Group 7:
Ryno Lawson
Vasileios-Evripidis Kyriakidis

Comparative Genomics
Practical 1 Report : Basic genome analysis

Summary

Our main goal on this practical is to compare the characteristics and features of BLAST and HMMER, understand how they work and find out which one is suitable and for what purpose. By utilising BLAST with five different genome sequences we identify and characterize the genomes according to their size, gene contents and the kingdom to which they belong.

BLAST

BLAST is a database searching tool which is used for finding sequence homologues. It performs a local search with a quick k-Tuple heuristic problem solving technique for aligning protein and nucleotide sequences. The heuristic approach seeks a solution which is good enough, but not the most optimal solution. Its k-Tuple feature is likened to a searching using ”words” in order to make a “word” list, which are short segments of the larger sequences instead of single residues. Although this improves the speed of computation, it reduces the accuracy of the matching query sequences. Such matching of words is done with the use of dynamic programming algorithms, such as Needleman-Wunsch or Smith-Waterman. Needleman-Wunsch is a global alignment and optimal matching technique for sequences of roughly equal size, which are suspected to have high similarity. Whereas, Smith-Waterman looks to achieve local alignment for dissimilar sequences which have regions of similarity.

While both are known to be slow due to their demand on processing resource, they are more sensitive to the precise match, mismatch scores and gap penalties. BLAST speeds this up by reducing the number of comparisons and uses the heuristic approach of Smith-Waterman local alignment. This results in trading accuracy for speed through the use of “words” instead of single residues together with Smith-Waterman.

This comparison of a “word“ list looks for the high-scoring segment pairs, giving each alignment a neighborhood score to be compared to the user-defined Thresholds(*T*).

HMMER

HMMER is a software used for homologs searching and sequence alignments (for either proteins or nucleotides) based on sequence databases. It does so by using the statistical tool of probabilistic Hidden Markov Model (HMM) and by creating HMM profiles which are used to be compared to sequences.

The HMM does this by labeling states and setting the parameters of emission and transition probabilities to each state. The transition probability is the likelihood of moving between state, while the emission probability is the likelihood of observing an output at a giving state.

The main advantage of HMMER is its ability to detect homologs as sensitive as possible and in a speed almost equal to BLAST.

Apart from having almost the same computational time, BLAST and HMMER share the ability to work with query sequences, but nevertheless it wouldn't work good if we attempted to run HMMER in parallel because it would take too long for a whole genome sequence comparison.

Viterbi Algorithm

On this exercise, it was also important for us to familiarize with the Viterbi algorithm. The Viterbi Algorithm is a forward-dynamic programing algorithm which is maximizing the probability essential, use of compound decision theory by combining Needleman-Wunsch and HMM. It is widely used for gene prediction and it provides an efficient way of finding the most likely state sequence in the maximum a posteriori probability sense of a process assumed to be a finite-state discrete-time Markov process. HMMs are used in biological research to resolve three essential problems: Evaluation, Decoding and Learning. The Decoding problem is our attempt to induce the most likely hidden states when given a model and a sequence of observations. Thus when our goal is to find the sequence of internal states that has, as a whole, the highest probability, the most used algorithm is the Viterbi algorithm.

Discussion

While BLAST is known to be of good speed of computing over other methods, there have been improvements of other methods in recent years and the programmes have been updated and technology of computing have improved. However, the HMM and Viterbi are slower due to their demand on computing resources such as memory and computation associated with forward dynamic programming, which requires the storage of states, emission probabilities, transitions probabilities, MAP, transition lengths and etc

Characterization of query genomes with BLAST using NCBI (https://blast.ncbi.nlm.nih.gov/)

File #	Organism	Size	Kingdom	# of genes
05.fa.txt	*Chlamydia trachomatis*	1042588	Bacteria	977
09.fa.txt	*Escherichia coli*	5277676	Bacteria	341
11.fa.txt	*Geobacter sulfurreducens*	4566144	Bacteria	4172
12.fa.txt	*Gloeobacter violaceus*	4659019	Bacteria	4430
28.fa.txt	*Saccharomyces Cerevisiae* (Chr VIII)	562643	Fungi	6445(Chr VIII 205)

Discussion

After running BLAST on the genome sequences as seen on the table above, we came up to a few interesting observations. First of all we can see that the number of genes is not proportional to the size of the genomes. For example, *E.coli* has the largest genome size of all the genomes we investigated, but the smallest number of genes. The only eukaryotic genome we had in our dataset, *Saccharomyces Cerevisiae*, seems to be more gene dense compared to the bacterial genomes.

BLAST search was fairly quick, taking approximately 5 minutes per search. If we were asked to run HMMER for the same genome sequences the computational time would be much slower than BLAST, but this might lead to better results.

References

Lobo, I. (2008) Basic Local Alignment Search Tool (BLAST). Nature Education 1(1):215
https://www.nature.com/scitable/topicpage/basic-local-alignment-search-tool-blast-29096

What is a hidden Markov model? (2004) - Sean R Eddy.
http://www.nature.com/nbt/journal/v22/n10/abs/nbt1004-1315.html

https://en.wikipedia.org/wiki/BLAST#Process

https://en.wikipedia.org/wiki/Heuristic_(computer_science)

https://www.quora.com/How-does-the-Smith-Waterman-alignment-algorithm-differ-from-the-Needleman-Wunsch-algorithm

https://sequencebase.com/smith-waterman-vs-blast/

http://www.gen.tcd.ie/molevol/nwswat.html

http://www.cim.mcgill.ca/~latorres/Viterbi/va_alg.htm

http://scialert.net/fulltext/?doi=jas.2012.1518.1525