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 permitting 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 subtle 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.
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.

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