Spatio-temporal model for multiple ChIP-seq experiments

Ranciati, Saverio; Viroli, Cinzia; Wit, Ernst

Published in:
Statistical applications in genetics and molecular biology

DOI:
10.1515/sagmb-2014-0074

Publication date:
2015

Saverio Ranciati*, Cinzia Viroli and Ernst Wit

Spatio-temporal model for multiple ChIP-seq experiments

Abstract: The increasing availability of ChIP-seq data demands for advanced statistical tools to analyze the results of such experiments. The inherent features of high-throughput sequencing output call for a modelling framework that can account for the spatial dependency between neighboring regions of the genome and the temporal dimension that arises from observing the protein binding process at progressing time points; also, multiple biological/technical replicates of the experiment are usually produced and methods to jointly account for them are needed. Furthermore, the antibodies used in the experiment lead to potentially different immunoprecipitation efficiencies, which can affect the capability of distinguishing between the true signal in the data and the background noise. The statistical procedure proposed consist of a discrete mixture model with an underlying latent Markov random field: the novelty of the model is to allow both spatial and temporal dependency to play a role in determining the latent state of genomic regions involved in the protein binding process, while combining all the information of the replicates available instead of treating them separately. It is also possible to take into account the different antibodies used, in order to obtain better insights of the process and exploit all the biological information available.

Keywords: ChIP-seq; Markov random field model; MCMC; mixture distributions.

DOI 10.1515/sagmb-2014-0074

1 Introduction

In the context of genetic analysis, data from a biological technique known as Chromatin ImmunoPrecipitation-sequencing (*ChIP-Seq*) are becoming increasingly more frequent. These experiments arise from the combined process of Chromatin ImmunoPrecipitation (ChIP) and massive parallel DNA sequencing (Seq), providing as a final result the number of tags (reads) or fragments of DNA aligned to each region of the strand of genome inspected (Kharchenko et al., 2008). This technique is employed to provide insight about DNA methylation, chromatin and histone modifications and the interactions between proteins and DNA: in particular, certain proteins are known to be participating in various biological processes and thus being able to detect which regions are "activated" by the former can explain better the role of the latter into the process itself. The aim of the statistical analysis is thus to distinguish between enriched regions (bound by the protein) from those who are not. There are inherent features of the problem outlined that have to be taken into consideration. The chromatin modifiers and transcription factors involved in the experiment interact with broad regions of the DNA: this highlights the spatial dependency component that characterizes the phenomenon and also means that the usual peak-detection algorithms will not be able to retrieve correctly the enriched state of the locations. The ChIP phase of the experiment is performed with specific antibodies that carry different efficiencies, where the concept of efficiency is defined as the capability of

***Corresponding author: Saverio Ranciati,** Department of Statistics, University of Bologna, Bologna, Italy; and Johann Bernoulli Institute, University of Groningen, Groningen, The Netherlands, e-mail: saverio.ranciati2@unibo.it.
http://orcid.org/0000-0001-7880-9465
Cinzia Viroli: Department of Statistics, University of Bologna, Bologna, Italy
Ernst Wit: Johann Bernoulli Institute, University of Groningen, Groningen, The Netherlands

interacting only with the protein analyzed and providing the bits of DNA effectively bound by the transcription factor or chromatin modifier. Thus, the amount of background noise and quality of the signal in the data are affected by the ImmunoPrecipitation (IP) efficiencies of these antibodies. The spatial dependency component in the observations is further enhanced due to a pre-processing of the data: the raw results of the ChIP-seq are summarized in fixed-width windows and the original counts for the tags at a higher resolution are summed over into smaller regions that may contain more than one gene. In this experimental setting, it is possible to have observations from two or more different time points (e.g., immediately after the protein-DNA interaction and after 30 min): this naturally introduces a temporal dependency structure that needs to be modeled in order to correctly perform statistical inference about the enrichment behavior of the process analyzed. Multiple technological or biological replicates are also available, demanding for a tool that can take into account jointly all the information available instead of relying on separate analysis for each instance. Given the features of the problem, it is a common approach to model the data using a mixture of discrete distributions. There are already a number of methods that adopted this procedure, using different parametric densities for the mixture components: (Kuan et al., 2001; Spyrou et al., 2009). They employ Generalized Poisson or Negative Binomial densities (Hilbe, 2011) to model the data, without though including spatial and temporal dependency. A few exceptions to this common approach are proposed by (Bao et al., 2014) and (Zeng et al., 2013): in this works, the spatial structure feature has been considered, as it has been already done for other ChIP experiments in the literature, by using a Hidden Markov Model framework (Spyrou et al., 2009; Qin et al., 2010; Bao et al., 2014). Most of them however do not take into account the joint information of multiple biological or technical replicates: instead, they just analyze each experiment and retain only the common regions found to be enriched: they exploit a control sample in order to compare the separate results obtained from the replicates available and they asses a differential enriching behavior between them (see Bardet et al., 2012). There is, thus, a lack of a statistical approach to jointly use the information from all the replicates with the aim to obtain a more robust inference procedure. The model proposed by (Bao et al., 2014) is able to tackle all the aforementioned issues, except for the temporal aspect of the problem: this model allows to take into account both the spatial dependency and the different IP efficiencies belonging to each individual replicate and/or antibody used in the experiment. One missing aspect remains the temporal dependency structure arising when the experiment is performed at multiple time points. In this work we propose a more general version of the model showed in (Bao et al., 2014), in which the novelty is the introduction of the temporal dependency directly into the latent structure that characterizes the underlying biological process of ChIP-seq experiments. In order to do so, we employ a Markov random field (Kindermann and Snell, 1980) to describe the enrichment behavior and a flexible mixture model as a measurement model to account for the specific features and IP efficiencies of the antibodies used while jointly considering all the technical and biological replicates available. In Section 2, the model is described and the statistical aspects are discussed while describing the main quantities used in the implementation; in Section 3, results from a simulation study are showed to asses the performances of our model with respect to other existing algorithms; in Section 4, we summarize the output after applying our model to a real dataset; in Section 5, we discuss the features of the method proposed and the further developments.

2 Model and methods

We propose a more general version of the model used in (Bao et al., 2014) by extending the latent variables structure, allowing it to take into account also a temporal dependency between the same region at different time points. Let Y_{mtr} be the number of reads (*tags*) for the m-th bin (m=1, ..., M), at time point t (t=1, ..., T) for the replicate r (r=1, ..., R_t). The subscript r includes also the specification of the antibody or the conditions/treatments that share the same latent structure (e.g., R_1=2, for two replicates with the same antibody at the first time point; R_2=6, for two replicates for each of three different antibodies used at the second time point). We consider a joint mixture distribution as follows:

$$Y_{mtr} \sim p_t f^S(y|\theta^S_{tr}) + (1-p_t) f^B(y|\theta^B_{tr}) \tag{1}$$

where $p_t = P(X_{mt}=1)$ is the probability of region m to be enriched at time point t, with respect to a latent binary variable X_{mt} that represent the underlying biological process of protein binding. The vectors θ^B_{tr} and θ^S_{tr} contain the mean and dispersion parameters corresponding to the specific background or signal component of the mixture. We use two discrete densities f^B and f^S respectively to model the background and the signal components of the mixture in (Eq. 1): for the background component f^B we consider a Zero-Inflated Poisson (ZIP) or a Zero-Inflated Negative Binomial (ZINB), in order to account for both the presence of overdispersion in the data and abundance of zeros, features that are quite common in this context; as for the signal component f^S, we employ a Poisson (P) or negative binomial (NB) distribution. A further latent variable Z is introduced to represent the zero-inflated density f^B as a mixture itself of a zero-mass distribution and a discrete density as the ones previously described. In the case of a ZINB, we have the following representation:

$$f^B(y_{mtr}|\pi_{tr}, \mu_{tr}, \phi_{tr}) = \begin{cases} (1-\pi_{tr}) + \pi_{tr}\left(\dfrac{\phi_{tr}}{\phi_{tr}+\mu_{tr}}\right)^{\phi_{tr}} & \text{if } y_{mtr}=0 \\ \pi_{tr}\dfrac{\Gamma(y_{mtr}+\phi_{tr})}{\Gamma(\phi_{tr})\Gamma(y_{mtr}+1)}\left(\dfrac{\mu_{tr}}{\phi_{tr}+\mu_{tr}}\right)^{y_{mtr}}\left(\dfrac{\phi_{tr}}{\phi_{tr}+\mu_{tr}}\right)^{\phi_{tr}} & \text{if } y_{mtr}>0 \end{cases}$$

with μ^B_{tr} being the mean parameters and ϕ^B_{tr} the dispersion parameters of the negative binomial distribution. The parameters $\pi_{tr} = P(X_{mt}=0, Z_{mtr}=1)$ and $(1-\pi_{tr}) = P(X_{mt}=0, Z_{mtr}=0)$ are the weights of the mixture representing the zero-inflated distribution: the latter is the proportion of the background which is due to an abundance of zeros, while the former is related to the raw noise in the data. The conditional distribution of $\mathbf{Y}$, given both the latent binary variables $\mathbf{X}$ and $\mathbf{Z}$, is then the following:

$$\begin{aligned} Y_{mtr}|X_{mt}=0, Z_{mtr}=0 &\sim \mathbb{1}(y=0) \\ Y_{mtr}|X_{mt}=0, Z_{mtr}=1 &\sim Poisson(\lambda^B_{tr}) \text{ or } NB(\mu^B_{tr}, \phi^B_{tr}) \\ Y_{mt}|X_{mt}=1 &\sim Poisson(\lambda^S_{tr}) \text{ or } NB(\mu^S_{tr}, \phi^S_{tr}) \end{aligned}$$

The latent structure $\{X_{mt}\}$, with $X_{mt}=1$ for enriched region m at time point t and $X_{mt}=0$ otherwise, can be represented as an undirected graph with nodes corresponding to each bin at a specific time point or also as a lattice with a generic site (m, t) (Figure 1). The edges of the graph connect each X_{mt} only to adjacent regions

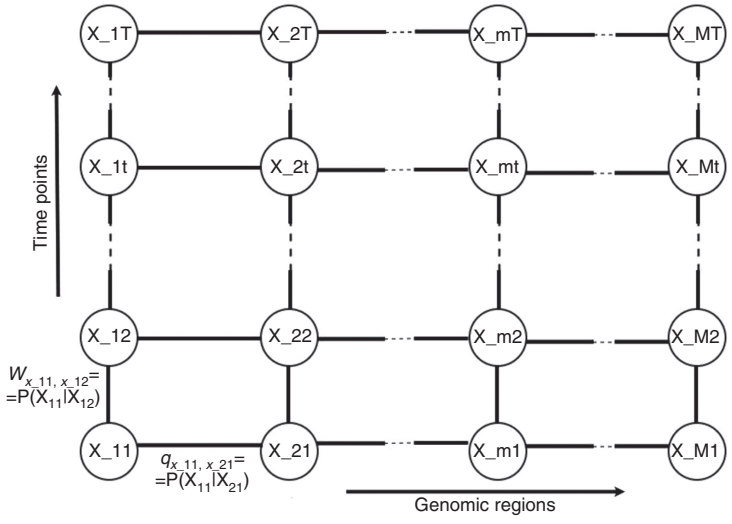

Figure 1 The undirected graph representing the latent structure **X** of the model.

in the genome and one-step backward/forward time points and they also are the *cliques* used to factorize the whole graph (Lauritzen, 1996). Each node represent a *separator* of these cliques, connecting two or more conditional relationships between neighboring variables.

These assumptions can be jointly expressed as a first-order Markovian property as follows:

$$P(X_{mt}=s|X_{(-m),(-t)})=P(X_{mt}=s|N(X_{mt})) \tag{2}$$

with $s=\{0,1\}$ and $N(X_{mt})=\{X_{m-1,t}, X_{m+1,t}, X_{m,t-1}, X_{m,t+1}\}$ being the neighborhood of the generic site (or node) (m, t), where $X_{(-m),(-t)}$ is the set of nodes except for the (m, t) one. This kind of latent structure is also called a Markov random field. The joint probability of the latent structure, given the assumption in (Eq. 2), can be factorized into:

$$\begin{aligned} P(\mathbf{X})=P(X_{11}, X_{21}, \ldots, X_{MT})&=\frac{\prod_{C\in cliques}P(X_C)}{\prod_{S\in separators}P(X_S)}= \\ &=P(X_{MT})\prod_{m=1}^{M-1}\prod_{t=1}^{T-1}P(X_{m+1,t}|X_{mt})P(X_{m+1,t+1}|X_{m+1,t})\prod_{m=1}^{M-1}P(X_{mT}|X_{m+1,T})\times \\ &\times\prod_{t=1}^{T-1}P(X_{1t}|X_{1,t+1})\prod_{m=2}^{M}\prod_{t=2}^{T}\frac{1}{P(X_{mt})} \end{aligned} \tag{3}$$

We introduce some notation to rewrite the conditional dependencies in (Eq. 3):

$$\begin{aligned} \delta_{j,k}&=P(X_{mt}=j, X_{m+1,t}=k) & \tau_{l,h}&=P(X_{mt}=l, X_{m,t+1}=h) \\ n_{j,k}&=\#\{X_{mt}=j, X_{m+1,t}=k\} & u_{l,h}&=\#\{X_{mt}=l, X_{m,t+1}=h\} \end{aligned}$$

$$q_{j,k}=\frac{\delta_{j,k}}{\delta_j} \quad w_{j,k}=\frac{\tau_{l,h}}{\tau_l}$$

$$q_0=q_{0,1} \quad q_1=q_{1,1} \quad w_0=w_{0,1} \quad w_1=w_{1,1}$$

and $j, k, l, h\in\{0,1\}$.

We assume that:

$$\begin{aligned} P(X_{mt}=1)&=\frac{q_0}{1-q_1+q_0}=\frac{w_0}{1-w_1+w_0} \\ P(X_{mt}=0)&=\frac{1-q_1}{1-q_1+q_0}=\frac{1-w_1}{1-w_1+w_0} \end{aligned}$$

and:

$$\begin{aligned} q_{0,1}&=q_{1,0} \quad w_{0,1}=w_{1,0} \\ q_0&=1-q_1 \quad w_0=1-w_1 \end{aligned}$$

We do so in order to parametrize the factorization of the latent structure $\{X_{mt}\}$ in (Eq. 3) without employing too many parameters; we assume that the spatio-temporal first-order markov dependency is the same for all the time points and regions considered: this means that the parameters q_1 and q_1 do not have any subscript and they are the same across the whole graph. More precisely, we want to capture the spatial dependency with the parameter q_1 while, conversely, retain the temporal information in the data through w_1. These two quantities represent the probability of a region remaining in the same state (bound or not by the protein) while moving respectively along the genome or from a time point to next/previous one. Their one's complements $1–q_1$ and $1–w_1$ measure the probability of the latent state to switch from 0 (not enriched) to 1 (enriched) and viceversa when moving from a node in the graph to its neighbors. In the following, we will only show the derivations for the negative binomial distribution used in modelling both the background and the signal

component. The joint complete likelihood for this model, considering a ZINB for background and NB for signal, is given by:

$$
\begin{aligned}
P(\mathbf{X},\mathbf{Z},\mathbf{Y}|\Theta) &= P(\mathbf{X}|\Theta)P(\mathbf{Z}|\mathbf{X}=0,\Theta)P(\mathbf{Y}|\mathbf{X},\mathbf{Z},\Theta) \\
&\propto q_1^{n_{1,1}+n_{0,0}}(1-q_1)^{n_{1,0}+n_{0,1}} w_1^{u_{1,1}+u_{0,0}}(1-w_1)^{u_{1,0}+u_{0,1}} \\
&\times \prod_{t=1}^{T}\prod_{r=1}^{R} \pi_{tr}^{\sum_{m=1}^{M}\mathbb{1}(X_{mt}=0,\,Z_{mtr}=1)}(1-\pi_{tr})^{\sum_{m=1}^{M}\mathbb{1}(X_{mt}=0,\,Z_{mtr}=0)} \\
&\times \prod_{t=1}^{T}\prod_{r=1}^{R}\prod_{m=1}^{M}\left[\frac{\Gamma(y_{mtr}+\phi_{tr}^{B})}{\Gamma(\phi_{tr}^{B})\Gamma(y_{mtr}+1)}\left(\frac{\mu_{tr}^{B}}{\phi_{tr}^{B}+\mu_{tr}^{B}}\right)^{y_{mtr}}\left(\frac{\phi_{tr}^{B}}{\phi_{tr}^{B}+\mu_{tr}^{B}}\right)^{\phi_{tr}^{B}}\right]^{\mathbb{1}(X_{mt}=0,\,Z_{mtr}=1)} \\
&\times \prod_{t=1}^{T}\prod_{r=1}^{R}\prod_{m=1}^{M}\left[\frac{\Gamma(y_{mtr}+\phi_{tr}^{S})}{\Gamma(\phi_{tr}^{S})\Gamma(y_{mtr}+1)}\left(\frac{\mu_{tr}^{S}}{\phi_{tr}^{S}+\mu_{tr}^{S}}\right)^{y_{mtr}}\left(\frac{\phi_{tr}^{S}}{\phi_{tr}^{S}+\mu_{tr}^{S}}\right)^{\phi_{tr}^{S}}\right]^{\mathbb{1}(X_{mt}=1)}
\end{aligned}
\tag{4}
$$

where the first term $P(\mathbf{X}|\Theta)$, representing the latent structure, is shared among the different replicates that are jointly considered in the model proposed. The features of each experiment are captured in the measurement model, which comprise of as many mean and dispersion parameters as antibodies used and it is characterized by $P(\mathbf{Z}|\mathbf{X}=0,\Theta)$ and $P(\mathbf{Y}|\mathbf{X},\mathbf{Z},\Theta)$. Inference is performed in a Bayesian framework: after choosing priors, the posterior distributions for all the parameters are derived along with the sampling schemes for the latent variables $\mathbf{X}$ and $\mathbf{Z}$. When the full conditional of the parameter is a non-conjugate distribution, a Metropolis-Hastings sampling procedure is used; otherwise, conjugate prior distributions and Gibbs samplings are involved. First, the latent variable X_{mt} is sampled from its full conditional distribution, according to the position of the node on the graph, the probability of the corresponding generic node (m, t) having state s being:

$$
P(X_{mt}=s|\ldots) \propto q_{X_{m-1,t},s}\; q_{s,X_{m+1,t}}\; w_{X_{m,t-1},s}\; w_{X_{m,t+1},s} \prod_{r=1}^{R_t} P(Y_{mtr}|X_{mt}=s)
$$

Similar derivations are obtained for nodes that lie on the borders and corners of the undirected graph (see Supplementary Material). Given the value of latent state $X_{mt}=0$, the latent variable Z_{mtr} is then sampled with a Gibbs method from its full conditional distribution:

$$
P(Z_{mtr}=s|X_{mt}=0,\ \ldots) \propto P(y_{mtr}|X_{mt}=0,Z_{mtr}=s,\Theta)P(Z_{mtr}=s|X_{mt}=0)
$$

The inflation parameter π_{tr} is sampled from its posterior Beta distribution

$$
\pi_{tr} \sim Beta\left(A_{\pi_{tr}}+\sum_{m=1}^{M}\mathbb{1}(X_{mt}=0,Z_{mtr}=1),\, B_{\pi_{tr}}+\sum_{m=1}^{M}\mathbb{1}(X_{mt}=0,Z_{mtr}=0)\right)
$$

and it represents the proportion of the background density component, f_B, which do not consist of the zero mass (inflation) distribution. The posterior densities for the transition probabilities are:

$$
\begin{aligned}
q_1 &\sim Beta(A_{q_1}+n_{1,1}+n_{0,0},\quad B_{q_1}+n_{1,0}+n_{0,1}) \\
w_1 &\sim Beta(A_{w_1}+u_{1,1}+u_{0,0},\quad B_{w_1}+u_{1,0}+u_{0,1})
\end{aligned}
$$

The mean and overdispersion parameters of the background and signal distributions are estimated through a Metropolis-Hastings procedure. A truncated normal distribution is used as a proposal to generate new values: this choice ensures that candidate values are always positive. The ratio of the normalizing constants has to be considered into the formula for the acceptance ratios (see the Supplementary Material for further derivations).

3 Simulation study

We assess the performance of the model proposed (*stMRF*) and compare it to other existing methods (*iSeq* Bardet et al., 2012; R package: iSeq, *MRF* Bao et al., 2014; R package: enRich). The scenarios used arise from the combination of many characteristics in the simulated data, such as the proportion of zero-inflation in the background component (π), the propensity to binding (low or high values of the transition probabilities), the difference in the mean level of the background and signal (low, high). We simulated four and two time points (T=4) and 2000 regions (M=2000) and three replicates for each temporal instance (R=3); 5000 MCMC iterations are performed, with a 2500 burn-in window. All the details are available in the Supplementary Material. Given that *MRF* and *iSeq* do not allow for more than one time point to be used, the two models are evaluated for each of the four simulated instances. The latter (*iSeq*), also, does not account for replicates and only the best result is showed for it (chosen between the three replicates available). The mean and overdispersion parameters of the negative binomial (μ, ϕ), along with the proportion of inflation (π), are well estimated by *MRF* and *stMRF*, without any pattern associated to the propensity of binding or degree of distance between background and signal components used in the scenarios. *iSeq* only allows for a Poisson distribution as background and signal and is unable to retrieve the true value of the mean parameters without introducing some bias. Between *MRF* and *stMRF*, the latter shows the lowest posterior standard deviations in all the scenarios. The transition probabilities are comparable only between *MRF* and *stMRF*, because *iSeq* do not account for the spatial dependency in the same way as the other two and employs a different parametrization. Both *MRF* and *stMRF* show a posterior mean distributions for q_1 close to the true values used in the simulation process and our model is also able to retrieve the parameter w_1 with a very low posterior standard deviation. The advantage of taking into account the temporal dependency in the model *stMRF* is clear when comparing the observed *false non discovery rate* (FNDR), which is the fraction of regions that were enriched but classified as not enriched, at a fixed estimated *false discovery rate* (FDR) (Benjamini and Hochberg, 1995) (see Table 1). In order to determine if a region m is enriched or not we put a cut-off value on the posterior probability $\hat{P}(X_{mt}=1)$ while controlling for the FDR to be 5%. If it is not possible to find such a cut-off point, we use a naive approach setting 0.5 as the threshold.

The *stMRF* model outperforms, in term of observed FNDR, the other two in every scenario simulated (three of which are presented in Table 1). *iSeq* cannot take into account all the replicates simultaneously, nor the temporal dependency, and thus has poor performances; *MRF* is able to perform as good as *stMRF* only at some instances of specific scenarios: this is due to the capability of the former to jointly use the replicates available. However, given that for *MRF* all the time points are treated as separate conditions/experiments with their own transition probabilities and it cannot model the temporal structure of the simulated data, it has generally higher values of observed FNDR.

Table 1 Observed FNDR (in percentage) at a fixed 5% estimated FDR for the three models at different scenarios, for all the time points.

Scenario	t	*iSeq*	*MRF*	*stMRF*
(12b)	1	35.71	3.87	1.68
π=0.9	2	35.80	6.56	2.87
Binding propensity: *high*	3	39.24	19.62	3.80
Ratio signal/background: *low*	4	33.03	1.73	1.12
(5b)	1	26.00	0.10	0.00
π=0.5	2	27.57	0.30	0.10
Binding propensity: *low*	3	28.72	0.81	0.40
Ratio signal/background: *High*	4	23.28	0.00	0.00
(8b)	1	29.80	0.08	0.00
π=0.5	2	40.50	0.00	0.00
Binding propensity: *high*	3	44.75	3.83	0.00
Ratio signal/background: *high*	4	42.71	0.00	0.00

4 Genome-wide assessment of differential roles for p300 and CBP in transcription regulation

Transcription coactivators p300 and CBP are known to partecipate in the regulation of genes responsible for many roles, especially in embryogenesis. Most of the genes related to the process regulated by these two TFs are bound by both, but (Ramos et al., 2010) showed that some of them may be preferentially bound to one or the other, following a different binding pattern. In particular, CBP has been found to regulate more genes that are involved in the negative transcription. Mostly, p300 and CBP have been showed to take a prominent role in important biological process of the cell such as proliferation, differentiation, and DNA repair mechanism. Also, some studies suggested that p300 and CBP may be involved in the development of cancer or other diseases (Ramos et al. 2010). The binding sites for the two transcription coactivators appear to be altered when the cell is stimulated or not, even if they retain an overlapping number of regions bound by both, and thus they are analyzed at different conditions and time points. We use our model to analyze the example dataset available in the package enRich (Bao et al., 2014) which contains data for p300 and CBP ChIP-seq experiments. We select data only for the protein p300 at time "zero" as our first time point and the two replicates obtained after 30 min as our second time point. The bins are already summarized from the processed data using a 1000bp window. The dataset contains 33,916 observations, each one corresponding to a region on the same chromosome 21. We run the MCMC algorithm with 10,000 iterations and a burn-in window of 5000, using a negative binomial distribution for both the background and signal component.

The posterior mean of π is close to one in both the time points analyzed: this means that a negative binomial distribution for the background could be potentially enough to capture the overdispersion in the data, without relying on a zero-inflated version of the model. The posterior standard deviations for the parameters of the measurement model (μ, ϕ) and the transition probabilities (q_1, w_1) are small. The posterior means for q_1 and w_1 are 0.96 and 0.99, the latter meaning that if a bin is enriched at t=1 it will very likely be enriched also at the next time point. There seems to be a strong dependency, thus, in both dimensions of the binding process (spatially and temporally). To decide if a bin m is enriched, we set a 0.5 cut-off to the posterior probabilities obtaining an estimated FDR and the number of enriched regions found by the algorithm. For both time points, the estimated FDR of *stMRF* is equal to 0.024 and the number of regions detected as bound by the protein is 3024 out of the total 33,196 observations. The number of regions bound by the protein identified by *MRF* is 3098, controlling for a fixed estimated FDR of 5%: if the same criteria is used for *stMRF*, the number of regions found as enriched is 3160 in both time points, which is slightly higher than the previous result (see Table 2). We show some of the bins analyzed (see Figure 2): the black dots are the observed counts of the dataset, while the red lines and squared dots represent the latent state of the corresponding bin estimated by the algorithm (model *stMRF*). To regions found to be enriched by our model correspond red dots aligned at the middle of the plot while, conversely, bins which are not bound by the protein (according to the algorithm) have red dots lying on the X axis. As in Figure 2, a genomic location could be labelled as not enriched by the algorithm even when presenting a count of tags as high as (spatially) neighboring regions: this is related to a correction effect that the temporal dependency structure can induce on each time point, allowing the detection of enriched locations while avoiding spurious binding that may occur due to the antibody used or other noise in the process. This can help understanding, through a validation process, the dynamics of regulation

Table 2 Number of enriched regions and estimated false discovery rate for *MRF* and *stMRF*.

Model	No of enriched regions	$\hat{FDR}$
MRF	3089	5%
([a]) stMRF	3160	5%
([b]) stMRF	3024	2.4%

[a]Cut-off on the posterior probability of X=1 equal to 0.5; [b]Cut-off on the posterior probability of X=1 equal to 0.24.

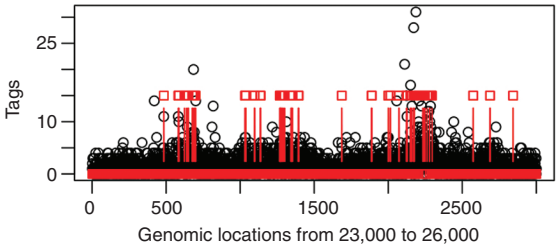

Figure 2 Plot of observed counts for the first time point along the chromosome 21. Estimated latent states of the bins are overplotted.

and activation of genes known to be related to this transcription factor (Ramos et al. 2010), while also pointing out new locations that could need a more detailed and specific re-sequencing after the experiments are performed and analyzed.

5 Conclusion and discussion

We have proposed a model that extended the one introduced by (Bao et al., 2014), incorporating in a parsimonious way a new dependency structure that spans through the time dimension, enabling it to account for the temporal aspect of the binding process. This has been done in order to exploit the information available at a specific time point to strengthen the inference on the other ones, while keeping a separate measurement model to better capture the inherent features of each experiment performed. We allowed the model to account for many characteristics that arise in the context analyzed, such as the inflation of zeros, the overdispersion in both the background (noise) and the signal of the data, the presence of different antibodies used and many replicates available for each experimental condition. We showed in a simulated environment the performances of our algorithm in comparison to other existing methods, assessing that it has a better degree of classifying the latent states of the regions as bound or not by the protein, scoring lower values of misclassification error in terms of observed False Non Discovery Rate. We applied then the model to real data and compared the results to the ones obtained with (Bao et al., 2014), showing a greater capability of detecting enriched bins and providing lower misclassification error in the process. Exploiting the temporal information can help with the detection of the regions that are enriched because bound by the protein and not due to a bias effect induced by a prolonged activity of the antibody used for the experiment. Also, replicates with low efficiency and ratio between signal and background can still be used to strengthen the inference about the parameters of the model, without having a negative huge impact on the recovering of the latent state of the genomic location, enabling the use of all the information available. Further developments include the multivariate modelling of the multiple antibodies, through the use of multivariate discrete distributions or other approach that allows for the specification of a dependency structure (similarity) between antibodies and/or transcription factors. Extending the mixture distribution to incorporate more components to better characterize the signal in the data can also be done. A more flexible parametrization of the spatio-temporal dependency could be introduced, using more than two transition probabilities in order to relax some of the initial assumptions of our model. The pre-processing of the data, such as the choice of the width of the bins used to summarize the counts, could be incorporated at a modelling level. The MCMC algorithm can also be improved in terms of computational speed.

References

Bao, Y., V. Vinciotti, E. Wit and P. T. Hoen (2014): "Joint modelling of ChIP-seq data via a Markov Random Field model," Biostatistics, 15(2), 296–310.

Bardet, A. F., Q. He, J. Zeitlinger and A. Stark (2012): "A computational pipeline for comparative ChIP-seq analyses," Nature Protocols, 7(1), 45–61.

Benjamini, Yoav and Yosef Hochberg (1995): "Controlling the false discovery rate: a practical and powerful approach to multiple testing," Journal of the Royal Statistical Society. Series B (Methodological), 57(1), 289–300.

Hilbe, J. M. (2011): Negative binomial regression, Cambridge University Press, Cambridge, England, UK.

Kharchenko, Peter V., Michael Y. Tolstorukov and Michael Y. Park (2008): "Design and analysis of ChIP-seq experiments for DNA-binding proteins." Nature Biotechnology, 26, 1351–1359.

Kindermann, Ross and J. Laurie Snell (1980): "Markov Random Fields and Their Applications," American Mathematical Society, Providence, USA.

Kuan, Pei Fen, Dongjun Chung, Guangjin Pan, James A. Thomson, Ron Stewart and Sündüz Keleş (2001): "A statistical framework for the analysis of ChIP-Seq data," Journal of the American Statistical Association, 106(495), 891–903.

Lauritzen, Steffen L. (1996): Graphical models. Oxford University Press, Oxford, England, UK.

Qin, Zhaohui S., Jianjun Yu, Jincheng Shen, Christopher A. Maher, Ming Hu, Shanker Kalyana-Sundaram, Jindan Yu and Arul M. Chinnaiyan (2010): "HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data," BMC Bioinformatics, 11(1), 369.

Ramos, Yolande F. M., Matthew S. Hestand, Matty Verlaan, Elise Krabbendam, Yavuz Ariyurek, Michiel van Galen, Hans van Dam, Gert-Jan B. van Ommen, Johan T. den Dunnen, Alt Zantema and Peter A. C. 't Hoen (2010): "Genome-wide assessment of differential roles for p300 and CBP in transcription regulation," Nucleic Acids Research, 38(16), 5396–5408.

Spyrou, C., R. Stark, A. G. Lynch and S. Tavar (2009): "BayesPeak: Bayesian analysis of ChIP-seq data," BMC Bioinformatics, 10(1), 299.

Zeng, Xin, Rajendran Sanalkumar, Emery H. Bresnick, Hongda Li, Qiang Chang and Sündüz Keles (2013): "jMOSAiCS: joint analysis of multiple ChIP-seq datasets," Genome Biology, 14(4), R38.

Supplemental Material: The online version of this article (DOI: 10.1515/sagmb-2014-0074) offers supplementary material, available to authorized users.