Syntenic Reconstruction of Microbial Chromosomal Mutations

R.W. Cutler and A. Gambis

Departments of Biology and Computer Science
Bard College, Annandale NY 12504 USA cutler@bard.edu

ABSTRACT

In this paper, we propose a method to reconstruct the evolutionary history of chromosomal mutation events in microbial genomes using conservation of synteny. Starting from the most closely related species, as defined by the currently established taxonomy for microbial species, a minimal parsimony structure of chromosomal mutations can be built. This structure contains regions that are completely conserved, regions with some form of mutation, for example gene duplication or loss, translocation, or chromosomal inversion. From these structures an ancestral genome arrangement can be determined for clades of species which can then be aligned to resolve regional ambiguities. As a proof of concept, this technique was tested on the five closely related completely sequenced Chlamydia species. By sequencing regions spanning mutation events from closely related out-group species, in many cases it is possible to reconstruct the ancestral Chlamydia genome by progressively resolving the chromosomal ambiguities.

Keywords: Chromosomal Mutation, Synteny, Microbial Evolution.

1 INTRODUCTION

Completely sequenced genomes offer a starting point from which to study a myriad of fundamental biological processes from determining regulatory pathways, finding proteins with unique functions, and organizing protein families, to such areas as revealing evolutionary relationships and species modifications. Since the publication of the first completely sequenced prokaryote in 1995, the number of completely sequenced microbial genomes has doubled roughly every eleven months. As of August 2002, GenBank contained 81 sequenced and annotated genomes, 20 sequenced genomes currently being annotated, and 114 ongoing sequencing projects. In August of 2002 alone, thirteen new genomes were added to the publicly accessible databases.

The data from the large number of completely sequenced genomes and soon-to-be completely sequenced genomes offers a level of detail about chromosomal structure that had before now been unattainable. As increased information about the structure and composition of Microbial Genomes is accumulated, it will become increasingly feasible to determine the evolutionary history of disparate lineages of microbial genomes. Regions between two microbial genomes that have a loss of synteny reveal a mutation event that occurred at some point in the history of one or both of the organisms. We present a method to reconstruct chromosomal mutation ambiguities which could thereby increase the evolutionary depth which could be resolved, or alternatively this reconstruction allows a view of the least common ancestor of modern lineages. One way to resolve specific ambiguities is to choose a nearby outgroup species and sequence the boundary region of the ambiguity. Other methods to align the genomes of species such as Clusters of Orthologous genes (COG) [1,2] and MUMmer [3] have focused on building classes of proteins with shared functionality. Although this approach is similar, oftentimes multiple regions within the same genome are aligned, and these approaches will categorize genome segments that share functionality but are not necessarily descended from a common ancestor. Likewise proteins from a common ancestor that have diverged in function will not be aligned.

To unambiguously align the species, we performed an all against all blast search. Only those sequences, which we call anchors, that uniquely align between different genomes are grouped. Pairs of these adjacent proteins then mark a syntenically conserved region. The proteins between these anchors are then aligned increasing the significance due to the reduced sample space. Finally boundaries between the contiguous regions are categorized and the regional structures are determined.

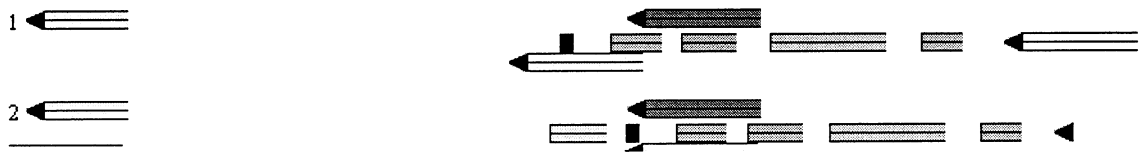

Figure 1: A ribbon flanked by two breaks. The ribbon extends from SSU1011 (turquoise) to SSU1013 (red). Two breaks occur on either side of the inversion. The figure above illustrates the case of a 'perfect invert' in which an inversion occurred between the blue and yellow proteins in two of the genomes lineages.

2 METHOD

To syntenically reconstruct the evolutionary history of the *Chlamydia* family, we started with a 5-way genome alignment, where conserved syntenic units known as ribbons (sets of unique proteins conserved across in this case all 5 species) and breaks were identified. Synteny, also known as genome collinearity, refers to the presence of 2 or more loci on the same chromosome that may or may not be linked closely. A ribbon is a large syntenic region containing at least two contiguous conserved proteins in all five species.

The method used to create a ribbon involved aligning unique proteins in the five species. Syntenic Units IDs called SSU (similar to COGs but having no functional constraints) were assigned to unique protein families in the five *Chlamydia* species. Under the condition that all successive unique SSU proteins are conserved, the ribbon is continuously extended until the condition is broken. Figure 1 shows an example of a ribbon region. Using a set of 455 unique proteins, a total of 65 ribbons were identified. The 145 identified breaks correspond to regions neighboring the ribbons, where synteny was disrupted. The ribbons and breaks were formed under the Synteny Database Creator. This is a set of Perl routines we developed to extract gene information from GenBank files, perform blast operations to identify unique proteins, generate anchor proteins, and print records of the ribbon and break information in htm and txt format for easy examination. As sequences are added to the database, ancestral sequence positions are automatically reconstructed, increasing the evolutionary depth the analysis can extend. In addition, ambiguous mutational event sequences are presented and can be corrected with sequences from appropriate outgroup species.

2.1 Container Classes

Using our dataset of ribbons and break locations, we then proceeded to the reconstruction of chromosomal mutation events by identifying inversions. These inversions can then be "reverted" to the ancestral form thereby forming larger contiguous ribbon regions and reducing the number of mutation events needing to be reconstructed. The technique used in this reconstruction is a bottom to top approach: determine the smallest unambiguous inversions, resolve the inversions, form new larger conserved ribbon regions and repeat. The concept implemented to deal with these ambiguities was to identify a specific mutational event and 'correct' the event by creating a new ribbon region spanning the known mutation event. Eventually as each inversion is corrected, small ribbons fuse to make larger ribbons. These larger ribbons can then be used to help reconstruct even larger or more ancestral chromosomal events. Extending this methodology backward could eventually reconstruct the ancestral genome of the Chlamydia clade. Of the many types of chromosomal mutations that can occur, several have a clear structure which we have characterized and will mention below.

2.2 Chromosomal Inversions

We can use Figure 1 as an illustration of the reconstruction of an inversion. In order to resolve this inversion, we could re-invert the turquoise red (TR) region in *C. trachomatis* and *C. muridarum* to obtain the Blue-red-turquoise-yellow (BRTY) configuration of genes found in the C. pneumoniae genomes. However, during the reconstruction, we do not have any preexisting information about the ancestral genome and usually there are several possible

"optimal" solutions. In this case, if we did not know that the ancestral ribbon was (BRTY), the ancestral chromosomal positioning could have either been (BRTY) by inverting the RT ribbon in *C. trachomatis* and *C. muridarum* or (BTRY) by inverting the RT ribbon in the C. pneumoniae species. To resolve the ambiguity it is necessary to examine the break region sequence in a closely related outgroup species. For this set of species, an appropriate outgroup species is the Parachlamydiaceae lineage. When aligned, *C. trachomatis* and *C. muridarum* show no breaks between shared unique proteins, and therefore there have not been any inversions or transpositions since speciation. Similarly, between the *C. pneumoniae* genomes there is a break count of 0, and they too have not undergone major chromosomal mutation events characterized by inversions or translocations. However, *C. trachomatis* and *C. muridarum*, when aligned to successive *C. pneunomaie* genomes, revealed respectively a break count ranging from 145 to 148 and a break count ranging from 149 to 157. The genome wide alignment found from this break data is shown in Figure 2 below.

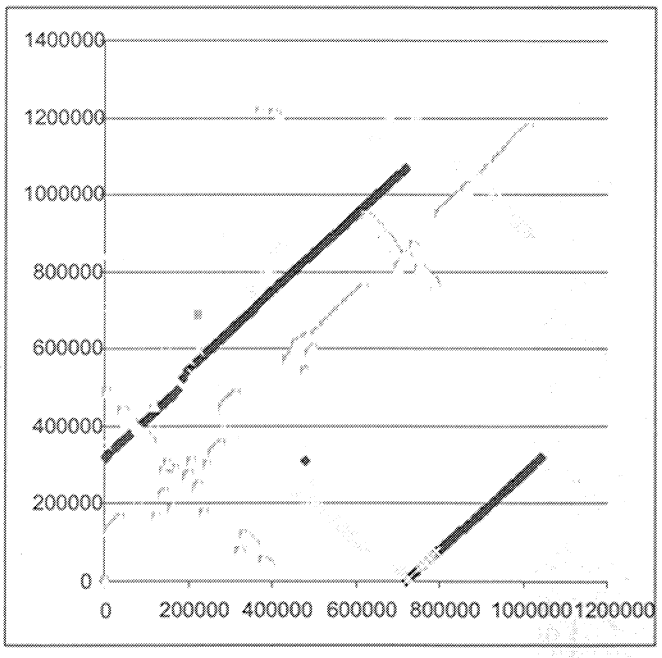

Figure 2. The ribbon alignment of the complete genomes of *C. pneunomaie* (light blue, yellow, and red) and *C. trachomatis* (navy blue) against *C. muridarum.. The Genbank genome sequence for C. pneumonaie AR39 was stored in the reverse direction with respect to the other two* C. pneumonaie species which accounts for the mirrored positioning of the light blue proteins.

3 CONCLUSIONS

In modern phylogenetic research, the basic unit of comparison is no longer the nucleotide in the gene but the gene in the genome. The phylogenetic reconstruction we have described here emphasizes gene order structures rather than individual gene sequences. After the complete genome sequences of the two chlamydian parasites – *C. trachomatis* and *C. pneumonia* were published [4], novel phylogenetic schemes were devised in an attempt to understand genome expansion, rearrangement, and transfer of genes between dipartite genomes. The 'proposed reconstruction model' attempts to tackle those questions. While the idea of reordering chromosomal mutation events in a parsimonious fashion is crucial to the model, it is important to also emphasize the syntenic extension aspect, where rearrangements once resolved are appended to existing ribbons. The rationale behind this is to work from bottom to top; in other words, we first look at low-order chromosomal events, resolve them, and then high-order (and often more important in terms of speciation) chromosomal events become apparent.

The ability to resolve these types of events will increase as more microbial sequence data becomes available. The methodology and results presented here can currently be implemented, but in a few years, this type of analysis will probably be of fundamental importance due to the ability to use all of the genomic data available to make phylogenetic predictions.

REFERENCES

[1] Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. **Science** 1997 Oct 24;278(5338):631-7.

[2] Tatusov RL, etal. The COG database: new developments in phylogenetic classification of proteins from complete genomes. **Nucleic Acids Res** 2001, 29(1):22-8.

[3] Delcher, A.L., Phillippy, A., Carlton, J., and Salzberg, S.L. Fast algorithms for large-scale genome alignment and comparison **Nucleic Acids Research** 2002, 30(11) 2478-2483.

[4] Kalman,S. etal. Comparative genomes of C. pneumoniae and C. trachomatis. **Nature Genetics** 1999, 21(4), 385-389