An Interactive Tool for Visualization of Complex Genetic Interactions

Amber Rosado
A/Prof Michael Charleston
School of Information Technologies
FACULTY OF ENGINEERING & INFORMATION TECHNOLOGIES

Introduction

- DNA sequencing is faster and easier than ever, and much research involves analysis of relationships between different parts of genetic sequences.
- These relationships may be of a variety of types depending on the goal of the research, for instance, finding identical or very similar blocks of genes within an entire genome to attempt to investigate evolutionary relationships between species.
- Visualization has been important to the interpretation of the results from the beginning, due to the scale of the data involved.[1]
- Because of the extremely large size of the data sets involved, visualization of the data in an interactive way is computationally challenging.

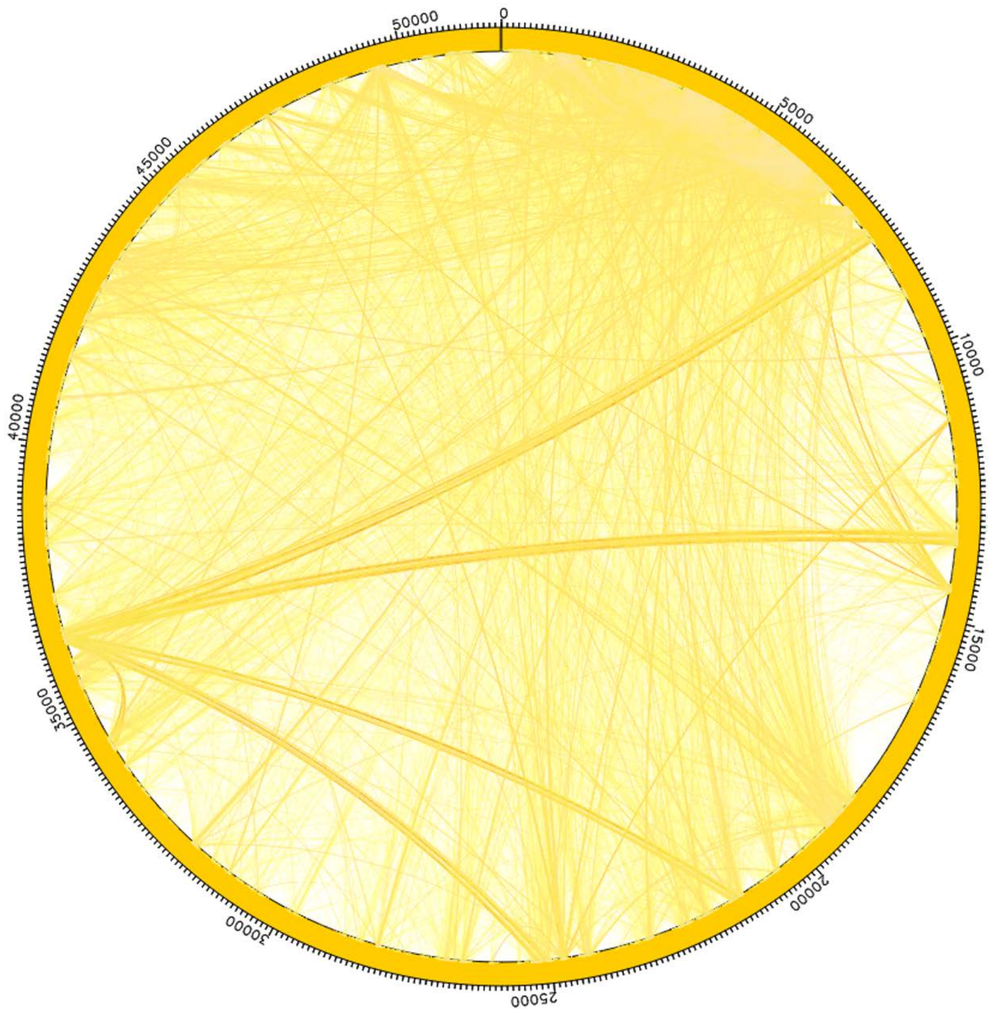

Figure 1: Interaction data for human BIM gene, showing 1 673 921 links displayed in bundles. No filtering has been done for this image, and maximum link opacity is very low.

Table 1: Trends over time in development of visual presentation and interactive capability of genomic comparison visualization applications.

Name	Year	Dot Plot	Linear	Circular	3-D	Interactive
MATCH[1]	1970	✓				
Dotter[2]	1995	✓				✓
PipMaker[3]	2000	✓				
REPvis[4]	2001		✓			✓
Cinteny[5]	2007		✓			✓
Circos[6]	2007			✓		
MizBee[7]	2009		✓	✓		✓
Gremlin[8]	2010		✓			✓
CMap3D[9]	2010		✓		✓	✓
New Tool	2013			✓		✓

Existing Visualization Tools

- Many tools have been developed to visualize genetic sequence and comparison data (see Table 1).
- Different methods of visually representing the data have developed over time. Initially, two sequences being analyzed were visualized as the two axes of a scatter plot, with dots indicating related regions.
- The type of visualization in widest use today presents sequences as coloured blocks on the circumference of a circle, with links across the circle indicating related regions.
- Circos[6] is the current standard for this type of visualization, but lacks interactivity and *cannot produce* visualizations of data beyond a modest size.
- MizBee[7] and Gremlin[8] have been developed more recently and include various interactive features, but Gremlin displays data in a linear arrangement (and also has yet to be released since [8] was published in 2010) and MizBee lacks Circos' flexibility in terms of input data type.

Development Process

- Based on my survey of existing software and related literature a set of basic requirements was identified for the first implementation of a new interactive visualization tool.

Basic Requirements

- Colour of links is determined by their weight (Fig. 1).
- Users can limit view of links to those with weights within a specified range (Fig. 2).
- Users can limit displayed links to those within specified sections of the genome (Fig. 3).

Additional Features

- After the first requirements were implemented, in consultation with Andrian Yang and Leslie Burnett regarding their use of this tool to aid in their research, the ability to set the maximum opacity of the links (compare Figs. 1 & 2) and several labelling features were added to the program.

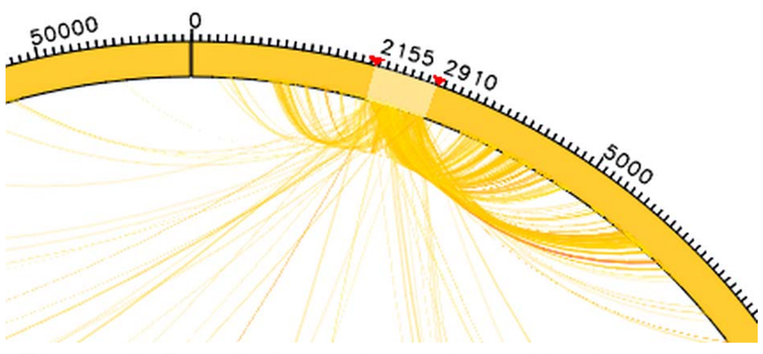

Figure 3: Detail of a region selected by the user in the data in Fig. 1 by clicking and dragging on the edge of the circle. Only links with at least one end between the 2155th and 2910th base pairs are displayed.

Implementation Details

Link Bundling

- To conserve memory and computation during rendering, links are stored in *bundles*.
- Links are stored in the same bundle if they meet certain criteria, the main being that if they were rendered separately, they would be overlapping at both ends, and the difference between their weights is below a certain threshold.

Data Structure

- Bundles are stored in a three-dimensional *quadtree*.
- *Bundles* are rendered rather than individual links.
- Utilizing this type of structure allows range queries to be performed on the links for filtering *O(d log n)* time[10] where *d* is the number of links in the query range.

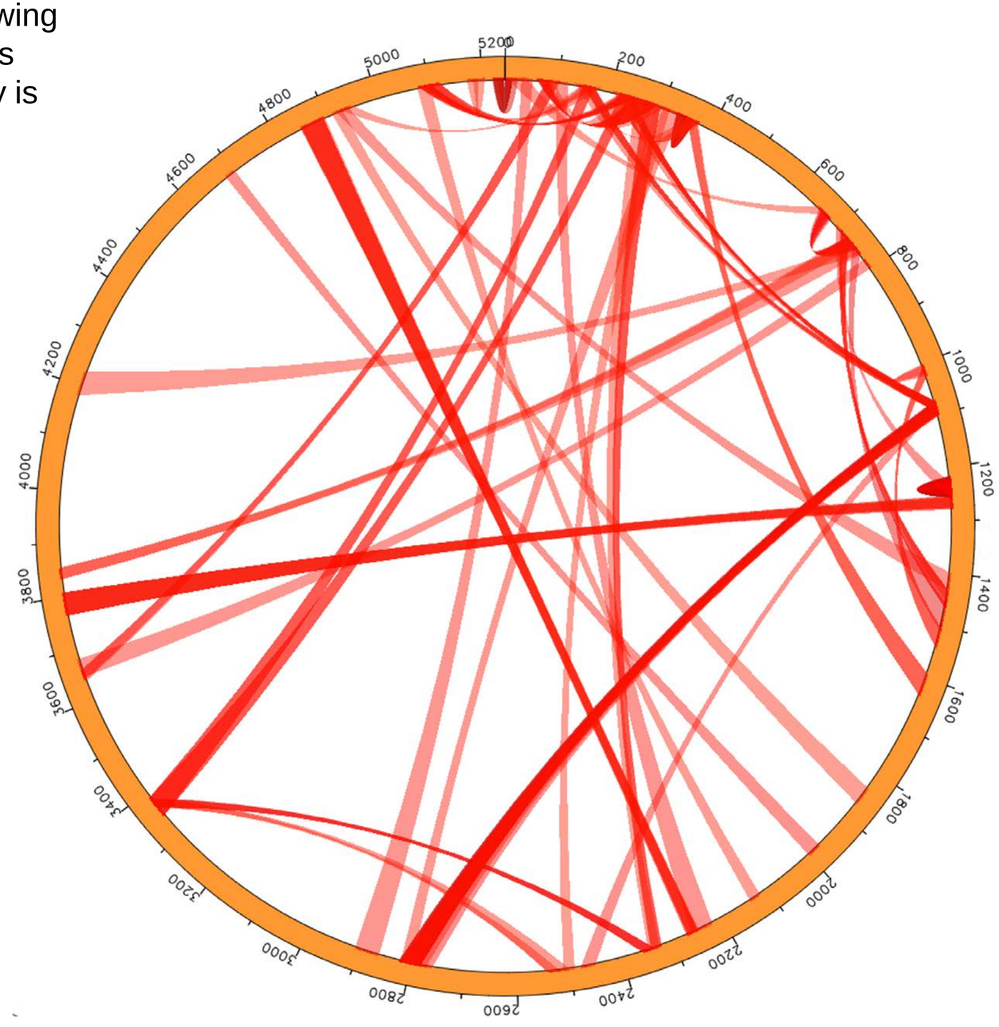

Figure 2: Interaction data for the SV40 virus, with links below a set threshold weight filtered out of the view. SV40 is much smaller than BIM so links are drawn wider in this image. The data contained 1 352 371 links, despite the shorter genome length.

Discussion and Future Work

Evaluating This Tool

- This stage of development has focused on creating a program which can be used interactively with data at a very large scale.
- Initially loading and processing the data in Fig. 1 took 17.82 seconds on our machine. Fig. 2 took 14.31 seconds.
- Direct comparison of running time against Circos is challenging because Circos only produces static images. In any case, Circos was not able to render the data seen in Fig. 1 at all.

Further Development

- Changes to how and when data is rendered should address the issues with interactivity because of the size of the data.
- The development process has been iterative, and should continue with more focus on refining the tool and extending its features and range of uses.

References

[1] Gibbs & McIntyre, *European Journal of Biochemistry,* 16:1-11,1970
[2] Sonnhammer & Durbin, *Gene,* 167:1-10, 1996
[3] Schwarts et. al., *Genome Research,* 10:577-586, 2000
[4] Kurtz et. al., *Nucleic Acids Research,* 29(22):4633-4642, 2001
[5] Sinha & Meller, *BMC Bioinformatics,* 8:82, 2007
[6] Krzywinski et. al., *Genome Research,* 19:1639-1645, 2009
[7] Meyer et. al., *IEEE Transactions on Visualization and Computer Graphics*, 15(6):897-904, 2009
[8] O'Brien et. al., *IEEE Transactions on Visualization and Computer Graphics,* 16(6):918-926, 2010
[9] Duran et. al., *Bioinformatics*, 26(2):273-274, 2010
[10] Finkel & Bentley, *Acta Informatics*, 4:1-9, 1974.