Exercises in MBV-INF 4410/9410/9410A - Sequence alignment and searching

Tuesday 27 November 2012

In these exercises we will start with a bacterial DNA repair protein called Nth and identify its homologs in different species including humans using BLAST and PSI-BLAST, and then identify conserved sequence motifs using multiple alignments. You should create a report document in Word (or a similar editor) were you describe briefly what you do, include figures you have made and answer the questions that are asked.

Exercise 1 - Sequence retrieval and multiple sequence alignment.

a) Find the protein sequence of the Endonuclease III (Nth) protein from the bacterium *Escherichia coli*, strain K-12, substrain MG1655, in the NCBI protein database (www.ncbi.nlm.nih.gov).

b) Copy the sequence, in FASTA format, into your report.

c) Retrieve the sequences of the Nth protein from *Mycobacterium tuberculosis* strain H37Rv (gi 57117142), *Bacillus antracis* strain Ames (gi 30261643), *Neisseria meningitidis* strain MC58 (gi 15676439), and *Streptococcus pneumoniae* strain R6 (gi 15903200) in FASTA format, and copy them into your report.

d) Edit the sequence titles to contain only the name of the bacteria and replace the spaces with the underscore character ("_"), but keep the initial larger-than character (">"), e.g. ">Escherichia_coli".

e) Start the JalView program (http://www.jalview.org/, click "Launch Jalview Desktop"). Close all the demo windows that the program initially opens automatically. Use "Input alignment from text box" to enter the five bacterial Nth sequences. Take a screenshot of Jalview with the input sequences using Alt-Print Screen and paste the image your report.

f) Click on "New Window". The use the embedded MUSCLE-algorithm (MuscleWS) (found under "Web service" - "Alignment") (with "Default settings") to generate a sequence alignment. Colour the amino acids according to "Percentage identity". Reformat the alignment to make it more compact ("Format - Wrap"). Remove extra information by un-hooking "Show annotations". Sort the sequences by pairwise similarity (Calculate>Sort>By Pairwise Identity). Adjust the width of the window appropriately. Export the alignment in png format, and import it into PowerPoint (or a similar program). Indicate the residues involved in the helix-hairpin-helix (HhH) motif (LxGVGxK) and the [4Fe–4S] (iron sulphur) cluster motif (CxxxxxxCxxCxxxxxC). See figure 3 in the article below for more information about these motifs. Copy the resulting figure into your report. Are both motifs fully conserved in all sequences?

Nora Goosen & Geri F. Moolenaar (2008) Repair of UV damage in bacteria. DNA Repair, Volume 7, Issue 3, Pages 353–379
http://dx.doi.org/10.1016/j.dnarep.2007.09.002

Exercise 2 - BLAST searches - alignments - phylogenetic trees

a) Using the sequence from *Escherichia coli* Nth as query, perform a protein BLAST search (http://blast.ncbi.nlm.nih.gov) against the Reference proteins database (Refseq protein). Limit the search to protein sequences from vertebrates. Set the max target sequences options to 500 under algorithm parameters.

b) From the resulting hits, select the following sequences: endonuclease III-like (Nth) (approx. 280-320 amino acids) and A/G-specific adenine glycosylase (MutY) (approx. 510-550 amino acids) from man (*Homo sapiens*), mouse (*Mus musculus*), cow (*Bos taurus*) and chicken (*Gallus gallus*). If there are several isoforms of the proteins, choose the one with the lowest isoform number. Also, if there are several entries for the same protein, select the one who has an accession starting with "NP_", or alternatively with "XP_". Retrieve the sequences in FASTA format, and paste them into the report. Shorten the titles of the sequences to contain only the protein name and the species (e.g. Nth_Homo_sapiens or MutY_Mus_musculus) (Keep the initital ">".)

c) As you did for the bacterial sequences, use JalView to generate a sequence alignment of the ten vertebrate Nth and MutY sequences, but this time use the MAFFT algorithm. Colour by percentage identity, turn on wrapping, turn off annotations, and sort the sequences by pairwise similarity. Import this alignment into PowerPoint or a similar program, and indicate the two sequence motifs. Copy the resulting figure into your report. Are both motifs fully conserved in all sequences?

d) From JalView, generate a phylogenetic tree from the alignment of the ten Nth and MutY proteins (Choose "Average distance using BLOSUM62"). Save the tree in png format, and import it into your report. Which proteins are most similar: human Nth and human MutY, or human Nth and chicken Nth?

Exercise 3 – Iterative BLAST searches and comparison of different alignment algorithms

a) Using the sequence of *E. coli* Nth as query, perform an **iterative** protein **PSI-**BLAST search against the NCBI Reference protein sequence database (Refseq protein). Before doing the search, limit the search to vertebrate sequences, and change the "PSI-BLAST threshold" from the default value of 0.005 to 0.00005 (=5e-5). After convergence (or at least five iterations), reformat the results to include only human (*Homo sapiens*) sequences. From the results, select sequences corresponding to the four human homologs denoted Endonuclease III-like protein 1 (NTHL1) (312 aa), A/G-specific adenine DNA glycosylase isoform 1 (MUTYH) (546 aa), N-glycosylase/DNA lyase isoform 1a (OGG1) (345 aa) and methyl-CpG-binding domain protein 4 (MBD4) (580 aa). The MBD4 protein may have an E-value worse than the PSI-BLAST threshold, but is still a homolog. Give the sequences short names.

b) Make a multiple sequence alignment of the four sequences, using the MUSCLE program from JalView. Format the alignment as earlier. Then try the MAFFT and ClustalW programs. Import the three sequence alignments into your report.

c) Are the HhH motif and the [4Fe–4S] cluster motif present in all four sequences? Note that the first 400 residues in the N-terminal of MBD4 are unrelated to the other proteins, and any similarity to that N-terminal part of the MBD4 protein is invalid.

d) Judging from the proper alignment of residues in the two motifs, which of the programs has produced the worst alignment?

e) Finally, make MUSCLE and MAFFT alignments where you also include the bacterial and vertebrate Nth and MutY sequences from Exercise 1 and 2, without duplicating the human Nth and MutY. Format the alignment as earlier. Include the alignments in your report, but crop the images so that roughly only the region with the HhH motif is shown. Is any of the programs able to correctly align the HhH motif when all sequences are included? Which program performed best?