Large-scale Dynamic Visualization of Multiple Comparative Genomic Data

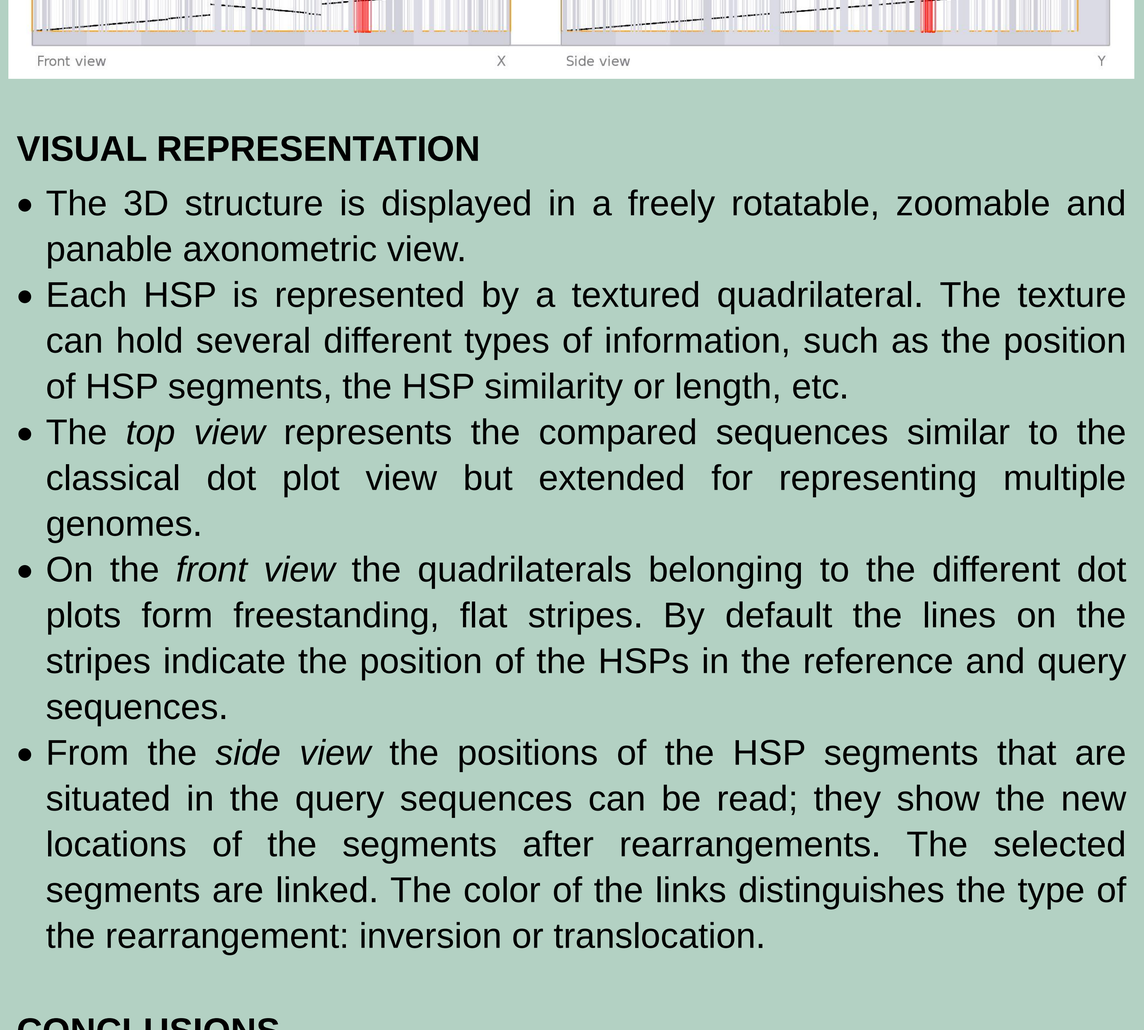
Balázs Tukora, Christoph Anthes, Paul Heinzlreiter, Dieter Kranzlmüller

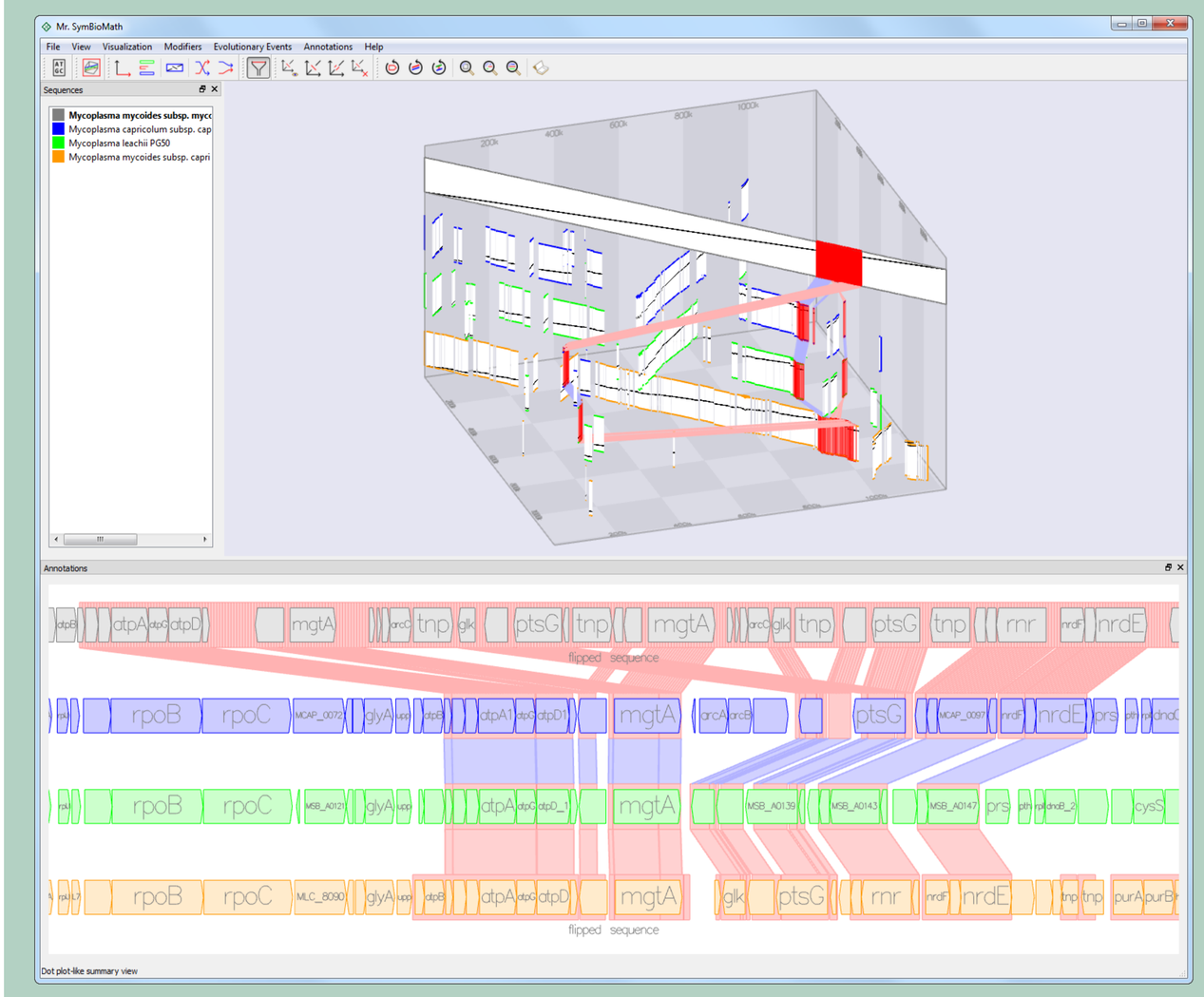
ABSTRACT

A novel visualization technique facilitates the comparative analysis of large-scale genome data. It reveals the relationships among multiple compared genome sequences by combining three traditional and well-established representations of comparative genomic data: the dot plot, the gradient view and the linear representation of inter-genome connections. By integrating them into a single three dimensional structure and offering dynamic exploration, a significantly more powerful visualization is provided. Being able to rotate the axonometric representation of the three-dimensional data freely enables the user to easily link semantics of different representations with each other and therefore getting a deeper understanding of the sequences and their relations.

DATA REPRESENTATION

- High-scoring segment pairs (HSP) – the fundamental output unit of BLAST-like algorithms – are arranged in a 3D structure for displaying.
- The position of the HSPs is defined in a right-handed orthogonal coordinate system.
- The position of an HSP along the *x*-axis corresponds to the position of the contained segment that is situated in the reference sequence.
- The position of an HSP along the *y*-axis corresponds to the position of the contained segment that is situated in the query sequence.
- The HSPs that belong to the same query sequence are placed at the same height, while different heights are assigned to the other comparisons.





VISUAL REPRESENTATION

- The 3D structure is displayed in a freely rotatable, zoomable and panable axonometric view.
- Each HSP is represented by a textured quadrilateral. The texture can hold several different types of information, such as the position of HSP segments, the HSP similarity or length, etc.
- The *top view* represents the compared sequences similar to the classical dot plot view but extended for representing multiple genomes.
- On the *front view* the quadrilaterals belonging to the different dot plots form freestanding, flat stripes. By default the lines on the stripes indicate the position of the HSPs in the reference and query sequences.
- From the *side view* the positions of the HSP segments that are situated in the query sequences can be read; they show the new locations of the segments after rearrangements. The selected segments are linked. The color of the links distinguishes the type of the rearrangement: inversion or translocation.

INTERACTION

- Rotation from view to view in three degrees of freedom is seamlessly controllable by the user. The displayed visualization is freely panable and zoomable as well.
- To reduce visual cutter interactive filtering functions can be easily inserted and combined in the program code and controlled with the GUI.
- Single or multiple HSPs can be selected from any points of view and at any zoom level by capturing them in a single selection rectangle or subsequent selection rectangles on the screen.
- By extracting the set of selected HSPs and the linkage that connects them we can sort and display detailed information about the compared regions.
- Additional annotation files can be loaded and displayed that hold information about the genes that appear in the selected regions.

CONCLUSIONS

The merit of the introduced 3D structure is that its projections from several points of view give three traditional representations of the comparative genomic data: the dot plot, the gradient view and the linear representation of inter-genome connections. Therefore the technique reveals the connections in the data represented by the three visualization methods, thus supporting the discovery of inter-genome relationships much better than using the three different views separately.

The project *High Performance, Cloud and Symbolic Computing in Big-Data Problems applied to Mathematical Modeling of Comparative Genomics* (Mr.SymBioMath) running from February 2013 to January 2017 is being funded from the European Union's Seventh Framework for research, technological development, and demonstration as an Industry-Academia Partnerships and Pathways (IAPP) project under grant agreement number 324554.

http://www.mrsymbiomath.eu