Genomic Spring-Synteny Visualization with IMAS

Chris D. Shaw
School of Interactive Arts and Technology, Simon Fraser University Surrey
shaw <AT> sfu.ca

Abstract

This paper presents a synteny visualization and analysis tool developed in connection with IMAS – the Interactive Multigenomic Analysis System. This visual analysis tool enables biologists to analyze the relationships among genomes of closely related organisms in terms of the locations of genes and clusters of genes. A biologist starts IMAS with the DNA sequence, uses BLAST to find similar genes in related sequences, and uses these similarity linkages to create an enhanced node-link diagram of syntenic sequences. We refer to this as SpringSynteny visualization, which is aimed at helping a biologist discover similar gene ordering relationships across species. The paper describes the techniques that are used to support synteny visualization, in terms of computation, visual design, and interaction design.

CR Categories and Subject Descriptors: J.3 [Life and Medical Sciences] Biology and genetics, I.3.3 [Computer Graphics]: Picture/Image Generation - Viewing Algorithms; I.3.6 [Computer Graphics]: Methodology and Techniques - Interaction Techniques.

Additional Keywords: Bioinformatics, Visual Analytics

1. Introduction

Discovering the structure and function of a newly-sequenced bacterial genome consumes a significant amount of effort. Once the DNA molecule has been sequenced into one or more contiguous stings of DNA letters (contigs), the biologist must then discover what genes are present in the organism, what they do, and how they interact with other parts of the organism. To infer genomic function bioinformatically, there are a number of sequence analysis tools that enable researchers to compare their own sequences to those that are present in the international public databases (NCBI, EMBL, DDBJ).

Although a certain amount of automated function determination may be possible, the results of such an automated process must necessarily rely on very strong sequence similarity to available database sequences, and on high-quality annotations of these similar sequences. Without manual examination of these analyses, there is the risk that the quality of the analysis will suffer due to the varying quality of the data available in the international databases, such as may occur when automatic annotations are built upon other automatic annotations.

The current tools to search public data are another source of difficulty for the biologist. The primary means of interaction with sequence data is for the biologist to maintain an ad hoc database of sequence information that is analyzed by cutting and pasting sequences into web sites that support sequence search and analysis. The advantage of this approach is that the biologist does not have to maintain a database of known sequences, but the disadvantages are that the data management task becomes quite significant, and the analyses themselves are offered on a remote server that can take anywhere from minutes to days to deliver results.

Our prior work introduced IMAS (Interactive Multigenomic Analysis System) [12], which combines sequence analysis tools such as Glimmer, Blast, and Clustal-W into a visualization and analysis workbench that enables rapid analysis of genomic sequences without the heretofore standard difficulties of having to manage a large number of printed outputs from websites.

The focus of this paper is a new analysis-based visualization technique that enables a biologist to visualize similarity of gene order across sequences of multiple related genomes. We refer to this display as *spring synteny visualization*, which is aimed at helping a biologist discover similar gene ordering relationships across species. The paper describes the techniques that are used to support synteny visualization, in terms of computation, visual design, and interaction design.

1.1 Biological Sequence Analysis

One significant job for a Biologist interested in the analysis of a newly-sequenced bacterial genome to do is to analyze the structure and function of the genes that are to be found in the organism. This entails a number of steps of analysis.

First, once the fragments of DNA that are detected by the basic DNA sequencing process must be strung together into one long contiguous sequence of DNA from the fragments derived from the sequencer.

Next, contig(s) are analyzed to find *genes*, which are stretches of contiguous DNA that are transcribed and translated by the organism's cellular machinery to express a string of Amino Acids (AA) that folds up into a protein molecule. The control of this transcription and translation process is often exercised by fragments of surrounding DNA. There are well-known programs for the discovery of genes in bacteria, such as Glimmer [5], which locate stretches of DNA that are bounded by in-phase *start* and *stop codons*. Each *codon* – a non-overlapping segment of 3NTs – encodes one of the 20 possible AAs in the protein sequence.

Next, the putative function of identified genes must be discovered by searching for similar DNA or AA sequences in the national public databases. The BLAST[1] tool is commonly used to search for sequences stored in a local database, or at the National Center for Biotechnology Information. BLAST takes as a query the NT or AA sequence of interest. Each sequence found by BLAST is accompanied by a pairwise alignment of the query sequence to the found (*subject*) sequence. The details of such an alignment are quite informative, and are generally indicative of the quality of the alignment and therefore the degree of similarity of function between the two sequences.

The fundamental purpose sequence alignment in this context is to enable the biologist to infer that the query gene and the subject have similar function because the sequences have sufficiently similar sequences. In about 50 percent of cases, the query coding sequence is so similar to a well-known gene that the analysis for it need go no further.

For the remaining putative genes, there may be sequences found by BLAST that the are sufficiently similar to warrant further investigation by aligning the putative DNA or AA sequence with each of the found sequences in a *multialignment*. Tools such as the Clustal-W multiple alignment program [14] are used to optimally align the multiple sequences found by BLAST into a single display of sequence similarity across all sequences. Such displays allow a biologist to observe regions where AAs are conserved across all of the sequences, and to observe if the query sequence fits into the conservation pattern.

Genes alone are not the only aspect of conserved DNA structure. That is, it is frequently the case that sets of genes are expressed together according to a coordinated unit of gene expression called an *operon* [2]. An operon consists of regulatory regions and protein-coding regions that are typically co-located along a contiguous stretch of DNA. Thus an observed consistency of the order of genes between one genome and another is another source of confirmation that putative genes in such a structure are real, even if an individual gene's function may itself be unknown.

As originally defined, *Synteny* is the condition where two genes are located on the same chromosome [11]. In this paper, we take another common usage of the term *synteny* to mean the possibly conserved location and ordering of genes between species. That is, the genes are conserved across species, and their neighborhoods are also conserved. Task we are supporting in this paper is the ability for a Biologist to observe instances of conserved or non-conserved gene neighborhoods across related species – that is, *synteny* from genome to genome.

2. Review of Syntenic Displays

The standard method for displaying syntenic relationships across genomes is to use the spatial layout of the genes along a chromosome of primary interest, and display similarity links to one or two secondary genomes of interest.

A bottom-up approach is to use the *Dot Plot* [6], which is a scatterplot technique that displays a NT sequence along the X compared to a NT sequence along the Y axis. A dot is drawn where there is sufficient sequence similarity between the two corresponding horizontal and vertical NT fragments. Identical sequences yield a dense display of dots along the positive diagonal. This allows one to infer similar structure, but does not typically encode meaning at the gene level.

GeneOrder 3.0 [4] extends this idea by comparing *genes* across genomes, enabling the user to infer structure more clearly. A small set of glyphs is drawn to display level of similarity.

In SyntenyVista [8], Hunt et al developed methods of displaying similar genes and other loci between chromosomes of human, rat, and mouse. The overview shows a primary organism with each chromosome assigned a unique color, and the secondary chromosomes are displayed with corresponding color patches. As single secondary chromosome may have multiple color patches from the primary genome, indicating that genes have moved from chromosome to chromosome. As the user drills down, pairs of chromosomes are lined up vertically, with line segments drawn between each to show similarity.

Cinteny [13] uses a similar color-coding approach to overview, and a line-segment linkage for similarity at higher levels of detail.

SynBrowse [10] presents genomes in parallel horizontal tracks, with similarity linkages from genome to genome drawn as line segments when zoomed out, and as pairs of line segments showing the range of similarity when zoomed in.

YGOB, the Yeast Gene Order Browser [3] displays similarities among yeast genomes, where the sequence is laid out horizontally with an equal-sized box per gene. The user focuses on a particular gene of interest and its 6 neighbors to the left and right are drawn in sequence. This establishes a grid in which each related organism is assigned its own track, and in each vertical slice is placed the corresponding gene to the column's primary gene. Only the primary organism's genome is presented in order, so similar secondary genes are colored according to their source chromosome.

Whole-genome browsers like the UCSC genome browser [9] and others tend to have a similar organizational idea. The primary sequence creates the order, and the secondary sequences are presented in the same order as the primary.

The advantage of this spatial layout is that the ordering and size relationships can be directly visualized (exception for YGOB). If the secondary sequence is presented in primary order, the assumption is that gene order is not as important as the direct examination of sequence alignment. For the Synteny viewers like SyntenyVista, Cinteny and SynBrowse, secondary order is important, so similarity linkage is drawn using color patches or line segments or a combination of both. For a pair of genomes, this works well, but it does not scale well to multiple genomes. Also, reversals in gene order cause an unfortunate crossing of many linkage line segments, which can be very hard to follow, except as an overview. The DotPlot handles this issue, but only works for a pair of genomes.

3. Syntenic Display as Graph View

Our synteny display, *SpringSynteny*, is aimed at helping a biologist discover similar gene ordering relationships across species. The basic idea of SpringSynteny is to dispense with the requirement to order genes along a line segment, and to instead display gene similarity as a graph, where nodes are genes and edges represent linkage between genes. Unlike existing synteny displays, SpringSynteny enables users to more directly observe reversals and other genome-scale reorderings of genes. In many cases, it is possible to see the differing order relationships among many sequences that have been aligned along the common backbone of the primary sequence under analysis, as in Figure 1.

The analytical process that supports SpringSynteny is as follows. First, the user runs IMAS and opens a *Primary Sequence* of DNA to analyze. The gene finder *Glimmer 3.02* can be run by the user to find putative genes in the Primary Sequence. In IMAS, Glimmer is run as an external process, and IMAS reads the Glimmer output and places the putative genes in the Genomic display. The user can analyze putative genes by running BLAST against sequence databases of interest. The resulting BLAST hits have the disadvantage of appearing out of their native sequence context.

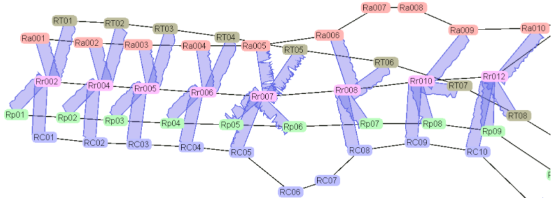

Figure 1. SpringSynteny of 5 Rickettsia Genomes. Each node is a gene. Common node color denotes common genome. The sawtooth links between genes indicate sequence similarity.

To better manage the sequence information that can be searched by BLAST, we have developed a subsystem in IMAS that manages the collection of published sequences from organisms related (or not) to the primary sequence. These *Secondary Sequences* are a collection of sequence data files that have been downloaded from NCBI and formatted into BLAST-readable NT and AA databases. Having a Secondary Sequence allows IMAS to reliably determine what neighbor genes may exist to genes found by BLAST. The user can then link the appropriate BLAST-aligned secondary genes to the primary sequence.

SpringSynteny is separate module that reads an IMAS linkage XML file on startup and draws relationships between the Primary genome and the secondary genomes. These linkage diagrams are displayed using a force-directed layout.

3.1 Edge Types

In the SpringSynteny linkage diagrams, there are two types of edges. The first is an edge that represents the location relationship between genes in a single organism. As shown in figure 1, these edges are drawn with a black line segment. For a single organism, the result would simply be a chain of nodes, labeled by their gene locus tag.

The second edge type is the *Similarity* link, which is created in IMAS when a BLAST hit in a secondary sequence is linked by the user to the primary gene. User-supplied linkage assures some degree of similarity quality and trust in the linkage.

Our first design iteration simply set the edge length to be equal for all edges. This had the advantage of implementation simplicity, but did not display useful information about gene location, and results in a confusing display where nodes tend to appear very close together and overlap their edges. Following an edge chain of neighbor links was somewhat difficult in this display.

The next design iteration for similarity links was to set edge lengths to a fixed value per organism. Thus links between primary sequence and Secondary Sequence A got a distance of 50, secondary B got 60, and so on. This yielded somewhat less node overlap.

For neighbor links, we next set edge length between each pair of genes (nodes) equal to the number of NTs between the left edges of each gene, thus directly displaying location information. This resulted in edge lengths being too long, so we scaled the length down to 1/20. This linear scale works well when there is strong similarity between gene positions, but as shown in figure 2, small changes in gene order in one genome yield graphs with acute angles at nodes with long neighbor edges.

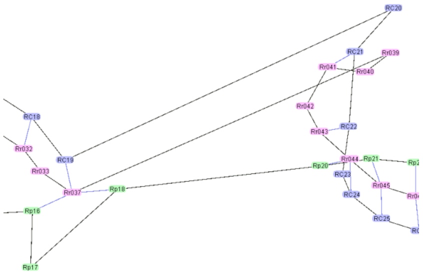

Figure 2. Linear scaling of spring length to NT length yields strong distortions.

The design next iteration on lengths was to devise an edge length that grows more slowly than a linear function of NT distance. We settled on

ln(length) * ln(length),

which clearly shows differences in magnitude, but does not grossly warp the graph when there is a long stretch on NTs in one genome that is not present at the same location in the others, as shown in figure 3.

The drawback of this technique is that a collection of small genes will yield a longer total link length than one gene over the same length of nucleotides.

3.2 Edge Rendering

Detailed sequence similarity between similar genes can be quite informative. We first attempted to display similarity by coloring the edge according to the overall percent identity between genes, but this approach yields a significant amount of the same color on each link, plus similarity color ranges tend to interfere cognitively with the node rendering colors. Nodes are colored according to organism, and edge colors interfere.

To better display the level of similarity, we developed an edge rendering technique that draws a line graph of percent identity between each linked primary and secondary gene. The source of identity information is a character-by-character scoring of similarity of the pairwise BLAST alignment. For an AA alignment, AA pairs are scored 3 at a time using a sliding non-overlapping scoring window. The result is a sequence of values between 0 and 1 that are then plotted along the edge linking two gene nodes in the graph. To score AA letter pairs, identical AAs score 1, and differing AAs scored by using the BLOSUM62 matrix [6]. The two AAs are used to look up the substitution log-odds probability *subst*, which is plugged into the formula

score = (subst – min) / (max – min).

This results in a vector of similarity scores along the sequence, which we plot on a similarity edge. The baseline of the graph is drawn along the line segment connecting the centers of the two nodes, and the line graph is plotted above the baseline at approximately twice the node height. Since the edge can be at any orientation, the line graph is rotated about the querying sequence node.

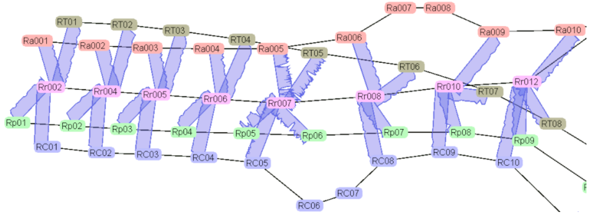

Figure 3. SpringSynteny of 5 Rickettsia Genomes: *R.Typhi* (RT - Olive) *R.Akari* (Ra –Orange), *R.Rickettsii* (Rr- Magenta), *R.Prowazekii* (Rp – Green), and *R.Conorii* (RC – Blue). The Black edges link Genes along a genome from left to right.

The first design iteration for edge rendering used a dark blue polyline without a baseline, which resulted in no discernable difference between high and low-quality alignments, although segments of poor identity in a high-quality alignment were visible. The next iteration used a polyline with baseline, which helped in understanding the height of the graph, since the baseline enables the user to estimate height. The addition of a topline helped further by allowing the range of height values to be directly perceived. However, edge rotation makes the use of line segments and polylines difficult to readily distinguish from the neighbor edges. We then drew the graph as a filled polygon with a solid blue color. This helps the user distinguish similarity edges from neighbor edges. The problem with solid color is that all similarity edges emerge from Primary sequence nodes, so there would be significant overdraw and visual confusion near primary nodes. We chose a transparency level of 30% for the graph polygon, which enables edge distinction, and supports the ability to read overlapping graphs at Primary nodes.

3.3 Interaction

The SpringSynteny visualization is implemented using the Prefuse toolkit [7], which supplies an interactive force-directed layout for graphs. When SpringSynteny starts up, the nodes start bunched together, and are separated by repeated iterations of the force-directed layout.

Users interact with the layout by dragging nodes to clarify the layout of graph nodes and edges. Users may drag the nodes in the display to enable detailed examination of individual parts of the sequence network. Users may zoom out to see an overview or zoom in to examine details. There is a dialog in the application that enables to user to select which secondary sequence to draw. At startup, all available sequences are linked and drawn together. Hiding a secondary sequence eliminates its nodes and similarity edges from the display, allowing the user to view a more limited set of genes and similarity links. This enables the user to examine a small number of sequences aligned with the primary sequence.

4. Case Study: *Rickettsia* Genomes

SpringSytnteny has been used to visualize the similarity relationships among 5 closely-related species of *Rickettsia: Rickettsia Rickettsii* (Rr)*, Rickettsia Prowazekii* (Rp). *Rickettsia Conorii* (RC)*, Rickettsia Typhi* (RT)*,* and *Rickettsia Akari* (Ra). These organisms are bacteria that are the cause Rocky Mountain Spotted Fever, Typhus, Mediterranean Spotted Fever, Murine Typhus, and Rickettsialpox, respectively. These organisms have about 1.2 million nucleotides each, and a significant amount of similarity across their genomes. We have analyzed the first 85000NT of *R.Rickettsii (Rr)* and displayed the genes occurring in first 85000NT of *Rp, RC, RT, and Ra.* In this subset, we have detected the closest alignment of each gene in *Rr* to any gene in the first 85000NT of the other sequences.

Figure 3 shows the similarity relationships among the first 8 genes of *Rr* and the most similar genes in the other 4 genomes. One can immediately see that *Ra* (orange nodes) and *RC* (blue nodes) each have a pair of genes that do not have an equivalent magenta *Rr* gene (Ra007, Ra008, RC06, RC07). One can also infer that these genes are insertions compared to *Rr*, since the neighbor edges bulge out from the parallel chains of the other genomes. Also, the set of genes linked to Rr007 have split into 2 in the case of Rp05 and Rp06, since there is good similarity to all genes linked to Rr007, and the similarity graphs linking Rr007 to Ra005 are similar to each other.

Another interesting point is that the similarity links between *Rr* and *RC* are very significant throughout. There are no points where the underlying sequence similarity drops significantly. *Rr*'s links to *Ra, RT,* and *Rp* are a little less similar, particularly at *Rr007*.

Figure 4 shows an overview of the first 80 putative genes of *Rr*. Also displayed are 70 *Ra* genes, 40 *Rp* genes, 65 *RC* genes and 56 *RT* genes. A number of linkage patterns are evident in the visualization.

First, there is a long stretch of *Ra* genes starting at *Ra014* and ending at *Ra064* that occur in reverse order to *Rr*. This is evident in the large loop structure created by the layout, in which the *Ra* strand runs parallel to the others in the upper left, then crosses over to the upper right and links in the opposite direction.

Second, there is a long stretch of *RT* genes that link to nothing in *Rr*. This is the olive set of links at the bottom of the image. There is good alignment from *Rt01* to *RT15*, then no link until a modest-quality link between *RT52* and *Rr074* on the lower right.

Third, there is an insertion of 6 genes starting with *Rr014* . These genes have no similar genes in the other sequence subsets.

Fourth, there are 5 genes present in *Rr* and *RC* that are not present in *Rp* (upper right) in Figure 4, preceded by 4 other *Rr* genes (5 *RC* genes) that also have no *Rp* equivalents.

This case study shows that the relative gene order relationships among these 5 genomes can be readily explored and analyzed.

5. Conclusions

We have introduced a visualization technique called SpringSynteny that enables users to visualize examine gene ordering and length relationships among a few related genomes. It takes the approach of transforming the numerical gene location into an ordinal value, and uses these ordering and similarity relationships between genes to display a node-link graph of similarity relationships. This transformation from numerical to ordinal allows the graph to be laid out with force-directed layout and enables the user to readily see gene insertion, deletion and reorderings in related genomes.

References

[1] Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman, "*Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*", Nucleic Acids Research, 1997, 25(17):3389-3402.

[2] JM Berg, JL Tymoczko, L Stryer, "*Biochemistry, 5th Ed.*", W.H. Freeman and Company, New York, 2002.

[3] Byrne, K.P., and K.H. Wolfe, "Visualizing syntenic relationships among the hemiascomycetes with the Yeast Genome Order Browser", Nucleic Acids Research, Vol 34 DB issue, 2006, pp. D452-D455.

[4] Celamkoti, S, S Kundeti, A Purkayashtha, R. Mazumderm, C. Buck, D. Seto, "GeneOrder3.0: Software for comparing the order of genes in pairs of small bacterial genomes", *BMC Bioinformatics*, 2004, 5:52.

[5] Delcher, AL, D Harmon, S Kasif, O White and SL Salzberg, "Improved microbial gene identification with GLIMMER", *Nuc.Acid.Res*, 27(23):4636-4641.

[6] Gibas, Cynthia, & Per Jambeck, Developing Bioinformatics Computer Skills, 2001, O'Reilly Press.

[7] J Heer, SK Card, JA Landay, "Prefuse: a toolkit for interactive information visualization" ACM CHI 2005, pp. 421-430.

[8] Hunt. E., N. Hanlon, D.P. Leader, H. Bryce, and A. F. Dominaczak, "The visual language of synteny", *OMICS*, 2004 Winter;8(4):289-305

[9] Karolchik, D, R Baertsch, M Diekhans, TS Furey, A Hinrichs, YT Lu, KM Roskin, M Schwartz, CW Sugnet, DJ Thomas, RJ Weber, D Haussler and WJ Kent. The UCSC Genome Browser Database. In *Nucleic Acids Research*, 2003, 31(1):51-54.

[10] Pan, X, L Stein, V Brendel, "SynBrowse: a synteny browser for comparative sequence analysis", *Bioinformatics*21(7):3461-3468.

[11] Renwick, J.H. Annu. Rev. Genet. 5, 81-120 (1972).

[12] Shaw, CD, G.A Dasch, and ME Eremeeva, IEEE Visual Analytics Science and Technology 2007, Sacramento, CA, Oct 30-Nov 1, 2007, pp 59-66.

[13] Sinha, AU, and J Meller, "Cinteny: flexible analysis and visualization of synteny and genome rearrangements in multiple organisms". *BMC Bioinformatics,* 2007, 8:82.

[14] Thompson, JD, DG Higgins, TJ Gibson. "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice". *Nuc.Acid.Res.*, Nov. 1994, 22(22):4673-4680.

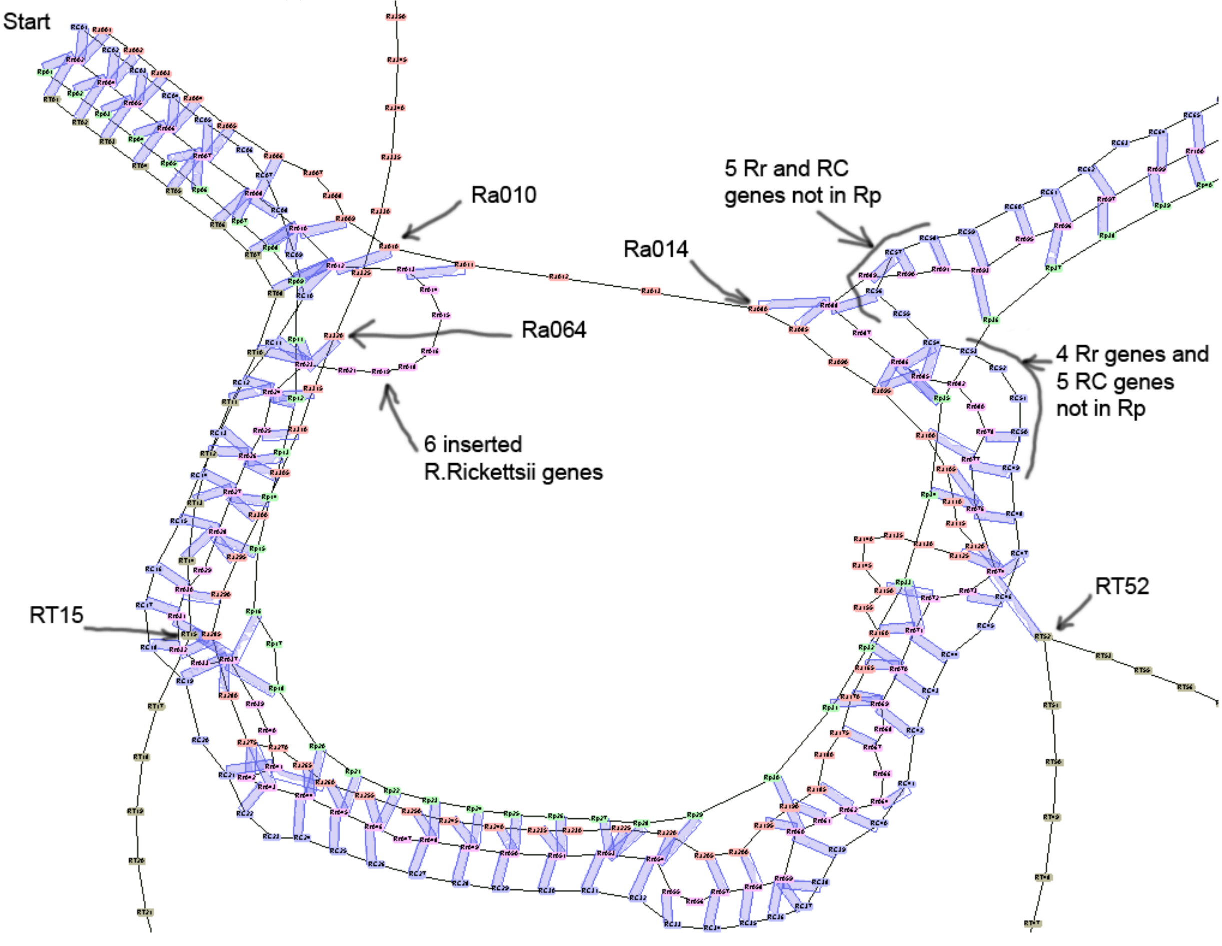

Figure 4. Overview of 4 sets of genes linked to Rr. The most salient two features are that *Ra* (orange) has an order reversal running from *Ra014* to *Ra064*. This causes the big loop structure in the visualization. The second feature is the loop below the diagram of *RT* genes, where there is sequence similarity to *RT* only up to *RT15.* Other image annotations are discussed in section 4.