**EnteriX 2003: visualization tools for genome alignments of Enterobacteriaceae**

Liliana Florea*, Michael McClelland1, Cathy Riemer2, Scott Schwartz2 and Webb Miller2

Informatics Research, Celera/Applied Biosystems, 45 W. Gude Drive, Rockville, MD 20850, USA, 1Sidney Kimmel Cancer Center, 10835 Altman Row, San Diego, CA 92121, USA and 2Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA 16802, USA

Received February 15, 2003; Revised and Accepted March 27, 2003

**ABSTRACT**

We describe EnteriX, a suite of three web-based visualization tools for graphically portraying alignment information from comparisons among several fixed and user-supplied sequences from related enterobacterial species, anchored on a reference genome (http://bio.cse.psu.edu/). The first visualization, Enteric, displays stacked pairwise alignments between a reference genome and each of the related bacteria, represented schematically as PIPs (Percent Identity Plots). Encoded in the views are large-scale genomic rearrangement events and functional landmarks. The second visualization, Menteric, computes and displays 1Kb views of nucleotide-level multiple alignments of the sequences, together with annotations of genes, regulatory sites and conserved regions. The third, a Java-based tool named Maj, displays alignment information in two formats, corresponding roughly to the Enteric and Menteric views, and adds zoom-in capabilities. The uses of such tools are diverse, from examining the multiple sequence alignment to infer conserved sites with potential regulatory roles, to scrutinizing the commonalities and differences between the genomes for pathogenicity or phylogenetic studies. The EnteriX suite currently includes >15 enterobacterial genomes, generates views centered on four different anchor genomes and provides support for including user sequences in the alignments.

**INTRODUCTION**

The enterobacterial system of microbial genomes provides one of the largest available collections of sequences from related species (Table 1). The availability of such a large volume of data opens up the opportunity for comparative studies to identify and interpret the common and divergent features among species and their phenotypic impact and brings the challenge of effectively organizing and presenting the data in a fashion that is both concise and informative. We describe EnteriX, a web visualization system that answers this need, consisting of three tools for graphically portraying alignment information from comparisons between a reference sequence [either *Escherichia coli* K-12 (ECO), *E.coli* O157:H7 (ECH), *Salmonella typhimurium* LT2 (STM) or *Salmonella typhi* CT18 (STY)] and several related genomes (Table 1), together with integrated data and annotations of genomic re-organization events and functional sites.

**RESULTS**

**Enteric**

The first visualization component, Enteric, presents pairwise alignments between a reference genome and each of the related bacteria, in a 20 Kb region centered at a user-specified address or gene in the reference genome. Alignments are represented schematically as PIPs (Percent Identity Plots; Fig. 1). A PIP is a 2D plot in which positions along the horizontal axis correspond to locations in the reference genome, and coordinates on the vertical axis correspond to alignment percent sequence identity levels, restricted to the 50–100% range. The ungapped segments within each alignment are represented as horizontal lines spanning the corresponding range in the reference genome and at a vertical position equal to the ungapped alignment's percent sequence identity value. The ends of alignments in Enteric PIP views are marked with color-coded bars that indicate deletion, insertion and rearrangement events between the genomes. Additional information, such as the length of deletion or the location of the nearest neighbor in the other genome, is revealed by placing the mouse pointer on the feature. Genes annotated in the reference genome are shown with arrows above the PIPs. Using an embedded-hyperlink mechanism, their names contain links to the associated COG category pages maintained at NCBI (11) in the ECO and ECH centered views or to a page containing information on *E.coli* orthologs of *S.typhimurium* genes at the web site of the Washington University Salmonella Sequencing Center, for the STM-centered views. The output is presented in PDF format. The alignments of fixed genomes are pre-computed using a locally developed program called blastz (12,13), an independent implementation of the Gapped BLAST algorithm.

*To whom correspondence should be addressed. Tel: +1 240 453 3736; Fax: +1 240 453 3324; Email: liliana.florea@celera.com

Published by Oxford University Press 2003
Table 1. List of enterobacterial genomes included in the comparative views

<table>
<thead>
<tr>
<th>Species name</th>
<th>Abbr.</th>
<th>Number of contigs</th>
<th>Ref.</th>
<th>Sequencing center and data source</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>E. coli</em> K-12</td>
<td>ECO</td>
<td>1 chrom</td>
<td>(1)</td>
<td>U. of Wisconsin, Madison GenBank A#: U00096</td>
</tr>
<tr>
<td><em>E. coli</em> 0157:H7 EDL933</td>
<td>ECH</td>
<td>1 chrom</td>
<td>(2)</td>
<td>U. of Wisconsin, Madison GenBank A#: NC_0022655</td>
</tr>
<tr>
<td><em>E. coli</em> CFT073</td>
<td>ECU</td>
<td>1 chrom</td>
<td>(3)</td>
<td>U. of Wisconsin, Madison GenBank A#: AE014075</td>
</tr>
<tr>
<td><em>Shigella flexneri</em></td>
<td>SHG</td>
<td>1 chrom</td>
<td>(4)</td>
<td>Microbial Genome Center of Chinese Ministry of Public Health GenBank A#: AE005674</td>
</tr>
<tr>
<td><em>S.typhimurium</em> LT2</td>
<td>STM</td>
<td>1 chrom 1 plasmid</td>
<td>(5)</td>
<td>Washington U. St. Louis GenBank A#: AE006468</td>
</tr>
<tr>
<td><em>S.typhimurium</em> SL1344</td>
<td>SSL</td>
<td>426</td>
<td></td>
<td>Sanger Center, UK <a href="http://www.sanger.ac.uk/Projects/Salmonella">http://www.sanger.ac.uk/Projects/Salmonella</a></td>
</tr>
<tr>
<td><em>S.typhi</em> CT18</td>
<td>STY</td>
<td>1 chrom 2 plasmids</td>
<td>(6)</td>
<td>Sanger Center, UK GenBank A#: 5L153382</td>
</tr>
<tr>
<td><em>S.typhi</em> Ty2</td>
<td>STT</td>
<td>1 chrom 1 plasmid</td>
<td>(7)</td>
<td>U. of Wisconsin, Madison GenBank A#: NC_004631</td>
</tr>
<tr>
<td><em>S.paratyphi</em> C</td>
<td>SPC</td>
<td>2289</td>
<td></td>
<td>U. of Calgary, Canada</td>
</tr>
<tr>
<td><em>S.dublin</em></td>
<td>SDU</td>
<td>2442</td>
<td></td>
<td>U. of Illinois Urbana-Champaign <a href="http://salmonella.utmem.edu/">http://salmonella.utmem.edu/</a></td>
</tr>
<tr>
<td><em>S.enteritidis</em></td>
<td>SEN</td>
<td>2303</td>
<td></td>
<td>U. of Illinois Urbana-Champaign <a href="http://salmonella.utmem.edu/">http://salmonella.utmem.edu/</a></td>
</tr>
<tr>
<td><em>S.hongori</em></td>
<td>SBO</td>
<td>66</td>
<td></td>
<td>Sanger Center, UK <a href="http://www.sanger.ac.uk/Projects/Salmonella">http://www.sanger.ac.uk/Projects/Salmonella</a></td>
</tr>
<tr>
<td><em>S.diarizonae</em></td>
<td>SDI</td>
<td>576</td>
<td></td>
<td>Sanger Center, UK <a href="http://www.sanger.ac.uk/Projects/Salmonella">http://www.sanger.ac.uk/Projects/Salmonella</a></td>
</tr>
<tr>
<td><em>Yersinia pestis</em></td>
<td>YPE</td>
<td>1 chrom 3 plasmids</td>
<td>(8)</td>
<td>Sanger Center, UK GenBank A#: NC_001435</td>
</tr>
<tr>
<td><em>Vibrio cholerae</em></td>
<td>VCH</td>
<td>2 chrom</td>
<td>(9)</td>
<td>TIGR GenBank A#: NC_025055 and NC_02506</td>
</tr>
<tr>
<td><em>Pseudomonas aeruginosa</em></td>
<td>PAE</td>
<td>1 chrom</td>
<td>(10)</td>
<td>Pseudomonas Genome Project GenBank A#: AE004091</td>
</tr>
</tbody>
</table>

algorithm that was specifically designed to compare two long sequences, and stored for fast retrieval. The server also has support for incorporating one user-supplied sequence, for which the alignment and annotations are computed on-the-fly.

**Menteric**

The second tool, Menteric, computes and displays 1 Kb views of annotated multiple alignments of the same sequences, starting at a user specified address or gene in the reference genome, shown at nucleotide-level resolution (Fig. 2). Known functional sites and other characteristic regions are marked on the alignment with a combination of graphical symbols. The user can select from five different criteria for determining conserved regions in the multiple alignment, ranging from consensus majority rule, phylogenetic distance, information content and distance from a fixed or unknown center sequence (14). The conserved regions thus identified are shown enclosed in boxes. In addition, known or predicted regulatory sites are marked with color-coded underlays (light chocolate, ORFs; green, promoters; red, regulatory protein binding sites) and may contain links to the annotation data source. Currently, only the GenBank annotation is available (1,2,5,6). Links embedded in the sequence labels on the right-hand side of the multiple alignment can be used to download sequence data in embedded in the sequence labels on the right-hand side of the multiple alignment views, respectively, and adds interactive zoom-in capabilities. The Maj wide view (Fig. 3A) uses Enteric’s paradigm for presenting alignments and associated information. However, unlike Enteric, information about the various features that was previously provided via labels and embedded pseudo-links is now displayed in the two message boxes at the top of the window. The top box displays details about the mouse location (PIP coordinates, contig name, properties associated with the color bands), while the bottom one is used to show information about the local alignment that the user has selected by clicking on its horizontal line in the PIP. Maj’s close-up view (Fig. 3B) emulates Menteric’s nucleotide-level multiple alignment views, but the information is organized somewhat differently. The multiple alignment is now shown in a scrollable bar at the bottom of the frame, while the main panel displays interactive PIPs of pairwise alignments projected directly from the multiple alignment. Using a Java applet, Maj allows the user to ‘zoom in’ on a sub-region of the view, selected by dragging the mouse in any PIP panel. It also provides the ability to toggle between the wide and close-up views for the current region, using the View button located at the top of the window.

**New features**

Since its public release in 2000 (15), the EnteriX suite has expanded to include >15 enterobacterial species, including the completely sequenced *E.coli* K-12, 0157:H7 and CFT073, *S.typhimurium* LT2 and *S.typhi* strains CT18 and Ty2, *Vibrio cholerae*, *Pseudomonas aeruginosa*, *Yersinia pestis* and a number of partially sequenced Salmonella and Klebsiella species. It has also been adapted to present alternative reference genomes (*E.coli* K-12 and 0157:H7, *S.typhimurium* LT2 and *S.typhi* CT18). With the increase in the number of genomes scheduled to be partially or completely sequenced over the next years, particularly from among the Salmonella species, storing, organizing and presenting the information efficiently will become increasingly difficult. To answer the need for flexibility and compactness, EnteriX now provides the user with the ability to select the genomes to
be included in the views from among those available in the data store. Perhaps the most notable new feature is that the Enteric and Menteric tools have acquired the capability to include a user-provided sequence in their comparative views. The third tool, Maj, is currently being updated to incorporate some of these new features as well.

**AVAILABILITY**

The EnteriX servers are available from http://bio.cse.psu.edu and from the Salmonella Sequencing Center site at Washington University, St Louis (http://genome.wustl.edu/projects/bacterial/styphimurium/).

Figure 1. (A and B) Enteric display at the S.typhimurium acnB gene location. The P.aeruginosa genome is used as the user-supplied sequence (USR). Alignments were pre-computed using the program blastz with parameters match = 10, mismatch = -10, gap_open = 60, gap_ext = 2, starting from exact words of size w = 8 (12 for genomes from the same species). Gap-free alignment segments are shown as horizontal bars in the PIPs, at a vertical coordinate corresponding to the alignment segment’s percent sequence identity, and genomic re-arrangement events are marked with color bars at the ends of alignments. Multiple horizontal matches at the same location in the reference genome indicate paralogues in the other sequence (e.g. the 174–174.8 K region in the STM PIP).
DISCUSSION

Recent years have brought a tremendous increase in the amount of sequence data from various genome sequencing projects, an increase that is projected to accelerate over the next years. As a result, the task of organizing and summarizing the data to extract the most informative features has become a challenging yet critical endeavor. Visualization is an effective way of structuring and presenting such information effectively, in a concise and eloquent fashion. The software we describe, EnteriX, has been developed to present alignment information and inferred or associated properties in an integrated framework, as an instrument for discovery and analysis.

To aid in the processes of structural and functional annotation, as well as in selecting the most promising candidates for experimental validation, integration of data from various complementary resources is essential. The PDF files produced by Enteric and Menteric, as well as Maj's Java views, contain hyperlinks to related repositories of information on the internet, such as GenBank entry pages, COG and the Washington University list of orthologous ECO and STM genes. Using the same mechanism, hyperlinks incorporated in Menteric's sequence labels allow one to download contig sequence data for further analyses.

To increase the applicability of these tools, we plan to incorporate additional reference genomes and to provide more extensive access to external sources of data, such as RegulonDB's database of experimentally validated E.coli regulatory sites, using our established mechanisms of embedded hyperlinks. With the ongoing effort to sequence a number of Salmonella serovars and related genomes, including S.paratyphi A and K.pneumoniae, at the Washington University Genome Sequencing Center in St Louis, it is anticipated that

Figure 2. Menteric-type view in the common regulatory region of the E.coli K-12 araC and araBAD operons. Color-coded underlays are used to mark known or predicted functional elements in the E.coli genome (light chocolate, ORFs; green, promoters; red, binding sites of regulatory proteins). The user sequence (usr) is the concatenation of S.bongori contigs.
Figure 3. (A and B) Maj wide and close-up views centered at the carAB operon in E.coli.
EnteriX will provide a complex and multi-faceted view of the genomics of the Enterobacteriaceae and will prove a valuable resource in the area of visualizing integrated annotation for the bacterial genomics community.

ACKNOWLEDGEMENTS

This work was supported in part by grant HG-02238 from the National Human Genome Research Institute to W.M. and grant AI34829 to M.M.

REFERENCES


