Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements

Dreycey Albin (da39)

April 4, 2019

Scientific question

The scientific question and goal for this paper focuses on *potential to use decomposition based methods, commonly used in natural language processing, to evaluate genome similarity.* This research extends the current techniques that were available at the time by decreasing memory costs for metagenomic genome assembly. To test the tool, there were several different analysis performed, among testing spiked datasets and benchmarking the program against the other tools that were being used at the time of publication.

Hypothesis tested

The hypothesis tested in this research is that *a unique hashing method along with SVD may create for a novel way to cluster then assemble genomes.* They tested this hypothesis by seperating different clustered principle components using partitioning, then testing the sequences associated with each partition.

Methods/Results

In my opinion, there are several ways the methods can be split up: (1) The algorithm; (2) Memory testing; (3) Testing the ability to cluster genomes. The algorithm works by using a special hash function to map kmers of each sample to specific hash bins (titled hyperplane hashing). This allows for a matrix to get built that has the kmers represented in the columns, and the samples represented in each of the rows. Once the kmers have been mapped, there is a two step conditioning process. First, the local weight is calculated, then the global weighting is calculated. After conditioning, this matrix then represents the k-mer abundance matrix. Now regular techniques to applying SVD are used to calculated the eigen vectors for the matrix (they used "Genism").

Memory tesing was done by testing the ability for their program, Latent Strain Analysis (LSA), to work on larger databases. This would usually become the bottleneck for other tools using similar techniques to LSA. To test this ability, they used several differently sized databases, ranging from 10GB, to 300 GB, to 4TB. Due to memory constraints of some of the previous programs, testing was not able to be completed for the latter two databases.

Lastly, they aimed to characterize the different clusters arising in their method. Doing so showed that the LSA program was able to cluster underrepresented genome within a big population of other genomes. While this worked very well, it was also indicated that some of the background genomes, from the Human Microbiome Project (HMP), would sometimes semi-randomly be shown to be similiar (or partition together). One of the techniques performed was to asses how well the program could partition similiar related genomes. They also convincingly showed that LSA has a more memory efficient technique than the other programs.

Key implications of the results

The key implication from these data is that the natural language processing techniques may be applied to the problem of metagenomic assembly. Using a unique hash function, conditioning, and followed by decomposition have the ability to give principle components, or eigen vectors, as the right most singlular vectors. In addition, they showed very precisely that this technique has the ability to uniquely map the kmers.