Neurocomputing 275 (2018) 1490–1499

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Predicting gene expression from genome wide protein binding profiles

Mohsina M. Ferdous[a,*], Yanchun Bao[b], Veronica Vinciotti[c], Xiaohui Liu[a], Paul Wilson[d]

[a] Department of Computer Sciences, Brunel University London, Uxbridge UB8 3PH, UK
[b] Institute for Social and Economic Research (ISER), University of Essex, Colchester CO4 3SQ, UK
[c] Department of Mathematics, Brunel University London, Uxbridge UB8 3PH, UK
[d] Computational Biology, GlaxoSmithKline Medicine Research Centre, Stevenage SG1 2NY, UK

ARTICLE INFO

Article history:
Received 27 June 2017
Revised 24 September 2017
Accepted 28 September 2017
Available online 6 October 2017

Communicated by Dr. Nianyin Zeng

Keywords:
ChIP-seq
Epigenetics
Gene expression
Markov random field
Machine learning

ABSTRACT

High-throughput technologies such as chromatin immunoprecipitation (IP) followed by next generation sequencing (ChIP-seq) in combination with gene expression studies have enabled researchers to investigate relationships between the distribution of chromosome-associated proteins and the regulation of gene transcription on a genome-wide scale. Several attempts at integrative analyses have identified direct relationships between the two processes. However, a comprehensive understanding of the regulatory events remains elusive. This is in part due to the scarcity of robust analytical methods for the detection of binding regions from ChIP-seq data. In this paper, we have applied a recently proposed Markov random field model for the detection of enriched binding regions under different biological conditions and time points. The method accounts for spatial dependencies and IP efficiencies, which can vary significantly between different experiments. We further defined the enriched chromosomal binding regions as distinct genomic features, such as promoter, exon, intron, and distal intergenic, and then investigated how predictive each of these features are of gene expression activity using machine learning techniques, including neural networks, decision trees and random forest. The analysis of a ChIP-seq time-series dataset comprising six protein markers and associated microarray data, obtained from the same biological samples, shows promising results and identified biologically plausible relationships between the protein profiles and gene regulation.

1. Introduction

Chromatin immunoprecipitation combined with massively parallel DNA sequencing (ChIP-seq) is a method used to identify the binding sites of chromosome-associated/'epigenetic' proteins (Note that the term epigenetic will be used in its broadest sense throughout this manuscript.). ChIP-seq in combination with gene expression data enables researchers to investigate relationships between chromosomal-bound protein regulatory mechanisms and gene expression responses on a genome-wide scale. However, despite falling costs, next generation sequence data remains too expensive to be generated on a large scale, and it is generally considered logistically unfeasible to generate next generation data from clinical trials where thousands of samples are involved. It is therefore currently not possible to determine how modification of the epigenetic landscape regulates changes in gene expression within large patient cohorts. Conversely, there are many studies where ChIP-seq data is in the public domain but the corresponding gene expression data is not available: and again, it is not possible to understand how epigenetic modifications dictate gene expression responses [8]. We propose that machine learning data models could be used to address such situations, by modelling the mechanistic relationships between observed gene expression responses and the corresponding epigenetic modifications. Once the association between gene expression and epigenetic regulatory events is defined, it should then be possible to predict one from the other and extrapolate this information into a deeper understanding of gene regulation mechanisms.

The computational biology community has proposed several methodologies that integrate protein binding and gene expression data to identify causal relationships between the two events. However, existing studies of high-throughput data have adopted relatively simple methods for the analyses of ChIP-seq data, which do not fully leverage all the information that this technology can offer. Furthermore, such studies generally restrict an investigation to the relationship between a single chromatin protein and a gene expression profile [24].

* Corresponding author at: Wolfson Institute of Preventive Medicine, Queen Mary University of London, Charterhouse Square, London EC1M 6BQ, UK.
E-mail address: m.ferdous@qmul.ac.uk (M.M. Ferdous).

https://doi.org/10.1016/j.neucom.2017.09.094

A further potential limitation of current integrative methods is that studies tend to primarily focus on protein bindings sites located only within gene promoter regions and transcriptional start sites [5,17]. Likewise, classification techniques have been used to elucidate relationships between epigenetic mechanisms and gene expression while focusing on a single genomic feature [4,12]. For example, the linear model named GEMULA [4] models gene expression as a function of predicted transcription factor binding to promoter regions. However, several reports have proposed that additional genomic regions, such as introns and exons in combination with distal enhancers, play important roles in gene regulation, and that both the number and length of exons and introns influence gene transcription [7,15,16]. Furthermore, first exons are reported to be enriched for regulatory signals, and conservation of the first intron has been reported to be positively correlated with gene expression [16]. Such observations suggest that, in order to fully understand epigenetic transcriptional regulatory mechanisms, protein binding data associated with both exons and introns, along with promoters, should be included in molecular models of transcription.

Further understanding of epigenetic regulatory mechanisms is also complicated by the dynamic nature of gene expression and the binding profiles of DNA-associated proteins, both of which change markedly in response to different biological stimuli and with time [12,22].

The primary objective of this study was to explore how genomic protein binding profiles could be predictive of gene expression and help elucidate epigenetic regulatory mechanisms. However, prior to identifying such associations another important goal was to better characterise the complex characteristics of ChIP-seq data and use this information to determine the most appropriate means of data pre-processing and modelling. We also considered that the protein profiles may prove more informative if our model included details of the genomic features where binding occurred (e.g. promoter, exon etc.) and how these changed with time and treatment.

A recently developed Markov random field model, that incorporates complex characteristics of ChIP-seq data, such as spatial dependencies and different immunoprecipitation (IP) efficiencies across replicates and biological conditions, was used to identify ChIP-seq binding regions [1]. The enriched binding regions were used to create protein profiles with respect to the genomic features. And the predictive power of the respective profiles was evaluated using advanced machine learning techniques; including neural networks, decision trees and random forest.

The described method clearly illustrates how the interactions of regulatory binding proteins, gene expression, time and treatment can be integrated into a unified model that is predictive of biologically plausible relationships between the protein profiles and gene expression.

2. The method

Fig. 1 contains a flowchart of the proposed method.

In this model, microarray technology is used to identify a set of genes of interest at a biological or experimental condition. In parallel, ChIP-seq data is used to create binding profiles of a set of proteins of interest under the same condition. The binding profiles indicate whether the proteins bind at those genes of interest and if so, which genomic features (e.g. a promoter) do they bind to. The binding profile and gene status data are integrated and modelled using different classification techniques. How accurately the protein binding profiles predicts gene expression is then quantified. Comparing performance of the different predictive models identifies both the proteins and genomic features that are most predictive of gene transcription.

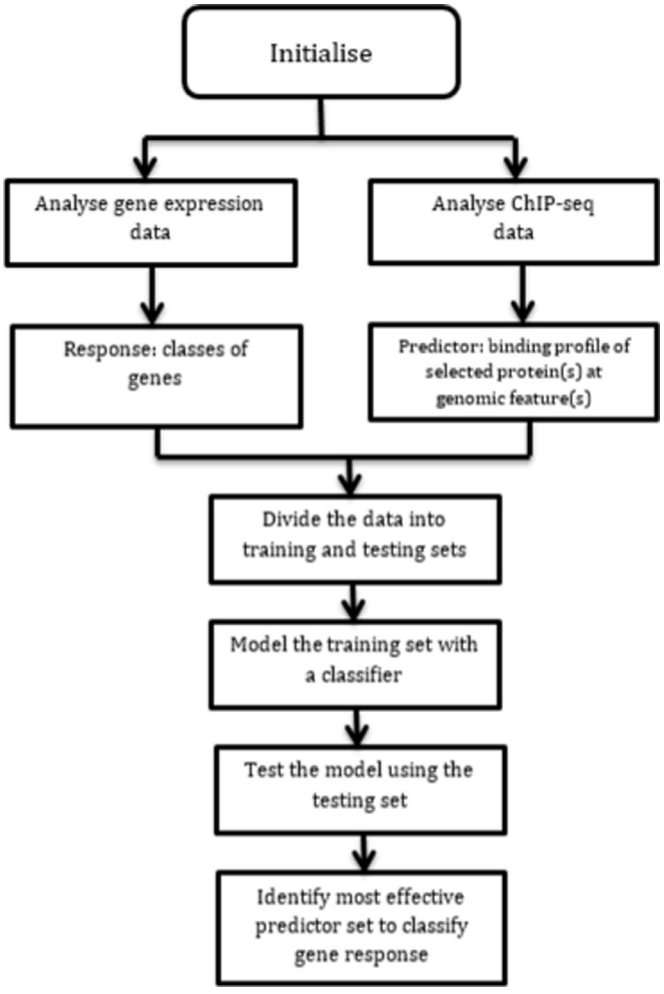

Fig. 1. Simple schematic of the process used to model and predict gene response using the epigenetic protein binding profiles in combination with different genomic features.

2.1. Microarray analysis

Genome-wide gene expression levels are determined using microarray technology. Expression datasets can be obtained from different biological conditions (e.g. treated/untreated), and each condition ideally represented by several biological/technical replicates. A collective of the average gene expression level for all genes under one condition is called the sample expression profile. This profile can be represented using one of several descriptors, for example, absolute measurement, expression ratio, or a discrete value, and each gene is classified as 'active' or 'inactive' depending on the observed value. Furthermore, classification or clustering techniques can be used to classify the genes, or, group the genes using an agreed expression value. However, given that the dataset used in this study details several biological conditions, differential expression analysis between conditions of interest was considered the most pertinent method of classifying gene status. Note that changes in gene expression are coordinated in biological systems (i.e. not truly independent). However, due to our limited understanding of transcriptional mechanisms, differential expression is measured per gene using appropriate statistical tests and differentially expressed genes selected using a stringent fold-change criterion.

In the popular R limma package [19], an empirical Bayes approach is implemented via a global variance estimator s_0^2 which is

computed using all genes' variances. The resulting test statistic is a moderated t-statistic, where instead of the single-gene estimated variances s_g^2, a weighted average of s_g^2 and s_0^2 is used. Under certain distributional assumptions, this test statistic can be shown to follow a t-distribution under the null hypothesis with the degrees of freedom dependent on the data being analysed [19].

Following differential analysis, a gene may be classed as 'up-regulated' or 'downregulated' between different biological or experimental conditions and labelled 0 or 1 respectively. Note that gene status is used as the response variable in the machine learning analysis model.

2.2. Analysis of ChIP-seq data

The proposed model can incorporate details of any number of transcription factor or proteins. However, the epigenetic and expression data must both be generated from the same biological condition. A peak calling method is used to locate the genomic regions that are bound by the protein in each ChIP-seq sample. The peak calling method ideally incorporates all the characteristics of ChIP-seq data, such as spatial dependency of regions along the genome, IP efficiency of individual experiments, and excess zeroes of the resulting count data. To account for all these features, we have used a recently developed Markov random field (MRF) model, proposed by Bao et al. [1], to analyse the ChIP-seq data.

Given count data, reporting the number of fragments aligned to consecutive regions of the genome of a pre-defined fixed size (i.e. bins), the MRF model evaluates the distribution of the counts across the genome in question and assigns each region a probability of it being enriched or not. Additional factors, such as the enrichment score of neighbouring regions is also considered while calculating this probability (i.e. to account for spatial dependencies). A brief overview of the model is summarised below. For a more detailed description see Bao et al. [1].

Let M be the number of total bins in a particular chromosome. Let Y_{mcr} be the counts in the mth bin, $(m = 1,\ 2,\ 3,\ \ldots, M)$, under condition c (time points or control) and replicate r. The counts can be from either background (non-enriched region) or from the signal (enriched regions). Our goal is to infer the state of the latent variable X_{mc}, which is defined as 1 if region m is enriched in condition c, and zero otherwise. The joint mixture model for Y_{mcr} can be written as follows:

$$Y_{mcr} \sim p_c f\left(y, \theta_{cr}^S\right) + (1 - p_c) f\left(y, \theta_{cr}^B\right) \tag{1.1}$$

where $p_c = P(X_{mc} = 1)$ is the mixture portion of the signal component and $f(y, \theta_{cr}^S)$ and $f(y, \theta_{cr}^B)$ are the signal and background densities, respectively.

One of the attractive characteristics of this model is that the probability p_c of a region being enriched does not depend on ChIP efficiencies. However, the parameters signal and background distributions θ_{cr}^S and θ_{cr}^B do depend on ChIP efficiencies of the individual replicates r and are therefore allowed to be estimated uniquely for each replicate. Typically, a ChIP-seq signal $f(y, \theta_{cr}^S)$ is modelled as a negative binomial distribution, whereas the background signal $f(y, \theta_{cr}^B)$ is modelled as a zero-inflated negative binomial distribution to account for the excess number of zeros. This leads to:

$$Y_{mcr} | (X_{mc} = 0) \sim\ ZINB(\pi_{cr}, \mu_{0cr}, \phi_{0cr}) \tag{1.2}$$

$$Y_{mcr} | (X_{mc} = 1) \sim\ NB(\mu_{1cr}, \phi_{1cr}). \tag{1.3}$$

The latent variable, X_{mc}, which represents the binding profile, is assumed to satisfy 1D Markov properties

$$p(X_{mc} | X_{-mc}) = p(X_{mc} | X_{m-1,c},\ X_{m+1,c}), \tag{1.4}$$

that is, the enrichment of a region given all the other regions depends only on the state of the two adjacent regions. All the parameters in this model are estimated using a Bayesian approach and are implemented in the R package enRich [1].

Using this model, the enriched regions can be detected by setting a threshold on the posterior probabilities of enrichment. One way to set this threshold is by fixing an acceptable false discovery rate (FDR). If D is the set of declared enriched regions corresponding to a particular cut-off on the posterior probabilities, then the estimated false discovery rate for this cut-off is given by

$$\widehat{FDR} = \frac{\sum_{m \in D} \hat{P}(X_{mc} = 0 | Y)}{|D|}. \tag{1.5}$$

2.3. Creating the binding profile of proteins

After the protein binding regions are identified and annotated, a binding profile of the proteins, required for integration with the expression data, is generated. The method for creating the protein binding profile is as follows. Let us assume that differential expression analysis identifies a set of m genes for a biological condition c, and that the annotated binding regions of p proteins are identified in the ChIP-seq analysis step. Each binding site is annotated with a gene symbol, of the closest gene, and the genomic feature in which the binding site is located (e.g. promoter, exon etc.). Let f be the number of genomic features to be included in the study. For each condition c, the binding profiles of p proteins for m genes and f genomic features are stored in an $m \times pf$ matrix where X_{ijkc} represents the binding status (1 or 0) of protein j to the feature k of gene i at biological condition c.

2.4. Classification model selection and evaluation

For classification purposes, the binding profile of the p proteins for m genes and f genomic features are used as the predictor. The expression status of the m genes is considered as the response variable. Note that this model can be extended to include more than one biological condition. In such a scenario, assume we have c biological conditions and from each of these we identify a set of genes along with their activity status. For each set s_l, where l represents the experimental condition, the binding profile of the proteins must be created from the same biological condition l. The binding profile is then integrated with the associated gene status, for classification. This data model is attractive in that it can be implemented with common classification techniques, such as neural network, random forests and decision trees.

10-fold cross validation is used on each of the feature selection and classification methods, to identify the most descriptive model. Classification accuracy is used to evaluate the performance of the models.

3. Experimental results

3.1. Summary of the data

All data values were collected from murine bone-marrow derived macrophages (BMDMs), stimulated with lipopolysaccharide (LPS), and from LPS stimulated BMDMs treated with a synthetic compound (I-BET). As I-BET mimics acetylated histones, I-BET's presence disrupts the chromatin complexes that regulate expression of key inflammatory genes in activated BMDMs. Data were collected at three time points: 0, 1 and 4 h. The epigenetic data was generated from a ChIP-seq time-series dataset that included quantification of bromodomain-containing protein 4 (Brd4); acetylated histone H4 (H4ac); histone H3 lysine 4 tri-methylation (H3K4me3); RNA polymerase II (RNA PolII); subunit of RNA polymerase II (RNA PolII S2); and cyclin-dependent kinase 9 (CDK9)

Table 1
Summary of the number of 200 bp binding regions identified at 5% FDR for each of the six proteins of interest and the three biological conditions investigated.

Proteins	LPS stimulated at 0H	LPS stimulated at 4H	IBET treated at 4H
	Number of 200 bp enriched regions at 5% FDR		
RNA polymerase II	1,132,284	705,177	625,282
RNA polymerase II S2	1,020,916	1,282,471	666,159
H3K4me3	293,266	327,854	318,679
H4ac	170,087	218,960	166,806
Brd4	151,048	135,101	38,831
CDK9	166,600	105,905	122,004

proteins/markers. Gene expression data was generated using the Illumina Genome Analyser II. See Nicodeme et al. [13] for further experimental details. Note that all data used in this study is publicly available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE21910.

3.2. Data pre-processing

3.2.1. Microarray data analysis

Gene expression data generated using Illumina bead array technology was pre-processed using the R package beadarray [3] and analysed by linear modelling to detect differential expression using the limma package [20]. To determine the effect of LPS, the LPS stimulated expression profile obtained at the 4-h time point was compared to the 0-h expression profile. To quantify the effect of IBET treatment on LPS induced genes, LPS + IBET treated samples at the 4-h time point were compared to the 4-h LPS only samples. A 2-fold change in expression, in association with a corrected *p*-value of less than .05, was considered differentially expressed. Using this threshold, 652 genes were defined as up regulated by LPS between 0 and 4 h time points. 183 of these genes were observed to be downregulated in response to IBET treatment. To facilitate the integrative analysis, each microarray probe was annotated with its respective Entrez gene symbol and these gene symbols were used to integrate the expression data with the ChIP-seq data.

3.2.2. ChIP-seq data analysis

The ChIP-seq reads were aligned to the mouse genome (obtained from the UCSC Genome Browser version mm9, released in 2007) using Bowtie [9] and only the uniquely mapped reads were retained for further analysis. As the distribution of the counts of sequences varies by chromosome, all chromosomes were modelled separately. After the alignment process, data from each of the 19 autosomal chromosomes were collected and the sequence counts per 200 bp region of each chromosome were determined. Regions found to be enriched at 5% FDR (Table 1) were selected for further analysis.

3.2.3. Annotation of the peaks

After the ChIP-seq datasets were analysed using the MRF model, a list of bound genomic regions (200 bp long) was obtained. These regions were then annotated with the nearest gene names using the R package ChIPseeker [25]. Note that the input for the annotation package is the binding locations of the ChIP-seq data in BED format. Peaks are annotated with the gene symbol, gene name and genomic feature. For example, if a peak was located within the 5′UTR of a gene, it was annotated as 5UTR and the gene symbol of that gene. The genomic features considered in this study were: promoter, exon, 5′UTR, 3′UTR, intron, and distal intergenic. The R package TxDb.Mmusculus.UCSC.mm9.knownGene [2] encoded a TxDb object detailing mouse genome mm9 build. The TxDb object contained the transcript-related features used to retrieve annotations from both the UCSC and BioMart data resources.

Table 2
Correlation values of binding profile of the six epigenetic proteins at different genomic features given gene status (4 h post LPS stimulation).

Proteins	Promoter	Distal intergenic	Exon	Intron
RNA PolII	0.525	0.429	0.525	0.524
RNA PolII S2	0.491	0.384	0.483	0.465
H3k4me	0.274	0.184	0.198	0.275
H4ac	0.384	0.273	0.222	0.290
Brd4	−0.002	−0.003	0.039	0.029
CDK9	0.029	−0.003	0.047	−0.013

The percentages of protein peaks located within the genomic features of interest were plotted as simple pie charts (Fig. 2). It is apparent that H3K4me3 and H4ac are mostly bound within promoter regions, while RNA PolII and RNA PolII S2 are often located within intron regions. In contrast, CDK9 and Brd4 are predominantly bound at distal intergenic regions.

3.2.4. Generation of protein binding profile and integration of both datasets

After annotating the peaks, the binding profile of each of the six proteins for four genomic features under the three biological conditions were generated (see Section 2.3). 652 unique genes were classified as expressed and up-regulated by LPS at 4-h time points. The expression values of these proteins were at the upper end of the ranked profile (i.e. >9.52) and these were assigned to be class 1. Further 609 genes with lower expression values (i.e. <5.72) were selected as low/non-expressed and assigned to be class 0. The binding profile for these 1261 genes was associated with the annotated peak file (see Section 2.4) for each of the four genomic features. Pearson correlation coefficients were then calculated between the respective input and output variables. The resulting correlation value (Table 2) was interpreted as indicative of how the binding or non-binding of a protein at a specific genomic feature affected gene regulation. RNA polymerase was reported to be most correlated with expression status at all genomic features. H3k4Me and H4ac were weekly correlated while both Brd4 and CDK9 were not correlated at any of the four genomic features evaluated.

This result indicates that the binding of RNA polymerase II at promoter, intron, exon and distal intergenic regions is significantly correlated with gene regulation response, and that Brd4 and CDK9 binding contribute the least.

3.3. Predicting gene statuses with machine learning approaches

3.3.1. Neural networks

Integrated ChIP-seq and microarray data were modelled using neural networks to evaluate whether a protein binding profile could predict gene expression status. nnet [18] is an R package that implements feed-forward neural networks with a single hidden layer. In this study, the R package e1071 and its wrapper function were used to model the data, and the performance of the

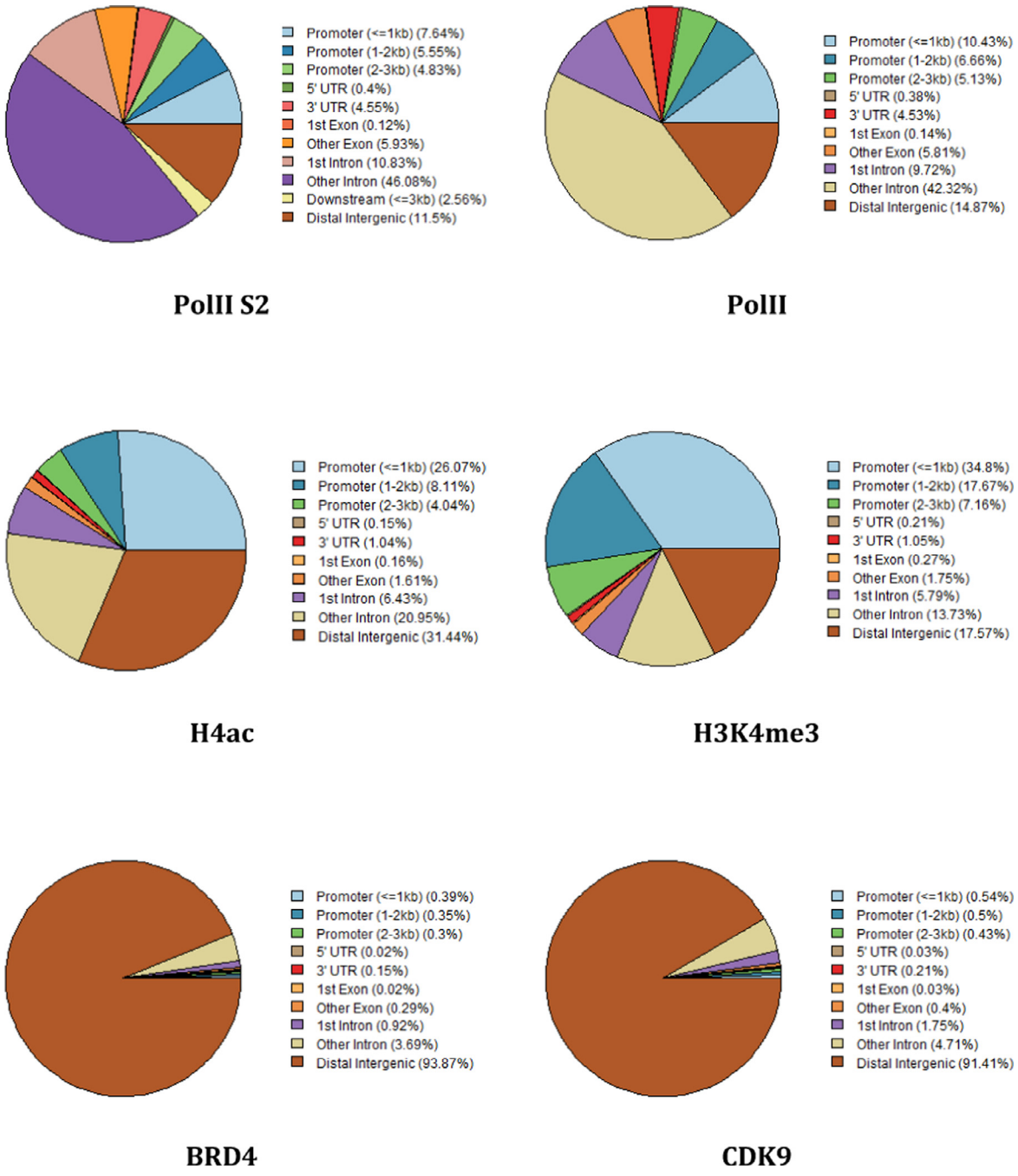

Fig. 2. Graphical summaries of the feature distribution of the six bound proteins derived from the I-BET treated samples at the 4-h time point. The numerical values report the percentage of features bound by the specific protein.

classifier was assessed by 10-fold cross validation. Different combinations of proteins were selected as predictors of gene status and those combinations reporting the highest accuracy were summarised in Tables 3 and 4.

The LPS stimulation results (Table 3) indicated that the binding profiles of RNA PolII, RNA PolII S2 and H4ac bound to the promoter region predict the expression data most accurately, while the binding profile at distal intergenic regions is the least predictive feature. Neither Brd4 nor CDK9 performed as strongly as any of the four other proteins.

Comparing the I-BET inhibition profile results indicated that most of the models report similar accuracies (Table 4) and that RNA PolII, RNA PolII S2 and H4a all are valuable predictors of expression. However, since the different combinations of predictors bound to different features produce equivalent results, it is unclear which proteins or features perform best. Again, Brd4 and CDK9 did not feature in the ranked best performing models.

3.3.2. Decision trees

The R package rpart [23] was used to fit recursive partitioning and regression trees to the data. As described above, the class detailing the 1261 LPS gene profile was used as the response variable and the binding profile of the six proteins at a promoter for those genes as the predictors. The R package rpart creates the tree with only important variables that can classify the response well. In this experiment, three datasets were used as input for the classifier: (1) the binding profile of all proteins at promoter (see Fig. 3), (2) the binding profile of all proteins at four genomic features (see Fig. 4)

Table 3
Performance of neural network predicting gene expression using various combinations of the epigenetic binding profile 4 h post LPS stimulation (i.e. this defines the LPS response). Numerical values are the percent accuracy after 10-fold cross validation.

Combination of variables	Genomics features			
	Promoter	Exon	Intron	Distal intergenic
PolII + PolII_S2 + H4ac	83.16	81.90	82.18	80.35
PolII + PolII_S2 + H3K4me + H4ac	82.36	82.35	82.30	80.37
PolII + H4ac	82.48	82.16	81.96	80.07
PolII + PolII_S2	82.67	81.76	82.03	80.40
PolII + PolII_S2 + H3K4me	81.82	81.33	80.20	79.87
PolII_S2 + H4ac	81.97	80.19	80.88	79.44
H3K4me + H4ac	79.99	76.76	78.39	77.18

Table 4
Performance of neural network when predicting up and downregulation of gene expression using binding profile of proteins as predictors in terms of accuracy (%) after 10-fold cross validation. This analysis was completed using 183 genes up-regulated at 4H (LPS only) and downregulated by I-BET at 4H (i.e. LPS + I-BET).

Combination of variables	Genomic features			
	Promoter	Exon	Intron	Distal intergenic
PolII + PolII_S2 + H4ac	77.20	78.34	78.98	76.36
PolII + PolII_S2 + H3K4me + H4ac	77.39	78.56	77.02	75.46
PolII + H4ac	75.34	77.89	76.91	76.07
PolII + PolII_S2	77.40	78.90	78.04	76.80
PolII + PolII_S2 + H3K4me	77.09	78.68	78.45	76.73
PolII_S2 + H4ac	78.90	78.23	78.67	76.05
H3K4me + H4ac	75.74	75.67	75.08	75.86

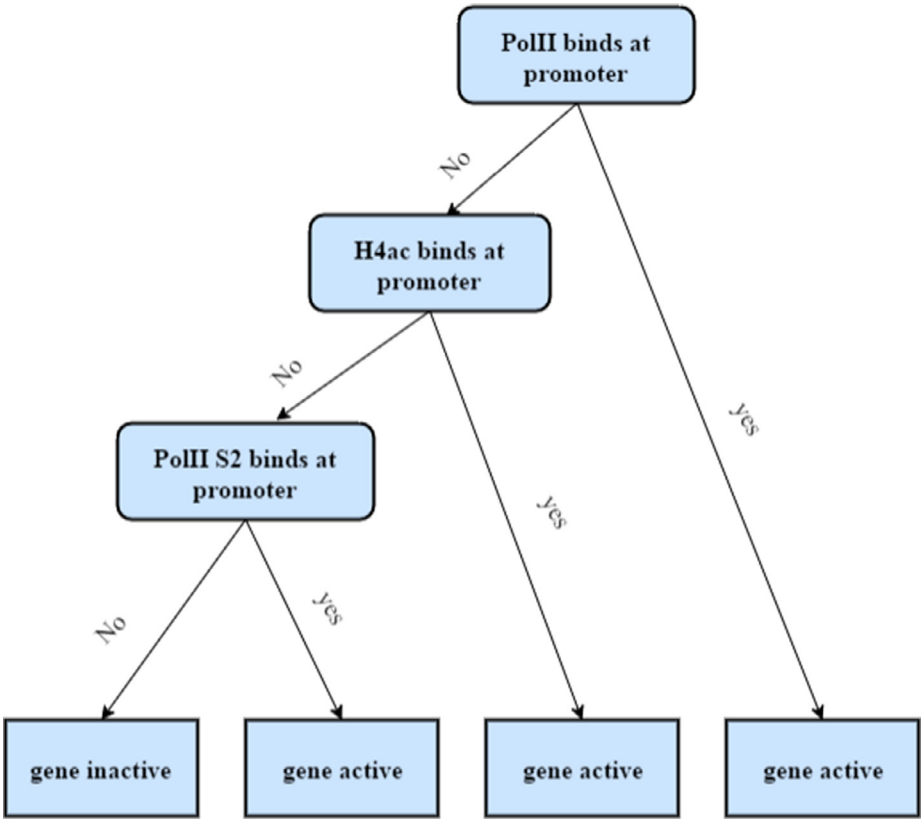

Fig. 3. Decision tree generated using rpart and the LPS-only profile. Leaf nodes represent the gene classification while the root and internal nodes represent binding of a protein at a promoter region.

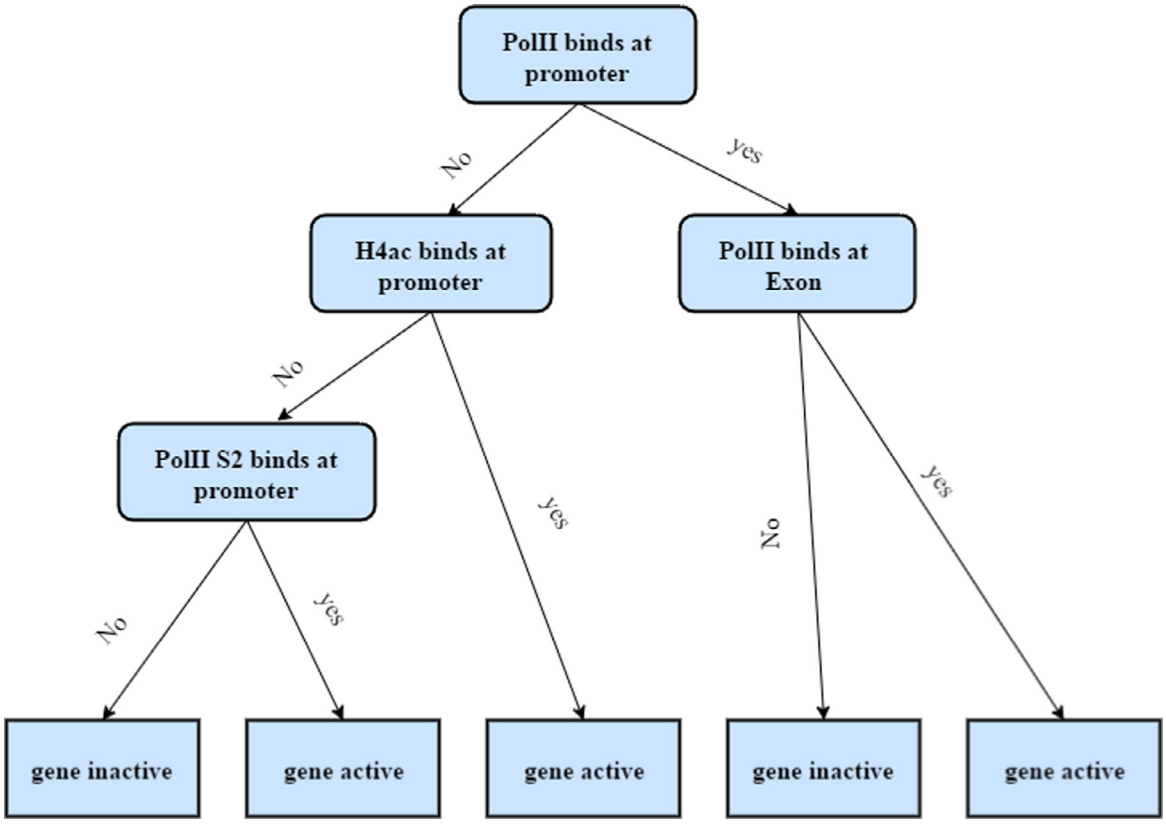

Fig. 4. Decision tree where leaf nodes represent gene classes while the root and internal nodes represent binding of protein at different genomic regions (promoter, exon etc.).

and (3) the binding profile of all proteins to a promoter at different time points (see Fig. 5). The tree constructed with the binding profile of all proteins bound to a promoter is depicted in Fig. 3.

The resulting tree indicated that if RNA PolII binds at a gene promoter, the gene will be active. However, if it does not and H4ac binds at the promoter, that gene will be active, and if PolII S2 binds at the promoter the gene will also be active. The gene is classified as inactive for other statuses of the protein. The accuracy for 10-fold cross validation was 83.94%.

Next, the profile of all the proteins bound at different genomic features (i.e. promoter, exon, intron, distal intergenic) for all 1261 genes were combined and the data modelled using rpart. The reported tree (Fig. 4) describes the bindings of RNA PolII, H4ac, RNA PolII S2 at promoter regions and RNA PolII at exon. This tree appears similar to that presented in Fig. 3. However, the branching on the right side of the tree indicates that RNA PolII bindings at promoter and exon would classify a gene as active.

We next investigated how time affects both protein binding and gene expression. Note that in this scenario, a gene expression response may occur at a later point, i.e. in response to the observed epigenetic event, rather than simultaneously. For this model, only the promoter feature was selected but we incorporated the profiles of all six proteins at the three time points (i.e. 0H, 1H and 4H respectively). The resulting tree (Fig. 5) indicates that when RNA Polll binds at a promoter at 4H or 1H h, or PolII S2 binds at a promoter at 1H, or H4ac binds at a promoter at the 4H time point, the gene will be classed as active; if not, the gene will be inactive at the 4H time point.

3.3.3. Random forest

The R package randomForest [10] was used to implement the random forest classification method.

Data were prepared as described above with gene classes used as the response variable and the binding profile of the six proteins at different genomic locations as the predictor variables. Fig. 6 summarises the importance of different variables obtained by the random forest method. The prediction models indicated that when only binding at a promoter is considered, RNA PolII, H4ac and RNA PolII S2 are reported to be the most important predictors of gene expression, in terms of mean accuracy and mean Gini. However, when the binding profile of all the variables are aggregated, RNA PolII, H4ac and RNA PolII S2 promoter binding and PolII binding at exon, are selected as the most important features. Again, both Brd4 and CDK9 contribute least to the prediction. These results concur with the features selections reported by the decision tree in the previous section.

3.3.4. Comparative performance of the three classification methods

After evaluating the performance of the individual classifiers, a combinatorial analysis of the variables previously reported as important was performed on the LPS stimulated profile. The combinations used for this analysis included:

1. RNA PolII, RNA PolII S2 and H4ac at promoter (pr);
2. RNA PolII, RNA PolII S2 and H4ac at promoter and RNA PolII at exon (ex);
3. RNA PolII and H4ac at promoter at 4 h time point (4H) and RNA PolII and RNA PolII S2 at promoter at 1 h time point (1H).

Comparing performances of the three classification methods (Table 5) indicated that, for all combinations of variables, the decision tree and neural network methods out-performed the random forest method. Note that decision trees are a popular choice of classification method as the output model is readily interpretable, and in this instance, it is clear which of the genomic variables are most predictive of gene activation.

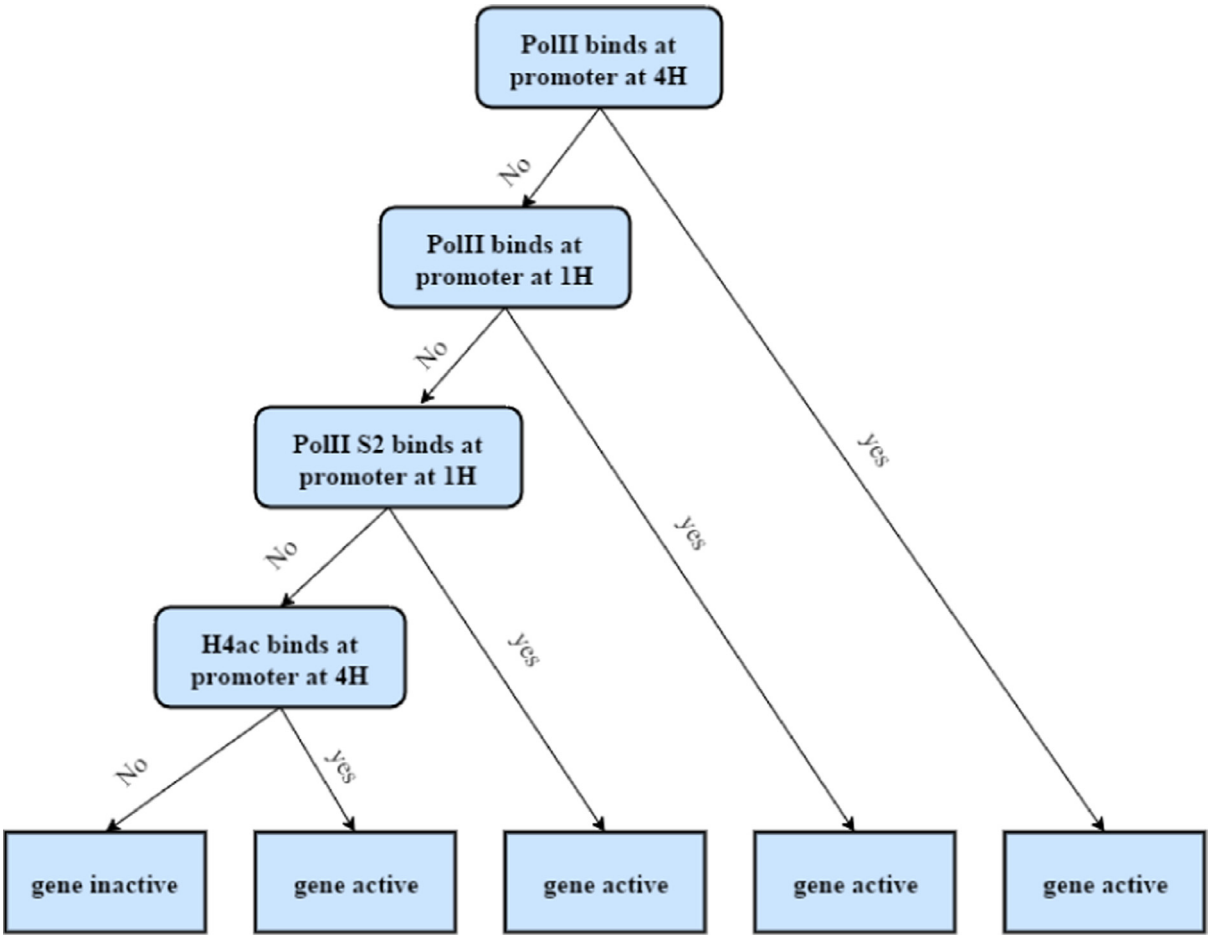

Fig. 5. In this decision tree where leaf nodes represent the class of the genes and the root node and internal nodes represent binding of a protein at a promoter at different time points (0H, 1H and 4H). As for the neural network model, neither Brd4 nor CDK9 were reported as contributing to the prediction.

Table 5
The performances of three different classifiers in terms of 10-fold cross validation accuracy. PolII, PolII S2 and H4ac are reported as the most informative proteins, while the promoter region is the most significant genomic feature. The early time point appears to have little predictive value irrespective of the machine learning method used.

Predictors	Neural network Accuracy (%)	Decision tree	Random forest
PolII_pr + H4ac_pr + PolII_S2_pr	80.02	84.08	78.83
PolII_pr + PolII_ex + H4ac_pr + PolII_S2_pr	82.93	84.75	78.43
PolII_1H + PolII_4H + H4ac_4H + PolII_S2_1H	83.48	84.70	80.41

Furthermore, decision tree models often simplify model interpretation as the classification model automatically selects those features that are important for the prediction and omits those features that are not. In contrast to this, a neural network based model uses all the input features, unless a user manually implements feature selection as part of the data pre-processing.

The random forest method also implements the feature selection steps. For this reason, when the binding profiles of all proteins at all the genomic features and the protein binding profiles at promoters at different time-points have been used for prediction, we have used only decision tree and random forest.

As the random forest method also implements the feature selection steps, a further comparative performance analysis compared decision tree and random forest classification. The analyses completed were:

1. The binding profile of all proteins at promoters;
2. The protein binding profiles of all protein at all the genomic features;
3. The binding profiles of all proteins at promoters at different time-points.

Table 6
The performance of decision tree and random forest in terms of 10-fold cross-validation accuracy. The analysis indicates that the promoter and time predictive model are the most accurate of the combinations evaluated.

Predictors	Decision tree Accuracy (%)	Random forest
Promoter only	83.94	79.38
All genomic features combined	83.80	79.70
Promoter at different time-points	85.01	80.73

The accuracy results (Table 6) obtained indicated that the decision tree classifier out-performed the random forest classifier in each instance.

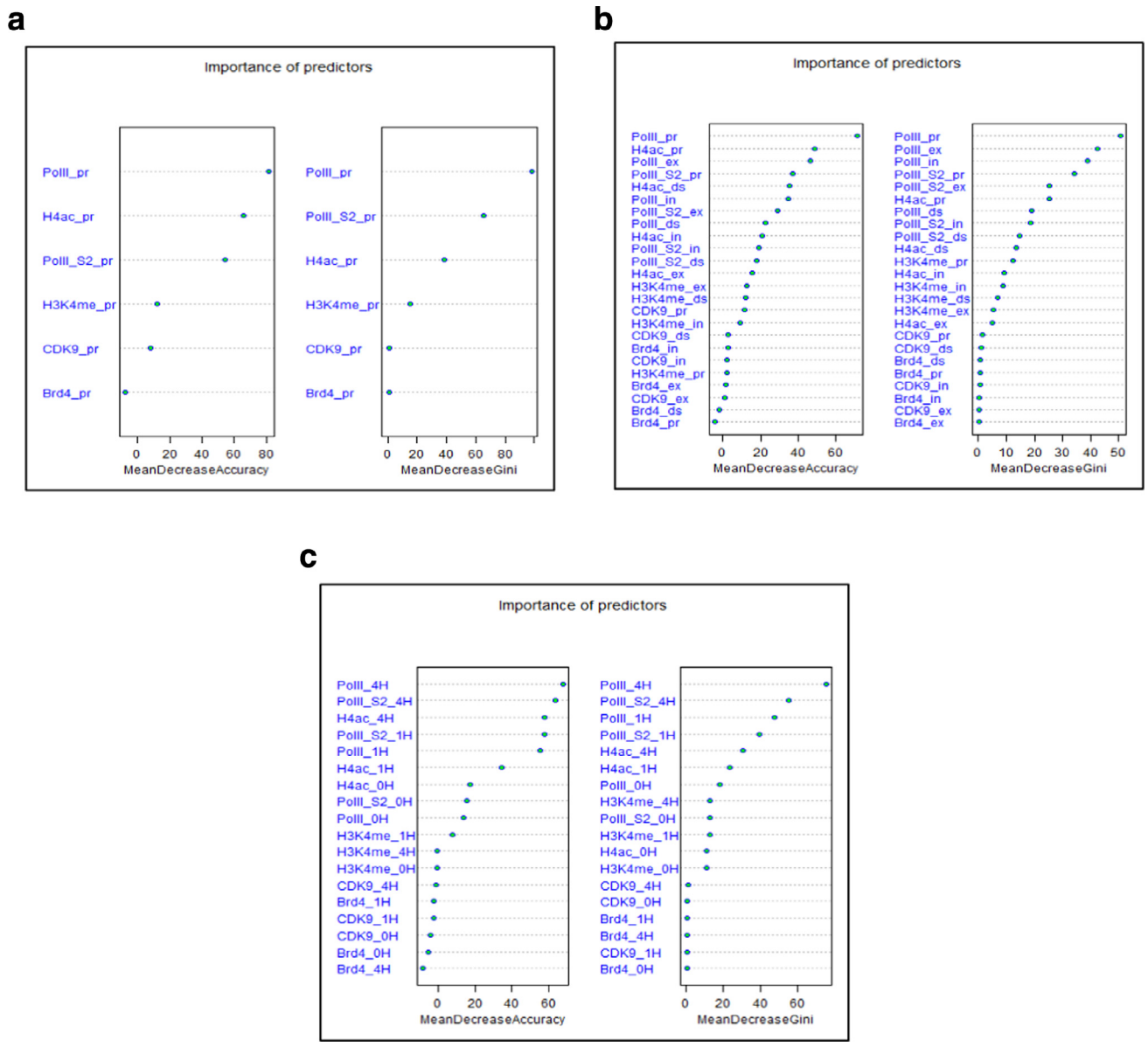

Fig. 6. Importance of variables as detected by a random forest model (a) from protein binding profile at promoter; (b) from protein binding profile at different features, exon (ex), intron (in), promoter (pr), distal intergenic (ds); (c) from protein binding profile at promoter at different time points (0H, 1H and 4H).

4. Conclusion

In this paper, we have demonstrated the application of machine learning techniques to predict gene transcriptional response, using the genomic binding profile of proteins believed to play a role in the regulation of gene expression. The method uses an advanced Markov random field model to detect enriched regions from ChIP-seq data, and adopts advanced machine learning methods for predictive modelling. We successfully applied the method to an integrated dataset comprising a ChIP-seq time-series dataset of six protein markers and the associated gene expression data obtained under several biological conditions.

Our results indicate that the combined binding profiles of several proteins at different genomic features accurately predict variations in gene expression. This combinatorial model concurs with current understanding of transcriptional gene regulation networks which are known to involve multi-factorial mechanisms participating in a huge complex of components and interactions spanning many genomic loci [11,21]. Of the six proteins investigated, RNA PolII, RNA PolII S2, H3K4me3 and H4ac significantly contributed to the accuracy of the expression profile. The individual protein profiles we report as predictive of gene expression have been validated experimentally, for example, RNA PolII is known to bind at promoter regions and then recruit other transcription factors to create a large multiprotein complex that initiates transcription [6,14]. Likewise, the histone proteins have also been confirmed to play several critical roles in transcription. In addition, both the decision tree and random forest models report the later time point profiles as accurate predictors of gene expression which emphasises the temporal aspect of the regulatory mechanism [12]. However, the sparsity of time point data available limits any further understanding of the protein profile responses over shorter and longer time frames.

Our analysis indicates that neither the CDK9 or Brd4 protein profiles were accurate predictors of gene expression. This is somewhat surprising as CDK9 is known to act as an elongation factor for RNA polymerase II-directed transcription, while Brd4 is reported to participate in the core binding of RNA polymerase II. However, this finding agrees with the original analysis of this dataset [13] and most likely reflects the time point used to

collect samples (e.g. responses may occur prior or post the 4-h window).

In summary, we have successfully demonstrated how machine learning techniques can be used to predict gene expression responses using ChIP-seq protein profiles and that the predictive models can be expanded to include additional descriptors of experimental factors. Further advances in our understanding of the complex regulatory mechanisms could be explored if datasets that spanned greater time frames and included additional experimental metrics (e.g. HiSeq long distance chromosome contacts, micro-RNA interactions with nuclear transcription factors, long noncoding RNA binding etc.) were made available.

Acknowledgements

This work was funded by EPSRC (EP/J501864) and GSK (STU100020510). Yanchun Bao was supported by the Economic and Social Research Council (ESRC) through the Research Centre on Micro-Social Change (MiSoC) at the University of Essex (grant number ES/L009153/1).

Conflict of interest

No conflict of interests.

References

[1] Y. Bao, V. Vinciotti, et al., Joint modeling of ChIP-seq data via a Markov random field model, Biostatistics 15 (2) (2014) 296–310.
[2] M. Carlson, B.P. Maintainer, TxDb.Mmusculus.UCSC.mm9.knownGene: Annotation Package for TxDb Object(s), 2015 R package version 3.2.2.
[3] M.J. Dunning, M.L. Smith, M.E. Ritchie, S. Tavaré, beadarray: R classes and methods for Illumina bead-based data, Bioinformatics 23 (August(16)) (2007) 2183–2184.
[4] G. Geeven, R.E. van Kesteren, A.B. Smit, M.C. de Gunst, Identification of context-specific gene regulatory networks with GEMULA–gene expression modeling using LAsso, Bioinformatics 28 (2) (2012) 214–221.
[5] D. Guan, J. Shao, Z. Zhao, P. Wang, et al., PTHGRN: unraveling post-translational hierarchical gene regulatory networks using PPI, ChIP-seq and gene expression data, Nucleic Acids Res. (2014) W130–W136.
[6] S. Hahn, Structure and mechanism of the RNA polymerase II transcription machinery, Nat. Struct. Mol. Biol. 11 (5) (2004) 394–403.
[7] P. Heyn, A.T. Kalinka, P. Tomancak, K.M. Neugebauer, Introns and gene expression: cellular constraints, transcriptional regulation, and evolutionary consequences, BioEssays 37 (2) (2014) 148–154.
[8] P.J. Hurd, C.J. Nelson, Advantages of next-generation sequencing versus the microarray in epigenetic research, Brief Funct. Genomics Proteomics 8 (2009) 174–183.
[9] B. Langmead, C. Trapnell, M. Pop, S.L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, Genome Biol. 10 (2009) R25.
[10] A. Liaw, M. Wiener, Classification and regression by randomForest, R News 2 (3) (2002) 18–22.
[11] J. Majewski, J. Ott, Distribution and characterization of regulatory elements in the human genome, Genome Res. 12 (2002) 1827–1836.
[12] F. Markowetz, K.W. Mulder, E.M. Airoldi, I.R. Lemischka, et al., Mapping dynamic histone acetylation patterns to gene expression in nanog-depleted murine embryonic stem cells, PLoS Comput. Biol. 6 (12) (2010) e1001034.
[13] E. Nicodeme, K.L. Jeffery, U. Schaefer, S. Beinke, Suppression of inflammation by a synthetic histone mimic, Nature 468 (7327) (2010) 1119–1123.
[14] D.B. Nikolov, S.K. Burley, RNA polymerase II transcription initiation: a structural view, PNAS 94 (1) (1997) 15–22.
[15] A. Nott, S.H. Meislin, M.J. Moore, A quantitative analysis of intron effects on mammalian gene expression, RNA 9 (2003) 607–617.
[16] S.G. Park, S. Hannenhalli, S.S. Choi, Conservation in first introns is positively associated with the number of exons within genes and the presence of regulatory epigenetic signals, BMC Genomics 15 (1) (2014) 526. http://doi.org/10.1186/1471-2164-15-526.
[17] J. Qin, M.J. Li, P. Wang, M.Q. Zhang, et al., ChIP-Array: combinatory analysis of ChIP-seq/chip and microarray gene expression data to discover direct/indirect targets of a transcription factor, Nucleic Acids Res. 39 (S2) (2011) W430–W436.
[18] B. Ripley, W. Venables, (2015) Package "nnet". Available at https://cran.r-project.org/web/packages/nnet/index.html (Accessed at January 16, 2016).
[19] G.K. Smyth, Linear models and empirical Bayes methods for assessing differential expression in microarray experiments, Stat. Appl. Genet. Mol. Biol. 3 (1) (2004) Article 3.
[20] G.K. Smyth, Limma: linear models for microarray data, Bioinformatics and Computational Biology Solutions Using R and Bioconductor, R. Springer, New York, 2005, pp. 397–420.
[21] B.V. Steensel, Mapping of genetic and epigenetic regulatory networks using microarrays, Nat. Genet. 37 (2005) S18–S24.
[22] P.B. Talbert, S. Henikoff, Histone variants on the move: substrates for chromatin dynamics, Nat. Rev. Mol. Cell Biol. 18 (2) (2017) 115–126.
[23] T.M. Therneau, B. Atkinson, (2009) Package: rpart Available at http://cran.r-project.org/web/packages/rpart/rpart.pdf (Accessed January 16, 2016).
[24] K. Xu, W. Xiong, et al., Integrating ChIP-sequencing and digital gene expression profiling to identify BRD7 downstream genes and construct their regulating network, Mol. Cell. Biochem. 411 (1–2) (2016) 57–71.
[25] G. Yu, L.G. Wang, Q.Y. He, ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization, Bioinformatics 31 (14) (2015) 2382–2383.

Dr. Mohsina Mahmuda Ferdous received her first degree in computing and IT from Goldsmiths, University of London and obtained her M.Sc. in computer Science from Imperial College London. She got her Ph.D. degree in bioinformatics in 2016 from Brunel University London. Since then she has been working as a postdoctoral researcher at the Centre for Cancer Prevention, Queen Mary, University of London. Her research interest includes discovering biomarkers in different disease states using advanced machine learning techniques and statistical algorithms.

Dr. Yanchun Bao got her Ph.D. at University of Manchester and did three post-doc jobs in University of Manchester, Brunel University and University College of London. She is a research fellow in Institute for Social and Economic Research (ISER), University of Essex. She is interested in bioinformatics research areas and statistical methods for social-health research, including hidden Markov model (HMM) for ChIP-sequence and methylation data, instrumental variables methods for Mendelian randomisation, survival analysis and longitudinal analysis.

Dr. Veronica Vinciotti received the first degree in mathematics from the University of Perugia, Italy, in 1999, and the Ph.D. degree in statistics from Imperial College London, UK, in 2002. Since then, she has worked at Brunel University London, where she is now a reader. Her research interests include high dimensional regularised approaches for regression and classification and network modelling, with applications in finance, health and biology.

Dr. XiaoHui Liu received the B.Eng. degree in computing from Hohai University, Nanjing, China, in 1982 and the Ph.D. degree in computer science from Heriot-Watt University, Edinburg, UK, in 1988. He has been professor of computing at Brunel University since 2000.

Paul Wilson has 16 years' experience as a computational biologist with GlaxoSmithKline. He has supported a number of systems biology focussed initiatives and has expertise with a range of 'omics' analysis methods, including RNA-seq and pathway analyses. He has an avid interest in the perturbation of inflammatory signalling mechanisms, and currently supports a number of epigenetic focussed projects that attempt to integrate NGS data with clinical readouts to identify differentially regulated transcription networks, that may be causative in disease progression.