A Survey of Statistical Models to Infer Consensus 3D Chromosomal Structure from Hi-C data

MEDHA UPPALA, University of California Los Angeles

The spatial organization of the genomic material leads to interactions between far placed loci, affecting the functions of the chromosomes. To understand this effect of the 3D structure of chromosomes, we need to infer the structure from collected data. Recent, high throughput data collection methods like 5C and Hi-C offer genomic level mapping of the 3-dimensional structure. This paper surveys two, currently existing statistical models, that apply to Hi-C data to infer the consensus 3D chromosomal structure. The first model called MCMC5C can be applied to both 5C and Hi-C data, while the second model called BACH can only be applied to Hi-C data. We learn that while both models have certain drawbacks, given the scope of Hi-C data, they open the door to a new field of genomics that requires development of exclusive statistical techniques.

Categories and Subject Descriptors: H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Evaluation / methodology*; H.1.2 [**Models and Principles**]: Statistical models—*Bayesian Models, MCMC*

General Terms: Hi-C, 3D chromatin structure, Model-based, MCMC

ACM Reference Format:

Medha Uppala. A Survey of Statistical Models to Infer Consensus 3D Chromosomal Structure from Hi-C data. 2014 *ACM Trans. Appl. Percept.* 0, 0, Article 0 (June 2014), 7 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

It has been shown that a single diploid cell contains almost 2 meters of DNA. The chromatin is a tightly packed, 3-dimensional, knot-free fractal globule of DNA. This 3-dimensional structure is known to affect the functions of the chromosomes. The spatial organization of the genomic material, or even a single chromosome, affects functions such as gene regulation, DNA replication, epigenetic modification etc. Many diseases like cancer have also been characterized by alterations in spatial organization (Hu *et al.* [2]). This not only motivates the study of chromatin 3D structure, but also exposes an uncharted field of computational biology, providing developmental opportunity for statistical and computational methods. Chromosome capture conformation technologies that have grown in popularity and capacity in the last decade provide different kinds of high throughput data such as 3C, 4C, 5C, Hi-C etc. With the increase in computational power, these data-collection methods provide a 2-dimensional, numerical rendering of the 3D structure that can be analyzed with statistical or computational models. This paper aims to review the only two existing Bayesian models that aim to infer the consensus 3D chromosomal structure from Hi-C data. There exist other optimization based models that aim to do the same, but will not be covered in this paper.

As we study these models that infer the 3D chromatin structure, it is important to acknowledge the highly dynamic nature of genetic material; the chromatin structure is not constant even within cells of a same kind. This is why the two models we cover in this paper aim to infer a 'consensus' 3D structure of the chromosomes or genome. However, "whether chromatin exhibits a consensus local 3D chromosomal structure is still under debate" (Hu *et al.*[1]). This is a cautionary reminder that inferring a consensus structure is ineffectual if there is no knowledge about the structural variance. This hurdle along with other drawbacks and achievements will be addressed later in the paper.

The paper is structured as follows: Section 2 provides a general overview of Hi-C data output required to understand the model applications to this data. As the goal of this paper is to review the statistical models at use, we will not review the systemic method of the Hi-C data collection process. Section 3 reviews the framework of the two Bayesian models formulated to infer the consensus 3D structure. Section 4 will provide an overview of the model results and evaluate them with respect to model features. Section 5 will conclude the paper with a discussion of future research avenues for chromosome 3D structure inference given the state of models discussed.

2. HI-C DATA

Hi-C data method is one of the latest chromosome capture conformation methods that allows for a genome-wide mapping of chromatin interactions. Tho whole genome is divided at regular intervals into separate loci or fragments of certain length. The paired-end reads between any two genomic loci are captured during the collection method; the contact frequencies are inversely proportional to the spatial distances between the genomic loci. The preprocessed reads are summarized in a Hi-C contact map, where the off diagonal entries represent the paired-end read counts between all pairwise combinations of genomic loci. However, an important aspect of Hi-C data is that it is conducted on a population of cells rather than a single cell. As a result, the contact matrix entries should be interpreted as population average reads between two loci, rather than a count from a single cell (Fig 1). The human genome is known to have around 10^12 genomic loci, with 6bp restriction enzyme. However, a typical Hi-C data collection process can only capture reads for about 10^8 of the loci (Hu *et al.* [1]). Imposing this complete loci matrix on the smaller matrix results in a highly sparse dataset; in remedy, some of the loci are grouped together into bins, allowing for dimension reduction and resulting in a lower resolution dataset. Moreover, it has been shown that the systemic collection process of Hi-C introduces biases in three ways: restriction enzyme cutting, GC content and sequence uniqueness (Hu *et al.* [2]). This systemic bias combined with the sparseness of the contact matrix produces a noisy and possibly biased dataset. As a result, it is crucial to take into account the shortcomings of the dataset when modeling to infer from it.

3. STATISTICAL MODELS

Given the high throughput data of Hi-C and the knowledge that spatial organization affects chromosomal functioning, the next step is to actively discern the 3D chromosomal structure through modeling and computational methods. As mentioned before, there are models that infer the consensus 3D chromosomal structure by solving it as a constrained optimization problem. However, we will only discuss probabilistic models specified with distributional assumptions and performed using Markov-chain Monte Carlo methods. Figure 2 illustrates the full circle of collecting the data from chromosomes and inferring back the 3D structure from the contact matrix.

3.1 MCMC5C

The MCMC5C method introduced by Rousseau *et al.* in 2011, was formulated to be used on both 5c and Hi-C data, with a few procedural variations. Each chromosome(or a region of a chromosome) is modeled as a string of uniform length fragments or loci, each represented as a spherical dot as seen in Figure 2. As a result, the chromosome is a string of pearls assuming a structure in a 3-dimensional space. The location of a fragment i is represented by the cartesian coordinates $S_i = (x_i, y_i, z_i)$; the model aims to infer $S = S_i, \ldots, S_n$ given the Hi-C contact matrix $\mathbf{M}$.

The main assumption in this model is that the paired-end read counts are inversely proportional to the spatial, euclidean distance between two loci. A further assumption made is that these read counts are gaussian distributed. The model specification is as follows: the probability of observing a certain

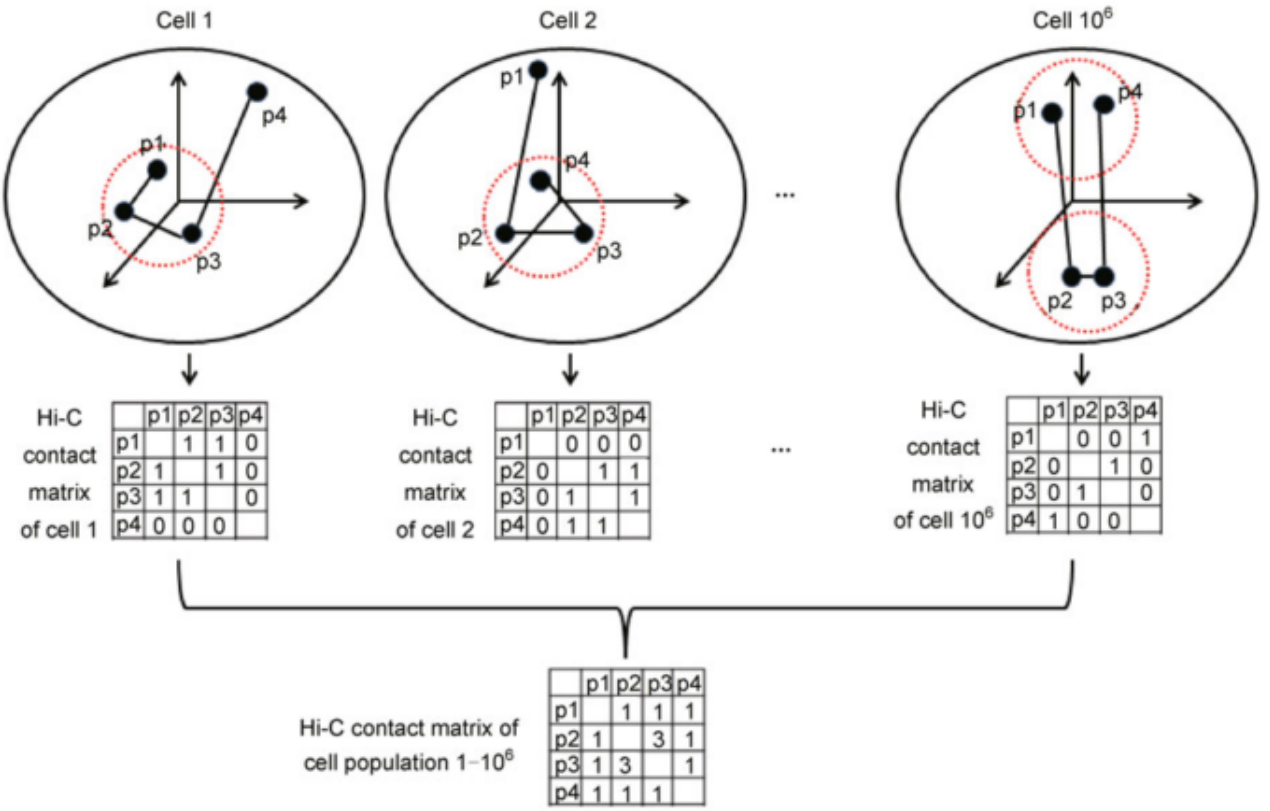

Fig. 1. Illustration of Hi-C data aggregation [2]

read count ($r(i,j)$) between a pair of loci (i,j), is binomial with probability

$$\hat{p}(i,j) = \frac{r(i,j)}{\sum_{a,b} r(i,j)}$$

as suggested by Duan *et al*. This probability of observing a read count for a specific loci pair is then approximated to be a gaussian distribution for computational convenience. This results in

$$P[(i,j)|\mathbf{S}] = N(\hat{p}(i,j), \hat{p}(i,j) + \kappa)$$

Here, the mean is approximated as the binomial probability and the variance is approximated as the binomial probability plus a small constant κ. While the variance estimate seems ineffectual, the small constant is to make sure that pairs with low read counts will not have too low a variance. The choice of variance in this model will be discussed in more detail in Section 4. Given this normal approximation to the probabilities of pairwise read counts, the posterior distribution of $\mathbf{S}$ given the can be approximated as follows:

$$P(\mathbf{S}|\mathbf{M}) = \frac{\prod_{ij} P[(i,j)|\mathbf{S}] \cdot P[\mathbf{S}]}{P[\mathbf{M}]}$$

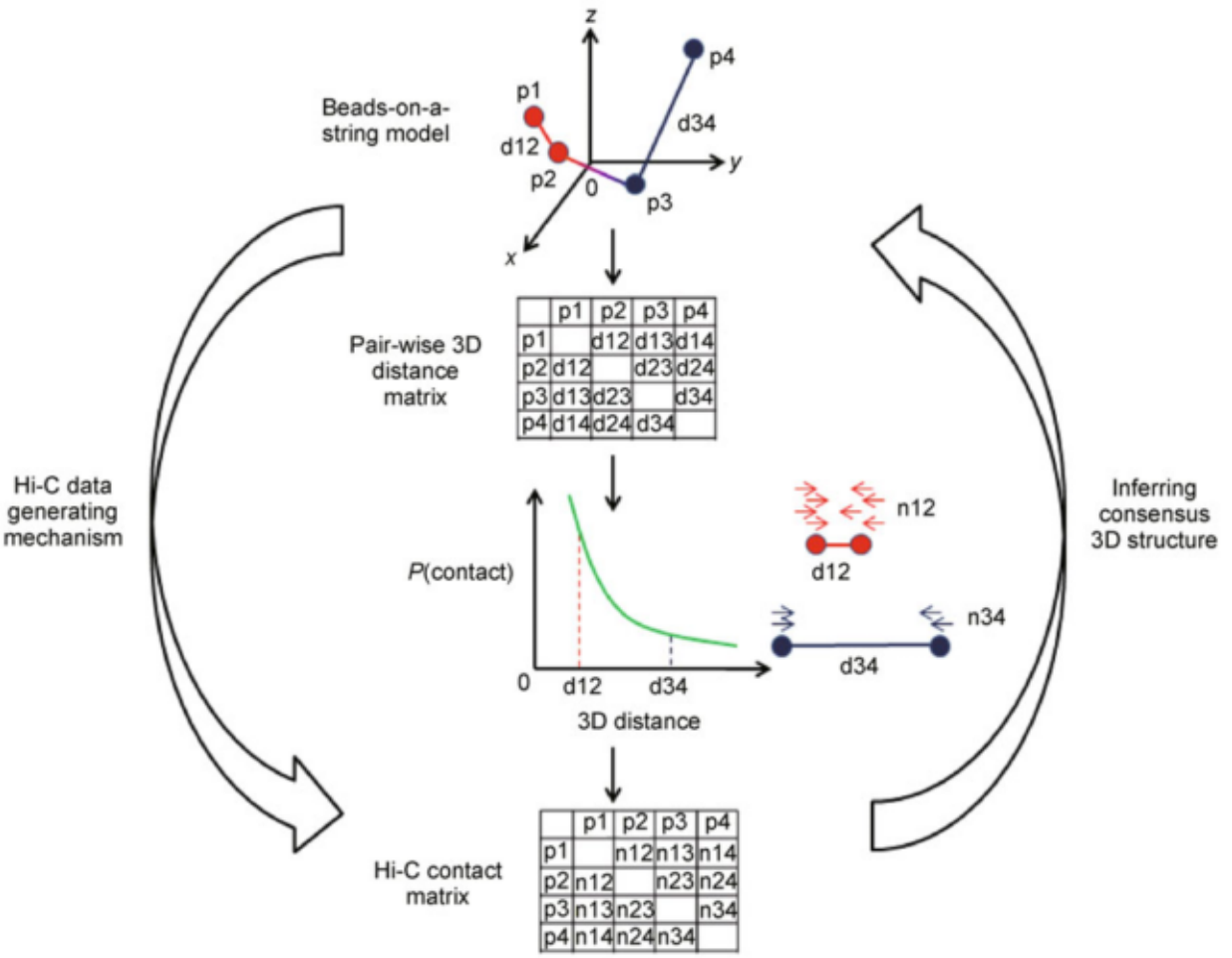

Fig. 2. Data collection and Modeling [2]

As there are no constraints on **S**, the probability of the observed data **M** is constant with respect to **S** (Hu *et al.*)

$$P(\mathbf{S}|\mathbf{M}) = \zeta \cdot \prod_{ij} P[(i,j)|\mathbf{S}]$$

which defines the posterior probability distribution of the space of structures in which **S** exists. As this posterior distribution cannot be easily evaluated, Markov Chain Monte Carlo methods are used to sample from this distribution. The authors of MCMC5C, as the name indicates, use the Metropolis-Hastings algorithm to sample unique structures from the posterior distribution. Once the algorithm is implemented–the authors sample 250 structures–the sample of structures are clustered into groups using hierarchical clustering. These clusters are then further analyzed based on characteristics such as looping density, compactness etc.

3.2 Bayesian 3D Constructor for Hi-C data(BACH)

The other model-based approach used to analyze Hi-C data was introduced by Hu *et al.* [1] in 2013 is called the Bayesian 3D Constructor for Hi-C data (BACH). The assumptions of this model are: the read counts between a pair of loci are inversely proportional to the spatial distance between the two loci and the read counts follow a poisson distribution. Moreover, the authors specifically state the assumption of an existing consensus 3D structure, which according to them is still up for debate. The chromosome is once again modeled as a piecewise linear curve in a 3D space, specified by a spherical dot for each fragment or locus. For consistency sake, we will use the same notation of $S_i = (x_i, y_i, z_i)$ for a single fragment position and $S = S_i, \ldots, S_n$ representing the structural positions of the n fragments.

Given the Hi-C contact matrix $\mathbf{M}$, the read counts between a loci pair are assumed to follow a poisson distribution:

$$\hat{p}(i,j) = Poisson(\theta_{ij})$$

where the parameter θ_{ij} is estimated using a log-linear model. This log-linear model is an attempt to reduce the bias introduced to the data through systemic factors like enzyme restriction, the GC content etc. The authors, while not rationalizing the choice of a log-linear model, endorse it as a fair bias reduction technique. The following is the log-linear model adopted:

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

where β_1 measures the negative association between the read counts and the euclidean distance between (i,j) and β_{enz},β_{gcc} and β_{map} are the estimated coefficients for the enzyme effect, GC content and mappability effect, respectively [1]. Model covariate data such as $(e_i e_j)$, $(g_i g_j)$ is not described nor their source mentioned by the authors in the paper. The model will be further evaluated in Section 4. Given the estimates for θ_{ij}, the authors continue to adopt a simple bayesian view to derive a posterior distribution (m_{ij} is the read count between (i,j)):

$$P(\mathbf{M}|\mathbf{S},\beta) = \prod_{ij} \hat{p}((i,j)|\theta_{ij}) = \prod_{ij} \frac{e^{-\theta_{ij}} \theta_{ij}^{m_{ij}}}{m_{ij}!}$$

$$P(\mathbf{S},\beta|\mathbf{M}) \propto P(\mathbf{M}|\mathbf{S},\beta) \propto \prod_{ij} e^{-\theta_{ij}} \theta_{ij}^{m_{ij}}$$

Once again, as the above posterior distribution is a complex space, the authors use sequential importance sampling (SIS) to generate an initial structure and refine the sample further using Gibbs sampler and adaptive rejection sampling (ARS) [1]. Further analysis is done on the sampled structure based on a statistic called the HD (height-diameter) ratio, formulated by the authors.

4. EVALUATION

The study of the chromosomal 3D structure has revealed the presence of spatially correlated topological domains that present different levels of gene concentration and transcription. It has been shown that mammalian genome is comprised of two general compartments A & B, that are gene rich and gene poor respectively [2]. These discoveries provide promising avenues for further advancements in genomics. However, specific biological results from these models are beyond the capacity of this survey paper and can be accessed at [1; 3].

The main focus of this section will be in evaluating the model specifications of MCMC5C and BACH; this paper will also attempt to suggest appropriate improvements to the models. The authors of MCMC5C begin by building upon an established binomial model (Duan *et al.*) and then approximating

it to a normal. While the normal approximation itself is not problematic, the variance assumed under this approximation is not very well supported. It's important to note that MCMC5C is a model that was formulated to be used on both 5C and Hi-C data. While 5C data provides variance estimates for the recorded interaction frequencies, Hi-C data does not provide any variance estimates for the read counts. As a result, a normal approximated variance is feasible for 5C data, but a similar normal approximated variance for Hi-C data is not. Moreover, the authors offer no statistical rationalization for how the variance for the normal model was chosen.

Another point of contention is the choice of the model. Setting aside the difficulty of variance specification, the authors of MCMC5C state that the binomial model was approximated to a normal for computational efficiency [3]. BACH authors, on the other hand, employ a poisson model that seems more appropriate for count data. However, given the scale of Hi-C data, a poisson could also be approximated to a normal model; as a result the choice of the type of model may be trivial if the parameters are approximated well. One way to test the model fit would be to validate results from both the models with FISH data.

Another important aspect that needs consideration is bias reduction when working with Hi-C data. While there is no mention of any bias reduction techniques in the MCMC5C model specification or paper, the BACH model employs a log-linear model to account for bias in the data. The authors of BACH do not explain the choice of a log-linear model nor do they present any details about the covariate information necessary for it. It is unclear as to why a log-linear model was the choice and the authors could have elaborated more on the statistical suitability in the context of Hi-C data.

As is the general case in the field of computational biology, both these models require a relatively high amount of computational power. But BACH seems to fare slightly better as it samples using multiple MCMC methods to gain refinement. The MCMC5C model was run on Hi-C data of the human chromosome 14, which consists of 89 fragments; mixing was achieved after 4×10^7 iterations and 250 structures were samples in approximately 2.5 hours [3]. An attempt to run MCMC5C on data from all 23 human chromosomes failed to achieve mixing after 24 hours of execution [3]. BACH, on the other hand, achieved mixing in 8 minutes for a genomic location consisting of 13 fragments; it is supposed to take on run time quadratically with the number of fragments. Computational time will still be a challenge, albeit a declining one, in this field.

A final, but crucial, remark is on the dynamic nature of the genomic material. Both the above models attempt to infer the consensus 3D structure of the genomic material; while this is a stepping stone, it is important to acknowledge that very little can actually be gained if there is no measure of the dynamic variance of the genomic material from its supposed consensus structure. Without the knowledge of how the structure intrinsically varies, knowing the consensus structure might not allow for much insight. One model that begins to address this is a generalized version of BACH called BACH-MIX. BACH-MIX allows for a particular genomic region to be divided into n substructures and models the variance between the multiple constructions of the 3D substructures [1]. More details of BACH-MIX can be found in Hu *et al.* [1].

5. CONCLUSION

While both the models have their drawbacks, there are a stepping stone to developing rigorous models to infer the chromosomal 3D structure. These models could gain to understand the model fit by validating their results through, what is considered as gold standard, Fluorescence in situ hybridization (FISH) data. The models, given their bayesian framework, could also gain from adopting more informed priors and knowledge from other genetic studies. It would also be statistically and biologically interesting to run these models on Hi-C data at different resolutions. Finally, given the scope of

Hi-C data, the authors have opened the door to a highly exciting niche in genomics that is awaiting further development of novel statistical methods.

REFERENCES

Ming Hu, Ke Deng, Zhaohui Qin, Jesse Dixon, Siddarth Selvaraj, Jennifer Fang, Bing Ren, and Jun S Liu. 2013b. Bayesian inference of spatial organizations of chromosomes. PLoS computational biology 9, 1 (Jan. 2013), e1002893. DOI:http://dx.doi.org/10.1371/journal.pcbi.1002893

Ming Hu, Ke Deng, Zhaohui Qin, and Jun S. Liu. 2013a. Understanding spatial organizations of chromosomes via statistical analysis of Hi-C data. Quantitative Biology 1, 2 (July 2013), 156174. DOI:http://dx.doi.org/10.1007/s40484-013-0016-0

Mathieu Rousseau, James Fraser, Maria a Ferraiuolo, Jose e Dostie, and Mathieu Blanchette. 2011. Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling. BMC bioinformatics 12, 1 (Jan. 2011), 414. DOI:http://dx.doi.org/10.1186/1471-2105-12-414