JP Journal of Biostatistics
© 2018 Pushpa Publishing House, Allahabad, India
http://www.pphmj.com
http://dx.doi.org/10.17654/JB015010039
Volume 15, Number 1, 2018, Pages 39-64 **ISSN: 0973-5143**

PROBABILISTIC GRAPHICAL MODELLING OF CAUSAL EFFECTS AMONG THE OCCURRENCES OF TRANSCRIPTION FACTORS IN DNA SEQUENCE

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

Genome mapping of transcription factor targeted by ChIP jointly with microarrays or sequencing procedures is a powerful instrument for laying a foundation for understanding transcriptional regulatory networks. Hence the need for computational methods that can form the basis of experimental verification of these networks. We employ a probabilistic graphical model of the form of linear Gaussian Bayesian network to model causal effects between transcriptional factors (TFs) in two genome datasets. The *bnlearn R* statistical package is used for learning the network structure of the ENCODE pilot data and Mouse Embryonic Stem Cell data. Our results show that the Bayesian network efficiently model the causal effects between TFs, handle uncertainty with respect to probability theory and establish indirect with direct causation. Finally, an integrated Bayesian network model

Received: March 1, 2018; Accepted: April 21, 2018

Keywords and phrases: probabilistic graphical model, Bayesian network, causal effects, transcriptional factors.

[*]Corresponding author

which can predict the TFs and identify experimentally verifiable relationships as well as missed relationships between TFs computationally is fitted to the genome data.

1. Introduction

The study of genetics and molecular biology has become increasingly accessible and affordable due to the microarray technology and gene regulatory networks from temporal gene expression data which has received growing interest. Every organism's hereditary information is contained in the DNA and a comprehensive analysis of an organism's genetic information has led to a satisfactory understanding of some phenotypic characteristics such as diseases (Bremer [4]). The unearthing of the molecular structure of deoxyribonucleic acid (DNA) by (Watson and Crick [31]), has progressively uncovered the information storage, extraction and conversion into proteins in the DNA. This has explained how certain variations in cells lead to abnormalities of individuals.

Genomes form thousands of protein-coding and non-coding RNA genes, most of which are differentially expressed at various locations and times throughout development, or in response to environmental indications (Martinez and Walhout [21]). Differential gene expression is carried out through complex regulatory networks that are controlled in part by two types of trans-regulators: transcription factors (TFs) and microRNAs (miRNAs). TFs bind to cis-regulatory DNA elements that are often found in or near their target genes, while miRNAs hybridize to cis-regulatory RNA elements generally located in the untranslated region of their target miRNAs (Filipowicz et al. [8]).

Understanding gene regulation process is of great interest in science, since a massive high-throughput data is being generated due to modern techniques of microarray technology such as ChIP-chip and ChIP-seq. In gene regulatory network the set of genes interact with one another through other genes, transcription factors, and protein products. The interactions between these elements contribute to the regulation of gene transcription and translation, and the structure of gene regulatory networks play a vital role in cell behaviour and structure (Rau et al. [26]).

The aim of this paper is to determine the complex interactions among genes and transcription factors along a DNA sequence or genome sequence through the use of Bayesian networks. The concept of gene networks is fundamental in system biology and we view networks as encompassing of nodes (genes or transcription factors) and the links (chemical reactions) between them. The networks describe the idea of stability and interconnections of molecular reactions and the challenge is to give these a statistical interpretation (Lotsi and Wit [20]).

System biology needs a flexible statistical method which can computationally and efficiently infer the complexity, the dependence structure of the network topology and the functional relationship between the genes or transcription factors. Several literature in bioinformatics have considered static networks (Maucher et al. [22]; Friedman et al. [11, 12]) and signal pathways of genes and transcription factors (Chen et al. [6]). Also, a lot of literature exist for dynamic networks modelling of genes or transcription factors by state space models (Lotsi and Wit [20]; Rau et al. [26]; Beal et al. [2]; Fang-Xiang et al. [7]).

We present a probabilistic graphical model which is based on Bayesian networks (Koller and Friedman [18]), and considers (conditional) probabilistic relationships in a set of random variables. Bayesian networks are useful tools for probabilistic inference among set of variables modelled using directed acyclic graph (DAG). The rest of the paper is organised as follows: Section 2 presents methods and materials of the study, Section 3 also presents analysis and results of the two datasets employed and Section 4 presents the conclusion of the study.

2. Methods and Materials

2.1. Bayesian networks

Bayesian networks are probabilistic graphical models that can be represented as set of causality relationships in a set of random variables (Koller and Friedman [18]). Bayesian networks have become a widespread technique used for inference of gene regulatory networks, due to their flexibility and intuitive interpretation (Lotsi and Wit [20]; Young et al. [33]; Rau et al. [26]). A Bayesian network can be described as, $\mathcal{B} = (\mathcal{X}, \mathcal{G}, \Theta)$, where $\mathcal{X}$ is the set of random variables $\{X_1, X_2, X_3, ..., X_p\}$, $\mathcal{G}$ is a directed acyclic graph that forms the structure of the network, (i.e. the graph that represents the conditional independences between the variables) and Θ is the set of parameters that specify the conditional probability distributions of the variables (Franzin et al. [9]; Koller and Friedman [18]). The graph $\mathcal{G}$ stands for conditional independence assumptions that allow the joint distribution to be decomposed and economising on the number of parameters. The graph $\mathcal{G}$ encodes the Markov assumption; each variable is independent of its non-descendants, given its parents in $\mathcal{G}$. Applying the chain rule of probabilities and properties of conditional independence, any joint distribution that satisfies the Markov assumption, can be decomposed into the product form as

$$\begin{aligned} P(X_1, X_2, ..., X_p) &= \prod_{i=1}^{p} P(X_i \,|\, Pa(X_i)) \\ &= P(X_i \,|\, X_1, X_2, ..., X_{i-1}), \end{aligned} \tag{2.1}$$

where $Pa(X_i)$ is the set of parents of X_i in $\mathcal{G}$. The individual factors $P(X_i \,|\, Pa(X_i))$ are the conditional probability distributions or local probabilistic models for each variable X_i (Koller and Friedman [18]) and are denoted as Θ. Several representations can be employed when specifying the conditional probability distributions of $\mathcal{G}$. We consider two of the most commonly used representations and the choice of representation depending on the type of variables we are dealing with (i.e. discrete or continuous).

Local probabilistic model for discrete-valued random variables of a Bayesian network are represented as *tabular conditional probability* *distributions*, where $P(X \mid Pa(X))$ is encoded as a table that contains entry for each joint assignment to X and $Pa(X)$, Conditional probability distributions table contains nonnegative values and that for each value $Pa(X)$, we have

$$\sum_{x \in Val(X)} P(X \mid Pa(X)) = 1. \tag{2.2}$$

A conditional probability distribution (CPD) needs to specify a conditional probability $P(X \mid Pa(X))$ for every assignment of values $Pa(X)$ and X, but it does not have to do so by listing each such value explicitly. We should view CPDs not as tables listing all the conditional probabilities, but rather as functions that given $Pa(X)$ and X, return $P(X \mid Pa(X))$. This implicit representation suffices to specify a well-defined joint distribution as a Bayesian network to avoid problems that might arise with the discrete-valued random variables (see Koller and Friedman [18]).

Most variables are best modelled as values in some continuous space and a tabular representation of the CPDs may not be possible. One common solution is to circumvent the entire issue by discretising all continuous variables. When continuous variables are discretised, we lose much of the structure that characterise the Bayesian network. There are many possible models one could use to model the continuous variables and the most commonly used parametric form for continuous density functions is the Gaussian distribution (Koller and Friedman [18]). Suppose *Y* is a continuous random variable with continuous parents nodes, $X_1, X_2, \ldots, X_p$. The variable *Y* has a linear Gaussian model if there are parameters $\alpha, \beta_1, \beta_2,$ $\ldots, \beta_p$ and σ^2 such that

$$(Y \mid \mathbf{X}) \sim \mathbf{N}(\alpha + \beta^T \mathbf{X}, \sigma^2). \tag{2.3}$$

This can be represented as $Y = \alpha + \sum_{i=1}^{p} \beta_i X_i + \varepsilon,$ a linear function of the variables $X_1, X_2, \ldots, X_p$, with the addition of the Gaussian noise (ε) of mean 0 and variance σ^2. This simple model captures many interesting dependencies. However, there are certain sides of the situation such as interaction (e.g. the variance of the child variable Y cannot depend on the actual values of the parents) might not be captured. The linear Gaussian model is a very natural one, which is a useful approximation in many practical applications. Bayesian networks based on the linear Gaussian models provide us with an alternative representation for multivariate Gaussian distributions, one that directly reveals more of the underlying structure. Situations where dependencies occur on a continuous variable with continuous and discrete parents or a discrete variable with continuous and discrete parents lead to Hybrid Bayesian network. Suppose X is a continuous variable, and $U = (U_1, U_2, \ldots, U_k)$ denotes its discrete parents and $Y = (Y_1, Y_2, \ldots, Y_p)$ denotes its continuous parents. Then X has a conditional linear Gaussian (CLG) conditional probability distribution, if for every value $u \in U,$ we have a set of $p + 1$ coefficients $\alpha_{u,0}, \alpha_{u,1}, \ldots, \alpha_{u,p}$ and a variance σ_u^2 such that

$$(X \mid (u, y)) \sim \mathbf{N}\left(\alpha_{u,0} + \sum_{i=1}^{p} \alpha_{u,i} y_i, \sigma_u^2\right). \tag{2.4}$$

A conditional Bayesian network $\mathcal{B}$ over Y given X is defined as a directed acyclic graph $\mathcal{G},$ whose nodes are $X \bigcup Y \bigcup Z,$ where X, Y, Z are disjoint. The variables in X are inputs, the variables in Y are outputs and the variables in Z are encapsulated. The variables in X have no parents in $\mathcal{G}.$ The variables in $Y \bigcup Z$ are associated with a conditional probability distribution. The network defines a conditional distribution using a chain rule

$$P_{\mathcal{B}}(Y, Z \mid X) = \prod_{X \in Y \bigcup Z} P(X \mid Pa(X)). \tag{2.5}$$

The distribution $P_{\mathcal{B}}(Y \mid X)$ is defined as the marginal of conditional Bayesian network as

$$P_{\mathcal{B}}(Y \mid X) = \sum_{Z} P_{\mathcal{B}}(Y, Z \mid X). \tag{2.6}$$

If Y is a random variable with k parents $X_1, X_2, ..., X_k$, then the CPD $P(Y \mid X_1, X_2, ..., X_k)$ is an encapsulated CPD if it is represented using a conditional Bayesian network over *Y* given $X_1, X_2, ..., X_k$.

2.2. Inference of Bayesian networks

Once the Bayesian network is constructed, there is the need to estimate the various probabilities or the causal effects from the model. For instance, in this paper, we intended to determine the causal probability effect of TFs on a particular TF in the model. The computation of these causal probability effects from the model is known as probabilistic inference. In this subsection, we describe probabilistic inference in Bayesian network since the network of variables of *X*, determines a joint probability distribution for *X*. In principle, we use the Bayesian network to compute any probability of interest. Generally, given a Bayesian network that specifies the joint probability distribution in a factored form, one can evaluate all possible inference queries by marginalisation of the variables or nodes. There are most often two types of inference support namely, predictive support for node X_i, based on evidence nodes connected to X_i through its parent nodes (also called top-down reasoning), and diagnostic support for node X_i, based on evidence nodes connected to X_i through its children nodes (also called bottom-up reasoning).

Generally, the inference for discrete variables or nodes is given by the following. Suppose that $X_1, X_2, X_3, ..., X_p$ are discrete variables for p nodes, then the probability of X_i given its parents nodes $Pa(X_i)$, is computed as

$$P(X_i \mid X_1, X_2, ..., X_k)$$

$$= \frac{P(X_i, X_1, X_2, ..., X_k)}{P(X_1, X_2, ..., X_k)} = \frac{P(X_i, X_1, X_2, ..., X_k)}{\sum_{X_i^*} P(X_i^*, X_1, X_2, ..., X_k)}, \tag{2.7}$$

where $k < p$, and $i \neq j$ $(j = 1, 2, ..., k)$ and $\sum_{X_i^*} P(.)$ is the sum of the probability over all possible values of X_i^*, whereas X_i assumes only one value of all the possible values of X_i^*.

This method becomes impractical when there are a lot of variables. However a conditional independence encoded in Bayesian network is employed to make this computation more efficient.

The conditional distribution of a continuous node X_i given its parents $Pa(X_i)$ is specified by a Gaussian function if the variables are normally distributed. If $X_1, X_2, X_3, ..., X_p$ are normally distributed random variables, then the conditional probability function of X_i given its parents nodes $Pa(X_i)$ is

$$f(X_i \mid x_1, x_2, ..., x_k) \sim N(U_i, \sigma^2), \tag{2.8}$$

where $U_i = \mu_i + \sum_{j \in Pa(X_i)} \beta_{ji}(x_j - \mu_j)$, the β_{ji} are the weights or the regression coefficients on the directed arcs to node *i* from its parents, $k < p$ and $i \neq j$ $(j = 1, 2, ..., k)$. Equivalently, we may write

$$X_i = \mu_i + \sum_{j \in Pa(X_i)} \beta_{ji}(x_j - \mu_j) + \sigma_i W_i, \tag{2.9}$$

where $W_i \sim N(0, 1)$ is a white noise random variable. Alternatively,

$$X_i = \sum_{j \in Pa(X_i)} \beta_{ji} x_j + C_i, \tag{2.10}$$

where $C_i \sim N(\mu_i, \sigma_i^2)$ is a coloured noise term.

The joint probability distribution has size $O(2p)$, where *p* is the number of nodes for a binary case. In general, the full summation over discrete (or integration in the case of continuous) variables is the *exact inference* and known as an *NP-hard problem*. Several researchers have developed some efficient probabilistic inference algorithms for Bayesian networks with discrete variables that exploit conditional independence (Ben-Gal [3]). One of the most popular algorithms is the message passing algorithm that solves the problem in $O(p)$ steps (linear in the number of nodes) for polytrees, where there is at most one path between any two nodes (Pearl [23]; Pearl and Russel [25]). Lauritzen and Spiegelhalter [19]) extended the algorithm to general networks. Other exact inference methods include the cyclecutset conditioning (Pearl [23]), variable elimination and clique trees (Koller and Friedman [18]). Approximate inference methods have also been proposed in the literature such as, *Monte Carlo sampling* that gives gradually improving estimates as sampling proceeds (Pearl [24]). A diversity of standard techniques such as Markov chain Monte Carlo (MCMC) methods, including the *Gibbs sampling* and the *Metropolis-Hastings algorithm*, have been used for approximate inference (Griffiths and Yuille [13]; Koller and Friedman [18]). Methods such as the *loopy belief propagation* and *variational methods* (Jordan et al. [16]) which uses the law of large numbers to approximate large sums of random variables by their means and Bayesian networks with other distributions, like the generalized linear regression model, have also been developed (Saul et al. [27]; Jaakkola and Jordan [15]).

2.3. Learning in Bayesian networks

In many practical settings, the Bayesian network is unknown and one needs to learn it from the data. The task of constructing a model from a set of instances, such as data and prior information to estimate the graph topology and parameters of the joint probability distribution of a Bayesian network is model learning. Learning the Bayesian network structure is considered a harder problem than learning the parameters and moreover, other hurdles arise in situations of partial observability when nodes are hidden or when data is missing. In general, four Bayesian network learning problems arise of which different learning methods are proposed (Table 2.1).

Table 2.1. The four problems that arise in Bayesian network learning

Case	Bayesian network structure	Observability	Learning method
1	Known	Full	Maximum likelihood estimation
2	Known	Partial	EM (or gradient ascent), MCMC
3	Unknown	Full	Search through model space
4	Unknown	Partial	EM + search through model space

Source: Ben-Gal [3].

The first case is the simplest and the goal of learning is to determine the values of the Bayesian network parameters (in each CPD) that maximize the log-likelihood of the training dataset (**D**). Given a training dataset $\mathbf{D} = (X_1, X_2, ..., X_m)$, where $X_l = (x_{l1}, x_{l2}, ..., x_{ln})^T$, and the parameter set $\Theta = (\theta_1, \theta_2, ..., \theta_n)$, where θ_i is the vector of parameters for the conditional distribution of variable X_i (represented by one node in the graph). The log-likelihood of the training dataset is a sum of terms, one for each node given as

$$\log L(\Theta | \mathbf{D}) = \sum_m \sum_n \log P(x_{li} | Pa(X_i), \theta_i). \tag{2.11}$$

The log-likelihood scoring function decomposes according to the graph structure, hence the contribution to the log-likelihood of each node is maximised independently (Aksoy [1]). Alternatively, we assign a prior probability density function to each parameter vector and use the training data to compute the posterior parameter distribution and the Bayes estimates. The zero occurrences in **D** of some sequences can be compensated with an appropriate conjugate prior distribution like the Dirichlet prior for the multinomial cases or the Wishart prior for the Gaussian case. This method results in a maximum *a posteriori* estimate or *equivalent sample size* (ESS) method (Ben-Gal [3]).

In general, the other learning cases are computationally inflexible. The second case where the structure is known but partially observable calls for the EM (expectation maximization) algorithm to find a locally optimal maximum likelihood estimate of the parameters (Griffiths and Yuille [13]). MCMC is an alternative approach that has been used to estimate the parameters of the Bayesian network model. In the third case, the goal is to learn a $\mathcal{G}$ that best explains the data. This is an *NP-hard* problem, since the number of $\mathcal{G}$s on *p* variables is super-exponential in *p*. One method is to proceed with the simplest assumption that the variables are conditionally independent given a class, which is represented by a single common parent node to all the variable nodes. This structure corresponds to the naive Bayesian network, which surprisingly is found to give realistically good results in some practical problems. To compute the Bayesian score in the fourth case with partial observability and unknown graph structure, the hidden nodes and the parameters have to be marginalised out. Since this is usually intractable, it is common to use an asymptotic approximation to the posterior which is the Bayesian information criterion (BIC). In this case, one considers the trade-off effects between the likelihood term and a penalty term associated with the model complexity. An alternative approach is to conduct a local search steps inside the *M* step of EM algorithm known as structural EM, that presumably converges to a local maximum of the BIC score (Friedman et al. [10]). In this paper, we apply the third learning method of which the graph $\mathcal{G}$ is searched through the model space since the structure is unknown and we have full observability of the data.

2.4. Materials

This paper employs two different set of data namely; ENCODE pilot data and Mouse Embryonic Stem Cell data. The ENCODE data is a ChIP-chip data directed by Affymetrix for the ENCODE pilot project presented in Carstensen et al. [5]. The data contains regions with binding sites for ten different transcriptional regulatory elements (8 transcription factors and 2 histone modification) in retinoic acid stimulated HL-60 cells reaped after 0, 2, 8 and 32 hours. The ChIP-chip regions of the data have a mean length of about 400 base pairs which are the enhanced regions of DNA with regulatory elements. To determine the causal effect among the TREs, we compute the number of occurrences of TREs in each 21 chromosomes across the different garnered times of TREs.

The Mouse Embryonic Stem Cell ChIP-seq data consists of thirteen sequence-specific TFs (NANOG, OCT4, STAT3, SMAD1, SOX2, ZFX, C-MYC, N-MYC, KLF4, ESRRB, TCFCP2L1, E2F1, and CTCF), two transcription regulators (P300 and SUZ12) and 17442 genes. The analysis of the core transcriptional network of the data is presented in Chen et al. [6]. Embryonic stem cells are obtained from the initial preimplantation embryos, and they can be kept for extended periods of time in culture through self-renewing partition (Smith [28]). To determine the causal effect among the TREs, we use the TRE-gene association scores computed by Chen et al. [6], which is a supplementary material of their study and details of the derivation of the scores is presented in their study.

3. Analysis and Results

The paper considers the two set of genome data described in Subsection 2.4 and used by researchers for biological experimental analysis (Chen et al. [6], ENCODE [30]). Since both datasets are continuous, we fit a linear Gaussian Bayesian network to estimate the causal effects between TFs. In this network, each node represents a univariate normal distribution with a standard deviation and a mean of linear combination of an unconditional mean and the values sampled from the parent vertices. We study the structure of a Bayesian network to help us reveal the important relationships within the dataset and use the model for prediction.

3.1. Results of ENCODE data

To model the causal effects among the TREs of the ENCODE pilot data, the number of occurrences of the TREs along each chromosome of the ten TREs across the four different times that the TREs harvested was recorded and transformed to follow a Gaussian distribution. The TREs (variables) represent the nodes in the Bayesian network model graph and a sample size of 84 (21×4 chromosomes) was used for the study. For continuous networks, linear Gaussian networks provide a more appropriate technique for representing continuous probability distributions within the Bayesian network (Hellman et al. [14]).

Figure 3.1 displays a linear Gaussian Bayesian network graph of causal effect among the ten TREs of the ENCODE data. The nodes or vertices represent the TREs and each edge or arc with arrows corresponds to a dependence relationship between two TREs. The lines with arrows indicate significant causal effect between TREs. Using the Score-Base algorithm of heuristic search (hill-climbing), twenty-three significant causal effects or relationships are obtained between TREs after 270 tests conducted using BIC score. The H3K27 TRE (histone modification) is predicted by four TREs namely, P300, CTCF, H4KAC4 and BRG1. However, it is worth noting that H3K27 itself does not significantly predict any TRE. We also observed that BRG1 predicts directly the conditional distribution of five TREs; H3K27, RARA, CTCF, H4KAC4 and CEBPE. However, BRG1 is itself (and the only TRE) that is not predicted by any element. The second histone modification H4KAC4, significantly predicts two TFs (P300 and CTCF) and the other histone modification element (H3K27) but is predicted by three TREs (PU1, CEBPE and BRG1).

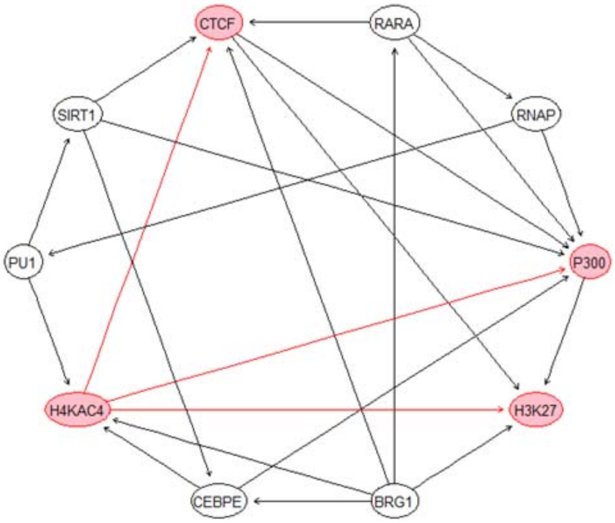

Figure 3.1. ENCODE data directed acyclic graph network plot.

The coefficients of the linear Gaussian Bayesian network model are displayed in Table 3.1. There are nine TREs been predicted conditionally by the combination of nine TREs directly. There are nine linear Gaussian models predicting nine TREs of which the intercepts and the standard errors are the means and standard deviations of the predicted (child) TREs, respectively. It is generally observed that the intercepts of the linear Gaussian models are very negligible. This may be due to the transformation of the data. The coefficients of the independent variables (parent TREs) are known as the weights of the TREs in the linear Gaussian network and are estimated through least squares method. P300 TRE is predicted significantly by six TREs with two TREs negatively depending on it and the other four TREs positively depending on P300 directly. It is observed that CTCF and H3K27 TREs are each predicted by four TREs and both are negatively influenced by H4KAC4. Four elements, namely, PU1, RARA, RNAP and SIRT1 are each predicted by one TRE while CEBPE and H4KAC4 are predicted directly by two and three TREs, respectively. It is generally observed that the linear Gaussian network models with many parent nodes (e.g. P300) records small standard deviation of the residuals. From Table 3.1, CTCF, for example, follows a Gaussian distribution as

$$\begin{aligned} CTCF \sim N(&-3.4e^{-11} + 0.499 BRG1 + 0.339 RARA \\ &+ 0.526 SIRT1 - 0.265 H4KAC4,\ 0.049). \end{aligned}$$

Table 3.1. Model coefficients of the conditional distributions of the ENCODE data TREs

Model	CEBPE	CTCF	P300	PU1	RARA	RNAP	SIRT1	H3K27	H4KAC4
Intercept	$5.1e^{-12}$	$-3.4e^{-11}$	$-4.3e^{-11}$	$6.6e^{-11}$	$-2.2e^{-11}$	$2.6e^{-11}$	$-4.6e^{-11}$	$1.3e^{-11}$	$3.1e^{-12}$
BRG1	0.618	0.499	0.000	0.000	0.955	0.000	0.000	0.297	–0.387
CEBPE	0.00	0.000	0.909	0.000	0.000	0.000	0.000	0.000	1.054
CTCF	0.000	0.000	0.202	0.000	0.000	0.000	0.000	0.237	0.000
P300	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.812	0.000
PU1	0.000	0.000	0.000	0.000	0.000	0.000	0.732	0.000	0.217
RARA	0.00	0.339	–0.417	0.000	0.000	0.852	0.000	0.000	0.000
RNAP	0.000	0.000	–0.262	0.815	0.000	0.000	0.000	0.000	0.000
SIRT1	0.362	0.526	0.184	0.000	0.000	0.000	0.000	0.000	0.000
H4KAC4	0.000	–0.265	0.221	0.000	0.000	0.000	0.000	–0.229	0.000
STD. Error	0.560	0.222	0.174	0.582	0.299	0.527	0.685	0.220	0.393

The residual analysis plots of the linear Gaussian Bayesian network model fits are displayed in Appendix A (Figures A1-A3). Figure A1 shows the residual density plots of the TREs, the residual *q*-*q* plots and the residuals against the fitted values plot of each TREs are displayed in Figures A2 and A3, respectively. The plots indicate a good fit to the data as they depict that assumptions for model fit are generally valid.

3.2. Results of mouse embryonic stem cell data

In modelling the causal effects between the transcription factors (TFs) of the Mouse Embryonic Stem Cell data, we use the TF-gene association scores presented in Chen et al. [6]. The TF-gene association scores were computed through a technique of assigning association score based on the genome location of the binding site that is closest to the transcription start site (TSS). The association score was computed using the distribution of the nearest site-to-TSS distances in the genome and the scores range from 0 – 1. A higher score suggests a higher possibility of the gene being the target of the TF and a zero score implies there is no possibility of the gene being a target of the TF. In all, there are 14 TFs and 17442 TF-genes association scores denoting the sample size. We fit linear Gaussian network model to this data, which is capable of displaying correctly identified network interactions between the TFs given that there is TF-gene association score of a gene with the TFs.

A linear Gaussian Bayesian network of directed acyclic graph displaying the causal effect among the 14 TFs of Mouse Embryonic Stem Cell data is shown in Figure 3.2. The lines with arrows indicate significant causal effects between TFs. The score base search (hill-climbing) produce sixty-three significant direct causal effects among the 14 TFs after performing 1092 BIC score tests. The C-MYC TF, for example, is a parent node for nine TFs (or directly predicts the conditional density of nine TFs): KLF4, OCT4, STAT3, ESRRB, CTCF, SUZ12, E2F1, ZFX and N-MYC. However, it is the only TF that it is itself not predicted by any other TF. We observe that the linear Gaussian model of SMAD1 and TCFCP2L1 TFs are predicted directly by 5 TFs and 10 TFs, respectively. However, the two are the only TFs that do not predict any other TF. The remaining eleven TFs serve as parent nodes and child nodes for TFs as they predict TFs or serve as independent variables in linear Gaussian network models and also at other instances serve as dependent variables in the model.

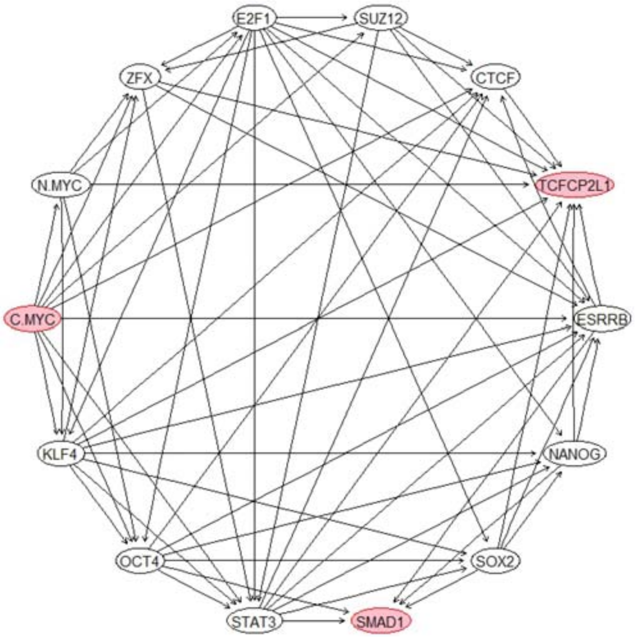

Figure 3.2. Mouse Embryonic Stem Cell data directed acyclic graph network plot.

Table 3.2 shows the coefficients of the linear Gaussian Bayesian network models of mouse embryonic stem cell data. It is observed that of the thirteen linear Gaussian network models, the model with the highest number of covariates (10 TFs) is one predicting TCFCP2L1 and the model with least number of covariates (one TF) is one predicting N-MYC. It is observed that almost all the weights of the covariates measuring the causal effect in all the models are positive except for STAT3 and ZFX TREs models, in which the coefficient of SUZ12 is negative in both cases, and SUZ12 TF model recording negative coefficients with only two covariates. It is worth noting that the SMAD1 TRE linear Gaussian network model records the smallest standard error (0.082) whilst E2F1 TRE model records the largest standard error (0.401). The 13 linear Gaussian Bayesian network models of the TFs are said to follow a normal distribution with the usual parameter estimates.

Table 3.2. Model coefficients of the conditional distributions of Mouse Embryonic Stem Cell TFs

Model	**NANOG**	**OCT4**	**SOX2**	**SMAD1**	**STAT3**	**N-MYC**	**KLF4**	**ESRRB**	**TCFCP2L1**	**ZFX**	**E2F1**	**SUZ12**	**CTCF**
Intercept	0.015	0.066	0.042	0.001	0.060	0.121	0.096	0.074	0.074	0.076	0.252	0.143	0.150
NANOG	0.000	0.000	0.000	0.051	0.000	0.000	0.000	0.133	0.099	0.000	0.000	0.000	0.000
OCT4	0.075	0.000	0.155	0.031	0.133	0.000	0.000	0.057	0.000	0.000	0.000	0.000	0.066
SOX2	0.238	0.000	0.000	0.124	0.000	0.000	0.000	0.102	0.067	0.000	0.000	0.000	0.000
STAT3	0.034	0.000	0.105	0.067	0.000	0.000	0.000	0.135	0.121	0.000	0.000	0.000	0.095
C-MYC	0.000	0.078	0.000	0.000	0.044	0.705	0.075	0.039	0.000	0.091	0.246	-0.045	0.041
N-MYC	0.000	0.040	0.000	0.000	0.000	0.000	0.135	0.000	0.042	0.170	0.504	0.000	0.000
KLF4	0.013	0.059	0.015	0.000	0.046	0.000	0.000	0.112	0.094	0.102	0.000	0.000	0.035
ESRRB	0.000	0.000	0.000	0.011	0.000	0.000	0.000	0.000	0.115	0.000	0.000	0.000	0.046
ZFX	0.000	0.000	0.000	0.000	0.014	0.000	0.000	0.060	0.042	0.000	0.000	0.000	0.000
E2F1	0.030	0.049	0.018	0.000	0.025	0.000	0.199	0.048	0.114	0.306	0.000	-0.116	0.019
SUZ12	0.000	0.000	0.000	0.000	-0.018	0.000	0.000	0.077	0.089	-0.037	0.000	0.000	0.046
CTCF	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.034	0.000	0.00	0.000	0.000
STD Error	0.142	0.212	0.131	0.082	0.190	0.336	0.355	0.304	0.337	0.361	0.401	0.257	0.258

It is generally observed from the residual analysis plots of the embryonic stem cell network (see Appendix A: Figures A4-A6) model that the residuals deviate from normality assumptions. The poor fit of the residuals may be as a result of very large sample observations.

3.3. Discussion of results

The first part of the analyses is based on ENCODE pilot, which covers only 1% of the human genome, and is a pilot study. The interactions between the transcriptional factors in this study corresponding to ENCODE data are not experimentally verified at present. However, our earlier computational study (Kallah-Dagadu et al. [17]) shows significant interactions among the TREs using multivariate linear Hawkes model. We therefore interpreted the findings of the analysis with some caution. The computational analysis often provides the basis for the starting points of experimental studies and validation for the specific causal effects found in the linear Gaussian Bayesian network. Nevertheless, the analysis of the Mouse Embryonic Stem Cells data shows that Bayesian network can find causal effects or direct interactions that can be experimentally verified. The study on the transcriptional regulatory networks with Mouse Embryonic Stem Cell by Chen et al. [6] locates specific genome regions extensively targeted by different TFs. Their study found 44 directed arcs or relationships between TFs whilst our study with the aid of Bayesian network significantly identify 63 directed arcs. This study shows an integrated network of causal effects among the fourteen TFs which is different from the study of Chen et al. [6] which established a network of transcriptional regulation integrated among eleven TFs. Chen et al. [6] shows two key signalling pathways which integrated to the OCT4, SOX2, and NANOG circuitries through SMAD1 and STAT3. Our study is consistent with the earlier studies (Chen et al. [6]; Ying et al. [32]) of multiple transcription factor-binding interaction among SMAD1, OCT4, SOX2, NANOG, KLF4, E2F1, ESRRB and TCFCP2L1. Strikingly, most TFs in the network are associated with TCFCP2L1 or ESRRB. The C-MYC TF is predominantly localised with OCT4, STAT3, N-MYC, KLF4, ESRRB, ZFX, E2F1, SUZ12 and CTCF. Thus, the linear Gaussian Bayesian network model has been able to establish the causal effects between transcriptional factors and capable of predicting accurately other TFs in the sequence.

4. Conclusion

We have presented a probabilistic graphical model which displays the causal relationships between transcriptional factors clearly and intuitively. We have shown that by learning Bayesian network structure, we can efficiently represent the causal effect between TFs, handle uncertainty through the established probability theory and represent indirect with direct causation. Furthermore, we have established an integrated Bayesian network model of the ENCODE pilot data which is capable of predicting the TREs and also serve as the basis for experimental verification. The linear Gaussian network model employed in this study can detect experimentally verifiable relationships as well as missed relationships between TFs of the Mouse Embryonic Stem Cell data computationally. Finally, the paper has shown that there exist causal effects between TREs and has determined a linear Gaussian Bayesian network models for predicting the TREs in the two genome data investigated.

Acknowledgement

We are grateful to the Association of African Universities of the “Small Grants for Theses and Dissertations Programme” support towards this study.

References

[1] S. Aksoy, Parametric models: Bayesian belief networks, Department of Computer Engineering, Bilkent University, Unpublished Lecture Notes, 2006.

[2] M. J. Beal, F. Falciani, Z. Ghahramani, C. Rangel and D. L. Wild, A Bayesian approach to reconstructing genetic regulatory networks with hidden factors, J. Bioinformatics 21(3) (2005), 349-356.

[3] I. Ben-Gal, Bayesian networks, Encyclopaedia of Statistics in Quality and Reliability, F. Ruggeri, R. S. Kenett and F. Faltin, eds., John Wiley & Sons, Ltd., Chichester, 2007.

[4] M. M. Bremer, Identifying regulated genes through the correlation structure of time dependent microarray data, Purdue University, Unpublished Ph.D. Thesis, (2006).

[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] X. Chen, H. Xu, P. Yuan, F. Fang, M. Huss, V. B. Vega, E. Wong, Y. L. Orlov, W. Zhang, J. Jiang, L. Yuin-Han, H. C. Yeo, Z. X. Yeo, V. Narang, K. R. Govindarajan, B. Leong, A. Shahab, Y. Ruan, G. Bourque, S. Wing-Kin, N. D. Clarke, W. Chia-Lin and N. Huck-Hui, Integration of external signalling pathways with the core transcriptional network in embryonic stem cells, J. Cell 133 (2008), 1106-1117.

[7] W. Fang-Xiang, Z. Wen-Jun and J. K. Anthony, Modelling gene expression from microarray expression data with state-space equations, Biocomputing 9 (2004), 588-592.

[8] W. Filipowicz, S. N. Bhattacharyya and N. Sonenberg, Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? Nature Reviews 9 (2008), 102-114.

[9] A. Franzin, F. Sambo and B. D. Camillo, bnstruct: an R package for Bayesian network structure learning in the presence of missing data, Bioinformatics 33(8) (2016), 1250-1252.

[10] N. Friedman, D. Geiger and M. Goldszmidt, Bayesian network classifiers, J. Machine Learning 29 (1997), 131-163.

[11] N. Friedman, T. Hastie and R. Tibshirani, Sparse inverse covariance estimation with the graphical lasso, Biostatistics 9(3) (2009), 432-441.

[12] J. Friedman, T. Hastie and R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, J. Statistical Software 33(1) (2010), 1-22.

[13] T. L. Griffiths and A. Yuille, A primer on probabilistic inference, Trends in Cognitive Sciences 10(7) (2006), 111 (Supplement to special issue on probabilistic models of cognition).

[14] S. Hellman, A. McGovern and M. Xue, Learning ensembles of continuous Bayesian networks, an application to rainfall prediction, University of Oklahoma, Norman, 2012, Unpublished.

[15] T. Jaakkola and M. Jordan, Recursive algorithms for approximating probabilities in graphical models, Proc. 10th Conference on Neural Information Processing Systems (NIPS), 1996, pp. 487-493.

[16] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola and L. K. Saul, An introduction to variational methods for graphical models, Learning in Graphical Models, M. I. Jordan, ed., Kluwer Academic Publishers, Dordrecht, 1998.

[17] G. Kallah-Dagadu, B. K. Nkansah and N. Howard, Flexible statistical modeling of occurrences of transcription factors, (2018) (in press under review).

[18] D. Koller and N. Friedman, Probabilistic Graphical Models: Principles and Techniques, Massachusetts Press, London, Cambridge, 2009.

[19] S. L. Lauritzen and D. J. Spiegelhalter, Local computations with probabilities on graphical structures and their application to expert systems (with discussion), J. Royal Statistical Society, Series B 50(2) (1988), 157-224.

[20] A. Lotsi and E. Wit, State-space modelling of replicated dynamic genetic networks, British J. Appl. Sci. Tech. 17(4) (2016), 1-18.

[21] N. J. Martinez and A. J. M. Walhout, The interplay between transcription factors and microRNAs in genome-scale regulatory networks, BioEssays 31 (2009), 435-445.

[22] M. Maucher, B. Kracher, M. Kühl and H. A. Kestler, Inferring Boolean network structure via correlation, Bioinformatics 27(11) (2011), 1529-1536.

[23] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann, San Francisco, Califomia, 1988.

[24] J. Pearl, Evidential reasoning using stochastic simulation of causal models, Artificial Intelligence 32(2) (1987), 245-258.

[25] J. Pearl and S. Russel, Bayesian networks, Report (R-277), November 2000, Handbook of Brain Theory and Neural Networks, M. Arbib, ed., MIT Press, Cambridge, 2001.

[26] A. Rau, F. Jaffrézic, J.-L. Foulley and R. W. Doerge, An empirical Bayesian method for estimating biological networks from temporal microarray data, Stat. Appl. Genet. Mol. Biol. 9(1) (2010), Art. 9, 28 pp.

[27] L. Saul, T. Jaakkola and M. Jordan, Mean field theory for sigmoid belief networks, J. Artificial Intelligence Research 4 (1996), 61-76.

[28] A. G. Smith, Embryo-derived stem cells: of mice and men, Annual Review of Cell Development Biology 17 (2001), 435-462.

[29] A. Suzuki, A. Raya, Y. Kawakami, M. Morita, T. Matsui, K. Nakashima, F. H. Gage, C. Rodriguez-Esteban and J. C. Izpisua-Belmonte, Nanog binds to Smad1 and blocks bone morphogenetic protein-induced differentiation of embryonic stem cells, Proceedings National Academic of Science, USA 103 (2006), 10294-10299.

[30] 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.

[31] J. D. Watson and F. H. C. Crick, A structure for deoxyribose nucleic acid, Nature 171 (1953), 737-738.

[32] Q. L. Ying, J. Nichols, I. Chambers and A. Smith, BMP induction of Id proteins suppresses differentiation and sustains embryonic stem cell selfrenewal in collaboration with STAT3, Cell 115 (2003), 281-292.

[33] W. C. Young, A. E. Raftery and K. Y. Yeung, Fast Bayesian inference for gene regulatory networks using ScanBMA, BMC Systems Biology 8 (2014), 47.

Appendices

Appendix A. Residual plots of TREs/TFs for ENCODE pilot and Mouse Embryonic Stem Cell data

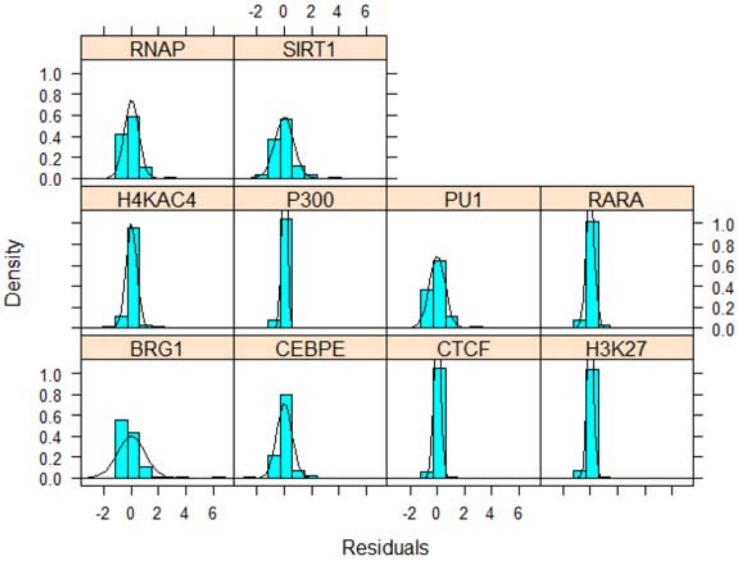

Figure A1. Histogram plots of TREs residuals of the ENCODE pilot data.

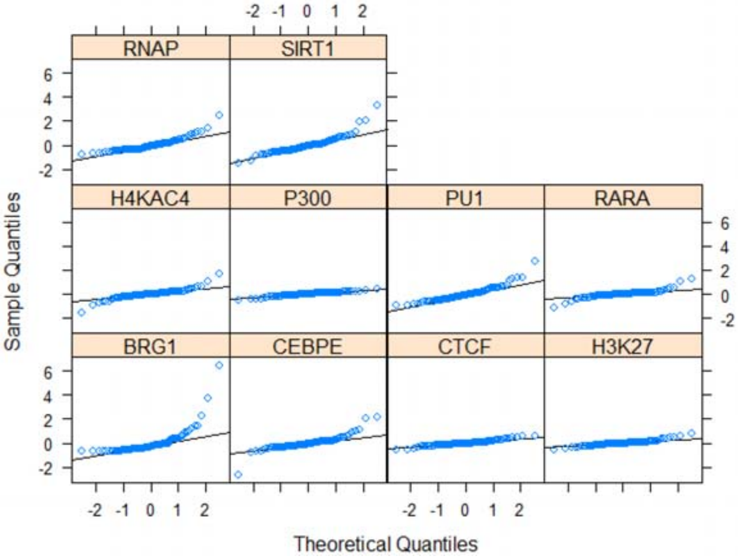

Figure A2. *q-q* plots of TREs residuals of the ENCODE pilot data.

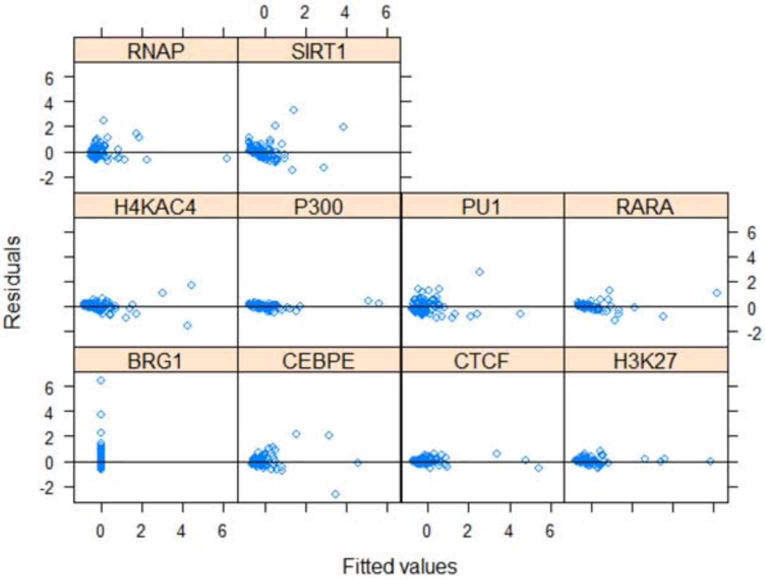

Figure A3. The plots of fitted values of TREs against their residuals of the ENCODE pilot data.

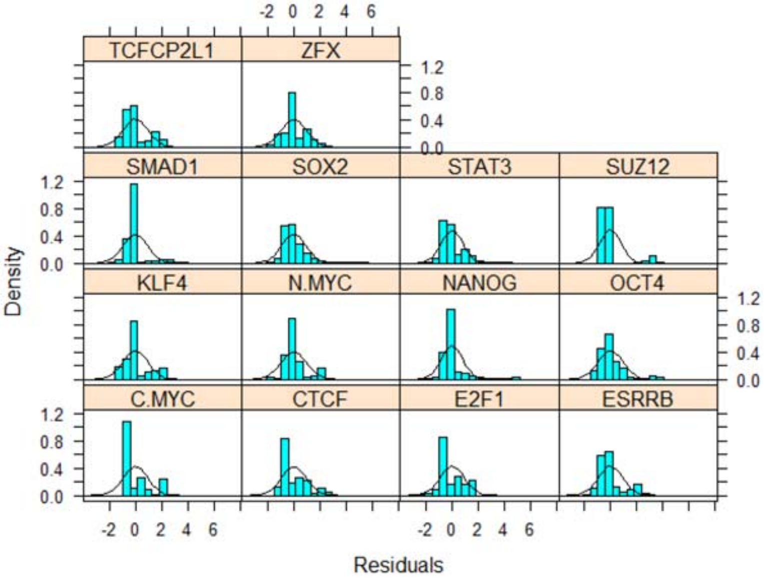

Figure A4. Histogram plots of TREs residuals of mouse embryonic stem cell data.

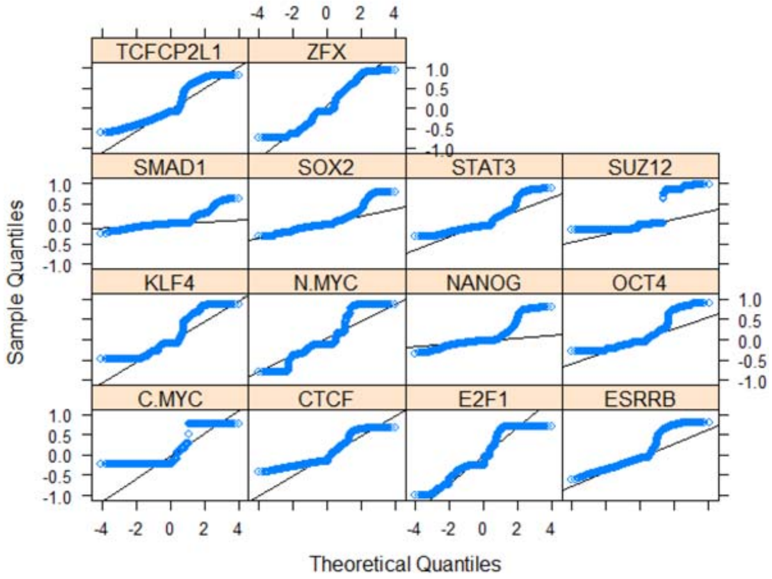

Figure A5. *q*-*q* plots of TREs residuals of mouse embryonic stem cell data.

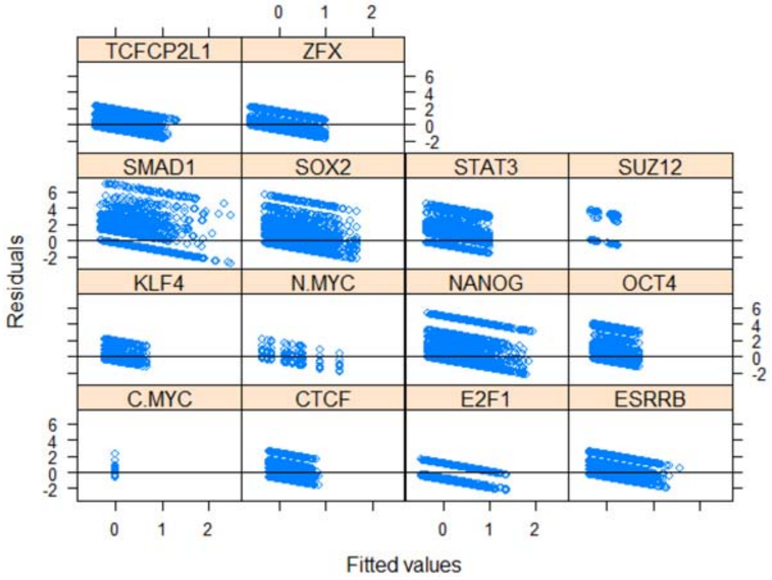

Figure A6. The plots of fitted values of TREs against residuals of mouse embryonic stem cell data.

Appendix B. List of Abbreviations

AIC	:	Akaike information criterion
BIC	:	Bayesian information criterion
DNA	:	deoxyribonucleic acid
TRE	:	transcriptional regulatory element
TF	:	transcription factor
BRG1	:	SWI/SNF related, matrix associated, acting dependent regulator of chromatin, subfamily a, member 4
CEBPE	:	CCAAT/enhancer binding protein (C/EBP), epsilon
CTCF	:	CCCTC-binding factor (zinc finger protein)
H3K27me3 (H3K27T)	:	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
SMAD1	:	MAD homolog 1
ZFX	:	zinc finger protein X-linked
SOX2	:	SRY-box containing gene 2
STAT3	:	signal transducer and activator of transcription 3

SUZ12	:	suppressor of zeste 12 homolog
NANOG	:	Nanog homeobox
C-MYC	:	myelocytomatosis oncogene
E2F1	:	E2F transcription factor 1
N-MYC	:	v-myc myelocytomatosis viral related oncogene, neuroblastoma derived
OCT4	:	POU domain, class 5, transcription factor 1