Advances and Applications in Statistics
© 2018 Pushpa Publishing House, Allahabad, India
http://www.pphmj.com
http://dx.doi.org/10.17654/AS053060659
Volume 53, Number 6, 2018, Pages 659-691 **ISSN: 0972-3617**

FLEXIBLE STATISTICAL MODELLING OF THE OCCURRENCES OF TRANSCRIPTION FACTOR BINDING SITES ALONG A DNA SEQUENCE

G. Kallah-Dagadu[1,*], B. K. Nkansah[2] and N. K. Howard[2]

[1]Department of Statistics and Actuarial Science
University of Ghana
Ghana

[2]Department of Mathematics and Statistics
University of Cape Coast
Ghana

Abstract

Experimental studies used in investigating the binding of specific factors to DNA along the genome are time consuming and expensive. Additionally, the increasing amount of data being produced with binding sites for different transcriptional regulators call for modern computational techniques for analysing binding patterns of several factors. In this paper, flexible statistical modelling techniques in the form of multivariate Hawkes processes have been used to model the occurrences of transcriptional regulatory elements (TREs) and their interaction along DNA sequence using 1% human genome ENCODE pilot data. We employed statistical procedures and techniques to model the transcription factor binding sites of ten TREs through favoured or avoided distances. It is generally revealed that there is significant interaction among transcription factor binding sites. In

Received: July 20, 2018; Accepted: September 30, 2018

Keywords and phrases: multivariate Hawkes model, transcription factors binding sites, point processes, interaction among transcription factors.

[*]Corresponding author

addition, similar patterns of interaction effects of each TREs on the others are observed. In all cases, the Hawkes log kernel model gives a better fit. The model, which is also in terms of histone modification elements, adequately captures the extreme inter-distances that usually characterise the transform point processes.

1. Introduction

The accurate execution of biological processes such as development, propagation, apoptosis, aging, and variation requires a precise and carefully orchestrated set of steps that depend on the proper spatial and chronological expression of genes (Maston et al. [19]). A fertilized egg of mammals becomes a baby after some months and in the next years grows to become an adult. Without the regulation of genes, the monumental development from a fertilized egg to an adult containing millions of differentiated specialized cells, all containing the same genetic information, would be impossible (Carstensen [4]). How genes are being regulated need to be understood, since for all living organisms, hereditary information is passed from parents to their offspring. This information is stored in the DNA, DeoxyriboNucleic Acid, found in the nucleus of a living organism's cell. Statistical analysis of transcription factor (motif) occurrences along a DNA sequence is an important task in computational molecular biology. Motifs are short sequences which allow interactions between DNA and proteins and initiate biological processes like gene transcription, restriction, DNA repair, replication, recombination and more (Gusto and Schbath [9]).

Most of the functional motifs of the sequence of these DNA bases have been identified and some unanimity patterns have been proposed by researchers. It will be beneficial to determine functional motifs based on the statistical properties of their occurrences. The unexpected number or frequency of the occurrences of these motifs is proposed to follow Markov models (Reinert et al. [23]). Also, Robin and Daudin [25] and Stefanov [30] have used statistical methods to identify the rich and poor regions of motifs concentration in the DNA sequence. Gusto and Schbath [9] show that, for a common biological process, the space between the occurrences of any two motifs (not necessarily consecutive) should be somehow fixed-favoured or avoided distances using Hawkes model. Transcriptional regulatory elements (TREs) are promoters or enhancers that bind to DNA sequence to enhance or repress gene expression. These elements interact and there exists dependence between the positions of the motifs or TREs.

Experimental studies used in investigating the binding of specific factors to the DNA along the genome are time consuming and expensive (Carstensen [4]). However, an increasing amount of data is being produced with binding sites for different transcriptional regulators, for instance by ChIP-seq or ChIP-chip methods (Park [22]). Statistical computational methods for analysing binding patterns of different factors or the occurrences of different motifs (TREs) and the dependencies between the motifs or transcription factors along the genome can be useful for gaining insight into the complex nature of transcriptional regulation, even though they cannot completely replace the experimental validation of interactions between events (Carstensen [4]). Modelling the joint occurrences and dependencies among events in a real line is a common problem in statistics.

The ultimate difficulty in gene regulation is the inadequate information in the DNA binding revealed by transcription factors, which leads to several untrue positives when predicting sites in genome sequences (Wasserman and Sandelin [32]). The combinations of binding sites produce additional information which are rich and responsible for tissue-specific expression, which can also be used for predictions (Krivan and Wasserman [16]; Wasserman and Fickett [31]). The advancement of experimental techniques has made it conceivable to measure the binding of DNA-binding proteins over a whole or partial genome; that is, Chromatin Immuno-Precipitation (ChIP) followed by sequencing, which is known as ChIP-seq (Park [22]) or ChIP followed by hybridization to DNA probes covering the genome, which is known as ChIP-chip (Buck and Lieb [3]). This leads to new ways of analysing gene regulation and especially the interaction between regulators or transcription factors. Though there have been improvement in computational and experimental developments in analysing regulatory factors, there is still more to do computationally in the area of multivariate statistical analysis of the joint binding sites of multiple transcription factors and occurrences of other transcriptional regulatory elements.

The paper's goal is to use ENCODE pilot project data to model the dependence between transcription factors in order to identify favoured or avoided distances between TREs and also model the ground conditional intensity rates of these TREs. We employ statistical techniques to study the occurrences of transcriptional regulatory elements (TREs) along genomes. The paper uses the Hawkes process for modelling the joint occurrences of multiple TREs along the genome, which is flexible and capable of capturing the dependencies among elements involved in transcriptional regulation.

2. Methods

This section presents the methods and data used for analysis in this paper. The paper begins with introduction to point processes and then continuous with Hawkes process, considering the parameterisation of the conditional intensities, estimation of model parameters and finally, models diagnostic techniques for goodness-of-fit tests and the datasets are presented.

2.1. Introduction of point processes

The simplest point process is the Poisson process, where points arrive independently of each other. Point processes considered in the literature and relevant to this paper are all of a special kind called *Hawkes processes* (Hawkes [15]). A Hawkes process is a special point process of which the occurrence of a point affects the probability of occurrences of other points in a specific direction given by a function of the distance to the point (Hawkes [15]). A point process provides a model for points or events that occur randomly in time or time and space. The paper considers point processes occurring along the DNA sequence, where events can be represented as points on a one-dimensional space. It is assumed that only one point or event occurs at a time, giving a simple point process. The events of interest along the DNA sequence are the binding sites of transcription factors and could also represent the position of any feature such as transcription start sites.

A simple point process on $\mathbb{R}_+$ consists of a sequence of points $t(i)_{i \in \mathbb{N}}$, where $0 \leq t(1) < t(2) < \ldots$. The corresponding counting process to this simple point process is given by $N(t) = \sum_{i \in \mathbb{N}} 1_{[t(i) \leq t]}$, $t \in \mathbb{R}_+$ and the common point process is the Poisson process, where points occur completely at random on a given time interval. However, a homogeneous Poisson process arises when points occur with intensity λ, the mean number of points arriving in an interval $(0, s]$ is equal to λs. Generally, the theory of point process on the real line shows that, points do not occur completely at random but for a given position *t*, the information about previous points are contained in the history of the process. The intensity of point process is in general case dependent on the history of points before *t*. The intensity is a generalized form of the hazard function known from survival analysis, and a large intensity at a given position means that there is a relatively large probability that a point occurs immediately after that position.

2.2. Hawkes processes

Hawkes processes are point processes equivalent to autoregressive models. In seismology, Hawkes processes have been used to model earthquakes and their aftershocks, favoured or avoided distances between occurrences of motifs or transcriptional regulatory elements (TREs) on the DNA sequence and social interactions or financial phenomena (Hansen et al. [11]). However, there exist many equivalent forms of definition for Hawkes point process, the standard Hawkes process can be defined as a temporal point process with long memory, branching effect and self-exciting properties. Hawkes process is characterized by its associated conditional intensity process which describes the underlying dynamics of the process in a convenient system.

Definition 2.1. The univariate unmarked Hawkes process $N(t)$ with conditional intensity $\lambda^*[t \mid \mathcal{H}(t)]$ ($\mathcal{H}(t)$, the history of the occurrence of events), is defined for all values of *t* and $h \to 0$ by

$$P\{N[t, t+h) = k \mid \mathcal{H}(t)\} = \begin{cases} \lambda^*[t \mid \mathcal{H}(t)]h + o(h), & \text{for } k = 1, \\ o(h), & \text{for } k > 1, \\ 1 - \lambda^*[t \mid \mathcal{H}(t)]h + o(h), & \text{for } k = 0. \end{cases} \tag{2.1}$$

The conditional intensity is defined for all values of *t* as

$$\lambda^*[t \mid \mathcal{H}(t)] = \begin{cases} \mu + \int_{(-\infty, t)} h(t-s)dN(s), & \text{for continuous case,} \\ \mu + \sum_{i:t_i < t} h(t - t_i), & \text{for discrete case,} \end{cases} \tag{2.2}$$

where $\mu > 0$, $h(s) \geq 0$ for $s \geq 0$ and zero otherwise. If the Hawkes process is assumed to be stationary, then $\int_0^\infty h(s)ds < 1.$ The parameter μ and the function $h(\cdot)$ are known as the background intensity and excitation function, respectively. The structure of conditional intensity is flexible, and it involves the description of the background intensity $\mu > 0$ and the excitation function $h(\cdot)$. The common excitation function used is an exponential decay function (Hawkes [15]; Hautsch [14]; Laub et al. [17]). In this situation, $h(t) = \alpha e^{-\beta t}$, where the constants $\alpha, \beta > 0$ are the parameters and hence

$$\lambda^*[t \mid \mathcal{H}(t)] = \begin{cases} \mu + \int_{-\infty}^{t} \alpha e^{-\beta(t-s)} dN(s), & \text{for continuous case,} \\ \mu + \sum_{t_i < t} \alpha e^{-\beta(t-s)}, & \text{for discrete case.} \end{cases} \tag{2.3}$$

The parameters α and β are interpreted as; each arrival in the system instantaneously increases the arrival intensity by α, then over time this arrivals influence decays at rate β.

Definition 2.2. Suppose a counting process $N(\cdot)$ has a conditional intensity function of the form

$$\lambda^+[t \mid \mathcal{H}(t)] = \phi\left(\int_{-\infty}^{t} h(t-s)dN(s)\right), \tag{2.4}$$

where $\phi : \mathbb{R} \to \mathbb{R}_+$ and $h : \mathbb{R}_+ \to \mathbb{R}$. Then $N(\cdot)$ is known as a nonlinear Hawkes process.

A particular case is Hawkes's self-exciting point process, for which $h(t)$ is non-negative and $\phi(x) = \mu + x$, where $\mu > 0$, reduces $N(\cdot)$ to the linear Hawkes process of Definition 2.1 (Daley and Vere-Jones [6]). This univariate nonlinear Hawkes intensity function, for instance, can be explained with regards to seismological purposes as: μ describes the spontaneous occurrences of real original earthquakes while the function $h(\cdot)$ models the self-interaction after a shock at time s, we observe an aftershock at time t with large probability if $h(t - s)$ is large.

Definition 2.3. Suppose $\mathcal{N}(t) = \{N_1(t), N_2(t), ..., N_d(t)\}$ is a multivariate point process, then the conditional intensity of the nonlinear multivariate Hawkes process is of the form

$$\lambda_j^+[t \mid \mathcal{H}(t)] = \phi_j\left(\sum_{\ell=1}^{d} \int_{-\infty}^{t} h_{j\ell}(t - s)dN_j(s)\right), \tag{2.5}$$

where the symbols denote their usual meaning as explained in Definition 2.2. Bremaud and Massoulie [2] state the conditions on the functions ϕ_j (namely, Lipschitz properties) and on the functions $h_{j\ell}(\cdot)$ to obtain existence and uniqueness of a stationary version of the associated process. Suppose that for any $j \in \{1, 2, ..., d\}$, the ϕ_j identity function form is

$$\phi_j(x) = (\mu_j + x)_+, \tag{2.6}$$

where $\mu_j > 0$ and $(\cdot)_+$ denotes the positive part. Now, introducing a linear predictable transformation, $\psi(f) = (\psi_1(f), ..., \psi_d(f))$, for any j and any t, we obtain

$$\psi_{t,\,j}(f_0) = \mu_j + \sum_{\ell=1}^{d} \int_{-\infty}^{t} h_{j\ell}(t - s)dN_j(s), \tag{2.7}$$

where $f_0 = (\mu_j, (h_{j\ell})_{\ell=1,2,\ldots,d})_{j=1,2,\ldots,d}$ and $\lambda_j^+[t \mid \mathcal{H}(t)] = (\psi_{t,j}(f_0))_+$. In this paper, $\lambda_j(\cdot)$ is the intensity of a realization of the point process (i.e., a chromosome, in this study) with the mark type *j* (TRE). The $h_{j\ell}$-functions represent the effect of the occurrence of points of type *j* (TRE) on subsequent points of type ℓ (TRE). The multivariate point process associated with this setup is called the *multivariate Hawkes self-exciting point process* (Hawkes [15]) and in this case, *d* is fixed and asymptotic properties are obtained when *t* tends to infinity.

2.3. Parameterisation and estimation of conditional intensity

In this paper, we have considered a multivariate Hawkes process which is the generalised linear Hawkes process. The conditional intensity, $\lambda_j[t \mid \mathcal{H}(t)]$ from now on shall be written as $\lambda_j(t)$. The parameterisation of its conditional intensity is of the form

$$\lambda_j(t) = \phi\left(\mu_j^T \mathbf{X}(t) + \sum_{\ell=1}^{d} \int_{-\infty}^{t} h_{j\ell}(t-s)\, dN_j(s)\right), \tag{2.8}$$

where $\phi : I \to [0, \infty)$, $I \subseteq \mathbb{R}$, is a given function, the function $h_{j\ell}(\cdot)$ is modelled as cubic *B*-spline basis functions and the process $\mathbf{X}(t)$ is an auxiliary, $\mathrm{d}(0)$-dimensional observed processes (at least discretely). In principle, the function $h_{j\ell}(\cdot)$ could be arbitrary functions where the parameter space could be infinitely dimensional. The $h_{j\ell}(\cdot)$ function is modeled as a linear combination of spline functions given as

$$h_{j\ell}(t) = (\beta_{j\ell})^T B(t) = \sum_{k=1}^{\mathrm{d}(\ell)} \beta_{j\ell}^k B_k(t), \tag{2.9}$$

where $\beta_{j\ell}$ is parameter vector and $\beta_\ell \in \mathbb{R}^{\mathrm{d}(\ell)}$, the $B_k(\cdot)$s are cubic *B*-spline basis functions such that $h_{j\ell}(\cdot)$ is a cubic spline (Green and Silverman [8]). The cubic *B*-splines with fixed equidistant knots are used to model $h_{j\ell}(\cdot)$

and the value of the largest knot gives the maximum range within which the dependencies are detected. The number of knots determines how detailed the description of the dependencies can be. Choosing too many knots will lead to over-fitting of the model and base on literature, we choose six knots as breakpoints for the cubic *B*-spline and log and identity link functions (kernels) are considered in estimating the parameters in this paper.

The vector of parameters, $\theta_j = (\mu_j, \beta^k{}_{j\ell}) \in \Theta \subset \mathbb{R}^p$ is given by

$$\theta_j = (\mu_j^1, \mu_j^2, ..., \mu_j^{\mathrm{d}(0)}, \beta_{j1}^1, \beta_{j1}^2, ..., \beta_{j1}^{\mathrm{d}(1)}, ..., \beta_{jd}^1, \beta_{jd}^2, ..., \beta_{jd}^{\mathrm{d}(d)}). \quad (2.10)$$

Now, suppose we have observations point processes of the multivariate point process such that, $s_1^\ell < s_2^\ell < \cdots < s_{n_\ell}^\ell < t$, for $\ell = 1, 2, ..., d$, then intensity $\lambda_j(\cdot)$ can be expressed as

$$\begin{aligned}\lambda_j(t) &= \phi\left(\mu_j^T \mathbf{X}(t) + \sum_{\ell=1}^{d}\sum_{i=1}^{n_\ell} h_{j\ell}(t - s_i^\ell)\right)\\ &= \phi\left(\mu_j^T \mathbf{X}(t) + \sum_{\ell=1}^{d}\sum_{k=1}^{\mathrm{d}(\ell)} \beta_{j\ell}^k \sum_{i=1}^{n_\ell} B_k(t - s_i^\ell)\right)\\ &= \phi(\theta^T \mathbf{Z}(t)), \qquad (2.11)\end{aligned}$$

where $\mathbf{Z}(t)$ is a process of dimension $p = \mathrm{d}(0) + \mathrm{d}(1) + \cdots + \mathrm{d}(d)$. Each of the linear filter components $\sum_{i=1}^{n_\ell} B_k(t - s_i^\ell)$ is computable from the observations and the fixed choice of basis. The minus logarithm-likelihood function for the *j*th mark type conditional intensity process is of the form

$$l_j(\theta) = \int_0^t \phi(\theta^T \mathbf{Z}(s))ds - \sum_{i=1}^{n} \log\phi(\theta^T \mathbf{Z}(t_i)). \quad (2.12)$$

The integral part of the minus log-likelihood function expression is not in general, analytically computable. The computation can be done by discretising the limit interval $(0 - t)$, to form a total of T_0, time points. The values of $\mathbf{Z}(t)$ at this discretised points are computed to form a matrix, **Z** of dimension $T_0 \times p$. If we let Δ denote a p dimensional vector of inter-distances from the discretisation, then an approximation to the integral part of $l_j(\theta)$ is given by Hansen [12], as

$$\int_0^t \phi(\theta^T \mathbf{Z}(s))ds \approxeq \Delta^T \phi(\mathbf{Z}\theta). \tag{2.13}$$

The approximated minus log-likelihood function is

$$l_j(\theta) \approxeq \Delta^T \phi(\mathbf{Z}\theta) - \sum_{i=1}^{n} \log \phi(\theta^T \mathbf{Z}(t_i)), \tag{2.14}$$

where the derivatives in the same manner approximated.

All the inter-distances of Δ may be equal but generally it is advantageous to have different values for the inter-distances. This is especially situations where the intensity is constant over a region and fluctuates over other regions. The approximated expression gives a fast computation of an approximation to the minus log-likelihood function as well as its first and second derivatives. The major problem with this approach is, however, that the matrix **Z** may become extremely large for large datasets or models with many parameters (Carstensen [4]). Hawkes point processes are usually fitted with both parametric and non-parametric estimation techniques. The standard method of estimating the parameters of the Hawkes process is the use of maximum likelihood estimation techniques. In general, there is no closed form for the maximum likelihood estimates and the likelihood function has to be optimized by standard numerical maximisation algorithms (Carstensen et al. [5]; Embrechts et al. [7]).

2.4. Diagnostic methods

Numerical maximisation algorithms are employed to search over the parameter space for the set of parameters that can maximise the log-likelihood since there is no close form. Therefore, assessing the goodness-of-fit of the model parameters of some dataset is an important practical consideration. In performing this assessment the point process' compensator is essential, as is the random time change theorem (Laub et al. [17]).

2.4.1. Transformation to a Poisson process

The random time change theorem states that if $\{t_1, t_2, ..., t_n\}$ is a realisation over a time interval $[0, t]$ from a point process $N(\cdot)$ with conditional intensity function $\lambda(\cdot)$, and if $\lambda(\cdot)$ is positive over $[0, t]$ and $\Lambda(t) < \infty$, *a.s.*, then the transformed points $\{\Lambda(t_1), \Lambda(t_2), ..., \Lambda(t_n)\}$ form a Poisson process with unit rate. The random time change theorem is fundamental to the model fitting procedure known as residual analysis.

The residual analysis states that if there is an unbounded, increasing sequence of time points $\{t_1, t_2, ...\}$ on $\mathbb{R}_+$, and a monotonic continuous compensator $\Lambda(\cdot)$ such that $\lim_{t-\infty} \Lambda(t) = \infty$, *a.s.*, then the transformed sequence $\{t_1^*, t_2^*, ...\} = \{\Lambda(t_1), \Lambda(t_2), ...\}$, with counting process $N(t)$ is a realisation of a unit Poisson process if and only if the original sequence $\{t_1, t_2, ...\}$ is a realization from the point process defined by $\Lambda(\cdot)$. Therefore, with a close form of the compensator, statistical inference can be conducted to test for Poisson process fitness.

2.4.2. Test for Poisson process

There are various techniques for testing whether a series of points form a Poisson process. To examine whether the transformed point process follow a homogeneous Poisson process, one can run the basic test by examining the hypothesis that $\sum_i 1_{\{t_i < t\}} \sim Pois(t)$. If this basic test succeeds, then the inter-arrival times $\{\tau_1, \tau_2, \tau_3, ...\} = \{t_1^*, t_2^* - t_1^*, t_3^* - t_2^*, ...\}$, will be *i.i.d.* exponentially distributed with unit rate. A quantile-quantile (q-q) plot can be employed to graphically investigate the fitness or a Kolmogorov-Smirnov test which can quantitatively test the fitness of the exponential distribution to the data. The test for independence is employed to examine the autocorrelation in the τ_i sequence. Understandably, the zero autocorrelation does not imply independence, but a non-zero amount would certainly imply a non-Poisson model. A graphical examination of the autocorrelation is conducted by plotting the points (U_{i+1}, U_i), where $U_i = 1 - e^{-\tau_i}$. If there are noticeable patterns, then the τ_i are autocorrelated. Otherwise the points should look evenly scattered. The transformed points can be extended to a multivariate point process, $(N_j)_{j \in \{1, 2, ..., d\}}$, given as the random time changes

$$t_i^{j*} = \int_{t_0}^{t_i^j} \lambda_j(s) ds, \quad j \in \{1, 2, ..., d\}, \tag{2.15}$$

where t_i^{j*} transformed the multivariate point process, $(N_j(\cdot))_{j \in \{1, 2, ..., d\}}$, into a multivariate Poisson process with independent variables each having unit rate (Daley and Vere-Jones [6]).

2.5. Description of dataset

The paper aims at employing computational techniques to model occurrences and interactions between transcription factor binding sites along a genome or DNA sequence. It employs a dataset known as ENCODE pilot data. The ENCODE data is a pilot project of 1% human genome (The ENCODE Project Consortium [27]). This 1% human genome is a ChIP-chip data which is directed by Affymetrix and studied by several laboratories and researchers (Carstensen et al. [5]). The data consists of 44 regions where transcriptional regulatory elements (TREs) concentrations are located and studied by several different laboratories. Our study considers the ChIP-chip data steered by Affymetrix of the ENCODE pilot project. The data contains locations of binding sites, within the ENCODE pilot regions, for eight different TREs and two histone modifications in retinoic acid stimulated human HL-60 cells gathered after 0, 2, 8 and 32 hours. ENCODE data contains locations of binding sites for ten different TREs, namely: BRG1:- SWI/SNF related, matrix associated, acting dependent regulator of chromatin, sub-family a, member 4; CEBPE:- CCAAT/enhancer binding protein (C/EBP), epsilon; CTCF:- CCCTC-binding factor (zinc finger protein); H3K27me3 (H3K27):- Histone H3 tri-methylated lysine 27; H4KAC4 (HisH4):- Histone H4 tetra-acetylated lysine; P300:- E1A binding protein p300; PU1:- Spleen focus forming virus proviral integration oncogene; RARA (RARecA):- Retinoic Acid Receptor-Alpha; RNAP:- RNA polymerase II; SIRT1:- sirtuin (silent mating type information regulation 2 homolog) 1.

3. Results and Discussions

In this section, we present results and discussions on the analysis of ENCODE pilot phase data of 1% of the human genome. The section begins with preliminary analysis for the occurrences of ENCODE data as a point process and later continues with the modelling of TFBS as multivariate Hawkes process. Due to the huge values of the occurrence times of this transcription factor binding sites, which requires super computers with a minimum of 30 gigabytes (GB) RAM for running the analysis, the data points are transformed by computing the square-root of the points and then divided the result by 100.

3.1. Preliminary analysis

Table 1 shows a summary of the dataset employed. It shows the number of TREs occurrence in each chromosome and the total number of occurrence of the transcription factor binding sites (TFBS). It is observed that, H4KAC4 recorded the highest number of occurrences of TFBS along the sequence, and SIRT1 recorded the least number of occurrences of TFBS along the genome sequence under study. It is also observed that Chromosome 7 recorded the highest number of TFBS whilst the lowest is observed in Chromosome 9 with none occurrence of SIRT1 TRE.

Figure 1 shows a detailed occurrences of the transcription factor binding sites (TFBS) of TREs along the first six chromosomes of the ENCODE sequence. A combined occurrence of the TFBS along a sequence is displayed on the first two sequences on top of each chromosome in Figure 1. The bottom part of the six chromosomes graphs display the individual occurrences of transcriptional regulatory elements along the genome sequence. Generally, TFBS can be observed to occur in a genome sequence as displayed in the two combined sequences on the top part of the six magnified chromosome graphs. The plots show the occurrences of the TFBS as point processes with their corresponding times. Time of occurrences is measured in units of base pair (bp) units.

Table 1. Frequency distribution of the events of TFBS in each chromosome

Chrom.	BRG1	CEBPE	CTCF	H3K27	H4KAC4	P300	PU1	RARA	RNAP	SIRT1	TOTAL
chr1	15	28	9	24	56	15	16	31	40	7	241
chr2	92	58	56	112	106	76	29	109	60	5	713
chr4	22	32	20	34	29	36	2	43	25	13	256
chr5	49	59	65	82	176	84	37	126	59	13	750
chr6	92	84	76	106	186	67	37	134	63	17	862
chr7	311	342	234	543	596	324	100	642	303	61	3456
chr8	25	22	21	59	18	24	15	39	27	1	251
chr9	5	6	4	10	34	6	6	7	9	0	87
chr10	39	33	31	31	21	35	2	54	31	11	288
chr11	73	95	61	109	150	81	36	128	91	8	832
chr12	31	48	32	40	84	39	19	77	23	12	405
chr13	23	36	18	40	65	26	13	34	29	2	286
chr14	34	42	32	52	45	34	20	37	33	4	333
chr15	22	19	14	21	50	19	20	36	34	1	236
chr16	60	54	38	59	47	64	20	55	45	7	449
chr18	66	91	53	80	75	85	38	162	71	19	740
chr19	17	44	9	22	53	25	28	37	48	4	287
chr20	17	22	10	16	51	18	9	31	39	6	219
chr21	70	147	82	100	302	89	58	229	183	15	1275
chr22	46	47	31	65	112	56	33	67	73	8	538
chrX	45	42	28	67	114	48	27	66	75	12	524
Total	1154	1351	924	1672	2370	1251	565	2144	1361	236	13028

Chromosomes 2 and 6 show evenly arrivals of TREs between 148.30-148.40 and then clustering is observed afterwards. However, the occurrences of the TFBS in Chromosomes 4 and 7 shows complete clustering over the magnified period. It can be observed that in Chromosomes 1 and 5 the case is different; the occurrence of the TFBS is a mixture of clustering and evenly arrivals over the magnified range. The occurrences of the TFBS along the chromosomes as displayed in Figure 1 demonstrate that the TFBS occur in batches (clustering). This shows the tendency of interaction among themselves leading to biological and chemical reactions in the nucleus of the cells.

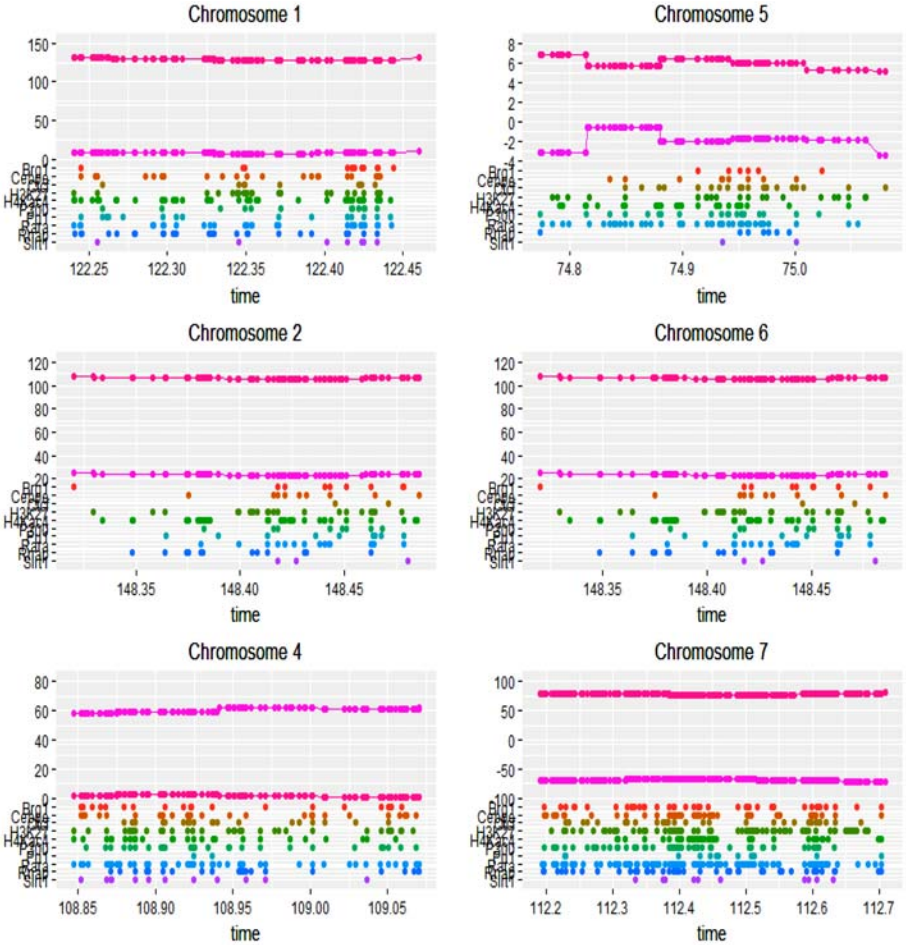

Figure 1. Magnified detail occurrences plot of TREs along the chromosomes with respect to time.

To assess whether the occurrences of the TFBS are Poisson distributed, we employ a graphical test known as empirical cumulative distribution plot and a one-sample Kolmogorov-Smirnov test. Figure 2 displays the empirical cumulative distribution plots of the transformed times (points) of the TFBS which are assumed to be uniformly distributed on the interval $[0, 1)$. The empirical cumulative distribution of $U(i)$ for the ten TFBS with 95% confidence limits for Kolmogorov distribution which assume that the $U(i)$

are uniformly distributed on $[0, 1)$ is displayed in Figure 2, where $U(i)$ is explained in the methods.

The empirical cumulative distribution plots examine whether the transformed point processes of the TREs are uniformly distributed on interval zero to one. When the cumulative distribution plots produce a straight diagonal line and the points are lying within the confidence bounds plotted, then we have a good fit to the data. A good fit to the data means that the transformed points are Poisson distributed with unit rate and implies Hawkes process.

It is observed from Figure 2 that RARA, RNAP, PU1, SIRT1 and H4KAC4 empirical cumulative plots lie within the 95% confidence bounds even though the curves of the plots do not exactly lie on the line $y = x$ on the euclidean plane. This suggests a uniform distribution of the transformed points of the $U(i)$ at 95% confidence level for these five TREs. However, the remaining five TREs empirical cumulative distribution plots show deviation from uniform distribution since the curves of the plots fall outside the 95% confidence bounds at the lower part of the graphs. Generally, there are some deviations observed, especially at the lower and middle portions in almost all of the empirical cumulative distribution plots with respect to the reference line, $y = x$. This can be attributed to too many small values as compared to large values of the $U(i)$. This implies that many of the points occur with a smaller inter-distances than what the model would predict if Poisson distribution is assumed for occurrences of the TREs. This would lead to smaller intensities being predicted in some parts of the genome. It is also probable that the shape of these curves and the deviances from the expected values indicate that the linear form of the intensities may not be the best possible choice. The empirical cumulative distribution plots of the uniformly transformed points have shown that the occurrences are not homogeneous Poisson distributed.

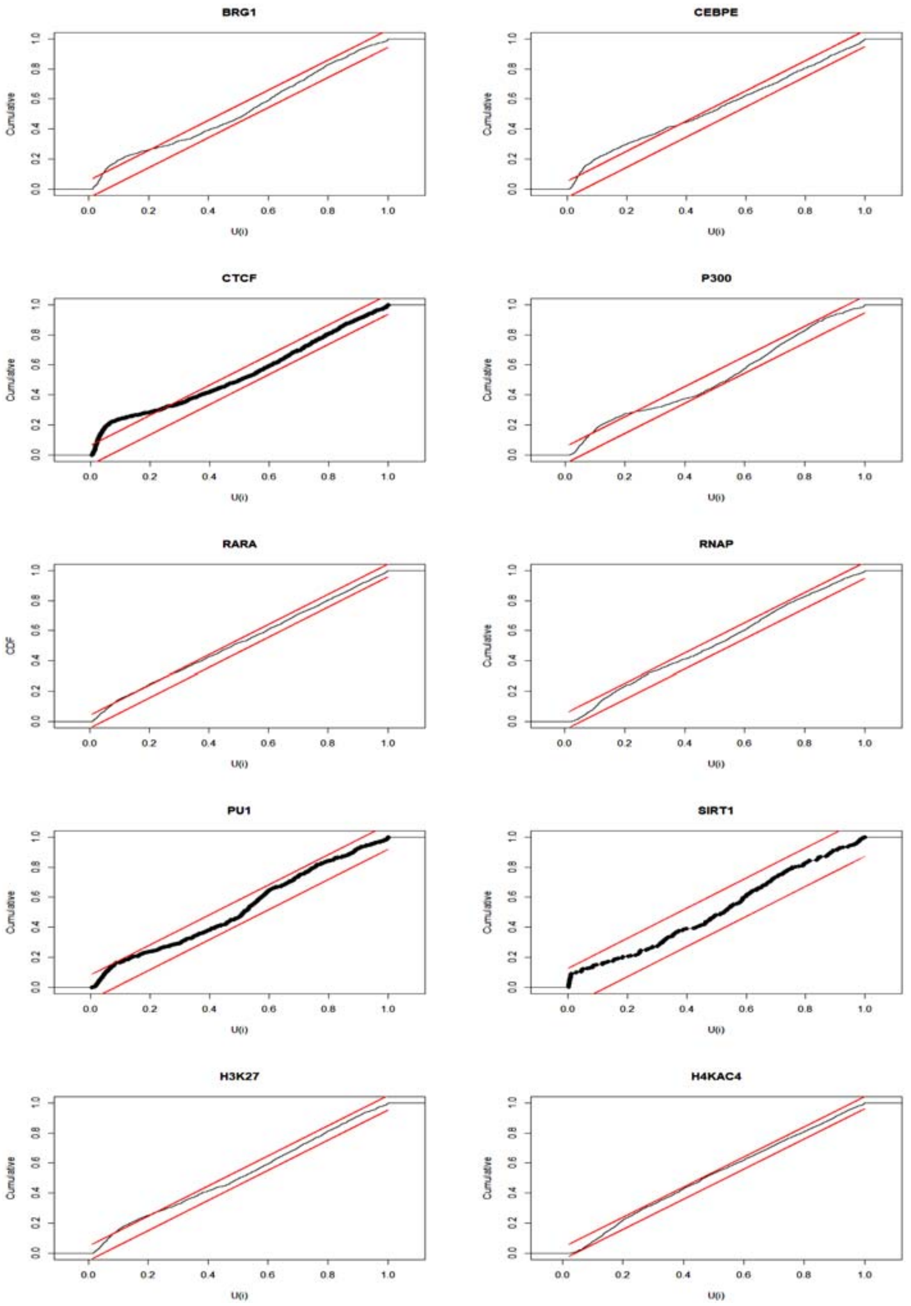

Figure 2. Empirical cumulative distribution plots of ten (10) TREs with 95% confidence bounds.

The quantitative goodness-of-fit test (Kolmogorov-Smirnov test) for the transformed points $\Lambda(t)$ and $U(i)$ of all the transcriptional regulatory elements is shown in Table 3. The transformed points also known as the compensator by the random time change theorem should form a Poisson process with unit rate whilst the cumulative distribution of the inter-arrival times (τ) is uniformly distributed on the interval $[0, 1)$ if Poisson assumption is satisfied.

It is generally observed that the transformed points of all the TREs are significantly different from the homogeneous Poisson process assumption at 5% level of significance (Table 3). This is as result of the *p*-values in the two Kolmogorov-Smirnov tests been less than $\alpha = 0.05,$ the significance level. However, it is observed that at 5% significance level, the inter-arrival times of only SIRT1 TRE is uniformly distributed on the interval $[0, 1).$ Whilst the remaining nine TREs show a significant departure from the uniform distribution and thus the TREs are not distributed as homogeneous Poisson process. This may be as a result of the deviations of the $U(i)$ at the extreme ends of the fits (Figure 2) and may be due to too many shorter and longer distances between the TFBS points.

It is observed from the tests (both graphical and quantitative) that most of the TREs have too many small inter-distances between successive transformed points which may be due to the clustering nature of the occurrences of these TREs along the genome sequence as shown in Figure 1. This implies that an ordinary Hawkes model may not be good at capturing the extreme values for the inter-distances and a log linear Hawkes model may be preferred. The study will therefore incorporate (in later section) histone modifications as covariates in fitting a multivariate Hawkes model to adjust the baseline intensities rates for this poor fit as suggested by Carstensen et al. [5].

Table 2. K.-S. test of homogeneous Poisson process for transformed times

TREs	Test statistic, $\Lambda(t)$	p-values	Test statistic, $U(i)$	p-values
BRG1	0.995	$2.2e^{-16}$	0.096	$9.8e^{-10}$
CEBPE	0.995	$2.2e^{-16}$	0.108	$1.8e^{-14}$
CTCF	0.995	$2.2e^{-16}$	0.151	$2.2e^{-16}$
P300	0.995	$2.2e^{-16}$	0.082	$9.5e^{-8}$
RARA	0.997	$2.2e^{-16}$	0.053	$1.4e^{-5}$
RNAP	0.994	$2.2e^{-16}$	0.041	0.0208
PU1	0.990	$2.2e^{-16}$	0.077	0.0023
SIRT1	0.979	$2.2e^{-16}$	0.082	0.0829
H3K27	0.996	$2.2e^{-16}$	0.064	$2.4e^{-6}$
H4KAC4	0.997	$2.2e^{-16}$	0.036	0.0043

We next investigate the independence among the transformed points of the TFBS. A graphical autocorrelation test is employed in the form of a scatter plot of the transformed points $U(i+1)$ against $U(i)$ to determine whether there is independence among $U(i)$. The scatter plots of $U(i+1)$ against $U(i)$ is displayed in Figure 3, showing the test for autocorrelation among the TFBS. It is generally observed that transformed points are evenly dispersed in almost all the plots. However, there exist some clustering of the transformed points at the lower parts of the scatter plots towards zero on the $U(i)$ (or horizontal) axis in BRG1, CEBPE, CTCF and H4KAC4. This shows some dependencies between the $U(i)$ at these parts of the plots for these TREs. The TREs P300, RARA, RNAP, PU1, SIRT1 and H3K27 are observed to exhibit independence between the inter-arrival times whilst the remaining TREs show incomplete independencies among $U(i)$. The incomplete independencies between the inter-arrival times in the four TREs may result in the baseline or the ground intensity not able to capture the active regions in the model.

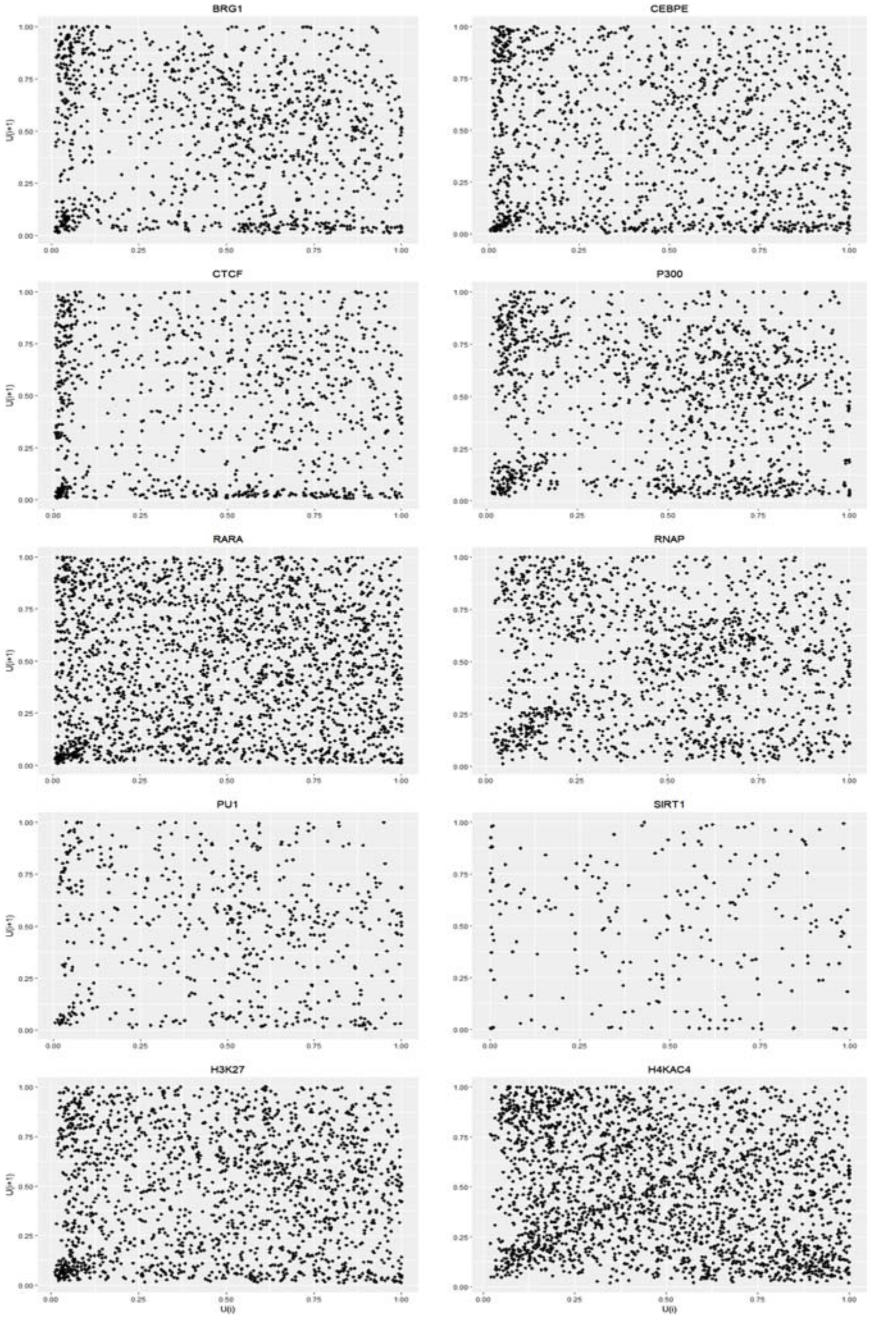

Figure 3. A graphical autocorrelation test of independence among transformed points, $U(i)$.

3.2. Multivariate modelling of transcriptional factor binding sites

In this section, we employ multivariate Hawkes model to determine the dependence effect of one transcription factor binding site on the occurrence of the other TFBS along a genome sequence. A generalised linear Hawkes model is fitted to the ENCODE pilot data through the use of cubic *B*-spline basis functions to determine the effect of one transcription factor binding site on the other TFBS. The $h(\cdot)$-function described in the methods and stated as equation (2.8), is modelled using a spline basis with six equidistant fixed knots, resulting into two estimates for each TFBS of the Hawkes generalised linear model. The knots are placed at equidistant points to capture the occurrences of the TFBS and also limit the range of dependence among the occurrences of TFBS to the maximum. Results from the preliminary analysis shows that the conditional intensities of the TREs possess some properties of Hawkes processes and therefore can be modeled with multivariate generalised Hawkes model to capture the interactions among the TFBS.

In this paper, we consider two kernels (Hawkes identity kernel and Hawkes log kernel) as link functions for modelling the cubic spline basis of the $h(\cdot)$-functions of equation (2.8) in other to model dependence effect of one TRE on the other TREs. The kernel density estimation is employed to estimate the $h(\cdot)$ function of the multivariate Hawkes model. It is a fundamental data smoothing method where inferences about the population are made, based on a finite sample data. Due to the unavailability of the closed form for the log-likelihood of the multivariate Hawkes model as described in the methods, the cubic *B*-spline functions are used to model the $h(\cdot)$ with the application of the kernel estimation. The bandwidth of the kernel estimation used is 0.01 with 0-150 support.

The coefficients of these models are displayed in Tables 3, 4 and Appendix A. The effects are estimated in a generalised multiple Hawkes model as coefficients using Hawkes identity kernel and Hawkes log kernel. The coefficients represent favoured distances for significant estimates and avoided distance for insignificant coefficients. The effects against the same

TRE in the model, estimates the self-excitation or inhabitation of the TRE if significant. By considering a log kernel process, it turns to obtain a more suitable model for extreme inter-distance of which the transformed points experienced poor fits (Figure 2).

First of all, before we examine the models in the tables, there is the need to explain the intensity effect of an observed TRE, k_1 given a TRE, k_2 at a distance s to a reference fixed knot and thus relevance in this study. The intensity effect of TRE k_1 for observing TRE, k_2 at a distance s to a fixed knot is given as a baseline intensity times $h_{k_1, k_2}(s)$. If $h_{k_1, k_2}(s) = 1$, then the occurrences of TRE k_2 is not associated to the occurrences of TRE k_1 at the distance s to the fixed knot. Figure 4 shows sixty-four multivariate plots of the spline basis functions of the $\log h_{k_1, k_2}(\cdot)$ with 95% confidence intervals. The $h(\cdot)$ functions are modelled with a cubic basis spline functions with a reference to the fixed knots. However, Figure 4 shows that some of the spline bases are above and below the zero horizontal line. A spline basis function below the zero horizontal line means that the significant number events of the observed TRE is below reference fixed knot and a positive spline basis function implies a significant number of events above the reference knot. Therefore, a negative spline function results in negative effect between the two TREs with the reference fixed knot at a distance, s. Since the study data was transformed before using it for analysis, it is therefore required that the transformation is reverse before making any judgement on the absolute interaction effect between two TREs at an observed distance. The spline basis graphs of BRG1 TRE on the other eight TREs including itself are shown in first row of Figure 4. It is observed that BRG1 against the others TREs produces negative spline basis functions in all the plots except with PU1 that record both negative and positive spline basis function over the support interval considered. It is generally observed that the spline functions plots under each TRE show similar characteristics of the spline functions lying below the zero horizontal line.

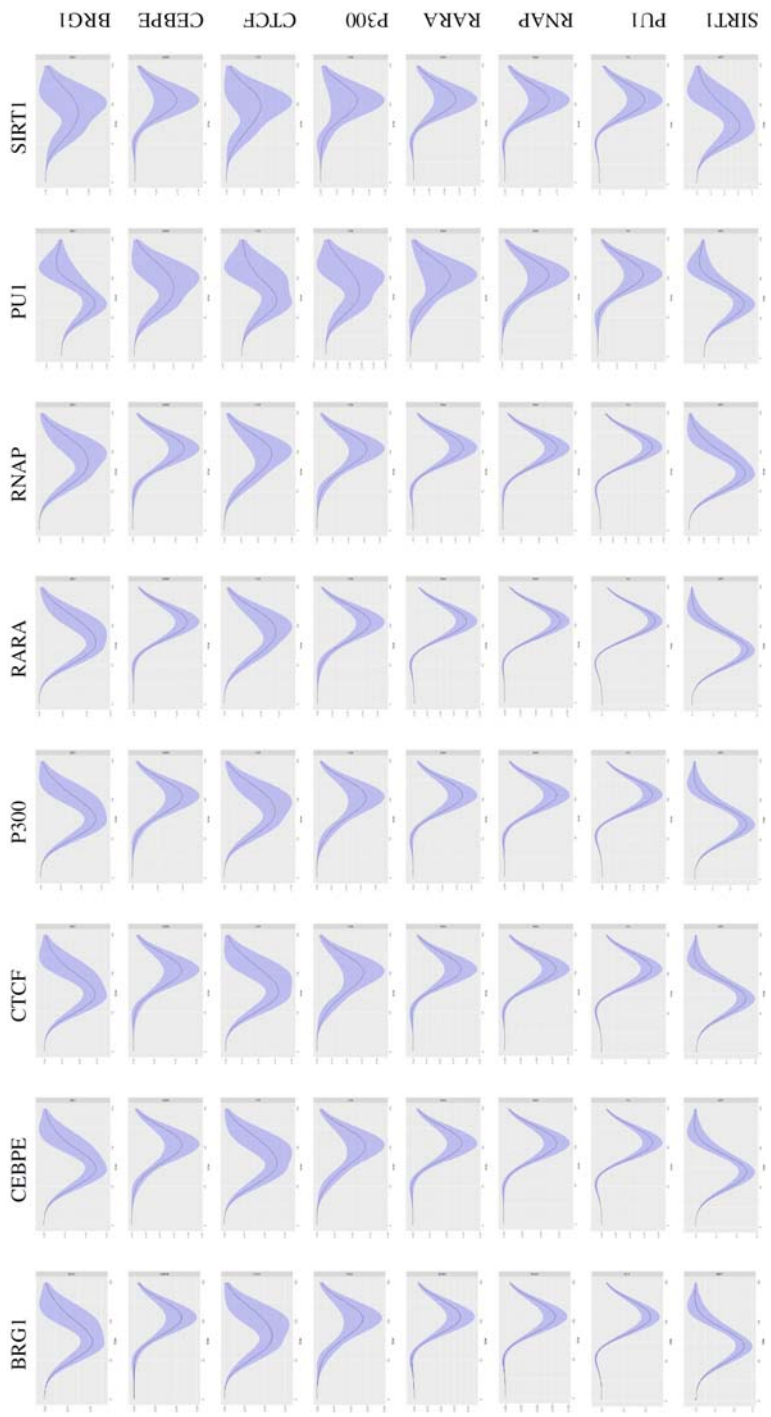

Figure 4. A graphical plot of the multivariate Hawkes model for the spline basis *h*-functions of the TREs.

Table 3 presents the coefficients of the generalised Hawkes model for BRG1 TRE as response variable with the other TREs including the two histone modification (H3K27 and H4KAC4) elements as explanatory variables. The histone modification elements are used as covariates in the model to increase the effect of the baseline intensity for the TFBS in the model in order to capture the extreme inter-distances among the TREs observed in the previous section. Table A6 shows two sets of models namely; Hawkes identity kernel and Hawkes log kernel models. It is observed that the Hawkes log kernel model provides a better fit to the data, since its AIC (2310.80) is smaller. It is observed that there is significant negative effect (–0.98) of BRG1 TRE on itself at the first estimate and also a significant positive effect (8.54) for the second estimate. This means that there is a significant interaction within the transcription factor binding sites of BRG1 itself at the breaks or fixed knots over the support interval. It can be deduced that the equidistant fixed knots have been able to capture the interactions among the TFBS of BRG1 against itself. The sign of the coefficients indicate whether the $h(\cdot)$ function is maximum or minimum between two successive fixed knots. A greater magnitude of the coefficient, results in a strong effect or dependency between the TREs over the interval. Table A6 shows that BRG1 TRE model records negative coefficients with CEBPE, CTCF, P300, RARA, RNAP and SIRT1 at the second estimate of each TREs and positive coefficients with CTCF (first estimate), PU1 and SIRT1 (second estimate) TREs.

Table 4 presents CEBPE TRE Hawkes model with the other TREs as explanatory variables. Similarly, the table shows the coefficients of the Hawkes process model for the CEBPE TFBS with the other TREs including the two histone modification elements. Again, it is observed that the Hawkes log kernel model provides a better fit to the data, even though the identity kernel model records significant coefficients for all the TREs as compared to the log kernel model, which records three insignificant effects or interactions. It is however observed that CEBPE recorded significance effect with itself at both estimates with the $h(\cdot)$ function being minimum at both intervals between the fixed knots that the two coefficients are estimated. It is observed that almost all the coefficients of Table 4 are negative implying that the cubic *B*-splines of $h(\cdot)$, are minimum in between the fixed knots over the support interval.

Table 3. Multivariate Hawkes model coefficients of BRG1 TRE on other TREs

Model variables	Hawkes identity kernel		Hawkes log kernel	
	Estimate	*p*-values	Estimate	*p*-values
Intercept	0.30	0.00	-1.33	0.00
bSpline(BRG1)1	-0.98	0.00	-1.42	0.00
bSpline(BRG1)2	8.54	0.00	9.71	0.00
bSpline(CEBPE)1	-0.08	0.02	-0.27	0.05
bSpline(CEBPE)2	-5.84	0.00	-7.39	0.00
bSpline(CTCF)1	1.56	0.00	2.82	0.00
bSpline(CTCF)2	-4.53	0.00	-8.69	0.00
bSpline(P300)1	-0.13	0.00	-1.08	0.00
bSpline(P300)2	-1.64	0.00	-0.27	0.71
bSpline(RARA)1	-0.49	0.00	-0.58	0.00
bSpline(RARA)2	-3.85	0.00	-2.44	0.01
bSpline(RNAP)1	-0.03	0.26	-0.24	0.00
bSpline(RNAP)2	-5.51	0.00	-3.39	0.00
bSpline(PU1)1	0.27	0.00	1.90	0.00
bSpline(PU1)2	0.89	0.03	-6.09	0.00
bSpline(SIRT1)1	0.55	0.00	2.45	0.00
bSpline(SIRT1)2	-2.69	0.00	-10.36	0.00
bSpline(H3K27)1	0.06	0.00	-0.06	0.25
bSpline(H3K27)2	5.11	0.00	7.17	0.00
bSpline(H4KAC4)1	0.26	0.00	0.21	0.00
bSpline(H4KAC4)2	3.40	0.00	3.08	0.00
AIC	2821.20		2310.80	

Furthermore, Tables A1-A6 of Appendix A show the coefficients of the other multivariate Hawkes linear models of both the log kernel and identity kernel of each TREs and the corresponding explanatory variables (TREs). Generally, it is observed from Tables A1-A6 that CTCF and RARA TRE models depict approximately the same pattern of negative and positive coefficients for the explanatory TREs. The AICs shows that the Hawkes log kernel model produces a better fit to the data in all the situations. It is observed that there are significant effects or interaction among almost all the TREs at both the two estimates each.

Table 4. Multivariate Hawkes model coefficients of CEBPE TRE on other TREs

Model variables	Hawkes identity kernel		Hawkes log kernel	
	Estimate	*p*-values	Estimate	*p*-values
Intercept	0.32	0.00	-1.08	0.00
bSpline(BRG1)1	-1.17	0.00	-1.34	0.00
bSpline(BRG1)2	11.51	0.00	10.58	0.00
bSpline(CEBPE)1	-0.24	0.00	-0.46	0.00
bSpline(CEBPE)2	-4.42	0.00	-6.65	0.00
bSpline(CTCF)1	1.74	0.00	2.29	0.00
bSpline(CTCF)2	-2.90	0.00	-6.24	0.00
bSpline(P300)1	-0.12	0.00	-0.76	0.00
bSpline(P300)2	-4.63	0.00	-3.03	0.00
bSpline(RARA)1	-0.42	0.00	-0.27	0.00
bSpline(RARA)2	-3.61	0.00	-2.80	0.00
bSpline(RNAP)1	-0.08	0.01	-0.08	0.27
bSpline(RNAP)2	-6.42	0.00	-5.07	0.00
bSpline(PU1)1	0.29	0.00	1.73	0.00
bSpline(PU1)2	2.98	0.00	-5.37	0.00
bSpline(SIRT1)1	0.92	0.00	2.12	0.00
bSpline(SIRT1)2	-16.55	0.00	-10.45	0.00
bSpline(H3K27)1	0.09	0.00	-0.06	0.20
bSpline(H3K27)2	4.73	0.00	6.98	0.00
bSpline(H4KAC4)1	0.34	0.00	0.06	0.10
bSpline(H4KAC4)2	2.71	0.00	3.73	0.00
AIC	3422.20		2642.20	

4. Conclusions

The paper obtains model for the conditional intensity of TFBS using the ENCODE pilot project data of 1% human genome. Eight TREs are identified along several chromosomes in addition to two histone modification elements in retinoic acid simulated human HL-60 cells gathered after 0, 2, 8, and 32 hours. It is revealed that the TREs occurrences along the DNA sequence can be modelled using a generalised linear Hawkes model. Additionally, the paper reveals that the multivariate Hawkes process is able to capture and explain the interactions between a set of transcription factor binding sites occurring along a DNA sequence. It is also identified that the log Hawkes kernel model is the most suitable for modelling the conditional intensity of the multivariate Hawkes linear model. The model thus adequately captures extreme inter-distances that are initially found to characterise the transform point process. With massive genomes data being sequenced and more experimental transcription regulation data becoming available, the paper has employed advance statistical techniques for analysing these data to serve as a confirmation of the experimental analysis. The study recommends that mathematical and statistical models such as Hawkes processes and Bayesian networks should be employed in providing the foundation for experimental determination of interactions between transcriptional regulatory elements or factors in DNA sequences.

References

[1] A. R. Bansal, V. P. Dimri and K. K. Babu, Epidemic type aftershock sequence (ETAS) modeling of northeastern Himalayan seismicity, J. Seismology 17(2) (2013), 255-264. DOI: https://doi.org/10.1007/s10950-012-9314-7.

[2] P. Bremaud and L. Massoulie, Stability of nonlinear Hawkes processes, The Annals of Probability 24(3) (1996), 1563-1588.

[3] M. J. Buck and J. D. Lieb, ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments, J. Genomics 83(3) (2004), 349-360.

[4] L. Carstensen, Hawkes processes and combinational transcriptional regulation, Ph.D. Thesis (unpublished), University of Copenhagen, 2010.

[5] L. Carstensen, A. G. Sandelin, O. Winther and N. R. Hansen, Multivariate Hawkes process models of the occurrence of regulatory elements, BMC Bioinformatics 11 (2010), 456. doi:10.1186/1471-2105-11-456.

[6] D. J. Daley and D. Vere-Jones, An Introduction to the Theory of Point Processes I, Springer, New York, 2003.

[7] P. Embrechts, J. T. Liniger and L. Lu, Multivariate Hawkes processes: an application to financial data, J. Applied Probability 48 (2011), 367-378.

[8] P. J. Green and B. W. Silverman, Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach, Chapman and Hall, 1994.

[9] G. Gusto and S. Schbath, FADO: a statistical method to detect favored or avoided distances between occurrences of motifs using the Hawkes' model, Statist. Appl. Genetics and Molecular Biology 4(1) (2005), 1-28. DOI:10.2202/1544-6115.1119.

[10] S. Hannenhalli, Eukaryotic transcription factor binding sites-modeling and integrative search methods, Bioinformatics 24(11) (2008), 25-31.
DOI: 10.1093/bioinformatics/btn198.

[11] N. R. Hansen, P. Reynaud-Bouret and V. Rivoirard, Lasso and probabilistic inequalities for multivariate point processes, Bernoulli 21(1) (2015), 83-143.

[12] N. R. Hansen, Point Process Statistics (ppstat) R-package for multivariate point processes on the line, 2013. http://www.math.ku.dk/richard/ppstat.

[13] D. Harte, PtProcess: an R package for modelling marked point processes indexed by time, J. Statistical Software 35(8) (2010), 1-32.

[14] N. Hautsch, Econometrics of Financial High-frequency Data, Springer Science & Business Media, 2011.

[15] A. G. Hawkes, Spectra of some self-exciting and mutually exciting point processes, Biometrika 58(1) (1971), 83-90.

[16] W. Krivan and W. W. Wasserman, A predictive model for regulatory sequences directing liver-specific transcription, Genome Research 11 (2001), 1559-1566.

[17] P. J. Laub, T. Taimre and P. K. Pollett, Hawkes processes, Statistical Finance Applications; 1-28, 2015. arXiv:1507.02822.

[18] B. Lemon and R. Tjian, Orchestrated response: a symphony of transcription factors for gene control, Genes Development 14 (2000), 2551-2569.

[19] G. A. Maston, S. K. Evans and M. R. Green, Transcriptional regulatory elements in the human genome, Annual Review of Genomics and Human Genetics 7 (2006), 29-59.

[20] Y. Ogata, Seismicity analysis through point-process modeling: a review. Pure and Applied Geophysics 155(2-4) (1999), 471-507.

[21] Y. Ogata, Detection of precursory relative quiescence before great earthquakes through a statistical model, J. Geophysics Research 97 (1992), 19845-19871.
doi:10.1029/92JB00708.

[22] P. J. Park, ChIP-seq: advantages and challenges of a maturing technology, Nature Reviews Genetics 10(10) (2009), 669-680.

[23] G. Reinert, S. Schbath and M. S. Waterman, Probabilistic and statistical properties of words: an overview, J. Computational Biology 7(2) (2000), 1-46.

[24] P. Reynaud-Bouret and S. Schbath, Adaptive estimation for Hawkes processes; application to genome analysis, Ann. Statist. 38(5) (2010), 2781-2822.

[25] S. Robin and J. J. Daudin, Exact distribution of word occurrences in a random sequence of letters, J. Applied Probability 36(1) (1999), 179-193.

[26] S. J. Sanders and C. E. Mason, Genomics, Circuits, and Pathways in Clinical Neuropsychiatry: The Newly Emerging View of the Genome, Academic Press (Elsevier), London, 2016.

[27] The ENCODE Project Consortium, Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project, Nature 447 (2007), 799-816.

[28] The ENCODE Project Consortium, The ENCODE (ENCyclopedia of DNA Elements) project, J. Science 306 (2004), 636-640.

[29] F. Schoenberg, Transforming spatial point processes into Poisson processes, Stochastic Processes and their Applications 81 (1999), 155-164.

[30] V. T. Stefanov, The intersite distances between pattern occurrences in strings generated by general discrete- and continuous-time models: an algorithmic approach, J. Applied Probability 40 (2003), 881-892.

[31] W. W. Wasserman and J. W. Fickett, Identification of regulatory regions which confer muscle-specific gene expression, J. Molecular Biol. 278 (1998), 167-181.

[32] W. W. Wasserman and A. Sandelin, Applied bioinformatics for the identification of regulatory elements, Nature Reviews Genetics 5(4) (2004), 276-287.

[33] T. W. Whitfield, J. Wang, P. J. Collins, E. C. Partridge, S. F. Aldred, N. Trinklein, R. M. Myers and Z. Weng, Functional analysis of transcription factor binding sites in human promoters, Genome Biology 13 (2012), 1-16.

[34] Z. D. Zhang, A. Paccanaro, Y. Fu, S. Weissman, Z. Weng, J. Chang, M. Snyder and M. B. Gerstein, Statistical analysis of the genomic distribution and correlation of regulatory elements in the ENCODE regions, Genome Research 17 (2007), 787-797.

Appendix A

Multivariate Hawkes models coefficients TREs

Table A1. CTCF TRE model

Model variables	Hawkes identity kernel		Hawkes log kernel	
	Estimate	*p*-values	Estimate	*p*-values
Intercept	0.3	0.00	-1.36	0.00
bSpline(BRG1)1	-0.51	0.00	-1.22	0.00
bSpline(BRG1)2	3.54	0.00	6.98	0.00
bSpline(CEBPE)1	-0.04	0.13	-0.03	0.87
bSpline(CEBPE)2	-2.74	0.00	-5.93	0.00
bSpline(CTCF)1	0.62	0.00	2.26	0.00
bSpline(CTCF)2	-0.11	0.68	-4.19	0.00
bSpline(P300)1	-0.21	0.00	-1.22	0.00
bSpline(P300)2	-0.92	0.00	-0.21	0.76
bSpline(RARA)1	-0.17	0.00	-0.3	0.01
bSpline(RARA)2	-3.17	0.00	-3.37	0.00
bSpline(RNAP)1	-0.09	0.00	-0.44	0.00
bSpline(RNAP)2	-2.59	0.00	-2.19	0.00
bSpline(PU1)1	-0.04	0.37	1.25	0.00
bSpline(PU1)2	2.13	0.00	-1.72	0.16
bSpline(SIRT1)1	0.43	0.00	1.4	0.00
bSpline(SIRT1)2	-2.45	0.00	-11.87	0.00
bSpline(H3K27)1	0.11	0.00	-0.09	0.17
bSpline(H3K27)2	2.74	0.00	5.93	0.00
bSpline(H4KAC4)1	0.22	0.00	0.35	0.00
bSpline(H4KAC4)2	1.54	0.00	2.01	0.00
AIC	2773		2411.9	

Table A2. P300 TRE model

Model variables	Hawkes identity kernel		Hawkes log kernel	
	Estimate	*p*-values	Estimate	*p*-values
Intercept	0.36	0.00	-1.15	0.00
bSpline(BRG1)1	-0.86	0.00	-1.16	0.00
bSpline(BRG1)2	7.55	0.00	11.23	0.00
bSpline(CEBPE)1	-0.14	0.00	-0.62	0.00
bSpline(CEBPE)2	-5.07	0.00	-8.27	0.00
bSpline(CTCF)1	1.46	0.00	3.06	0.00
bSpline(CTCF)2	-2.92	0.00	-8.96	0.00
bSpline(P300)1	-0.21	0.00	-1.64	0.00

bSpline(P300)2	-1.83	0.00	-2.02	0.01
bSpline(RARA)1	-0.40	0.00	-0.27	0.01
bSpline(RARA)2	-4.37	0.00	-2.66	0.00
bSpline(RNAP)1	-0.10	0.00	-0.23	0.00
bSpline(RNAP)2	-5.00	0.00	-4.55	0.00
bSpline(PU1)1	0.14	0.00	2.09	0.00
bSpline(PU1)2	1.82	0.00	-8.59	0.00
bSpline(SIRT1)1	0.49	0.00	2.55	0.00
bSpline(SIRT1)2	-2.93	0.00	-9.06	0.00
bSpline(H3K27)1	0.10	0.00	-0.08	0.17
bSpline(H3K27)2	4.89	0.00	8.36	0.00
bSpline(H4KAC4)1	0.29	0.00	0.12	0.00
bSpline(H4KAC4)2	3.03	0.00	4.22	0.00
AIC	2999.70		2474.70	

Table A3. RARA TRE model

Model	Hawkes identity kernel		Hawkes log kernel	
variables	Estimate	*p*-values	Estimate	*p*-values
Intercept	0.53	0.00	-0.68	0.00
bSpline(BRG1)1	-1.55	0.00	-1.33	0.00
bSpline(BRG1)2	13.59	0.00	9.53	0.00
bSpline(CEBPE)1	-0.30	0.00	-0.36	0.00
bSpline(CEBPE)2	-8.67	0.00	-6.53	0.00
bSpline(CTCF)1	2.27	0.00	2.47	0.00
bSpline(CTCF)2	-2.16	0.01	-7.71	0.00
bSpline(P300)1	-0.19	0.00	-0.81	0.00
bSpline(P300)2	-4.20	0.00	-1.15	0.05
bSpline(RARA)1	-0.48	0.00	-0.35	0.00
bSpline(RARA)2	-10.49	0.00	-2.63	0.00
bSpline(RNAP)1	-0.14	0.00	-0.27	0.00
bSpline(RNAP)2	-10.62	0.00	-3.85	0.00
bSpline(PU1)1	0.08	0.32	1.87	0.00
bSpline(PU1)2	5.79	0.00	-5.89	0.00
bSpline(SIRT1)1	0.87	0.00	1.93	0.00
bSpline(SIRT1)2	-4.42	0.00	-9.57	0.00
bSpline(H3K27)1	0.11	0.00	-0.10	0.01
bSpline(H3K27)2	9.37	0.00	6.97	0.00
bSpline(H4KAC4)1	0.51	0.00	0.14	0.00
bSpline(H4KAC4)2	6.03	0.00	3.13	0.00
AIC	2466.70		1536.20	

Table A4. RNAP TRE model

Model variables	Hawkes identity kernel		Hawkes log kernel	
	Estimate	*p*-values	Estimate	*p*-values
Intercept	0.35	0.00	-0.97	0.00
bSpline(BRG1)1	-1.16	0.00	-1.37	0.00
bSpline(BRG1)2	8.06	0.00	8.93	0.00
bSpline(CEBPE)1	0.27	0.00	0.13	0.28
bSpline(CEBPE)2	-5.08	0.00	-5.51	0.00
bSpline(CTCF)1	0.71	0.00	1.51	0.00
bSpline(CTCF)2	3.87	0.00	-2.47	0.05
bSpline(P300)1	-0.13	0.00	-0.16	0.30
bSpline(P300)2	-4.86	0.00	-3.66	0.00
bSpline(RARA)1	-0.18	0.00	-0.27	0.00
bSpline(RARA)2	-8.79	0.00	-2.72	0.00
bSpline(RNAP)1	0.26	0.00	-0.02	0.75
bSpline(RNAP)2	-8.41	0.00	-6.12	0.00
bSpline(PU1)1	-0.49	0.00	0.85	0.00
bSpline(PU1)2	4.59	0.00	0.01	1.00
bSpline(SIRT1)1	0.78	0.00	0.96	0.00
bSpline(SIRT1)2	-2.51	0.00	-9.78	0.00
bSpline(H3K27)1	0.23	0.00	0.00	1.00
bSpline(H3K27)2	6.37	0.00	5.10	0.00
bSpline(H4KAC4)1	0.13	0.00	-0.01	0.71
bSpline(H4KAC4)2	4.78	0.00	3.31	0.00
AIC	4815.90		3015.80	

Table A5. PU1 TRE model

Model variables	Hawkes identity Kernel		Hawkes log Kernel	
	Estimate	*p*-values	Estimate	*p*-values
Intercept	0.13	0.00	-1.89	0.00
bSpline(BRG1)1	-0.46	0.00	-1.50	0.00
bSpline(BRG1)2	2.82	0.00	9.11	0.00
bSpline(CEBPE)1	0.01	0.59	0.02	0.91
bSpline(CEBPE)2	-1.46	0.00	-6.52	0.00
bSpline(CTCF)1	0.45	0.00	2.02	0.00
bSpline(CTCF)2	0.81	0.04	-5.51	0.00
bSpline(P300)1	-0.19	0.00	-0.36	0.11
bSpline(P300)2	-1.45	0.00	-1.99	0.05
bSpline(RARA)1	-0.09	0.00	-0.63	0.00
bSpline(RARA)2	-2.71	0.00	-1.59	0.16
bSpline(RNAP)1	0.06	0.00	0.19	0.12

bSpline(RNAP)2	-2.38	0.00	-4.91	0.00
bSpline(PU1)1	-0.16	0.00	0.63	0.02
bSpline(PU1)2	0.78	0.01	-2.56	0.24
bSpline(SIRT1)1	0.25	0.00	1.59	0.00
bSpline(SIRT1)2	-0.86	0.00	-8.87	0.00
bSpline(H3K27)1	0.15	0.00	0.15	0.06
bSpline(H3K27)2	1.94	0.00	5.20	0.00
bSpline(H4KAC4)1	0.12	0.00	0.11	0.06
bSpline(H4KAC4)2	1.38	0.00	3.20	0.00
AIC	3015.10		2378.40	

Table A6. SIRT1 TRE model

Model variables	Hawkes identity kernel		Hawkes log kernel	
	Estimate	*p*-values	Estimate	*p*-values
Intercept	0.07	0.00	-2.81	0.00
bSpline(BRG1)1	-0.18	0.00	-1.27	0.00
bSpline(BRG1)2	1.16	0.00	6.87	0.01
bSpline(CEBPE)1	-0.03	0.06	-0.19	0.53
bSpline(CEBPE)2	-0.70	0.00	-6.37	0.00
bSpline(CTCF)1	0.24	0.00	2.32	0.00
bSpline(CTCF)2	-0.60	0.01	-9.47	0.00
bSpline(P300)1	-0.04	0.02	-0.86	0.02
bSpline(P300)2	-0.05	0.72	2.94	0.13
bSpline(RARA)1	-0.05	0.01	-0.41	0.06
bSpline(RARA)2	-0.63	0.00	-2.96	0.14
bSpline(RNAP)1	-0.02	0.14	-0.29	0.08
bSpline(RNAP)2	-0.47	0.00	-0.88	0.54
bSpline(PU1)1	0.03	0.17	1.51	0.00
bSpline(PU1)2	-0.06	0.80	-9.62	0.02
bSpline(SIRT1)1	0.16	0.00	1.53	0.03
bSpline(SIRT1)2	-2.28	0.00	-11.31	0.04
bSpline(H3K27)1	0.01	0.08	-0.04	0.72
bSpline(H3K27)2	0.83	0.00	7.17	0.00
bSpline(H4KAC4)1	0.06	0.00	0.24	0.01
bSpline(H4KAC4)2	0.27	0.00	2.22	0.02
AIC	1417.30		1276.90	

G. Kallah-Dagadu: gkallah-dagadu@ug.edu.gh