Reproducing an Alignment-Free Analysis of Phylogenetic Relationships

Chris Henn

7 May 2018

A wide variety of approaches exist for computing alignments of fragmented genomic data, but none are perfect and most require unwieldy amounts of computational power. These difficulties motivate so-called "alignment-free" approaches to analyzing DNA; analyses concerning DNA sequences which don't require an alignment and can instead operate directly on fragmented reads produced by a sequencing machine.

A successful example of an alignment-free analysis is given in [2]. The authors use a k-mer based statistic called d_2^S to compute the strength of phylogenetic relationships between 144 bacteria [3]. They analyze this data with an intuitive graph visualization, which successfully reproduces known phylogenetic relationships between the bacteria.

For this project, I have attempted to reproduce the result in [2] via a Python implementation of the d_2^S statistic. I encountered several difficulties, which I detail below.

Data

The authors of [2] used the 144 bacteria listed in Supporting Table 2 of [1]. This table lists the "NCBI RefSeq Accession Number" for each bacteria, as well as the corresponding domain and phylum. I scraped the IDs of this data into a text file, which then allowed me to download the complete genome for each organism via the NCBI Batch Entrez tool. I also scraped the domain and phylum data into a Python file (`phylum_data.py`), for use later in generating visualization data.

I then used the pyfasta[1] Python library to load FASTA data into memory for analysis. The particular steps I took in this analysis is described in the following section and codified in the project `analysis.py` file. In the end, this analysis produced a JSON file suitable for consumption by a JavaScript-based visualization.

Steps Taken

I implemented the d_2^* statistic described in [3] as a Python program (`d2.py`). In the end, this naive implementation ended up being far too computationally demanding, and thus I implemented the much simpler (but less useful) d_2 statistic as well.

Using this d_2 implementation, the `analysis.py` file performs the following steps:

[1] `https://github.com/brentp/pyfasta/`

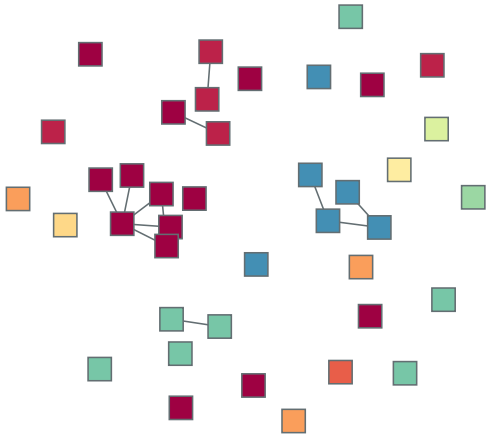

Figure 1: A static frame chosen from the animated and interactive visualization available online. In this graph, each node represents an organism. Nodes are color-coded by phylum. Edges in this graph represent similarity between organisms beyond a controllable threshold strength; this similarity is computed using the d_2 statistic. A low-quality clustering of nodes by phylum is apparent in the visualization.

1. Read all bacteria genomic data into memory.
2. Compute the d_2 statistic for each possible pairing of genomes.
3. Write this data to a JSON file, along with the some metadata concerning each genome (identifiers, corresponding phylums, and so on).

The JSON file is consumed by a JavaScript-based visualization. This visualization is a graph with nodes corresponding to organisms and weighted edges corresponding to the computed d_2 statistic between each organism. The visualization provides controls to filter the visible edges by weight (d_2 statistic), and organizes nodes with a force-based layout to facilitates clustering.

Discussion

The visualization of our analysis is available online.[2] In a successful visualization, we would expect to be able to cluster nodes by phylum effectively by choosing an approriate d_2 similarity threshold. While some amount of clustering is visible in our visualization (see Figure 1), it is generally weak and falls short of the results in [2]. There are a few contributing factors to our result:

- Our naive implementation of d_2^*, the statistic used in [2], was far to slow to be of any use. I used the d_2 statistic instead, which is known to be inferior.

[2] `https://beta.observablehq.com/@chnn/producing-phylogenetic-graphs-with-the-tex-d_2-statistic`

- Again for performance reasons, I did not compute the d_2 similarity between all possible pairings of the 144 bacteria sequences. Instead, I choose an arbitrary subset of 40 sequences for which to compute the d_2 statistic. With more computation time, a more complete visualization could be obtained.

Clearly, the failure to reproduce the desired results stems from incidental difficulties in my method. It does not indicate any fundamental flaw with the methodology in [2].

Given an opportunity for further study, I still believe that this alignment free approach to genomic analysis could answer some interesting questions. Perhaps most importantly, I wonder whether it can represent phylogenetic relationships between other groups of organisms (in primates for example).

References

[1] Robert G. Beiko, Timothy J. Harlow, and Mark A. Ragan. "Highways of gene sharing in prokaryotes". In: *Proceedings of the National Academy of Sciences* 102.40 (2005), pp. 14332–14337. ISSN: 0027-8424. DOI: 10.1073/pnas.0504068102. eprint: http://www.pnas.org/content/102/40/14332.full.pdf. URL: http://www.pnas.org/content/102/40/14332.

[2] Cheong Xin Chan Guillaume Bernard Mark A. Ragan. "Recapitulating phylogenies using k-mers: from trees to networks". In: *F1000Research* (2016). URL: https://f1000research.com/articles/5-2789/v2.

[3] Gesine Reinert et al. "Alignment-Free Sequence Comparison (I): Statistics and Power". In: 16 (Dec. 2009), pp. 1615–34.