**Practical 1: Basic Genome Analysis**

Group no. 11

Fuqi Xu, Shuhan Xu, Milda Valiukonytė

**Exercise 1.**

1. 09.fa.txt: *Escherichia coli*

17.fa.txt: *Streptomyces coelicolor*

24.fa.txt: *Saccharomyces cerevisiae*

49.fa.txt: *Rubrobacter xylanophilus*

51.fa.txt: *Spiribacter curvatus*

1. 09.fa.txt: 5277676 bp

17.fa.txt: 356023 bp

24.fa.txt: 1531933 bp

49.fa.txt: 3225748 bp

51.fa.txt: 1926631 bp

1. 09.fa.txt: 5373

17.fa.txt:356

24.fa.txt: 799

49.fa.txt: 3281

51.fa.txt: 1912

1. 09.fa.txt: Prokaryotic

17.fa.txt: Prokaryotic

24.fa.txt: Eukaryotic

49.fa.txt: Prokaryotic

51.fa.txt: Prokaryotic

**Exercise 2.**

1. blastn – *nucleotide vs. nucleotide* (searches nucleotide databases using nucleotide query)

blastp – *protein vs. protein* (searches protein databases using protein sequence query)

blastx – *nucleotide (translated) vs. protein* (uses nucleotide query to identify potential protein products encoded by it and then searches protein databases)

tblastn – *protein vs nucleotide (translated)* (uses protein query to search nucleotide database for sequences encoding similar proteins)

tblastx – *nucleotide (translated) vs. nucleotide (translated)* (uses nucleotide query to identify potential protein products encoded by it and uses them to search nucleotide database for sequences encoding similar proteins)

1. Ran blastp with refseq database
2. Protein family is a group of proteins with evolutionary related sequences (e.g. mitochondrial carrier protein family). Protein superfamily is a larger set of protein families grouped together based on sequence or mechanistic similarity (e.g. alkaline phosphatase superfamily, a group of families which share an αβα sandwich structure and and perform reactions by a common mechanism).
   1. Yes, this protein belongs to CNH superfamily because it has CNH protein domain (also found in NIK1-like kinase, mouse citron and yeast ROM1, ROM2 proteins).
   2. Similarity – two sequences are said to be similar if they share a large enough number of identical nucleotides/amino acid residues. Homology – shared ancestry between sequences.
   3. Sequence similarity does not mean homology. Homology can be inferred from high sequence similarity but sequences might also be similar because of convergent evolution.
   4. The main aim of running a BLAST search is to find homologous sequences by ranking similar sequences based on sequence identity.
3. This protein family is spread only in animal kingdom among mammals.
4. Substitution matrices give scores between amino acids which show the relative ease one amino acid can mutate into another during evolution. They are used to score sequence alignments.
   1. PAM matrices are derived from evolutionary closely related proteins (>85% sequence identity) while BLOSUM matrices are based on highly conserved stretches of distantly related proteins.
   2. The original matrix is PAM1 and higher number matrices are derived from PAM1. One PAM unit distance between sequences means that on average 1point mutation has been accepted among 100 residues (accepted means it has changed the amino acid sequence of the protein). PAM distance does not equal sequence identity. At PAM1, two proteins are 99% identical, but at PAM250, there are 80 differences per 100 residues (only 20% sequence identity). It is because more than one point mutation can occur at a single position. In BLOSUM matrices, higher numbers denote higher sequence similarity and therefore smaller evolutionary distance.
   3. These substitution matrices are used to score and rank sequence alignments.
5. Top three homologs: transforming growth factor-beta receptor-associated protein 1 [Mus musculus]; PREDICTED: transforming growth factor-beta receptor-associated protein 1 isoform X2 [Mus musculus]; PREDICTED: transforming growth factor-beta receptor-associated protein 1 isoform X1 [Mus musculus].
   1. Identities: best hit 100%, worst hit 83%. Gaps: best hit 0%, worst hit 0 %. Both of the alignments produced full coverage (0% gaps) however there is a difference in sequence identities indicating the best hit is more likely to be a homologous sequence.
   2. Sequence coverage 100% and high E-value (0.0).
   3. Expect value
   4. 

m and n are lengths of aligned sequences;

S is the score of the alignment, which is given using a substitution matrix; E-value decreases exponentially with increasing S values;

K and λ are constants which depend on the substitution matrix used.

**Exercise 3.**

1. It is used for homology searching.
2. There are four types of searches: phmmer – uses a protein sequence as query and searches it against a protein database; hmmsearch – uses an HMM profile as query and searches it against a protein database; hmmscan – uses a protein sequence as query to search HMM profile database; jackhmmer – uses a protein sequence to iteratively search a protein database.
   1. HMMER uses a profile HMM instead of single sequence information which is constructed from a multiple sequence alignment. The advantage of this is that it captures more information about each residue position by calculating probabilities of insertion/deletion at each position. Hence, when searching for homologs each insertion and deletion is penalized differently depending on profile HMM in contrast to BLAST, which gives the same penalties to all insertions/deletions. As a result, HMMER is more accurate and more able to detect true homologs. However, profile HMM needs to be computed first which decreases the speed of the search.
   2. Historically HMMER was much slower than BLAST but now the algorithm is almost as quick as BLAST.
   3. HMMER would be more suitable to detect more distant homologs whereas BLAST would be useful for quick initial homology search.

**Exercise 4.**

1. Jackhmmer is slightly slower than phmmer (19.91 secs compared to 19.83 secs) but they both produced the same results (as jackhmmer only ran 1 iteration). A different reference database was used compared to BLAST (results of which included various isoforms of the same protein, predicted proteins, etc.)
2. No, matches were also found not just in animal, but other kingdoms as well, such as fungi.