

Part 2

Comparative sequence analysis of a gene across different animals (or other biological units) is a widely used approach for providing understanding about the gene. For example, mutations in the sequence of a gene which is known to be a disease gene: what kind of mutations are 'natural consequences' of evolutionary processes, i.e. what kind of differences exist among different species?

BRCA1 is a human cancer gene which has been widely studied also in terms of evolutionary sequence changes. Much sequence data from different animals exist in EMBL/Genbank database.

Compare human BRCA1 sequences with non-human sequences.

Start with the sequence U14680.

Find out, by BLASTing (<http://blast.ncbi.nlm.nih.gov/Blast.cgi>), what kind of information exists⁽¹⁾.

Collect 5 different⁽²⁾ human sequences and ~50 non-human⁽³⁾ sequences.

Construct a FASTA-file⁽⁴⁾ from the sequences.

Align the sequences by using Clustal and perform neighbor-joining clustering (<http://www.ebi.ac.uk/Tools/clustalw2/index.html>)⁽⁵⁾.

Write a short report which includes some comments on the basis of the alignment. For example, are those nucleotide sites which are different among human sequences, also variable among different animals, or is it so that no evident patterns emerge. Don't include alignment files in your report, send the clustering picture ('phylogenetic tree') as this evidences your performance and shows what kind of a set of sequences you have collected⁽⁶⁾.

(1) If you blast with the default 'general parameters', you don't get answer to this. The window 'max target sequences' behind 'algorithm parameters' is 100, which gives you a list mainly consisting of human sequences and some non-humans. Set 250 (or more), so you also get the full list of non-humans. Note also that there are several BLAST-algorithms.

(2) Different = not identical or nearly identical. Inspect the 'scores' and choose those sequences which form 'as diverse as possible' collection of human sequences.

Human = *Homo sapiens*

(3) What is behind the binomial latin names? For example *Loxodonta africana*, Google this (if don't find it behind the respective accession number, it is an elephant.)

(4) A sequence retrieved in FASTA-format:

```
>gi|13195216|gb|AF284021.1|AF284021 Loxodonta africana BRCA1 gene, partial cds
ATGTGGCACAAATACTCATGCCAGCTCATTACAGCATAAGAACAGCAGTTTATTACTCACTAAAGACAGA
ATGAATGTAGAAAAGGCTGAATTCTGTAACAAAAGCAAACAGCCTGGCTTAGCAAGAAGCCAACAGAGCA.....
```

The title row is too long, a more convenient title (which subsequently appears in the alignment and clustering picture) would be >Loxodonta_AF284021, or >elephant_AF284021

(5) The most convenient way to use EBI-server is by e-mail, i.e. you get the results when they are ready, "interactive" communication is often very frustrating.

(6)and also indicates that you have performed a personal work.

