Analysis Report on MizBee: A Multiscale Synteny Browser

The research paper focuses only analyzing about evolution and genomic function by comparing the genomes of species to find regions of shared sequences. The aim is to achieve aesthetic visualizations to display multiple types of relationships at multiple scales between the species. The goal is achieved using MizBee, a multiscale synteny browser with the unique property of providing interactive side-by-side views of the data across the range of scales supporting exploration of all these relationship types. The structure of the analysis report will be based on the What, Why and How framework.

The data used for the analysis is gathered by conducting a series of interviews with two target users. The dataset structure contains the data in the form of blocks which is divided into three main layers of scale.

- The first or the highest level is the genome which comprises of the list of chromosomes.
- The second level is chromosome containing a list of blocks whose locations are definite in terms of the chromosome sequence coordinate system.
- The last block level comprises of list of conserved features. The features in this level are chromosome id, coordinate along the sequence, length, orientation, tag, match on another chromosome, and similarity score.

The ordered attribute is position in chromosome sequence coordinates, categorical attributes are orientation and matching feature, and quantitative attributes are length and similarity score.

The 'Why' framework in the research paper intends decent usage of visualization tools. It analyzes similarity relationships between the genomic features. The high-level biology questions regarding evolution and function of genomic sequence are stated. The research also consists of some low-level data-centric questions which are based on algorithm improvement. For an instance, which chromosomes share conserved blocks?

One of the main purpose of 'Why' framework is the task abstraction. The research paper mainly focuses on block reliability, high-level science and algorithms implementation. The questions raised in the research are analyzed using the four relationship types which are proximity, size, orientation and similarity, and four data scales which are genome, chromosome, block and feature. According to me, the research paper raises some of the bright enquiries such as Do neighboring blocks go to the same: chromosome? relative location on a chromosome? and Are similarity scores alike: with respect to neighboring blocks? within a block? These types of specific questions significantly help in deep analysis and clears the user intentions to use visualization tools.

The 'How' part of the framework helps to create reasons about similarities and differences between different visualization. The important contribution of the research is a classification of the design space that can be used to generate effective visual encodings. It represents conserved features as a segment on a chromosome with a matching segment on another chromosome. Another crucial contribution of the research is the design of MizBee which allows to explore many kinds of conserved synteny relationships with linked views at the genome, chromosome, and block levels. One of the idiom design choice is made by encoding match relationships between the segments of chromosome with features like color and connection marks. The next one is Genome view which provides a high-level overview of the many-to-many relationships between chromosomes. The other one is the Block view which provides details about the conservation relationships of features within the selected block.

Due to the large size of the dataset, validation becomes very crucial part of the design. The validation is performed by the authors at Domain level, Idiom level, Qualitative result image analysis and the Data Abstraction level. The authors make sure that the data is gathered after a series of interviews with the biologists which validates the Domain level. They implemented block computation algorithm for a stickleback-pufferfish dataset which used Mizbee and the algorithm results were optimal for data cleaning. Also, the authors identified important questions that biologists ask which justifies at data abstraction level.

The other validation that would benefit the proposed visualization system is the innermost algorithm level. It consists of the memory and the system time for the computation. The interaction idioms are ineffective if the computation takes more time than expected i.e. instead the data should be precomputed for fast response.