Exploring the link between gene expression and protein binding by integrating mRNA microarray and ChIP-seq data

Mohsina M Ferdous[1], Veronica Vinciotti[2], Xiaohui Liu[1], and Paul Wilson[3]

[1] Department of Computer Sciences, Brunel University London, Uxbridge UB8 3PH, UK
mohsina.ferdous@brunel.ac.uk
[2] Department of Mathematics, Brunel University London, Uxbridge UB8 3PH, UK
[3] GlaxoSmithKline Medicine Research Centre, Stevenage, SG1 2NY, UK

Abstract. ChIP-sequencing (ChIP-seq) experiments are now routinely used to study genome-wide chromatin marks in epigenetic research. However, due to the high cost and complexity associated with this technology, it is of great interest to investigate whether the results produced by the low-cost option of mRNA microarray experiments can be used in place of ChIP-seq data and what advantages can be achieved if both data sources are combined together. Most comparative or integrated analyses to date do not consider important features of ChIP-seq data, such as spatial dependencies of counts for neighbouring regions of the genome and the different efficiencies of individual ChIP-seq experiments. These, if not accounted for, could lead to misleading results. In this paper, we address these issues by applying a Markov random field model to ChIP-seq data. We then investigate the correlation between the enrichment probabilities around transcription start sites, estimated by the model, and microarray gene expression values. In particular, we focus on the protein Brd4 for which count data from ChIP-seq experiments as well as mRNA microarray data are available at different time points at drug and control conditions. The aim is to elucidate whether the binding of the protein Brd4 at the transcription start site affects the mRNA expression of the associated gene. Our preliminary results suggest that binding of the protein is associated with lower gene expression, however, differential binding across different conditions does not show an association with differential expression of the associated genes.

Keywords: protein binding, gene regulation, Markov random field

1 Introduction

The development and maintenance of any organism is regulated by a set of chemical reactions that switch specific loci of the genome off and on at strategic times and locations. Epigenetics is the study of these reactions that control gene expression levels and the factors that influence them. Although the relationship

between epigenetics and phenotypes is not always straightforward, studying tissues of affected and unaffected subjects and maintaining the study prospective may help identify the differences between causal associations and non-causal associations [1]. DNA microarray and ChIP-seq technologies play a crucial role in genome research understanding this relationship, by investigating structural and functional characteristics of genomes. DNA microarray technology, which enable measurement of expression level of a large number of genes simultaneously, has been used in functional genomic studies, system biology, epigenetic research and so on. ChIP-seq, which is a comparatively new technology, has been used to describe the locations of histone post-translational modifications and DNA methylation genome-wide in many studies and to study alterations of chromatin structure which influence gene expression levels.

Next generation sequencing has undoubtedly several advantages over microarray experiments and it is often the choice for many studies. However, microarray experiments still have a place in bioinformatics, due to the cost-effectiveness and relative simplicity of this technique [2]. Hurd *et al.* [3] has predicted that in the near future, these two technologies may also complement each other and form a symbolic relationship. Integration of results from these two technologies might open new doors for epigenetic research.

Several attempts have been made to combine protein binding and mRNA expression data over the years. Markowetz *et al.* [4] have explored how histone acetylation around Transcription Start Sites (TSSs) correlates with gene expression data. In their study, ChIP-ChIP is used for measuring acetylation levels. Qin *et al.* and Guan *et al.* [5, 6] have proposed a web-based server to analyse interactions between transcription factors and their effect on gene expression, by using information on bound and non-bound regions. Other attempts have also been made to infer relationships between gene expression and histone modification where absolute tag counts around a feature, such as promoter, is considered. Hoang *et al.* [7] has shown how, incorporating the spatial distribution of enrichment in the analysis, can improve the result. In general, it is absolutely vital to measure the level of acetylation and probability of enrichment accurately in order to find possible relationships between ChIP-seq and gene expression data. There are several characteristics of ChIP-seq data that are needed to be considered while modelling such data before we attempt to combine it with gene expression data.

In a typical ChIP-seq experiment, an antibody is used in the immunoprecipitation step to isolate specific DNA fragments that are in direct physical contact with a protein of interest. Figure 1 [14] gives an overview of how ChIP-seq technology works. Those fragments are called reads/tags. The reads are then mapped back to the reference genome and the resulting mapped reads are further analyzed to find out peaks or enriched regions where the protein in question is actually bound. It is common to divide the genome into fixed sized windows/bins and then summarize the counts per bin. Finally, a statistical model is used to detect the windows with a significant number of counts, that is the regions that are bound by the protein in question. While generating the data, some random

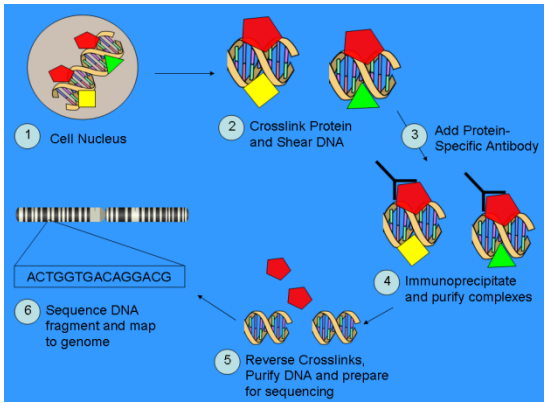

Fig. 1. Schematic representation of ChIP-seq technology

DNA sequences are also collected with the bound sequences. These are usually scattered across the genome and form a background noise. Due to the particular antibody used and to the difficult protocol that each experiment needs to follow, it is common to observe varying background to signal ratios for different experiments. This poses an issue when multiple ChIP-seq experiments need to be modelled together and when comparative analyses with other data sources need to be carried out. Bao *et al.* [9] have proposed a mixture model where multiple experiments can be modelled together while taking into account the efficiency of individual experiments. However, there are other issues related to ChIP-seq data. Due to an often ad-hoc division of the genome in fixed-size windows, it is possible for an enrichment profile to cross neighbouring regions. This induces spatial dependencies in the count data, which is often observed for ChIP-seq data. All these issues are addressed in the approach proposed by Bao et al. [8]. In this proposed approach, a Markov random Field (MRF) model has been implemented that accounts for the spatial dependencies in ChIP-seq data as well as the different ChIP-seq efficiencies of individual experiments. In this paper, we have adapted this model for the analyse ChIP-seq data for Brd4 protein.
Investigating enrichment around a feature in the genome such as promoter, TSS etc is very common while studying relationships between binding of a protein/TF and gene regulation. TSS is where transcription of the genes into RNA begins, therefore it is often considered in comparative analyses of binding and expression data. After analysing the ChIP-seq data using the MRF model, we have used the estimated probability of enrichment around the transcription start (TS) and performed comparative analysis on the associated gene expression data generated in the same biological condition.
In Section 2, we describe the data that has been used for this paper. We also give a brief overview of the MRF model for ChIP-seq data and how the parameters are estimated, as well as the differential expression analysis of the microarray

data. In Section 3, we show our current results in comparing ChIP-seq and microarray data. Finally, we draw some conclusions in Section 4.

2 Data and Methods

2.1 Description of the Data

In this study, we have used the ChIP-seq data for the Brd4 protein provided by Nicodeme *et al.* [11]. Illumina beadarray technology was also used to collect gene expression data on the same experimental conditions as ChIP-Seq. The data was collected from samples that are treated with a synthetic compound (I-BET) that, by 'mimicking' acetylated histones, disrupts chromatin complexes responsible for the expression of key inflammatory genes in activated macrophages (Drug data) and also from sample simulated with lipopolysaccharide (LPS) (control data). The ChIP-seq data was collected at three time points: 0, 1 and 4 hours (0H, 1H, 4H) and microarray data at four time points (0H, 1H, 2H and 4H). For the ChIP-seq data, one replicate is available for each condition, whereas three replicates per condition are available in the microarray study.

2.2 Analysis of ChIP-seq data

The ChIP-seq reads are aligned against the mouse genome (version mm9) using bowtie [10] and only uniquely mapped reads were retained for further analysis. The reference genome was obtained from UCSC Genome Browser. The percentage of reads that are aligned ranges from 60.86% to 78.03%. In this experiment, for simplicity, we have considered only Chromosome 1. So, we have selected only those reads that are found in Chromosome 1 of the mouse genome. We have divided the length of Chromosome 1 into 200bp windows and generated count data per windows. These count data are then supplied as the input for MRF model, described in the next section.

2.3 A brief description of MRF model

We have followed the methodology proposed by Bao *et al.* [8] for the analysis of ChIP-seq data. Given the data, the model associates to each window a probability of being enriched or not. Additional information such as enrichment information of neighbouring regions is also considered while calculating this probability. A brief overview of the model is given below.
Let M be the total number of bins and Y_{mcr} the counts in the mth bin, $m = 1, 2, \ldots, M$, under condition c and replicate r. In our case, the condition c stands for a particular protein and/or a particular time point, and $r = 1, \ldots, R_c$ is the number of replicates under condition c. The counts Y_{mcr} are either drawn from a background population (non-enriched region) or a from a signal population (enriched region). Let X_{mc} be the unobserved random variable specifying if the

mth bin is enriched ($X_{mc} = 1$) or non-enriched ($X_{mc} = 0$) under condition c. A mixture model for Y_{mcr} is defined as follows [9]:

$$Y_{mcr} \sim p_c f(y|\theta^S_{cr}) + (1 - p_c) f(y|\theta^B_{cr}),$$

where $p_c = P(X_{mc} = 1)$ is the mixture portion of the signal component and $f(y, \theta^S_{cr})$ and $f(y, \theta^B_{cr})$ are the signal and background densities for condition c and replicate r, respectively. An attractive feature of this model is the fact that the probability p_c of a region being enriched does not depend on ChIP efficiencies. However the parameters, signal and background distributions θ^S_{cr} and θ^B_{cr} depend on ChIP efficiencies of replicates r. This allows to combine multiple ChIP-seq experiments, while accounting for the individual ChIP efficiencies.

As the signal and background densities can take any form, the signal can be modelled using Poisson or Negative Binomial and their zero-inflated extensions to account for the excess number of zeros typical of this type of data. So for the mixture components $f(y, \theta^S_{cr})$ and $f(y, \theta^B_{cr})$, we consider:

$$Y_{mc}|X_{mc} = 0 \sim ZIP(\pi_c, \lambda_{0c}) \quad \texttt{or} \quad ZINB(\pi_c, \mu_{0c}, \phi_{0c}),$$

$$Y_{mc}|X_{mc} = 1 \sim Poisson(\lambda_{1c}) \quad \texttt{or} \quad NB(\mu_{1c}, \phi_{1c})$$

In our study, we have used zero inflated negative Binomial for modelling the background and Negative binomial for modelling the signal for all our ChIP-seq datasets.

In order to account for spatial dependencies, the latent variable X_{mc}, which represents the binding profile, is further assumed to satisfy one-dimensional first order Markov properties. Given the adjacent bins states, $X_{m-1}, c = i$, and $c = j$, with $i, j \in \{0, 1\}$

$$Y_{mcr}|X_{m-1,c} = i, X_{m+1,c} = j \sim p_{c,ij} f(y, \theta^S_{cr}) + (1 - p_{c,ij}) f(y, \theta^B_{cr})$$

Thus, the enrichment of a region depends on the state of the two adjacent regions. All the parameters in this model are estimated using a Bayesian approach, which is implemented in the R package `enRich`. The method returns the posterior probability of enrichment for each region of the genome.

Finally to decide whether a region is enriched or not, a threshold is set on these probabilities. Different criteria can be used to set this cut-off. In our study, we set a cut-off corresponding to a chosen 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

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

In our study, we used this approach for all 200bp regions in Chromosome 1. We then further refine the output to only consider the regions that contain TSs.

2.4 Analysis of Microarray data

Microarray data have been preprocessed using the R package `beadarray` [12]. Then the processed data has been normalised and analysed for differential expression using the package `limma` [13]. This returns an adjusted p-value for differential expression between drug and control using an empirical Bayes method. We use these p-values to select the differentially expressed genes.

2.5 TSS selection

We have downloaded TSS information of the mouse genome (chromosome 1) using NCBI mm9 assembly. Each txStart (Transcription start) and txEnd (Transcription end) coordinates are then linked with the associated genes. Many genes have several TSSs, and also some txStarts are at the same co-ordinate and others may reside within 200 bp regions to each other. Firstly, we remove the duplicate TSSs from the list. As we select enrichment probability within regions of 200bp, for each gene we select only one TSS within this window. From UCSC we downloaded 55419 TSSs and retained 38156 after this selection. As we consider only transcription start point for this experiment, we then retrieve the estimated probability of enrichment from the ChIP-seq analysis per TS.

3 Results and Discussion

3.1 ChIP-seq analysis

We have analysed the ChIP-seq data with both the latent mixture model and the MRF model. For each condition, Table 1 shows the number of regions bound by Bdr4 at 5% FDR. The efficiency for each experiment estimated by the model is also given in the fourth column.

Table 1. comparison of mixture model and MRF model in terms of number of regions bound by Brd4 at 5% FDR

Conditions	MRF model	Mixture model	IP efficiency
0H control	3394	1475	0.8201
0H drug	3185	930	0.8501
1H control	3161	614	0.8937
1H drug	3265	926	0.8937
4H control	3354	1345	0.8347
4H drug	2810	281	0.7809

At 5% FDR, the MRF model produces more enriched regions for each condition than the mixture model. By inspection of the regions detected by MRF

but not by the mixture model, we have found out that MRF can assign a high probability to a region that has relatively low tag counts but has neighbouring regions with a large number of counts, as it incorporates spatial dependency in the model. On the other hand, the mixture model will assign a very low enrichment probability to those regions, thus discarding potentially useful information.

3.2 Expression data versus enrichment probability

Nicodeme *et al.* [11] suggests that the impact of the drug I-BET on LPS-inducible gene expression is highly selective and it has no effect on the expression of housekeeping genes. Our key interest has been to investigate whether differential binding or differential enrichment of the protein Brd4 around TSS between drug and control data is associated with differential expression of the corresponding genes.

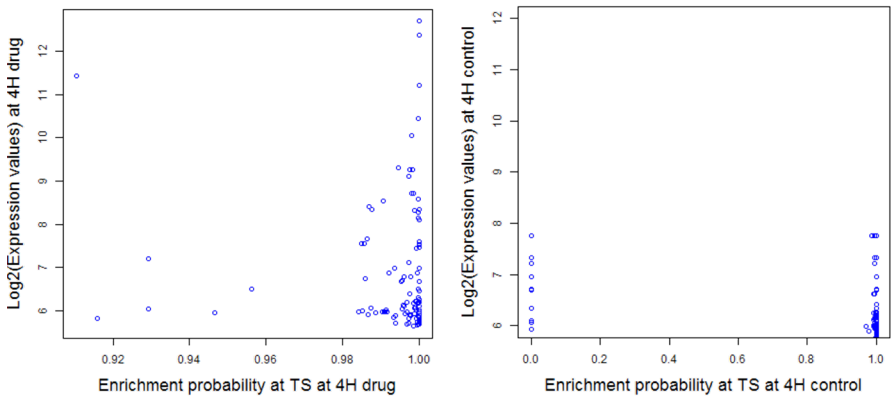

Fig. 2. Investing correlation between differential binding with differential expression result. (Left) At time point 4H and for the drug condition, TS regions with very high probabilities of enrichment are plotted versus the corresponding gene expression (log2) values. (Right) The low expressed genes found in the left plot are investigated in control data to check the effect of differential binding on regulation of genes.

At time point 4H, we select the TSs that have high probabilities of enrichment in the drug condition (at 5% FDR) and isolate 115 regions. The left plot in Figure 2 shows the gene expression values (in the log scale) versus the probabilities of enrichment for these regions. The plot shows a cluster of 84 genes in the bottom right corner that have very low expression values (below the median overall expression of 6.22). This was observed also at different time points. To find out whether binding of Brd4 in those regions play any role in the down-regulation of genes, we consider the binding and expression of these genes on the control data.

The Right plot in Figure 2 shows that there is a small number of non-bound regions. However, these genes do not have a significantly higher expression value than in the drug samples. Thus, in this study, we found that differential bindings did not play a direct role in down-regulation of genes between drug and control experiments.

To investigate whether differential acetylation levels is associated with differential expression, we have selected differentially expressed genes at 4H between drug and control with a 1% cutoff on the adjusted p-values. We have subtracted expression values (log2) of control data from drug data (i.e. taking log ratios) and have done the same with enrichment probabilities. In Figure 3, Left plot shows the differential expression versus differential probability. Overall, few changes are found in the probabilities of enrichment between different conditions, suggesting that the genes are bound or not bound in both conditions and that the probabilities are either close to 0 or close to 1. Therefore, the plot does not show any association between differential probabilities of enrichment and differential expression. However, we are considering using different measures of acetylation levels than posterior probabilities. Similar results were obtained when comparing two time points (1H and 4H, respectively), as shown in the right plot of Figure 3. Here there are some regions with different probabilities of enrichment, but with no associated down or up regulation.

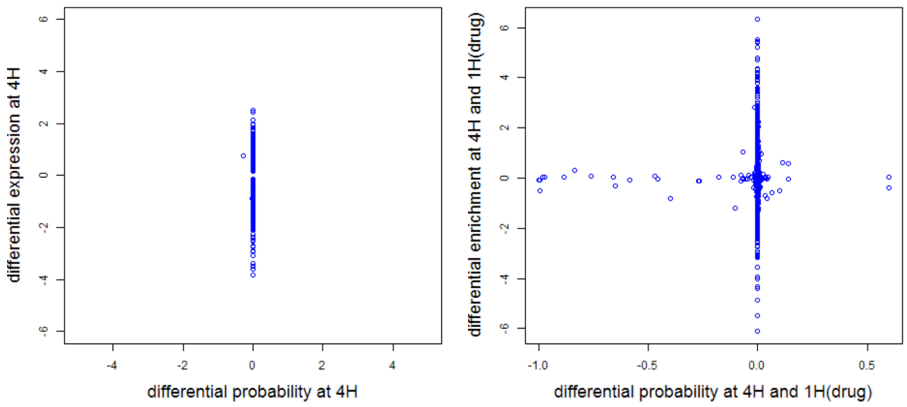

Fig. 3. Investing correlation between differential probability with differential expression result. (Left) Plot for differential probability versus differential expression(log2) between drug and control data at 4H. (Right) Plot for differential probability versus differential expression(log2) between 4H and 1H for drug data.

4 Conclusion

In this study, we have investigated a possible association between gene expression and protein binding data, from mRNA microarray and ChIP-seq data respectively. We have emphasized the need to account for important features of ChIP-seq data in the detection of the enriched regions and have therefore opted for the use of a Markov random field model for the analysis of ChIP-seq data. Our results show that protein binding is associated with lower expression values, but that differential binding between different conditions, e.g. drug and control or different time points, is not associated with up or down regulation of the corresponding genes.

A number of steps will be considered in the future to consolidate the research in this study. Firstly, we will extend the analysis from Chromosome 1 to the whole genome, to check whether the results generalise to this case, as well as to different proteins (the data in [11] is for five proteins). Secondly, we will consider different ways of measuring acetylation levels, while still considering the issue of different ChIP efficiencies of individual experiments. Finally, we will consider other chromatin markers, such as promoters, to explore a possible association between these and gene regulation, as well as possible combinatorial patterns between chromatin markers and gene regulation.

References

1. Petronis A, (2010) Epigenetics as a unifying principle in the aetiology of complex traits and diseases. *Nature* 465, 721727
2. Xiao H, et al. (2009) Perspectives of DNA microarray and next-generation DNA sequencing technologies. *Science in China Series C: Life Sciences*, Volume 52, Issue 1, pp 7-16
3. Hurd PJ, et al. (2009) Advantages of next-generation sequencing versus the microarray in epigenetic research. *Brief Funct Genomics Proteomics* 8: 174-183
4. Markowetz F, et al. (2010) Mapping Dynamic Histone Acetylation Patterns to Gene Expression in Nanog-Depleted Murine Embryonic Stem Cells. *PLoS Comput Biol* 6(12): e1001034
5. Qin J, et al. (2011) ChIP-Array: combinatory analysis of ChIP-seq/chip and microarray gene expression data to discover direct/indirect targets of a transcription factor. *Nucl. Acids Res*
6. Guan D, et al. (2014) PTHGRN: unraveling post-translational hierarchical gene regulatory networks using PPI, ChIP-seq and gene expression data. *Nucl. Acids Res*
7. Hoang SA, et al. (2011) Quantification of histone modification ChIP-seq enrichment for data mining and machine learning applications. *BMC Research Notes*, 4:288
8. Bao Y, et al (2014) Joint modeling of ChIP-seq data via a Markov random field model. Biostat (15 (2): 296-310.
9. Bao Y, et al (2013) Accounting for immunoprecipitation efficiencies in the statistical analysis of ChIP-seq data. *BMC Bioinformatics*, 14:169
10. Langmead B, et al. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, 10, R25.
11. Nicodeme E, et al (2010) Suppression of inflammation by a synthetic histone mimic. *Nature* 23;468(7327):1119-23.

12. Dunning MJ, et al. (2007) beadarray: R classes and methods for Illumina bead-based data. *Bioinformatics*, 23(16):2183-4
13. Smyth GK (2005). Limma: linear models for microarray data. In Gentleman R, Carey V, Dudoit S, Irizarry R and Huber W (eds.), *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, pp. 397420. Springer, New York.
14. wikipedia entry. available at www.wikipedia.org