Comparative evaluation of clustering methods to better understand bacterial evolution

Lydia Paradiso, Ananth Kalyanaraman, Shira Broschat

REU in Plant Genomics and Biotechnology, EECS

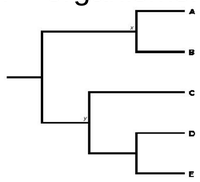
Washington State University — World Class. Face to Face.

NJIT New Jersey's Science & Technology University

Introduction

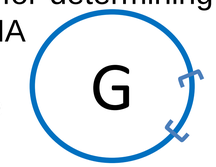
Today, thousands of genomes are available, and tools are constantly being developed with faster and more effective ways to mine this wealth of information.

Phylogeny is the study of the evolutionary relationship among groups of organisms. A phylogenetic tree is a graphical representation of this relationship. Evolutionary relationships can be Determined through genome comparison and analysis.

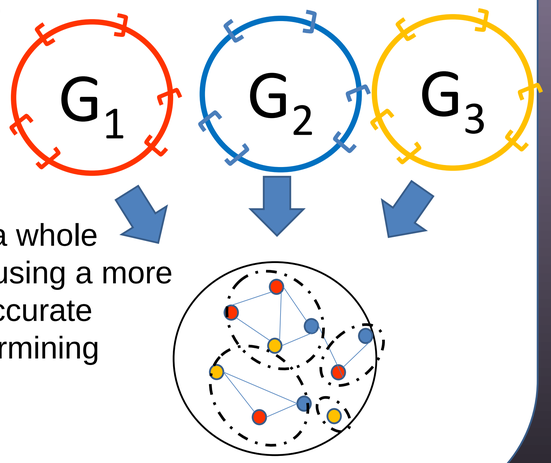


The understanding of bacterial evolution is important for several reasons, with one of the most important applications being drug design and vaccine development.

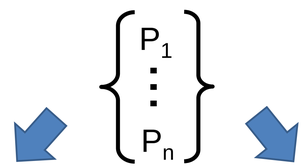
The current predominant method for determining microbial evolution is through 16s RNA profiling. The 16s RNA gene is a conserved gene for which the degree of conservation is hypothesized to reflect the degree of evolutionary relationship. Comparison of the gene across several organisms can lead to a proposed phylogenetic tree. However, this method only considers one gene out of many in its attempt to infer the evolutionary relationship among bacterial genomes.



We propose a method of inferring microbial evolution by comparing the species at a whole genome level, using a more scalable and accurate method of determining clusters.



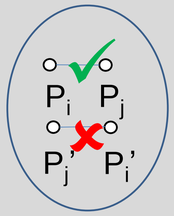
Methods



Current method

"self" BLAST
Input n protein sequences
$\{P_1 \ldots P_n\}$ vs $\{P_1 \ldots P_n\}$

-suboptimal alignment heuristic
-limited scalability: long runtime, i/o intensive, hard to parallelize

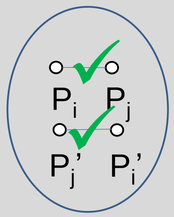


Proposed method

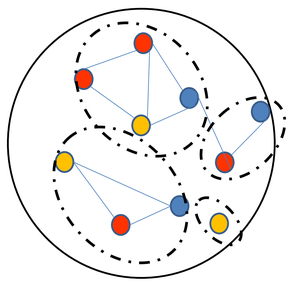
Filter pairs based on exact matches

Smith Waterman dynamic programming alignment

-optimal alignment heuristic
-filtering for best pairs saves time



CLUSTERING



Results

Four bacterial genomes were used as the dataset, *Bartonella quintana*, *Bartonella henselae*, *Rickettsia typhi,* and *Brucella melitensis*. Altogether, they contain 6602 protein sequences.

The self BLAST method took significantly longer to obtain results and was more i/o intensive. The BLAST output used to determine clusters was 18 GB, while the proposed method of filtering pairs prior to alignment reduced this size significantly.

Our proposed method produced 4.5x more edges than with self BLAST and placed about 50% of the protein sequences in clusters as opposed to about 25% with self BLAST.

Overall, the proposed method formed almost twice as many clusters, providing a larger dataset to determine evolutionary lineage.

Method	Runtime	Edges	Vertices	Clusters
Self BLAST	5h22m	1952	1538	703
S.W. w/ filter	5min	8763	3827	1253

Conclusions

Our proposed method for inferring bacterial evolution is more accurate, more efficient, and faster than the current 'self' BLAST method. Unlike BLAST, Smith Waterman alignment guarantees finding the best possible match. This is evident in the greater number of edges and vertices produced by the proposed method.

Considering the entire genomes of organisms when attempting to reconstruct evolutionary lineages is a more accurate method than considering only a small part of the genome, and our proposed method makes it possible to obtain the most accurate results.

Acknowledgements

This REU project was supported by the National Science Foundation grant DBI-1156880.

References

C. Wu,and A. Kalyanaraman, "An efficient parallel approach for identifying protein families in large-scale metagenomic data sets," *Proc. ACM/IEEE conference on Supercomputing (**SC'08**)*, Austin, TX, November 15-21, 2008, pp. 1-10, ISBN 978-1-4244-2835-9, IEEE Press, Piscataway, NJ, USA.

S.F. Altschul, T.L. Madden, A.A. Schäffer, 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*, vol. 25, 1997, pp. 3389 –3402.

ftp.ncbi.nlm.nih.gov/genomes/Bacteria