torognes@ifi.uio.no jonkl@medisin.uio.no

# Sequence searching and sequence alignments – MBV-INFX410

***NEW TASK 32! DO THIS INSTEAD OF THE TASK 32 IN THE ORIGINAL EXERCISE.***

32. 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 (taxid: 7742), chose to "Exclude" "Models (XM/XP)" (check this box), set the max target sequences options to 1000 under algorithm parameters, and keep the "PSI-BLAST threshold" default value of 0.005. Every time you have run one iteration, you have to click "Go" at "Run PSI-Blast iteration *n* with max 1000" to run the next iteration, iteration *n*. After convergence (or at least four 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). Give the sequences short names. After each iteration, check how many hits you have.

    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.

NB! 2016 - no new sequences were found. Copy the fasta sequences on next page for excercise 33

**The 1st iteration is identical to ordinary blastp and you get 11 hits (6 in human, both NTHL1, NP_002519, and MUTYH, NP_036354, the same as you got in the blastp task earlier). The 2nd iteration gives 14 hits (3 new, but no new human sequences). The 3rd iteration gives 22 hits (8 new, all of the isoforms of human OGG1, for example NP_002533). The MBD4 sequence (NP_003916) is further down (E-value WORSE than the threshold). The 4th iteration does not give anything new and PSI-BLAST has converged.**

**You see that running PSI-BLAST (2 iterations or more) finds more remote homologs of *E. coli* Nth. Human OGG1 will not be detected as a homolog with blastp only. Same with MBD4, but it is actually always "under the treshold" even in PSI-BLAST. We still include it here since it is a known DNA repair protein and possibly a homolog anyway. When we align all the sequences later in the exercise we see that this was correct. All these 4 sequences are homologs to *E. coli* Nth.**

***CONTINUE IN THE ORIGINAL EXERCISE***

**APPENDIX 6:**

**4 human homologs, original headers**

>gi|4505471|ref|NP_002519.1| endonuclease III-like protein 1 [Homo sapiens]
MCSPQESGMTALSARMLTRSRSLGPGAGPRGCREEPGPLRRREAAAEARKSHSPVKRPRKAQRLRVAYEG
SDSEKGEGAEPLKVPVWEPQDWQQQLVNIRAMRNKKDAPVDHLGTEHCYDSSAPPKVRRYQVLLSLMLSS
QTKDQVTAGAMQRLRARGLTVDSILQTDDATLGKLIYPVGFWRSKVKYIKQTSAILQQHYGGDIPASVAE
LVALPGVGPKMAHLAMAVAWGTVSGIAVDTHVHRIANRLRWTKKATKSPEETRAALEEWLPRELWHEING
LLVGFGQQTCLPVHPRCHACLNQALCPAAQGL
>gi|6912520|ref|NP_036354.1| A/G-specific adenine DNA glycosylase isoform 1 [Homo sapiens]
MTPLVSRLSRLWAIMRKPRAAVGSGHRKQAASQEGRQKHAKNNSQAKPSACDGMIAECPGAPAGLARQPE
EVVLQASVSSYHLFRDVAEVTAFRGSLLSWYDQEKRDLPWRRRAEDEMDLDRRAYAVWVSEVMLQQTQVA
TVINYYTGWMQKWPTLQDLASASLEEVNQLWAGLGYYSRGRRLQEGARKVVEELGGHMPRTAETLQQLLP
GVGRYTAGAIASIAFGQATGVVDGNVARVLCRVRAIGADPSSTLVSQQLWGLAQQLVDPARPGDFNQAAM
ELGATVCTPQRPLCSQCPVESLCRARQRVEQEQLLASGSLSGSPDVEECAPNTGQCHLCLPPSEPWDQTL
GVVNFPRKASRKPPREESSATCVLEQPGALGAQILLVQRPNSGLLAGLWEFPSVTWEPSEQLQRKALLQE
LQRWAGPLPATHLRHLGEVVHTFSHIKLTYQVYGLALEGQTPVTTVPPGARWLTQEEFHTAAVSTAMKKV
FRVYQGQQPGTCMGSKRSQVSSPCSRKKPRMGQQVLDNFFRSHISTDAHSLNSAAQ
>gi|4505495|ref|NP_002533.1| N-glycosylase/DNA lyase isoform 1a [Homo sapiens]
MPARALLPRRMGHRTLASTPALWASIPCPRSELRLDLVLPSGQSFRWREQSPAHWSGVLADQVWTLTQTE
EQLHCTVYRGDKSQASRPTPDELEAVRKYFQLDVTLAQLYHHWGSVDSHFQEVAQKFQGVRLLRQDPIEC
LFSFICSSNNNIARITGMVERLCQAFGPRLIQLDDVTYHGFPSLQALAGPEVEAHLRKLGLGYRARYVSA
SARAILEEQGGLAWLQQLRESSYEEAHKALCILPGVGTKVADCICLMALDKPQAVPVDVHMWHIAQRDYS
WHPTTSQAKGPSPQTNKELGNFFRSLWGPYAGWAQAVLFSADLRQSRHAQEPPAKRRKGSKGPEG
>gi|4505121|ref|NP_003916.1| methyl-CpG-binding domain protein 4 [Homo sapiens]
MGTTGLESLSLGDRGAAPTVTSSERLVPDPPNDLRKEDVAMELERVGEDEEQMMIKRSSECNPLLQEPIA
SAQFGATAGTECRKSVPCGWERVVKQRLFGKTAGRFDVYFISPQGLKFRSKSSLANYLHKNGETSLKPED
FDFTVLSKRGIKSRYKDCSMAALTSHLQNQSNNSNWNLRTRSKCKKDVFMPPSSSSELQESRGLSNFTST
HLLLKEDEGVDDVNFRKVRKPKGKVTILKGIPIKKTKKGCRKSCSGFVQSDSKRESVCNKADAESEPVAQ
KSQLDRTVCISDAGACGETLSVTSEENSLVKKKERSLSSGSNFCSEQKTSGIINKFCSAKDSEHNEKYED
TFLESEEIGTKVEVVERKEHLHTDILKRGSEMDNNCSPTRKDFTGEKIFQEDTIPRTQIERRKTSLYFSS
KYNKEALSPPRRKAFKKWTPPRSPFNLVQETLFHDPWKLLIATIFLNRTSGKMAIPVLWKFLEKYPSAEV
ARTADWRDVSELLKPLGLYDLRAKTIVKFSDEYLTKQWKYPIELHGIGKYGNDSYRIFCVNEWKQVHPED
HKLNKYHDWLWENHEKLSLS