Large-scale Dynamic Visualization of Multiple Comparative Genomic Data

Balázs Tukora*
Leibniz Supercomputing Centre
University of Pécs

Christoph Anthes†
Leibniz Supercomputing Centre
Ludwig-Maximilians-Universität München (LMU)

Paul Heinzlreiter‡
Leibniz Supercomputing Centre
RISC Software GmbH

Dieter Kranzlmüller§
Ludwig-Maximilians-Universität München (LMU)
Leibniz Supercomputing Centre

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.

Keywords: Comparative genomics, Bioinformatics, Information visualization.

1 INTRODUCTION

One application area for visualization tools within bioinformatics is displaying the results of large-scale comparison of genomic sequences to answer questions about evolution and sequence functionality.

A novel visualization approach is presented producing a spatial arrangement of the results of large-scale genome comparisons in a 3D structure. This data structure can be perceived from different views implemented as 2D projections.

The orthogonal views of this structure produce views similar to traditional visualization techniques like dot plots [2], gradient views [4] and linear representations of the inter-genome connections [6] respectively. This arrangement greatly helps to understand the interconnections between the visual information represented by the different techniques, by clearly illustrating that they are different representations of the same underlying data. This technique of combining three different views into one 3D structure lets the user observe, select, highlight and manipulate the whole or subsets of the data dynamically, from any point of view.

2 VISUALIZATION TECHNIQUE

2.1 Spatial Data Representation

High-scoring segment pairs (HSP) [3] – the fundamental output unit of BLAST-like [1] algorithms – are arranged in a 3D structure for display. We define the position of the HSPs in a right-handed orthogonal coordinate system. The *x*-*y* plane is considered as horizontal, the vertical positions of the elements are represented along the *z*-axis. The spatial position of the HSPs is determined according to the following criteria: The DNA coordinates of the reference genome are mapped onto the *x*-axis, thus the position of an HSP along this axis corresponds to the position in the reference sequence. The DNA coordinates of the query genome are mapped onto the *y*-axis, thus the position of an HSP along this axis corresponds to the position in the query sequence. The HSPs resulting from distinct pairwise comparisons are grouped along the *z*-axis: The HSPs that belong to the same query sequence are placed at the same height, while different heights are assigned to the other comparisons.

2.2 Visual Representation

In the 3D structure each HSP is represented by a textured quad which is parallel to the *z*-axis. The texture of the quads can hold several different types of information. By default the DNA positions of the fragment that belongs to the HSP and is situated in the query sequence are indicated on it by a straight line crossing the quad: The vertical positions of its tips correspond the DNA position of the beginning and the end of the fragment in the query sequence, while they are normalized to the length of the query sequence. The position of the quads along the *x* and *y*-axes corresponds the DNA coordinates of the related HSP as defined in the previous section.

When a group of HSPs are selected, whose coordinates along the *x*-axis are located in the same area but they belong to different query sequences, they become linked by colored quads: The blue color indicates that the two fragments in the compared sequences that are referred to by two linked HSPs are on the same strand, while the red color indicates different strands – this points to a reversion.

The structure is displayed in a freely rotatable, zoomable and panable axonometric view as shown in Fig. 1, top right. The reason of applying parallel projection instead of perspective is that the orthographic views that are constructed by the rotation of the axonometric view preserve the scale of the represented data independently from their depth along the foreshortened axis. Thus the values presented by positions along the visible coordinate axes on the screen can be compared directly.

The orthographic projection onto the *x*-*y* plane is referred to as *top view*. It represents the compared sequences similar to the classical dot plot view but extended for representing multiple genomes. The comparisons are plotted as graphs by mapping multiple query sequences onto the *y*-axis and one reference sequence onto the *x*-axis thereby producing one graph of a unique color for each query sequence as shown in Fig. 1 (top left).

The orthographic projection onto the *x*-*z* plane is referred to as *front view*. Due to the projection the quads belonging to the different dot plots form freestanding, flat stripes as shown in Fig. 1 (bottom left). On the stripes rising and sloping lines can be seen: these lines are formed by a number of lines that are drawn on the quad textures. Their vertical positions indicate the DNA positions of the related HSP in the query sequence.

The orthographic projection onto the *y*-*z* plane is referred to as *side view*. In Figure 1 it is shown at lower right, while the axono-

*e-mail: balazs.tukora@lrz.de
†e-mail: christoph.anthes@lrz.de
‡e-mail: paul.heinzlreiter@risc-software.at
§e-mail: kranzlmueller@ifi.lmu.de

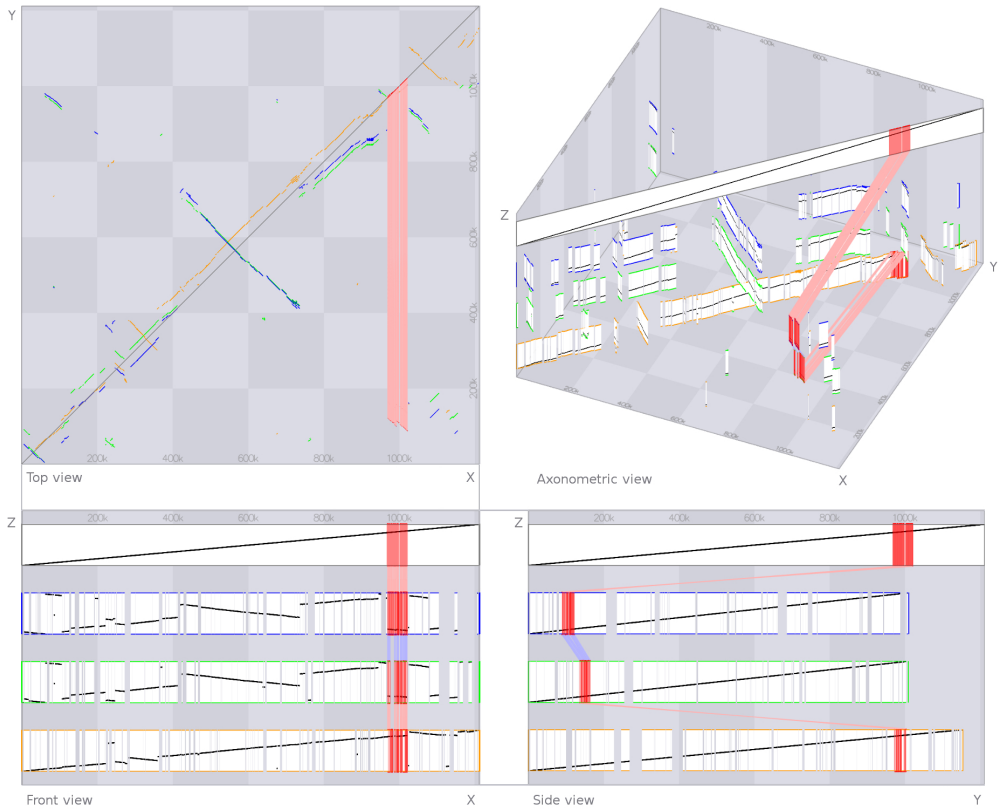

Figure 1: Main views with selected fragments. The top view (top left) provides a similar impression than multiple dot plots aligned in a single coordinate system. The axonometric view (top right) gives a three-dimensional overview on the whole set of compared genomes. The front view (bottom left) displaying the data structure by a projection on the *x-z* plane, shows multiple comparisons with the reference genome. The side view (bottom right) represents the rearrangements.

metric view (top right) shows the underlying connection between the front and side views. A group of HSPs, whose DNA positions in the reference sequence are located in a given area, are selected in all of the sequences and are thus linked. The front view shows only vertically linked quads. As the quads have different coordinates along the *y*-axis – according to their DNA coordinates in the query sequences – and the horizontal axis of the side view is the *y*-axis, the positions of the selected fragments in the query sequences can be read from the side view. The color of the links distinguishes the type of the rearrangements: reversions or translocations, while the positions of the fragments in the query sequences show their new locations after the rearrangement. This type of visualization of the comparative genomic data shows the traditional linear representation of inter-genome connections.

2.3 Interaction

According to [5] we implemented the features overview, zoom, filter, details-on-demand, and relate. These operations work on the complete data structure independent of the selected view.

Rotation in three degrees of freedom is seamlessly controllable by the user, while the displayed visualization is freely panable and zoomable as well. Though the data is pre-filtered already within the computational process producing it [3], additional filtering can be applied to reduce visual clutter.

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. By extracting the set of selected HSPs and the linkage that connects them we can sort and display short-scale 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.

This work was supported in part by a grant from the project *High Performance, Cloud and Symbolic Computing in Big-Data Problems applied to Mathematical Modeling of Comparative Genomics* (Mr.SymBioMath), being funded by the European Union's Seventh Framework Programme for research, technological development and demonstration as an Industry-Academia Partnerships and Pathways (IAPP) project under grant agreement number 324554.

References

[1] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

[2] A. J. Gibbs and G. A. Mcintyre. The diagram, a method for comparing sequences. *European Journal of Biochemistry*, 16(1):1–11, 1970.

[3] A. R. Moreno, Óscar Torreño Tirado, and O. Trelles. Out of core computation of hsps for large biological sequences. *Advances in Computational Intelligence*, 7903:189–199, 2013.

[4] D. R. Riley, S. V. Anqiuoli, J. crabtree, J. C. D. Hotopp, and H. Tettelin. Using sybil for interactive comparative genomics of microbes on the web. *Bioinformatics*, 28(2):160–166, 2012.

[5] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *IEEE Symposium on Visual Languages (VL '96)*, pages 336–343, 1996.

[6] J. Yang, J. Wang, Z.-J. Yao, Q. Jin, Y. Shen, and R. Chen. Genomecomp: a visualization tool for microbial genome comparison. *Journal of Microbiological Methods*, 54:423–426, 2003.