Implementation of a workflow for the processing and analysis of genome interaction datasets

Anandashankar Anil
anandashankar.anil@ist.utl.pt

Instituto Superior Técnico, Lisboa, Portugal

December 2016

Abstract

Enhancers and their interactions with specific promoters play an important role in gene transcription and consequently, phenotype expression. The distal regulation of genes by enhancers has recently been identified to play a role in diseases such as cancer with almost 80% of disease-associated sequence variants located within enhancers. With the HiCap method, which combines a 4-cutter restriction enzyme Hi-C with sequence capture of promoter regions, promoter-anchored enhancer interactions can be easily identified. Biases inherent in Hi-C may carry over into HiCap but bears investigation as HiCap has a higher selectivity and resolution. Also, structural and functional interactions are not differentiated in the HiCap output. In this project, the HiCap output from a line of THP-1 cells with Lipopolysaccharide stimulation was evaluated for inherent biases. Certain Machine Learning tools were used to try to differentiate between structural and functional interactions in HiCap output using datasets from ChIP-Seq studies on enhancers as reference. It was found that the biases in Hi-C were not carried over to the HiCap output. The results from the machine learning techniques suggest that more data parameters may be required to definitively distinguish structural and functional interactions.
Keywords: Enhancers, HiCap, Bias identification, Machine Learning, One Class Classification, Gaussian Mixture Models

1. Introduction

The complete sequencing of the Human genome in 2003 led to an explosion of genomic data and techniques to process and extract information from this data which ushered in a new era in genetics and genomics called the genomic era[4]. The advances in genomics has helped with the rapid identification of newly discovered pathogens, gene-expression profiling to assess risk of disease and guide therapy, improve understanding of the role of specific genes in the causation of common conditions and so on. A part of studies in genomics focuses on the transcriptional control of genes. The type and amount of ribonucleic acid (RNA) transcribed from genes control the phenotypic expression of the cell. Transcriptional regulation thus becomes important in studying disease phenotypes. A kind of regulatory element is cis-regulatory modules called enhancers. These short deoxyribonucleic acid(DNA) segments, which may be situated many thousands of bases away from the genes they act on, can boost the transcription from the promoter of the target gene to a great degree[16]. The number of enhancers in eukaryotic genomes correlates with the complexity of the organism.

HiCap is a technique that has been formulated recently in[15] which is based on Hi-C and consequently on Chromosome Conformation Capture (3C), which can be considered as a part of the advances made in the genomic era. HiCap selects for promoter sequences and generates genome-wide maps of chromatin interactions where one of the interactors is a promoter sequence. It can generate fragments short enough for single-enhancer resolution[15]. Due to the novelty of the technique, a lot of scope in the interpretation of the data generated by it exists. This thesis explores some of the facets of the HiCap output including biases which may be inherited from Hi-C and the question of classifying different types of chromatin interactions captured by HiCap.

This paper will investigate how clustering with gaussian mixture models(GMM) and classification with One Class Support Vector Machines (SVM) performs in distinguishing structural and functional interactions in the HiCap output from a line of THP-1 cells with LPS stimulation. Different variables inherent in the HiCap data were used as parameters in the machine learning techniques. The variables used are not exhaustive and further vari-

ables could be included in the future. The investigation is limited by the quality of the enhancer information in the reference datasets as well as by the assumptions made about data in HiCap output, like that it follows a Gaussian distribution. Biases in Hi-C as evaluated in [21] will also investigated if they exist in HiCap data generated from the line of THP-1 cells with LPS stimulation.

2. Background

The human genome contains various regulatory sequences such as enhancers, promoters and untranslated regions are considered gene associated. The promoter sequence is where the RNA polymerase binds to begin transcription of the gene, so by definition is situated close to the transcription start sites (TSS) of genes. Enhancers, which are also called activators[7] or cis-regulatory modules[18] are a kind of genomic regulatory element which influences the intensity of genomic expression. They are distinct regions in the genome which contain binding site sequences for transcription factors (TFs). Transcription factors are proteins that bind to specific DNA sequence motifs. By forming complexes with TFs and gene promoters, enhancers can up or down regulate the transcription of a target gene[18]. Enhancers can be located at any distance from the target gene in the linear DNA sequence and come into spatial proximity by the looping of chromatin[18]. Enhancers can be found both upstream and downstream of their target genes and may regulate multiple genes[12]. Multiple enhancers may also regulate the same gene. The location and spatio-temporal activities of most of the enhancers are either not known with confidence or are unknown[11]. Also, predicting enhancers and their activity states from their DNA sequences is difficult as the TF binding motifs may be many and varied[18]. The activity of enhancers are also influenced by the openness of chromatin, i.e. chromatin where nucleosomal histone proteins are modified (notably by monomethylation of lysine 4 and acetylation of lysine 27 in histone H3)[16]. As recent studies have confirmed a role of mutations in distant cis-regulatory elements underlying various human diseases[12], the importance of identifying enhancers and their target genes can be understood. Out of the several computational and experimental approaches that have been developed to determine enhancer elements[18], a few of the methods used capture them by taking a snapshot of the spatial organization of the chromosomes in the nucleus.

2.1. HiCap

HiCap is an extension of Hi-C, which is based on Chromosome conformation capture (3C), which is a technique used to detect the spatial organization of chromosomal DNA[2]. It substituted a 4-cutter restriction enzyme instead of the 6-cutter usually used in Hi-C and introduced the sequence capture of promoter regions[15]. Similar to Hi-C, HiCap also generates genome-wide maps of chromatin interactions with the added functionality of selecting for promoter-anchored interactions. The mean fragment size in HiCap is around 699 bp which gives it close to single-enhancer resolution. By fixing one interaction partner through sequence capture of promoter regions, HiCap has a higher sensitivity than Hi-C. This means that HiCap gives higher sensitivity with lower sequencing depth [15]. This method follows the same steps as Hi-C till ligation and purification of fragments using the beads. Then labelled capture probes are added to further selectively purify the hybridised fragments. The fragments captured by hybridisation are then analysed and identified. These steps are illustrated in Figure 1.

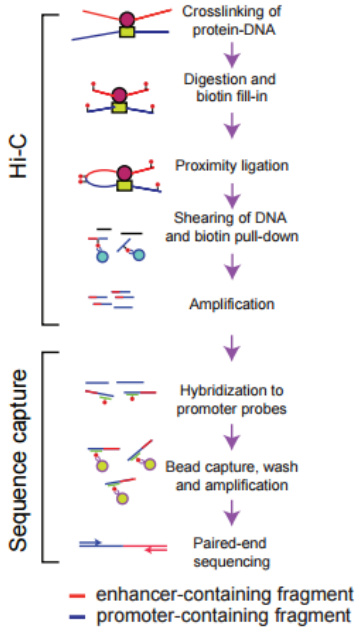

Figure 1: Simplified overview of HiCap.

Studies have found various systematic biases in Hi-C read counts[21, 5]. These include biases due to GC content of fragment ends, distance between restriction enzyme cut sites and uniqueness/mappability of fragment ends of short sequence reads. These results were obtained with a six-cutter restriction enzyme in the genome cutting step of Hi-C[21]. It needs to be examined if the same biases are carried over unchanged into HiCap as it uses a four cutter restriction enzyme[15]. Another difference between Hi-C and HiCap is the targeted capturing of promoter sequences.

An interaction in HiCap output is a ligated pair of two fragments. One fragment contains the pro-

moter and is selected for by a probe designed for it and the other fragment contains a sequence which putatively interacts with the promoter on the first fragment. A particular pair of fragments ligate only if they are in spatial proximity at the time of DNA crosslinking. As this spatial proximity may not mean that the two fragments actually interact, the interactions called in HiCap can be divided into 2 cases - Functional and Structural interactions. Functional interactions are those in which the two fragments on the ligated pair actually interact. Structural interactions are those in which the two fragments on the ligated pair do not interact and were simply spatially adjacent at the time of DNA cross linking. Discriminating between structural and functional interactions based on just the number of supporting pairs for the interactions, as is currently done, might lead to the exclusion of functional interactions which have a low number of supporting pairs. As the DNA crosslinking captures a temporal snapshot of the cell(s) in the experiment, certain structural interaction might have supporting pairs higher than an arbitrarily fixed threshold and lead to the exclusion of functional interactions which have a low number of supporting pairs.

3. Implementation

The workflow proposed for the processing of data and subsequent bias evaluation and model generation using machine learning algorithms is as illustrated in Figure 2.

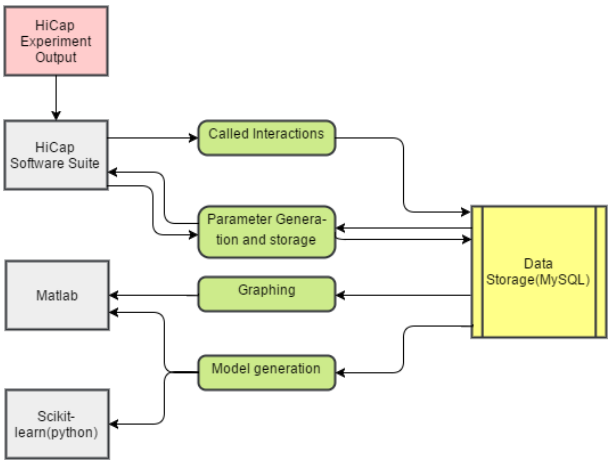

Figure 2: The proposed workflow for processing data from HiCap experimental output and subsequent bias evaluation and model generation using machine learning algorithms.

The input is the experimental output that is obtained from the HiCap method. This HiCap output is then passed to the HiCap Software Suite modules which call the interactions, parse and store them in the data repository. The various further paramaters needed for bias evaluation and model generation are then generated using the requisite modules of the HiCap Software Suite and stored in the data repository. The plots and models are then generated using MATLAB and scikit-learn by drawing the data from the repository. The various parts of the workflow are explained below.

3.1. Data and Database Design

There are two different kind of datasets used in this project. First is the input dataset obtained from the HiCap procedure on a THP-1 cell line and calling the resulting interactions using the HiCap interaction-calling developed in association to [15]. The second dataset is a combined reference dataset of ChIP-Seq results of the enhancer landscape from various published studies in [14, 13, 6, 3, 19, 20].

THP-1 is a human leukemia monocytic cell line, and is a common model to estimate modulation of monocyte and macrophage activities[1]. The input dataset contains data from two experiments with THP-1 cell lines - one in which the cells were stimulated with inflammatory lipopolysaccarides(LPS) and the other which was not stimulated with LPS. LPS evokes strong immune responses in animal cells and is usually found in the outer membrane of gram-negative bacteria.

In each experiment, a number of interactions are 'called'(or taken to be true interactions) depending on whether the interaction had at least three supporting read pairs in each biological replicate[15]. An interaction is a ligated pair of probe-selected promoter fragment and a distal interacting fragment. Each interaction has different measured parameters which are as given below.

1. Probe related data: A unique identifier for each probe designed against promoters in the experiment. Each probe would therefore have an associated gene, chromosome number and other chromosome related information.

2. Interactor Chromosome related data: Data on which chromosome the interacting distal region(the putative enhancer) is situated and the start and end positions of the fragment on the chromosome.

3. Interactor Expression data: Data related to interactor expression which are as below

   FPKM: FPKM (or Fragments per kilobase of exon per million fragments mapped) value for the interaction.
   Distance: The distance between the restriction enzyme fragments involved in the interaction which is the distance from the promoter with which the distal region is interacting. The distance can be a positive or negative value de-

pending on whether the distal region is situated upstream or downstream of the promoter.
Number of Supporting Pairs: The number of supporting pairs in the which support that particular interaction.
p-value: The p-value based on background frequency of the observed interaction.
Strand Combination: The number of supporting pairs for each combination in which ligation of the two fragments in the interaction can take place. Ligation can occur in the forward-forward, forward-reverse, reverse-forward or reverse-reverse manner. This parameter is presented in the '$x\_y\_z\_a$' format where x, y, z and a represent the number of supporting pairs for each combination respectively. This parameter gives a measure of entropy in the interaction.

The reference dataset contains called peak information from ChIP-Seq, which includes the chromosome number, start and end positions on the chromosome in the BED format.

The total number of interactions in the HiCap output for THP-1 with LPS stimulation dataset number a total of 9,264,115. This dataset was then filtered using the number of supporting pairs higher than 3 and the p-value for the interactions lesser than 0.05 as filters down to 809,520 interactions. Out of these filtered interactions, 593,781 overlapped with ChIP-Seq peaks in the reference dataset, which are all assumed to be functional interactions. The rest of the non-overlapping interactions numbering 215,739 may contain a mixture of structural and functional interactions. The training and test set each contain 269,840 interactions which is composed of 197,927 reference overlapping interactions and 71,913 interactions which do not.

The database schema was designed as shown in Figure 3 to store the data used in the work. The Probe table includes all the Probe related information. A unique integer identifier for each probe called probeID was used as index to the table.

The Interactor table stores all the information related to the distal interacting fragments that is common to both experiments - with LPS and without. The probeID from the Probe table is used to connect which Probe is connected to which distal fragment. The InteractWLPS and InteractNLPS tables stores the interacting fragment information specific to the experiments with LPS and without respectively. The unique interactID identifier is used to connect these tables to the main Interactor table. The ChipseqOverlap table stores the information about the distal interacting fragments which overlap with ChIP-Seq peaks from the reference dataset. The flags isTraining, isTest and isValid indicate whether the data is used in Training sets, Test Sets or Validation sets respectively. The FilteredInteract table stores the information about the distal interacting fragments which do not overlap with ChIP-Seq peaks from the reference dataset. The flags isTraining, isTest and isValid indicate whether the data is used in Training sets, Test Sets or Validation sets respectively. The information in the fields GCContent, RECount, mappability and repeatOverlaps in the Probe and Interactor tables were generated from the respective start and end positions of the chromosomes as part of this project.

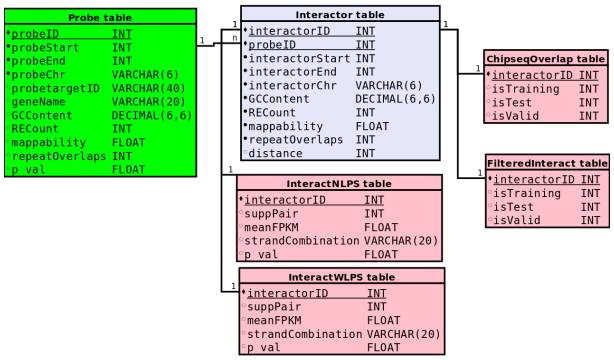

Figure 3: Database Schema implemented in a relational database system used to store the Probe and Interactor information to be used in this project.

3.2. Methodology

The objectives of this paper were two-fold:

1. to check whether the biases from the Hi-C method, on which HiCap is based, is carried over to HiCap.

2. to find a way to improve the calling of interactions and see if structural interactions can be differentiated from functional interactions.

3.2.1. Objective 1: Discovering Biases

Hi-C has number of documented systematic biases including biases due to GC Content, distance between restriction enzyme cut sites, mappability of fragments. The repeat overlap bias of fragments is a measure similar to mappability that was evaluated in the HiCap output data. The repeat overlap is something that arises when a sequence pattern repeats in two different distal fragments which do not interact with the same promoter sequence. This means that there is an ambiguity in mapping the fragments uniquely to a promoter.

For finding the GC content bias, a normalisation process was undertaken. The percentage GC content of probe fragments was binned, normalised by dividing with the total number of probes and then plotted. The probes themselves are only 120 bp long, so the restriction fragments which the probes select for were chosen. The number of interactions

with a certain number of supporting pairs falling into specific percentage GC bins divided by the total number of interactions with that number of supporting pairs was also plotted. The equation used in the normalisation process is as shown in Equation 1.

$$\begin{aligned} x &= a/b \\ y &= c/d \end{aligned} \tag{1}$$

where n is a percentage GC bin range, a denotes the number of probes with GC_n, b the total number of probes, c is the number of interactions with Supporting pair =(1, 2, 3, 4, ...) and GC_n and d is the total number of Interactions with Supporting pair =(1, 2, 3, 4, ...).

For the Restriction enzyme cut sites, the chromosome sequence around a 10 kilobase region of the distal interacting fragments were searched for the recognition site of the 4-cutter restriction enzyme used in HiCap(DpnII).

3.2.2. Objective 2: Improving Interaction Calling

A way to discriminate between structural and functional interactions could be to use more parameters than just the number of supporting pairs of each interaction. From the experiment, the distance, FPKM, p value and strand combination can also be included as parameters to decide whether an interaction is structural or functional.

A way to find patterns in the interactions is to use machine learning techniques. From the nature of the data, two approaches can be used. As a sample of what structural or functional interactions does not exist, the first approach is to use an unsupervised clustering technique that uses all the parameters as input. The second approach is to use the reference dataset of enhancers identified from ChIP-Seq peak studies as a model to verify what actual functional interactions look like. The reference dataset is intersected with the input dataset which yields a subset of interactions in the input dataset which are assumed to be functional interactions. As the rest of the interactions may be of either type, a negative class of structural interactions can not be defined. This means that the conventional supervised binary classification techniques can not be used. In this case, a one class classification technique can be used in which a model is constructed where only the target class is defined. The rest of the data is classified as either of the target class or not.

Strand Combination to interaction entropy As the parameter 'strand combination' cannot be used as such in the input of algorithms of either of the approaches, a new derived parameter called 'interaction entropy' was defined. This was based on the 'tissue specificity index' as defined in [22] and expanded in [9]. The value of interaction entropy must give a sense of whether an interaction prefers a specific strand combination. The equation for interaction entropy is defined as given in Equation 2

$$e = \frac{\sum_{j=1}^{4}(1 - [(c_j)/(c_{max})])}{3} \tag{2}$$

where j is the specific value of strand combination, c_j denotes the number of supporting pairs of the j^{th} strand combination and c_{max} is the maximum value of supporting pairs in a strand combination.

Thus the value of interaction entropy ranges between 0 and 1. The closer the entropy value of an interaction is to 1, the likelier is it to favour a particular strand combination. If the value is 1, the interaction has supporting pairs of only one of forward-forward, forward-reverse, reverse-forward or reverse-reverse combinations. If the value of entropy is 0, the number of supporting pairs for all the four combinations are equal.

Tools Used A modified version of the software used in [15] implemented in C++ was compiled on gcc version 4.8.4 and run on Ubuntu 14.04. The results were stored on and retrieved from a relational database implemented on mysql Version 14.14 Distribution 5.7.15 for Linux. Matlab(R2016a) was used to aggregate data and generate plots to visualise the results for bias discovery. For clustering, a gaussian mixture model was used. The Matlab implementation 'fitgmdist'[10] was used to model a gaussian mixture with two components. The reference dataset was intersected with the input data using the 'intersect' option of bedtools v2.17.0. For one class classification, the implementation of The One-Class SVM (as in[17]) of scikit-learn was used with the RBF kernel.

4. Results

4.1. Bias Discovery

4.1.1. Repeat Overlap Bias

The distribution of repeat overlaps around the distal interacting regions(interactors) is as shown in Figure 4. The interactors were grouped on the number of supporting pairs its corresponding interaction had in HiCap output. Most of the interacting regions map to very low numbers of repeat overlaps. This means that the distal interacting regions can be uniquely mapped to the promoter regions and no bias with respect to repeat overlaps could be seen in the HiCap output data.

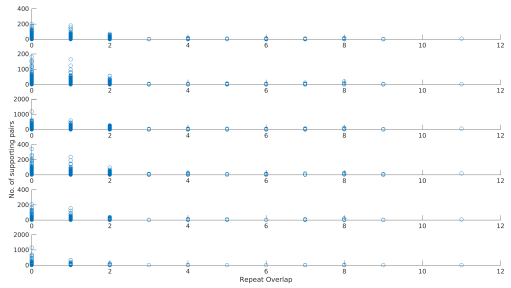

Figure 4: Plot of Repeat Overlaps around distal interacting regions for different counts of supporting pairs. From the top, the plots show repeat overlaps against interactions with 1 supporting pair, 2 supporting pairs, 3 supporting pairs, 4 supporting pairs, 5 supporting pairs, and more than 5 supporting pairs.

4.1.2. Restriction Enzyme Site Bias

The distribution of the cut site counts of the restriction enzyme used in HiCap(DpnII) in a 10kb region around the distal interacting regions is as shown in Figure 5. As in the case in section **??**, the interactors were grouped on their number of supporting pairs. The plots show an enrichment in the number of supporting pairs around 20 to 35 restriction enzyme cut sites. The recognition sequence of the four-cutter DpnII is 'GATC'. Assuming that the bases in the genome are uniformly distributed, the expected number of occurences of the 4-mer 'GATC' in a 10kb region is approximately 39. This shows that the enrichment found could be due to the normal density of restriction cut sites in the genome and no significant bias selecting for or against restriction enzyme cut sites in the HiCap method could be found.

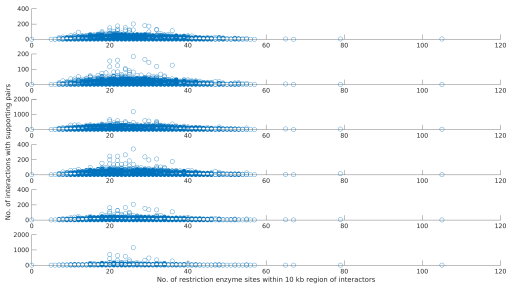

Figure 5: Plot of Restriction Enzyme sites 10 Kb around distal interacting regions for different counts of supporting pairs. From the top, the plots show restriction enzyme sites against interactions with 1 supporting pair, 2 supporting pairs, 3 supporting pairs, 4 supporting pairs, 5 supporting pairs, and more than 5 supporting pairs.

4.1.3. GC Content Bias

A normalized mapping of GC content of interactors for each grouping of support pairs and of GC content of probe containing restriction enzyme fragments was done with respect to the Equation 1 and can be seen in Figure 6. The GC content was binned with bin size corresponding to 10 % GC content. As the probes themselves are only 120 bp in length, the GC content of the restriction enzyme fragment which the probe would select for was used instead. These fragments are variable sized. The figure shows a shift of the peak of the GC content of the distal interacting regions with respect to Probe fragment GC content to regions of lower GC. This could be a result of promoter regions having a higher GC content compared to the rest of the genome. As the probes select for promoters, consequently the probe containing regions have higher GC as well. In the case of the interactors, HiCap seems to pick fragments with GC content that is in the normal range of the human genome except for the fragments that have more than 5 supporting pairs which have a slight enrichment of GC content. This may be a natural consequence of the fact that cis-regulatory elements were found to be enriched in GC nucleotides[8].

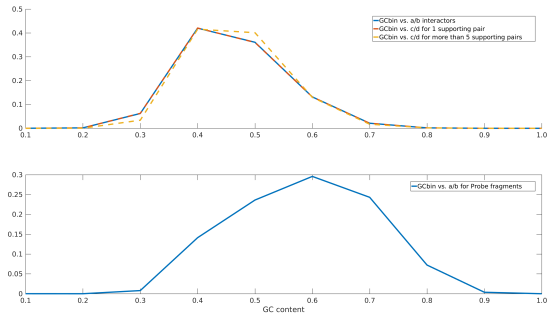

Figure 6: Normalized GC Content measure binned into ten equal sized bins. The first subplot shows the plot of normalized GC content of all interactors against the GC content bins and the plots of normalized GC content interactors with 1 and more than 5 supporting pairs against the GC content bins

The GC content of probes and interacting distal regions also showed very low correlation with Spearman's $\rho = 0.1948$ and Pearson's r = 0.1989. This also seems to indicate that the GC content of the probes may not affect the selection of distal interacting regions.

4.2. Improving Interaction Calling

4.2.1. Approach 1: Modeling with Gaussian Mixtures

The input dataset was labeled with 'ChipSeqOverlapSet' or 'NegSet' depending on whether a partic-

ular interaction intersected with ChIP-Seq peaks in the reference dataset or not. It was assumed that the interactions overlapping with ChIP-Seq peaks are true functional interactions. The 'NegSet' may include structural and functional interactions. All parameters were normalised to the range [0, 1] if it did not include negative values or to the range [-1, 1] if it did. The normalisation is required as the values for certain parameters had very big ranges; for instance, 'distance' had a range of approximately [-2^8, 2^8] which skews the fitted distribution if used without normalisation.

As there were five parameters in the input data, the 'fitgmdist' MATLAB procedure was run once with all five parameters and a second time after using Principal Component Analysis (PCA) to convert the five parameters to two with the added advantage that the clusters can also be visualised. In each run of the 'fitgmdist' procedure, the EM algorithm was repeated 20 times, and the largest log-likelihood is chosen from all the repeats. The runs were also repeated with different values of regularization. The regularization term controls for the complexity of the fitted model and is a means to reduce overfitting in case of noisy data.

The results of clustering will be presented in the format as in Table 1. The table does not correspond to a traditional confusion matrix for classifiers, nor do the calculated values correspond to sensitivity of the classification. This is because a properly defined example negative class to exemplify structural interactions does not exist. The variable 'n' indicates the number of interactions clustered into a specific cluster given by the superscript which is either C1 for Cluster 1 and C2 for Cluster 2.

The variable 'N' indicates the total number of interactions in various cases as given by the subscript - CS gives total number of ChIP-Seq overlapping interactions in the procedure input, NCS gives the total number of non-ChIP-Seq overlapping interactions in the procedure input, C1 and C2 gives the total number of interactions clustered into Cluster 1 and Cluster 2. It is important to note that the clustering of interactions into clusters 1 and 2 are not connected with the labels in any manner.

Case 1: No Regularization The results obtained are as shown in Table 2. The visualisation of the clustering after the five parameter input was converted to 2 parameters with PCA is as shown in Figure 7.The green region contains the datapoints clustered into Cluster 1 and the red region those clustered into Cluster 2. The contour curves of one of the fitted gaussians can also be seen as dashed lines. The datapoints of the interactions that intersect with ChIP-Seq peaks are shown as red dots and the datapoints of the interactions that do not intersect with ChIP-Seq peaks are shown as blue dots. It can be seen from the Table 1 that without PCA approximately 68% of the ChIP-Seq overlapping interactions are clustered into one cluster and with PCA, it rises to 75%. As can be seen from the figure, the dataset is very grainy.

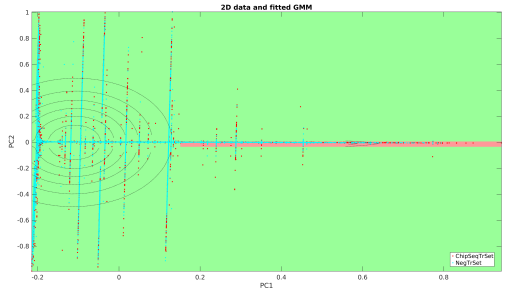

Figure 7: Visualisation of GMM clusters with no regularization of the with PCA run as in Table 2. The green region contains Cluster 1 datapoints and the red region contains Cluster 2 datapoints

Case 2: With Regularization=0.0001 The results are as shown in Table 3.The visualisation of the clustering after the five parameter input was converted to 2 parameters with PCA is as shown in Figure 8. The green region contains the datapoints clustered into Cluster 1 and the red region those clustered into Cluster 2. The contour curves of both of the fitted gaussians can also be seen as dashed lines, one fully and the other partially. It can be seen from the Table 3 that without PCA approximately 82% of the ChIP-Seq overlapping interactions are clustered into one cluster and with PCA, it rises to 83%. This seems to indicate that the regularization is effective to smoothen out the noisy data. The corresponding values for the non ChIP-Seq overlapping set seems to rise as well.

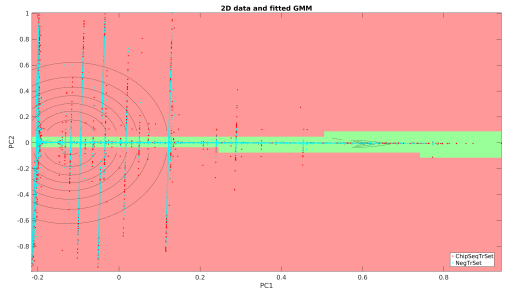

Figure 8: Visualisation of GMM clusters with regularization=0.0001 of the with PCA run as in Table 3. The green region contains Cluster 1 datapoints and the red region contains Cluster 2 datapoints.

	Label	No PCA		with PCA	
		Cluster I	Cluster II	Cluster I	Cluster II
Training data					
	ChIP-Seq overlap	$\frac{n_{CS}^{C1}}{N_{CS}}$	$\frac{n_{CS}^{C2}}{N_{CS}}$	$\frac{n_{CS}^{C1}}{N_{CS}}$	$\frac{n_{CS}^{C2}}{N_{CS}}$
	No overlap	$\frac{n_{NCS}^{C1}}{N_{NCS}}$	$\frac{n_{NCS}^{C2}}{N_{NCS}}$	$\frac{n_{NCS}^{C1}}{N_{NCS}}$	$\frac{n_{NCS}^{C2}}{N_{NCS}}$
Test Data					
	ChIP-Seq overlap	$\frac{n_{CS}^{C1}}{N_{CS}}$	$\frac{n_{CS}^{C2}}{N_{CS}}$	$\frac{n_{CS}^{C1}}{N_{CS}}$	$\frac{n_{CS}^{C2}}{N_{CS}}$
	No overlap	$\frac{n_{NCS}^{C1}}{N_{NCS}}$	$\frac{n_{NCS}^{C2}}{N_{NCS}}$	$\frac{n_{NCS}^{C1}}{N_{NCS}}$	$\frac{n_{NCS}^{C2}}{N_{NCS}}$

Table 1: Format of presentation of results of clustering and classification.

	Label	No PCA		with PCA	
		Cluster I	Cluster II	Cluster I	Cluster II
Training Data					
	ChIP-Seq overlap	0.6774	0.3226	0.7585	0.2415
	No overlap	0.6534	0.3466	0.7047	0.2953
Test Data					
	ChIP-Seq overlap	0.6756	0.3244	0.7561	0.2439
	No overlap	0.6527	0.3473	0.7052	0.2948

Table 2: Gaussian Mixture Model fitted for training data with no regularization

	Label	No PCA		with PCA	
		Cluster I	Cluster II	Cluster I	Cluster II
Training Data					
	ChIP-Seq overlap	0.1798	0.8202	0.1617	0.8383
	No overlap	0.2208	0.7792	0.2000	0.8000
Test Data					
	ChIP-Seq overlap	0.1829	0.8171	0.1646	0.8354
	No overlap	0.2199	0.7801	0.1988	0.8012

Table 3: Gaussian Mixture Model fitted for training data with regularization=0.0001

Case 3: With Regularization=0.001 The visualisation of the clustering after the five parameter input was converted to 2 parameters with PCA is as shown in Figure 9. The green region contains the datapoints clustered into Cluster 1 and the red region those clustered into Cluster 2. The contour curves of both of the fitted gaussians can also be seen as dashed lines, one fully and the other partially. It was found that without PCA approximately 86% of the ChIP-Seq overlapping interactions are clustered into one cluster and with PCA, it rises to 89%. As a very high percentage of ChIP-Seq overlapping interactions seem to be in a single cluster and assuming that these are all functional interactions, it can be seen that a high percentage of the non-overlapping interactions are clustered into the same cluster. This seems to indicate that the number of structural interactions in the dataset may be

low.

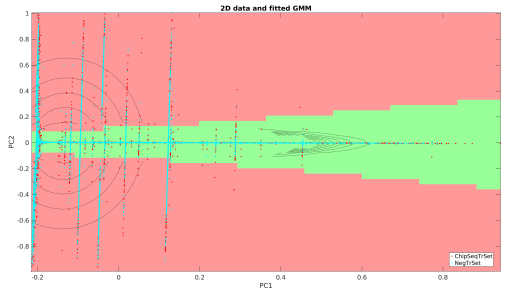

Figure 9: Visualisation of GMM clusters with regularization=0.001 of the 'with PCA' run. The green region contains Cluster 1 datapoints and the red region contains Cluster 2 datapoints

Case 4: With Regularization=0.01 The results are as shown in Table 4. The visualisation of the clustering after the five parameter input was converted to 2 parameters with PCA is as shown in Figure 10. The green region contains the datapoints clustered into Cluster 1 and the red region those clustered into Cluster 2. The contour curves of both of the fitted gaussians can also be seen as dashed lines, one fully and the other partially.It can be seen from the Table 4 that without PCA approximately 94% of the ChIP-Seq overlapping interactions are clustered into one cluster and with PCA, it is also around 94%. As only very low number of non overlapping interactions are clustered into the second cluster, it may not be clear if the model becomes too simplified with this value of regularization.

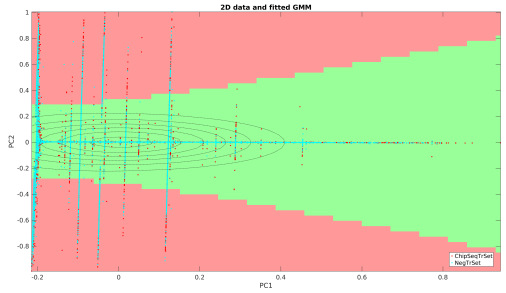

Figure 10: Visualisation of GMM clusters with regularization=0.01 of the 'with PCA' run as in Table 4. The green region contains Cluster 1 datapoints and the red region contains Cluster 2 datapoints

4.2.2. Approach 2: Classification with One Class SVM

The OneClassSVM procedure of sklearn was run with the rbf kernel with a gamma of 0.1. The results are as shown in Table 5. The visualisation of the clustering after the five parameter input was converted to 2 parameters with PCA is as shown in Figure 11. The thick red line indicates the class boundary learned by the algorithm. The yellow region inside the boundary contains the datapoints clustered into the Class and the region outside the boundary contains the datapoints rejected as being dissimilar to and not belonging with the datapoints in the learned class. From the Table 5, it can be seen that for both ChIP-Seq overlapping and non-overlapping interactions around 90% of the datapoints are accepted by the classifier as part of the learned class. From the figure 11, it can be seen that the data is very grainy and the datapoints of the ChIP-Seq overlapping interactions are widely distributed when compared to the non-overlapping set.

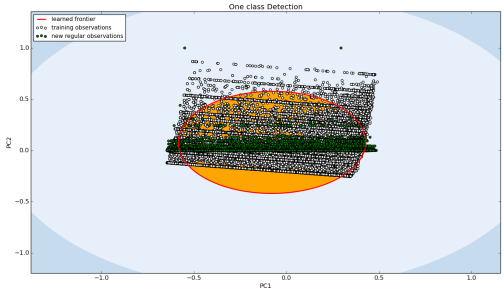

Figure 11: Visualisation of learning boundary with training and test data of the 'with PCA' run as in Table 5. The datapoints inside the learning boundary are accepted by the One class SVM, while the ones outside are rejected.

5. Conclusions

In the evaluation of biases in HiCap, It was found that most of the biases in Hi-C data may not be carried over into HiCap. It is hypothesized that this may be due to the higher selectivity and resolution inherent in the HiCap method. This does not discount the fact that this selectivity of HiCap may dispose it towards biases that are not found in Hi-C data. This is an area that bears more investigation.

In the case of differentiating between structural and functional interactions in HiCap output, it was found that although the techniques separated the interactions into two, more work may be needed to increase the confidence in the classification. This may include filtering the reference datasets for high confidence enhancer peaks, increasing the number of clusters with cross-validation and using binary

	Label	No PCA		with PCA	
		Cluster I	Cluster II	Cluster I	Cluster II
Training Data					
	ChIP-Seq overlap	0.9418	0.0582	0.9441	0.0559
	No overlap	0.9366	0.0634	0.9326	0.0674
Test Data					
	ChIP-Seq overlap	0.9403	0.0597	0.9428	0.0572
	No overlap	0.9353	0.0647	0.9312	0.0688

Table 4: Gaussian Mixture Model fitted for training data with regularization=0.01

	Label	No PCA		with PCA	
		Learned Class	Rejected	Learned Class	Rejected
Test Data					
	ChIP-Seq overlap	0.8995	0.1006	0.9002	0.0998
	No overlap	0.9005	0.0994	0.9024	0.0976

Table 5: One Class SVM fitted for training data with the rbf kernel

classification methods for comparison against the results of the one-class classification.

Acknowledgements
The author would like to thank Dr. Pelin Sahlén for her guidance and the Science for Life Laboratory, Stockholm and the KTH Royal Institute of Technology, Stockholm allowing access to their resources which were used in the completion of this work. The author also thanks Dr. Isabel Sá-Correia of the Instituto Superior Técnico, Lisbon for the opportunity to pursue this work. The author also thanks the euSYSBIO Erasmus Mundus programme for providing the funding which allowed work on this paper.

References

[1] W. Chanput, J. J. Mes, and H. J. Wichers. Thp-1 cell line: An in vitro cell model for immune modulation approach. *International Immunopharmacology*, 23:37–45, Nov. 2014. doi: 10.1016/j.intimp.2014.08.002.

[2] N. F. Cope and P. Fraser. Chromosome conformation capture. *Cold Spring Harbour Protocols*, 29, Feb. 2009. doi: 10.1101/pdb.prot5137.

[3] M. B. Gerstein, A. Kundaje, M. Hariharan, S. G. Landt, K.-K. Yan, C. Cheng, X. J. Mu, E. Khurana, J. Rozowsky, R. Alexander, R. Min, P. Alves, A. Abyzov, N. Addleman, N. Bhardwaj, A. P. Boyle, P. Cayting, A. Charos, D. Z. Chen, Y. Cheng, D. Clarke, C. Eastman, G. Euskirchen, S. Frietze, Y. Fu, J. Gertz, F. Grubert, A. Harmanci, P. Jain, M. Kasowski, P. Lacroute, J. J. Leng, J. Lian, H. Monahan, H. O'Geen, Z. Ouyang, E. C. Partridge, D. Patacsil, F. Pauli, D. Raha, L. Ramirez, T. E. Reddy, B. Reed, M. Shi, T. Slifer, J. Wang, L. Wu, X. Yang, K. Y. Yip, G. Zilberman-Schapira, S. Batzoglou, A. Sidow, P. J. Farnham, R. M. Myers, S. M. Weissman, , and M. Snyder. Architecture of the human regulatory network derived from encode data. *Nature*, 489(7414):91–100, Sept. 2012. doi: 10.1038/nature11245.

[4] A. E. Guttmacher and F. S. Collins. Welcome to the genomic era. *New England Journal of Medicine*, 349:996–998, Sept. 2003. doi:10.1056/NEJMe038132.

[5] M. Hu, K. Deng, S. Selvaraj, Z. Qin, B. Ren, and J. S. Liu. Hicnorm: removing biases in hi-c data via poisson regression. *Bioinformatics*, 28:3131–3133, Sept. 2012. doi: 10.1093/bioinformatics/bts570.

[6] M. J. Iglesias, S.-J. Reilly, O. Emanuelsson, B. Sennblad, M. P. Najafabadi, L. Folkersen, A. Mälarstig, J. Lagergren, P. Eriksson, A. Hamsten, and J. Odeberg. Combined chromatin and expression analysis reveals specific regulatory mechanisms within cytokine genes in the macrophage early immune response. *PLoS One*, 7(2):91–100, Feb. 2012. doi: 10.1371/journal.pone.0032306.

[7] G. Khoury and P. Gruss. Enhancer elements. *Cell*, 33:313–314, June 1983. doi:10.1016/0092-8674(83)90410-5.

[8] C. M. Koch, R. M. Andrews, P. Flicek, S. C. Dillon, U. Karaz, G. K. Clelland, S. Wilcox, D. M. Beare, J. C. Fowler, P. Couttet, K. D. James, G. C. Lefebvre, A. W. Bruce, O. M. Dovey, P. D. Ellis, P. Dhami, C. F. Langford, Z. Weng, E. Birney, N. P. Carter, D. Vetrie, and I. Dunham. The landscape of histone modifications across 1in five human cell lines. *Genome Research*, 17:691–707, June 2007. doi:10.1101/gr.5704207.

[9] B.-Y. Liao, N. M. Scott, and J. Zhang. Impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate of mammalian proteins. *Molecular Biology and Evolution*, 23:2072–2080, Aug. 2006. doi: 10.1093/molbev/msl076.

[10] G. McLachlan and D. Peel. *Finite Mixture Models*. John Wiley and Sons, 2000. doi: 10.1002/0471721182.

[11] J. O. Y. nez Cuna, E. Z. Kvon, and A. Stark. Deciphering the transcriptional cis-regulatory code. *Trends in Genetics*, 29:11–22, Jan. 2013. doi:10.1016/j.tig.2012.09.007.

[12] L. A. Pennacchio, W. Bickmore, A. Dean, M. A. Nobrega, and G. Bejerano. Enhancers: five essential questions. *Nature Reviews Genetics*, 14:288–295, Apr. 2013. doi:10.1038/nrg3458.

[13] T.-H. Pham, J. Minderjahn, C. Schmidl, H. Hoffmeister, S. Schmidhofer, W. Chen, G. Längst, C. Benner, and M. Rehli1. Mechanisms of in vivo binding site selection of the hematopoietic master transcription factor pu.1. *Nucleic Acids Research*, 41(13), May 2013. doi: 10.1093/nar/gkt355.

[14] M. E. Reschen, K. J. Gaulton, D. Lin, E. J. Soilleux, A. J. Morris, S. S. Smyth, and C. A. O'Callaghan. Lipid-induced epigenomic changes in human macrophages identify a coronary artery disease-associated variant that regulates ppap2b expression through altered c/ebp-beta binding. *PLoS Genetics*, 11, Apr. 2015. doi: 10.1371/journal.pgen.1005061.

[15] P. Sahlén, I. Abdullayev, D. Ramsköld, L. Matskova, N. Rilakovic, B. Lötstedt, T. J. Albert, J. Lundeberg, and R. Sandberg. Genome-wide mapping of promoter-anchored interactions with close to single-enhancer resolution. *Genome Biology*, 16:156, Aug. 2015. doi: 10.1186/s13059-015-0727-9.

[16] W. Schaffner. Enhancers, enhancers - from their discovery to today's universe of transcription enhancers. *Biological Chemistry*, 396(4):311327, Feb. 2015. doi:10.1515/hsz-2014-0303.

[17] B. Schölkopf, R. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt. Support vector method for novelty detection. In *Advances in Neural Information Processing Systems 12*, page 582588. MIT Press, 2000.

[18] D. Shlyueva, G. Stampfel, and A. Stark. Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics*, 15:272–286, Mar. 2014. doi:10.1038/nrg3682.

[19] J. Wang, J. Zhuang, S. Iyer, X. Lin, T. W. Whitfield, M. C. Greven, B. G. Pierce, X. Dong, A. Kundaje, Y. Cheng, O. J. Rando, E. Birney, R. M. Myers, W. S. Noble, M. Snyder, and Z. Weng. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Research*, 22(9):17981812, Sept. 2012. doi:10.1101/gr.139105.112.

[20] J. Wang, J. Zhuang, S. Iyer, X.-Y. Lin, M. C. Greven, B.-H. Kim, J. Moore, B. G. Pierce, X. Dong, D. Virgil, E. Birney, J.-H. Hung, and Z. Weng. Factorbook.org: a wiki-based database for transcription factor-binding data generated by the encode consortium. *Nucleic Acids Research*, 41:D171D176, Jan. 2013. doi:10.1093/nar/gks1221.

[21] E. Yaffe and A. Tanay. Probabilistic modeling of hi-c contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature Genetics*, 43:10591065, Oct. 2011. doi:10.1038/ng.947.

[22] I. Yanai, H. Benjamin, M. Shmoish, V. Chalifa-Caspi, M. Shklar, R. Ophir, A. Bar-Even, S. Horn-Saban, M. Safran, E. Domany, D. Lancet, and O. Shmueli1. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics*, 21(5):650–659, Sept. 2005. doi: 10.1093/bioinformatics/bti042.