DEVELOPING A NEW APPROACH TO VISUALISING CHROMOSOMAL STRUCTURE

ALEX OLSON

4th Year Project Report
Artificial Intelligence and Computer Science

SCHOOL OF INFORMATICS
UNIVERSITY OF EDINBURGH
2018
Acknowledgements

I would like to thank my supervisor, Benjamin Bach, for all of his help in developing this project, and providing guidance and support throughout.

Many thanks to Peter Adams, for providing us with data to test our system, providing extremely helpful feedback, and also supporting me throughout this project.

Thanks to Neil Robertson for his guidance in developing the Hi-C pipeline, and giving advice and ideas to grow this project.

I am grateful to Tamir Chandra for offering us his time and expertise for feedback on our system.

Thank you to Ronell Sicat for all of his help and advice.

I greatly appreciate the help and support of my friends, Arthur Wilcke, Tom Bock, Andreea Cucu, and many many more.

Finally, thank you to my parents for their support.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(Alex Olson)
Abstract

In genetics, capturing the structure of chromatin is an area of active research in which new techniques are rapidly improving the status quo. Unlike other types of biological visualisation, the cutting-edge in chromatin visualisation is simplistic and does not take advantage of many useful techniques.

At the same time, the potential applications of stereoscopic visualisation and augmented reality are not yet explored fully, and their objective merits uncertain.

As a result, we seek to combine these questions into a project which introduces a new, stereoscopic 3D model of chromatin, developed in collaboration with geneticists from the Cancer Research UK Beatson Institute and the MRC Institute of Genetics and Molecular Medicine. This model employs augmented reality in order to allow geneticists to view chromatin models in stereoscopic 3D without interrupting the rest of their workflow.

In order to examine both the utility of the system for exploring chromatin models, and to investigate whether augmented reality and stereoscopic perception enhance understanding of chromatin structure, we developed and piloted a complete study which compares traditional Hi-C heat maps to the new model, while simultaneously comparing on-screen visualisations to mixed reality.

After demonstrating the completed system to our geneticist collaborators, there is clear indication that the produced system is capable of improving understanding of complex structures within chromatin. Our collaborators highly approved of the system and were able to quickly grasp its functionality.
2.1.4 Existing Hi-C Visualisation Methods ...................................... 21
2.2 Visualisation ................................................................. 21
  2.2.1 Terms ......................................................................... 21
    Curve .............................................................................. 21
    Mixed Reality ................................................................. 21
    Stereoscopy .................................................................... 21
  2.2.2 Augmented Reality ...................................................... 22
  2.2.3 Microsoft HoloLens ................................................... 22
  2.2.4 Matrix and Curve Visualisations ................................... 23
  2.2.5 Perception of 3D Curves .............................................. 24
2.3 Prior Work ......................................................................... 25
  2.3.1 Processing Hi-C Data .................................................. 25
    Pipeline .......................................................................... 25
    Algorithm ....................................................................... 25
3 Methods ................................................................................. 29
  3.1 Interface ........................................................................... 29
    3.1.1 Heat Map ................................................................. 29
    3.1.2 Screen ...................................................................... 31
      Colour Scale ................................................................... 31
      Curve Smoothing .......................................................... 32
    3.1.3 Augmented Reality .................................................. 33
3.2 Study ................................................................................... 33
  3.2.1 Data ............................................................................ 33
  3.2.2 Quantitative Study ..................................................... 34
  3.2.3 Techniques ............................................................... 35
    Colour Selection ............................................................. 36
3.2.4 Study Tasks ................................................................. 37
  Point Distance .............................................................. 37
  Larger TAD ................................................................. 38
  Contact Points .............................................................. 38
  Triple Points ................................................................. 39
  Segment Distance .......................................................... 39
  Curve Comparison .......................................................... 39

4 Results ................................................................. 41
  4.1 Qualitative Feedback .................................................. 41
    4.1.1 First Meeting .................................................... 41
    4.1.2 Second Meeting .................................................. 41
    4.1.3 Third Meeting ..................................................... 42

5 Conclusion ............................................................... 43
  5.1 Discussion ............................................................. 43
  5.2 Future Work ............................................................ 44

Bibliography ............................................................. 45
INTRODUCTION

In many fields, a major challenge of large-scale data is its representation in ways that enable comprehension and analysis by users. Stereoscopic perception has enormous potential to enhance our traditional understanding of highly complex, or abstract, datasets, such as those associated with human genome sequencing. Through the use of stereoscopic perception techniques, it is possible that biomedical researchers may have new insights with respect to large datasets.

Nevertheless, there is still a debate about the best way to apply stereoscopic perception techniques to facilitate research and discovery. Although some applications, such as medical visualisations, require 3D in order to fully represent the information, frequently 3D is employed as an alternative choice to a potentially more complex 2D representation. This particularly is the case in techniques such as 3D scatter plots and bar charts, in which the data could be represented using more traditional methods. 3D is currently accepted more as a method for interaction than as a way to faithfully reproduce data for visualisation.

In other areas of genetics and biology at large, there are many examples of 3D visualisation. At the molecular level, applications such as Jmol are widely used to produce 3D images. For protein, the National Center for Biotechnology Information provides an online visualisation tool, iCn3D, which can display pre-generated 3D models of protein. These approaches do not include stereoscopy, but focus solely on a screen projection of the models. In this project, we focus on the visualisation of chromatin (described in Section 2.1.2), the structure of which is determined using a process called Hi-C (described in Section 2.1.3).

When 3D methods are considered, they are typically limited to what is likely a subsection of the full range of their potential application. One reason for this is because such methods often take the form of a 3 dimensional model displayed on a screen. When this is not the case, recent studies have looked at 3D bar charts in physicalised form, or scatter plots in augmented reality. Such studies are invalu-

---


able but fall short of being decisive on the benefit of their respective techniques. The benefit of ‘realism’ added by approaches such as physicalisation and augmented reality for 3D models in information visualisation remains unclear. By investigating a series of approaches that are linked by an intent to supplement or transcend the common screen environment, this project aims to complement this discussion. This investigation is fuelled by rapidly advancing technology for immersive augmented and virtual reality, both of which are intrinsically linked to 3D space.

By applying these novel questions in the field of information visualisation to a specific biological situation, we aim to obtain results with apparent real-world value. This is a critical distinction from prior work in the field, which frequently relies on synthetic data.

In order to ensure that our research followed lines currently of interest in genetics, we have collaborated with several researchers in the field of genetics. Professor Peter D Adams, at the Cancer Research UK Beatson Institute and the Sanford Burnham Prebys Medical Discovery Institute, is interested in the effects of chromatin on cellular senescence (see Section 2.1.1), particularly with regards to ageing and cancer. Professor Adams consulted with us on the nature of his research and the areas in which biological visualisation could assist, which is described fully in Section 4.1. He also provided invaluable feedback on the direction of the project at numerous stages in development. In Professor Adams’ Glasgow laboratory, computational biologist Neil Robertson assisted in designing the method outlined in Section 2.3.1, as well as bridging the gap between biological and computational disciplines. Dr. Tamir Chandra, at the MRC Institute of Genetics and Molecular Medicine, is interested in a variety of mechanisms of ageing, particularly cell senescence, and their impacts on a range of age-related disease including neurodegeneration, cardiovascular disease and cancer. Dr. Chandra examined the completed exploratory environment, providing us with detailed feedback and relating the potential of the project to research in his own laboratory.

Based upon discussions with our geneticist collaborators, we present a high-level overview of the information in which they are interested when observing chromatin data. Our interactive system allows for visualising chromatin in augmented reality using both the matrix
and the 3D structure. In Section 3.2, we describe two user studies. In the first, we record observations and findings from having the geneticists explore data using our system. In the second, we design an evaluation of the two representations: matrix and 3D structure, and the two environments: screen and augmented reality. All of our studies are designed to employ real-world data.

In this project, we wished to determine whether stereoscopic perception can aid in understanding information in complex, irregular 3D structures. Stereoscopic vision has been shown to aid in finding missing links in 3D cave systems (Figure 1.2). In our case, we have investigated 3D chromatin structures (the nature of which is outlined in Section 2.1.2). These structures are highly complex, but understanding their structure is critical to enable understanding specific cell functions and phenotypes.

We are interested in investigating augmented reality additionally as it provides several benefits not seen in other mixed reality techniques (see Section 2.2.1). As augmented reality devices overlay virtual objects into the real world, this allows our geneticist collaborators to view the model while talking to colleagues, performing related experiments, or, critically, while comparing the stereoscopic model to the heat map, viewed on a standard computer screen.

We present a basic interactive system to visualise chromatin structures in immersive augmented reality. We then report on the design of a user study on the perception of information in these 3D chromatin structures. This study, described in Section 3.2, has been fully developed and piloted. Though not yet performed due to constraints in obtaining the desired real-world Hi-C data, this study is due to be carried out this summer and is entirely prepared. We are interested in two research questions:

---


Figure 1.2: Cave survey data from Schuchardt et al. 2007
• How do the two representations - matrices and curves - compare with respect to common tasks?

• Does the stereoscopic representation of the data facilitate the investigation of the data better than existing 2D representations?

Methods presenting Hi-C data are still new and in the cutting edge of the field. There is still a significant amount of progress to be made, particularly in terms of the working resolution of the data. The data used here is at the resolution of 100,000 base pairs, but after consultation with our biological collaborators we have ascertained that certain types of structures are expected to be visible only with much finer grain resolutions.

Current approaches to reconstructing the 3D structure of chromatin are far from perfect, and the resulting curves can be seen as approximations of the true spatial positions of genomic loci. Our studies therefore investigate the potential of 3D chromatin structures for a specific set of tasks, with the hope of foreseeing the potential of such methods once future improvements in computation and data recording allow for an extension of the sets of patterns visible in such structures. Moreover, while AR and VR equipment becomes increasingly affordable and reliable, it is pertinent to examine whether such technology is useful in the bioinformatics space. We believe this project contributes towards advancing the current state of the art in biological data visualisation.
2

BACKGROUND

In this chapter, the background for the project is detailed. As this project is multi-disciplinary, there is a wide range of material to be covered in order to fully understand the contributions detailed in the later chapters.

The first section details the genetics background to the project. We first describe the nature of chromatin, which is the specific complex investigated in this research. The method by which our processed chromatin structure data is obtained is outlined here in full, including our own contributions to this process in Section 2.3.1.

The second section details the visualisation background. This includes both an outline of the state of mixed reality (see Section 2.2.1) techniques, and an explanation of relevant approaches in embedding data into 3D and stereoscopic visualisations. Simultaneous to this project, we are writing a paper on the approaches outlined in this project. Sections 2.2.4 and 2.2.5 have been taken from this paper, which is co-written by Benjamin Bach.

2.1 Biology

2.1.1 Terms

Loci

A genomic locus is a specific section of a chromosome. Hi-C measures interactions between different loci. For example, a locus might here be recorded as chr19 − 42400000 − 42500000, which refers to the region of base pairs between the 4,240,000th and 4,250,000th base pairs in chromosome 19.
Transcription

Transcription is the first stage in gene expression, during which a DNA template is replicated as an RNA strand.

Enhancers

An enhancer is a small region of DNA. When bound by proteins that recognise specific DNA sequences, it increases the likelihood that a corresponding gene is transcribed. They are not necessarily located near the gene they regulate in terms of base pairs, but due to chromatin folding they can be physically close.

Epigenetics

Epigenetics is the study of all mechanisms that affect gene function without modifying the actual sequence of DNA. Chromatin folding is within the study of epigenetics, as it is thought to significantly impact gene expression without modifications of the DNA itself.

Nucleosomes

A nucleosome is a core unit of packaged DNA. It refers to a section of DNA wrapped around a set of histone proteins. This structure is the basic building block of chromatin.

Replicative Senescence

Senescence is a stage in the life cycle of a cell. A senescent cell no longer divides, although it remains active in other aspects. Senescence can be triggered by cell damage or ageing, which typically leads to shortened telomeres.

Cross Linking

Cross-linking DNA refers to a chemical process that bonds two fragments of DNA together.

Ligation

While cross-linking bonds DNA at any sites, ligation refers to the process of joining two fragments of DNA end-to-end, so that a new, complete fragment is formed.
A topologically associated domain is a region in space in which the chromatin within the region interact substantially more with themselves than with chromatin outwith the region.

2.1.2 Chromatin

In animal cells, DNA exists within a structure called chromatin, a mass of compacted DNA and protein. In chromatin, DNA is wound around proteins called histones (Figure 2.1), which allows DNA to be packaged into a very small space - end to end, the human genome within a single cell would be roughly 2 metres long.

The exact nature of this packaging is very important to the functionality of the genome. Enhancers can be separated by as many as 2 or 3 million base pairs from the genes they affect. Due to the structure of chromatin, however, these distant sections can be geographically local due to loops in the compacted DNA.

Understanding this structure is critical to understanding the genome, and is currently a widely researched area. Chromatin structure plays a significant role in epigenetics. Modifications to histones are known to contribute to tumour suppression. Chromatin is also important in understanding how ageing relates to cancer, with recent research revealing how during senescence, chromatin fragments activate pathways causing chronic inflammation associated with cancer.

2.1.3 Hi-C

Although the existence of chromatin has been known since the 1800s, the understanding of its complex structure has been limited by technology. Over the last fifteen years, techniques designed to determine the spatial structure of chromatin have been developed, based on chromosome conformation capture (3C). These techniques share a core process: first, DNA strands are cross-linked. The genome is then cut into small fragments, resulting in many pairs of loci which interact with each other, regardless of their distance in DNA. These pairs are then joined end-to-end, and the crosslinks reversed. These ligated pairs are then analysed using a technique-specific method, in order to quantify the level of interaction between regions of the chromosome (Figure 2.2).
Unlike earlier methods, which are only able to examine interactions within a subset of the genome, Hi-C operates genome-wide. Such approaches have been previously limited due to the high data throughput required. By analysing the entire genome at once, Hi-C removes prior methods’ necessity for targeting specific areas of the genome, and enables both a more exploratory approach as well as the possibility of recovering the full structure of the chromatin.

Read Mapping

Ligation fragments, also called ‘reads’, are first matched against fragments of a sequenced genome in order to identify them. As each end of the read is a different section of the genome, both ends must be separately matched.

Fragment Assignment & Filtering

When both ends of a read are matched to the same DNA fragment, it is possible that this is a single, un-ligated fragment instead of two joined loci. It is also possible for a fragment to be ligated with itself, into a circle. In order to remove these fragments, which interfere with the data, all read pairs that fall into the same restriction fragment are filtered.

Binning

The raw resolution of Hi-C at this stage is very high - typically the human genome is split into sections of 6 base pairs. While theoretically possible, in practice sequencing at this resolution is uncommon, both due to the sheer size of the data produced as well as noise from the experimental process. Instead, the fragments are binned into larger sections, reducing the resolution but also the effect of noise and the size of the data.

Bin Level Filtering

At this stage, specific bins are removed from the data, such as bins that are located around areas of the chromosome which are difficult to map, namely the centromere. This is because such areas contain...
short, repeated sequences which vary in length with age, preventing easy matching with a reference genome.

**Normalisation**

As the likelihood of successful mapping varies along the length of the genome, each bin has a separate bias that must be accounted for. Many methods exist to eliminate this bias, including probabilistic modelling\(^9\).

**Result**

The end result of this process is a set of filtered interaction frequencies or contact probabilities. These values indicate the likelihood that two loci interact with each other, and are proportional to distances between loci.

Processed Hi-C data is typically rendered as a two dimensional ‘contact matrix’ (Figure 2.4). Each cell of this heat map represents contact between two genomic loci, and the intensity of the cell colour corresponds to the likelihood of contact. Darker cells indicate higher spatial proximity between the base pairs in the loci indicated by the corresponding row and column. These matrices are advantageous in that they show patterns in the chromosome in a manner that is straightforward to generate and is typically clutter-free. Cutting-edge software allows exploration of this heat map representation\(^10\), by selecting and identifying patterns in the matrix. Such methods support pan-and-zoom interaction, or interactive analytical methods for the exploration of individual patterns that have been extracted by prior analysis.

Contact matrices support geneticists in exploring specific patterns, of which the most common are TADs and loops. Within matrices, these structures are most evident as specific patterns of matrix cells, as in Figure 2.3.

However, relating the matrix pattern to real-world chromatin structural features is difficult. When consulting with our geneticist collaborators, they cited situations with more than two genomic loci interacting as being especially difficult to relate to the heat map representation. As such, visualising the structure in 3D is expected to reveal a view of the chromatin that is complimentary to the standard heat map, exposing higher-level and more complex structures than are easily seen in 2D.

Separate to the heat map representation of Hi-C data, it is also possible to generate a 3D chromosome by relating the contact probabilities to distance. This approach requires some manipulation of the data, as discussed in section 2.3.1.

---


\(^10\) Fritz Lekschas, Benjamin Bach, Peter Kerpedjiev, Nils Gehlenborg, and Hanspeter Pfister. HiPiler: Visual Exploration of Large Genome Interaction Matrices with Interactive Small Multiples, 2017. ISSN 10772626
Figure 2.3: This specific heat map pattern corresponds to a simple ‘double loop’ structure, in which the curve loops back on itself two times. It is immediately apparent that the heat map representation does not lend itself well to appreciating and identifying such structures.

Figure 2.4: Juicebox Hi-C heat map from Durand et al. 2016. This is an example of the standard representation of processed Hi-C data.
2.1.4 Existing Hi-C Visualisation Methods

In the past two years, several approaches to generating a 3D chromosome from Hi-C data have been put forward. Due to the size of the human genome, and the optimisation problem posed by relating contact probability to distance, each method has its own approach to producing a coherent output. In some approaches, such as Chrom3D\(^{11}\) and TADbit\(^{12}\), the chromosome is reduced into units of around 1 Mbp, based on TADs. As TADs are areas of the chromosome in which a significant portion of interactions are internal to the domain, it is possible to ignore these internal interactions and focus on the intra-TAD interactions, which greatly reduces the optimisation problem. Other approaches employ simulated annealing to convert the entire Hi-C dataset from contact probability to distance\(^{13}\).

For visualising the matrix output, many methods exist, including Juicebox\(^{14}\) (Figure 2.4), as well as HiGlass\(^{15}\), which is a web-based approach.

2.2 Visualisation

2.2.1 Terms

Curve

We use the term *curve* to refer to the 3D chromatin structure. This is based upon earlier visualisation research (e.g. Bach et al. [2016]) in which the term is used to describe similar constructs.

Mixed Reality

Mixed Reality is a term used to describe the entire category of technology which seeks to merge elements of the real and the virtual\(^{16}\). Both Virtual Reality and Augmented Reality fall into this category.

Stereoscopy

Stereoscopy refers to techniques which employ two images of the same scene, taken from slightly offset positions, in order to infer depth. This technique is employed in human vision, and can as such be used in headsets such as the HoloLens to provide a convincing depth effect.


\(^{16}\) Paul Milgram and Fumio Kishino. Taxonomy of mixed reality visual displays. *IEICE Transactions on Information and Systems*, 1994. ISSN 0918-8552. doi: 10.1109/51.4646
2.2.2 Augmented Reality

The field of augmented reality technology arguably began in earnest with Ivan Sutherland’s 1968 head mounted display\(^\text{[7]}\), which achieved virtual 3D objects using head-mounted cathode ray tubes (CRTs) connected to a mechanical arm attached to the ceiling of the room (Figure 2.5). By using a separate CRT per eye, Sutherland was able to generate stereoscopic 3D images, the position of which was updated relative to the headset. By combining these techniques, the illusion of a virtual object being ‘in the room’ was achieved for the first time.

The state-of-the-art quickly advanced. VIDEOPLACE by Krueger et al. in 1985\(^\text{[18]}\) utilised multiple desktop computers connected to a video feed of the user’s surroundings, allowing for rudimentary interaction with virtual reality in a way less spatially constrained than with Sutherland’s headset, which limited movement to where the mechanical arm could reach. However, as with Sutherland, this approach required a significant amount of custom-built technology, which prevented it from being applied to practical applications at scale. Krueger primarily utilised his VIDEOPLACE as an art installation, although he did discuss at length possible applications to video games and telecommunications.

With ARToolKit, released in 2002\(^\text{[19]}\), open-source video tracking to overlay virtual images onto a live feed became possible. With smartphones, this technology became mobile, enabling the wide-scale accessibility lacking in previous technologies. However, this came at the cost of stereoscopy, which was prohibitively computationally expensive until recently. Google Cardboard headsets now allow for smartphones to display stereoscopic 3D images. These can be overlaid onto a live video feed of the user’s environment.

2.2.3 Microsoft HoloLens

Microsoft HoloLens\(^\text{[20]}\) uses a series of cameras mounted on a wearable headset to map the user’s environment, and subsequently track their position. This, in addition to sensors including a gyroscope and magnetometer\(^\text{[21]}\), enables the headset to keep virtual objects in place within the environment. This set of sensors allows the device to maintain stable room tracking, keeping objects in the virtual environment exactly in place even when the user is not looking at them. Two separate, transparent screens allow for stereoscopic 3D images to be overlaid on the user’s environment directly, without the need for a video feed. Unlike other devices that require continuous connection to a desktop computer, the HoloLens contains an on-board computer for processing 3D environments. Unlike augmented reality approaches using mobile phones or tablets, which require the user to hold the device in front of them, the HoloLens rests on the user’s head, allowing for full use of their hands.
Devices such as the HoloLens go further than to simply combine the most important features for an immersive AR experience. For applications of the technology, far more important is mass production. Prior attempts at AR took the form of custom-built, unique devices. While viable for proof of concept, such devices naturally lacked the support necessary for third party development. With HoloLens, the technology is now sufficiently accessible for practical applications to be feasible. Now that augmented reality in the workplace is viable, research into its utility is increasingly important\textsuperscript{22}. It is necessary to investigate for each potential application, whether augmented reality provides a tangible improvement over 3D projections or 2D visualisations.

\textsuperscript{22} Benjamin Bach, Ronell Sicat, Johanna Beyer, Maxime Cordeil, and Hanspeter Pfister. The Hologram in My Hand: How Effective is Interactive Exploration of 3D Visualizations in Immersive Tangible Augmented Reality?, 2017

Figure 2.5: Sutherland’s augmented reality headset tracked head position using a mechanical arm attached to the ceiling of the room.

2.2.4 Matrix and Curve Visualisations

The matrix and curve visualisation are linked by the intrinsic function of embedding. An embedding assigns data points a position in one-, two-, or three-dimensional space, depending on some encoding
of their specific nature. Curves can be described as an embedding technique as they use space to indicate the relationships between elements. Other embedding methods in information visualisation include node-link diagram layouts, as well as multidimensional scaling methods.

Matrix representations are often regarded as a convenient method, allowing for both an overview of a dataset and a more detailed inspection of its local patterns. For example, an array of research has been carried out into the use of matrices for network visualisation. Ghoniem et al. reported on a study on unordered matrices, which showed that such a representation supported most tasks on networks better than equivalent node-link diagrams. Alper et al. [2013] conclude similar results for graph comparison tasks. Consequently, a variety of sophisticated visualisation techniques have been created for matrix visualisations, such as NodeTrix [23], MatrixWave [24] and Multi-Piles [25].

Currently, matrices are the standard visualisation for depicting similarity and correlation. For time-varying self-similarity, i.e. showing the amount of change in a dataset over time, time curves [26] have been investigated as an alternative to matrices. Time curves embed versions of a dataset with multidimensional scaling using a corresponding similarity matrix to generate distances. As versions (represented as distinct beads) are ordered in time, an adapted bezier spline (a similar technique as described in Section 3.1.2) is used to connect the points in this order, which results in highly informative visual patterns. As reported by the authors, the emerging patterns describe an ‘interesting space’ of patterns that are difficult to detect in a matrix.

2.2.5 Perception of 3D Curves

3D curves are widely employed in scientific visualisation, representing wind flows [27] or blood vessels [28, 29].

Curves are perhaps some of the most difficult shapes to understand, because of their complex shapes. The intrinsic question when visualising 3D curves is how well patterns in complex structures are perceived. Generally speaking, some structures are more difficult to understand in 3D than others. For example, the 3D structure of a chair or a human body can be understood almost immediately, as there is a depth of experience in recognising these objects which the brain is able to rely upon. The same holds for simple abstract structures such as cubes or grids. However, more complex abstract structures, or irregular structures, are significantly more difficult to immediately understand. These cases necessarily require more time to investigate, but most importantly rely on active exploration involving viewpoint changes through rotation and zoom. A well-known example of this effect is the dazzling camouflage used by British war-
ships during World War One\(^3\) - by introducing complex abstract shapes into the hull of a ship, onlookers were unable to immediately recognise it and process the size or direction of travel correctly.

### 2.3 Prior Work

#### 2.3.1 Processing Hi-C Data

**Pipeline**

We have developed a pipeline to process raw Hi-C data and convert it into pairwise contact frequencies. This pipeline uses runHiC\(^3\) and hiclib\(^3\) to perform the analysis described in Section 2.1.3. The raw Hi-C data is mapped to the human genome\(^3\), then filtered and binned as described. Bias is then corrected for, and the resultant data is converted into pairwise contact frequencies.

**Algorithm**

Separate to existing methods, we have developed our own approach to produce a model of chromatin from the processed Hi-C data described above. This data does not directly describe the distances between parts of the chromosome, which is needed in order to produce a physical model. Instead, the data comprises contact frequencies between genomic loci. In order to reconcile this with the required spatial distance information, contact frequency is treated as directly proportional to proximity between loci. Such an assumption is well-established in existing approaches, and is known as "wish-distances\(^3\).

As wish-distances do not necessarily directly correspond to true distances, there is no guarantee that points will "fit together" when using these values naïvely. As such, any method employing wish-distances must introduce a method to place genomic loci such that the difference between the true distances and the wish-distances is minimised. Such a problem is immediately recognisable as an optimisation problem.

In order to minimise the difference between the recovered distances and the wish distances, other approaches have employed simulated annealing, treating each locus at the resolution of the dataset as a distinct 'bead'. (e.g. Adhikari et al. [2016] and Paulsen et al. [2017]).

Our approach instead employs a genetic algorithm. This is in part to enable comparison of these two approaches' performance on the task, and is motivated by results in other fields which indicate that a genetic algorithm approach can often be superior to simulated...
The algorithm used to recover distances is described below. The implementation is written in C++ and uses a number of optimisations to reduce the running time of the program.

---


begin
  \( \text{generation} := \text{makeFirstGeneration}(\text{popSize}) \)
  \( \text{mutationRate} := \text{initialValue} \)
  \( \textbf{while} \ \text{mutationRate} \geq \text{finalValue} \ \textbf{do} \)
  \( \quad \text{best} := \text{rankGeneration}(\text{generation}) \)
  \( \quad \text{generation} := \text{makeGeneration}(\text{best}, \text{mutationRate}) \)
  \( \quad \textbf{if} \ \text{best} \approx \text{lastBest} \)
  \( \quad \quad \text{mutationRate} := \text{reduceMutation}(\text{mutationRate}) \)
  \( \textbf{end} \)
  \( \textbf{end} \)

proc \text{makeFirstGeneration}(\text{popSize}) \equiv 
  \( \text{generation} := [] \)
  \( \textbf{for} \ i \ \textbf{to} \ \text{popSize} \ \textbf{step} \ 1 \ \textbf{do} \)
  \( \quad \textbf{for} \ j \ \textbf{to} \ \text{numberOfPoints} \ \textbf{step} \ 1 \ \textbf{do} \)
  \( \quad \quad \text{generation}[i][j] := \text{random3dPoint()} \)
  \( \textbf{end} \)
  \( \textbf{end} \)

proc \text{rankGeneration}(\text{generation}) \equiv 
  \( \text{sumError} := 0 \)
  \( \textbf{for} \ i \ \textbf{to} \ \text{popSize} \ \textbf{step} \ 1 \ \textbf{do} \)
  \( \quad \textbf{for} \ j \ \textbf{to} \ \text{popSize} \ \textbf{step} \ 1 \ \textbf{do} \)
  \( \quad \quad \text{sumError} = \text{sumError} + \text{abs}(\text{wishDistances}[i][j] - \text{distance}(i, j)) \)
  \( \textbf{end} \)
  \( \textbf{end} \)

proc \text{makeGeneration}(\text{best}, \text{mutationRate}) \equiv 
  \( \text{generation} := [] \)
  \( \textbf{for} \ i \ \textbf{to} \ \text{popSize} \ \textbf{step} \ 1 \ \textbf{do} \)
  \( \quad \textbf{for} \ j \ \textbf{to} \ \text{popSize} \ \textbf{step} \ 1 \ \textbf{do} \)
  \( \quad \quad \textbf{for} \ k \ \textbf{to} \ \text{numberOfPoints} \ \textbf{step} \ 1 \ \textbf{do} \)
  \( \quad \quad \quad \text{r} := \text{randomChoice}(\text{best}[i][k], \text{best}[j][k]) \)
  \( \quad \quad \quad \text{generation}[i][k] := \text{mutate}(\text{r}, \text{mutationRate}) \)
  \( \quad \textbf{end} \)
  \( \textbf{end} \)
  \( \textbf{end} \)

proc \text{mutate}(\text{r}, \text{mutationRate}) \equiv 
  \( \textbf{for} \ i \ \textbf{to} \ 3 \ \textbf{step} \ 1 \ \textbf{do} \)
  \( \quad \text{r}[i] := \text{r}[i] + \text{random}(-\text{mutationRate}, \text{mutationRate}) \)
  \( \textbf{end} \)
  \( \textbf{end} \)

end
In this chapter, the methods developed for this project, as well as
the approach for the study, are detailed. The visualisation software
has all been written using C# in Unity 2017.2.1f1, the latest version
supported by Microsoft’s Mixed Reality Toolkit. Unity was used due
to its cross-platform support, particularly for mixed reality head-
sets such as the Microsoft HoloLens. With Unity, the same under-
lying software could be deployed both for the desktop environment
and the HoloLens, ensuring maximal similarity between the environ-
ments.

As with the Background chapter, this chapter has been written
concurrently with the paper, co-written with Benjamin Bach, and so
there is some overlap. In particular, Sections 3.2.1, 3.2.3, and 3.2.4
will also appear in the paper.

3.1 Interface

3.1.1 Heat Map

To provide a baseline for comparison, a contact matrix visualisation
was developed in Unity. As described in Section 2.1.3, each cell in a
matrix represents the contact probability between two genomic loci.
Each axis runs through the same chromosome, start to end, and the
intensity of each cell increases with higher probability. The system
is able to generate a heat map from processed Hi-C data, at the orig-
inal resolution. From the data, a texture is generated where each
pixel corresponds to a cell in the contact matrix. This texture is then
stretched to the size of the screen while retaining its original aspect
ratio. Scroll and zoom functionality is available with mouse and key-
board, and arrow keys can be used to switch the chromosome being
displayed.
30 DEVELOPING A NEW APPROACH TO VISUALISING CHROMOSOMAL STRUCTURE

Figure 3.1: Human chromosome 16 heat map in the Unity application described in Section 3.1.1.

Figure 3.2: In order to fully understand the challenges presented by each combination of model and interface, it is critical to distinguish between the different levels at which dimensionality is at play. At a conceptual level, the heat map representation operates on two dimensions. It is also rendered at this same level of dimensionality - it is not a projection onto a lower dimension as with the screen representation of the curve, for example. Finally, the ‘pan’ and ‘zoom’ operations available also operate at the level of two dimensions.
3.1.2 Screen

The 3D projection accepts the finished output of the optimisation procedure described in Section 2.3.1. If running the code directly on the HoloLens, which is limited in rendering capability due to the onboard computer, the structure of the chromosome is rendered as a near 2D line through 3D space. When using the Holographic Remoting protocol, which allows for each frame to be rendered on a linked computer instead of on the HoloLens itself, a more complex version of the model is shown. On the computer, a mouse can be used to rotate the chromosome around its centre, with the scroll wheel allowing for zoom functionality. Right-clicking on the chromosome brings up a window displaying the identifier for the loci closest to the mouse, e.g. chr16-100000-200000. As with the 2D heat map, using the arrow keys switches the chromosome displayed.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{screen_representation.png}
\caption{The screen representation projects a three-dimensional curve onto the two-dimensional screen space. The pan functionality on the mouse translates two-dimensional mouse movements into three-dimensional mouse rotation.}
\end{figure}

Colour Scale

Histone modifications can be mapped directly onto the curve. When this is done, colours are mapped to the raw values using a method designed to highlight differences, while not being distorted by outliers:

\begin{verbatim}
begin
    sortedValues := sort(rawValues)
    lastValue := 0
    gradients := []
    for i to sortedValues.count() do
        gradients.add(sortedValues[i] - lastValue)
        lastValue := sortedValues[i]
    end
    colorMap := generateColorGradient()
    modifiedColorMap := []
    for i to gradients.count() do
        modifiedColorMap.add(colorMap[gradients[i]])
\end{verbatim}
This method ensures that the difference between neighbouring values is clear, even when large outlier values exist that would normally force the rest of the values closer together in the colour scale.

The colour scale is displayed with the curve, with the width of the scale corresponding to the number of points mapped to that section of the scale. This allows for an intuitive understanding of the method without a detailed explanation of its functionality.

Curve Smoothing

In order to improve the appearance of the 3D curve, splines are used to smooth the line, such as centripetal Catmull-Rom splines, to interpolate between existing points. This works by using three points on either side of each given point to generate new intermediary points. First, the coefficients of the cubic polynomial are calculated:

\[
\begin{align*}
a &= 2p_1 \\
b &= p_2 - p_0 \\
c &= 2p_0 - 5p_1 + 4p_2 - p_3 \\
d &= -p_0 + 3p_1 - 3p_2 + p_3
\end{align*}
\]

Second, for a time value \( t \), between 0 and 1, which represents the point along the line from \( p_0 \) to \( p_3 \), we calculate the cubic polynomial:

\[
p_{\text{new}} = 0.5(a + bt + ct^2 + dt^3)
\]

We can modify the value of \( t \) in order to introduce as many intermediary points as we require. However, this comes at a substantial computational cost, as adding additional points significantly increases the complexity of the curve.

3.1.3 Augmented Reality

The mixed reality environment is based upon the same visualisation as the 3D projection. Microsoft’s Mixed Reality Toolkit\(^3\) enables HoloLens support. In the mixed reality environment, the chromosome is affixed in the space the user begins the application in, and can be viewed from any angle by physically moving around the room. Unlike on the desktop, it is not possible to switch chromosome during operation.

3.2 Study

To understand the potential of curve visualisations for the analysis of Hi-C data, we planned to spend several hour-long interviews and discussions with two genetic research groups. During the discussions, we showed representations of matrices and 3D curves in both environments- on-screen and in augmented reality.

3.2.1 Data

For the user study we elected to use real-world data, in order to maintain difficulty and ensure validity of the results. Data was anticipated to be obtained from one of our collaborators (see examples of our in-development data in Figures 3.6(b)-3.6(f)). Each trial was designed to use a different dataset. This data would comprise Hi-C captures of Human chromosomes 10-22 and chromosome X. For each chromosome, there would be a proliferating and senescent version.

Our in-development captures are processed at a resolution of 100,000 base pairs, resulting in approximately 1,500 loci for the largest, chromosome X, and approximately 350 loci for the smallest, chromosome 22. This data was sufficient for testing the system in development, but insufficient for performing the study itself, which is why we organised with our geneticist collaborators to obtain more recent, and higher resolution, data.
In the following sections, as well as in the study description given to participants, we used a simplified terminology referring to the visual elements on the screen, rather than describing the biological concepts. None of our participants was expected to have any greater knowledge of chromatin data, as in Boy et al. [2014].

3.2.2 Quantitative Study

To better understand the importance of matrix representations vs. curve renderings, as well as to assess the impact of virtual reality, we planned to conduct a controlled user study.

Based upon our initial discussions with the geneticists, we selected four tasks together with the geneticists. Out of the large variety of possible tasks, we selected those that we thought would challenge the matrix visualisation. This is due to the fact that we believe the Hi-C matrix to have many advantages and support many straightforward visualisation tasks well. However, in this project we are interested in the benefits of curve renderings and the impact of stereoscopic holograms. Thus, as we believe that matrices and curves are highly complimentary techniques, our study aimed to challenge Hi-C matrix representations and provide evidence for or against curve visualisation on-screen and in augmented reality.

Each user would be asked to perform a sequence of these tasks, sequentially in each environment. The order of introduction for the environments would be varied. The time taken for the user to arrive at an answer, indicated by pressing a button specific to each platform, would be recorded. Whether the user arrived at the correct answer would also be recorded.

We decided on an in-between subjects design, where each participant would be running through all the techniques and tasks. Trials were blocked by technique. As solving tasks was quite different between the Matrix view and the two curve views, we were concerned about a possible learning bias if participants were to perform the CurveScreen and AR views in sequence. To lower any eventual training bias, we decided on two groups for the technique ordering, with the Matrix view always as the second technique:

- The AR group would perform the following order:
  CurveScreen
  Matrix
  AR

- The CurveScreen group would perform the following order:
  AR
  Matrix
  CurveScreen
This way, should either group (AR group or CurveScreen group) perform better in their respective final technique (AR or CurveScreen respectively), this effect could be assumed a consequence of training.

The study was planned to be conducted in a quiet room during the day. Participation would be voluntary and participants could skip the study at any point if they so desired. Upon arrival, participants would receive an introduction to the study and its objectives, and would be shown both the HoloLens and desktop computer. During all techniques, participants would be asked to remain seated, with the possibility to stand during the AR technique. After each technique, participants would switch environment and could take a break if desired. From previous studies on HoloLens, no side-effects of dizziness had been reported.

3.2.3 Techniques

All techniques were implemented and run from Unity (Section 3.1). The study would involve two setups:

- The desktop setup was specified by a 22-inch monitor attached to a MacBook Pro (2016). Participants would interact using the keyboard and an attached mouse (not the laptop trackpad).

- Augmented Reality was achieved using the Microsoft HoloLens headset. To improve rendering performance, frames were rendered on a Lenovo Thinkpad laptop and transmitted to the HoloLens headset through the Holographic Remoting Player. The HoloLens and the laptop shared a private Wi-Fi router with no other devices connected, in order to ensure a stable connection with high bit rate.

Based on these two environments, we planned to test the following three techniques:

- **Matrix**: This technique consisted of the Hi-C matrix rendered in full screen in the Desktop environment. The interface provided pan and zoom operations (Section 3.1.1). Cell shading was inspired by HiGlass, using a yellow-to-orange colour scale (Section 3.2.3). At the start of each trial, the matrix would be shown at the lowest possible zoom level, i.e. fully visible on the screen. Participants would interact with pan (mouse drag) and zoom (mouse wheel).

- **CurveScreen**: This technique rendered the 3D chromatin curve on a 2D screen in the Desktop environment. Participants could rotate the curve by mouse.

- **AR**: In this technique, the 3D chromatin curve was rendered and perceived in the Windows Mixed Reality holographic environment. Participants could rotate the curve using the computer.
mouse. This condition was run using Holographic Remoting, performing the entire rendering cycle on the laptop computer, while streaming images to the HoloLens.

We did not plan to test any matrix conditions in the AR environment, as we believe that the desktop screen and AR environments would be functionally identical for viewing a 2D matrix. Visualisation sizes were kept consistent across all techniques in order to reduce bias.

**Colour Selection**

In order to ensure that the specific colours used in the study do not interfere with the participant’s judgement, or lead them to favour any certain conclusion, standardised colours were used. These colours have been specifically selected to ensure a balanced transition according to human perception\(^7\).

3.2.4 Study Tasks

In this section we outline the tasks selected for the trial. Additional tasks were developed fully and piloted - namely, the Triple Points, Segment Distance, and Curve Comparison tasks. These tasks are outlined briefly at the end of the section, with an accompanying explanation as to why they were not selected for the study.

Point Distance

Which of the two pairs of coloured segments are closer to each other: the red pair, or the blue pair?

In the CurveScreen and AR views, 2 pairs of beads (genomic loci) were coloured on the curve (Figure 3.6(b)). One pair was coloured orange, the other blue. Participants would be asked to assess which pair of segments was closer to each other not by physical space, but by length along the curve itself. This task was inspired by the distance task in Bach et al. [2017], and would serve as a general spatial perception task about distances in the three testing conditions. In the Matrix view, loci were coloured at the margin of the matrix (Figure 3.6(a)). Boundaries of the segment were highlighted through fine lines in the matrix. Participants would be asked to compare the proximity values for the highlighted segments - i.e. the darkness of the shaded pixels that indicate the distance between the two segments in each colour pair, respectively. Starting from the original real-world data, we generated the data for this task as follows:

1. A raw Hi-C file is randomly selected.
2. From this file, a line is randomly selected. Each line contains two loci and the corresponding interaction frequency between them.
3. If the physical distance between the two loci is within the toler-
ance, then this pair of points is selected for the blue points. Otherwise, new lines are randomly selected until the condition is met.

4. The remaining lines in the file are shuffled into a random order.

5. For each line in the shuffled file, the colour distance (see Section 3.2.3) is calculated. If this value is within the tolerance, and the physical distance between the two new loci is also within tolerance, the line is selected for the red points.

This process ensures that the tasks fall within the level of difficulty required, but still allow the study to rely upon real-world data.

Larger TAD

Which of the two visible clusters is packed less densely in space?

This topological task asks for a global characteristic (volume). Volume is an intrinsically spatial property that is only represented implicitly in matrices - light shading indicates larger distances while dark shading indicates closer distances. Thus in the Matrix view, participants would have to read density from the shading of the blocks along the diagonal of each TAD - darker blocks indicating more tightly packed TADs due to the higher proximity between points, and lighter blocks indicating the less tightly packed TADs spanning a larger volume. We worked to ensure that the number of points within each TAD was not correlated with the volume of the cluster as this would have meant that large blocks of any colour on the matrix would indicate large volume.

Contact Points

How many of the red and blue segments are in close proximity to each other?

This task is similar to the Point Distance task (Section 3.2.4), but expanded to a set of 3 to 6 pairs of mutually coloured segments: each pair consists of one red and one blue point. The participants’ task would be to find the pairs of points where the two points were nearly touching. This task combines topology with attributes and represents cases where geneticists are interested in finding touching points with distinct attributes, such as gene promoters and enhancers.

In the CurveScreen and AR views, the curve showed coloured segments (Figure 3.6(f)), whereas in the matrix colours were again indicated at the margin of the matrix (Figure 3.6(e)). Solving the task in the CurveScreen and AR views required rotating the curve as coloured segments were distributed across the entire structure.
**Triple Points**

In this task, rather than determining the number of pairs of points which were sufficiently close together as in the Contact Points task, participants were asked to identify the number of loops in the chromosome which doubled back on themselves an additional time, to form two loops. This would result in three points in very close contact. However, this task was nearly impossible on the heat map (see Figure 2.3), and on the CurveScreen and AR views it was very similar to the Contact Points task.

**Segment Distance**

This task was nearly identical to the Point Distance task, except rather than highlighting only the pairs of points which the participant would need to examine, the entire length of the curve between each pair was highlighted. However, this was deemed to be too straightforward, and was already mutually exclusive with Point Distance for inclusion.

**Curve Comparison**

Here, participants were asked to examine an initial “reference” curve, and two additional curves. The participants were tasked with identifying which of the two additional curves was most similar to the reference curve. This task was possible with entirely synthetic data, but there was no way to control for the similarity or dissimilarity of two real-world chromosomes. As a result, it was impossible to ensure a consistent level of difficulty across examples.
4

RESULTS

4.1 Qualitative Feedback

This project was carried out with the consultation of several geneticists with research interests directly involving chromatin, in order to evaluate the efficacy of the applications. Professor Peter D Adams and Neil Robertson, from the Cancer Research UK Beatson Institute, provided feedback at several stages in the project, ensuring close ties with the practical applications of the project. Dr. Tamir Chandra provided feedback at the conclusion of the project, evaluating the completed system.

The results of the qualitative analysis of the tool is also set out in this section.

4.1.1 First Meeting

The first consultation was carried out in advance of the project. The outline for the project was discussed, and contact between enhancer and promoter regions was identified as an important application of the visualisation. Additionally, visualising related attributes was considered a potential issue with the 2D heat map visualisation.

4.1.2 Second Meeting

The second meeting took place after the initial development of the HoloLens application was completed. Demonstrating this and the other visualisations represented a large portion of the consultation.

It was quickly found that the ability to directly interact with the chromosome in the HoloLens was missing. An immediate suggestion was to be able to ‘pull apart’ the chromosome, which was discussed as being something slightly separate to a zoom mechanic. This latter option was also something thought to be useful.
We investigated the idea of pulling apart the chromosome - due to variable certainty about the level of interaction between points, the elasticity could inversely correlate with the certainty of the data. That is to say, when the chromosome is pulled on, the distances which have little data are less rigid, and areas where there is more confidence are more rigid. Upon release, the chromosome would return to the original shape. We were concerned that this would add additional computational burdens to the HoloLens.

We also touched on the issues apparent with large clouds of points. One option discussed was that transparency could correlate with the density of points. Areas of low density would be less transparent, and high density areas more opaque, so that TADs could be more easily identified.

Professor Adams was additionally very keen on applying this technology to the growing area of mapping gene-enhancer interactions. There is a lot of active study in this area, particularly as the resolutions necessary to obtain the necessary detail are only recently available. Context was important - heat maps are only capable of showing the relationship between two dimensions at a time. With the structure, it is possible to look at the interactions of the entire structure at once.

4.1.3 Third Meeting

The third meeting took place after the completion of the development of the tool. This meeting took place with Dr. Tamir Chandra, allowing us to observe the areas in which there was agreement with Professor Adams’ lab. The finalised tool was presented. By using the final tool, the geneticists validated many of our previous discussions.

As with the second meeting, there was agreement that the curve allowed for the identification of more complex structures than that which could easily be seen on the matrix. These structures were related to secondary structures in protein, which are three dimensional structures in local segments of protein - as opposed to the global structure. These structures, which most commonly take the form of alpha helices and beta sheets, do not have a direct equivalence in chromatin structure. However, such an equivalent could be discovered once sufficiently high-resolution Hi-C data is used in combination with a 3D structure recovery algorithm.

It was also highlighted that in protein structure research, it is well-established that protein structures are better represented as 3D models than as a heat map.


5

CONCLUSION

5.1 Discussion

We have here presented an entire system to translate raw Hi-C data into an interactive 3D model, which can be used in an exploratory environment on either a computer screen or in a Windows Mixed Reality environment. This system is highly stable, and has been verified to work on real-world data provided by our geneticist collaborators.

This environment has been demonstrated to geneticists from two laboratories, who have provided detailed feedback and confirmed the utility of the system in allowing for the exploration of chromatin models. From both laboratories, there was particular interest in the potential for this system to provide an avenue to investigate gene-enhancer interactions, which was identified as a potentially critical area of research. There was some concern that the current resolution of data available is insufficient to allow for visualisation of such smaller structures, but it is clear that the technology developed for this project will support such data should it become available.

A complete study into the efficacy of the 3D model, as well as into the potential benefits of augmented reality systems over traditional screen interfaces, was also developed. This study was fully realised, with tasks prepared and consulted against. It was also piloted using existing datasets, in order to ensure that the tasks were viable, with four participants. However, due to a delay in obtaining the final dataset, and owing to a longer-than-expected processing timeline, it was not possible to complete the study within the course of this project.

Valuable feedback from our geneticist collaborators lends credence to this project’s potential. With some further investigation, it is highly possible to show conclusively that the presented environment can aid geneticists in exploring the structure of Hi-C data.
5.2 Future Work

Chief in further work required is the completion of the study, the design of which has been presented in its entirety here. This requires little more than obtaining data of a quality sufficient to meet the standard outlined here, and then assessing this data using the developed study code.

Although possible to perform the study with the data used for piloting, we have designed the tasks with the data described in mind and as such elected not to perform the study. We expected to receive processed Hi-C data from our geneticist collaborators shortly after the final feedback meeting with them in March. However, this data was only received weeks later and, instead of processed Hi-C data we received the raw values. Although it is possible to perform the entire processing pipeline on this data, this entire pipeline would be expected to take around a week per chromosome. As such it was impossible to perform this analysis in the time remaining.

Further developments to the environment which would likely be of benefit would include tracking the curve in augmented reality onto a movable marker. This would allow the user to rotate the 3D model in an intuitive way. This was touched on in Section 4.1.2, and could be expanded upon further.

Additionally, in the on-screen representation, it would be valuable to investigate alternative methods of interacting with the curve than using a traditional mouse interface. This is because, as indicated in Figure 3.3, the mouse interface uses fewer dimensions than the 3D curve, which can cause difficulty for the user in understanding how a 2D transformation is mapped into 3D space.

In this study we set out to demonstrate that 3D visual perception techniques could assist biomedical scientists in the analysis of massive and complex datasets. Furthermore, we sought to create a 3D visual perception system that could assist these scientists, leading to discovery, which has been previously hindered by the use of traditional 2D visualisation techniques.

Through close discussion with these geneticists, we have shown that 3D visual perception techniques are likely a valid and useful tool for the investigation and visualisation of massive scientific datasets.


Ann Dean. On a chromosome far, far away: LCRs and gene expression, jan 2006.


Todd Holmdahl. A closer look at the Microsoft HoloLens hard-
30/build-2015-a-closer-look-at-the-microsoft-hololens-
hardware/.

Maxim Imakaev: hiclib: a collection of tools to map, filter and analyze

Peter Kerpedjiev, Nezar Abdennur, Fritz Lekschas, Chuck McCal-
hum, Kasper Dinkla, Hendrik Strobelt, Jacob M Luber, Scott B
Ouellette, Alaleh Ahzir, Nikhil Kumar, Jeewon Hwang, Burak H
Alver, Hanspeter Pfister, Leonid A Mirny, Peter J Park, and Nils
Gehlenborg. HiGlass: Web-based Visual Comparison And Explo-

Ivan Krivega and Ann Dean. Enhancer and promoter interactions-
long distance calls, apr 2012. ISSN 0959437X. URL http://

Myron Krueger, Thomas Gionfriddo, and Katrin Hinrichsen. VIDEO-
PLACE - An Artificial Reality. ACM SIGCHI Bulletin, 16(4):35–40,
1985. ISSN 07366906. doi: 10.1145/1165385.317463. URL http:
//dada.compart-bremen.de/docUploads/ch85{-krueger.pdf.

Bryan R Lajoie, Job Dekker, and Noam Kaplan. The Hitchhiker's
guide to Hi-C analysis: Practical guidelines. Methods, 72(C):65–75,

Kai Lawonn, Sylvia Glaßer, Anna Vilanova, Bernhard Preim, and To-
bias Isenberg. Occlusion-free Blood Flow Animation with Wall
Thickness Visualization. IEEE Transactions on Visualization and
Computer Graphics, 22(1):728–737, 2016. ISSN 10772626. doi:
10.1109/TVCG.2015.2467961.

Fritz Lekschas, Benjamin Bach, Peter Kerpedjiev, Nils Gehlenborg,
and Hanspeter Pfister. HiPiler: Visual Exploration of Large
Genome Interaction Matrices with Interactive Small Multiples,
2017. ISSN 10772626.

Guoliang Li, Liuyang Cai, Huidan Chang, Ping Hong, Qiangwei
Zhou, Ekaterina V Kulakova, Nikolay A Kolchanov, and Yijun
Ruan. Chromatin interaction analysis with paired-end tag (ChIA-
PET) sequencing technology and application. BMC Genomics, 15
(Suppl 12):S11, dec 2014. ISSN 14712164. doi: 10.1186/1471-2164-
15-S12-S11.

Wentian Li and Jan Freudenberg. Mappability and read length.
Frontiers in Genetics, 5(NOV):1–1, 2014. ISSN 16648021. doi:
10.3389/fgene.2014.00381.

Jon Lyons, Matt Zeller, and Brandon Bray. Holographic Remoting
mixed-reality/holographic-remoting-player.


1.1 Left: 2D projection of our 3D curve model from the completed system. Right: The same chromatin structure represented as a heat map. (12)

1.2 Cave survey data from Schuchardt et al. 2007 (13)

2.1 DNA strands wind around histones. Each of these units is called a nucleosome, and are connected by linker DNA. (17)

2.2 The common process of chromosome conformation capture. After reverse cross-linking, additional steps are applied which vary depending on method. For Hi-C, a 5-step workflow is employed to obtain interaction data for the entire genome. (18)

2.3 This specific heat map pattern corresponds to a simple ‘double loop’ structure, in which the curve loops back on itself two times. It is immediately apparent that the heat map representation does not lend itself well to appreciating and identifying such structures. (20)

2.4 Juicebox Hi-C heat map from Durand et al. 2016. This is an example of the standard representation of processed Hi-C data (20)

2.5 Sutherland’s augmented reality headset tracked head position using a mechanical arm attached to the ceiling of the room. (23)

3.1 Human chromosome 16 heat map in the Unity application described in Section 3.1.1. (30)

3.2 In order to fully understand the challenges presented by each combination of model and interface, it is critical to distinguish between the different levels at which dimensionality is at play. At a conceptual level, the heat map representation operates on two dimensions. It is also rendered at this same level of dimensionality - it is not a projection onto a lower dimension as with the screen representation of the curve, for example. Finally, the ‘pan’ and ‘zoom’ operations available also operate at the level of two dimensions. (30)
3.3 The screen representation projects a three-dimensional curve onto the two-dimensional screen space. The pan functionality on the mouse translates two-dimensional mouse movements into three-dimensional camera rotation. (31)

3.4 Graphical depiction of the method described in Section 3.2.3. First, the raw data values are sorted. Then, the difference between each pair of values is calculated. A standard colour scale is generated, and points are sampled from it evenly. These points are then mapped to the new colour line at a spacing corresponding to the gradient values calculated in step 2. As can be seen in the ‘before’ and ‘after’ figures, this approach ensures that the difference between any two distinct values is clear, while still representing the numerical distance using the intensity of the colour. (32)

3.5 The virtual reality condition operates entirely in three-dimensional space. Using stereoscopy, the user views a three-dimensional curve, and is able to navigate around it in three-dimensional space. This is in contrast to the on-screen variation of the same visualisation, in which the curve exists in three conceptual dimensions, but is projected onto the two dimensions of the screen, and using a mouse is interacted with in fewer than three dimensions. (33)

3.6 Example trials from our study: Distance, Volume, Segments (36)

3.7 The colour scale used for the study. Taken from ColorBrewer. (37)