Comparative Genomics 2018

Practical 1: Basic Genome Analysis

Group 11
Shuhan Xu, Fuqi Xu, Milda Valiukonyte

Summary

In this practical, we practiced how to understand and use BLAST, interpret the BLAST and compared BLAST with a new sequence alignment tool, HMMER. We tried to predict genes in genomes, predict the protein one gene corresponds and find the homologous proteins of target genes. Also, we explained how BLAST and HMMER work and the evaluation of BLAST result.

In exercise 1, we found that the number of genes varies in different reference genomes, and we prefer to use the NCBI top hit genome since it is based on the BLAST result and more widely accepted. In exercise 3, we compared JackHMMER, HMMER, and BLAST. In general, HMMER is slower but more accurate and sensitive than BLAST

Key questions to answer

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

2.
09.fa.txt: 5277676 bp from top hit
17.fa.txt: 8667507 bp from reference genome
24.fa.txt: 1531933 bp from top hit
49.fa.txt: 3225748 bp from top hit
51.fa.txt: 1926631 bp from top hit

3.
09.fa.txt: 5373 genes from top hit (5358 genes from reference genome)
17.fa.txt: 7910 genes from reference genome
24.fa.txt: 799 genes from top hit (799 genes from reference genome)
49.fa.txt: 3281 genes from top hit (3257 genes from reference genome)
51.fa.txt: 1912 genes from top hit (1908 genes from reference genome)

4.
09.fa.txt: prokaryotic
17.fa.txt: prokaryotic

24.fa.txt: eukaryotic
49.fa.txt: prokaryotic
51.fa.txt: prokaryotic

Exercise 2

1.

blastp: searches amino acid sequence against protein sequence database
blastn: searches nucleotide sequence against nucleotide sequence database
blastx: translates nucleotide sequence in all 6 reading frames and searches against protein database
tblastn: searches protein sequence against nucleotide sequence database translated in all 6 reading frames
tblastx: translates nucleotide sequence in all 6 reading frames and searches against nucleotide sequence database translated in all 6 reading frames.

2.

run blastp against refseq

3.

A protein family is a group of closely related homologs which perform the same function.
Protein families are then grouped into superfamily in which all members are homologs but may have different functions from one another.

3.1.

There is a superfamily for this protein.

3.2.

HSP90 superfamily

4.1

Similarity is a measure of relationship between aligned sequences based on the degree of identity or similarity of amino acids at the matched positions.

Homology means that sequences are similar in structure or sequence due to a common ancestry.

4.2

Homology is different from similarity. Homology comes from evolutionary relationship which may not be entirely evident from sequences, e.g. sequences which are distantly related may have low identity or similarity or sequences may be similar due to convergent evolution. Similarity is a quantitative method used to measure relationship between sequences based on alignment. Highly similar sequences with Evalue below a threshold are inferred to be homologous.

4.3

When we are performing BLAST search, we are looking for highly similar sequences. We assume that highly similar sequences with Evalues below a threshold are homologous.

5.
In eukaryota, there are 555 hits and 464 organisms.
In bacteria, there are 463 hits and 291 organisms.

The taxonomic spread is extended to both eukaryote (464) and bacteria (291). However, this protein family is not represented in Archaea.

6.
The scores in substitution matrices represent the likelihood/feasibility with which one residue (nucleotide or amino acid) may mutated into another. Substitution matrices are used to align sequences. The alignment score is calculated from the summation of the substitution scores of all pairs of matched residues in an alignment (if there are no gaps).

6.1
The BLOSUM matrix is generated based on mutation information from highly conserved regions in sequences. And the PAM matrix is based on the substitution frequency of homologous protein sequences.[1]

6.2
PAM-N represents the substitution rate and the evolutionary distance. The larger N is, the lower the similarity is. For example, PAM 250 is used for sequences with around 20% similarities.

In BLOSUM, N represents the percentage identity between sequences, the larger N is, the higher the similarity is. For example, BLOSUM 62 is used to compare sequences with 62% sequence similarities.

6.3
BLAST uses these substitution matrices together with gap initiation and extension penalty to align sequences and calculate alignment score. The optimal alignment is the alignment which maximizes the alignment score. We measure similarity between sequences based on their alignment score with more similar sequences have higher score.

7.
Top three homologues:
heat shock protein 75 kDa, mitochondrial [Mus caroli]
heat shock protein 75 kDa, mitochondrial isoform X1 [Mus pahari]
heat shock protein 75 kDa, mitochondrial precursor [Rattus norvegicus]

7.1
Identities: best hit 100%, worst hit 86%
Similar matches: best hit 100%, worst hit 91%
Gaps: best hit 0%, worst hit 3%
The difference indicates that both coverage and sequence identity are important to score higher in BLAST search. The best hits in BLAST have both a better coverage of the target sequence and higher sequence identity whereas worse scoring hits cover less and have lower sequence identity. It indicates that the best hit might have closer evolutionary relationship than the worst hit.

7.2
The Max target sequences (100) and expect threshold (10) select the worst hit. The worst hit is the hit with the lowest Max alignment score within the 100 hits accepted by the expect threshold.

8.1
E value stands for the number of hits one would expect to get by chance when searching a particular database. A low E value indicates that the similarity is less likely to happen by chance, and the alignment is better.

8.2
E = k*n*m*exp(-lamda*S)
n: the length of query, m: the size of the sequence.
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.

Section 3
1.
HMMER can search for homologous protein or DNA sequences using either a single query sequence or an alignment of sequences. It can also perform sequence alignments.

2.
Phmmer: searches protein sequence against protein sequence database
Hmmscan: searches protein sequence against profile-HMM database
Hmmsearch: searches profile-HMM (created from protein sequence alignment) against protein sequence database
Jackhmmer: searches protein sequence or profile-HMM (created from protein sequence alignment) iteratively against protein sequence database

3.1
HMMER advantage:

1. With a multiple sequence alignment as an input, HMMER is able to search for distantly related homologs.
2. It allows for automatic annotation of protein domain within the sequence.
3. It uses probabilistic model (profile hidden Markov model) to determine how residues and gaps should be scored at each position instead of a position-independent scoring system
4. Instead of reporting single best-scoring alignment, it considers the complete ensemble of alignments and gives a confidence score at each aligned residue [2]
5. Its consideration of alignment uncertainty increases the accuracy of its homologue search.

HMMER disadvantage:

1. It does not consider how the identity of a residue at a particular position may influence the identity of residues in other positions. For instance, it cannot capture base-pair interactions.
2. Currently it is not able to translate the nucleotide query or nucleotide database in the 6 reading frames and perform searches.
3. HMMER3 is limited to local alignment

BLAST advantage:

1. BLAST is faster than HMMER, especially in nucleotide sequence searches;

BLAST disadvantage:

1. It is less sensitive and accurate in detecting distantly related homologs.
2. Less accurate in detecting distant homologue as it does not take into account alignment uncertainty.

3.2

BLAST is faster than HMMER, especially in nucleotide sequence searches.

3.3

When the query is a single sequence, I would use BLAST since BLAST is faster HMMER and HMMER does not have obvious advantage over BLAST.

When the query is a multiple alignment of sequences and that I want to find distant homologue, I would use HMMER since its use of profile hidden Markov Model makes it more sensitive and accurate in the search for distant homologue.

Exercise 4

3.

Speed:

pHMMER: 19.63s

JackHMMER (iteration 1): 19.55s

BLAST: about 40s

The speed of pHMMER and JackHMMER are approximately the same but faster than BLAST. The

Output:

pHMMER top 3 hits:

TRAP1_MOUSE – Heat shock protein 75 kDa, mitochondria, Mus musculus
TRAP1_RAT – Heat shock protein 75 kDa, mitochondria, Rattus norvegicus
A0A1A6GXM4_NEOLE – Uncharacterized protein, Neotoma lepida

JackHMMER (iteration 1) top 3 hits:

TRAP1_MOUSE – Heat shock protein 75 kDa, mitochondria, Mus musculus
TRAP1_RAT – Heat shock protein 75 kDa, mitochondria, Rattus norvegicus
A0A1A6GXM4_NEOLE – Uncharacterized protein, Neotoma lepida

BLAST top 3 hits:

heat shock protein 75 kDa, mitochondrial precursor [Mus musculus]
heat shock protein 75 kDa, mitochondrial [Mus caroli]
heat shock protein 75 kDa, mitochondrial isoform X1 [Mus pahari]

The outputs of pHMMER and JackHMMER are quite similar.

The outputs of BLAST are proteins from more closely related species (from the same genus Mus) while the outputs of pHMMER and JackHMMER are proteins from more distantly related species (from different genus but the same order Rodentia)

4. In pHMMER and JackHMMER, there are 5081 hits in Bacteria, 4132 hits in Eukaryote, 14 hits in Archaea and 12 unidentified sequences. Compared to BLAST, pHMMER and JackHMMER able to identify 10-fold more hits in bacteria and eukaryote and also identify hits some hits in Archaea which BLAST did not. HMMER methods seem to suggest that the protein family is present in all 3 domains of life

Reference:

1. Zvelebil, Marketa J., and Jeremy O. Baum. Understanding bioinformatics. Garland Science, 2007.
2. HMMER Web Server: Interactive Sequence Similarity Searching. R. D. Finn, J. Clements, S. R. Eddy. Nucleic Acids Research, 39:W29-37, 2011.
3. Sean R. Eddy, Travis J. Wheeler, HMMER3 User guide Version 3.1b2; February 2015