Comparative Genomics 2018

Practical 1: Basic Genome Analysis

Group number: 6

Group members: Kyle Kimler, Kajetan Juszczak

**Summary**

Sequence alignment was one of the first applications of computation in the study of biology. It remains today the method by which we learn about the evolution of genes and proteins. Nowadays many web tools are made readily available by the USA NIH or by other laboratories around the world. BLAST and HMMER, the former an old heuristic search method to find local (contiguous) alignments of small query sequences, the latter a newer tool that utilized machine learning to produce a profile of likely amino acid changes in evolution, are two such web tools that can also be downloaded and run on your computer for free to run large batches.

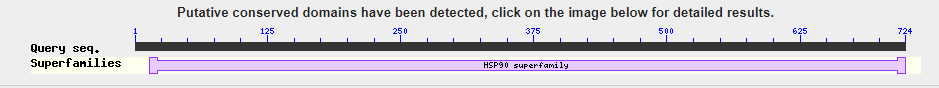
In this lab report we explored these two commonly used tools for sequence alignment.

**Key Questions to Answer**

Ex.1 : BLAST

1. 03.fa
   1. Bacteroides thetaiotaomicron strain 7330
   2. 6487685bp
   3. 5145 genes
   4. prokaryotic
2. 09.fa
   1. Escherichia coli HUSEC2011
   2. 5277676 bp
   3. 5766 genes
   4. prokaryotic
3. 20.fa
   1. Thermotoga maritima strain Tma100
   2. 1869610 bp
   3. 1928 genes
   4. prokaryotic
4. 24.fa
   1. Saccharomyces cerevisiae S288C
   2. 1531933 bp
   3. 799 genes
   4. eukaryotic
5. 51.fa
   1. Spiribacter curvatus strain UAH-SP71
   2. 1926631 bp
   3. 1912 genes
   4. prokaryotic

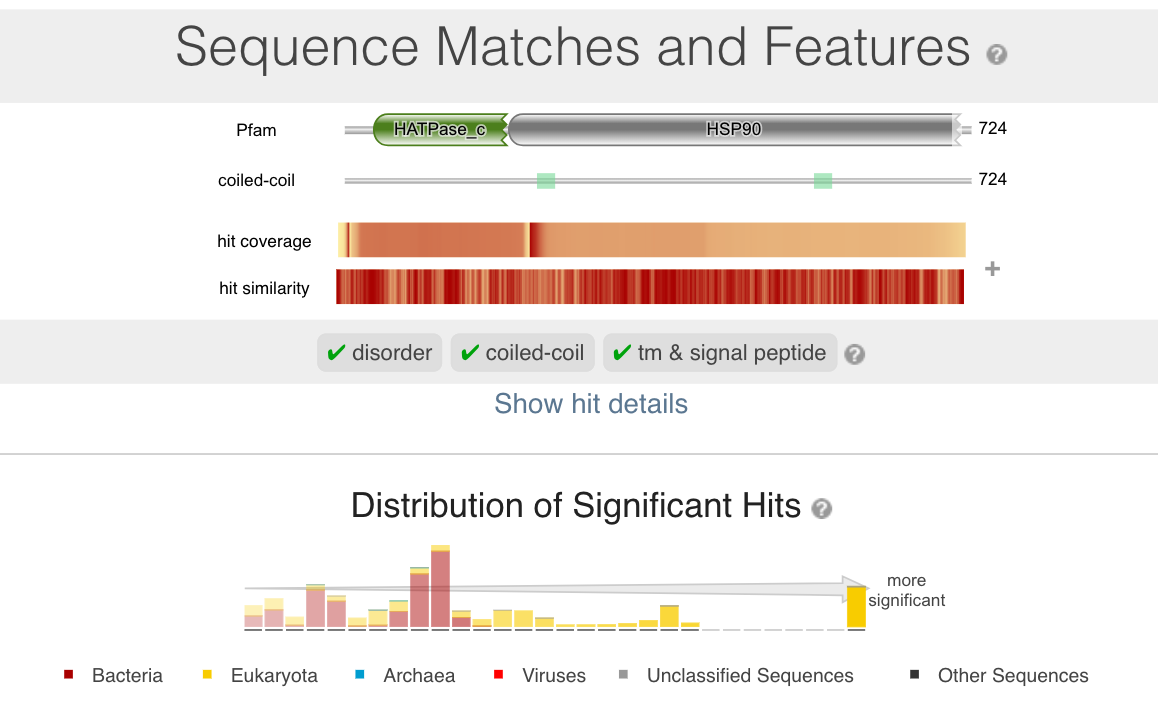
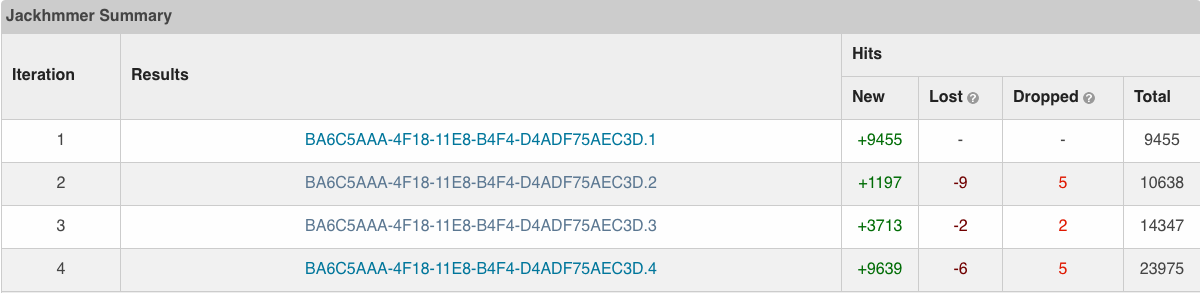
Ex.2 : BLAST

1. Types of BLAST
   1. BLASTp – performs alignment of protein query against protein database
   2. BLASTn – performs alignment of nucleotide query against nucleotide database
   3. BLASTx – takes nucleotide sequence as query and translate it and performs alignment against protein database.
   4. tBLASTn – takes protein sequence as query and performs alignment against translated nucleotide database.
2. HSP90 - beta
3. It belongs to HSP90 superfamily
4. Sequence similarity is defined by percent amino acid identity between two sequences – homology is defined by evolutionary distance between two sequences, which isn’t necessarily the same as percent identity. Homology includes measures of similarity but also takes into account evolutionary distance between two proteins. BLAST uses a rough heuristic to search for sequences of high similarity and can miss distant homologies - often more refined searches are needed to find orthologs.
5. Homologs of HSP90 exist in all eukarya families but also in bacteria. Apparently they’re absent in archaea, where HSP60 fulfills a similar function.
6. BLOSUM is based on substitution frequency between sequences with a certain number of substitutions – so BLOSUM62 has 62% similarity between sequences. PAM matrices are based on sequences that diverged from the same common ancestor and differ by a certain amount of mutations, so PAM250 has 250 substitution events per 100bp (a total similarity of about 22%, because sub events happen more frequently at certain positions or on certain amino acids). BLAST is a rougher search tool that uses a heuristic to find sequences of great similarity to the query sequence, and you can configure it to use different substitution matrices to rank search results – I think the command line program includes BLOSUM62 and PAM250 as options among a few others.
7. Top 4: mus musculus 100%, cricetulus griseus 99%, mustela putorius furo 99%, bos Taurus 99%. Worst hit is chrysochloris asiatica – only 98% identity. It has full sequence alignment with the query sequence but it is changed on one end – only 1442 out of 1480 align. Query cover of 100% still allows 98% identity matches in this way.
8. E-value stands for expect value – the expected number of random hits for a search for a match of score S.  
     
   E = Kmne^(\*S) where m and n are the lengths of the query sequence and the potential match, K and lambda are parameters chosen depending on the size of the search (database) and the scoring algorithm respectively, and S is the score under scrutiny.

Ex. 3 : HMMER

1. HMMER is a hidden-markov based alignment program used to find distantly related homologs
2. You can do HMMER searches on multiple sequence alignment files or single fasta files on proteins or nucleotides. You can also do jackhmmer on proteins to iteratively search a homology space, using an iteration’s scoring/sub table as an MSA for the next iteration, similar to psi-blast
3. HMMER creates a “profile HMM” which categorizes and scores certain substitutions based on a training dataset for best identification of distantly homologous sequences. HMMER comes with a default profile HMM that was found to be best suited to the job of sequence alignment, and it’s been optimized to run quickly, so the latest version of HMMER actually runs as fast as BLAST (which purposely uses a quick and dirty “heuristic” search) for proteins and only half the speed for nucleotide searches. A quick run of HMMER on the attached fasta picked out a mosquito version of HSP90 as the top hit, so apparently it doesn’t take into account either total identity or some other measures into its e-value – maybe BLAST is better for species identification while HMMER is best to find family information.

Ex. 4 : HMMER

1. 
2. 
3. Phmmer took 25 seconds. BLAST took 45 seconds. Jackhmmer takes 25 seconds for first iteration, 35 for second, 38 seconds for third, and so on.
4. Taxonomic spread is similar to BLAST, but jackhmmer found 10 archaea variants with very low e-values!

**References**

HMMER downloadable command-line tool userguide.pdf

HMMER.org

<https://www.ncbi.nlm.nih.gov/blast/html/sub_matrix.html>

biostars.org

“Comparison of the PAM and BLOSUM Amino Acid Substitution Matrices”**.** Cold Spring Harbor Protocols 2008 Jun 1; 2008:pdb.ip59. doi: 10.1101/pdb.ip59.