Practical 1: Basic Genome Analysis

Group number: 2
Group members: Maximilian Senftleben, Zhong Hao Daryl Boey

Summary: This practical covered the basics of performing sequence alignments using 2 programs, BLAST and HMMER, and learning about the features both programs offer. By using and contrasting BLAST and HMMER, we understand the basic principles employed in both, and the different purposes they may serve a researcher. The practical also goes into the details of extracting critical information from both and discerning the limitations of such programs, so that we know when the information given is trustworthy, or if certain parameters should be tweaked to further improve the accuracy of the search.

Exercise 1 - Take the unknown genomes that are given to you and extract the following from the NCBI webpage (https://www.ncbi.nlm.nih.gov/):

1. Which organism does each given genome belong to?

   29.fa.txt: Saccharomyces cerevisiae S288C chromosome IX, complete sequence
   44.fa.txt: Leuconostoc gelidum JB7, complete genome
   47.fa.txt: Neisseria meningitidis alpha710, complete genome
   16.fa.txt: Rhodopirellula baltica SH1
   20.fa.txt: Thermotoga maritima strain Tma100

2. What is the size of this genome (in bp)?

   29.fa.txt: 439888
   44.fa.txt: 1893499
   47.fa.txt: 2242947
   16.fa.txt: 304850
   20.fa.txt: 1869610

3. What is the number of genes?

   29.fa.txt: 232
   44.fa.txt: 1,967
   47.fa.txt: 2,104
   16.fa.txt: 7406
   20.fa.txt: 5806

4. What type of organism is it (i.e., prokaryotic or eukaryotic)?

   29.fa.txt: eukaryotic

44.fa.txt: prokaryotic
47.fa.txt: prokaryotic
16.fa.txt: Prokaryotic
20.fa.txt: Prokaryotic

Exercise 2 - A protein name with a species specification is given to you. First, extract the amino acid sequence from NCBI in FASTA format, and then answer the following questions:

1. Familiarize yourself with the different types of BLAST (blastp, blastn, blastx, etc) and provide brief explanations for each BLAST type.

   blastn – blastn provides the best matching DNA sequence to the query sequence (query sequence = DNA)
   blastp – The input for blastp should be a protein sequence and the return will be the most similar protein sequence taken from the by the user chosen protein database.
   blastx – blastx creates a six-frame conceptual translation product of the given nucleotide sequence input and compares it to a chosen protein sequence database.

2. Run the suitable type of BLAST at NCBI on your extracted sequence (hint: in order to avoid duplicate sequences, run BLAST with a specific dataset (i.e., Reference proteins, refseq).

3. What is a protein family and superfamily?

   A super-family of proteins can be seen as the largest assembly of proteins together with their common protein ancestors. This concept of group is derived from a structural alignment or mechanistic similarity, not necessarily regarding their sequence similarity, where the proteins can be distantly related. It is more likely, that they have a structural similarity rather than a sequence similarity.

   The term protein family corresponds to a group of proteins with related functions and similarities concerning their structure or sequence. Their corresponding genes are usually also grouped together in gene families.

   a. 3.1. Is there a superfamily for this protein?

      Yes.

   b. 3.2. If so, what is it?

      The protein is in the p53-superfamily.

   c. 3.3. If not, what do you think the reason is?

4. Search for the terms similarity and homology;
    a. 4.1. Define the terms similarity and homology.

       Similarity refers to the shared identity between 2 sequences, while homology implies shared ancestry.

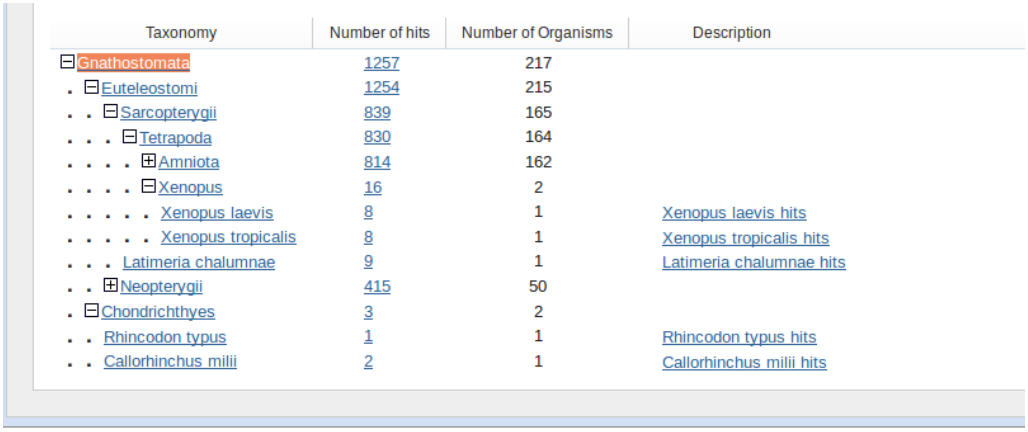
    b. 4.2. Briefly discuss the relationship between similarity and homology.

       2 similar sequences may not necessarily be homologous, but similarity is used as a unit of measure for homology.

    c. 4.3. Relate these terms for your BLAST search (discuss the main aim of running a BLAST search from the perspective of homology and similarity).

       A BLAST search works by aligning sequences between the query and databases, by using local sequence similarity. This is then extrapolated to calculate statistical significance of the matches, and excess similarity can be assumed to be an indicator of homology.

5. What is the taxonomic spread of this protein family? Do homologs exist in all major kingdoms?



Taxonomy	Number of hits	Number of Organisms	Description
⊟Gnathostomata	1257	217	
. ⊟Euteleostomi	1254	215	
. . ⊟Sarcopterygii	839	165	
. . . ⊟Tetrapoda	830	164	
. . . . ⊞Amniota	814	162	
. . . . ⊟Xenopus	16	2	
. . . . . Xenopus laevis	8	1	Xenopus laevis hits
. . . . . Xenopus tropicalis	8	1	Xenopus tropicalis hits
. . . Latimeria chalumnae	9	1	Latimeria chalumnae hits
. . ⊞Neopterygii	415	50	
. ⊟Chondrichthyes	3	2	
. . Rhincodon typus	1	1	Rhincodon typus hits
. . Callorhinchus milii	2	1	Callorhinchus milii hits

The taxonomic spread derived from the blast-search shows, that the p53 super family only occurs in the kingdom animalia, as the species *Gnathostomata* (jawed vertebrates) is part of the animal kingdom. The blast-search was done with the database refseq_protein, 20000 of maximum aligned sequences and an expected e-value threshold of 0.001.

6. Search for substitution matrices (i.e., BLOSUM and PAM) and explain for what reason and how they are used;
    a. 6.1. What are the differences between BLOSUM and PAM matrices?

Both the BLOSUM and PAM matrices are based on data from actual substitutions, while BLOSUM is based on local alignments, PAM is based on global alignments.

b. 6.2. What are the differences within the substitution matrices themselves (i.e., there are different BLOSUM and PAM matrices such as BLOSUM45, BLOSUM62, and PAM100, PAM250, etc.. You are expected to describe the differences among the BLOSUM matrices and among the PAM matrices.).

   PAM is measured in terms of PAM units, or Point Accepted Matrices, which states, per hundred residues, the number of acceptable point mutations. PAM100 means 100 PAM units, or 100 mutations per average per 100 residues, while PAM250 means 250 mutations per average per 100 residues. BLOSUM however is calculated based on variations in the cut-off percentage for similarity clustering, with BLOSUM45 obtained using a cut-off percentage threshold of 45%.

c. 6.3. In what way does BLAST use these substitution matrices?

   BLAST uses these matrices as scoring matrices, to score alignments of the sequences.

7. From your BLAST search, extract the top 3 homologous in all species.
   Mus musculus targeted non-conditional, lacZ-tagged mutant allele Wrap53:tm1e(EUCOMM)Wtsi; transgenic, Mus musculus targeted KO-first, conditional ready, lacZ-tagged mutant allele Wrap53:tm1a(EUCOMM)Wtsi; transgenic, Mouse DNA sequence from clone RP23-56I20 on chromosome 11.

   a. 7.1. Extract identities, similar matches and gaps in percentages of the best and the worst hits from your search and explain what does the difference indicate?

      Identities: 11515/11515(100%) vs 292/351(83%).
      Similar matches: 3 vs 1
      Gaps: 0/11515 (0%) vs 3/351(0%)

      Firstly the top hit has a significantly larger length of identity, compared to the bottom hit, this indicates a large continuous section of similar sequences, which signifies greater homology as opposed to a short section. The lack of gaps in the top hit also indicate a good match homology wise.

   b. 7.2. What search criterion selected the worst hit?

      Searching using the wrong nucleotide database (such as the Human G+T) returned the worst results.

8. E-value:

a. 8.1. What does the E-value stand for?

   Expect value. The E-value is used as a parameter to assess the level of noise and ensure that the hit is statistically significant.

b. 8.2. Write down its formula and explain it in detail (every parameter and what is it used for).

   $$E = Kmn\, e^{-\lambda S}$$

   S refers to the raw score from the scoring matrices, m and n refer to the sequence lengths, lambda refers to the natural scale for the scoring system, and K the natural scale for the search space size.

HMMER

In this part of the practical you are going to search the same query as in BLAST and compare the output of this search to the BLAST search.

Exercise 3 - The basics of HMMER:

5. What is HMMER used for?
   HMMER is similar to BLAST in function, in that its used to perform searches and sequence alignments of homologous sequences, between a query sequence and a database, using Hidden Markov Model (HMM) profiles instead of directly comparing sequences and scoring them, similar to PSI-BLAST. These profiles are generated using Position Specific Scoring Matrices (PSSM), which are a better marker of conserved positions in the sequence.

6. Write down and explain different types of HMMER searches.

   Hmmscan: using a protein sequence as a query against a database of HMM profiles.
   Hmmsearch: using a protein profile as a query against a database of protein sequences.
   Phmmer: using a protein sequence as a query against a database of protein sequences.
   Jackhmmer: repeated searching a sequence against a database of protein sequences.
   Nhmmer: using a DNA or RNA sequence as a query against a database of DNA or RNA sequences.
   Nhmmscan: using a DNA sequence as a query against a database of DNA profiles.

7. Compare HMMER to BLAST in the following ways:
   . 3.1. What are the advantages and disadvantages of each over the other?

     HMMER is more likely to detect distant evolutionary relationships of sequences compared to BLAST, as it uses conservation information in the form of PSSM profiles rather than simply searching using sequence matches. However HMMER requires more steps to generate profiles for either the query or database, as opposed to BLAST which does not have this complication.

   . 3.2. Which one is faster?

     BLAST.

   . 3.3. For what kind of search would you choose BLAST and for what other(s) HMMER? Why?

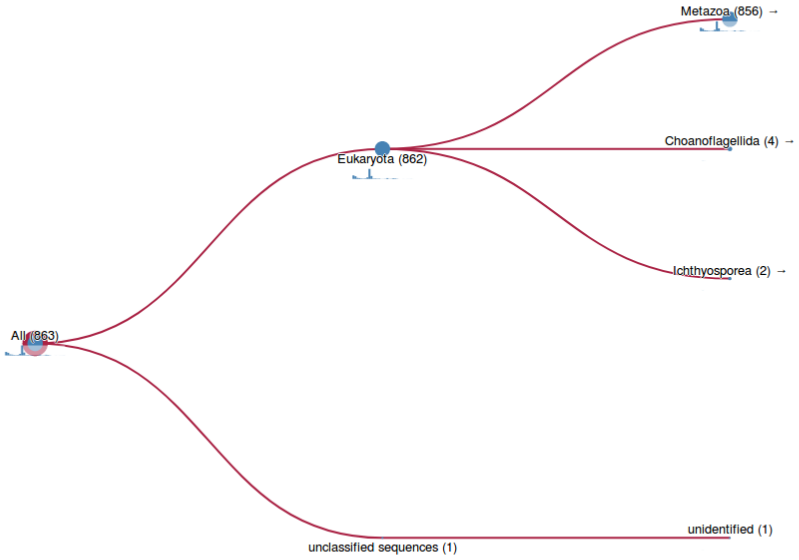
     For simple initial searches to determine the function or location of a new sequence, or to verify that an amplified sequence is correct, BLAST would be more appropriate. HMMER would be more useful when searching for evolutionarily related sequences or proteins.

Exercise 4 - HMMER searching

1. Run the same query as in exercise 2 using phmmer on the HMMER website.
2. Run the same search using jackhmmer.
3. Discuss the speed and output of phmmer and jackhmmer, and the difference to BLAST.

   Blast is slightly faster than phmmer and much faster compared to jackhmmer. Phmmer and jackhmmer both use the probabilistic method profile hidden Markov model for their search, whereas blast is using scoring matrices, such as matrices derived from the Needleman-Wunsch-algorithm.

4. Is the taxonomic spread the same as in BLAST?



The taxonomic spread for jackhmmer and phmmer is the same. Both show, that most of the homologs exist in the kingdom of metazoa (animalia). Additionally, two unranked kingdoms are shown (Choanoflagellida and Ichtyosporea), as they both contain species with the p53 superfamily, which is different to the blast search. The unclassified sequence can be disregarded.