, the size of the human genome.

After locating the 52 orthologs in the human genome, we found two conserved genomic regions, on the chromosomes 11 and 19.

After computing, using our compound Poisson approximations, the p-values of the different gene clusters, we obtain the following results.

The region #1, on the human chromosome 11 (60495750–133526846), contains 14 orthologs, of labels 2×1, $3\times\frac{1}{2}$, $\frac{1}{3}$, $2\times\frac{1}{7}$, $6\times\frac{1}{16}$, and so $h=3.494$. The normalized length is $r=\frac{140}{36396}$. The normalized length of this region is $r=\frac{998}{36396}$. We obtain a p-value of 0.0083; thus, this region is very significant (at the level $\alpha=0.01$).

The region #2, on the human chromosome 19 (40511919–60093650), contains 13 orthologs, of labels 3×1, $\frac{1}{2}$, $2\times\frac{1}{3}$, $\frac{1}{4}$, $2\times\frac{1}{7}$, $4\times\frac{1}{8}$; hence, $h=5.2024$. For this region $r=\frac{803}{36396}$ and we obtain a p-value of 1.7612×10^{-6}. This region is highly significant.

We identified, in the human genome, two very significant conserved regions orthologous to the reference region in Ciona. The first one is on the human chromosome 11 and the second one on the human chromosome 19. This shows the conservation of the immunoglobulin superfamily in human and in Ciona since their divergence from their common ancestor, more than 800 million years ago.

5.3. Comparison of Homo-Sapiens and Danio-Rerio

In this third example, we present another comparison involving the MHC human region on the chromosome 9. We are interested in finding orthologous regions for this human reference region in the genome of Danio-Rerio.

Danio-Rerio, commonly known as zebrafish, is a tropical freshwater fish very popular as an aquarium fish. It is also an important vertebrate model organism for biologists.

In this comparison, the reference region is the Homo-Sapiens MHC region on chromosome 9 (ENSG00000136895–ENSG00000159247), the numbers in the brackets being the gene identifiers, in the Ensemble database, respectively, for the starting and the ending gene in the region. When searching for orthologs for the genes from the reference region in the genome of Danio-Rerio, we find 20 genes in the human reference region which have at least one ortholog in the genome of Danio. Among these 20 genes, 16 have a single ortholog, two have two orthologs, one has three orthologs, and one has eight orthologs.

We thus have $m = 20$, $\phi' = (1, 2, 3, 8)$, $\boldsymbol{g} = (16, 2, 1, 1)$, and $n = 31$. The size of the genome of Danio-Rerio is $N = 28509$ genes.

After locating these 31 orthologs in the genome of Danio-Rerio, we identify a conserved genomic region on chromosome 5 (ENSDARG00000030173–ENSDARG00000068122) containing seven orthologs, of weights $5 \times 1, \frac{1}{2}, \frac{1}{8}$; hence, the weight of the cluster is $h = 5.625$. The region contains 38 genes in total; hence, its normalized length is $r = \frac{38}{28509}$. We obtain a p-value of 2×10^{-10}, and hence the region is highly significant.

With this example, we find another strong evidence for the conservation of the MHC region between human and fish.

ACKNOWLEDGMENTS

S. Grusea wishes to thank Malgorzata Roos for sending a copy of her Ph.D. thesis. This work was partially supported by the ANR MAEV (contract ANR-06-BLAN-0113). The authors also thank the two reviewers for their useful suggestions which improved the presentation of this article.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Abi-Rached, L., Gilles, A., Shiina, T., et al. 2002. Evidence of en bloc duplication in vertebrate genomes. *Nat. Genet.* 31, 100–105.

Barbour, A., Chen, L., and Loh, W. 1992. Compound Poisson approximation for nonnegative random variables via Stein's method. *Ann. Probabil.* 20, 1843–1866.

Barbour, A., and Xia, A. 2000. Estimating Stein's constants for compound Poisson approximation. *Bernoulli* 6, 581–590.

Bergeron, A., Corteel, S., and Raffinot, M. 2002. The algorithmic of gene teams. *Lect. Notes Comput. Sci.* 2452, 464–476.

Chen, L.H.Y. 1975. Poisson approximation for dependent trials. *Ann. Probabil.* 3, 534–545.

Danchin, E., and Pontarotti, P. 2004. Statistical evidence for a more than 800-million-year-old evolutionarily conserved genomic region in our genome. *J. Mol. Evol.* 59, 587–597.

Durand, D., and Sankoff, D. 2003. Tests for gene clustering. *J. Comput. Biol.* 10, 453–482.

Glaz, J., and Naus, J. 1983. Multiple clusters on the line. *Commun. Stat. Theor.* M. 12, 1961–1986.

Glaz, J., Naus, J., and Wallenstein, S. 2001. *Scan Statistics.* Springer Verlag, New York.

Grusea, S. 2008. Applications of probability calculus to the detection of conserved genomic regions [Ph.D. dissertation]. Univ. de Provence, Marseille.

Hoberman, R., and Durand, D. 2005. The incompatible desiderata of gene cluster properties. *Lect. Notes Bioinform.* 3678, 73–87.

Hoberman, R., Sankoff, D., and Durand, D. 2005. The statistical significance of max-gap clusters. *Lect. Notes Bioinform.* 3388, 55–71.

Huntington, R., and Naus, J. 1975. A simpler expression for *k*th nearest neighbor coincidence probabilities. *Ann. Probabil.* 3, 894–896.

Lopez Rascol, V., Levasseur, A., Chabrol, O., et al. 2009. CASSIOPE: an expert system for conserved regions searches. *BMC Bioinform.* 10:284.

Pyke, R. 1965. Spacings. *J. R. Stat. Soc. B* 27, 395–449.

Pyke, R. 1972. Spacings revisited. *Proc. Sixth Berkeley Symp. Math. Stat. Probabil.* 1, 417–427.

Raghupathy, N., and Durand, D. 2005. Individual gene cluster statistics in noisy maps. *Lect. Notes Bioinform.* 3678, 106–120.

Raghupathy, N., and Durand, D. 2009. Gene cluster statistics with gene families. *Mol. Biol. Evol.* 26, 957–968.

Roos, M. 1993a. Stein-Chen method for compound Poisson approximation [Ph.D. dissertation]. Univ. of Zurich.

Roos, M. 1993b. Compound Poisson approximations for the number of extreme spacings. *Adv. Appl. Probabil.* 25, 847–874.

Stein, C. 1972. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. *Proc. Sixth Berkeley Symp. Math. Stat. Probabil.*

Stein, C. 1986. Approximate computation of expectations. In: *IMS Lect. Notes. Monograph Series*, *Vol. 7.* IMS, Hayward, CA.

Zucchetti, I., De Santis, R., Grusea, S., et al. 2009. Origin and evolution of the vertebrate leukocyte receptors: the lesson from tunicates. *Immunogenetics* 61, 463–481.

Address correspondence to:
Dr. S. Grusea
INSA de Toulouse
Département GMM
135 avenue de Rangueil
31077 Toulouse, *France*

E-mail: grusea@insa-toulouse.fr