

Phylogenetics of Whole Genomes Using a Structured Subsequence Database Approach

Zayed Albertyn, Poh Yang Ming, Ali Reza Zamli, Ching Soo Meng and Robert G. Hercus

SynaCompare™, an application built upon SynaBASE™'s unique structured network database architecture, has been previously used for rapid whole genome alignments based on structured subsequences. SynaCompare was used to find all homologies across a dataset of nine Enterobacteria genomes in less than five minutes by deriving all multiple alignment anchors between them. A phylogenetic tree of the SynaCompare results was constructed to visualize the relationships between these taxa. The inferred whole genome phylogenies are consistent with previously published work by multiple genome alignment tools and confirm the accuracy and validity of SynaCompare's whole genome alignments. These results demonstrate SynaCompare's capabilities and potential in facilitating analysis of multiple genomes for their phylogenetic relationships in SynaBASE's high throughput sequence analysis environment.

Introduction

Various sequence alignment packages are available to align large eukaryote chromosomes or bacterial genomes to each other. The BLAST family of programs and FASTA are widely used for local alignments with modifications for larger input sequences [1-2]. These algorithms are generally quite sensitive but require expensive cluster hardware for optimal performance in high-throughput comparative genomics.

Pairwise sequence comparisons provide the basis for aligning multiple sets of genome sequences. Significant matches between two sequences can be used as a basis for identifying regions of homology and potential synteny between related species. The majority of all multiple genome alignment programs require anchors from pairwise comparisons to seed subsequent alignments. Patternhunter, LAGAN and Mummer are more recent examples of programs that

efficiently align mega base sequences to each other with moderate CPU requirements e.g. small servers or desktop workstations [3-5].

The objective of this study was to demonstrate that a database that stores genomic data as subsequences, in a novel structured network database - SynaBASE - can be used for very rapid and scalable chromosomal and genome level comparisons. SynaBASE automatically calculates and stores associations between sequences based on shared subsequences [6]. As unique subsequences are only stored once, SynaBASE becomes more efficient as more data is added. Whole genomes or multiples thereof can be stored in SynaBASE and queried in a fraction of the time taken by conventional methods.

Working in an environment that allows persistent storage of data while permitting rapid comparisons on the genome scale makes SynaBASE ideally suited for high-throughput and large scale multi-genome comparisons. Pattern-based alignments computed by SynaBASE serve as anchors that may be used for more detailed analyses such as multiple genome alignments and phylogenetic reconstruction. SynaCompare calculates pairwise sequence comparisons using subsequences from the data stored in SynaBASE.

The aim of this study was to explore the full capabilities of SynaCompare and qualify results by identifying homologous regions in genomes and representing them using a phylogenetic tree. A dataset of nine Enterobacteria genomes was used to demonstrate the arrangement of SynaCompare alignment anchors across multiple species.

Materials and Methods

Alignment of 9 Enterobacteria Genomes

Genome	Length (bp)	Reference
<i>E. coli</i> K12 MG1655	4,639,221	Blattner et al. 1997
<i>E. coli</i> O157:H7 EDL933	5,524,971	Perna et al. 2001
<i>E. coli</i> O157:H7 VT-2 Sakai	5,498,450	Hayashi et al. 2001
<i>E. coli</i> CFT073	5,231,428	Welch et al. 2002
<i>S. flexneri</i> 2A 2457T	4,599,354	Wei et al. 2003
<i>S. flexneri</i> 2A	4,607,203	Jin et al. 2002
<i>S. enterica</i> Typhimurium LT2	4,857,432	McClelland et al. 2001
<i>S. enterica</i> Typhi CT18	4,809,037	Parkhill et al. 2001
<i>S. enterica</i> Typhi Ty2	4,791,961	Deng et al. 2003

Table 1: Nine Enterobacteria genomes used for sequence comparisons with SynaCompare * See Reference [7] for these genome references.

The Enterobacteria genomes were chosen as a test dataset as they are known to have undergone significant genome rearrangements and contain numerous repetitive elements. Table 1 shows the bacterial genome used in this study. A SynaBASE of these 9 bacterial genomes was built to facilitate all genome comparisons.

A one versus all comparison with SynaCompare was done for each bacterial genome to find all local alignments

greater than or equal to 60 base pairs within a window size of 500 base pairs using a minimum match of 13 continuous bases. Alignments were calculated for each pair in the Enterobacteria genome set using SynaCompare. Distance measures were calculated as follows:

$$D = 1 - 2A/(Q+T)$$

Where A is the total pairwise alignment length, Q is the query length and T is the length of the target sequence - all obtainable from the SynaCompare application. Self distances were normalized to 0. The NeighborNet algorithm [8] was used to construct a phylogenetic tree from the pairwise genome distances.

Results

The time taken to build a SynaBASE of the 9 Enterobacteria genomes was 9 minutes and 28 seconds. However, this step is not required if a Bacterial SynaBASE already exists. Alignment of all the Enterobacteria genomes against each other was completed in 4 minutes and 12 seconds on a single Intel Itanium 1.3 GHz CPU processor with 1GB of RAM.

The ability to compare all genomes to each other allowed evaluation of the phylogenetic relationship of the 9 Enterobacteria genomes. The results of a one versus all "MultiCompare" against all these genomes were used for calculating distance measures between all pairs of aligned genomes. The distances were subsequently used to construct a phylogenetic tree showing the relationships between these nine Enterobacteria genomes (see Figure 2 on the next page).

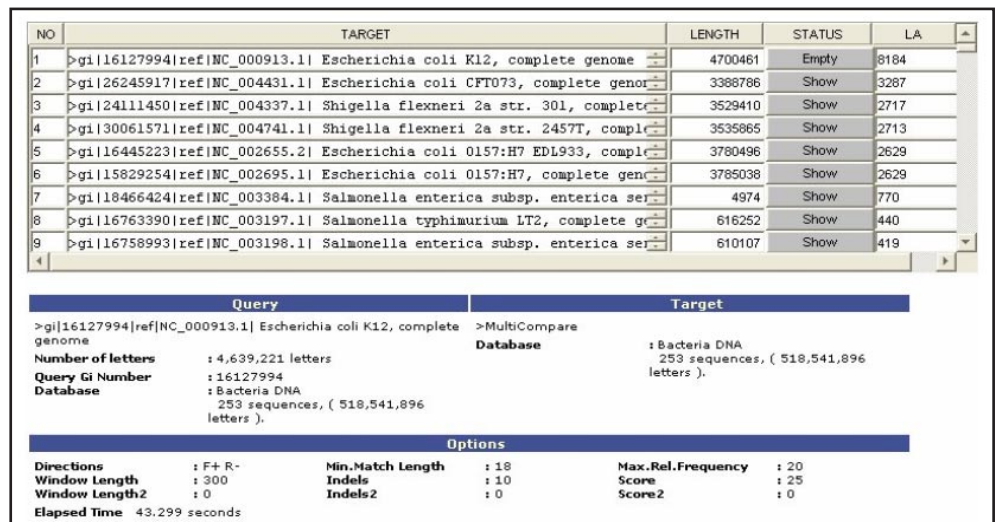


Figure 1: SynaCompare result for E. coli K12 compared to all the Enterobacteria genomes listed in Table 1. SynaCompare alignment time was 43 seconds for this comparison.

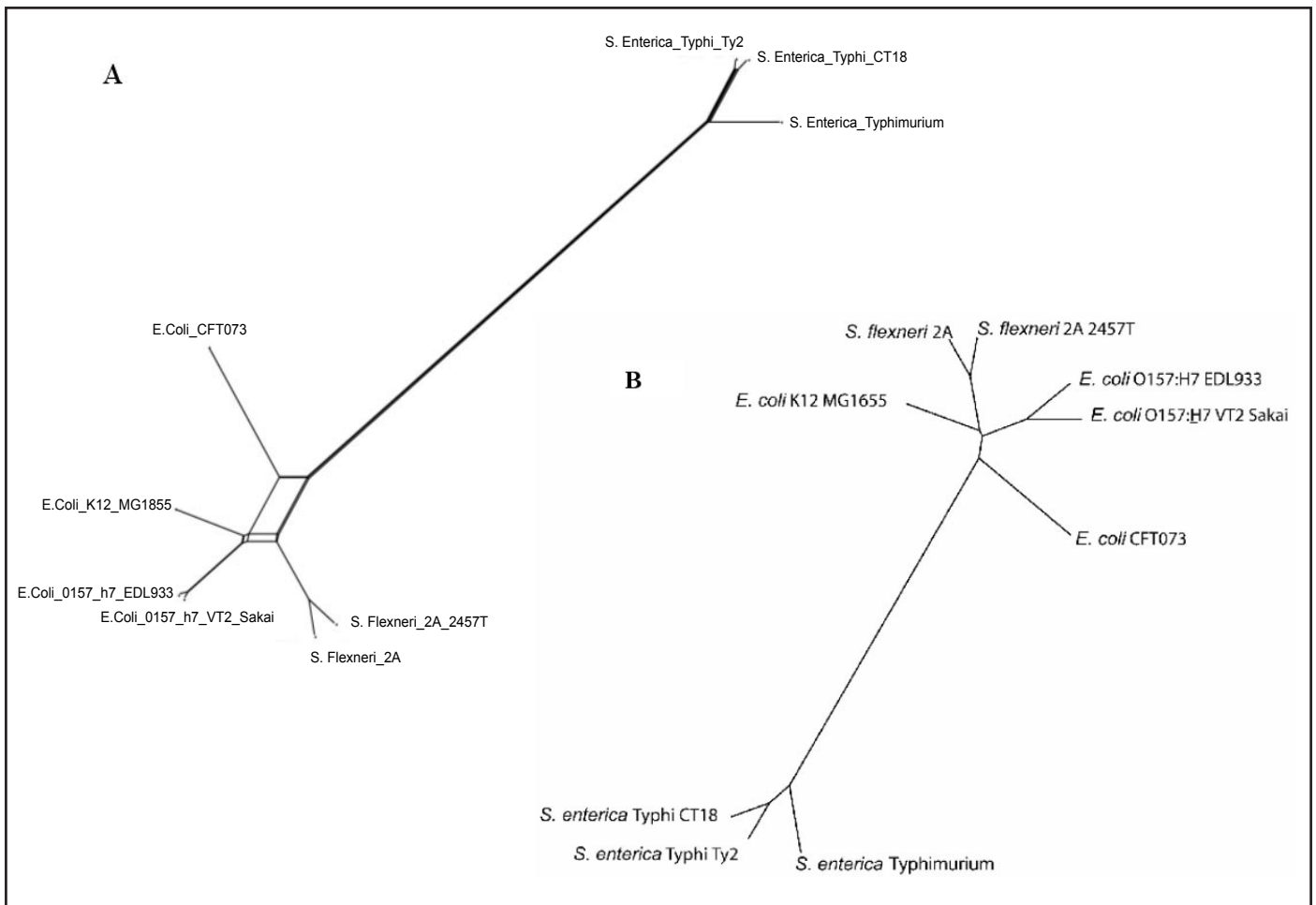


Figure 2: Unrooted phylogenetic trees relating the nine Enterobacteria genomes in Table 1. The tree in (A) was calculated with the NeighborNet algorithm using SynaCompare alignments and (B) shows a similar Neighbour joining tree produce by the Mauve multiple genome alignment tool (reproduced from Darling et al).

The tree results are consistent with those obtained in a previous study involving the multiple alignments of these 9 Enterobacteria genomes using the Mauve algorithm [7].

Conclusion

Rapid alignment with SynaCompare for whole genomes has been extended to identify the phylogenetic relationships between related Enterobacteria species and the results have been shown to correlate with previously published works. Current approaches are limited in that they require expensive computational resources to conduct exhaustive comparisons of genomes where the focus is on smaller isolated regions. With the ability to store and rapidly analyze datasets of whole genomes for local and global similarities, SynaBASE is ideally suited for building analysis pipelines to investigate sequence rearrangements and evolution based on multiple sequence alignment of whole genomes.

The capabilities of using SynaBASE for comparative genomics are not limited to pairwise alignments. In

this paper we have successfully applied our methods to analyze multiple Enterobacteria genomes by deriving all multiple alignment anchors between them and calculating a phylogenetic tree to visualize the relationships between these taxa.

References

1. Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ (1990). Basic local alignment search tool. *J Mol Biol* 215,403-10.
2. Pearson WR and Lipman DJ (1988). Improved tools for biological sequence comparison. *Proc Natl Acad Sci* 85, 2444-8.
3. B. Ma, J. Tromp, and M. Li. Super Seeds for Faster and More Sensitive Homology Search. *Bioinformatics* 18: 440-445. 2002.
4. Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Green ED, Sidow A, Batzoglou S and NISC Comparative Sequencing Program (2003). LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* 13,721-31
5. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL (2004). Versatile and open software for comparing large genomes. *Genome Biol.* 5,R12.
6. Albertyn, ZI, Wong CS, Tay L and Hercus, G (2004). A new approach to genome-wide annotation based upon calculation of significance from a structured pattern database. Available from www.synamatix.com
7. Darling AC, Mau B, Blattner FR, Perna NT (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* 14,1394-403.
8. Bryant D and Moulton V. (2004). Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Mol Biol Evol.* 21,255-65.

This Application Note is for distribution to Synamatix members, associate members and mailing list subscribers only. The contents are provided for personal, non-commercial purposes only and are protected by various national and international intellectual property laws, conventions and treaties. All title and intellectual property rights in and to Synamatix, SynaBASE, SynaCompare, and SynaSuite and the accompanying printed materials are owned by Synamatix sdn bhd. Other trademarks or names are used only in an editorial fashion and to the benefit of the respective trademark owner with no intention of the infringement of the trademark. All trademarks or service marks are the property of their respective owners.