A probabilistic method for quantifying chromatin interactions

Master’s Thesis
Espoo, September 29, 2015

Supervisor: Professor Harri Lähdesmäki
Advisor: Professor Harri Lähdesmäki
Chromatin interactions have an important role in transcription regulation and therefore they can affect the function of the whole cell and the organism. To study chromatin interactions for better understanding of gene regulation, a method called Chromosome Interaction Analysis using Paired End Tags (ChIA-PET) has been developed. ChIA-PET is a high-resolution next-generation sequencing method for finding chromatin interactions which involve a protein of interest.

ChIA-PET experiments give a list of putative interactions between two chromatin sites as a result. There are several experimental laboratory steps in ChIA-PET protocol which induce high level of background noise. The aim of this thesis is to construct a statistical model for identifying the true interactions from ChIA-PET interaction count data. First, the current methods for solving this task are reviewed. Then a new method combining a Bayesian mixture model with bias removal by Poisson regression is proposed. The model parameters are estimated by using Markov chain Monte Carlo methods. The new model is implemented on Matlab and tested on real ChIA-PET data sets.

The results suggest that the proposed mixture model can quantify chromatin interactions and make good use of incorporated bias correcting. Comparison with two other methods, ChIA-PET Tool and Mango, shows that the mixture model results are partially the same as for the other two methods but there also also some interactions only found by the mixture model. Annotation analysis revealed that the mixture model results are in line with earlier research results.

Keywords: ChIA-PET, chromatin interactions, statistical methods, Poisson regression, Bayesian mixture model

Language: English
Kromatiini-interaktiot ovat tärkeä tekijä geenien sääntelyssä ja tätä kautta kokon solun ja elon toiminnassa. Kromatiinin muodostamat silmukat tuovat transkription käynnistävät tekijät toistensa läheisille ja näin mahdollistavat proteiinien rakentamisen.

Kromatiini-interaktioiden tutkimiseen on kehitetty erilaisia NGS-menetelmiä, joista yksi on Chromatin interaction analysis using paired end tags eli ChIA-PET. Tässä menetelmässä kromatiisisilmukat lukuuta paikoilleen ja pilkotaan niin, että lopputuloksena on lista yhdessä esiintyneistä DNA:n kohdista. ChIA-PET tyyppinen data sisältää kuitenkin oikeiden interaktioiden lisäksi runsaasti satunnaisesti toisiinsa kiinnittyneitä pääkiitä, jotka tulisi erottaa oikeista havainnoista. Tämä työ esittelee jo olemassa olevat menetelmät tämän ongelman ratkaisemiseen. Sen jälkeen esitellään uusi menetelmä interaktioiden luokitteluun.

Uusi menetelmä yhdistää bayeslaisen mikstuurimallin ja Poisson regression virhelähteiden poistoon. Mallin parametrien estimointiin käytetään bayeslaisista analyysä ja Markov Chain Monte Carlo -menetelmää. Mallin toteutus tehtiin Matlabilla ja sitä testattiin ChIA-PET-aineistoon.

Tulokset osoittavat, että mikstuurimalli pystyy erottelemaan kromatiini-interaktioita käyttäen hyväksy virhelähteiden korjautta. Vertailtaessa tuloksia ChIA-PET Tool ja Mango-ohjelmistojen kanssa huomataan, että mikstuurimalli löytää osaksi samoja ja osaksi eri interaktioita. Annotaatioanalyysin perusteella mikstuurimallin tulokset ovat linjassa aiempien tutkimustuloksien kanssa.
Acknowledgements

First, I would like to thank my supervisor and advisor Professor Harri Lähdesmäki for an interesting thesis topic and all the guidance I have received. Second, I am also very grateful to my unofficial advisor Henrik Mannerström for his hard work for this thesis. Henrik has helped me in the many practical tasks this work has involved and without his help I would probably still be banging my head to the wall. I would like to thank the whole CSB research group and especially my room mates Henrik and Sini for the great work atmosphere I have had a privilege to enjoy. The computational resources provided by the Aalto Science-IT project were used when making this thesis.

I would like to thank my family for their support. Special thanks to Aunt Kati, who helped me to settle in Espoo when I started my studies.

Last but not least, I would like to thank all my dear friends for encouraging me and all the fun we have had! Especially Joonas, the fun conversations we have had have inspired me to go on during this master’s thesis process. Thanks to you, my study years have been a bizarre adventure! And Antti, thank you for always being there for me.

Espoo, September 29, 2015

Viiivi Halla-aho
Abbreviations and Symbols

ChIA-PET  Chromatin interaction analysis using paired end tags
DNA     Deoxyribonucleic acid
DNA-FISH DNA fluorescence in situ hybridization
EM      Expectation-maximization algorithm
HMC     Hamiltonian Monte Carlo
MCMC    Markov chain Monte Carlo
NGS     Next-generation sequencing
NUTS    No-U-turn-sampler
PET     Paired end tag
TFBS    Transcription factor binding site
TSS     Transcription starting site
$I_{AB}$ Interaction count for the interaction between anchor $A$ and $B$
$R_A$  Anchor region $A$
$c_A$  Marginal count for anchor $A$
$N$    Number of inter-ligation PETs
$n$    Number of fragment pairs
$\rho$  Expected number of interactions
$\rho_{AB}$  Expected interaction count for anchors $A$ and $B$
$\rho_A$  Expected marginal count for anchor $A$
$\omega_{AB}$  The non-centering coefficient for interaction between anchors $A$ and $B$
$I_i$  Interaction count for fragment pair $i$
$w_{0i}$  Mixing parameter for fragment pair $i$
$w_{1i}$  Mixing parameter for fragment pair $i$
$\lambda_{0i}$  Poisson rate parameter of the first component for fragment pair $i$
$\lambda_{1i}$  Poisson rate parameter of the second component for fragment pair $i$
$d_i$ Distance between the interacting anchors for fragment pair $X_i$

$m_{iA}$ Mappability of anchor $A$ of interaction pair $X_i$

$g_{iA}$ GC content of anchor $A$ of interaction pair $X_i$

$c_{iA}$ Coverage of anchor $A$ of interaction pair $X_i$

$\beta_i$ Coefficients of Poisson regression model

$K$ Number of explanatory variables in Poisson regression model

$k$ Truncation parameter

$D_i$ Addition to $\lambda_{ii}$ to get $\lambda_{1i}$

$\beta_j$ Regression coefficient, $j = 0, \ldots, K$

$\sigma_b$ Standard deviation of regression coefficient prior

$\mu_D$ Mean of $\mu_{D_i}$’s prior

$\sigma_D$ Standard deviation of $\mu_{D_i}$’s prior

$\mu_{\mu_D}$ Mean of $\mu_D$’s prior

$\sigma_{\mu_D}$ Standard deviation of $\mu_D$’s prior

$a_{\sigma_D}$ Gamma distribution shape parameter of $\sigma_D$’s prior

$b_{\sigma_D}$ Gamma distribution rate parameter of $\sigma_D$’s prior

$I$ Vector $I = (I_1, \ldots, I_n)$

$D$ Vector $D = (D_1, \ldots, D_n)$

$w$ Vector $w = (w_{01}, \ldots, w_{0n})$

$\beta$ Vector $\beta = (\beta_0, \ldots, \beta_K)$
## Contents

Abbreviations and Symbols

1 Introduction 1
  1.1 Background ............................................. 1
  1.2 Problem statement ..................................... 3
  1.3 Structure of this thesis ................................. 3

2 Background 5
  2.1 DNA Sequencing ......................................... 5
  2.2 ChIA-PET Sequencing .................................... 6
  2.3 Other experimental methods for identifying chromatin interactions ........................................ 9

3 Literature review 11
  3.1 Current ChIA-PET analysis methods ....................... 11
    3.1.1 Hypergeometric model ................................. 12
    3.1.2 Non-central hypergeometric model ..................... 13
    3.1.3 Bayesian mixture model ............................... 14
    3.1.4 SPROUT ............................................. 17
    3.1.5 Mango ................................................ 19
    3.1.6 Comparing the methods ............................... 20
  3.2 Improving the methods .................................... 22
    3.2.1 Bias correction ....................................... 23
    3.2.2 Including external input data ....................... 23
  3.3 Result validation ......................................... 24

4 Methods 26
  4.1 Preprocessing of the data ............................... 26
    4.1.1 Processing raw ChIA-PET data ....................... 27
    4.1.2 Anchor site finding .................................. 28
    4.1.3 Calculating anchor properties ....................... 29
4.1.4 Finding the corresponding anchors for interaction head
and tail sites ........................................ 30
4.2 Poisson mixture model with Poisson regression .......... 30
4.3 Bayesian inference and MCMC methods .............. 34
4.4 Distributions used in sampler ........................ 38
4.5 Sampler implementation ............................... 43

5 Results 47
5.1 The Mcf7 cell line data set ............................ 47
5.2 Modeling inter-chromosomal interactions ............ 48
5.3 Preliminary results from Poisson regression ........ 49
5.4 Full model ............................................. 53
5.5 Comparison with other methods ....................... 56

6 Conclusions 61

Bibliography 64
Chapter 1

Introduction

1.1 Background

Genes are segments of deoxyribonucleic acid, DNA, and they determine the primary structure of a particular protein or ribonucleic acid, RNA. These proteins and RNAs have their own functions in the organism and thus the genes determine how the organism works. The four bases adenine, cytosine, guanine and thymine work as building blocks of the DNA, which consists of genes and regulatory elements plus other nucleotide sequences. A new RNA can be produced in a process called transcription, where DNA is used as a template. RNAs can then be used for synthesizing proteins by translation (Cooper, 2000, Expression of Genetic Information). This process where genetic information of the genes is transformed into functional products, RNA and proteins, is called gene expression (Niu, 2012).

The entirety of DNA, both genes and the non-coding sequences, is called genome (Niu, 2012). Eukaryotes have a complex genome structure and organization. Eukaryotic genome consists of multiple chromosomes, which contain nucleic acids in a long linear string like structure. Together with different kinds of proteins the eukaryotic genome forms a complex which is called chromatin. One major type of proteins included in chromatin is histones, which are needed in packaging the genome. There is also a wide range of other chromosomal proteins with different purposes. The number and size of chromosomes vary between the wide range of eukaryotic species, but the basic structure is the same for all of them (Cooper, 2000, Chromosomes and Chromatin).

Cells can control gene expression through gene regulation (Niu, 2012). These mechanisms are very complex in eukaryotic organisms. The primary way of controlling gene expression is to regulate the starting of transcription. RNA polymerase is an enzyme needed in transcription of RNA. RNA poly-
Figure 1.1: Example of chromatin interaction where DNA loop brings enhancer region near the transcription starting site. Figure from Cooper (2000, Regulation of Transcription in Eukaryotes).

RNA polymerase needs to bind to a promoter DNA sequence for transcription to start. So the presence of a promoter sequence is needed for producing RNA. Other type of regulatory sequences stimulating the process is enhancers, which can be located quite far from the transcription initiation site. They bind transcription factor proteins, which then interact with RNA polymerase or other transcription factors in the promoter site. Chromatin looping enables this interaction by bringing the promoter and enhancer sequences close to each other. Example of such looping is presented in Figure 1.1. There are also other ways how the chromatin structure can prevent or enable transcription (Cooper, 2000, Regulation of Transcription in Eukaryotes). For example, looping can isolate genome regions from transcription (Reeder and Gifford, 2013). So in addition to the necessary transcriptional regulatory proteins binding to enhancers and promoters, the correct structure of the chromatin is essential for initiating the transcription (Cooper, 2000, Regulation of Transcription in Eukaryotes).

Chromatin structure has been studied using microscopes and many different kinds of DNA sequencing methods (de Wit and de Laat, 2012). The aim is to get a more comprehensive understanding of the structure and function of the genome. The most intuitive way to present the structure is a three-dimensional model and the current methods strive towards that goal (Li et al., 2014).

One of the most recent methods is chromatin interaction analysis using paired-end tags, in short ChIA-PET. It captures the pairs of looped chromatin parts which are brought together by a certain protein. The interacting
chromatin fragments which are extracted in the experiment are called tags. ChIA-PET is a de novo method with which genome-wide interactions can be analyzed with high throughput and high resolution (Li et al., 2014). This experiment produces data from which a list of putative chromatin interaction sites and their respective frequencies can be formed (de Wit and de Laat, 2012).

Better understanding of chromatin structure and its role in gene regulation might give insights about how diseases and cancer work. For example, the structure of the chromatin might be different in normal and cancer cells and so lead to anomalous function of the cancer cell. Comparisons like this have already been done (Niu et al., 2014). As the cost and time consumption of the sequencing methods are decreasing, chromatin structure analysis might become an important tool in personalized medicine.

1.2 Problem statement

The difficult part of analysing data from ChIA-PET experiment is to distinguish between true interactions and random noise. Some statistical methods have been proposed to answer this question and they are introduced in Chapter 3.

All of the methods are based on the assumption that the signal should be more powerful for the true interactions than for the random noise. Some of the methods also take the distance between the interacting chromatin parts into account as interactions are thought to happen more often between parts in close proximity. Along the distance information, the interaction analysis could be enhanced by adding external information about the chromatin parts in question into the model. For example, influence of known bias factors could be eliminated this way.

This thesis reviews the current methods of distinguishing true interactions from false ones and proposes improving the Bayesian mixture model first proposed by Niu and Lin (2015) by incorporating bias correction into the model. This model is then implemented in Bayesian framework using Markov chain Monte Carlo methods and tested on real data.

1.3 Structure of this thesis

The structure of this thesis is the following: first the background of the problem at hand is introduced and then the existing solutions to solve it are presented in Chapter 3 Literature review. In the end of Chapter 3, the potential solutions
for improving the present methods are reviewed. Chapter 4 starts by describing how ChIA-PET data is preprocessed before analysis. After this the statistical model proposed in this thesis is explained along with the theory behind it. Bayesian analysis and Markov chain Monte Carlo methods, which are used in estimating the model parameters are introduced as well. The details about the implementation of the proposed method can be read from end of the same Chapter 4. To evaluate how well the improved method works, the model is implemented and tested on a real data set. Then the results are compared with some other current methods. The results are presented in Chapter 5. Finally, the results are discussed and possible future research proposals are given before giving the conclusions of this thesis in Chapter 6.
Chapter 2

Background

The basics of DNA sequencing are presented in the first section of this chapter to provide understanding of how the data handled in this thesis has been produced. In the second section the focus is on chromosome interaction analysis using paired-end tags (ChIA-PET) method and the protocol is explained in detail. Different types of methods for chromatin interaction discovery are also introduced in the last section to show the different approaches for chromatin interaction analysis. The differences and advantages compared to ChIA-PET are also reviewed.

2.1 DNA Sequencing

The aim of DNA sequencing is to find out the exact order of the nucleotides adenine, guanine, cytosine and thymine in a DNA molecule coming from a molecular biology experiment. State-of-art methods are called next-generation sequencing (NGS). These NGS techniques are very efficient because the DNA to be sequenced is fragmented into small segments and the order of nucleotides is parallelly determined for these fragments. The strings of bases obtained as a result of sequencing are called reads (Grada and Weinbrecht, 2013). The details of the multiple available methods for the actual nucleotide order determining are not covered in this thesis. The next step after sequencing is to map the reads to a reference genome. This is how the genomic coordinates of the sequenced fragments are found out (Grada and Weinbrecht, 2013).

All types of experiments studying the chromatin biology are ridden with bias factors, some of them experiment-specific ones and some of them influence all experiments. There are bias sources in all the stages of the experiment starting from the structure of the chromatin to be investigated to the mapping of the reads. It is important that the bias factors are considered in the analy-
sis, otherwise they might be interpreted accidentally as biological phenomena (Meyer and Liu, 2014). Next some of the bias factors relating to ChIA-PET experiment are briefly introduced.

Experiments such as ChIA-PET include sonication of the chromatin to cut the whole thing into shorter fragments. The structure of the chromatin might make some parts more fragile than the other ones and when sonicated i.e. broken apart using sound energy, this results in fragments of different lengths. GC content, the proportion of guanines and cytosines in a certain part of DNA, can induce bias in multiple ways. For example in polymerase chain reaction (PCR) amplifying, where fragments to be sequenced are copied, the efficiency of amplification might differ between sequences. The efficiency is higher for fragments with high GC content and thus the GC rich fragments are amplified more than the others (Meyer and Liu, 2014).

When mapping the short NGS sequences into the reference genome, new challenges are encountered. It might be impossible to find a unique match from the whole reference genome for a short sequence. Repetitive elements or duplications make this task even harder. Genomic variation might even result in sequences which do not match the reference genome and cannot be mapped at all. In consequence, there are regions in the genome which are generally hard to map (Meyer and Liu, 2014). Quantity called mappability measures this property and it can be calculated for sequences of certain length for the whole genome (Derrien et al., 2012).

2.2 ChIA-PET Sequencing

ChIA-PET is a chromatin conformation capture (3C) based method first proposed by Fullwood et al. (2009) where chromatin interactions bound by a protein of interest are precipitated using chromatin immunoprecipitation (ChIP) technique. ChIA-PET can be used for finding genome-wide interactions in high resolution and it uses high-throughput sequencing. ChIA-PET has already been applied to different types of human and mouse cells, such as human cancer and T-cells and mouse embryonic stem cells (Li et al., 2014).

The steps of the ChIA-PET procedure and the different kinds of interactions captured by the method are illustrated in Figure 2.1. Figure 2.1a shows the wet-lab experiment in a simplified way. First, chromatin is cross-linked by using formaldehyde. After this, the DNA-protein complexes are chopped into smaller fragments with sonication. Chromatin immunoprecipitation is then used for picking the fragments bound by a protein of interest. In other words, a specific antibody for the protein of interest is used to capture the desired fragments. Next the picked chromatin fragments are split into two different
Figure 2.1: The steps of ChIA-PET protocol and the different types of interactions that can be identified. Figure from Li et al. (2010).
aliquots. Each of the aliquots includes a half-linker of its own and these linkers stick to the DNA-fragment ends. When the two aliquots are combined, half-linkers ligate with their nearest half-linker neighbors. As shown in Figure 2.1a, ligation can happen between two linkers stuck to the same chromatin fragment or two linkers coming from different chromatin fragments. The former type of ligation is called self-ligation and the latter is called inter-ligation. In these cases the linker is formed of the two ligated half-linkers of the same type. If proximity ligation happens between two fragments from different chromatin complexes, the linker will very likely consist of two different half-linkers. The DNA fragments along with the linkers are extracted from the DNA-protein complex with restriction enzyme *Mme*I and paired-end tags (PETs) are gotten as a result. These PETs are then sequenced with next-generation sequencing machines and paired-end reads are obtained as a result (Li et al., 2014).

In Figure 2.1b the architecture of the paired-end tags is shown in more detail. A single PET is constructed of two DNA fragment tags of length 20 base pairs. In between of the two tags lies two half-linkers, which are 19 base pairs long each. Together they form a whole linker with length 38 base pairs. PETs with different half-linkers are called chimeric PETs and the ones with a pair of linkers of the same kind are non-chimeric PETs. Coming from different DNA-protein complexes and from different aliquots, chimerically ligated fragments most likely have not been proximal before chromatin fragmentation. Chimeric PETs can thus be assumed to result from random collision events when the two aliquots are combined. Non-chimeric PETs represent the putative interactions and are used in later analysis of interaction quantifying (Li et al., 2010).

Self-ligations are separated from inter-ligation PETs based on the distance between the interacting sites. If it is under a certain cut-off value, e.g. 4600 base pairs as stated in Figure 2.1c, a PET can be called self-ligation. Other threshold values have also been used in different works, for example Paulsen et al. (2014) used threshold value of 8000 base pairs for separating the self- and inter-ligations. PETs with distance greater than this value are considered as inter-ligation PETs. Inter-ligations can again be divided into intra- and inter-chromosomal interactions, which is illustrated in Figure 2.1d. As the names suggest, the intra-chromosomal interactions are formed between two sites inside one chromosome considerably far away from each other and inter-chromosomal interactions happen between two genomic locations from different chromosomes (Li et al., 2010).

The quality of the data obtained from a ChIA-PET experiment depends mostly on the linker design and quality of the available DNA on which the experiments is done for (Li et al., 2014). The specificity of the antibody used in immunoprecipitation step of the experiment has also an effect on the quality.
CHAPTER 2. BACKGROUND

The generic bias factors of DNA sequencing were introduced before in the Section 2.1 and they apply also for ChIA-PET sequencing.

2.3 Other experimental methods for identifying chromatin interactions

In addition to ChIA-PET, there are many different methods for identifying chromatin interactions. Some of them have a slightly different scope and purpose. Chromatin conformation capture (3C) and its derivatives chromosome conformation capture-on-chip (4C), chromosome conformation capture carbon copy (5C), Hi-C and ChIA-PET all start with almost the same procedure as was depicted in previous Section 2.2 and in Figure 2.1. The differences between the methods lie in resolution and the range of genomic sites of between which the interactions can be identified. 3C can find only interactions between two predefined genomic sites, 4C screens the genome for sites interacting with one particular site, 5C can find interactions between many genomic sites and Hi-C can find all interaction between any genomic sites. ChIA-PET can also find interactions between any genomic sites as long as they are bound by a specific protein (de Wit and de Laat, 2012).

ChIA-PET method is the most complex method of the ones mentioned above (Li et al., 2014). It has high resolution and can map genome-wide interactions, which are its advantages over the rest of 3C based methods. On the other hand, it can only reveal interactions bound by a specific protein (de Wit and de Laat, 2012). Only capturing interactions bound by a specific protein can be thought as targeted approach and results in higher resolution than the more general approach of Hi-C method. 5C is also a method that can provide more high-resolution interaction frequency maps of the targeted range of genome locations, but without protein limitations (Meyer and Liu, 2014). Another disadvantage of ChIA-PET is the low signal-to-noise ratio: most of the putative interactions are just products of random ligation (de Wit and de Laat, 2012). Also, the shortness of the DNA fragments makes specific mapping difficult (Li et al., 2014).

Currently there are only very few tools available for comprehensive ChIA-PET data processing and analysis. ChIA-PET Tool pipeline (Li et al., 2010) provides the basic steps of processing raw data into lists of interaction clusters and simple visualizations. More comprehensive analysis tools would make ChIA-PET more applicable and attractive for prospective users (Li et al., 2014). Also, the difficulties in comprehensive result validation is a disadvantage of this method (Li et al., 2014). Some ways of validating the found interactions
are reviewed in Section 3.3. As the chromatin conformation capture methods cannot tell whether the found interactions actually are functional or not, the importance of validation through other experiments should not be underrated (Kwon et al., 2013).
Chapter 3

Literature review

The literature on current methods for statistical modeling of the chromatin interactions are reviewed in the first section of this chapter. First the different approaches are described and then they are briefly compared to each other based on results found in literature.

After discussing the available methods, the possibilities of refining them are reviewed in Section 3.2. As mentioned in Chapter 2.1, there are quite a few bias factors which should be taken into account when building a mathematical model. There are already a few methods for bias normalization for Hi-C type of data and using them also in statistical models targeted for ChIA-PET data are discussed in Section 3.2.1. The idea of combining different data types for more comprehensive modeling has also been proposed in earlier works. This possibility and some existing implementations are further described in Section 2.1.

Validating the results of chromatin interaction analysis is challenging, as the true chromatin structure is not known. The generally used validation methods found in literature are introduced in Section 3.3.

3.1 Current ChIA-PET analysis methods

As described in previous chapters, there are two kinds of observations that are obtained from ChIA-PET experiment: the ones representing the true interactions and the ones that have happened randomly and can be thought as noise. The interaction models try to classify the observations into these two categories.

Interaction frequencies, i.e. the number of times a certain interaction has been observed, is the key variable in these models. The general assumption is that the interaction frequency is expected to be higher for true interactions
CHAPTER 3. LITERATURE REVIEW

Figure 3.1: Illustration of two anchor regions $R_A$ and $R_B$ with interaction count $I_{AB} = 5$ and marginal counts $c_A = 9$ and $c_B = 7$. Figure from (Li et al., 2010).

than for the random ones. If the interaction indeed is biologically meaningful, it will probably occur more often than some random parts of chromatin sticking to each other and thus it will have a high interaction frequency. In most models the basic idea is that interaction frequency of the interactions is assumed to follow some statistical distribution and the noise interaction frequencies follow some other distribution. Some of the statistical models try to answer the question "which are true interactions and which are not?" by giving a threshold value, of which higher interaction counts are labeled as true interactions and lower are labeled as noise., whereas some models attempt to give a probability of belonging to the false and true groups.

The current methods for quantifying chromatin interactions and names for the software using the methods in question are hypergeometric model by Li et al. (2010) (ChIA-PET Tool), non-central hypergeometric model by Paulsen et al. (2014) (ChiaSig) and Bayesian mixture model by Niu and Lin (2015) (MDM), mixture model combining anchor finding and interaction classification by Reeder and Gifford (2013) (Sprout) and binomial model with bias correction by Phanstiel et al. (2015) (Mango). The hypergeometric and non-centered hypergeometric model along with the binomial model are based on calculating thresholds and $p$-values for the interactions, whereas Bayesian mixture model gives distributions for the probabilities of interaction belonging either to the true interactions or random interactions group. These methods are presented in more detail in the Sections 3.1.1-3.1.5. Comparing these different methods and their advantages is in focus in Section 3.1.6.

3.1.1 Hypergeometric model

A hypergeometric model was proposed by Li et al. (2010). This model is based on the assumption that each anchor site can interact randomly with other anchors with an equal probability. The interactions are also assumed to be independent. Thus the interaction frequency of these random interactions follows hypergeometric distribution. If an interaction is not a random, but a biologically meaningful interaction, it should have a higher interaction
frequency which does not fit to this distribution. This idea is utilized by constructing a statistical test of whether a certain interaction frequency comes from the hypergeometric distribution or not.

First the needed variables are defined. $I_{AB}$ is the notation used for interaction count for interaction between anchor regions notated as $R_A$ and $R_B$. $N$ is the number of inter-ligation PETs and $c_A$ and $c_B$ are the marginal counts for fragment ends from regions $R_A$ and $R_B$ respectively. Marginal count is calculated by summing together all frequencies of the interactions the anchor region in question has been a part of. Now, under the null hypothesis, the probability that interaction count $I_{AB}$ is observed comes from the hypergeometric density function

$$P(I_{AB}|N, c_A, c_B) = \frac{\binom{c_A}{I_{AB}} \binom{2N-c_A}{I_{AB}-I_{AB}}}{\binom{2N}{c_B}}.$$ (3.1)

As marginal counts $c_A$ and $c_B$ are included in the model, the different enrichments of the anchors have been taken into account. So the interaction count is always compared to the level of the marginal counts of the anchors.

To decide if the interaction between fragments $A$ and $B$ is a true interaction or if it happened randomly, it has to be checked whether the interaction count $I_{AB}$ is larger than expected by the above model. For this purpose, a $p$-value can be calculated and a cut-off false discovery rate value can be set to distinguish the too unlikely interactions.

ChIA-PET data processing pipeline ChIA-PET Tool uses this model when distinguishing between significant and insignificant interactions. The pipeline provides all the necessary steps of processing ChIA-PET data from the raw sequence reads to the visualization of the found interactions and is widely used (Li et al., 2010).

### 3.1.2 Non-central hypergeometric model

Paulsen et al. (2014) propose an extension to the previous model. The hypergeometric distribution presented by Li et al. (2010) can be replaced with non-central hypergeometric distribution if the interaction between anchors $A$ and $B$ are more likely by factor $\omega_{AB}$ than interactions between one of these anchors and the other anchors. Using the same notation as above the probability of observing interaction count $I_{AB}$ between anchor regions $R_A$ and $R_B$ by chance becomes

$$P(n_{ij}|N, c_A, c_B, \omega_{AB}) = \sum I'_{AB} \cdot \frac{\binom{c_A}{I'_{AB}} \binom{2N-c_A}{I'_{AB}-I_{AB}} \omega_{AB}^{I_{AB}}}{\omega_{AB}^{I_{AB}}}.$$(3.2)
where the non-centering coefficients $\omega_{AB}$ depend on the genomic distance between the two anchors $A$ and $B$. This non-central hypergeometric model is conditional on $N$, $c_A$ and $c_B$, the total number of interactions and marginal counts for both anchors $R_A$ and $R_B$.

Non-centering coefficients $\omega_{AB}$ can be estimated using equation

$$\omega_{AB} = \frac{\rho_{AB}(2 \rho - \rho_A - \rho_B + \rho_{AB})}{(\rho_A - \rho_{AB})(\rho_B - \rho_{AB})}, \quad (3.3)$$

where $\rho_{AB}$ is the expected number of interactions between anchors $A$ and $B$ and $\rho_A = \sum_B \rho_{AB}$. The expected total number of interactions is a sum of all expected numbers of interactions $\rho = \sum_{A<B} \rho_{AB}$. These values $\rho_{AB}$ are determined based on the distance between interacting anchors $A$ and $B$. This arises from the knowledge that proximal genomic regions interact randomly more often than regions far away from each other. Paulsen et al. (2014) propose that the genomic distances between all possible pairs of anchors are divided into 1000 quantiles and mean interaction count is calculated for each of these intervals. A function to compute the expected interaction frequency $\rho_{AB}$ is obtained by smoothing the distribution of quantile frequencies using cubic smoothing spline.

The interaction frequencies $I_{AB}$ for random interactions are assumed to follow non-central hypergeometric distribution presented above. Interaction count should be higher than the expected value for the true interactions. To test whether the observed value $I_{AB}$ is bigger than the expected one, the following formula for one-sided $p$-value can be used

$$P_{AB} = \sum_{x \geq I_{AB}} P(x|N, c_B, c_B, \omega_{AB}) = \frac{\sum_{x \geq I_{AB}} \binom{c_A}{x} \binom{2N - c_A}{c_B - x} \omega_{AB}^x}{\sum_x \binom{c_A}{x} \binom{2N - c_A}{c_B - x} \omega_{AB}^x}. \quad (3.4)$$

A cut-off value can again be chosen to divide the observations into random and non-random interactions.

Paulsen has released a free ChIA-PET data analysis software called Chi-aSig. It utilized the non-centered hypergeometric model for distinguishing significant and insignificant interactions from each other (Paulsen et al., 2014).

### 3.1.3 Bayesian mixture model

Niu (2012) proposes a Bayesian mixture model for identifying true interactions from random noise using Bayesian framework and Markov chain Monte Carlo methods. This model, also called Marginal Count - Distance Model, has been implemented into the freely available R package MDM which classifies interactions into false and true groups.
The mixture model consists of two components: the true interactions resulting from real chromatin interactions and false interactions due to random collision. The interaction counts for both components are assumed to be Poisson distributed, but with different Poisson rate parameters. Poisson distribution has been popular for modeling next-generation sequencing count data and it is a natural choice for this purpose as well. The mixture model can be written as

\[ X_i \sim \sum_{j=0}^{1} w_{ji} f(\cdot|\lambda_{ji}), \]  

(3.5)

independently for \( i = 1, 2, ..., n \) where \( f(\cdot|\lambda) \) is (\( k - 1 \))-truncated Poisson distribution mass function. \( k \)-truncated means that the distribution is scaled to support only values of \( k \) and higher. Truncation is done so that the observed interactions with observation count lower than \( k \) can be assumed to be noise and thus left out from the analysis. The probability mass function for \( k \)-truncated Poisson distribution is

\[ \text{Poisson}_{k-1}(x; \lambda) = f(X = x|\lambda) = \frac{\lambda^x}{x!(e^{\lambda} - (1 + \lambda + \frac{\lambda^2}{2!} + \cdots + \frac{\lambda^{k-1}}{(k-1)!}))} \]  

(3.6)

for \( x = k, k+1, ... \) and \( \lambda \) is the Poisson rate parameter, which is the mean and the variance of the distribution. It is also assumed, that \( \lambda_{0i} < \lambda_{1i} \) for \( 1 \leq i \leq n \). This corresponds to the assumption that the interaction count should be higher for true interactions than the random ones. This restriction also solves the fundamental problem of label switching which is common for mixture models.

For the mixture component proportions it holds that

\[ w_{0i} + w_{1i} = 1, \]  

(3.7)

for \( i = 1, ..., n \). The weights \( w_{0i} \) and \( w_{1i} \) can be interpreted as the probabilities that observation \( i \) comes from components 0 and 1.

As the name of the model suggests, the mixture model is then implemented in Bayesian framework. Useful information of distances between anchor sites and functional genomic locations can be included in the prior distributions used in Bayesian framework. It should also be noted in the model that two interactions sharing an anchor have correlated interaction counts.

The Poisson rate parameter of component 0 is assumed to follow gamma distribution with shape parameter \( b \) and rate parameter \( \frac{b}{r_0} \)

\[ \lambda_{0i} \sim \Gamma(b, \frac{b}{r_0}), \]  

(3.8)
independently for $i = 1, 2, ..., n$. Rate parameter for component 1 follows truncated normal distribution

$$\lambda_1|\lambda_0 \sim N(\lambda_0, \sigma^2) \cdot I(\lambda_1 > \lambda_0),$$

(3.9)

independently for $i = 1, 2, ..., n$ where $\lambda_1$ always gets bigger values than $\lambda_0$, and $\sigma^2$ is the variance of the distribution. The hyperparameters $b$, $r_0$ and $\sigma$ are given uniform priors where the constants are $A = B = C = 1000$.

$$b \sim \mathcal{U}(0, A),$$

(3.10)

$$r_0 \sim \mathcal{U}(0, B),$$

(3.11)

and

$$\sigma \sim \mathcal{U}(0, C).$$

(3.12)

As the two mixing parameters are linked together through Equation 3.7, there is need to set a prior only to one of them. Niu (2012) proposes that this prior should be used for including the knowledge of distances to interesting, functional genomic locations into the model. The idea is that an interaction should be more significant if the interacting anchors are both close to genomic locations that should be interacting with each other, like for example transcription factor binding sites and transcription starting sites. These minimums of the sums of distances from anchor sites to either TFBS or TSS are calculated for each interacting pair of anchors $i$. Then the prior for the mixing parameter of the true pairs component is proportional to this minimum value $\text{dist}_i$ through beta distribution

$$w_{1i} \sim \beta(rmc_d_i \cdot d, rmc_d \cdot d),$$

(3.13)

independently for $i = 1, 2, ..., n$, where

$$rmc_d_i = \frac{mc_i}{\text{dist}_i}.$$  

(3.14)

Here $mc_i$ is the marginal count information for pair $i$ with interaction count $X_i$ and is calculated as

$$mc_i = c_{iA} + c_{iB},$$

(3.15)

where $c_{iA}$ and $c_{iB}$ are the marginal counts for the head and tail anchors $R_iA$ and $R_iB$. The ratio of marginal count to distance $rmc_d_i$ is determined as follows

$$rmc_d = \frac{1}{n} \sum_{i=1}^{n} rmc_d_i,$$

(3.16)
where \( d \) is a hyperparameter with uniform prior

\[
d \sim U(0, D).
\] (3.17)

\( D \) is set as a large constant such as 10000. The parameter \( d \) has a function in balancing the difference between \( mc_i \) and \( dist_i \). The authors of the model have acknowledged that the interaction count for pair \( i, X_i \), is included twice in \( mc_i \). To be more accurate, they tested changing the formula for \( mc_i \) to more proper

\[
mc_i = mc_{iA} + mc_{iB} - X_i,
\] (3.18)

but the change in results was very small and thus the definition in Equation 3.15 was found to be satisfactory.

Latent indicator variables \( Z_i \) are introduced for making inference from the model and they follow Bernoulli distribution

\[
Z_i \sim Bernoulli(w_{1i}).
\] (3.19)

If the probability

\[
Pr(Z_i = 1|X),
\] (3.20)

is greater than the decided cut-off value, the interaction \( i \) is a true interaction. Again, if the probability is smaller, the interaction is a false interaction. The cut-off value can be set for example 0.5. In other words the model can be written as

\[
(X_i|Z_i = j) \sim f(\cdot|\lambda_{ji}),
\] (3.21)

for \( j = 0, 1 \) and \( i = 1, \ldots, n \).

Markov chain Monte Carlo (MCMC) methods, which are introduced in detail in Chapter 4, are used for calculating the posterior probabilities. The MCMC implementation is a combination of Gibbs sampler and Metropolis-Hastings algorithms.

Niu and Lin (2015) discuss improving the Bayesian mixture model by using ChIP-Seq and gene expression data instead of the distance information in the mixture weight prior distribution.

### 3.1.4 SPROUT

Reeder and Gifford (2013) propose a mixture model for identifying self-ligations, inter-ligations and noise from each other in ChIA-PET data and other suited data types such as Hi-C. In their model they combine the anchor finding and interaction classification step to make the model statistically stronger. In the previous models the anchor sites have already been discovered by utilizing
self-ligation PETs, which are omitted from the interaction analysis. This procedure is described in detail in Section 4.1.2. The data sets SPROUT can handle consist of both self- and inter-ligations.

The notation for this model differs quite a lot from the previous models. First, a set \( R = \{r_1, ..., r_N\} \) is defined, where \( r_i = \langle r_i^{(1)}, r_i^{(2)} \rangle \) is a pair of coordinates pointing to two genomic regions coming from a paired-end read produced by ChIA-PET experiment. These interactions are assumed to come from a mixture model with \( M \) components which each represent an anchor site and set \( l = \{l_1, ..., l_M\} \) tells the locations of the putative anchor sites. The total likelihood is

\[
Pr(R, \pi, \psi, \rho, l) = \prod_{i=1}^{N} \left[ \rho \left( \sum_{j=1}^{M} \pi_j Pr(r_i|l_j) \right) + (1 - \rho) \left( \sum_{j=1}^{M} \sum_{k=1}^{M} \psi_{j,k} Pr(r_i|l_j, l_k) \right) \right]
\]

(3.22)

where for mixing parameter \( \rho \) holds \( 0 \leq \rho \leq 1 \) and the relative weights \( \pi_i \) and \( \psi_{i,j} \) sum to one \( \sum_{i=1}^{N} \pi_i = 1, \sum_{i=1}^{N} \sum_{j=1}^{N} \psi_{i,j} = 1 \).

The aim is to find minimal number of anchors that explain the data well. The reads which do not correspond to any of these anchors come from the noise component from dummy location \( l_B \), where \( B \notin \{1, ..., M\} \). The distribution of the noise component \( Pr(r_i|l_B) \) is uniform.

Latent variables \( Z = \{z_1, ..., z_N\} \) where \( z_i = \langle z_i^{(1)}, z_i^{(2)} \rangle \) are also included in the model. A pair of anchor indices \( z_i \) tells from which anchor 1, ..., \( M \) the two reads come from. Special index \( B \) corresponds to noise distribution and notation \( z_i = \langle j, - \rangle \) means that the read pair \( i \) is a self-ligation.

The estimation of the model is done with expectation-maximization method. Priors are set to the parameters in the following way. The relative weights \( \pi \) and \( \psi \) have negative Dirichlet priors

\[
Pr(\pi|\beta) \propto \prod_{j=1}^{M} \pi_j^{-\alpha},
\]

(3.23)

and

\[
Pr(\psi|\alpha) \propto \prod_{j=1}^{M} \prod_{k=1}^{M} \psi_{j,k}^{-\beta},
\]

(3.24)

where parameters \( \alpha \) and \( \beta \) specify how many reads there has to be for a putative anchor or interaction for it to stay in the model.

There are two assumptions in relation to the anchor locations: anchors are located at a particular genomic coordinates and there can only be one anchor at one location. These assumptions are included in Bernoulli prior for the
anchor site locations $l$

$$\Pr(l|k) = \prod_{i=1}^{L} k_i^{1(i \in l)} (1 - k_i)^{1(1 \not\in l)}$$

(3.25)

$$\propto \prod_{j=1}^{M} \frac{k_{ij}}{1 - k_{ij}},$$

(3.26)

where $L$ is the number of genomic coordinates and $k$ is uniformly distributed.

Mixture parameter $\rho$ has a Beta prior

$$\Pr(\rho|a, b) \sim \rho^{a-1}(1 - \rho)^{b-1}.$$  

(3.27)

Reeder and Gifford (2013) set the parameters $a$ and $b$ to ones, so the prior is reduced to uniform distribution. Expectation maximization algorithm is used for finding the parameter values $\pi$, $\psi$, $\rho$ and $l$ that maximize the posterior model likelihood.

### 3.1.5 Mango

ChIA-PET analysis pipeline Mango, proposed by Phanstiel et al. (2015), is a recent, comprehensive tool which uses binomial distribution for modeling the interaction counts. In this model, the interaction counts are assumed to follow binomial distribution for which the probability of observing interaction with interaction count exactly $k$ is

$$P(I_i = k) = \binom{N}{k} P(i)^k (1 - P(i))^{N-k},$$

(3.28)

where $P(i)$ is the probability of observing a PET $i$ connecting anchor regions $R_{Ai}$ and $R_{Bi}$ with distance $d_i$ and joint peak depth $e_i$. This probability is calculated utilizing Bayes’ theorem in the following way

$$P(i) = \frac{P(i|d_i)P(i|e)}{P(C|d_i)P(C|e)C_T},$$

(3.29)

where $C_T$ is the number of all possible anchor pairs. $P(i|d_i)$ is the probability of detecting paired-end-tag separated by distance $d_i$ and $P(i|e_i)$ is the probability of detecting a PET with joint peak depth $e_i$. $P(C|d_i)$ and $P(C|e_i)$ are again the probabilities for detecting a pair of loci with these properties. All of these probabilities can be calculated from the interaction data by using spline fitting.
To calculate the p-value of interaction $i$ having interaction count $I_i = k$ or higher is

$$P(I_i \geq k) = \sum_{t=I_i}^{N} P(I_i = t). \quad (3.30)$$

Using these formulas, p-values are calculated for all possible interactions, including those that were not even observed. Benjamini-Hochberg method is used for multiple hypothesis testing correction. The true interactions are then determined from these results by using defined false discovery rate.

Mango pipeline is a useful tool, as it can be used for all the necessary steps of ChIA-PET analysis, starting from filtering of the linkers and aligning the sequencing data to a reference genome to detecting interactions. It utilizes Bowtie, MACS2 and Bedtools software packages in the preprocessing steps and is itself implemented in R language.

### 3.1.6 Comparing the methods

The five different models for ChIA-PET analysis were presented in the earlier sections and in this section their advantages and disadvantages are discussed along with comparisons between them found in literature.

The major disadvantage of the hypergeometric and non-central hypergeometric models is that the model assumptions are not satisfied by the data, which was criticized by Niu (2012). Niu claims that the wrong assumptions made in the model are the reason why hypergeometric model performs poorly. There is also a second problem in the hypergeometric model. If it happens so that some anchor regions only appear in one interaction, i.e. $c_A = c_B = I_{AB}$, and interaction count for that interaction $I_{AB}$ is much smaller than the total number of interactions the hypergeometric probability mass function reduces to the form

$$Pr(I_{AB}|N, c_A, c_B) = \frac{\binom{c_A}{I_{AB}} \binom{2N-c_A}{c_B-I_{AB}}}{\binom{2N}{c_B}} = \frac{\binom{c_A}{I_{AB}} \binom{2N-c_A}{c_B-I_{AB}}}{\binom{2N}{c_B}} = \frac{1}{\binom{2N}{c_A}} \quad (3.31)$$

which will be close to zero. This leads to always falsely accepting such interactions as true ones.

The advantage of hypergeometric models is that they are simple to use. The usage is based on calculating p-values from (non-central) hypergeometric model and a cut-off p-value is used to separate true interactions from the false ones. Non-central hypergeometric model makes use of distance between the interacting fragments and thus gives more sophisticated results. The simple hypergeometric model does not use any information to refine the analysis.
Even so, the hypergeometric model is the established method for quantifying chromatin interactions, as it is a part of the widely used ChIA-PET Tool pipeline.

Bayesian mixture model on the other hand uses distance information to the nearest transcription starting and transcription factor binding sites. This is a way to emphasize the interactions with known biological meaning and makes the results more reasonable. Of course the information of the TSS and TFBS locations must be available so that this can be done. The downside of using distance information is that detecting interactions between unknown TSS and TFBS sites becomes unlikely. So some completely new interactions might be passed over and the already known interaction sites get more attention. There seems to be no reason to suspect that the model assumptions would not hold up. It is also possible to expand the Bayesian mixture model to compare the intensities of interactions in different data sets as Niu et al. (2014) propose. Bayesian mixture model has been implemented as an R package MDM and is easily available. Calculating the distances to the TSS and TFBS locations is what makes using the model somewhat laborious.

Bayesian mixture model has been compared to the hypergeometric model (Niu, 2012; Niu and Lin, 2015). Niu and Lin (2015) performed the comparison using Mcf-7 cell line ChIA-PET data where estrogen-α bound interactions were detected. The hypergeometric model classified most of the putative interactions as true ones, whereas Bayesian mixture model was more conservative by accepting interactions with high interaction count or $rmcd$ value or both. Out of the few interactions accepted by the mixture model and rejected by hypergeometric model some were found out to connect estrogen or estrogen-α regulated genes and a genomic region that can bind estrogen-α. This shows that hypergeometric model can easily miss biologically meaningful interactions that can be detected by Bayesian mixture model.

SPROUT model by Reeder and Gifford (2013) is quite different from the ones mentioned above, as the anchor finding step is integrated into the interaction classification step. The authors claim that this makes the method statistically stronger compared to the other models, where the data has already been processed into anchors and interaction counts. They show that the results for CTCF data set are different in many ways when compared to the hypergeometric model used by Handoko et al. (2011). SPROUT outperforms ChIA-PET Tool in consistence across replicates and in number of interactions supported by motifs found in vicinity of the interacting regions. Combining both the anchor and interaction finding steps into one eliminates the possibility of cutting out some possible interactions in the anchor finding step as might happen in case of the other methods where the two steps are separated. On
the other hand, SPROUT does not use any other information such as distance information to the biologically meaningful sites to refine the model.

Mango has been compared to the non-centered hypergeometric model and ChIA-PET Tool by Phanstiel et al. (2015). In the comparison three different data sets were analyzed with all three approaches and the results were then compared with each other and with Hi-C data to evaluate the biological relevance of the found interactions. For each of the data sets the number of found interactions was much higher for ChIA-PET Tool than for non-centered hypergeometric model and Mango. ChIA-PET Tool does not correct genomic distance bias like the two other methods, which shows in the results: ChIA-PET Tool detects a large number of interactions with considerably short distance between the interacting anchors, which are not detected by the other two methods. Comparisons to Hi-C data show that results from Mango are supported the best by Hi-C data. One flaw of Mango’s is that it cannot be used for detecting inter-chromosomal interactions. The lack of implementing this feature was justified by Phanstiel et al. (2015) by referring to previous studies, which claim that most functional interactions span under only one megabase. Leaving inter-chromosomal interactions out of the analysis also makes the computations easier and reduces the effect of multiple hypothesis testing. As Mango is a comprehensive pipeline tool, it can be used for all the needed data processing steps needed when handling ChIA-PET data. It is also more easier to install and use than ChIA-PET Tool, which makes it more usable.

A thorough comparison between all the five models with multiple data sets, real and simulated, would help to understand the differences between the models and their performances.

### 3.2 Improving the methods

This section gives an overview on different methods found in recent literature for improving ChIA-PET or other NGS data analysis methods. One way to improve statistical analysis is to remove effects of bias factors. This kind of bias removal method has already been implemented for Hi-C data and it could be applied on ChIA-PET data as well. The details of this method are discussed in the first section. Another approach to improve ChIA-PET analysis is to include external information for more comprehensive understanding.
3.2.1 Bias correction

As discussed in Section 2.1, there are multiple bias factors in the ChIA-PET experiment and in the ways the data is being processed. These factors should be taken into account in the analysis to prevent false results and in worst case even interpreting the bias as biological phenomena.

Yaffe and Tanay (2011) corrected biases for Hi-C data using contact map normalization. Whereas Hu et al. (2013) present a Bayesian model for inferring the three-dimensional structure of chromatin from Hi-C experiment data. Hu et al. (2013) assume that the number of interactions between fragments $i$ and $j$, $u_{ij}$ ($i \neq j$) follows Poisson distribution with parameter $\theta_{ij}$. The rate parameter $\theta_{ij}$ depends on the mean GC contents of the fragment ends $g_i$ and $g_j$, the mean mappability scores of the fragment ends $m_i$ and $m_j$, the marginal counts of the fragment ends $e_i$ and $e_j$ and the distance between the interacting fragment ends $d_{ij}$. The relationship between Poisson rate parameter $\theta_{ij}$ and the linear predictor is the following

$$
\log(\theta_{ij}) = \beta_0 + \beta_1 \log(d_{ij}) + \beta_{enz} \log(e_i e_j) + \beta_{gcc} \log(g_i g_j) + \beta_{map} \log(m_i m_j),
$$

(3.32)

where $\beta_1$, $\beta_{enz}$, $\beta_{gcc}$ and $\beta_{map}$ are the coefficients for the predictor values and $\beta_0$ is the constant term. The terms in the right side of this equation form the linear predictor of this model.

The Poisson distribution with this rate parameter is then used in the joint likelihood function for the data. This likelihood is part of the model which tries to capture the three dimensional structure of the chromosome. Markov chain Monte Carlo methods are applied to draw samples from the posterior distribution.

3.2.2 Including external input data

Combining different data types in one analysis could result in more comprehensive view on chromatin structure as a whole. As Li et al. (2014) ponder: to truly understand transcription regulation would require investigating DNA, RNA and proteins as a whole, not just separately.

There have already been some methods where external information has been included in the ChIA-PET analysis. He et al. (2014) tried a different approach in ChIA-PET analysis in their work and combined multiple types of data to make a deeper ChIA-PET analysis. They used ChIP-Seq, Hi-C and some more data types together to provide a better understanding of long range interactions and present analysis concentrating on estrogen receptors in different cell types such as Mcf-7. He et al. (2014) also identified many such
long-range interactions from Chip-Seq data that were not found in original ChIA-PET analysis by Fullwood et al. (2009). The integrative analysis showed that it can overcome the limitations of the single methods and cast light on the hidden patterns of long range chromatin interactions. Although the results seem promising, the obvious drawback also exists: to be able to combine many different types of data, they have to be available for the cell type of interest. For example, multiple chromatin immunoprecipitation sequencing data sets were necessary for the analysis conducted in their work (He et al., 2014).

The Bayesian mixture model proposed by Niu (2012) already includes external information of transcription starting site and transcription factor binding site locations in the model. As discussed in Section 3.1.6, this approach has been proven to produce sophisticated results and reveal relevant chromatin interactions that would be discarded by more simple models such as hypergeometric model. Niu and Lin (2015) also discuss shortly about using data from ChIP-Seq and gene-expression experiments to improve the calculation of $dist_i$ in the Bayesian mixture model even further. The possible model or results have not yet been published.

### 3.3 Result validation

Validation of the found interactions from real data sets is hard because the true chromatin structure is not known. That is why error rates or such quantities often used for describing the performance of the methods cannot be calculated in a genomic-wide manner.

Interactions found in ChIA-PET data analysis and their functions can be validated by doing other wet-lab experiments. 3C experiment could be used for interactions between short genomic distance. As for long-range interactions, such as inter-chromosomal interactions, DNA Fluorescence in situ hybridization (DNA-FISH) could be used even though the resolution is limited. These techniques require the possibility of wet-lab experiments or ready-made data sets for the same cell type the ChIA-PET experiment was done for. Because of lack of large-scale validation methods, only a fraction of all the found interactions have been validated (Li et al., 2014).

After conducting an interaction analysis, it should also be checked that the found interactions are reasonable. This means that interactions detected from mis-mapped data or noise should be removed along with obviously false observations, e.g. interactions involving Y chromosome in breast cancer cells. The genome rearrangements in for example cancer cell-lines should also be taken into account (Li et al., 2010).

If interaction validation is not possible when comparing different methods,
only the numbers of found interactions and whether different methods found
the same interactions or not can be compared. One way to go around the
validation problem is to use simulated data. This is a simple way to make sure
a method works when it is still in development stage. Then again the results
might be only too good as the observations are simulated straightforwardly
from the model which is being tested. Simulations are still commonly used to
test the methods and to give error rates and such indicators of the goodness
of the performance. For example, Li et al. (2010), Paulsen et al. (2014) and
Niu (2012) all use simulated data to show how well their methods perform.
Chapter 4

Methods

This chapter describes the methods needed for analyzing ChIA-PET data. Data processing starts from taking the raw data that has been sequenced. As depicted in Figure 4.1, first step is to filter the PETs depending on the linkers and mapping the data to the reference genome. After sequences that were not mapped properly have been eliminated, the mapped PETs are divided into self- and inter-ligations. Self-ligations are used for finding anchor sites and inter-ligation PETs are utilized in interaction analysis. The results of the interaction analysis can later be visualized in the end if desired (Li et al., 2014).

The first section runs through the outline of data processing in such detail that reproducing the results is possible. After data preparation the actual model where Bayesian mixture model and Poisson regression for bias factors are combined is provided. Focus is given to the necessary theory along with the assumptions and prior information needed for constructing this model. The resulting model and Bayesian framework needed for posterior sampling is presented in section 4.2.

The estimation of model parameters is conducted using Bayesian methods. The principles of Bayesian analysis are presented in Chapter 4.3 together with a short explanation of how Markov chain Monte Carlo methods work. The Section 4.4 explains how the MCMC methods are utilized in a sampler that is used for estimating the parameters and the model distributions are derived in detail. The last Section 4.5 describes the implementation of the sampler.

4.1 Preprocessing of the data

This section elaborates the steps of preprocessing the ChIA-PET data prior to the actual interaction analysis. Preprocessing methods dictate how the final data set used for the analysis looks like and hence have a remarkable effect on
the result. It is important to make clear the choices made in this part of the process to enable repeating the analysis and results later and for understanding what the data in question tells about chromatin interactions.

The resulting file from the preprocessing steps will contain all the needed information for the model presented in Section 4.2. This includes interaction head and tail sites, interaction frequencies and marginal counts. GC content, coverage and mappability of each anchor site are also given in the file.

4.1.1 Processing raw ChIA-PET data

The data obtained from ChIA-PET sequencing needs to undergo certain preparation steps even before aligning to reference genome. Raw data files consist of paired-end reads where the two half linkers separate 20 base pairs long tags. The full-length linker formed by the two half-linkers is 38 base pairs long. The first step is to align the half-linkers to the chosen half-linker sequences. The PETs with non-readable or non-mappable linkers are dropped. Then it is checked whether the linkers in each PET are of the same kind or not, i.e. if the PETs are chimeric or not. Chimeric PETs are assumed to be caused by non-specific proximity ligation and should be removed in this phase. The paired-end tags with homogeneous linkers are then mapped to the reference genome. First it is checked, whether an exact match for a tag is found in the
reference genome. If none is found, mapping is tried with one mismatch allowed. In the end the PETs which cannot be mapped to the reference genome uniquely are removed from the data set (Li et al., 2010).

When the paired-end reads have been mapped to the reference genome, the data can be further divided into different kinds of interactions depending on the distance between the interacting anchor sites. As illustrated in Figure 2.1, the two main categories are self-ligation PETs and inter-ligation PETs. Some cut-off value is decided and the interactions with the distance under this threshold are considered self-ligations. Inter-ligations are thought to represent the putative interactions, whereas self-ligations are thought to appear on genomic locations where the chosen protein of interest possibly binds. As shown in the Figure 2.1, there should be a self-ligation PET cluster in the both ends of the interaction. Thus inter-ligation PETs are utilized later on in the actual interaction analysis step and self-ligation PETs are used in the next step of quantifying the anchor sites between which the interactions happen (Li et al., 2010).

4.1.2 Anchor site finding

The anchor finding principles described by Paulsen et al. (2014) are used as a basis in this phase. First, the self-ligations are extracted from the original data set as described in previous section. The cut-off distance between head and tail fragments used for dividing the intra-chromosomal interactions into self- and inter-ligations is 8000 base pairs as suggested by Paulsen et al. (2014). Next, self-ligations need to be transformed from paired-end data into single chromosome regions. This step is not specified by Paulsen et al. (2014), but the following transformation is used: each self-ligation interaction is transformed into a region, which starts from the head fragment starting point and ends to tail fragment end point. Model-based Analysis of ChIP-Seq 2 (Zhang et al., 2008), in short MACS2, analysis can be performed on the resulting data set. MACS2 searches for self-ligation clusters such as illustrated in Figure 2.1 and returns the coordinates for these clusters. In order to find self-ligation peaks, MACS2 slides a window of a certain size through the genome and finds locations with high self-ligation enrichment. More specifically, the enrichments are compared to a local Poisson rate parameter MACS2 estimates from a small region surrounding the putative peak. The null hypothesis is that self-ligation enrichment follows Poisson distribution with this local rate parameter. A p-value can then be calculated for each putative peak and if p-value is lower than the defined threshold value, null hypothesis is rejected. This means that the enrichment was higher than expected in this small region and there is a peak in this location (Zhang et al., 2008).
MACS2 reports peak summits’ genomic coordinates, which are expanded to form wider anchor regions. The peaks within 1000 base pairs from each other are combined by assigning the starting point of the first peak as the anchor starting point and end of the last peak as the anchor end site. The anchors with range under 1000 base pairs are expanded by 1000 base pairs to both directions. If some anchors overlap after the expansion step, they are combined in the same manner as the peaks were combined before.

4.1.3 Calculating anchor properties

Anchor GC content, mappability and coverage are needed for using the model specified in Section 4.2. In this section the computational tools used for calculating these quantities and their definitions are briefly explained.

The anchor site GC contents can be easily calculated using ready-made tool Bedtools by Quinlan and Hall (2010) which includes a useful command nuc for this purpose. The command nuc calculates the proportion of guanine and cytosine for given genomic intervals by using a reference genome. Anchor mappability is defined here as the mean of the mappability of all possible \( k \)-mers, i.e. sequences of length \( k \), that could begin from an anchor site. The mappabilities for \( k \)-mers starting from each point inside an anchor are calculated and the mean of these values is taken to obtain the anchor mappability. Mappabilities for \( k \)-mers starting from each site of anchor are calculated by using Bedtools intersect function to find the overlapping sites of the anchor file and a mappability track. Some mappability tracks are available in UCSC Genome Browser collections by Derrien et al. (2012). As building a new mappability track which would exactly fit the data in question is quite laborious, the mappability quantities are approximated by using a ready-made alignability track by Derrien et al. (2012). The track is aimed for reads of length 24 base pairs instead of 20 base pairs, which is the exact length of the tags in ChIA-PET data. The track used for this thesis is an alignability track, which tells how uniquely a \( k \)-mer sequence can be aligned to the reference genome.

Anchor coverage is also calculated for each anchor. Coverage means the number of self-ligations found in the anchor area. Bedtools command intersect is used to find the number of overlapping self-ligations for each anchor. Coverages are used as independent replacements for marginal counts in the model provided in Section 4.2. Using marginal counts as linear predictors of interaction counts is questionable, as marginal counts are sums of interaction counts, whereas coverages are highly correlated with marginal counts but are independent of interaction counts.
4.1.4 Finding the corresponding anchors for interaction head and tail sites

Interaction fragment ends were mapped to the found anchor sites. The overlapping anchors for head and tail sequences were found using again Bedtools intersect command (Quinlan and Hall, 2010). Only interactions where corresponding anchors were found for both head and tail sites are used in further analysis. The number of observed interactions between each anchor pair was counted and this number is called interaction count. Marginal counts were also computed at the same time. As earlier defined in Section 3.1.1, marginal count means the number of occurrence of a single anchor site and is the sum of interaction counts for all the interactions an anchor is a part of.

4.2 Poisson mixture model with Poisson regression

The aim is to model the interaction count data with a Poisson mixture model with two components. The model is based on the Bayesian mixture model proposed by Niu and Lin (2015) and Poisson regression model proposed by Hu et al. (2013). Poisson regression is included for predicting the rate parameters of the mixture component Poisson distributions.

The set $\mathbf{I} = \{I_i, i = 1, 2, ..., n\}$ of random variables represents the interaction frequencies for each observed anchor pair $i$ of the data set with $n$ unique anchor pairs in total. Interaction $i$ connects anchor sites $R_{iA}$ and $R_{iB}$. Only observations with count $I_i > k$ are taken into consideration in the analysis and the ones with lower count are thought to be just random interactions.

The first mixture component corresponds to the random interactions, the noise model, and the second component corresponds to the true interactions. The distribution of interaction frequency $I_i$ can be written as a sum of two probability mass functions

$$ I_i \sim w_{0i} f(\cdot|\lambda_{0i}) + w_{1i} f(\cdot|\lambda_{1i}), \text{ for } i = 1, ..., n, \tag{4.1} $$

where $w_{ji}, j = 0, 1$ is the mixing parameter, $f(\cdot|\lambda_{ji}), j = 0, 1$ is the Poisson distribution probability mass function and $\lambda_{ji}, j = 0, 1$ is the Poisson parameter. It holds for the mixing parameters that

$$ w_{0i} + w_{1i} = 1 \text{ for } i = 1, ..., n. \tag{4.2} $$

The mixing parameter $w_{ji}$ can be interpreted as the probability of observation $i$ coming from component $j$. 
A \((k - 1)\)-truncated version of Poisson distribution would be more proper instead of the usual Poisson distribution, as it is known that only interactions with minimum interaction frequency of \(k\) are included in the data set. Probability mass function for \((k - 1)\)-truncated Poisson distribution is defined in Equation 3.6.

Interaction frequencies are expected to be higher for true interactions than for the random ones. So the Poisson parameters of the first component, the noise component, should be restricted to be lower than the parameter for the second component representing the true interactions

\[
\lambda_{0i} < \lambda_{1i}, \quad (4.3)
\]

for \(i = 1, \ldots, n\).

Poisson regression is used in times when the realizations of the response variable to be explained are positive integers and follow Poisson distribution (Dunteman and Ho, 2006). In this model some bias factors along with anchor coverage, which is an independent replacement for marginal counts, are incorporated into the Poisson parameters through Poisson regression. This regression model is good for predicting the Poisson rate parameters of the first mixture component. If the bias factors used as linear predictors truly can explain variations in the data which is considered as noise, they can be used to predict the random proportion of the interaction frequencies.

In Poisson regression the linear predictors are used for predicting the expected interaction frequency

\[
E(I_i) = \lambda_{0i}, \quad i = 1, 2, \ldots, n, \quad (4.4)
\]

assuming the \(i\)th putative interaction is not a real interaction. The link function describes the connection between linear predictors and logarithm of the rate parameter. Link function is in the case of this model

\[
\log(\lambda_{0i}) = \beta_0 + \beta_1 \log(d_i) + \beta_2 \log(m_{iA}m_{iB}) + \beta_3 \log(g_{iA}g_{iB})
\]

\[
+ \beta_4 \log(c_{iA}c_{iB}), \quad \text{for } i = 1, \ldots, n, \quad (4.5)
\]

where \(d_i\) is the distance between the interaction head start site and tail start site, \(m_{iA}\) and \(m_{iB}\) are the mappabilities for the head (A) and tail (B) fragments, \(g_{iA}\) and \(g_{iB}\) are the GC contents of head and tail fragments and \(c_{iA}\) and \(c_{iB}\) are the coverages for head and tail fragments divided by the fragment lengths. Link function of the same form was used by Hu et al. (2013) for bias correction in three dimensional modeling of chromatin structure as discussed before in Chapter 3.2.1.
Link function can also be written with vector notation. Vector
\[ \mathbf{x}_i = (1, x_{i1}, x_{i2}, x_{i3}, x_{i4})^T, \]
(4.6)
is the explanatory variable vector with 4 variables. The first element 1 represent intercept (Dunteman and Ho, 2006). The explanatory variables here are \( x_{i1} = \log(d_i), \ x_{i2} = \log(m_iA m_iB), \ x_{i3} = \log(g_iA g_iB) \) and \( x_{i4} = \log(c_iA c_iB). \) The coefficient vector \( \beta \) is
\[ \beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)^T. \]
(4.7)

So the Poisson regression model can also be written in vector form as
\[ \lambda_{0i} = \exp(\mathbf{x}_i^T \beta). \]
(4.8)

The maximum likelihood estimation of the Poisson regression coefficients \( \beta \) is a non-linear optimization problem. In this thesis the estimation is done by using MCMC methods, but some preliminary results are also produced by using maximum likelihood estimates.

The Poisson parameter for the second mixture component is a sum of \( \lambda_{0i} \) defined above and \( D_i \)
\[ \lambda_{1i} = \lambda_{0i} + D_i, \text{ for } i = 1, \ldots, n, \]
(4.9)
where \( D_i \) represents the addition to the background model to obtain the true interaction distribution rate parameter. It cannot be expected that the amount of the addition would be the same for all the interactions so there is a \( D_i \) parameter for each interaction \( i \). To simplify the notations, from now on the rate parameter of the first component is written as \( \lambda_i \) and for the second component \( \lambda_i + D_i \).

The model hierarchy is illustrated in Figure 4.2, where circles represent model parameters to be estimated and boxes are the available data, which is used for the estimation. The model parameters are estimated by using MCMC methods. The prior distributions needed in the Bayesian approach are presented and justified next.

The prior distributions play an important part in Bayesian analysis, as useful prior information can be included into the model through them. For example, Niu (2012) incorporated the fragment distances to known functional loci and marginal counts into the priors of the mixing parameters.

Poisson regression coefficients and intercept have zero-centered normal priors. Because the expected signs or magnitudes of the coefficients are not known, the coefficient values should not be restricted heavily by the prior. The intercept \( \beta_0 \) and coefficients \( \beta_1, \ldots, \beta_4 \) have priors
\[ \beta_i \sim N(0, \sigma_b^2), \]
(4.10)
Figure 4.2: Model hierarchy. Circles illustrate model parameters and boxes are data or other prior information.

for $i = 0, ..., 4$. $\sigma_b^2$ is variance parameter, which should be given a moderate value that still enables covering the range of all likely values for the coefficients.

Prior distribution for $D_i$ is log-normal distribution

$$D_i \sim \log N(\mu_D, \sigma_D^2), \quad (4.11)$$

for $i = 1, ..., n$. Log-normal distribution supports only positive values, which goes along with that $D_i$ should only get positive values as it is expected that the second component rate parameter is as high or higher than the first component rate parameter. If $D_i$ is log-normal, then logarithm of $D_i$ is normal distributed

$$\log(D_i) \sim N(\mu_D, \sigma_D^2), \quad (4.12)$$

for $i = 1, ..., n$. This knowledge is exploited in the implementation because of computational advantages. The mean of log-normal distribution is calculated from parameters $\mu_D$ and $\sigma_D^2$ as follows

$$E[D_i] = \exp(\mu_D + \frac{\sigma_D^2}{2}), \quad (4.13)$$

and variance is

$$\operatorname{Var}[D_i] = (\exp(\sigma_D^2) - 1) \exp(2\mu_D + \sigma_D^2). \quad (4.14)$$

The hyperpriors for $\mu_D$ and $\sigma_D$ should be chosen to allow a reasonable range for the values of $D_i$. Mean $\mu_D$ is given a normal prior

$$\mu_D \sim N(\mu_{\mu_D}, \sigma_{\mu_D}^2), \quad (4.15)$$
where $\mu_{\mu_D} = 2$ and $\sigma_{\mu_D}^2 = 2^2$. Gamma prior is chosen for $\sigma_D$

$$\sigma_D \sim \text{Gamma}(a_{\sigma_D}, b_{\sigma_D}),$$  \hspace{1cm} (4.16)

where $a_{\sigma_D} = 2$ and $b_{\sigma_D} = 3$.

A large fraction of the data ChIA-PET experiment produces is due to random collisions, so it can be assumed that most of the observed interactions are noise. That is why the prior distribution of $w_{0i}$, the mixing parameter for the first component, should be skewed to being closer to 1 than 0. To represent this assumption, prior for mixing parameter $\theta_{0i}$ is chosen to be

$$w_{0i} \sim \text{Beta}(\alpha_{w}, \beta_{w}), \text{ for } i = 1, \ldots, n.$$  \hspace{1cm} (4.17)

Beta distribution was chosen as prior because its realizations limit between 0 and 1 and the parameters for the distribution can be chosen to represent the prior assumptions. For example, with parameter values $\alpha_{w} = 7$ and $\beta_{w} = 3$ beta distribution is strongly skewed towards values closer to one than zero. The mixing parameter for the second component $w_{1i}$ can be solved from Equation 4.2 in the following way

$$w_{1i} = 1 - w_{0i}, \text{ for } i = 1, \ldots, n,$$  \hspace{1cm} (4.18)

so there is no need to set a separate prior distribution for it.

The model distributions are written and computational aspects are discussed in more detail in section 4.4.

### 4.3 Bayesian inference and MCMC methods

Bayesian data analysis is a process, where a full probability model is set up and is conditioned on the observed data. The resulting posterior distribution is a conditional distribution for the unobserved quantities conditioned on the available data. The unobserved quantities of interest might e.g. be the model parameters or new data points. The idea is to quantify the uncertainty by using whole probability distributions to describe a quantity. In this section the above-mentioned general data analysis process is introduced along with the computational methods used when handling more complicated models. The book by Gelman et al. (2014) has been used as a main reference in this section.

First, a joint probability distribution for parameter or parameters $\theta$ and observed values $y$ is set up. This probability mass or density function $p(\theta, y)$
can be defined as a product of the prior and sampling distributions $p(\theta)$ and $p(y|\theta)$ in the following way

$$p(\theta, y) = p(\theta)p(y|\theta). \quad (4.19)$$

All the knowledge about the process in question should be integrated into the prior and sampling distributions. Prior distribution is a distribution that sums up the assumptions about the distribution of the parameter $\theta$. In the case there is no prior information about $\theta$ and its distribution, a non-informative prior should be used. Noninformative prior is often a flat distribution, which covers all the possible values of the quantity and does not restrict or give different weights to the values.

After this the posterior density can be calculated by using Bayes’ rule

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)}, \quad (4.20)$$

where $p(y)$ is either a sum over all possible values of $\theta$

$$\sum_{\theta} p(\theta)p(y|\theta), \quad (4.21)$$

for discrete $\theta$ or in case of a continuous parameters an integral

$$\int_{\theta} p(\theta)p(y|\theta)d\theta. \quad (4.22)$$

In simplified, unnormalized case $p(y)$ can be left out of the Equation 4.20 as a constant term, which leads to unnormalized posterior density

$$p(\theta|y) \propto p(\theta)p(y|\theta). \quad (4.23)$$

This form is often used because it is easier to compute. Computing with log-densities is also preferred over the usual density functions, as this substitution reduces numerical problems such as over- and underflows.

The posterior distribution cannot always be computed analytically in closed form. As this is only possible for nice and simple cases, simulation techniques are often used to estimate the posterior distributions in more complex situations. This is why simulation techniques are a fundamental part of applied Bayesian analysis. Simulation can also be used for computing the posterior predictive distribution, which means predicting new, observable data points conditioned on the available observations.

The simulation techniques used in this thesis are Markov chain Monte Carlo, in short MCMC, methods. Gibbs sampler and Metropolis algorithms
are the main methods used. Hamiltonian Monte Carlo and No-U-Turn sampler, which are extensions of Metropolis algorithm, are also tested for the estimation problem in this thesis.

MCMC methods are based on the idea that several independent sequences are produced by a Markov process with a certain stationary distribution. These Markov chains start from certain starting points and in each time step new values are drawn to the sequences from a transition distribution. Because the sequence is a Markov chain, the new value depends only on the previous sequence value. The aim is to choose the transition distribution so that in the end the chain will converge to its unique stationary distribution, which is the posterior distribution. The samples after convergence are from this posterior distribution of interest.

Two commonly used MCMC methods are Gibbs sampler and Metropolis algorithm. They are useful when handling relatively simple models. The general Gibbs sampling algorithm starts by dividing the parameter vector $\theta$ into $d$ subvectors: $\theta = (\theta_1, ..., \theta_d)$. Then in one iteration $t$, each of these subvectors $\theta^t_j$ are sampled in turn from conditional distribution, conditioned on all the current values of the other vectors and the observed values $y$.

$$p(\theta_j | (\theta^t_1, ..., \theta^t_{j-1}, \theta^t_{j+1}, ..., \theta^t_d), y),$$

where the components with upper index $t$ have already been updated and those with $t - 1$ will be updated after $\theta_j$.

Metropolis algorithm is a random walk algorithm, where next value is drawn from proposal distribution. An acceptance rule is used to decide whether the drawn value is accepted as a new value or not. The acceptance rule takes care of that the sequence converges to the posterior distribution. It can be proved that these sequences are Markov chains and their stationary distribution is the targeted posterior distribution. First, some starting points are either drawn from a starting distribution $p_0(\theta)$ or chosen according to some prior assumptions so that $p(\theta_0 | y) > 0$. The actual sampling starts after this first step, which is counted as time point $t = 0$.

For time steps $t = 1, ..., N$: New proposal $\theta^*$ is drawn from jumping distribution $J_t(\theta^* | \theta^{t-1})$. The ratio of model posteriors with parameters $\theta^*$ or $\theta^{t-1}$ and observed values $y$ are calculated

$$r = \frac{p(\theta^* | y)}{p(\theta^{t-1} | y)}.$$  

(4.25)

The proposed values are set as new current values $\theta^t$ with probability $\min\{r, 1\}$. If proposition is rejected, the current value $\theta^{t-1}$ will be set as $\theta^t$. The better the posterior probability with the new parameters compared to the previous values, the more likely proposition will be set as the new value.
Jumping distribution should be symmetrical in Metropolis algorithm, so that \( J_t(\theta_A | \theta_B) = J_t(\theta_B | \theta_A) \). If jumping distribution is asymmetric, it has to be taken into account when calculating the ratio \( r \). This version of the algorithm is called Metropolis-Hastings.

When using MCMC techniques, e.g. Metropolis algorithm, it is necessary to do convergence diagnostics for the drawn samples. As sequences’ first samples are often influenced by the chosen starting point for the sequence, they are usually removed before further inference is done based on the samples. A general, although quite conservative customary, is to remove the first half of the samples. This first half of the sequences is called warm-up. Sometimes the samples from a converged sequence are also thinned, meaning that only every \( k \)th sample is saved after warm-up phase while the others are discarded to reduce correlation between samples and to save space.

In MCMC methods it is usual to sample several sequences from the same model, starting from different initial points chosen to represent the posterior space as well as possible. This way the convergence of a single chain can be measured as the ratio of within-sequence variance and between-sequence variance. If the variances are the same, the chains have converged. Quantity called scale reduction, \( \hat{R} \), is used to describe this ratio. \( \hat{R} \) should be near 1 when the sequence is converged and generally the value 1.1 is used as threshold value. Also, checking the changes in the model likelihood gives an idea whether the target distribution has been found or not. If the likelihood values have no trend and they oscillate moderately around some value, the sequence has converged. Same kind of check should be done for sampled values. Oscillation of the likelihood and sampled values shows that the whole posterior distribution is being covered. The simulation draws obtained after warm-up are often autocorrelated. Effective sample size, \( n_{\text{eff}} \), is a measure of how many of the samples for a certain variable are independent.

When it has been made sure that the simulation method has converged to the posterior distribution, the samples obtained from the simulations must be summarized in some way to make inference on the results. Looking at a histogram of posterior samples gives a general overview on the distribution. From histogram it is possible to see the number of modes and levels of kurtosis, deviation and symmetry. Mean is usually a good point summary of a symmetric distribution. Other point summaries can be sample modes or quantiles.

Efficiency is an important aspect when choosing the sampling algorithm. Metropolis algorithm and Gibbs sampler might be too inefficient at times. They can be used as building blocks for more efficient methods and they can also be tuned. For example the Hamiltonian Monte Carlo (HMC) and no-U-turn-sampler (NUTS) can be much more efficient, tuned versions of Metropolis
algorithm. Hamiltonian Monte Carlo systematically explores the state space instead of random-walk by using Hamiltonian mechanics. (Brooks et al., 2011) NUTS is again based on HMC, but it differs from it by avoiding the difficulty of tuning the sampler parameters, which is the problem of Hamiltonian Monte Carlo (Hoffman and Gelman, 2014).

Checking the model rationality is an important part of Bayesian analysis, as it should be for all statistical analysis methods. After building a Bayesian model and computing the posterior distribution, it must be checked that the results fit to the data and to all the prior knowledge about the problem. This is also a part of model building, as bad fit is a sign that some part of the model might be wrong. The fault might be in prior distribution, hierarchical structures, explanatory variables in regression or other structures of the constructed probability model. It would be good to check that all the knowledge was properly captured by the model. Usually the rationality of the posterior inference can be assessed through the additional information the modeler has about the problem. If something strange is noticed, constructing a new, more accurate model should be considered.

There are many kinds of model diagnostics. Sensitivity analysis is for checking whether other plausible models would give the same posterior inferences. Other plausible models might have entirely different structure or differ in prior or sampling distribution or have different predictor variables. In statistical modeling it is not out of ordinary that multiple models are plausible.

The model can also be used to predict completely new observations, which were not used for constructing the model. The predicted and new data is then compared and if the model is a good one, it should fit well to the new observations. This is called external validation. Usually new data is not available and this technique must be approximated by using the data at hand. Replicated data can be generated from posterior predictive distribution and it should be similar to the observed ones. If there are some systematic differences between the data, it is a sign that the model does not fit well to the data it was created from. One easy way to make sure the model fits is to check whether the histogram of these samples resembles the posterior distribution density function.

4.4 Distributions used in sampler

In this section, the distribution needed in the MCMC sampler used to estimate the parameters of the model described in section 4.2 are derived and simplified for efficient calculations. Gibbs sampler and Metropolis algorithm introduced in Section 4.3 are used as building blocks for sampling from the
distributions of the different random variables. Gibbs sampler is used for the mixture weights $w_{0i}$ and Metropolis algorithm for $D_i$ and $\beta_i$. The actual distribution of $w_{0i}$ conditioned on the other variables and data has to be derived for Gibbs sampling. For the Metropolis algorithm it is sufficient to choose a reasonable jumping distribution and calculate the likelihood of the model given the data and current parameter values. Next these needed distributions are presented.

Conditional distribution for mixture parameter $w_{0i}$ is needed for Gibbs sampling. Beta distribution was chosen as a prior distribution for the weights $w_{0i}$, $i = 1, ..., n$

$$p(w_{0i}) = \text{Beta}(w_{0i}; a, b) = B(a, b)w_{0i}^{a-1}(1 - w_{0i})^{b-1}, \quad (4.26)$$

where $a$ and $b$ are the parameters of beta distribution and

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a + b)}, \quad (4.27)$$

is the beta function. $\Gamma$ function is defined as an improper integral

$$\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}dx. \quad (4.28)$$

The total probability of interaction count data point $I_i$ given the model and parameter values is

$$p(I_i|\lambda_i, D_i, w_{0i}) = w_{0i}\text{Poisson}(I_i|\lambda_i) + (1 - w_{0i})\text{Poisson}(I_i|\lambda_i + D_i). \quad (4.29)$$

Now the equations above can be used in Bayes’ rule given in Equation 4.20 for calculating the posterior distribution $p(w_{0i}|y_i, \lambda_i, D_i)$

$$p(w_{0i}|I_i, \lambda_i, D_i) = \frac{p(I_i|\lambda_i, D_i, w_{0i})p(w_{0i})}{p(I_i|\lambda_i, D_i)}, \quad (4.30)$$

where

$$p(I_i|\lambda_i, D_i) = \int_0^1 p(I_i|\lambda_i, D_i, w_{0i})p(w_{0i})dw_{0i}. \quad (4.31)$$

As the data points $I_i$ can be thought as constants and parameter $w_{0i}$ is integrated out in the equation, the factor $p(I_i)$ can be omitted. This gives unnormalized posterior density

$$p(w_{0i}|I_i, \lambda_i, D_i) \propto p(I_i|\lambda_i, D_i, w_{0i})p(w_{0i}). \quad (4.32)$$

By substituting the beta prior distribution density function given above in Equation 4.26 and taking advantage of the product of density function $\text{Beta}(w_{0i}; a, b)$
and \( w_{0i} \) and \( 1 - w_{0i} \) the posterior distribution can be written in the following way

\[
p(w_{0i} | I_i, \lambda_i, D_i) \propto \left[ w_{0i} \text{Poisson}(I_i; \lambda_i) + (1 - w_{0i}) \text{Poisson}(I_i; \lambda_i + D_i) \right] \text{Beta}(w_{0i}; a, b)
\]

\[
+ w_{0i} \text{Beta}(w_{0i}; a, b) \text{Poisson}(I_i; \lambda_i)
\]

\[
+ (1 - w_{0i}) \text{Beta}(w_{0i}; a, b) \text{Poisson}(I_i; \lambda_i + D_i)
\]

\[
\propto \text{Beta}(w_{0i}; a + 1, b) \frac{B(a + 1, b)}{B(a, b)} \text{Poisson}(I_i; \lambda_i)
\]

\[
+ \text{Beta}(w_{0i}; a, b + 1) \frac{B(a, b + 1)}{B(a, b)} \text{Poisson}(I_i; \lambda_i + D_i).
\]

(4.33)

So the posterior is a mixture of two beta distributions \( \text{Beta}(w_{0i}; a + 1, b) \) and \( \text{Beta}(w_{0i}; a, b + 1) \). The unnormalized weights for this beta mixture model are

\[
w_{\beta1UNi} = \frac{B(a + 1, b) \text{Poisson}(I_i; \lambda_i)}{B(a, b) p(I_i)},
\]

(4.34)

and

\[
w_{\beta2UNi} = \frac{B(a, b + 1) \text{Poisson}(I_i; \lambda_i + D_i)}{B(a, b) p(I_i)},
\]

(4.35)

where \( p(I_i) \) is the total probability of \( I_i \) given in Equation 4.31.

The weights should be normalized to sum to one as \( p(w_{0i} | I_i, \lambda_i, D_i) \) is a probability distribution. Next the weights are normalized by scaling them with the sum of the two weights.

\[
w_{\beta1} = \frac{w_{\beta1UN}}{w_{\beta1UN} + w_{\beta2UN}}
\]

\[
= \frac{B(a + 1, b) \text{Poisson}(I_i; \lambda_i)}{B(a + 1, b) \text{Poisson}(I_i; \lambda_i) + B(a, b + 1) \text{Poisson}(I_i; \lambda_i + D_i)},
\]

(4.36)

and respectively for \( w_2 \)

\[
w_{\beta2} = \frac{w_{\beta2UN}}{w_{\beta1UN} + w_{\beta2UN}}
\]

\[
= \frac{B(a, b + 1) \text{Poisson}(I_i; \lambda_i + D_i)}{B(a + 1, b) \text{Poisson}(I_i; \lambda_i) + B(a, b + 1) \text{Poisson}(I_i; \lambda_i + D_i)}.
\]

(4.37)

The formulas above can be simplified even further by writing

\[
B(a + 1, b) = \frac{\Gamma(a + 1) \Gamma(b)}{\Gamma(a + 1 + b)} = \frac{\Gamma(a) \Gamma(b)(a + 1)}{\Gamma(a + 1 + b)}
\]

(4.38)
and
\[ B(a, b + 1) = \frac{\Gamma(a)\Gamma(b + 1)}{\Gamma(a + b + 1)} = \frac{\Gamma(a)\Gamma(b)\Gamma(b + 1)}{\Gamma(a + b + 1)}. \] (4.39)

So the term \( \frac{\Gamma(a)\Gamma(b)}{\Gamma(a + b + 1)} \) is a common divisor and will be reduced when calculating normalized weights. The final forms of the reduced formulas are

\[ w_{\beta 1 i} = \frac{w_{\beta 1 UN i}}{w_{\beta 1 UN i} + w_{\beta 2 UN i}} \]
\[ = \frac{(a + 1)\text{Poisson}_0(I_i; \lambda_i)}{(a + 1)\text{Poisson}_0(I_i; \lambda_i) + (b + 1)\text{Poisson}_0(I_i; \lambda_i + D_i)}, \] (4.40)

and

\[ w_{\beta 2 i} = \frac{w_{\beta 2 UN i}}{w_{\beta 1 UN i} + w_{\beta 2 UN i}} \]
\[ = \frac{(b + 1)\text{Poisson}_0(I_i; \lambda_i + D_i)}{(a + 1)\text{Poisson}_0(I_i; \lambda_i) + (b + 1)\text{Poisson}_0(I_i; \lambda_i + D_i)}. \] (4.41)

The posterior probabilities conditioned on data \( I \) are needed in both Gibbs sampling and Metropolis algorithm. \( w \) and \( D \) are vector forms of model parameters \( w_{0i} \) and \( D_i \), \( i = 1, \ldots, n \). The vector \( \beta \) is defined in Equation 4.7. According to the Equation 4.20, the posterior density function without specifying the distributions is

\[ p(w, D, \beta | I) = \frac{\prod_{i=1}^n p(I_i | w_{0i}, D_i, \beta) \prod_{i=1}^n p(w_{0i}) \prod_{i=1}^n p(D_i) \prod_{j=1}^K p(\beta_j)p(\mu_D)p(\sigma_D)}{p(I)}. \] (4.42)

where \( I \) is a data vector of length \( n \). Unnormalized posterior density shown in Equation 4.23 can then be calculated as follows

\[ p(w, D, \beta | I) \propto \prod_{i=1}^n \left[ w_i \text{Poisson}_0(I_i; \lambda_i) + (1 - w_i) \text{Poisson}_0(I_i; \lambda_i + D_i) \right] \]
\[ \prod_{i=1}^n p(w_{0i}) \prod_{i=1}^n p(D_i) \prod_{j=1}^K p(\beta_j)p(\mu_D)p(\sigma_D). \] (4.43)

When the formula for Poisson distribution is substituted, the posterior can be written as

\[ p(w, D, \beta | I) \propto \prod_{i=1}^n \left[ w_i \frac{\lambda_i^{I_i}}{I_i!(\exp(\lambda_i) - 1)} + (1 - w_i) \frac{(\lambda_i + D_i)^{I_i}}{I_i!(\exp(\lambda_i + D_i) - 1)} \right] \]
\[ \prod_{i=1}^n p(w_{0i}) \prod_{i=1}^n p(D_i) \prod_{j=1}^K p(\beta_j)p(\mu_D)p(\sigma_D). \] (4.44)
As the log-probabilities are easier to handle computationally and is easier to calculate, logarithm is taken of the posterior and log-probability is obtained

\[
\log(p(w, D, \beta | I)) \propto \sum_{i=1}^{n} \log \left[ w_i \frac{\lambda_i^I \Gamma(I_i)}{\Gamma(I_i!)(\exp(\lambda_i) - 1)} \right. \\
\left. + (1 - w_i) \frac{(\lambda_i + D_i)^I_i}{\Gamma(I_i!(\exp(\lambda_i + D_i) - 1))} \right] \\
+ \sum_{i=1}^{n} \log(p(w_{0i})) + \sum_{i=1}^{n} \log(p(D_i)) + \sum_{j=1}^{K} \log(p(\beta_j)) \\
+ \log(p(\mu_D)) + \log(p(\sigma_D)).
\]  

(4.45)

To make the computations easier, the summand terms can be first calculated as log-probabilities, then exponentiated and summed. The final form is

\[
\log(p(w, D, \beta | I)) \propto \sum_{i=1}^{n} \log \left\{ \log \{w_i + I_i \log(\lambda_i) - \log(I_i!(\exp(\lambda_i) - 1))\} \right. \\
+ \log(1 - w_i + I_i \log(\lambda_i + D_i) - \log(I_i!(\exp(\lambda_i + D_i) - 1))) \left. \right\] \\
+ \sum_{i=1}^{n} \log(p(w_{0i})) + \sum_{i=1}^{n} \log(p(D_i)) + \sum_{j=1}^{K} \log(p(\beta_j)) \\
+ \log(p(\mu_D)) + \log(p(\sigma_D)).
\]  

(4.46)

The priors for \( w_i, \beta_j, D_i, \mu_D \) and \( \sigma_D \) are now added to the log-probability derived above. Mixture weight \( w_{0i} \) has a beta prior, which was defined above in Equation 4.26. To make the computations easier, \( \log(D_i) \) are sampled instead of \( D_i \). If \( D_i \) is assumed to follow log-normal distribution, then \( \log(D_i) \) follows normal distribution.

\[
\log(D_i) \sim N(\mu_D, \sigma_D^2).
\]  

(4.47)

The connection between the parameters of these two distributions is described in section 4.2. The distribution of \( D_i \) should be set to meet the prior expectations and the same parameter values can be used for the prior of \( \log(D_i) \). The priors for hyperparameters \( \mu_D \) and \( \sigma_D \) are defined in Equations 4.15 and 4.16.

The coefficients of the Poisson regression model \( \beta_0, ..., \beta_4 \) have a normal prior. As the effects of the variables are not exactly known, they are given a zero-centered prior, \( \mu_b = 0 \), with a reasonable variance \( \sigma_b^2 \) to enable large enough deviations, for example \( \sigma_b = 3 \). So for each \( j = 0, ..., 4 \)

\[
\beta_j \sim N(\mu_b, \sigma_b^2).
\]  

(4.48)
The priors are then substituted in the log-probability in Equation 4.46 and the final form of the model log-likelihood is obtained

\[
\log(p(w, D, \beta|I)) \propto \sum_{i=1}^{n} \log \left[ \exp(\log(w_{0i}) + I_i \log(\lambda_i) - \log(I_i!(\exp(\lambda_i) - 1))) 
+ \exp(\log(1 - w_{0i}) + I_i \log(\lambda_i + D_i) 
- \log(I_i!(\exp(\lambda_i + D_i) - 1)))) \right] 
+ \sum_{i=1}^{n} \log(N(\log(D_i); \mu_D, \sigma_D^2)) 
+ \sum_{j=1}^{K} \log(N(\beta_j; \mu_b, \sigma_b^2)) + \sum_{i=1}^{n} \log(Beta(w_{0i}; a, b)) 
+ \log(N(\mu_D; \mu_{\mu_D}, \sigma_{\mu_D}^2)) + \log(Gamma(\sigma_D; a_{\sigma_D}, b_{\sigma_D}))
\]

(4.49)

### 4.5 Sampler implementation

The MCMC sampler was implemented in high-level computing language Matlab. Metropolis algorithm and Gibbs sampler introduced in Section 4.3 are used as building blocks of the sampler. Gibbs sampler is used for drawing new weight values \( w^t \) from its distribution, whereas Metropolis is used for updating \( \log(D), \beta, \mu_D \) and \( \sigma_D \). The needed distributions for Gibbs sampler and Metropolis algorithm were derived in previous section.

Before actual sampling, jumping distribution parameters and starting values for the sampled variables are set and the first posterior density function value is calculated. Sampler rotates the update turns of the variables. First, new value of \( w^t \) is drawn by using Gibbs sampling. After that, the posterior probability is updated. Then new proposal \( \log(D)^t \) is drawn, each element from their respective jumping distributions. The proposal is accepted or rejected based on the Metropolis acceptance rule described in Section 4.3. Regression coefficients \( \beta^* \) and hyperparameters \( \mu_D \) and \( \sigma_D \) are also updated in Metropolis steps of their own. The elements of each vector \( w, \log(D) \) and \( \beta \) are updated all at once, which is computationally more efficient than updating them element-wise. Gibbs sampler is then used again and so on. This loop is repeated \( N_2 \) times. After this the jumping rules can be adjusted and the simulation continues from where it was left in the end of the last round. Drawing \( N_2 \) samples is repeated \( N_1 \) times.

To make the Markov chains converge better, efficient jumping rules are
The chosen jumping distribution for $\log(D)$ is

$$J(\log(D)^t \mid \log(D)^{t-1}) = N(\log(D)^t; \log(D)^{t-1}, I_{c_{log(D)}^2\sigma_{log(D)^t}}), \quad (4.50)$$

which is centered around the current value $\log(D)^{t-1}$ with variance $c_{log(D)}^2\sigma_{log(D)^t}^2$, which is a vector consisting of variances for each element of vector $\log(D)^t$. To get an approximation of the variance of $\log(D)^t$, sample standard deviation is calculated from the samples and the jumping rule can be updated by replacing the variance with a newer estimate. The jumping distributions for $\beta$ and $\mu_D$ are also normal distributions with means being the current values with covariance $c_{\beta}^2\Sigma_{\beta}$ and variance $c_{\mu_D}^2\sigma_{\mu_D}^2$. To make computations easier, jumping distribution is defined for $\log(\sigma_D)$ instead of $\sigma_D$. This jumping distribution is also a normal distribution with mean $\log(\sigma_{D}^{t-1})$ and variance $c_{\sigma_D}^2\sigma_{\log(\sigma_D)}^2$. Jumping distribution variances for $\log(D), \beta, \mu_D$ and $\sigma_D$ are updated in this way after $N_2$ samples. It has been proven that in the most efficient jumping rule, the parameter $c$ is

$$c \approx 2.4\sqrt{N_{dim}}, \quad (4.51)$$

where $N_{dim}$ is the number of dimensions of the normal distribution (Gelman et al., 2014). When calculating $c_{\log(D)}$ the number of dimension $N_{dim}$ is $N$ and in the case of $c_{\beta}$ the dimension is $K$. As hyperparameters $\mu_D$ and $\sigma_D$ are updated in the same Metropolis step, the number of dimensions $N_{dim}$ is 2 in their case.

The outline of the sampler is the following. For each round $r = 1, ..., N_1$ the following cycle is repeated $N_2$ times. This results in sequences of $N_2$ samples of each parameter in each round. Number of samples $N_2$ is multiplied for the last round $r = N_1$ to get a longer sequence for parameter estimation. For each $t = 1, ..., N_2$

**Step 1:** New weights $w^t$ are drawn using Gibbs sampler.

- Draw $w^t$ from its conditional distribution $p(w^t|I^{t-1}, \lambda^{t-1}, \log(D)^{t-1})$.

**Step 2:** Update posterior probability.

**Step 3:** Update $\beta^t$ using Metropolis algorithm:

- Draw proposal $\beta^*$ for $\beta^t$ from their jumping distribution $J_t(\beta^*|\beta^{t-1})$.
- Calculate new $\lambda$ using the updated $\beta$.
- Either accept the proposal or reject it and set $\beta^t = \beta^{t-1}$. Set $\lambda^t$ accordingly.
Step 4: Update posterior probability.

Step 5: Update $\mu_D^t$ and $\sigma_D^t$ using Metropolis algorithm:

- Draw proposal $\mu_D^*$ from $J(\mu_D^* | \mu_D^{t-1})$.
- Draw proposal log($\sigma_D^*$) from $J(\log(\sigma_D^*) | \log(\sigma_D^{t-1})$) and exponentiate to get $\sigma_D^*$.
- Accept the proposals or reject them and set $\mu_D^t = \mu_D^{t-1}$ and $\sigma_D^t = \sigma_D^{t-1}$.

Step 6: Update log($D^t$) using Metropolis algorithm:

- Draw proposal for log($D^t$) from $J(\log(D^*) | \log(D^{t-1})$).
- Accept or reject the proposal. If rejected, set log($D^t$) = log($D^{t-1}$).

Step 7: Values log($D^t$), $\beta^t$, $\mu_D^t$ and $\sigma_D^t$ are stored into matrices $M_r^t$ log($D$), $M_r^t$ $\beta$, and vectors $M_r^t$ $\mu_D$ and $M_r^t$ log($\sigma_D$) for covariance calculations. The posterior probability values are also stored.

After repeating the cycle $N_2$ times, the covariance matrices and covariances of the jumping distributions for $\beta$, log($D$), $\mu_D$ and $\sigma_D$ are updated. For log($D$) the update for each element of vector $\sigma_{\log(D)}$ is

$$\sigma_{\log(D)i}^{r+1} = \text{Std}(M_{\log(D)i}^r)$$

for each $i = 1, ..., n$ where $M_{\log(D)i}^r$ is a column vector of $M_{\log(D)}^r$ and consists of samples for log($D$)$_i$. For $\beta$ the updated covariance matrix is

$$\Sigma_{\beta}^{r+1} = \text{Cov}(M_{\beta}^r).$$

For $\mu_D$ and $\sigma_D$ the updates are

$$\sigma_{\mu_D}^{r+1} = \text{Std}(M_{\mu_D}^r)$$

and

$$\sigma_{\log(\sigma_D)}^{r+1} = \text{Std}(M_{\log(\sigma_D)}^r).$$

To make sure the covariances will never be zeros, a small positive quantity such as $1 \cdot 10^{-4}$ is added to these values.

After each round $r$ every 150th posterior probability and desired variable values are saved into a file. After the sampler has converged, the thinned sequences of all the variables are saved in the same manner. The sequences can be quite autocorrelated and so there is no need to save the whole sequence.
but a thinned one. Naturally the advance of thinning is that it saves space. Numbers of rejected samples in each Metropolis step can be saved to evaluate the performance of the sampler.

If the expected behaviour of the model is known from running multiple chains from the same model, the easiest way to check for convergence of a sequence is to plot the variable sequences and see whether there is still some trend in their behaviour. If the sequence seems to oscillate around a fixed value, the number of samples can be assumed to be large enough and convergence has been reached.

When the Markov chain has converged to its stationary distribution, the resulting samples are used for calculating the desired quantities. These quantities describe the distribution of the model parameters to be estimated. Before taking the mean, the burn-in samples are removed. In practice, the $N_2$ samples from the last sampling round $r = N_1$ are used for the calculations and the rounds $r = 1, ..., N_1 - 1$ are thought of as burn-in. To get a better picture of the posterior distribution, a histogram of the samples can also be plotted. Other visualizations are used as well for presenting the results and how well the model performs.
Chapter 5

Results

In this chapter the model proposed in this thesis is evaluated. First, the data set used in the development of the model is introduced. After this, the special measures needed for handling inter-chromosomal interactions are explained.

The focus is then given to the actual results from the model. First, preliminary results from stand-alone Poisson regression are presented to proof the relevance of making it a part of the mixture model. Then the results from the whole model are given and they are evaluated. Finally, the results from proposed mixture model is compared to results given by ChIA-PET Tool and Mango, which were introduced in Chapter 3.

5.1 The Mcf7 cell line data set

The data set used in the development of the Bayesian mixture model with Poisson regression was a ChIA-PET data set of Mcf7 breast cancer cell line. The data is freely available in UCSC ENCODE downloads. The data set was first introduced by Li et al. (2012). In (Li et al., 2012), the function of higher-order chromosomal organization was investigated by detecting chromatin interactions associated with RNA polymerase 2 enzyme in different cell lines. The exact used data set was replicate 1, and the ChIA-PET data had already been linker-filtered and aligned to human genome 19.

The data set included 35444479 PETs of which 5035994 were self-ligations and the rest 30408485 were inter-ligations. Peak calling resulted in 9956 anchor sites. Of all anchor site lengths around 98,5% are 2001 base pairs and the remaining 1,5% are between 2607 to 3994 base pairs long, with one exception of 5137 base pairs long anchor. After each inter-ligation start and end site were mapped to these anchors, there were 101801 putative interactions of which 8450 were intra-chromosomal and 93261 inter-chromosomal. Interaction counts for
these interactions ranged from 1 to 47, but most of interaction counts were ones.

5.2 Modeling inter-chromosomal interactions

The model presented in Section 4.2 includes the distances between the two interacting anchors. This distance is not defined for the inter-chromosomal interactions: there is no way to calculate how many base pairs there are between two locations in different chromosomes. As the numerous inter-chromosomal interactions should still be included in the analysis, some kind of substitute value has to be determined for the distance quantity.

As there is no common practise how to determine such quantity, a natural choice was to use maximum likelihood approach. A truncated Poisson regression model where distance was used to predict intra-chromosomal interaction counts was first estimated. The Poisson rate parameters were estimated from this model for a grid of distance values ranging from $10^4$ to $10^{10}$. Then likelihoods for the set of inter-chromosomal interactions counts were calculated using truncated Poisson distribution with the different rate parameters. The distance corresponding to the rate parameter which gave the highest likelihood was then chosen as the substitute distance.

The log-likelihood of the data with different distances are plotted in Figure 5.1. The log-distance value which gave the greatest log-likelihood was approximately 17.2, which is around $3 \cdot 10^7$ base pairs when exponentiated. This distance best describes the ratio of interaction count number according to the Poisson regression model. The range of intra-chromosomal interaction log-distances is from $8.4$ to $19.3$, which correspond to $4.4 \cdot 10^8$ and $2.4 \cdot 10^8$ base pairs in linear scale. Compared to these values the substitute inter-chromosomal distance is quite reasonable. Most of inter-chromosomal interactions are assumed to be random ligations and the interaction counts are quite low as they also are for intra-chromosomal interactions with long distance between the interacting sites.
Figure 5.1: Likelihood of the count data with different distance values. The peak of the curve, i.e. the maximum likelihood distance, is located around $\log(distance) \approx 17.2$.

5.3 Preliminary results from Poisson regression

In this section some preliminary results from Poisson regression are presented. These results show how well Poisson regression works for this data and whether the explanatory variables are significant.

The variables to be used in the Poisson regression are distance between interacting anchors, mappabilities, GC-contents and anchor coverages. To see how the three latter properties affect the marginal counts, plots in Figure 5.2 were produced and correlation coefficients were calculated. This figure shows that correlations between log-marginal counts and either mappability or GC-content were quite weak. Correlation between marginal counts and coverages was very strong instead. This can also be seen in Figure 5.2, where the exact correlation coefficients are also presented.

In Figure 5.3 the correlations between the interaction counts and interacting anchor pair properties are shown as scatter plots and correlation coefficients. The explanatory variables are now in the form given in Chapter 4 which means that logarithm is taken of the product of the two anchors properties.
Figure 5.2: Correlations between logarithms of anchor marginal counts and anchor properties. Pearson correlation coefficients $\rho$ are given in each figure.

Log-distance has the greatest correlation coefficient $-0.136$.

Figures 5.2 and 5.3 raise doubts of whether all the explanatory variables are significant or not. Distance between interacting anchors and anchor coverage seem to be reasonable regressors as their correlation coefficients to interaction counts are the highest. The distance variable scatter plot already clearly manifests that the shorter the distance the higher the interaction count is. Considering these results, it might be justified to leave the explanatory variables mappability and GC-content out of the regression model. The Figures 5.2 and 5.3 include all anchors and observed interactions.

To test how well the Poisson regression model can explain the variance of the interaction counts, the Poisson regression model described in Equation 4.5 is estimated using the whole data set. The maximum likelihood estimates for the regression coefficients are $\beta_0 = -0.111$, $\beta_1 = -0.287$, $\beta_2 = -0.2501$, 

\[ \hat{\beta}_0 = -0.111, \hat{\beta}_1 = -0.287, \hat{\beta}_2 = -0.2501, \]
\[
\beta_3 = -0.362 \quad \text{and} \quad \beta_4 = 0.224.
\]

The fitted interaction counts using this model are shown in Figure 5.4 and the residuals against the observed interaction counts in Figure 5.5.

The Poisson regression model fits seem to be systematically lower than the observed values. As around 95\% of the interaction counts in this data set are ones, it is quite reasonable that the model cannot explain the higher interaction counts very well. As the Poisson regression part of the mixture model is used for predicting the interaction frequency based on just the bias factors, it actually should account mostly for the noise part of the data. From these results it can be seen that Poisson regression model explains quite well the low interaction counts, which are expected to be noise for the most part.

It can be questioned whether the high counts should be included at all in
Figure 5.4: Observed interaction counts versus the values fitted using the maximum likelihood estimates of the Poisson regression model.

The estimation of the model parameters. If the model should account only for random noise, it should be estimated using only the assumed random interactions. Including the more infrequent high interaction counts, which are of a different magnitude compared to the noise counts, can lead to overfitting the parameters to these rare values. Even though there are only a few large values in the data set, they have an influence on the estimation of the parameters.

The bad fits and residuals which grow along with growing interaction counts are normally a sign that a regression model is missing something, perhaps another regressor or different functional form of the included regressors. But this time the model is not even expected to predict all interaction counts well. This is also why the explanatory variables mappability and GC-content are left in the model, even though they do not have high correlations with the counts. There is indeed something missing from the model, which is the "true interaction" component. This component will be added in the mixture model to the expected values given by the Poisson regression model here. These components $D_i$ will account for the additional interaction counts that are expected to be seen in case of true interactions.
CHAPTER 5. RESULTS

Figure 5.5: The residuals of the estimated Poisson regression model plotted against the observed interaction counts. This figure shows clearly how the pure regression model cannot predict high interaction counts very well.

5.4 Full model

The full Bayesian mixture model with Poisson regression was tested on the Mcf7 cell line data set introduced in Section 5.1. Only the 8450 intra-chromosomal putative interactions were analyzed to reduce the size of the data set. The sampler was started from initial values $\sigma_D = 0.9$, $\mu_D = 2$, $\beta_0 = 2$, $\beta_1 = -0.6165$, $\beta_2 = 0.09$, $\beta_3 = -0.2381$ and $\beta_4 = 0.5792$. The Poisson rate parameters were then calculated from the initial values of $\beta$ as described in Equation 4.8. The initial values for mixture weights vector $\log(w)$ was drawn from Beta distribution with parameters 0.7 and 0.3 and initialization for $\log(D)$ was drawn from normal distribution with mean $\mu_D$ and variance $\sigma_D^2$. The convergence of the sequences was monitored by checking the progress of log-posterior probabilities. The number of rounds $N_1$ was first set to 400. After these rounds the sequence had not still converged, so the number of rounds was increased to $N_1 = 600$. The number of samples $N_2$ was 15000 in all rounds except $r = 400$ and $r = 600$, where the length of the sequence was $5 \cdot 15000 = 75000$. The 75000 samples from last round $r = 600$ were then used for parameter estimation.

It took 5 days and 16 hours to complete the 600 rounds of sampling to
estimate the model parameters. This is a long time for completing just one step of the analysis, so there is still much room for improvement in the efficiency. Initializing the starting points closer to the true values could reduce the needed number of samples. Also, Matlab is not a very efficient programming language, and implementing the model in some other more efficient language could resolve this problem.

The log-posterior probability of the model is presented in Figure 5.6. The log-probability rises little by little as the number of samples goes up until it stabilizes in the very last sampling rounds. In the last 75000 sampling cycles the mean log-probability was $4.3 \cdot 10^4$.

The hyperparameter values were narrowed down quite well. The mean of the samples is 0.809 for $\mu_D$ and $2.467 \cdot 10^{-3}$ for $\sigma_D$. The small ranges of values of $\mu_D$ and $\sigma_D$ can be clearly seen from the histograms of their samples presented in Figure 5.7 and Figure 5.8. As the variance of $D_i$s is so small, it means that there should be only one $D$, which would be the same for all interactions. This is something which should be considered in further development. Setting the same $D$ to all interactions would mean significant simplification in the model and reduction of computation time, as there would be $n - 1$ less parameters to estimate in the model.

Of the 8540 putative intra-chromosomal interactions 852 has mean of weight $w_{bi}$ samples under 0.5 and 7688 has mean of the samples 0.5 or higher. The 852
Figure 5.7: Histogram of the thinned sample of $\mu_D$.

Figure 5.8: Histogram of the thinned sample of $\sigma_D$. 
are thus assumed to come from the second mixture component representing the true interactions and the rest 7688 interactions are from the false interactions component. The mean interaction count for the true interactions is 2.61 and 1.02 for the false interactions. This shows that the mixture model supports the assumption that the signal is stronger i.e. interaction frequency is higher for the true interactions and weaker for the false ones.

In Figure 5.9 the means of \( w_{0i} \) samples for each interaction \( i \) are plotted as function of mean of Poisson rate parameter \( \lambda_{0i} \) samples. Each interaction count is represented with a color of its own. The yellow dots, representing interaction count 1, form a cluster where all interactions have mean of mixture weight \( w_{0i} \) close to 0.8 and the mean Poisson rate parameter values \( \lambda_{0i} \) range from \( 10^{-4} \) up to \( 10^0 \). Interaction counts 2 and 3 have both interactions with mixture weight under and over 0.5, depending on the Poisson rate parameter. The ones with high rate parameter \( \lambda_{0i} \) values have higher mixture weight, which shows that the bias factors explain these interactions so well they cannot fit to the true interaction component. Most of the higher counts have mixture weight between 0.3 and 0.4. The effect of mixture weight’s beta prior shows in the distribution of \( w_{0i} \)'s, as the most of mixture weights are 0.8 and the most of the rest are between 0.3 – 0.4.

There are some exceptions too, where an interaction with high count has a high mixture weight mean. These interactions have high Poisson rate parameters. Most interesting of them is the rarely high interaction count of 47, which has rate parameter \( \lambda_{0i} \) mean of 37. This exceptional case raises the question of whether the Poisson regression model has been overfitted or not, yet for the most of the data points the regression works as expected. However, when inspecting the properties of this particular interaction with high interaction count, the coverage property is noticed to be very high. The mean of the coverage attribute is 10.4 for all intra-chromosomal interactions and for the interaction in question the coverage is 17.7. The estimate for regression coefficient \( \beta_4 \) is 0.83, which means that the expected number of interactions goes up as the coverage gets higher. The model thus suggests that the interaction count if 47 is not enough for an interaction with such a high coverage for it to be classified as a true interaction.

5.5 Comparison with other methods

The comparison of the proposed mixture model, Mango and hypergeometric model of ChIA-PET Tool was done with the same data set which was introduced in Section 5.1. Each method was used for finding interactions from this data, Mango and mixture model give only the detected intra-chromosomal
interactions, while ChIA-PET Tool’s results include both intra- and inter-chromosomal interactions. The results from the Bayesian mixture model with Poisson regression were introduced earlier in Section 5.4. Mango results were produced by using raw ChIA-PET experiment data downloaded from UCSC ENCODE website. The ready-made ChIA-PET Tool results for the experiment available in NCBI GEO data repository were used in this comparison.

The numbers of found interactions vary greatly between the compared three models. Table 5.1 presents the number of found interactions and gives the numbers of both intra- and inter-chromosomal interactions for ChIA-PET Tool. The number of interactions found by ChIA-PET Tool, 5893, is almost thirty times the number of interactions found by Mango, which is 223. The mixture model proposed in this thesis classifies 852 intra-chromosomal interactions to be true interactions, which is again considerably larger amount than in Mango’s results but much smaller than for ChIA-PET Tool. Phanstiel et al. (2015) suggest that inter-chromosomal interactions are only noise and are not worth analysing, which is why inter-chromosomal interaction analysis was not implemented in Mango. ChIA-PET Tool by contrast proposes that there are 6736 true inter-chromosomal interactions in the data set and does not ques-
tion the significance of inter-chromosomal interactions. The Bayesian mixture model with Poisson regression can be used for inter-chromosomal interaction analysis, as substitute value for distance between anchor sites was derived, which was described in Section 5.1.

The found intra-chromosomal interactions were compared to find out if the methods find the same interactions or not. For each found interaction it was checked whether the two other methods had found an interaction from the same chromosome with the same head and tail anchor location. Anchors were thought to be the same if they overlap or are up to 1000 base pairs from each other. The Table 5.5 shows the number of overlapping interactions between each of the three methods. ChIA-PET Tool reports interactions where two different interactions have head and tail anchors very close or even overlapping with each other. Overlapping interactions and anchors could have been merged to make the results less ambiguous. This unclarity makes further comparison between the results from different methods difficult and the principles used in ChIA-PET Tool seem to be questionable.

The same kind of overlap analysis as presented above was also performed for the interactions the mixture model classified as false interactions and the interactions found by Mango and ChIA-PET Tool. Of the 7688 interactions the new mixture model classified to be false interactions, 9 were classified to be true interactions by Mango and 439 by ChIA-PET Tool. The mean interaction count for the ones overlapped with Mango was 6.22 and with ChIA-PET Tool the mean was 1.14. The mean of interaction count is so high for Mango because it accepted the interaction with interaction count 47. This interaction was only just rejected by the mixture model and is a possible false negative. For the rest of the 9 interactions one interaction count is 2 and seven are 1. ChIA-PET Tool did not find the interaction with interaction count 47.

To compare the types of interactions the three models found, annotation analysis was done for the found interactions. An annotation was determined for each head and tail anchor using Homer software (Heinz et al., 2010). Then the numbers of interactions where both of the anchors are promoters, one of the anchors is a promoter and neither of them is a promoter were calculated.

Table 5.1: Number of found interactions for each method. Only ChIA-PET Tool was used for finding inter-chromosomal interactions.

<table>
<thead>
<tr>
<th>Method</th>
<th>Total number of interactions</th>
<th>Intra-chromosomal</th>
<th>Inter-chromosomal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mixture model</td>
<td>852</td>
<td>852</td>
<td>-</td>
</tr>
<tr>
<td>ChIA-PET Tool</td>
<td>12626</td>
<td>5893</td>
<td>6736</td>
</tr>
<tr>
<td>Mango</td>
<td>223</td>
<td>223</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 5.2: Number of overlapping intra-chromosomal interactions found by different methods. The total number of found intra-chromosomal interactions for each method is presented in the diagonal.

<table>
<thead>
<tr>
<th>Method</th>
<th>ChIA-PET Tool</th>
<th>Mango</th>
<th>Mixture model</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChIA-PET Tool</td>
<td>5890</td>
<td>216</td>
<td>349</td>
</tr>
<tr>
<td>Mango</td>
<td>216</td>
<td>223</td>
<td>70</td>
</tr>
<tr>
<td>Mixture model</td>
<td>349</td>
<td>70</td>
<td>852</td>
</tr>
</tbody>
</table>

Table 5.3: Proportions of different kinds of interactions found by each method.

<table>
<thead>
<tr>
<th>Method</th>
<th>Promoter-promoter</th>
<th>Promoter-other</th>
<th>Other-other</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChIA-PET Tool</td>
<td>4%</td>
<td>20%</td>
<td>76%</td>
</tr>
<tr>
<td>Mango</td>
<td>7%</td>
<td>16%</td>
<td>77%</td>
</tr>
<tr>
<td>Mixture model</td>
<td>35%</td>
<td>44%</td>
<td>21%</td>
</tr>
</tbody>
</table>

The proportions of each kind of interactions are presented in Table 5.5. For both ChIA-PET Tool and Mango the proportion of interactions where neither of the anchors is a promoter is as high as over 75%. The proportions of promoter-promoter and promoter-other interactions were also very similar between these two methods. The new mixture model again has found interactions of which 44% are promoter-other interactions, which is significantly higher than in case of Mango and ChIA-PET Tool which have proportions of 16% and 20%. The second largest group is promoter-promoter interactions with 35% and the smallest is other-other interactions with proportion of 21%. The proportions for the mixture model are in line with previous research, as Li et al. (2012) found a large number of promoter-promoter interactions in their study. They provide experimental evidence of many of these promoters cooperatively regulating the promoters they interact with. Their findings indicate that this mechanism is common in cells. By contrast, the interactions which do not include promoter regions are more difficult to interpret.

Overall it seems that the results from the mixture model are not exactly the same as the ones the two other model give, but not completely different either. The number of found interactions was huge for ChIA-PET Tool and very small for Mango. It seems that Mango might find some interactions which were missed by the mixture model, while ChIA-PET Tool easily accepts interactions despite their low interaction counts. Mixture model results show that the interactions it accepts have mean interaction count significantly higher than the ones it rejects. It also can reject interactions with very high interaction counts, if the bias factors explain the count well. The biological validation of the interactions by using annotation results shows, that the mixture model can
capture biologically meaningful interactions.

The three approaches for chromatin interaction analysis compared here have different preprocessing principles, which might affect the results of the comparison greatly. Especially the methods for finding and defining anchors have an influence on comparison results. The results from annotation analysis also depend on the anchors which were used by each method, so the annotation analysis results actually show how well the whole data analysis process from data preprocessing to actual interaction analysis performs. To purely compare the statistical methods used for the quantification of the interactions would require that the lists of putative interactions given to each analysis tool would be the same.
Chapter 6

Conclusions

In this thesis the current methods for ChIA-PET data analysis were reviewed and a new one was proposed. The new model consisting of Bayesian mixture model and Poisson regression for bias reduction was described in detail along with the implementation and Bayesian analysis and Markov chain Monte Carlo methods, which were used for parameter estimation. The implemented model is then tested on a real ChIA-PET data set and compared to two different analysis tools.

The results from the proposed method show that the model produces reasonable results which base on the interaction counts and bias factors. In addition, the results were compared to results from Mango and ChIA-PET Tool. Comparison showed that some of the found interactions were same for the mixture model and the other tools, but there were also interactions which were only found by the mixture model. To further evaluate the results, an annotation analysis was done for the found interactions. The most of the interactions the mixture model classified as true interactions included promoter regions. This result is supported by the research by Li et al. (2012). Externally validated interactions would be needed to further evaluate the performance of the model, but unfortunately none were available at the time this thesis was done.

Based on the results from the model it seems that the Poisson regression part of the model does not work optimally. The problem lies in the explanatory variables and the nature of the data. As most of the observed interaction counts are ones when setting the truncation to \( k = 1 \), there is not much variation to be explained by the regression model. On the other hand, the point of the Poisson regression model was not to completely explain the data, just predict the random noise part of the interaction counts. The regression works quite well for this purpose. From the results of the full model presented in Section 5.4 it could be seen that an interaction with interaction count as large as 47 was classified as a false interaction, because the estimated Poisson rate parameter
was so high. This invokes the question, whether the model is overfitted when estimating the parameters. But on the other hand, this result makes sense if the properties of this particular interaction are considered and the regression model seems to work as expected for most of the interactions.

One way to truly make the regression model represent the noise component in the data is to extract a subset of data from the analysis and estimate the regression coefficients separately based on this subset. This subset could then be selected to represent the data points that are most likely to be noise. However, this approach has its own difficulties such as how to select the subset to represent false interactions.

The results also suggest that there is no need for interaction specific $D_i$s as one $D$ for all interactions would be enough. Different forms of Poisson regression link function and other type of $D$ are also worth trying if further research is done. For example, $D$ could be multiplicative instead of additive.

The effect of the decisions made in the data processing step is quite significant. There are many ways of doing the preprocessing steps such as anchor site finding and in some cases different choices could lead to approximately same results. Preprocessing of the data was also a quite laborious phase of this work. As the number of steps is high in this process, there are many possibilities to make errors which can in the worst case accumulate during the analysis.

One of the hard decisions was how to model the inter-chromosomal interactions with the same model as the intra-chromosomal ones, as the distance between the interacting anchors is not defined for the inter-chromosomal ones. There is no correct answer to how to reason out a substitute distance value for them. In this thesis the problem was solved by using maximum likelihood estimation. Some of the other models using the distance between anchors in the analysis go around this problem just by discarding all the inter-chromosomal interactions. There has been no other method to model inter-chromosomal interactions with a model that takes the distance between interacting anchors into account. Leaving all the inter-chromosomal interactions out of the analysis is a quite a conservative choice, which leads to missing some possible interactions with biological relevance. This has been the common practise in non-central hypergeometric model and Mango, which were presented in Sections 3.1.1 and 3.1.5.

The downside of the method presented in this thesis is the intensiveness of the computations. As the number of observations in the data set grows larger, the number of parameters to be sampled increases and computation time goes up. Making the computations more efficient play an important role in making the model a useful tool for analysing ChIA-PET data. For example, changing
the parameterization of the model might help the MCMC methods to converge faster. At first implementing the model in probabilistic programming language Stan was tested, but the MCMC method no-U-turn sampler Stan utilizes failed to reach convergence. This phenomenon was quite surprising, as no-U-turn sampler should be an advanced MCMC method targeted for complex models, but this particular problem is apparently too complex for it. Stan would have provided easy ways of convergence analysis, which only have a small role in the Matlab implementation.

One flaw in the model development stage was that only one data set was used. The used data set was assumed to be representative of typical ChIA-PET data and most likely this assumption holds. Using other data sets as well would still most likely have been worthwhile. Of course the estimation of the model parameters will be done for the data set in question, but the structures of the model were determined from this one data set alone.

All in all, the Bayesian mixture model which incorporates Poisson regression for bias correcting seems to work reasonably, regardless of the computational challenges. The model can be used even for analysing inter-chromosomal interactions, which is not possible with all available tools. If the tool was to be developed further, computational efficiency and different model structures should be considered.
Bibliography


Guoliang Li, Melissa J. Fullwood, Han Xu, Fabianus Hendriyan Mulawadi, Stoyan Velkov, Vinsensius Vega, Pramila Nuwantha Ariyaratne, Yusoff Bin


Eitan Yaffe and Amos Tanay. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nature Genetics, 43(11), 2011. doi: 10.1038/ng.947.