Visualizing the Prevalence of Gene Duplication in Bacterial Genomes

By

Jer-Yee (John) Chuang

UC Berkeley School for Information Management and Systems

1. Introduction:

The goal of this project is to apply a visualization method to better understand data obtained from the analysis of gene duplication[1] across bacterial genomes. Gene duplication is thought to be a mechanism for evolving complexity in gene regulation. Understanding how such mechanisms evolved would shed light on how genomes evolved as a whole and the forces involved. Although, there have been numerous studies on gene duplication, none have focused specifically on proximally duplicated transcription factors (PD-TF). These are pairs of genes that lie in proximity to one another on the genome and express transcription factors. Research experiments into PD-TFs were conducted by the author back in 2004. The data used in this visualization study will be from those experiments.

Due to the complexity and multi-dimensional nature of this dataset, it is difficult to visualize in a meaningful and effective manner. Although there are a large number of bioinformatics visualization tools, they are usually not reusable for purposes different from those they were built for. This is a consequence of biology being predominantly a hypothesis driven science and the tools being built on a case-by-case basis. Back in 2004, the author made an initial attempt at visualizing this data but no concrete conclusions were drawn. Perhaps developing a different visualization tool would yield insight.

This visualization tool would be targeted towards evolutionary biologists interested in genomic studies. It would clearly present the multi-dimensional datasets by merging the statistical information among genes with the evolutionary information among genomes. More than just presentation, this tool would be invaluable in the analysis of the dataset leading to a greater understanding the role of duplication as a driving force for evolution of regulatory complexity in biological organisms. It would allow researchers to see patterns in the dataset that may help in answering questions regarding:

[1] We have assumed that the reader has a working knowledge of basic molecular biology concepts.

- The prevalence of a set of duplicated genes in all other bacterial genomes.
- The tracing of the evolutionary history of a gene.
- The likelihood gene acquisition through horizontal gene transfer.
- The possibility of ‘reprogramming’ bacterial by re-using transcription factors in one genome to regulate pathways in another genome (i.e., synthetic biology).

2. Dataset:

A. Description

The dataset used in this study will be a subset of that created by the author in the Spring of 2004 during his research on gene duplication in bacterial genomes at UCSF. Thirty pairs of genes (encoding transcription factors) were identified in E.Coli K12 (used as the reference genome). The two genes in each of these pairs were determined to be duplicates of each other. A search was then done for each of these 30 pairs in all other fully sequenced bacterial genomes. For each genome, we identified the presence of any gene sequence homologues for each gene in the pairs. For simplicity, we used reciprocal best hit (in BLAST[2]) as the selection criteria and percent identity as a filtering metric on low BLAST scoring hits.

B. Interpretation

For each pair of genes from our reference genome, we can expect one of five possible outcomes when we search for that pair in another genome. We have summarized those five outcomes in Figure 1 (yellow boxes represent genes). Since we are examining gene pairs, it is useful to distinguish between the two genes in the pair by denoting them as ‘left’ and ‘right’ respectively. We will use the term ‘both’ to refer to instances where we consider both left and right genes together. We also define as proximity as residing within two coding sequences either upstream or downstream.

[2] Sequence alignment tool developed by NCBI. See references link.

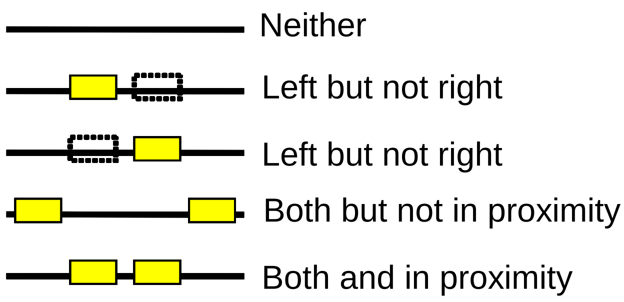


Figure 1: Summary of the Five Outcomes

For this visualization study, the dataset is comprised of four slices (i.e., matrices) each having the same dimensions of 56 rows by 27 columns. Each row represents a genome, and each column represents a gene pair from the reference genome. Hence, there are 56 genomes and 27 gene pairs represented. The first three slices correspond to the percent identity of the left, of the right, and of both genes respectively. The last slice corresponds to the outcomes: one of the five in the above figure for each “genome and reference gene pair” pair.

C. Challenges

There are two primary challenges in visualizing this dataset. The first involves answering what it means to say that a duplication of the reference gene exists in another genome. A statistical measure needs to be associated with that prediction. Secondly, the bacterial genomes that we are examining have some evolutionary relationship.

i. Percent Identity as Statistical Measure:

We chose to use percent identity as the statistical measure of sequence similarity between proteins. While high sequence similarity does not necessary guarantee similar protein function, there is a high correlation between these two characteristics. Additionally, percent identity is intuitive and extremely simple to compute. Consider the following example of two sequences that differ in the last two bases (bold/underlined).

Seq1: ggtagc**ca** Seq2: ggtagc**ga** Percent Identity: 6/8 bases = 75%

We can now associate a statistical measure for each of the five outcomes in our data. Obviously, if no genes are found then the associated percent identity is 0%. Otherwise, for the two cases where only one gene is found, we compute the percentage identity of that gene with the reference gene. For the two cases where both reference genes are found, we can compute the percentage identity of the two genes in addition to the percentage identity of each with their respective reference genes. In the end, for each data point, we have three associated percent identity values: (left, right, combined). For example, suppose a data point had an associated percentage identity of (25%, 45%, 15%). This would indicate that the left gene is 25% similar in sequence to the left gene in the reference pair, and the right gene is 45% similar in sequence to the right gene in the reference pair. These two genes are also 15% similar in sequence with each other.

ii. Dendograms and Evolutionary Relationships:

We chose to use a dendogram to illustrate evolutionary relationships among the genomes. This tree-like representation is very popular among evolutionary biologists in phylogenetic analyses. Since that community is one of our targeted audiences, we felt that using a dendogram would be our visualization tool more accessible. The dendogram's strength is that it concisely outlines the divergence path and evolutionary distance of genomes from their ancestors. Its primary disadvantage is its poor use of vertical nature which makes poor use of screen real estate. This is especially pronounced and problematic when viewing large data sets. There has been some work by other researchers to address this problem (see related work section).

3. Related Work:

Visualization tools developed for biological datasets are usually not reusable. This is a consequence of biology being predominantly a hypothesis driven science as opposed to being data driven. Hence, visualization tools that act on biological datasets usually cannot be generalized. Instead, they are specific and built on a case-by-case basis. There has been no previous work that specifically addresses our proposed research question (the prevalence of duplicated transcription factor pairs in bacterial genomes). Although, there have several studies on gene duplication, these have presented their results using tables. We found this approach ungainly and difficult for broad analysis. Ultimately, we chose to create our own, which is outlined below.

Visualizing Phylogentic Relationships:

Another component of our visualization is the need to visualize phylogenetic information. This is almost always done with some variant on tree visualization, since they naturally illustrate evolutionary relationships among genomes. The simplest of these are dendograms. We have already mentioned their strengths and weaknesses in the previous section. For most dendogram visualization needs, TreeExplorer[3] (shown in Figure 2) is sufficient. Since, our intended use of a dendogram is only for presentation and not analysis purposes, we will use TreeView to generate our dendograms.

[3] Developed by Koichiro Tamura (see reference).

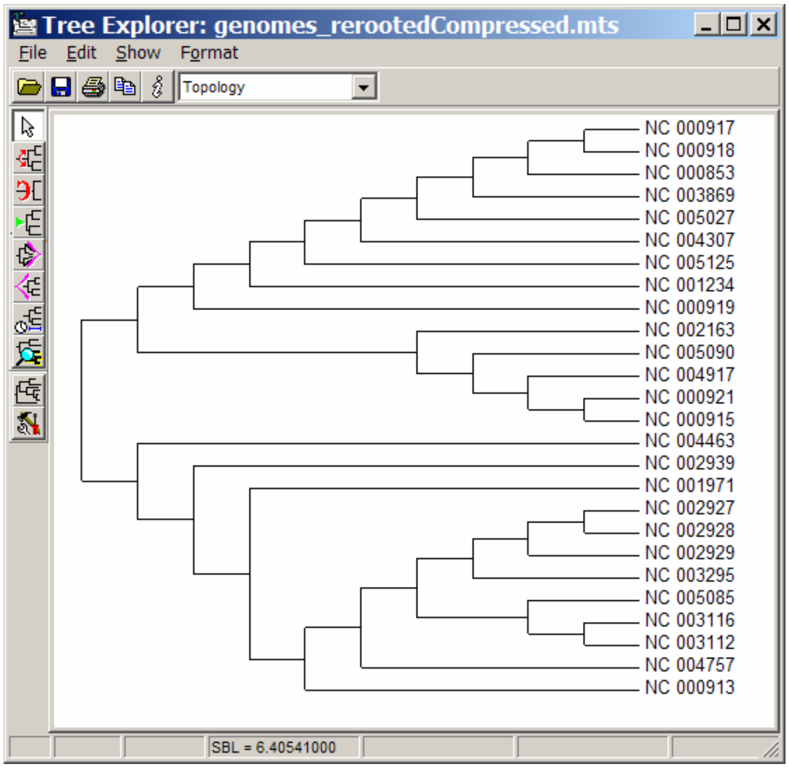


Figure 2: Treeview

However, as was pointed out previously, the main weakness is the inability of the dendogram to 'scale' gracefully for large number of nodes. One proposed solution is TreeJuxtaposer[4] (see Figure 3), which guarantees visibility of selected nodes. The downloadable software was problematic. It performed very slowly and was prone to crashing. The tree was also constantly being redrawn; though, this may be a graphics card issue. However, the software did require installation of the proprietary GL4Java library, which may have contributed to its instability. The mouse navigation was also non-intuitive. In short, while the concept is great, the implementation left more to be desired.

[4] Developed by Tamara Munzner (see reference).

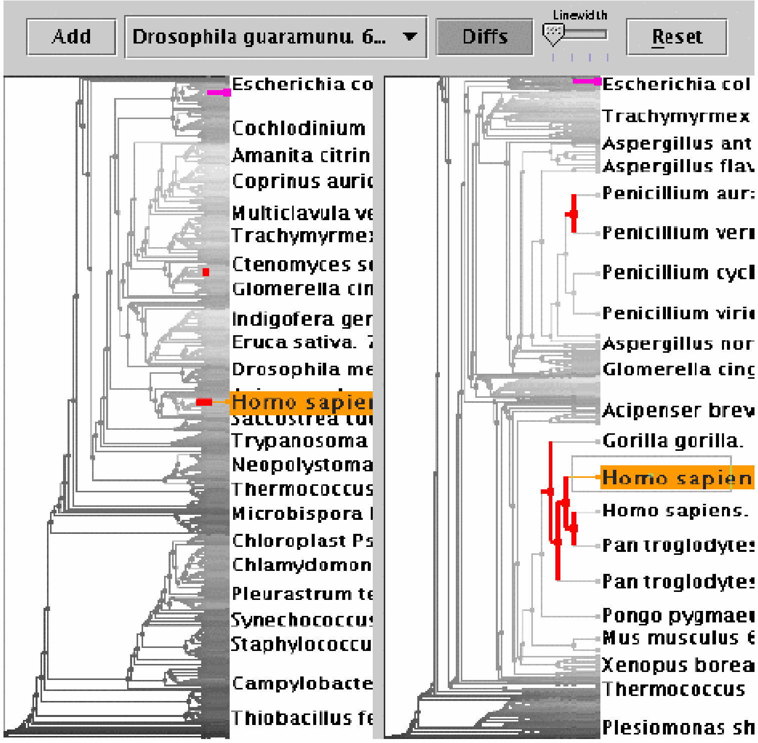


Figure 3: TreeJuxtaposer

Lastly, hyperbolic trees are another approach to visualizing phylogenetic information. Their strength is their ability to present a large number of nodes in a fixed amount of screen real estate. However, as we outlined above, there are dendogram based methods that offer similar feature. We believe that our target audience of mostly evolutionary biologists would prefer the use of the more familiar dendograms representation.

Previous Visualization Attempt:

Back when the dataset for this study was created by the author in 2004, he also attempted a visualization of it. The result was moderately successful but somewhat confusing. Only the outcomes slice of the dataset was visualized and none of the statistical information. Each of the five possible outcomes was encoded with a distinct hue (see Figure 4), as indicated by the color key at the lower left. The gene pairs reside in the columns and the genomes in the rows. The rows are ordered by taxonomy groups, but no dendogram is presented to show evolutionary relationship. There is no ordering for the gene pairs, though the numbering is misleading. Worse, the choice of hues is confusing since the blue/white colors encode data points that are more 'interesting' then their red/green counterparts. Yet, the red and green and more salient and plentiful- distracting our attention. Although this is not is not a microarray dataset, users who have experience working with those would also be inadvertently drawn to the red and green colors, which are the standard colors for microarray datasets. In the next section, we propose a visualization prototype that avoids these problems.

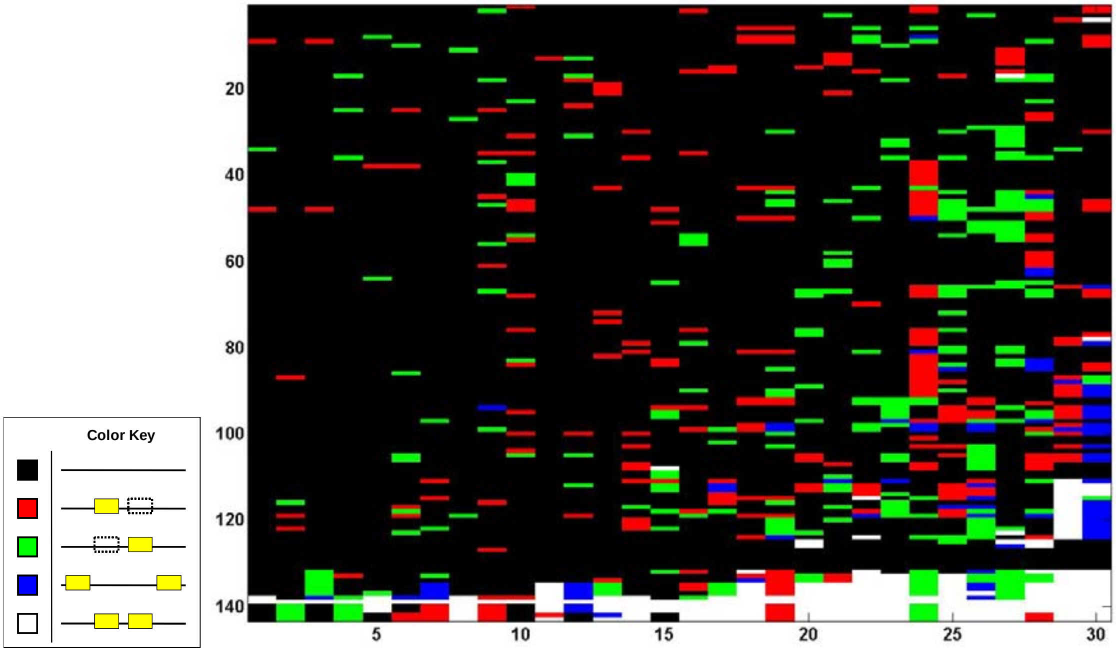
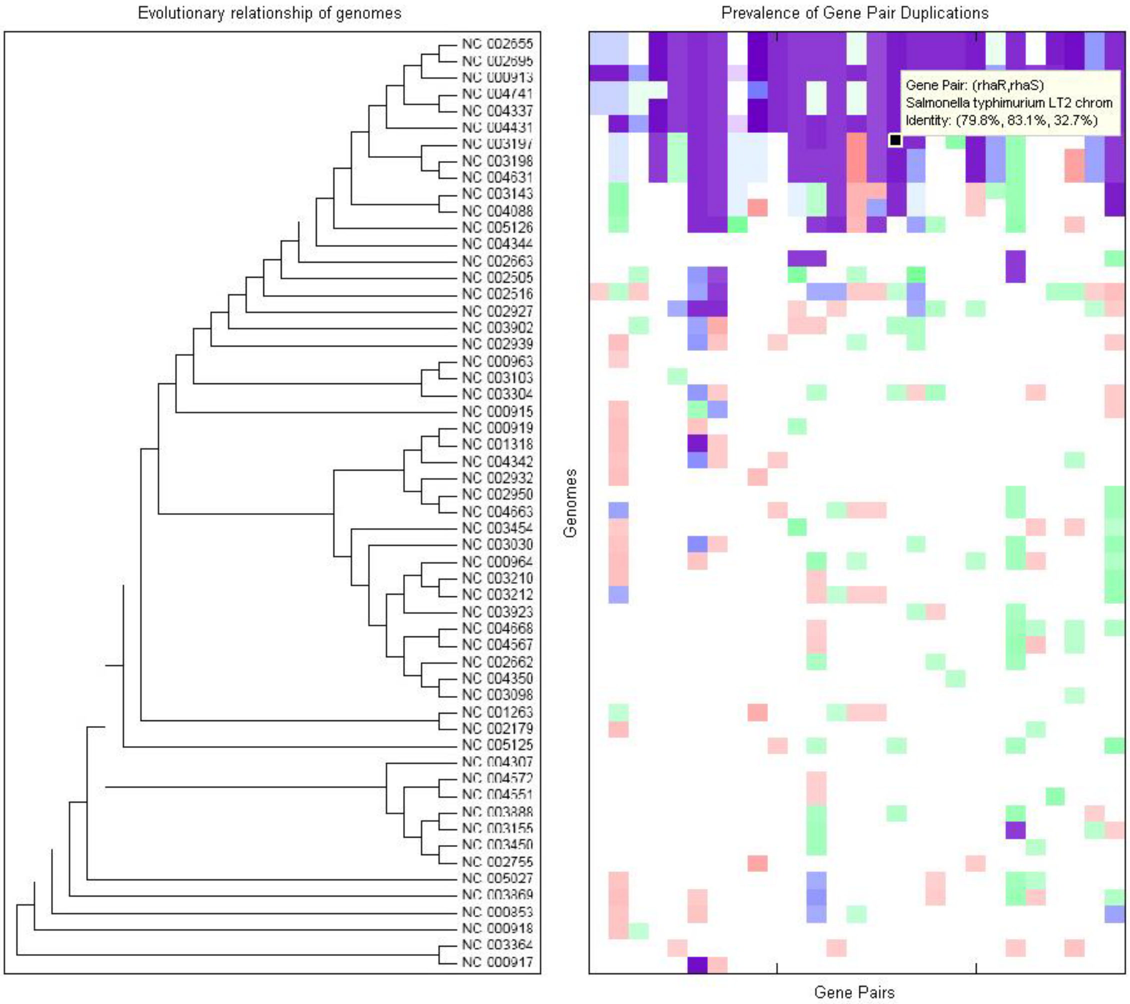


Figure 4: Previous Visualization Attempt

4. Proposed Prototype:

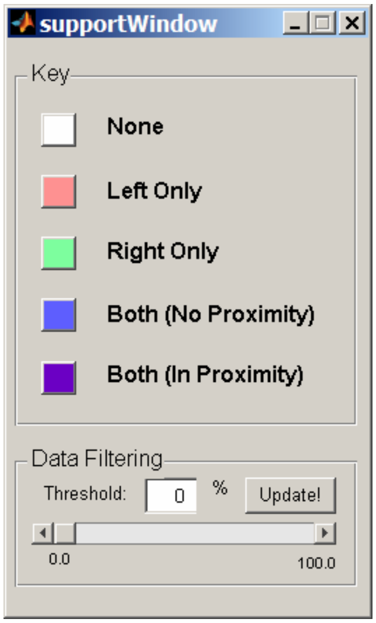
A. Design Specifications:

The need to couple statistical measures and evolutionary relationships within this multi-dimensional dataset makes this visualization particularly challenging. As discussed previously, we have chosen to use percent identity as the statistical measure and dendograms to illustrate evolutionary relationships. We merge these two aspects through a dual paneled layout (see Figure 5). The left panel contains the dendogram, while the right panel contains our dataset and statistical measures. Recall that for this dataset (right panel of Figure 5), the genomes are located in the rows while the gene pairs are in the columns. The color key and a slider bar for filtering the data is presented in a separate window (see Figure 6).



Gene Pair: (rhaR,rhaS)
Salmonella typhimurium LT2 chrom
Identity: (79.8%, 83.1%, 32.7%)

Figure 5: Dual Panel Visualization (Primary Window)



Threshold: 0 %

Figure 6: Color Key and Data Filtering Slider (Support Window)

i. Dendogram (Figure 5, left panel):

The genomes labeled in the dendogram are represented in the corresponding row directly to the right in the right panel. We have labeled the y-axis as ‘Genomes’ for the since it is shared between the two panels. This soft linking underscores the important fact that analysis of the data should take into account the inherent evolutionary relationships among the genomes represented.

ii. Coloring Scheme for Data Points (right panel Figure 5 and top panel of Figure 6):

We chose five distinct hues to represent the five possible outcomes in the dataset. These five hues are shown in the top panel of Figure 6. We wanted hues that were both soft to the eyes, but distinct enough to avoid confusion. The white hue was chosen for the case where neither gene in the pair was found (i.e., the background). For the case where only one gene was found, we have chosen a soft, pastel red and green. These are easy for viewing and less attention grabbing then the more bold colors used for the two remaining outcomes where both genes were found. Since the last two outcomes are distinguished only by whether the both genes in the pair are found in proximity, we believed it would be advantageous to choose to distinct but adjacent colors. These two outcomes also correspond to what biologist would find most interesting, so we wanted bolder, more salient colors. We used a dark sky blue hue and a purple hue. Since, the purple hue is more visible then the dark sky blue, we used it to encode the most interesting outcome- where both genes are present and in proximity.

Additionally, we wanted to encode the percentage identity associated with each data point. This was done by varying the saturation of each of the four non-white hues (Figure 5, right panel). The more saturated the hue, the higher the percentage identity value associated with that point.

iii. Details on Demand (Figure 5, right panel):

For increased effectiveness, we allowed the user to interact with the visualization by clicking on data points. Clicking on a data point brings up a tooltip with details. These details include: gene pair, genome, and percentage identity of the left, right, and both genes (if applicable). The visualization also supports simultaneous display of multiple tooltips (i.e., the user can click multiple data points and have the details of each displayed in separate tooltips). This allows for quick comparison among selected data points (Figure 7). Alternatively, the user can browse/explore the dataset by clicking a data point then dragging the mouse. The detailed tooltip is automatically updated for whatever data point the cursor is positioned over.

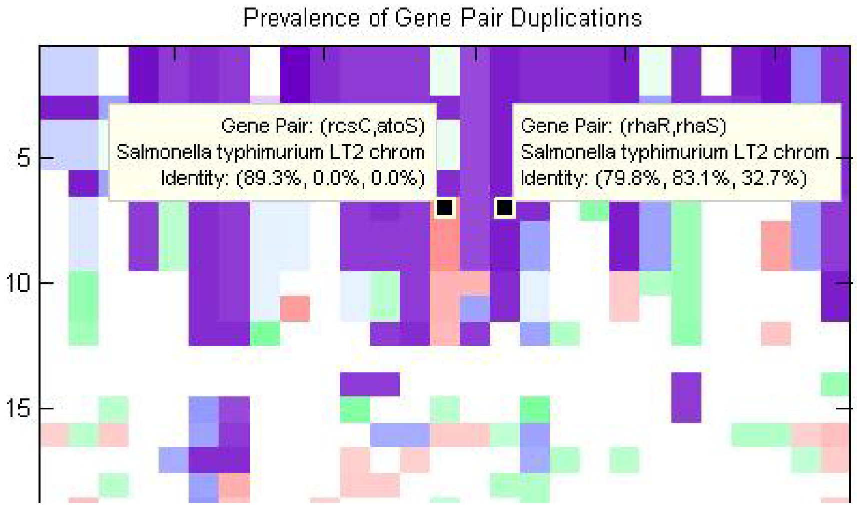


Figure 7: Support for Multiple Detailed Tooltips

iv. Data Filtering (Figure 6, bottom panel):

We can filter our data based on percent identity to only keep those that are high similarity. This is done using the slider in the bottom panel of Figure 6. All data points with percent identity values less than the threshold are filtered from the plot. The threshold value can either be set by the slider or direct entry into the editable textbox. The user then pushes the 'Update!' button to refresh the visualization in the Primary Window (i.e., Figure 5, right panel).

B. Implementation

We decided to build the visualization using Matlab. Most scientists already have experience using Matlab as a tool for mathematical/statistical analysis and manipulation of multi-dimensional datasets. This is crucial since many biologists are reluctant to learn yet another 'tool', but would rather use existing ones. Matlab also provides GUI building tools. Additional algorithms can be implemented and included with relative ease into the GUI. The drawback of using Matlab is that it is proprietary and does have a learning curve for new users.

The dendogram will be visualized using Treeview then exported as an image file. The image is then loaded into our Matlab based GUI for display. We opted not to implement the dendogram in Matlab since there are already many free packages that do a superb job at dendogram visualization. Implementing the dendogram directly in Matlab would have been tedious. Worse, it would have locked-in the user to that particular dendogram representation. Depending on the dendogram construction algorithm used and the node chosen as the root, there are numerous possible representations for a given dendogram. We would rather that the user be given a choice in deciding which representation is most suitable for his needs. The only drawback to this approach is the soft coupling between the dendogram and dataset panels. The rows of the dataset must be sorted to correspond to the sequence of nodes in the dendogram. This is currently handled through a Python script that pre-process the dataset before it is loaded into Matlab.

5. Evaluation:

A. Analysis:

This visualization prototype presents the dataset in a far clearer and more meaningful manner than the previous attempt (Figure 4). The use of judicious color choices coupled with good design principles such as details on demand and interactivity made the visualization more accessible while simultaneously encoding more information. This survey of the PD-TF 'duplication space' in bacterial genomes is actually quite insightful.

Some observations:

1. The high prevalence of both genes in the genomes near the top is expected since those genomes are similar to the reference genome (E.Coli K12). However, there are instances of blue/purple boxes near the bottom. These are very interesting since they imply the presence of genes from an evolutionary recent genome in an evolutionary ancient one.
2. Columns that contain blue/purple boxes at the top and bottom extremities (with few or none in between) may hint at gene acquisition via horizontal gene transfer.
3. There is a lot of variation in the fraction of reference gene pairs that are kept in the other genomes. This is probably a consequence functional evolution over time or differences in environment.
4. We can compare two related genomes (adjacent rows), and identity which gene pairs are in one but not the other. This may provide insight into the phenotypic variation observed on the clinical level.

B. User testing:

We enlisted three participants to test this visualization. All three participants are bioinformatics graduate students (mid 20s to early 30s), who are currently doing either genomics or proteomics research. They were each given a short verbal introduction of the dataset being visualized; and the design and features of the visualization tool. Each participant was then given three short questions to answer. Their answers to the questions and their general remarks were collected.

Questions presented to the participants:

1. Are there any genomes besides E.Coli K12 that contain the (appY,envY) gene pair? If so, how many?
2. Are there any genes pairs from E.Coli K12 (reference genome) found in Treponema pallidum? If so, how many?
3. Estimate the fraction of the genomes that contain at least one instance of a reference gene pair (regardless of whether the genes are in proximity)?

The first two questions test the participant's understanding of axes in this visualization. For the first question, the participant needs to identity the column corresponding to (appY,envY), then count the number of colored boxes in that column for the number of genomes that contain this gene pair. This correct number is actually one less, since one of the rows is the reference genome (E.Coli K12), which has been inserted as a control. For the second question, the participant needs to identity the row corresponding to Treponema pallidum then count the number of colored boxes in that row for the number of gene pairs that are in this genome. For the last question, the participant needs to realize that data points representing the outcomes where both genes are present are colored either blue or purple (obvious by looking at the Support Window with the color key). It is then easy to estimate by eye the fraction of rows that have at least one blue or purple box.

Response to questions by the participants:

All participants were able to successfully complete the three tasks outlined above. To answer the first two questions, the participants all browsed the data by click-and-dragging the mouse. This allowed them to locate the column/row corresponding to the question. It was easy to miss the fact that E.Coli K12 was included as a reference genome when browsing quickly. For third question, the participants had no problem realizing that they needed to use the color key and only consider rows with blue and purple boxes.

General remarks from the users:

1. The dual panel design is a bit awkward at first. Once explained, it works fine.
2. They liked the interactivity, especially having details on demand. The browsing by click-and-drag was a nice addition.
3. Four distinct hues good. The use of dark colors to highlight the important outcomes was good. However, at lighter saturations, these loose their ability to draw attention (e.g.., very pale light blue not as attractive as more saturated pale green).
4. Assumed that saturation corresponds with percent identity (more saturated means higher percentage identity).
5. The slider was very useful and intuitive as a filtering mechanism. It would be nice to have image update in real-time instead of hitting ‘Update!’ button, but still ok.
6. Desire to filter based on outcome in addition to percent identity (i.e., show me only those where the left but not the right gene is found- how much red is there?).
7. Possibly place gene pair and genome names along axes as a guide or have a grid in the background. This is useful when browsing narrow columns.

6. Conclusion:

We believe that this visualization tool has accomplished our goal of clearly presenting our multi-dimensional dataset and providing insight into patterns within the data. This was only made possible through the coupling of the statistical information among genes with the evolutionary information among genomes. In our user study, the participants had no problems understanding what was presented, including which dimensions of the dataset were encoded and were available.

We do not anticipate the use of this visualization tool for purposes other than analysis of gene duplication data. This is not a shortcoming of the tool, but rather a consequence of the specificity of biological datasets. We believe that evolutionary biologists would be greatly interested in using this visualization tool to complement their existing techniques.

Future Directions

Although the existing visualization tool is already quite useful, there are several directions we would like to extend it. These include the following:

1. More advanced filtering:
   Support for a dropdown menu on the PrimaryWindow (i.e., Figure 5, right panel) that allows users to select which 'slice' of the dataset to view. They can choose one of the following: left, right, both, or outcomes (default- current setting).
2. More visual cues when clicking data points:
   Highlight of corresponding genome and gene pair location on the axes, in addition to displaying the current detailed tooltip.
3. Selection via genome or gene pair:
   a. Genome: Selection of a genome on the y-axis highlights the affected gene pairs on the x-axis.
   b. Gene Pair: Selection of a gene pair on the x-axis highlights the affected genomes on the y-axis.
4. Re-ordering columns:
   It may be useful to allow the user to re-order the columns, since the adjacent gene pairs are not necessarily related.
5. Advanced sequence analysis tools:
   For a selected data point, the ability to pull the associated nucleotide and protein sequences from GenBank then run them though a sequence alignment program.

References:

BLAST:

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410.

http://www.ncbi.nlm.nih.gov/BLAST/

TreeJuxtaposer:

Tamara Munzner, Francois Guimbretiere, Serdar Tasiran, Li Zhang, and Yunhong Zhou. *SIGGRAPH 2003*, published as ACM Transactions on Graphics 22(3), pages 453-462.

http://olduvai.sourceforge.net/tj/index.shtml

Treeview (now incorporated into MEGA):

MEGA: Molecular Evolutionary Genetics Analysis software for microcomputers. Comput Appl Biosci. 1994 Apr;10(2):189-91.