Bioinformatics, 31(12), 2015, 2054–2055
doi: 10.1093/bioinformatics/btv078
Advance Access Publication Date: 5 February 2015
Applications Note

Data and text mining

Whiteboard: a framework for the programmatic visualization of complex biological analyses

Görel Sundström[1,†], Neda Zamani[1,2,†], Manfred G. Grabherr[1] and Evan Mauceli[3,*]

[1]Department of Medical Biochemistry and Microbiology, Bioinformatics Infrastructure for Life Sciences, Uppsala University, 75123 Uppsala, Sweden, [2]Department of Plant Physiology, Umeå University, 901 87 Umeå, Sweden and [3]Parabase Genomics, 27 Drydock Avenue, Boston MA 02210, USA

*To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Jonathan Wren

Received on November 10, 2014; revised on January 14, 2015; accepted on February 1, 2015

Abstract

Summary: *Whiteboard* is a class library implemented in C++ that enables visualization to be tightly coupled with computation when analyzing large and complex datasets.

Availability and implementation: the C++ source code, coding samples and documentation are freely available under the Lesser General Public License from http://whiteboard-class.sourceforge.net/.

Contact: emauceli@parabasegenomics.com.

1 Introduction

A key aspect in understanding complex biological or medical systems lies in visualizing the data, which allows for recognizing patterns that may explain the mechanisms underlying the evolution of phenotypes, or susceptibility to, and progression of, diseases. Especially for large datasets, as routinely generated by high throughput sequencing technologies, the challenge of graphically presenting the data goes well beyond drawing simple plots, often encompassing both figures that show how variables change across entire genomes, as well as graphs that highlight specific sites at a nucleotide or amino acid resolution. Moreover, a meaningful interpretation often involves mathematical or statistical algorithms, where, ideally, analysis and visualization are tightly coupled. Unfortunately, it is common practice for these to be decoupled, with the 'heavy-lifting' computation being done in one language, followed by the visualization being done with a graphical package like R (http://CRAN.R-project.org) or MATLAB (http://www.mathworks.com), or via a scripting language like python (www.python.org). Although this approach can be effective, it becomes necessary to reduce size of the analytical data, not because its best for the visualization, but to accommodate the limits inherent in the choice of graphical package. Here, we describe *Whiteboard*, a class library implemented in C++, allowing for easily graphing any kind of data into any format.

2 Using the whiteboard

Whiteboard provides all basic tool sets for drawing simple objects, as well as color manipulation methods, an extendable set of compound graphs, and a set of utilities such as a command line parser, file parsers etc. for data reading and processing. Figure 1 shows four examples of visualizing different datasets by software that integrate *Whiteboard*. For visualizing a summary of protein alignments, we color-code alignments by taxa first, and use color gradients to indicate local sequence similarity across the query sequence (Fig. 1a). For genome-wide synteny data computed from the medaka and zebrafish genomes, we show syntenic regions in zebrafish color coded by medaka chromosomes (Fig. 1b), where e.g. medaka chromosome 16 and zebrafish chromosome 16 share synteny over the entire length, whereas synteny in other chromosomes is broken up, e.g. zebrafish chromosome 18 (Fig. 1b). For clustering gene expression, we use color-coding to indicate absolute expression on a scale from black to yellow (Fig. 1c), which can be computed for all genes in the dataset (Fig. 1c shows a small subset). We indicate differential expression against the mean values in one point in a time series by color ranging from dark blue (lower) to dark red (higher), exemplified by HOX genes in eight clusters on the zebrafish genome (Fig. 1d). An animated image, computed from individual *Whiteboard* images, in which we interpolated values between time points, can be found on *Whiteboard's* web site.

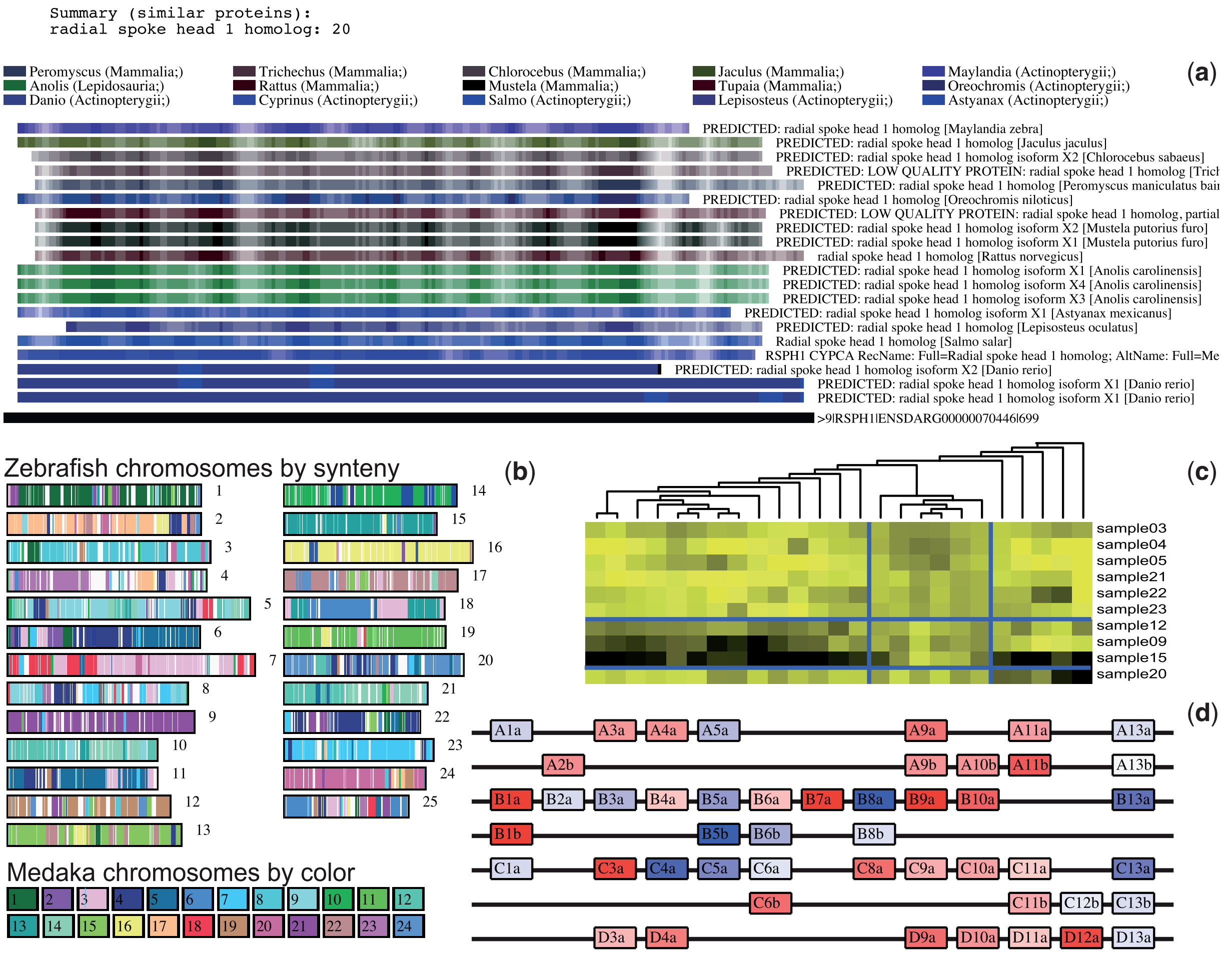

Fig. 1. Example graphs generated with Whiteboard. (**a**) Visualization of protein alignments, color coded by taxa and shaded by local sequence similarity; (**b**) Synteny between the zebrafish and medaka genomes, color coded by chromosomes (boxes to the left). Alignments were generated with Satsuma (Grabherr *et al.*, 2010); (**c**) A clustered heat map based on RNA-Sequencing expression values (yellow notes high expression, black low); (**d**) One time point of differential expression of HOX genes (data from Yang *et al.*, 2013) during the development of zebrafish (blue shades indicate lower than the mean, red higher). An animated image containing all time points can be found at http://whiteboard-class.sourceforge.net/

The *Whiteboard* interface is accessed through the 'whiteboard' class which is a collection of graphics objects to be drawn, and a display object on to which the graphics objects are displayed, either through a 2D or 3D geometry. Graphics objects are defined as derived classes of the abstract base class 'graphic' and provide the basic definitions of what is to be displayed. A toolkit of basic graphics objects is provided, including 'line', 'rect', 'arc', etc. as well as compound graphics, such as arrows and boxes. Display objects, represented by the abstract base class 'display_type', are used to transform the basic definitions provided by the graphic objects into the format of choice for the graphical output. A derived class, 'ps_display', is provided to generate figures in Post-Script (PS-Adobe-3.0 EPSF-3.0) displays, which are of production quality, as well as fully editable in graphics processing programs. The code is designed to be extensible, and other file formats can be accommodated by implementing a class for the desired output format. Detailed documentation about the interface, together with a set of example source code files, can be found here: http://whiteboard-class.sourceforge.net/.

In conclusion, *Whiteboard* provides a programmatic interface for visualizing large and complex datasets, which can easily be integrated into programs analyzing biological or medical processes.

Images can be generated in production quality and ready for publication, interactively, or even as animations. Implemented in C++, *Whiteboard* allows for adding extensions supplied by us, and its wider community of users.

Acknowledgements

We would like to thank the Bioinformatics Initiative for Life Sciences (BILS) in Sweden and the Department of Laboratory Medicine at Boston Children's Hospital.

Funding

The work was funded in part by salary support to E.M. from Boston Children's Hospital and Claritas Genomics and by a start-up grant to M.G.G. by the Science for Life Laboratories.

Conflict of Interest: none declared.

References

Grabherr,M.G. *et al.* (2010) Genome-wide synteny through highly sensitive sequence alignment: Satsuma. *Bioinformatics*, **26**, 1145–1151.

Yang,H. *et al.* (2013) Deep mRNA sequencing analysis to capture the transcriptome landscape of zebrafish embryos and larvae. *PLoS One* 8, e64058.