Modeling Spatial Genomic Interactions with the Hawkes model

Anna Bonnet
Univ Lyon, Université Lyon 1, CNRS, LBBE UMR 5558
F-69622 Villeurbanne, France
anna.bonnet@univ-lyon1.fr

Vincent Rivoirard
CEREMADE UMR CNRS 7534
Université Paris Dauphine, Paris, France

Franck Picard
Univ Lyon, Université Lyon 1, CNRS, LBBE UMR 5558
F-69622 Villeurbanne, France

November 7, 2017

Abstract

The spatial localization of many DNA-protein interactions is now available thanks to the development of ChIP-Seq, and their investigation calls for adapted statistical methods. Many methods were developped for peak calling, but few were proposed for the downstream analysis of peak-like data, whereas the spatial structure of such data may contain relevant biological information, like binding constraints for instance. Associations between the occurrences of two genomic features are usually assessed by overlaps, but here we propose a statistical model to precisely quantify the spatial interactions between the location of binding events. Our methodology relies on a multivariate spatial process, the Hawkes model, that can also be interpreted in terms of a graphical model to highlight spatial dependencies between genomic features. Using our method, we explore the chromatinian landscape of replication origins, and we highlight attractive and repulsive patterns that can be related to the regulation of the spatial program of replication.

1 Introduction

Next generation sequencing technologies allow to study genomes with an unprecedented resolution; molecular processes can be captured through the mapping of the whole genetic variants of a genome, the measure of the expression of all genes of a cell, and now the structure of chromatin and its modifications. Epigenetics has certainly been one of the most active fields for the last few years, thanks to the development of ChIP-Seq that allow the spatial localization of DNA-proteins interactions. One of the major challenges now is to extract relevant information from this amount of data, to better understand genomes regulation.
ChIP-Seq data provide maps of chromatin modifications along chromosomes, and we focus on the statistical modeling of this spatial specificity of the data. A first category of methods

is based on Hidden Markov Models (HMMs) and aims at detecting regions in the ChIP-Seq signal. The ChromHMM method (Ernst and Kellis, 2012) detects from the raw signal the presence or absence of each chromatin mark and segments the genome in regions characterized by specific combinations of marks. This segmentation defines a "chromatinian landscape", which describes the modifications of chromatin states and their associated biological functions. Wu and Zhaohui (2013) developed the GiClust multivariate procedure, also based on HMMs, that uses ChIP-seq data after peak detection. GiClust provides a segmentation in clusters on which the probability of occurrence for each feature is estimated.

Several methods were also developed to study the co-occurrences of a group of features by testing their pairwise associations. These methods are very general and provide co-occurrences patterns for any spatialized data described by points or intervals, in particular ChIP-Seq data after any peak detection method. Favorov et al. (2012) gathered and compared four standard methodologies (relative distance test, absolute distance test, projection test and Jaccard test), all implemented in the GenometriCorr package. D. Chikina and G. Troyanskaya (2012) noticed that most procedures used to compare sets of intervals considered the binary situation -with or without overlaps- and defined a metrics to assess the distance between two intervals lists. Wei and Wu (2016) proposed to use the ChromHMM segmentation to account for genome heterogeneities before looking for associations between genomic features.
In this work we focus on the modeling of spatial interactions between called ChIP-Seq peaks, and more generally between maps of genomic intervals structured along chromosomes. We propose a modeling based on point patterns to account for potential physical interactions that come from the spatial proximity of binding events. To propose a fine-scale modeling of spatial interactions, we focus on the spatial covariance between points, embedded within a multivariate model. This multivariate aspect of our model is central compared to existing pairwise strategies, since it allows to correct for spurious aliasing correlations. Then the results of our method can be interpreted in terms of spatial attraction/repulsion between genomic features, which could help in the identification of characteristic distances for instance. Our strategy is model based, and relies on the modeling of spatial interactions with the Hawkes model, for which non parametric estimation procedures were recently proposed by Reynaud-Bouret et al. (2014). The benefits of this method compared to standard approaches are plural, in particular it allows the reconstruction of sparse interaction functions, which is very convenient for interpretation. Moreover, based on these functions, one can reconstruct a graphical model with an edge for each significant interaction between pairs of features. Let us also note that defining a model could also be useful to generate realistic spatially interacting genomic features.
This method is used to explore the spatial interaction landscape of replication origins with their chromatin context. DNA replication is a biological process that is intrinsically spatial and the identification of replication origins at fine scales has revealed complex interplays with genetic and epigenetic features in the early steps of DNA replication (Picard et al., 2014; Julienne et al., 2015). Using our method, we reveal attractive effects between marks and origins, with estimated intensities and distances of interaction. We also show strong attractions and repulsions between marks, which confirms the necessity of including all marks in a multivariate model.

2 Model and method

2.1 Multivariate Hawkes process

We consider a dataset with M maps of ChIP-Seq peaks, or more generally of mapped genomic features. We denote by X_m the m-th map, such that $X_m = \{T_1^m, \dots, T_{n_m}^m\}$, with T_i^m the genomic coordinate of the peak i in map m.
In a first approximation, we reduce the ChIP-Seq intervals to single points. Then, we consider that these positions are random to account for biological and technical noise and we model $(X_1, \dots, X_M)$ as a multivariate point process.
To model the spatial interactions between genomic features, we introduce the conditional intensity of the process, which models the probability of occurrence of points, such that: for $1 \leqslant m \leqslant M$, for $t \in [T_1, T_2]$ a given genomic region,

$$\lambda^{(m)}(t) = \nu^{(m)} + \sum_{l=1}^{M} \sum_{T \in X_l} h_l^{(m)}(t - T) \tag{1}$$

- $h_l^{(m)}$ is a function defined on $(0, +\infty)$ that characterizes the influence of each occurrence $T \in X_l$ on $T' \in X_m$. For instance, if $h_l^{(m)}$ is a positive function on $(0, d)$ and is null elsewhere, the probability of observing points of X_m will increase within a distance of d after observing points of X_l.

- $\nu^{(m)}$ is a spontaneous rate, it corresponds to the part of the intensity $\lambda^{(m)}(t)$ that cannot be explained by the occurrences of $(X_1, \dots, X_M)$.

Remark 1. The function $h_m^{(m)}$ describes the self-interactions between the occurrences of the mth process. In particular, it can model regions with clusters of occurrences of the same process.

Remark 2. In Model (1), the interaction functions $h_l^{(m)}$ and $h_m^{(l)}$ can be (and generally are) different, so that we can differenciate the effect of X_m on X_l from the one of X_l on X_m. Model (1) can be interpreted as a directed graphical model with M nodes and edges for each non-zero interaction function.

Our goal is to propose an estimator of $\left(\nu^{(m)}, h_l^{(m)}\right)_{l,m}$ in order to understand the spatial interactions between all the point processes.

2.2 Method

Reynaud-Bouret et al. (2014) proposed a methodology to estimate the interaction functions, a brief description of which is given in this section. A more detailed version is given in the Appendix, section 5. The procedure is nonparametric, which means that there is no

assumption on the functions $\left(h_l^{(m)}\right)_{l,m}$ that we want to estimate.

The main idea is to find a candidate $g_l^{(m)}$ to estimate $h_l^{(m)}$ that can be decomposed on a histogram basis as follows:

$$g_l^{(m)} = \sum_{k=1}^{K} a_{m,l,k} \delta^{-1/2} \mathbb{1}_{((k-1)\delta, k\delta]},$$

where δ is the size of each bin and K the number of bins. The product $K\delta$ corresponds to the maximal distance between two occurrences that interact with each other. The coefficients $a_{m,l,k}$ can be interpreted is terms of spatial covariance between points of X_l and points of X_m at lag included in $((k-1)\delta, k\delta]$, corrected by other potential aliasing covariates.
The coefficients $(a_{m,l,k})_{l,m,k}$ are estimated thanks to a penalized least squares criterion with theoretically calibrated weights, which provides a sparse estimation of the interaction functions $\left(h_l^{(m)}\right)_{l,m}$.

3 Application

3.1 Data

The spatio temporal program of replication is tighly regulated in cells, as it allows the faithful duplication of the genetic material. Thanks to the recent identification of replication origins (Picard et al., 2014), much progress has been made in the identification of the genetic and epigenetic basis of these controls (Picard et al., 2014; Julienne et al., 2015). Current studies proposed to unravel the interactions between replication origins and chromatin marks using standard associations, by studying the overlap between genomic maps. However, the precise spatial interactions between origins and surrounding marks have never been investigated for lack of appropriate method. The data we study here are the positions of replication origins as well as the positions of five chromatin marks: H3k9ac, H3k4me3, H3k27me3, H3k9me3 and H4k20me1, for the Hela cell line. We deal separately with regions replicated early and late in the cell cycle, as regulation mechanisms were suggested to be different (Picard et al., 2014).

3.2 Results

The output of this model is the estimates of $\left(h_l^{(m)}\right)_{1\leqslant l,m\leqslant M}$. Notice that in Model (1), the interaction functions have support in $(0, +\infty)$, which means that an occurrence can affect another only if it is located before in terms of genomic coordinates. In the following results, the portion of the interaction functions with negative support are obtained by considering the genomic coordinates in decreasing order.

3.2.1 Influence of the marks on origins

Figure 1 displays the estimations of h_{mark}^{ori}, that is the influence of each mark on the presence of origins, for each of the five chromatin marks.
The interpretation for the value of $h_l^{(m)}$ on the interval $[(k-1)\delta, k\delta]$ for $k = [1, \ldots, K]$ is the mean number of additional points of X_m due to the presence of each point of X_l. For

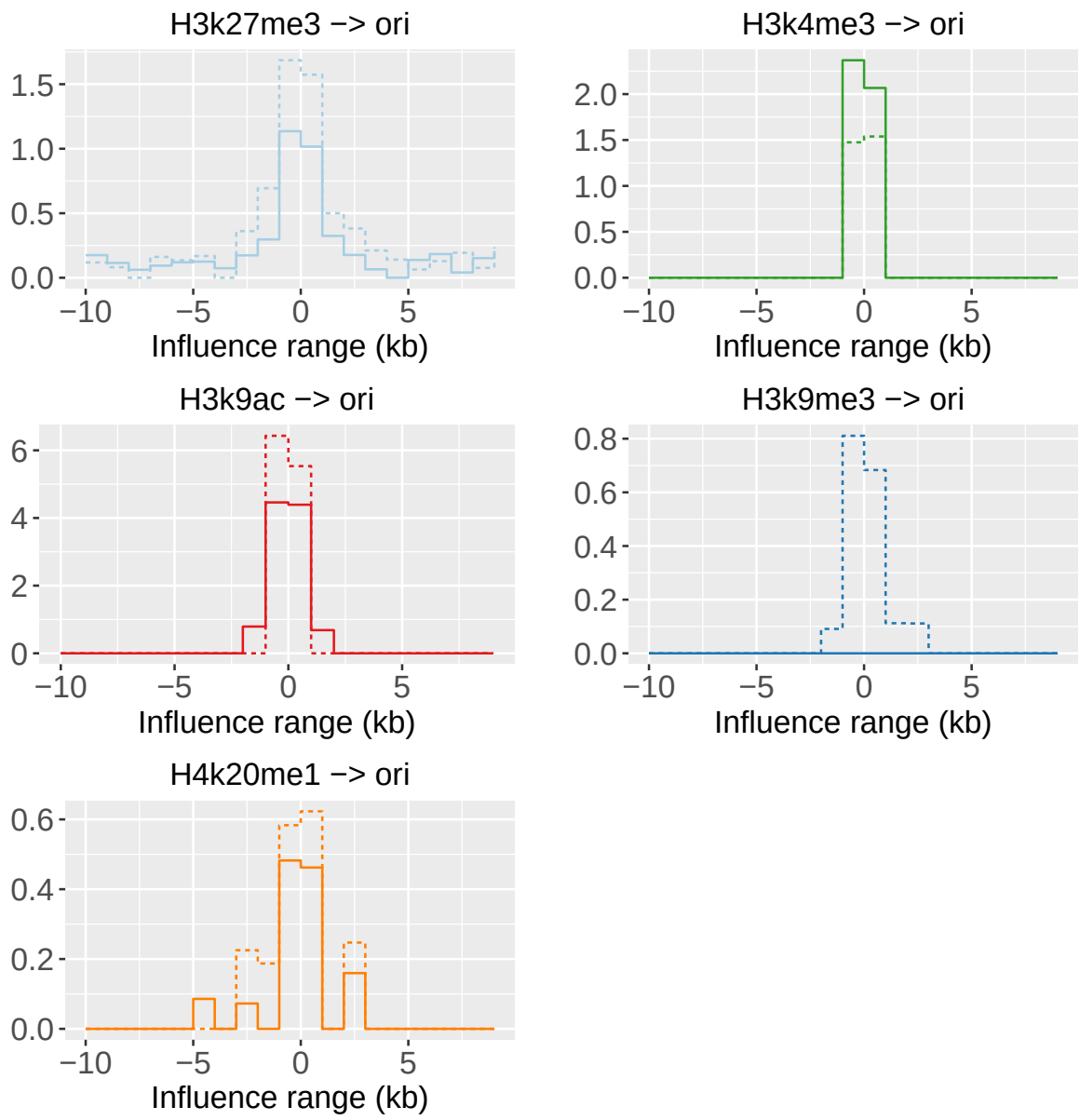

Figure 1 – Estimated interaction functions h^{ori}_{mark} for all chromatin marks, with $\delta = 1$kb and $K = 10$. The linetype indicates the timing regions: early (full line) and late (dotted line).

instance, if we look at the influence of H3k27me3 on the origins (Figure 1, first column, first row), $h^{ori}_{H3k27me3} \simeq 1$ on $[0, 1]$ for early regions: it means that each occurrence of H3k27me3 in early regions adds on average one origin within a distance of 1 kb compared to the number of origins we would observe without the presence of this mark.
We observe that the presence of most of the marks increases the number of origins in both early and late regions, within a small distance ($<$ 3kb, except for H3k27me3, whose effect is more diffuse). In particular, the influence of H3k9ac is very strong within a distance smaller than 1kb.
Let us also note that the difference between late and early regions is obvious for H3k9me3,

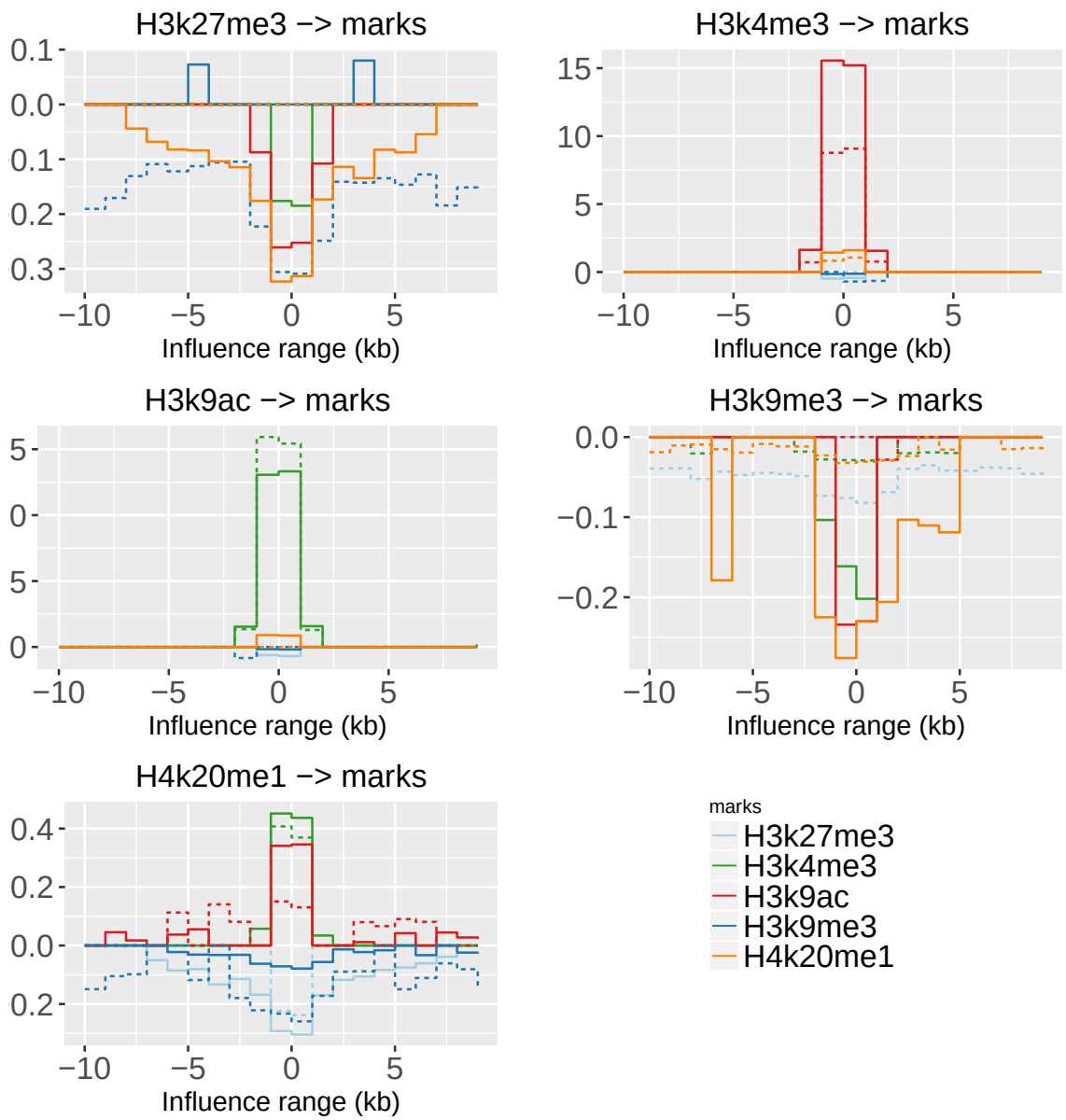

Figure 2 – Estimated interaction functions between marks for all chromatin marks, with $\delta = 1$kb and $K = 10$. The linetype indicates the timing regions: early (full line) and late (dotted line).

where the interaction exists only in late regions, which is consistent with the correlation study of Picard et al. (2014). However, the effect of the timing is less noticeable for the other marks, which might be explained by our division of regions between two groups (early and late) while Picard et al. (2014) compared the associations between origins and marks divided in 6 timing groups.

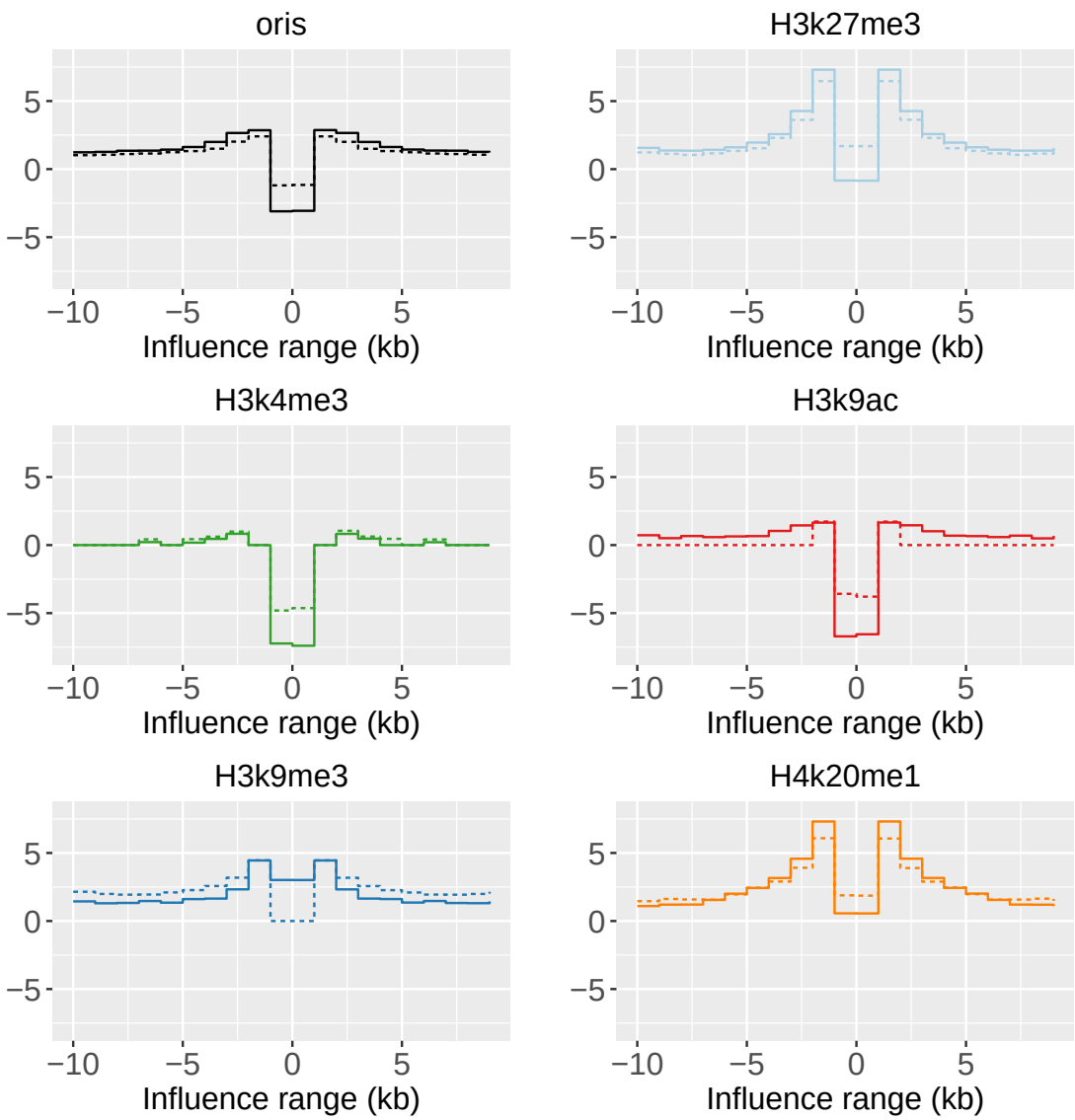

Figure 3 – Estimated self interaction functions between marks or origins, with $\delta = 1$kb and $K = 10$. The linetype indicates the timing regions: early (full line) and late (dotted line).

3.2.2 Interaction between marks

Figure 2 shows the interactions between chromatin marks. We observe a very strong attraction between H3k4me3 and H3k9ac which are both open-chromatin marks. These marks have also attractive interactions, though weaker, with H4k20me1. These three marks seem to have a repulsive interaction with H3k27me3, although they all have an attractive effect origins. Finally, H3k9me3 is also repulsed by the presence of the other chromatin marks.
As mentioned in Section 1, the multivariate approach that we use allows us to correct the pairwise interactions by taking into account the interactions with other features. Here, we see several interactions between marks which are likely to impact the associations mark/origins that we might detect without this correction.

3.2.3 Self interactions

Figure 3 shows the estimated $h_m^{(m)}$ functions of Model (1), that is the self interactions between occurrences of origins, and between the occurrences of each mark. We see that for origins and most marks, the interactions between the points of the same process are strong, which suggests that there are regions with clusters of points. We notice that the self interactions are weaker for the two open-chromatin marks H3k4me3 and H3k9ac. Let us also observe that for origins and several marks, the estimated $h_m^{(m)}$ take negative values on [-1,1kb] which is due to the modeling where we consider our observations as points when they are actually centers of intervals; the distance between two centers is indeed obviously at least larger than the length of the intervals.

3.2.4 Interaction graph

The estimated interactions are summarized in a graph (Figure 4), the edges of which have a width proportional to the integrated interaction function between two nodes. It confirms that there is groups of marks composed of H3k4me3, H3k9ac and H4k20me1 that have attractive effects on each other and on origins, and repulsive effects on the other marks. H3k27me3 and H3k9me3 attract only origins but repulse the other marks (with only a small attraction detected between H3k27me3 and H3k9me3 in early regions). Although the differences between the early and late graphs are quite small, we notice that the edge between H3k9me3 and origins appears only in the late graph.

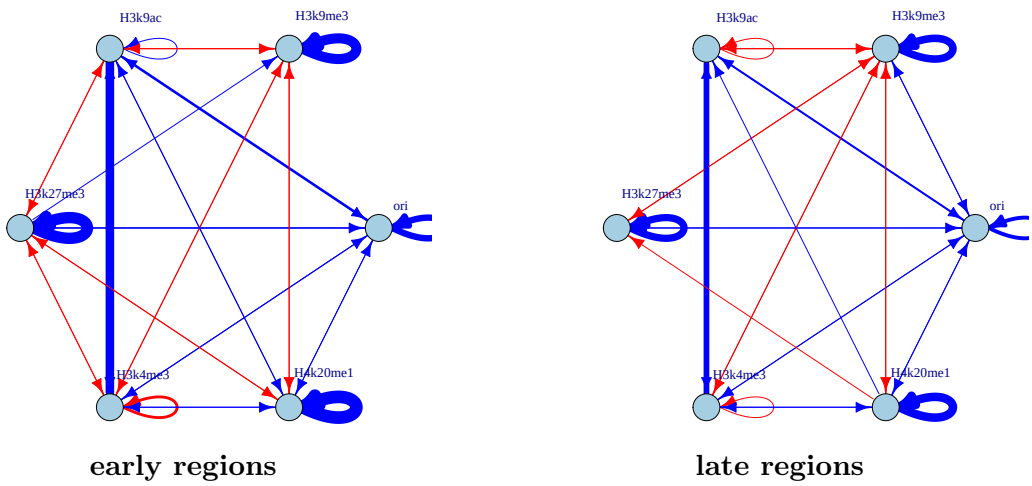

Figure 4 – Graph of interactions between marks or origins. The repulsive effects are in red, the attractive effects in blue. The width of arrows is proportional to the integrated interaction function.

4 Discussion

We propose in this paper a model to quantify spatial interactions between replication origins and chromatin marks. The results that we obtain describe these interactions with much

more precision than classical correlation or overlaps approaches. We highlight indeed several attraction/repulsion patterns with typical distances of interaction and estimated intensities. The interpretation of the Hawkes model as a graphical model allows us to represent our result in a very synthetical graph. Let us also note that the methodology we propose can be generalized to any set of spatially ordered genomic features.
The next step to improve this method is to take into account the variability in the detection of the positions of the origins and chromatin marks. The extension of the procedure is an ongoing work.

References

D. Chikina, M. and O. G. Troyanskaya (2012). An effective statistical evaluation of chipseq dataset similarity. Bioinformatics 28(5), 607–613.

Ernst, J. and M. Kellis (2012, March). Chromhmm: automating chromatin-state discovery and characterization. Nat Meth 9(3), 215–216.

Favorov, A. V., L. Mularoni, L. M. Cope, Y. A. Medvedeva, A. A. Mironov, V. J. Makeev, and S. J. Wheelan (2012). Exploring massive, genome scale datasets with the genometricorr package. PLoS Computational Biology 8(5).

Julienne, H., B. Audit, and A. Arneodo (2015). Embryonic stem cell specific "master" replication origins at the heart of the loss of pluripotency. PLoS Comput Biol 11(2), e1003969.

Picard, F., J.-C. Cadoret, B. Audit, A. Arneodo, A. Alberti, C. Battail, L. Duret, and M.-N. Prioleau (2014). The spatiotemporal program of dna replication is associated with specific combinations of chromatin marks in human cells. PLoS Genet 10(5), e1004282.

Reynaud-Bouret, P., V. Rivoirard, F. Grammont, and C. Tuleau-Malot (2014, April). Goodness-of-fit tests and nonparametric adaptive estimation for spike train analysis. Journal of Mathematical Neuroscience, 4:3.

Wei, Y. and H. Wu (2016). Measuring the spatial correlations of protein binding sites. Bioinformatics 32(12), 1766–1772.

Wu, H. and S. Q. Zhaohui (2013). Exploring the cooccurrence patterns of multiple sets of genomic intervals. BioMed Research International.

5 Appendix

We give in this section more details on the estimation method proposed by Reynaud-Bouret et al. (2014).
Let us denote $f = \left(\mu^{(m)}, g_l^{(m)}\right)_{l,m}$ a candidate to estimate $\left(\nu^{(m)}, h_l^{(m)}\right)_{l,m}$ and the corresponding estimator of $\lambda^{(m)}$:

$$\psi_t(f) = \mu^{(m)} + \sum_{l=1}^{M} \sum_{T \in X_l} g_l^{(m)}(t-T).$$

We want the intensity candidates to minimize the ℓ_2-distance to the intensites $\lambda^{(1)}, \ldots, \lambda^{(M)}$:

$$\|\psi(f) - \lambda\|^2 = \sum_{m=1}^{M} \int_{T_1}^{T_2} \left[\psi_t^{(m)}(f) - \lambda^{(m)}\right]^2 dt,$$

which is equivalent to minimizing the contrast $\tilde{\gamma}(f)$ defined by

$$\begin{aligned}\tilde{\gamma}(f) &= -2 < \lambda, \psi(f) > + \|\psi(f)\|^2 \\ &= -2 \sum_{m=1}^{M} \int_{T_1}^{T_2} \psi_t^{(m)}(f) \lambda^{(m)} dt + \sum_{m=1}^{M} \int_{T_1}^{T_2} \psi_t^{(m)}(f)^2 dt.\end{aligned}$$

Since a natural approximation of $\lambda^{(m)}(t)$ is the point measure $dN^{(m)}(t)$, we shall focus on minimizing $\gamma(f)$ defined by

$$\gamma(f) = -2 \sum_{m=1}^{M} \int_{T_1}^{T_2} \psi_t^{(m)}(f) dN^{(m)}(t) + \sum_{m=1}^{M} \int_{T_1}^{T_2} \psi_t^{(m)}(f)^2 dt. \tag{2}$$

The main idea is to find a candidate $g_l^{(m)}$ to estimate $h_l^{(m)}$ that can be decomposed on a histogram basis as follows:

$$g_l^{(m)} = \sum_{k=1}^{K} a_{m,l,k} \delta^{-1/2} \mathbb{1}_{((k-1)\delta, k\delta]},$$

where δ is the size of each bin and K the number of bins. The product $K\delta$ corresponds to the maximal distance between two occurrences that interact with each other. The coefficients $a_{m,l,k}$ can be interpreted is terms of spatial covariance between points of X_l and points of X_m at lag included in $((k-1)\delta, k\delta]$. Then, if we use this decomposition in the definition of $\psi_t(f)$,

$$\psi_t^{(m)}(f) = \mu^{(m)} + \sum_{l=1}^{M} \sum_{k=1}^{K} a_{m,l,k} \delta^{-1/2} N^{(l)}\left([t-k\delta, t-(k-1)\delta]\right),$$

where $N^{(l)}(I)$ denotes the number of points of X_l on interval I. If we replace $\psi_t^{(m)}(f)$ in (2), Reynaud-Bouret et al. (2014) show that the contrast can be rewritten as:

$$\gamma(f) = \sum_{m=1}^{M} -2(a^{(m)})' b^{(m)} + (a^{(m)})' G a^{(m)}, \tag{3}$$

where $\left(a^{(m)}\right) = \left(\mu^{(m)}, a_{m,1,1}, \ldots, a_{m,1,K}, a_{m,2,1}, \ldots, a_{m,M,K}\right)$ are the coefficients to estimate, and $b^{(m)}$ and G are defined by:

$$b^{(m)} = \left(N^{(m)}([T_1, T_2]), \delta^{-1/2}\mathbf{n}'_{m,1}, \ldots, \delta^{-1/2}\mathbf{n}'_{m,M}\right)$$

where

$$\mathbf{n}'_{m,l} = \left(\int_{T_1}^{T_2} N^{(l)}([t - k\delta, t - (k-1)\delta))dN^{(m)}(t)\right)_{k=1,\ldots,K}$$

and

$$G = \int_{T_1}^{T_2} (\mathbf{R}c_t)'\mathbf{R}c_t dt,$$

where

$$\begin{aligned}(\mathbf{R}c_t)' = (1,\ & \delta^{-1/2}N^{(1)}([t-\delta, t)), \ldots, \delta^{-1/2}N^{(1)}([t - K\delta, t - (K-1)\delta)), \ldots, \\ & \delta^{-1/2}N^{(M)}([t-\delta, t)), \ldots, \delta^{-1/2}N^{(M)}([t - K\delta, t - (K-1)\delta))\Big).\end{aligned}$$

The solution of the minimization of (3) can be easily obtained; in particular, if G is invertible

$$a^{(m)} = G^{-1}b^{(m)}.$$

Notice that the entries of $b^{(m)}$ correspond to pairwise correlograms, while the coefficients of $a^{(m)}$ are corrected by other potential aliasing covariates, the effects of which are quantified by the matrix G.
The procedure is achieved by thresholding the coefficients of $a^{(m)}$, as follows:

$$\hat{a}^{(m)} = \underset{a^{(m)}}{\operatorname{argmin}} \left\{ \sum_{m=1}^{M} -2(a^{(m)})'b^{(m)} + (a^{(m)})'Ga^{(m)} + (d^{(m)})'|a^{(m)}| \right\},$$

where the weights $(d^{(m)})_{1 \leqslant m \leqslant M}$ are theoretically calibrated. Once the support is determined by the previous procedure, the final estimation of the non-zero coefficients of $a^{(m)}$ is obtained thanks to a classical least-squares estimate on this support. This two-step approach allows to overcome the bias issue of the LASSO estimates.