UiT
THE ARCTIC
UNIVERSITY
OF NORWAY

Faculty of Science and Technology
Department of Computer Science

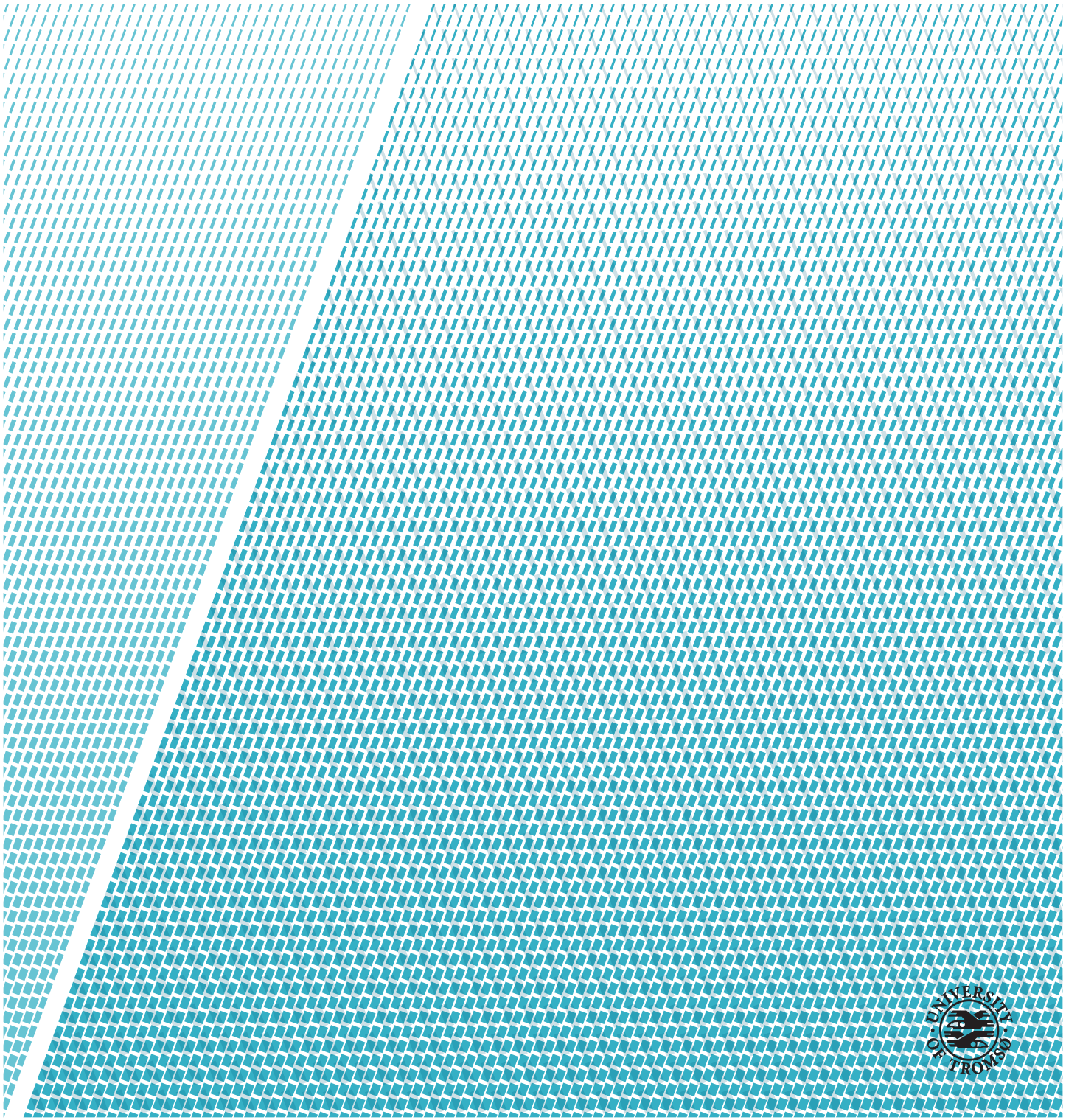
Virtual Reality for the visualization of high-dimensional relationships in bioinformatics

Subtitle

—

Álvaro Martínez Fernández

INF-3990 Master's thesis in Computer Science November 2020



Contents

List of Figures **iii**

1 Introduction **1**
1.1 Background 1
1.2 Challenges and research problem 3
1.3 Proposed solution 3
1.4 Significance and Contribution 3
1.5 Outline 4

2 Bioinformatics in VR **5**
2.1 VR 5
2.1.1 Software and frameworks for VR development 5
2.1.2 Locomotion and ergonomics 6
2.1.3 Clustering analysis 6

3 Related work **7**
3.1 BioVR 7
3.2 CellexaVR 7
3.3 BigTop 7

4 MIxT VR **9**

5 Evaluation and discussion **11**

6 Conclusion and future work **13**

List of Figures

1.1 Visualization for network biology. a A simple drawing of an undirected unweighted graph. b A 2D representation of a yeast protein-protein interaction network visualized in Cytoscape (left) and potential protein complexes 3D identified by the MCL algorithm from that network (right). c A 3D view of a protein-protein interaction network visualized by BiolayoutExpress. d A multilayered network integrating different types of data visualized by Arena3D. e A hive plot view of a network in which nodes are mapped to and positioned on radially distributed linear axes. f Visualization of network changes over time. g Part of lung cancer pathway visualized by iPath. i Remote navigation and control of networks by hand gestures. h Integration and control of 3D networks using VR devices. Figure adapted[8]. 2

1.2 Network view of the MIxT application where nodes repsent genes and the modules are repsented by colors. Relationships are represented by grey lines that connect a gene with another one. 4

1 Introduction

1.1 Background

Technological advancement has revolutionized the field of genomics, which has led to a cost-effective generation of big amounts of sequence data. The sequencing of the first human genome (2002) took around 13 years and cost over $3 million to complete. Nowadays we can resequence a human genome for $1000 and can generate more than 320 genomes per week[6]. This technological innovation leads to the accumulation of vast quantities of genomic data, posing a tremendous challenge to scientists for effective mining of data to explain a phenomenon of interest[12]. New ways of analysing the produced data have been therefore necessary in order to discover interesting patterns and make the most out of it. No matter how much resources we use into extracting the data if we don't get anything interesting out of it.

Some of the main problems that researchers face when analysing genomic data are information overload, data interconnectivity and high dimensionality. Visualization is one way of facing this problems. For this reason it is very important to implement efficient visualization technologies that can lead to find new patterns and the extraction of good conclusions of the data.

In the field of system biology there are usually network representations where the nodes or bioentities are connected to each other, where these edges represent associations. Because of the improvements in technology, these networks can increase dramatically in size and complexity. We need therefore better

visualization systems and more efficient alogirithms for the analysis of the data.

In 1.1 we can see a representation of the evolution for visualization of networks in system biology. From simple graphs in 2 dimensions, to 3D representations and nowadays also visualizations in virtual reality where we can interact directly with the data itself.

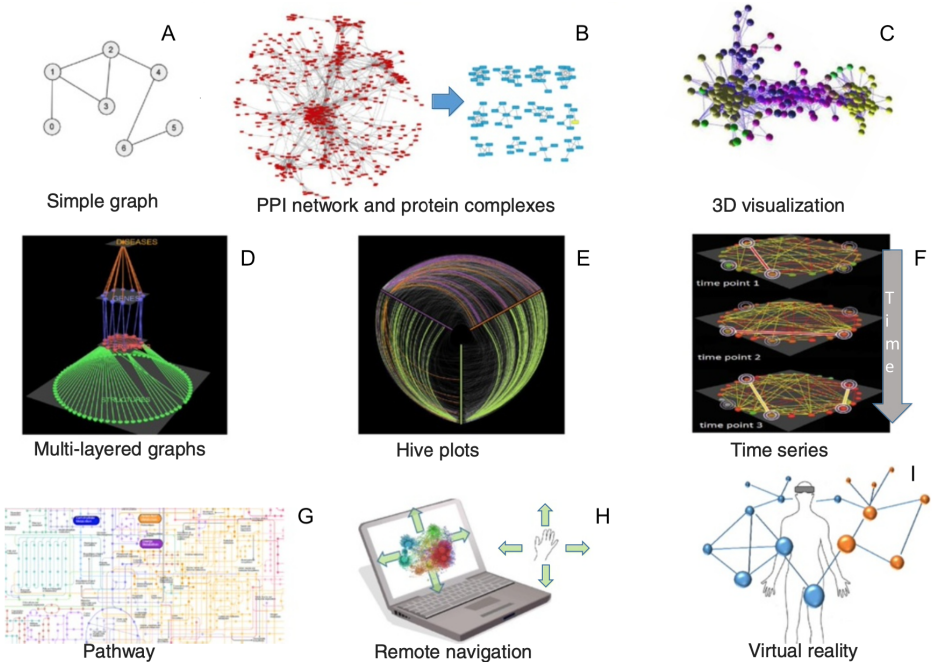


Figure 1.1: Visualization for network biology. a A simple drawing of an undirected unweighted graph. b A 2D representation of a yeast protein-protein interaction network visualized in Cytoscape (left) and potential protein complexes 3D identified by the MCL algorithm from that network (right). c A 3D view of a protein-protein interaction network visualized by BiolayoutExpress. d A multilayered network integrating different types of data visualized by Arena3D. e A hive plot view of a network in which nodes are mapped to and positioned on radially distributed linear axes. f Visualization of network changes over time. g Part of lung cancer pathway visualized by iPath. i Remote navigation and control of networks by hand gestures. h Integration and control of 3D networks using VR devices. Figure adapted[8].

Virtual reality (VR) is still a field under exploration and that can be of great help in network analysis. VR can be very powerful because it takes advantage of the way the human being perceives and analysis things. We as human beings have a great ability to discover patterns, however we are biologically optimized to see the world and the patterns in 3 dimensions. VR is one of the best ways then for better discovery in spatial dimensions. It has been demonstrated that

VR help scientists work more effectively in fields like medicine [5][10][2], biology[11][9] and neuroscience[1][7], to cite some examples.

1.2 Challenges and research problem

This project focus mainly on solving the problem of visualization of high dimenasional data from the MIxT project by using virtual reality. Furthermore the application allows the user to interact with the network created from the data in the virtual environment. It also allows the user compare the blood and biopsy networks at the same time in order to finde relationship, which wasn't possible in the MIxT web application as this only allows the user to visualize one network at a time.

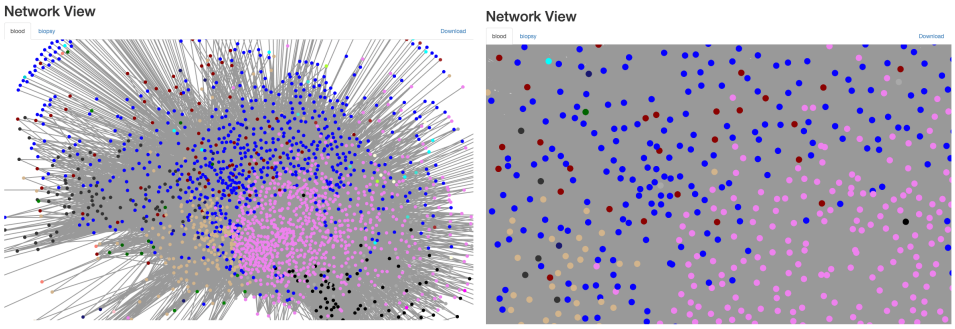
MIxT[4] is a web application for bioinformaticians. Among other tools, it offers a network visualization of genes which are represented as nodes in the network and where the edges represent statistically significant correlation in expression between two nodes. This tool was used in a study[3] that identifies genes and pathways in the primary tumor that are tightly linked to genes and pathways in the systemic response of a patient with breast cancer. When exploring a network in MIxT, it can be hard to understand the data and its relationships because there is too much data. This problem is easy to occur when there are too many node and edges. In figure 1.2 we can see an example of the network visualization from MIxT. As we can see in Figure 1.2a, there are many nodes and relationships among them and when we zoom in in the network, it becomes very difficult to understand the data and the relationships as shown in in Figure 1.2b.

The network is also in 2-dimensions and what we propose in this project is to use a virtual reality 3d visualization in order to cope better with this problem.

1.3 Proposed solution

1.4 Significance and Contribution

This project contributes in the exploration of the possibilities that Virtual Reality offers for visualization of big data in bioinformatics.



(a) Network with several modules. **(b)** Zoom in the network.

Figure 1.2: Network view of the MIxT application where nodes repsent genes and the modules are repsented by colors. Relationships are represented by grey lines that connect a gene with another one.

1.5 Outline

2

Bioinformatics in VR

2.1 VR

Todo: write about VR technologies (Oculus htc vive).

2.1.1 Software and frameworks for VR development

Unity3D[1] and Unreal Engine[2] are two popular programs for development of videogames and also virtual reality games and applications. They offer integrations for Oculus Quest and other VR devices in the market. In addition, Oculus Quest offers a devlopment mode that can be activated once the glasses are connected to the PC. In this way the VR application can be tested directly on the VR device.

Virtual Reality development can also be done for the browser. WebVR[3] is an open specification that makes it possible to experience VR in the browser, no matter what VR device is used. We can find many web frameworks to build VR applications for the web that are based on WebVR. Some of these frameworks

1. https://unity.com
2. https://www.unrealengine.com
3. https://webvr.info

are A-frame[4], React360[5] and three.js[6].

2.1.2 Locomotion and ergonomics

- Physical movement
- Script movement
- Avatar movement
- Steering motion
- World pulling
- Teleports

2.1.3 Clustering analysis

Cluster analysis is used to classify objects or cases into relative groups called clusters. Unlike supervised machine learning techniques, in cluster analysis, there is no prior information about the group or cluster membership for any of the objects. We can find many clustering approaches, two of the most comonly used ones are k-means and DBSCAN.

The k-means algorithm starts by choosing k random centers which can be manually set. Then the data points are assigned to the closest center based on their Euclidean distance.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is another algorithm that is based on the density of the data points. The algorithm identifies clusters and expands them by scanning neighborhoods. If it cannot find any points to add, it simply moves on to a new point hoping it will find a new cluster.

4. https://aframe.io
5. https://facebook.github.io/react-360
6. https://threejs.org

3

Related work

3.1 BioVR

3.2 CellexaVR

3.3 BigTop

4

MIxT VR

5

Evaluation and discussion

/6

Conclusion and future work

Bibliography

[1] Corey J. Bohil, Bradly Alicea, and Frank A. Biocca. "Virtual reality in neuroscience research and therapy." In: *Nature Reviews Neuroscience* 12.12 (2011), 752–762. DOI: 10.1038/nrn3122.

[2] "Commentary on Rose, F.D., Brooks, B.M., & Rizzo, A.A., Virtual Reality in Brain Damage Rehabilitation: Review." In: *CyberPsychology & Behavior* 8.3 (2005), 263–271. DOI: 10.1089/cpb.2005.8.263.

[3] Vanessa Dumeaux et al. "Interactions between the tumor and the blood systemic response of breast cancer patients." In: *PLOS Computational Biology* 13.9 (2017). DOI: 10.1371/journal.pcbi.1005680.

[4] Bjørn Fjukstad et al. "Building Applications for Interactive Data Exploration in Systems Biology." In: *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics - ACM-BCB 17* (2017). DOI: 10.1145/3107411.3107481.

[5] KE. Laver et al. "Virtual reality for stroke rehabilitation." In: *Cochrane Database of Systematic Reviews* 11 (2017). ISSN: 1465-1858. DOI: 10.1002/14651858.CD008349.pub4. URL: https://doi.org//10.1002/14651858.CD008349.pub4.

[6] Mike May. "LIFE SCIENCE TECHNOLOGIES: Big biological impacts from big data." In: *Science* 344.6189 (2014), pp. 1298–1300. ISSN: 0036-8075. DOI: 10.1126/science.344.6189.1298. eprint: https://science.sciencemag.org/content. URL: https://science.sciencemag.org/content/344/6189/1298.

[7] Matthias Minderer et al. "Virtual reality explored." In: *Nature* 533.7603 (2016), 324–325. DOI: 10.1038/nature17899.

[8] Georgios A. Pavlopoulos et al. "Visualizing genome and systems biology: technologies, tools, implementation techniques and trends, past, present and future." In: *GigaScience* 4.1 (2015). DOI: 10.1186/s13742-015-0077-2.

[9] David A. Thorley-Lawson, Karen A. Duca, and Michael Shapiro. "Epstein-Barr virus: a paradigm for persistent infection – for real and in virtual reality." In: *Trends in Immunology* 29.4 (2008), 195–201. DOI: 10.1016/j.it.2008.01.006.

[10] James Xia et al. "Three-dimensional virtual-reality surgical planning and soft-tissue prediction for orthognathic surgery." In: *IEEE Transactions on*

Information Technology in Biomedicine 5.2 (2001), 97–107. DOI: 10.1109/4233.924800.

[11] Yuting Yang et al. "Integration of metabolic networks and gene expression in virtual reality." In: *Bioinformatics* 21.18 (July 2005), pp. 3645–3650. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bti581. eprint: http://oup.prod.sis.lan/bioinformatics/article-pdf/21/18/3645/520673/bti581.pdf. URL: https://doi.org/10.1093/bioinformatics/bti581.

[12] Jimmy F. Zhang et al. "BioVR: a platform for virtual reality assisted biological data integration and visualization." In: *BMC Bioinformatics* 20.1 (2019). DOI: 10.1186/s12859-019-2666-z.