Bacterial Gene Neighborhood Investigation Environment: A Large-Scale Genome Visualization for Big Displays

Jillian Aurisano, Khairi Reda, Andrew Johnson and Jason Leigh

Fig. 1: BactoGeNIE enables large-scale comparisons across hundreds of gene neighborhoods on large, high-resolution environments. Similarities and differences in local gene content around target genes are highlighted through alignment, coordinated coloring of ortholog clusters, and directional color ramps. This image shows the neighborhood around a hypothetical protein in all draft *Escherichia coli* genomes from the PubMed database. Image courtesy of the UIC Electronic Visualization Laboratory (Photo: Lance Long, UIC).

Index Terms—Large and High-res Displays, Design studies

1 Introduction

Improvements in genome sequencing technology over the past decade have driven down sequencing costs faster than Moore's Law producing a genome sequencing boom [12]. Accelerated rates of complete genome sequence production are particularly evident in bacterial genomics, where small genome sizes enable rapid and inexpensive sequencing. These large volumes of complete genome sequences have given researchers a new approach to the long-standing challenge of identifying and characterizing novel bacterial genes: **comparative gene neighborhood analysis**. Due to unique properties in bacterial genome organization, researchers believe that it is possible to generate hypotheses around the function and pathway membership of novel genes by examining the neighborhood around **gene orthologs**, or genes with highly similar sequences. Visual approaches to this problem are necessary, since subtle patterns and relationships can be missed through automated approaches, but current comparative gene neighborhood visualizations are only designed to accomodate comparisons across 2-9 genomes in a single view ([11, 5, 8, 6, 3, 9, 4]).

In parallel, technical developments have enabled rapid advances in display and graphics hardware, enabling low-cost, high-resolution environments. Recent research has indicated that these environments present perceptual and cognitive benefits, such as allowing users to perform visual queries over a larger volume of data by scaling up perceptual processing [13] and permiting the use of embodied cognition, such as spatial memory and proprioception, in exploring a large and complex dataset [2]. While it is tempting to simply 'scale-up' current comparative gene neighborhood approaches to show more data on a larger display, we found that the underlying design of these approaches does not scale visually to these environments, due to encodings that assume low-information density and small-scale display environments.

In response to the aforementioned developments in bacterial genome sequencing and display technology, and the need for visually scalable approaches, we worked closely with a team of genomics researchers to develop a novel visualization approach and application called BactoGeNIE, which stands for Bacterial Gene Neighborhood Investigation Environment. Our design targets the patterns and relationships that are difficult to identify through automated methods alone in comparative gene neighborhood analysis, such as the identification of deletions, duplications, insertions and inversion events, as well as recognition of sublte patterns that are significant to experts. Unlike current comparative gene neighborhood, our design is oriented around enabling large-scale comparisons across big displays.

2 BactoGeNIE Design

Current comparative gene-neighborhood approaches do not scale effectively to big displays because they do not *use increased pixel density to show more entities or relationships, or show higher levels of detail* because their designs assume low-information density. To accommodate and leverage an increase in pixel density, our approach adopts a high-information density design (figure 2). We eliminate the comparative track, eschewing the orthology-line approach, and displays gene identities or other textual data on-demand in pop-up bubbles, rather than by-default. Orthology is shown using coordinated application of color across ortholog clusters. This color can be given to individual genes by a user, or through the ortholog targeting function, described below.

- *Jillian Aurisano, Khairi Reda and Andrew Johnson are with the Electronic Visualization Laboratory at the University of Illinois at Chicago E-mail: jauris2@uic.edu, mreda2@uic.edu, ajohnson@uic.edu.*
- *Jason Leigh is with the University of Hawaii at Manoa E-mail: leighj@hawaii.edu.*

Manuscript received 31 Mar. 2014; accepted 1 Aug. 2014; date of publication xx xxx 2014; date of current version xx xxx 2014.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

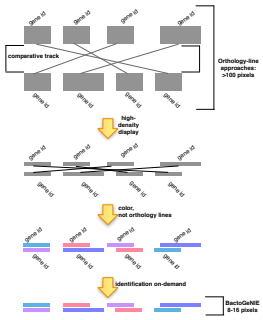
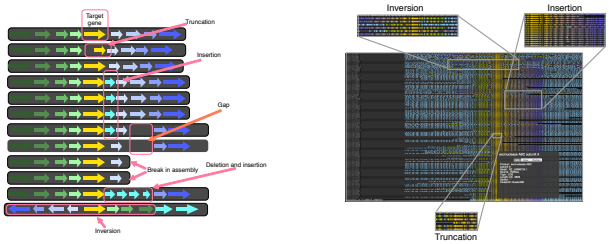


Fig. 2: BactoGeNIE compresses the space available for each genome, eliminates orthology lines and uses color to indicate orthologs across genomes.



(a) (b)

Fig. 3: The ortholog cluster targeting function, highlights differences between gene neighborhoods.

In addition, current comparative gene-neighborhood designs also do not *scale-up spatially across a big display* because their designs are oriented around small displays where users can see the entirety of the visualization at one level of detail in a single view. To better accomodate large display sizes, we adopted *clustering techniques*, by implementing *genome sorting* and *ortholog alignment*, which cluster gene neighborhoods by similarity and positions orthologs in a single region of the screen (figure 3). This, coupled with ortholog coloring, enables detail-oriented comparisons close to the display. Once aligned, the spatial distance between similar genes in distinct genomes and the target becomes immediately evident, with variations in distance signaling potentially meaningful variations.

Current comparative gene neighborhood approaches use encodings that allow for comparison through visual search, not pre-attentive processing. To enable rapid and immediate comparisons between gene neighborhoods, we developed an *ortholog cluster targeting function*, which combines genome sorting, alignment and coloring. A user-selected target gene is highlighted, centered on the display and a color gradient is applied to genes on either side, yellow to blue in the upstream direction, and yellow to green in the downstream direction, and this color is applied to gene orthologs in all genomes on display. Gene neighborhoods with identical gene content will have an identical color gradient. Genes not present in the targeting neighborhood will remain the default color, signaling to the user the occurrence of an insertion or deletion event. The gradient is directional, and as a result inversion events, where segments of a genome migrate onto a different strand, will be quickly visible by a reversal of the color gradient. This novel visualization algorithm, creates a view that enables high-density gene neighborhood analysis across hundreds of related bacterial genomes (figure 1).

3 Results and Conclusion

This approach accommodates more gene neighborhood comparisons than previous tools (Figure 4).

In addition, user feedback suggests that our design better addresses the domain problem than existing tools allowing for easier comparisons across genomes.

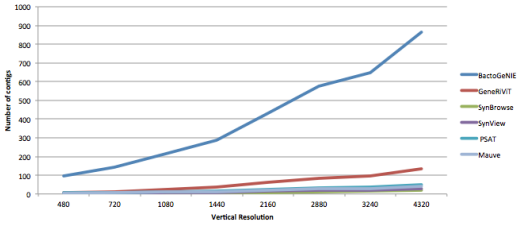


Fig. 4: BactoGeNIE is capable of displaying more gene neighborhoods simultaneously than other approaches.

To the authors knowledge, this is the first interactive, large-scale comparative gene neighborhood visualization for big displays. Since genome sequencing rates will likely continue to increase in the coming decade, and costs of large displays will likely continue to decrease, we believe that our design approach may inform future large-scale comparative genome visualizations. Future work will concentrate on expanded user feedback and case-studies to document the uses of this approach.

Acknowledgments

The authors wish to thank the computational biology team at Monsanto.

References

[1] D. Albers, C. Dewey, and M. Gleicher. Sequence surveyor: Leveraging overview for scalable genomic alignment visualization. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2392–2401, 2011.

[2] C. Andrews, A. Endert, B. Yost, and C. North. Information visualization on large, high-resolution displays: Issues, challenges, and opportunities. *Information Visualization*, 10(4):341–355, 2011.

[3] A. C. Darling, B. Mau, F. R. Blattner, and N. T. Perna. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome research*, 14(7):1394–1403, 2004.

[4] C. Fong, L. Rohmer, M. Radey, M. Wasnick, and M. J. Brittnacher. Psat: a web tool to compare genomic neighborhoods of multiple prokaryotic genomes. *BMC bioinformatics*, 9(1):170, 2008.

[5] S. McKay. Using the generic synteny browser. In *Plant and Animal Genome XX Conference (January 14-18, 2012)*. Plant and Animal Genome, 2012.

[6] M. Meyer, T. Munzner, and H. Pfister. Mizbee: a multiscale synteny browser. *Visualization and Computer Graphics, IEEE Transactions on*, 15(6):897–904, 2009.

[7] R. Overbeek, M. Fonstein, M. Dsouza, G. D. Pusch, and N. Maltsev. The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences*, 96(6):2896–2901, 1999.

[8] X. Pan, L. Stein, and V. Brendel. Synbrowse: a synteny browser for comparative sequence analysis. *Bioinformatics*, 21(17):3461–3468, 2005.

[9] A. Price, R. Kosara, and C. Gibas. Gene-rivit: A visualization tool for comparative analysis of gene neighborhoods in prokaryotes. In *Biological Data Visualization (BioVis), 2012 IEEE Symposium on*, pages 57–62. IEEE, 2012.

[10] R. A. Ruddle, W. Fateen, D. Treanor, P. Sondergeld, and P. Ouirke. Leveraging wall-sized high-resolution displays for comparative genomics analyses of copy number variation. In *Biological Data Visualization (BioVis), 2013 IEEE Symposium on*, pages 89–96. IEEE, 2013.

[11] H. Wang, Y. Su, A. J. Mackey, E. T. Kraemer, and J. C. Kissinger. Synview: a gbrowse-compatible approach to visualizing comparative genome data. *Bioinformatics*, 22(18):2308–2309, 2006.

[12] K. Wetterstrand. DNA sequencing costs: Data from the NHGRI large-scale genome sequencing program. `http://www.genome.gov/sequencingcosts` (Accessed May 7, 2014).

[13] B. Yost, Y. Haciahmetoglu, and C. North. Beyond visual acuity: the perceptual scalability of information visualizations for large displays. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 101–110. ACM, 2007.