David Bass
Professor Martin Wu

Spring 2022 semester research report

Abstract

In our research, we studied synteny blocks in a sample of *E. coli* genomes in preparation for phylogenetic inference. We used SibeliaZ (Minkin and Medvedev 2020) and MAFFT (Katoh and Standley 2013) to detect, extract, and align synteny blocks common to the chromosomes of 2,176 *E. coli* genomes obtained from the NCBI Reference Sequence Database (Tatusova et al. 2016). From the aligned synteny blocks, we constructed "core genomes," FASTA files for each sampled organism containing only the portions of its genome that appear in a synteny block. We determined SibeliaZ parameters and other synteny block conditions under which core genome length and quality is maximized. The final product, two collections of 2,176 core genomes, each prepared with slightly different specification, is prepared for use with phylogenetic inference software RAxML (Stamatakis 2014).

Introduction

The task of describing macroscale evolution in microorganisms is severely inhibited by the lack of microbial genomes in the fossil record (Knoll 2015). Patterns of evolution can be inferred in organisms by the presence of synteny blocks, which are large portions of the genome that are found replicated in other genomes or replicated in the same genome (Liu et al. 2018). In order to attempt to detect evidence of pulsed evolution, George Gaylord Simpson's model of evolution as "stasis interrupted by pulses of rapid change" (Landis and Schraiber 2017), in *E. coli*, we assembled core genomes from synteny blocks found in a sample of 2,176 *E. coli* genomes obtained from the NCBI Reference Sequence Database. We define a "core genome" as the subset of an organism's genome that appears in a given set of synteny blocks. We are especially interested in non-paralogous synteny blocks, i.e. those that appear no more than once in each genome in our sample. We call such synteny blocks "non-repetitive." We define the span of a synteny block to be the number of unique genomes in the sample in which it appears. We define the non-unique span of a synteny block to be the total number of times that it appears in the genome; the span of a synteny block equals its non-unique span if and only if it is non-repetitive. We define a full-span synteny block to be one that appears at least once in each genome in the sample. We define an ideal synteny block to be one that is both non-repetitive and full-span. Note that the definition and detection of synteny blocks is dependent on software and parameters used.

Unsurprisingly, very few synteny blocks are ideal; providing more lenient requirements for core genome construction, e.g. accepting non-repetitive blocks whose span is only 2,150 rather than 2,176, is necessary to obtain meaningfully long core genomes. We call synteny blocks that are "close enough" to being ideal according to a given threshold "sufficient."

Methods

We downloaded 2,179 complete *E. coli* genomes from the NCBI Reference Sequence Database as FASTA files. We removed all non-chromosomal sequences from each file because plasmids often originate from horizontal gene transfer, the existence of which can only inhibit phylogenetic inference. We also removed three genomes from the sample that contained drastically smaller chromosomes than the other 2,176 and were thus preventing SibeliaZ from detecting full-span synteny blocks. We used SibeliaZ on 40 CPU cores on UVA Research Computing's Rivanna supercomputer (https://rc.virginia.edu) with k=15 (as recommended for bacterial genomes) and varying values for m and a. We initially used maf2synteny (Kolmogorov et al. 2018) to reformat the GFF file output from SibeliaZ for readability, but ultimately opted to extract synteny blocks from the sample's FASTA files with a script reading coordinates directly from the GFF file. We created one FASTA file per synteny block, with lines in each file alternating between FASTA headers and synteny block sequences. We applied MAFFT with default settings to each of these files in order to align the appearances of each synteny block. Finally, we appended the appearances of each synteny block from each genome in order to form one core genome per initial genome. We did this twice, one time including only ideal synteny blocks and a second time deeming non-repetitive synteny blocks with span at least 2150 to be sufficient for inclusion in the core genomes.

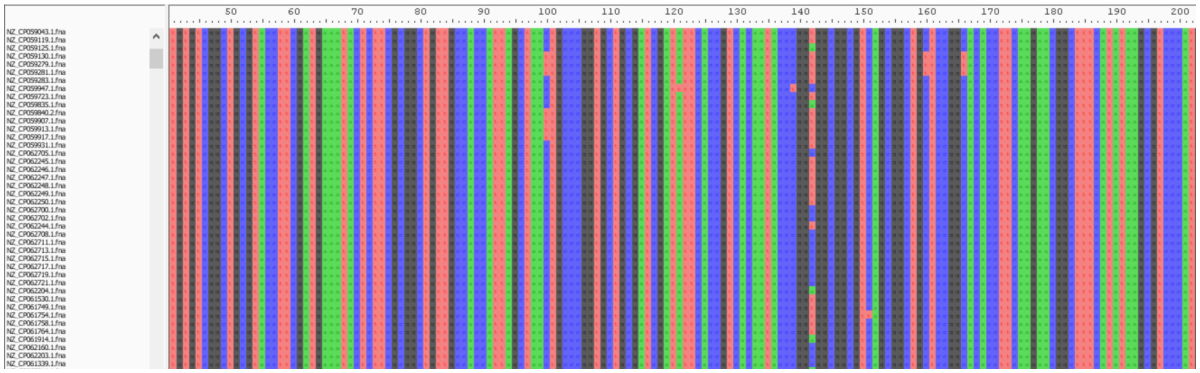

Figure 1: Alignment of appearances of an ideal synteny block with MAFFT, visualized in AliView (Larsson 2014). Note that this figure only shows some of the 2,176 appearances of the synteny block and only a fraction of its total length, between ~40 and ~200 base pairs. The appearances of the synteny block are extremely similar.

It is important to note that the core genomes are not in the same order as the initial genomes. When the synteny blocks are pieced together into core genomes, they are ordered according to SibeliaZ's ordering of the synteny blocks. For example, a synteny block may appear after all other synteny blocks in a genome, but appear as the first synteny block in the core genome because it appears before all other blocks in another genome, and thus gets ordered as the "first" synteny block. SibeliaZ does not report synteny blocks that overlap within a given genome.

All code used in this project can be found at our GitHub repository (https://github.com/DavidB256/Wu-Lab-Scripting). ZIP files containing the core genomes are

150 MB and 991 MB for the sufficiency thresholds of span 2176 and span 2150, respectively, so are too large to upload to the GitHub repository, but can be transferred via other conventional means.

Results

We find that optimal values for sufficiency threshold, core genome length, and block count are maximized with SibeliaZ with default parameters, except for k=15, a=4,352 (twice 2,176), and m=200, and sufficiency thresholds of 2,150 and 2,176. To clarify, the core genomes in the instance with threshold=2,150 are constructed from non-repetitive synteny blocks with span greater than or equal to 2,150, while the core genomes in the instance with threshold=2,176 are constructed only from ideal synteny blocks. We determine optimality first based on threshold, then based on core genome length, and finally based on block count, with greater values for each quantity being better. Each initial genome is approximately five million base pairs in length, so our assembled core genomes, which feature lengths of 1,479,347 and 223,751 base pairs at threshold=2,150 and threshold=2,176, respectively, cover approximately 29.6% and 4.5% of the initial genomes, respectively.

Increasing SibeliaZ's m parameter causes significant increases in runtime, but produces better results, up to a point. Thus, we chose m=200, as opposed to the default value of m=50, but not m=500, which caused significant decreases in block counts at all threshold levels.

As the lengths of appearances of a synteny block vary slightly and choosing a threshold lower than the sample size will cause the distribution of synteny block appearances to be non-uniform across genomes, core genome length inevitably varies among organisms. Thus, in Table 1 below, we simply give the length of the first core genome, as chosen from an arbitrary ordering of the genomes in the sample, as an estimate of the average core genome length.

threshold	m	block count	core genome length
2000	25	2841	1482518
2000	50	2603	1520341
2000	100	2357	1582569
2000	200	2019	1641617
2000	500	1612	1434069
2100	25	2538	1410927
2100	50	2346	1445045
2100	100	2131	1497118
2100	200	1818	1543882
2100	500	1484	1339615
2150	25	2325	1369642
2150	50	2170	1402064
2150	100	1978	1448092
2150	200	1679	1479347
2150	500	1343	1242271
2176	25	241	220766
2176	50	224	220503
2176	100	207	225838
2176	200	177	223751
2176	500	121	148093

Table 1: Table showing synteny block counts and core genome lengths at different levels of threshold and m parameter in SibeliaZ. Synteny blocks were detected by SibeliaZ with default parameters, except for k=15, a=4,352,

and m as given in each row. The highlighted rows represent parameter values for which core genomes were fully assembled Visualized in Google Sheets.

The two compilations of 2,176 core genomes, one at the stringent threshold of 2,176 and one at the moderate threshold of 2,150, represent a significant amount of genomic data processing. With the majorities of initial genomes that do not appear in sufficient synteny blocks removed, we will be able to perform future work predicting macroscale evolution patterns in *E. coli* with phylogenetic inference with software like RAxML.

Works Cited

- Landis, M. J., & Schraiber, J. G. (2017). Pulsed evolution shaped modern vertebrate body sizes. Proceedings of the National Academy of Sciences of the United States of America, 114(50), 13224–13229. https://doi.org/10.1073/pnas.1710920114
- Larsson, A. (2014). AliView: a fast and lightweight alignment viewer and editor for large data sets. Bioinformatics30(22): 3276-3278. http://dx.doi.org/10.1093/bioinformatics/btu531
- Liu, D., Hunt, M. & Tsai, I.J. Inferring synteny between genome assemblies: a systematic evaluation. BMC Bioinformatics 19, 26 (2018). https://doi.org/10.1186/s12859-018-2026-4
- Minkin, I., Medvedev, P. Scalable multiple whole-genome alignment and locally collinear block construction with SibeliaZ. Nat Commun 11, 6327 (2020). https://doi.org/10.1038/s41467-020-19777-8
- Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution, 30(4), 772–780. https://doi.org/10.1093/molbev/mst010
- Knoll A. H. (2015). Paleobiological Perspectives on Early Microbial Evolution. Cold Spring Harbor perspectives in biology, 7(7), a018093. https://doi.org/10.1101/cshperspect.a018093
- Kolmogorov M, Armstrong J, Raney BJ, Streeter I, Dunn M, Yang F, Odom D, Flicek P, Keane TM, Thybert D, Paten B., Pham S. "Chromosome assembly of large and complex genomes using multiple references" Genome research. 2018 doi:10.1101/gr.236273.118
- Alexandros Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies, Bioinformatics, Volume 30, Issue 9, 1 May 2014, Pages 1312–1313, https://doi.org/10.1093/bioinformatics/btu033
- Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, Lomsadze A, Pruitt KD, Borodovsky M, Ostell J. NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res. 2016 Aug 19;44(14):6614-24