Comparative Genomics 2018

Practical 1: Basic Genome Analysis

Assistants: Deniz Seçilmiş, Miguel Castresana, Stefanie Friedrich

All forms of plagiarism are forbidden, and if detected it will result in a lower grade.

This practical is divided into 2 categories as BLAST and HMMER; however, the main aim of both parts is the same, to make you familiarized with homology search. For this purpose, you will run several exercises until you get the feeling of covering the main principles of the basic genome analysis.

BLAST

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?
2. What is the size of this genome (in bp)?
3. What is the number of genes?
4. What type of organism is it (i.e., prokaryotic or eukaryotic)?

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.
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?
    - 3.1. Is there a superfamily for this protein?
    - 3.2. If so, what is it?
    - 3.3. If not, what do you think the reason is?
4. Search for the terms similarity and homology;
    - 4.1. Define the terms similarity and homology,
    - 4.2. Briefly discuss the relationship between similarity and homology.
    - 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).
5. What is the taxonomic spread of this protein family? Do homologs exist in all major kingdoms?

6. Search for substitution matrices (i.e., BLOSUM and PAM) and explain for what reason and how they are used;
    6.1. What are the differences between BLOSUM and PAM matrices?
    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.).
    6.3. In what way does BLAST use these substitution matrices?
7. From your BLAST search, extract the top 3 homologous in all species.
    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?
    7.2. What search criterion selected the worst hit?
8. E-value:
    8.1. What does the E-value stand for?
    8.2. Write down its formula and explain it in detail (every parameter and what is it used for).

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:

1. What is HMMER used for?
2. Write down and explain different types of HMMER searches.
3. Compare HMMER to BLAST in the following ways:
    3.1. What are the advantages and disadvantages of each over the other?
    3.2. Which one is faster?
    3.3. For what kind of search would you choose BLAST and for what other(s) HMMER? Why?

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

Reporting your exercises:

Use the template that we provided for you.