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

Sequence searching and sequence alignments – MBV-INFX410

In this exercise 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. It is a good idea to create a report document in Word (or a similar editor) were you describe briefly what you do, save the sequences that you work with and answer the questions that are asked. You must also save screen shots of what you do in your report. *Make sure you know how to do this!*

1. Find the RefSeq protein sequence of the Endonuclease III (Nth) protein from the bacterium *Escherichia coli*, strain K-12, substrain MG1655, using Entrez, in the NCBI protein database (www.ncbi.nlm.nih.gov). *First try yourself, without checking below!*

   For the rest of the exercise, it is a good idea to sign up for a "My NCBI" account and sign in. Follow the link "Sign in to NCBI" at the top/right hand side of the front page, to do this. When you are signed in, you can, for example, save your searches and pick them up at a later stage to do more work.

   There are many ways to find the correct Nth protein, but what we are looking for is the RefSeq sequence NP_416150. One possibility is to search for "Escherichia coli" AND "Endonuclease III" AND MG1655 in the Protein database and then filter for RefSeq in "Source databases". You will then have some 10s of candidates and among these the only one that is MG1655, "Endonuclease III", and RefSeq is NP_416150. *Make sure you understand how you find a sequence by searching like this!*

2. Get the FASTA sequence for the protein and paste it into your report document.

   >gi|16129591|ref|NP_416150.1| DNA glycosylase and apyrimidinic (AP) lyase (endonuclease III) [Escherichia coli str. K-12 substr. MG1655]
   MNKAKRLEILTRLRENNPHPTTELNFSSPFELLIAVLLSAQATDVSVNKATAKLYPVANTPAAMLELGVE
   GVKTYIKTIGLYNSKAENIIKTCRILLEQHNGEVPEDRAALEALPGVGRKTANVVLNTAFGWPTIAVDTH
   IFRVCNRTQFAPGKNVEQVEEKLLKVVPAEFKVDCHHWLILHGRYTCIARKPRCGSCIIEDLCEYKEKVD
   I

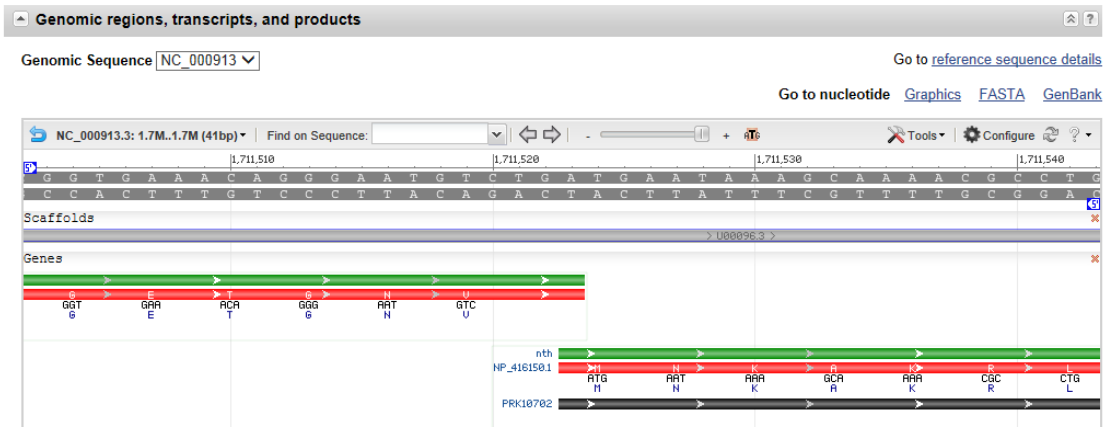
3. Follow the link to the corresponding Gene (in the list of "Related information" at the right hand side). What is the NCBI gene identifier (Gene ID) for the gene? What is the Swiss-Prot identifier for this protein? Which genes are found directly upstream and downstream of *nth*? Are the three genes transcribed in the same direction?

The Gene ID is 947122 and the Swiss-Prot ID is P0AB83. The two neighbouring genes are *rsxE* and *dtpA*. All three genes are transcribed in same direction.

4. In the simple genome browser on the NCBI Gene page ("Genomic regions, transcripts, and products" section) click and drag the genome to centre the region where you have the start of *nth* and the stop of *rsxE*. Then zoom in all the way to the highest possible magnification by using the slider and/or the "+" and "ATG" buttons. Make sure the start of *nth* stays in the middle of your browser by click-dragging, if necessary. What are the three nucleotides of the codon encoding the last (C-terminal) amino acid in *rsxE*? What are the nucleotides of the stop codon of *rsxE* and the start codon of *nth*? How many nucleotides are there between the start of *nth* and the stop of *rsxE*?

   The last amino acid of *rsxE* is Val, encoded by GTC, and the stop codon is TGA. The start codon of *nth* is encoded by ATG. In this case the A of TGA (stop) is the same as the A of ATG (start). The two genes overlap by a single nucleotide, and there are, obviously, no nucleotides between them.

   Notice how densely packed the genes are in bacteria compared to, for example, the vertebrates.



5. Get the homologous sequences of the Nth protein from *Mycobacterium tuberculosis* strain H37Rv (GI number 57117142), *Bacillus antracis* strain Ames (GI number 30261643), *Neisseria meningitidis* strain MC58 (GI number 15676439), and *Streptococcus pneumoniae* strain R6 (GI number 15903200) in FASTA format, and copy them into your report.
6. Edit the sequence titles to contain only the name of the bacteria and the RefSeq identifiers. Replace the spaces with the underscore character ("_"), but keep the initial larger-than character (">"). For your first sequence, the header will then be ">Escherichia_coli_NP_416150".

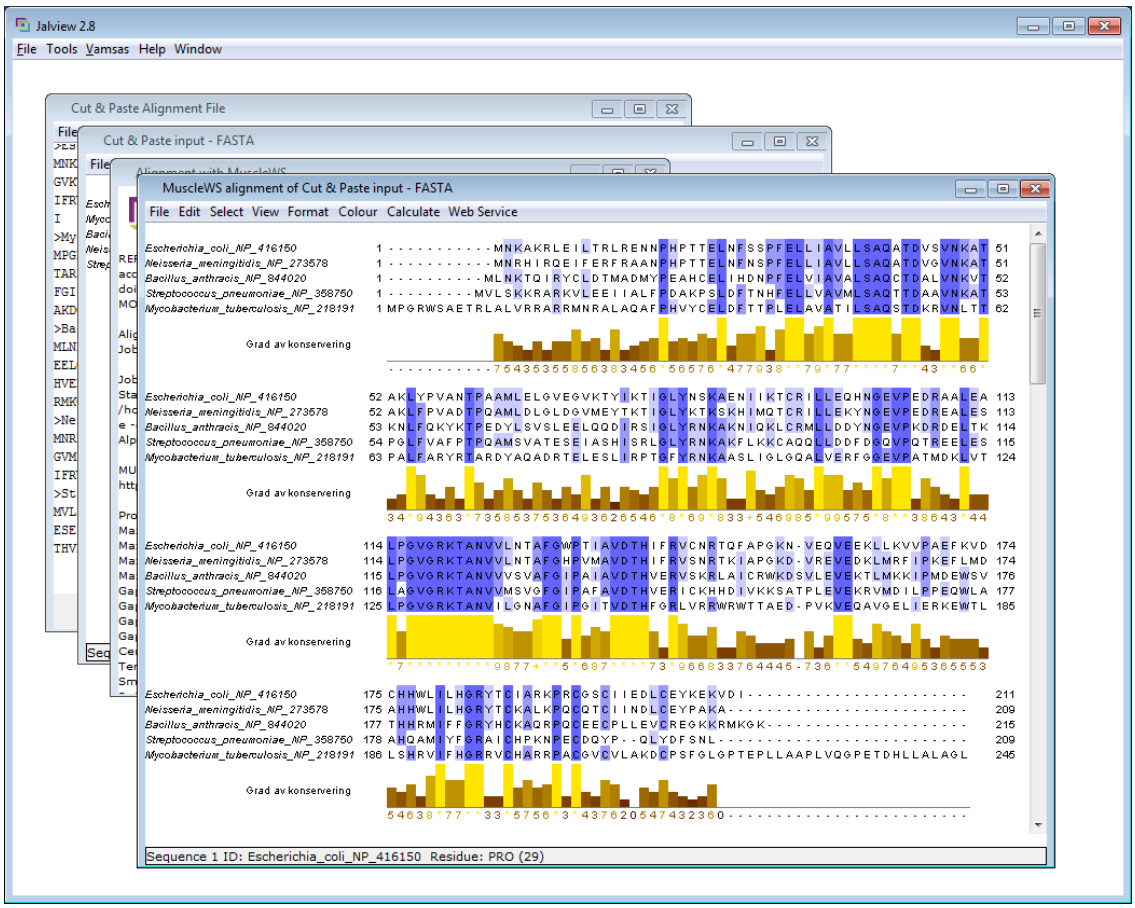
7. Start Desktop JalView. Use “File” → “Input Alignment” → “from Textbox” to enter the five bacterial Nth sequences by copying and pasting. Click on “New Window”. Take a screenshot of Jalview with the input sequences and paste the image into your report.

8. Use the MUSCLE-algorithm web service (found under “Web Service” → “Alignment”) (with “Muscle with Defaults”) to generate a multiple sequence alignment (MSA). Can you say anything about on which computer the MUSCLE-algorithm is running? Where on the planet? Hint: Look in “Tools” → “Preferences” → “Web Services”.

   The job is running as a web service on a server that at least contains the name “dundee”. The service is, very likely, running in Dundee, Scotland, on a server belonging to the group of Professor Geoff Barton. This is the group that is developing Jalview.

9. Colour the amino acids according to “Percentage identity”. Remove the “Quality” annotation information in the lower part of the window by right-clicking on the word “Quality” and choose “Hide this row”. Do the same for the “Consensus” annotation. Right-click on “Conservation” and choose “Edit Label/Description”. Change the “Annotation name” to your native language. For example, in Norwegian use the text “Grad av konservering”, then click “OK”. Sort the sequences by pairwise similarity (“Calculate” → “Sort” → “By Pairwise Identity”). Reformat the alignment to make it

more compact ("Format" → "Wrap"). Adjust the width of the window so that you get the MSA split into 4 lines/blocks. Also remove the tick mark at "Show Sequence Limits" under "Format".

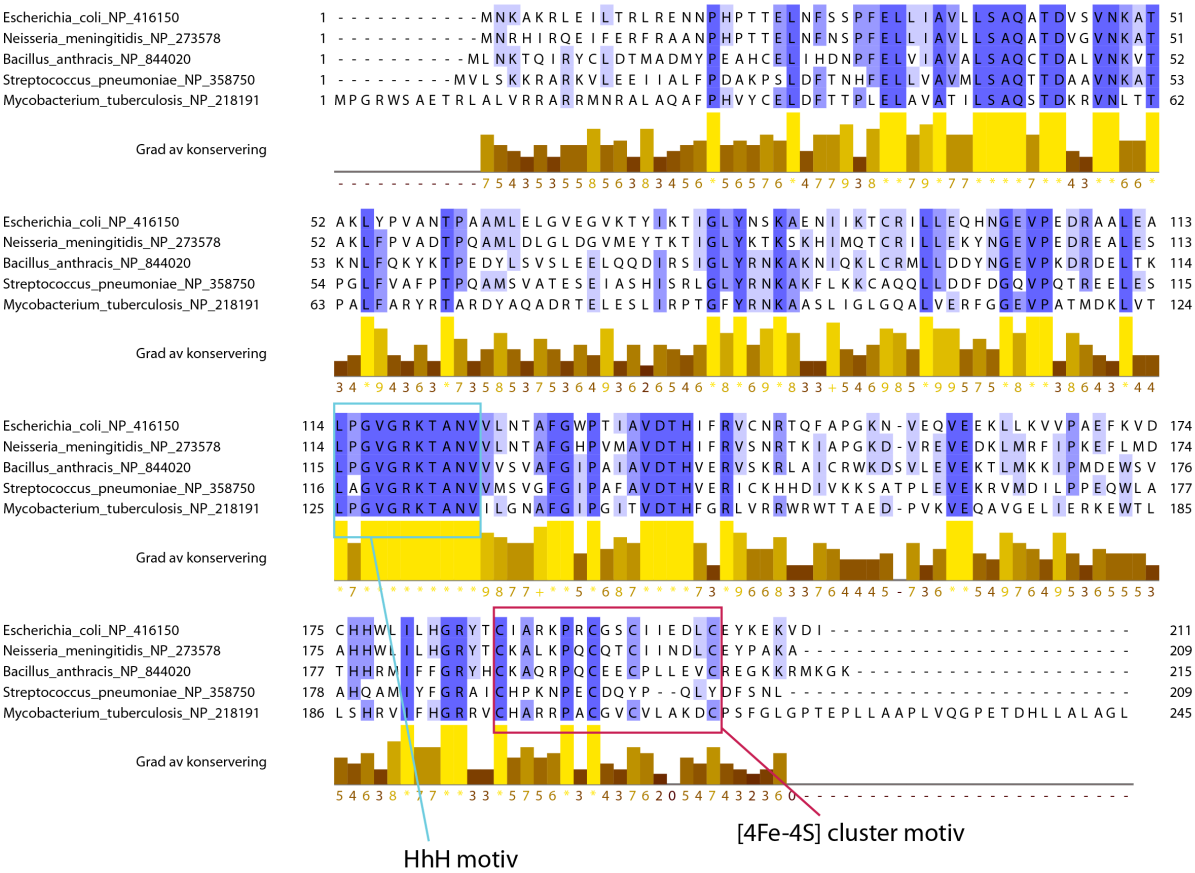


Export the alignment in PNG format, and import it into PowerPoint, Adobe Illustrator, or a similar program in order to add some extra information in the MSA. Indicate the residues involved in the helix-hairpin-helix (HhH) motif (LxGVGxK) and the [4Fe–4S] (iron sulphur) cluster motif (CxxxxxxCxxCxxxxxC). See Fig. 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?

N. Goosen & G.F. Moolenaar, "Repair of UV damage in bacteria", DNA Repair **7**, 353 (2008) **http://dx.doi.org/10.1016/j.dnarep.2007.09.002**

The HhH motif is conserved in all species, with GVGRKTANV being fully conserved. The [4Fe-4S] cluster is conserved in all species except Streptococcus, which lacks the two last cysteines.

Sammenstilling av fem Nth homologer fra bakterier



10. Find the percentage sequence identity between the five sequences. First select all the sequences in the Jalview MSA window, for example by typing <ctrl>-a. Btw, if you want to select nothing, press the <Esc> key. Try this. Select all sequences again and do "Calculate" → "Pairwise Alignments…". Look at the pairwise alignments and find which two sequences are the most similar. Which are they? What is the sequence identity between those two sequences?

    ***E. coli* and *N. meningitidis* are 72% identical while no other pairs are above 47%**

11. Using the sequence from *E. coli* Nth as query, perform a protein BLAST (blastp) search at the NCBI website (http://blast.ncbi.nlm.nih.gov). Use "Basic BLAST" and "protein blast" and search in the RefSeq protein database. Limit the search to protein sequences from vertebrates. Set the max target sequences options to 5000 under algorithm parameters. Also set "Word size" to 3. Why do we choose blastp in this case and not tblastn?

We are searching with a protein query sequence in a protein sequence database, hence blastp. tblastn is used for searching with a protein sequence in a translated nucleotide database.

12. How many hits do you get? The easiest way to find this out is *not* by counting, but by jumping down to "Descriptions" and click "All" in "Select: All None". How many homologs of *E. coli* Nth do you find in vertebrates?

    1015 hits Nov. 20 2016

    On Nov 14, 2014, there are 591 hits, but this number will most likely change, and grow, fast. There are *not* necessarily 591 homologs of *E. coli* Nth in vertebrates here since the maxium threshold for E-value was set to 10 (as default). Many of the hits are "random hits" with E-value approaching this value.

13. We could *define* an *E. coli* Nth homolog as a hit with E-value better (lower) than 0.01 (but we could also have chosen a different value). Do this, and check how many hits/homologs you find now. *Hint:* Use the "Formatting options" at the top and set "Expect Max:" to 0.01, press "Reformat", and now count the number of hits.

    On Nov 14, 2014, I get 477 hits with E-value better than 0.01. These are most likely homologs (with a common ancestor gene with *E. coli* Nth). 806 hits Nov. 20 2016

14. What is the top hit with the best E-value? Write the accession identifier in your document. Check also hit number 2 and 3 on the list, then 4 and 5. Which species are these sequences from? What is the sequence identity between *E. coli* Nth and these hits. What appears to be, very roughly on average, the sequence identity between *E. coli* Nth and vertebrate Nth-like proteins.

    Hits 1 to 3 are from *Pantholops hodgsonii*, the Tibetan antelope or chiru. The best hit has identifier XP_005981298. Number 4 is from *Elephantulus edwardii*, the Cape elephant shrew, and 5 from *Chrysochloris asiatica*, the Cape golden mole. Sequence identities between *E. coli* Nth and these homologs are 55%, 55%, 49%, 32%, and 34%. Most of the other vertebrate Nth-like homologs are roughly 30% identical to *E. coli* Nth.

15. From the resulting BLAST hits, select the following sequences: endonuclease III-like (Nth) (approx. 280-360 amino acids) and A/G-specific adenine glycosylase (also known as MutY) (approx. 510-720 amino acids) from man (*Homo sapiens*), mouse (*Mus musculus*), cow (*Bos taurus*), chicken (*Gallus gallus*), frog (*Xenopus tropicalis*), and the fugu pufferfish (*Takifugu rubripes*). 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

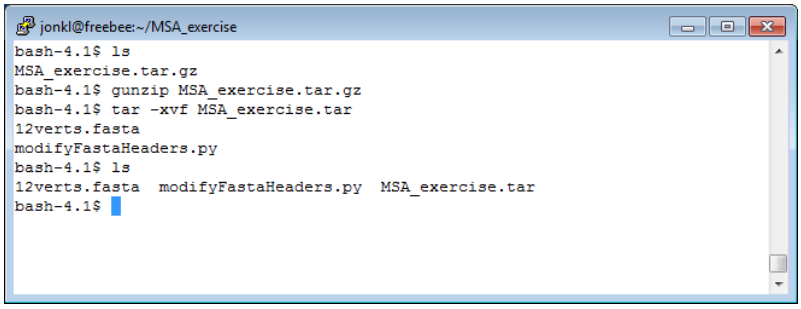
"NP_", or alternatively with "XP_". We do not have time to look very closely at all the sequences and their splice variants, but if we wanted to do serious work with these sequences, we would have to do that. We should, for example, have checked if there is something obviuosly wrong with the splicing of the sequences. Retrieve the sequences in FASTA format, and paste them into the report. Make sure you are able to do this properly, at least for human, mouse, and cow, before you continue below. Can you use some of the options under "Formatting options" (at the top of the page) to make this task easier? Why choose sequences with "NP_" identifier, rather than "XP_"?

If you, under "Formatting options" filter on "Organism", you get only a few hits for each organism and finding the relevant ones is much much easier than if you work with the full list. Sequences with "XP_" identifiers are "models" (see lecture notes from the first day of the course) and have almost certainly never been manually curated, while "NP_" sequences at least possibly have been checked a bit better.

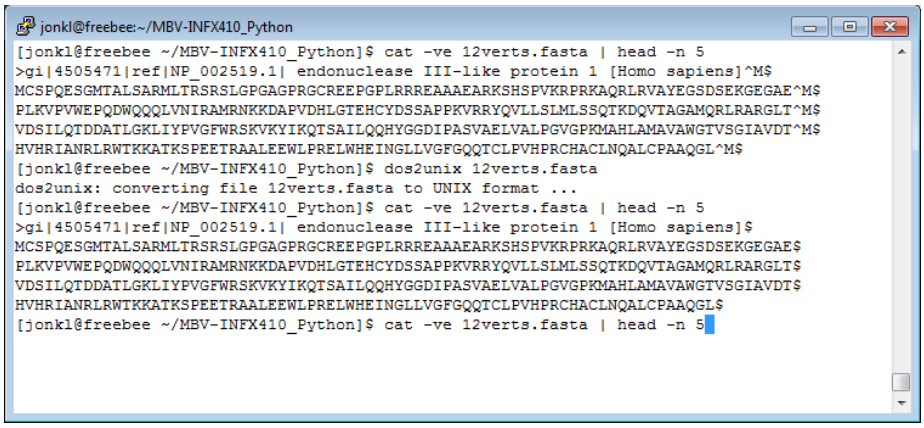
16. We want, as for the bacterial protein sequences, to shorten each sequence title to contain only the species name and RefSeq identifier. We could do this manually, in a text editor, as we did above for the bacterial homologs. However, the task here will be to use a little program or script to do this. If we had hundreds of sequences, making a script would certainly be quicker and less error prone. If we had even more sequences, a script would be the only option.

    Log onto freebee.abel.uio.no, and create a new directory that you will work in. Call it, for example, "MSA_Exercise". Download the file MSA_exercise.tar.gz from the wiki page and put it in the new directory. How you do this will depend on your laptop. When you have done this, make sure you understand what you did. We will do similar operations more times during the course *(and very likely for the exam…)*. *This is important!* https://github.com/jonbra/MBV-INFx410/raw/master/MSA_exercise.tar.gz
17. The file has a double ending, ".tar.gz". This indicates that this is a compressed file that has been compressed, or packed to save space, by the gzip software application (hence the ".gz"). It is also a "tar file", also known as a "tarball", which usually means that many files have been packed into a single file. This is often done to make file transfer and/or file storage easier. Use *ls -l MSA_exercise.tar.gz* to see the size of the compressed file.
18. Uncompress the file by running the command *gzip -d MSA_exercise.tar.gz* (*gunzip MSA_exercise.tar.gz* will do exactly the same and is possibly easier to remember). Do *ls -l* to see what you have now. Notice that the uncompressed file is much bigger than the "gzipped" version. *gzip* and other compression applications are very useful to save disk space and speed up file transfer.

19. Now pack out all the files in the tarball archive file by running *tar -xvf MSA_exercise.tar*. Here "-x" tells *tar* to "extract" all files in the archive, "-f" tells *tar* to extract them from the file MSA_exercise.tar (and not, for example, from a tape station), and "-v" tells *tar* to be "verbose" and print to the terminal what it is doing. Of course, you can read more about *gzip* and *tar* by using the *man* command. Now do *ls -l* to find out what you have in your current directory.



20. You find the 12 vertebrate Nth homologs in the file 12verts.fasta. Make sure the file has correct Unix format with Unix end-of-lines by using *cat -ve*. Use *man cat* to find out what the "*-ve*" is doing. If the end-of-lines are not correct, fix the problem with the command *dos2unix.*



21. Your task is to open the file 12verts.fasta and change all headers to the correct format (*e.g.* ">Homo_sapiens_NP_002519" for the human Nth homolog). Change all spaces to "_" and, of course, keep the initial ">". Then write out a new Fasta file, identical to the original one, but with modified and simplified Fasta headers. Call the new file 12verts_new.fasta.

```
>gi|4505471|ref|NP_002519.1| endonuclease III-like protein 1 [Homo sapiens]
MCSPQESGMTALSARMLTRSRSLGPGAGPRGCREEPGPLRRREAAAEARKSHSPVKRPRKAQRLRVAYEGSDSEKGEGAE
PLKVPVWEPQDWQQQLVNIRAMRNKKDAPVDHLGTEHCYDSSAPPKVRRYQVLLSLMLSSQTKDQVTAGAMQRLRARGLT
...
```

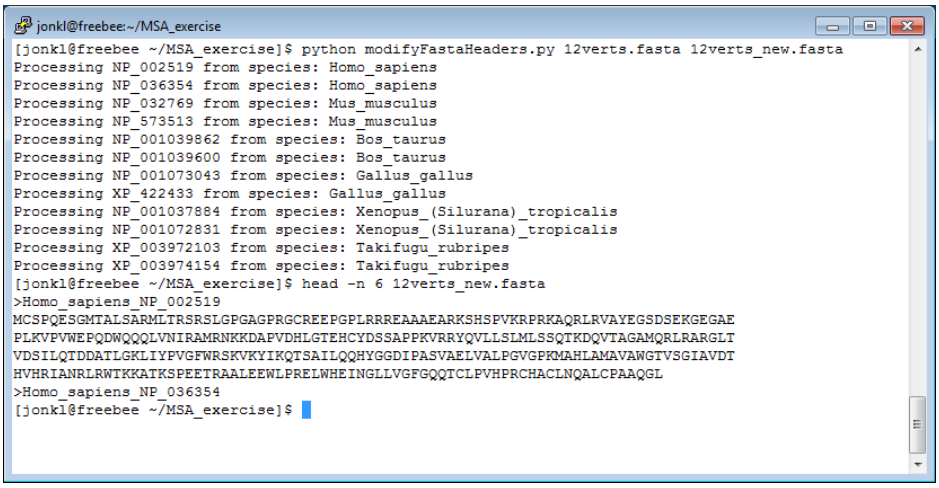
should become

```
>Homo_sapiens_NP_002519
```

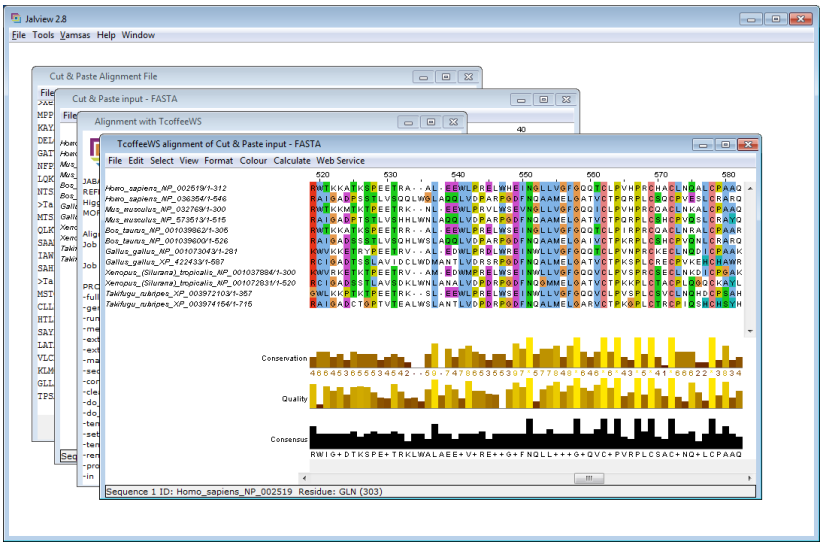
```
MCSPQESGMTALSARMLTRSRSLGPGAGPRGCREEPGPLRRREAAAEARKSHSPVKRPRKAQRLRVAYEGSDSEKGEGAE
PLKVPVWEPQDWQQQLVNIRAMRNKKDAPVDHLGTEHCYDSSAPPKVRRYQVLLSLMLSSQTKDQVTAGAMQRLRARGLT
...
```

and so on.

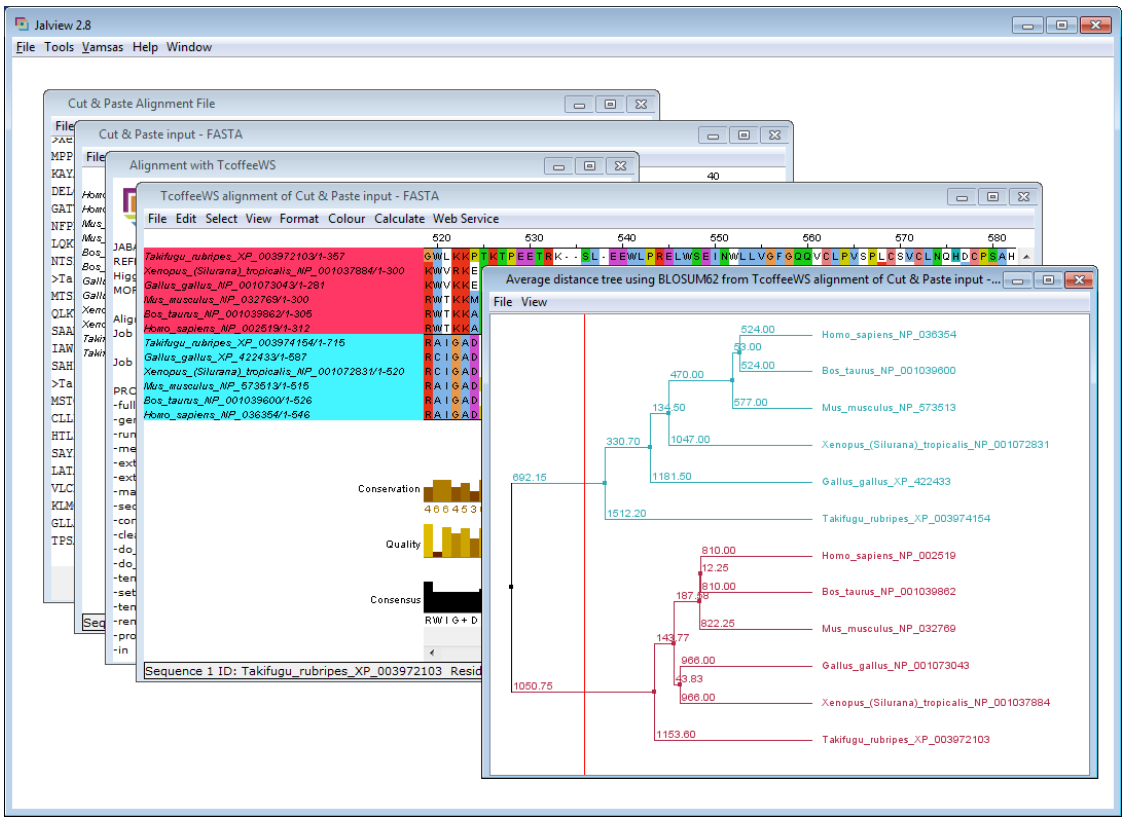
22. Write down the steps, or a little flowchart, that shows what the script has to do in order to solve the task.
23. If you have done any programming before *or* you want a challenge, choose (a) below, otherwise do (b).
    a. Make a script in a programming language of your own choice that does the file conversion described above
    b. Take a look at the python script modifyFastaHeaders.py you found in the tarball MSA_exercise.tar.gz. Go through it step by step and make sure you understand what it will do. Use this script to do the file conversion



24. As you did for the bacterial sequences, use Jalview to generate an MSA for the twelve vertebrate sequences, but this time use the T-Coffee algorithm (with default settings).



25. In Jalview, generate a phylogenetic tree from the alignment of the twelve proteins (Choose “Calculate” → “Calculate Tree” → ”Average Distance Using BLOSUM62”). Click in the tree window to get different colours on the two clades in the tree (See below). Then, in the MSA windows, do “Calculate” → “Sort” → “By Tree Order” and choose sorting according to the tree you just generated.



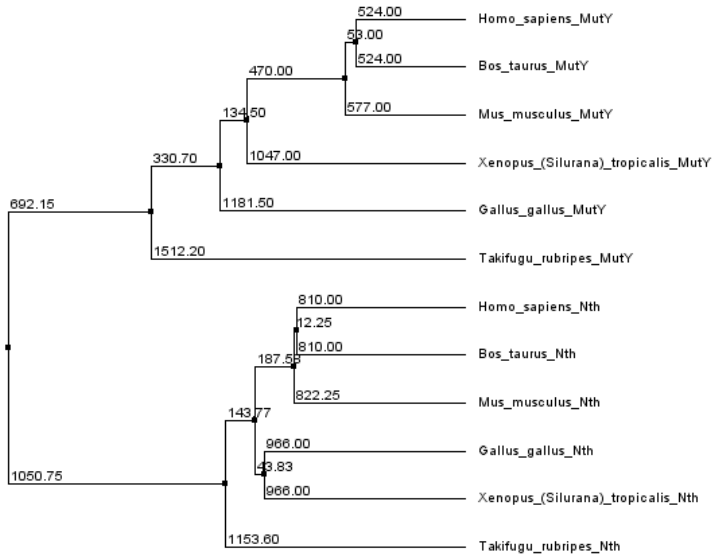
26. Use the terms homologs, paralogs, and orthologs to describe the relationships between these proteins/genes.

NP_036354 is human MutY (with official gene name and symbol “mutY homolog” and MUTYH) while NP_002519 is human Nth (with official gene name and symbol “nth endonuclease III-like 1 (E. coli)” and NTHL1). All the other “blue” sequences/nodes in the figure above are orthologs of human MUTYH. They are unique genes/proteins due to a speciation event. Similarly, all the nodes in the “red” clade are orthologs of NTHL1. NTHL1 and MUTYH are paralogs, due to a gene duplication. All the sequences are homologs.

27. We now change the names of the sequences a final time and put Nth in all the headers of the NTHL1 orthologs and MutY in all the headers of the MUTYH orthologs. Open the file 12verts_new.fasta in a text editor and change the headers so that all the Nth orthologs are named by their species and Nth (*e.g.* Homo_sapiens_Nth),

while all MutY homologs are named with MutY (*e.g.* Homo_sapiens_MytY). Do this manually, or with a script. Save the new Fasta file as 12verts_final.fasta.

28. Get the sequences from 12verts_final.fasta into Jalview and generate an MSA with the T-Coffee algorithm, as above. Also generate a phylogenetic tree and sort as above. Save the tree in PNG format, and import it into your report. Are all the clades as you would expect?

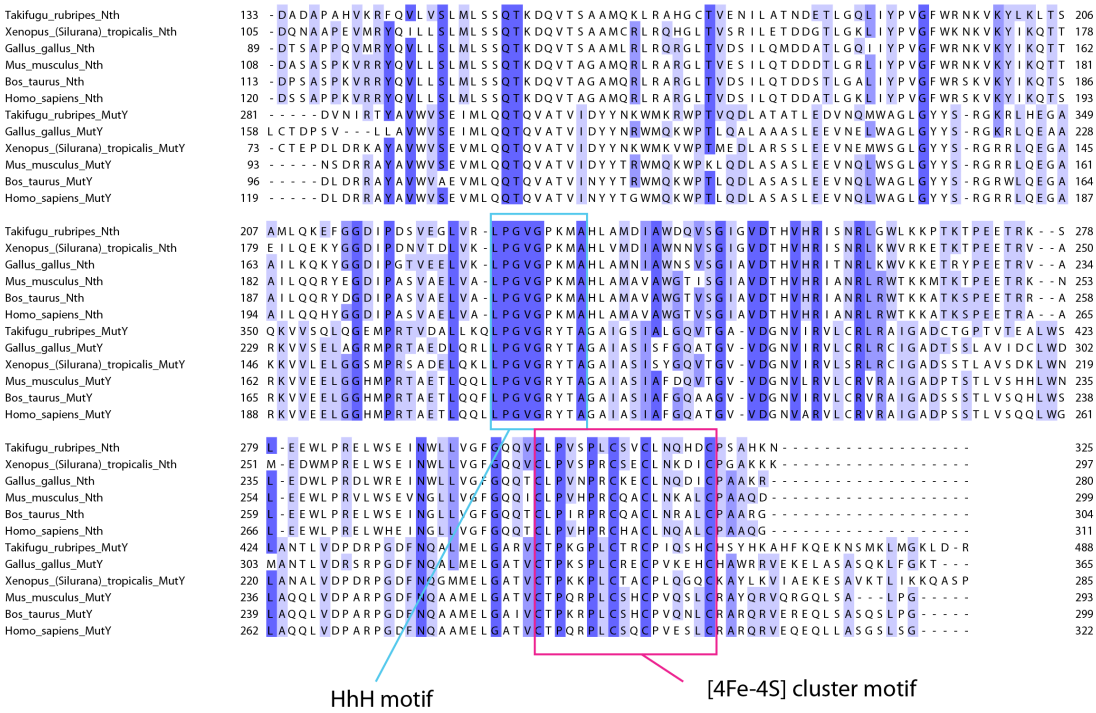


One would expect, in each of the two main clades, that the orthologs in human, mouse and cow were most similar, as they also are. Further out one would expect to encounter first chicken, then frog, and finally the fish orthologs. The fish clades are where they should be, but the frog/chicken is swapped (MutY) or in their own clade (Nth).

It is certainly not possible to make reliable phylogenetic species trees from just a single gene, but another problem here is that the phylogenetic tree building algorithms in Jalview are *very* simple. They can certainly not be used in a publication, for example!

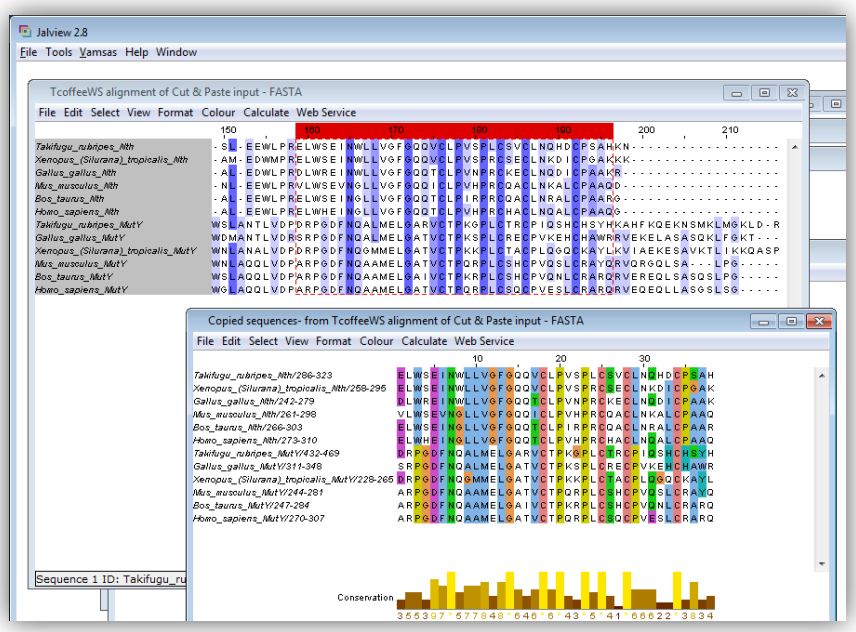
This is a good review on how to make reliable trees: Z. Yang & B. Rannala, "Molecular phylogenetics: principles and practice", Nat. Rev. Genet. 13, 303 (2012).

29. For the MSA, colour by percentage identity, turn off all annotations (remove the tick mark at “View” → “Show Annotations”), and use “Edit” → “Remove Left/Right” to trim the MSA and only keep the core part that is relatively conserved in all the sequences (roughly corresponding to human MutY residues 120 – 320). Turn on wrapping, and export the MSA as a PNG file. 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?

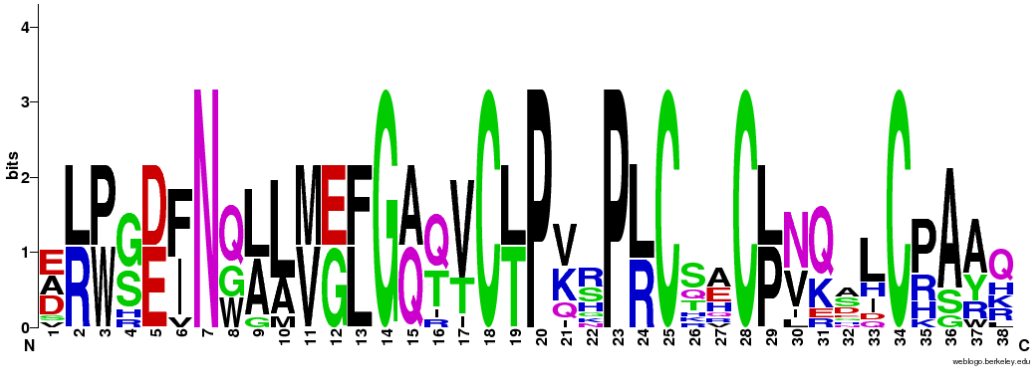


The 4 Cys residues of the [4Fe-4S] cluster motif are 100% conserved in all homologs. The motif LPGVGxxxA is conserved in all sequences where xxx is PKM or RYT in Nth and MutY, respectively.

30. Select a chunk of the MSA between human MutY residues 270 and 307 containing the [4Fe-4S] cluster motif. Do this by left-clicking just above the MSA, next to “160” (See below), and pull to the right while holding down the mouse button. Select the “red region” below. Copy this segment by pressing <ctrl>-c, and paste this into a new window by pressing <ctrl>-<shift>-v. See below. It should look something like this!:



31. Now let us make a sequence logo for this segment. Go to the following website, http://weblogo.berkeley.edu, and follow the link “create”. In Jalview, get the MSA for our [4Fe-4S] cluster motif segment in Fasta format by doing “File” → “Output to Textbox” → “FASTA”. Copy the Fasta format text into the window on the WebLogo website. Then press “Create Logo”. Put the logo in your report. Take a screen-shot, for example.



The logo gives a good illustration of which residues are conserved in this protein family and which are not.

32. NB! Check out new version of task 32 on the wiki pages

f *E. coli* Nth as query, perform an iterative protein PSI-BLAST CBI Reference protein sequence database (Refseq protein). rch, limit the search to mammalian sequences, set the max target sequences options to 1000 under algorithm parameters, and change the “PSI-BLAST threshold” from the default value of 0.005 to 0.0002. After convergence (or at least three 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.

MUSCLE:

```
NTHL1_Homo_sapiens     1 -------------MCSPQESGMTALSARMLT--RSRSLGPGAGPRGCREEPGPLRR-----------------------REAAAEARKSHSPV 55
MUTYH_Homo_sapiens     1 -------------MTPLVSRLSRLWAIMRK--PRAAVGSGH-----------RK-----------------------QAASQEGRQKHAKN 42
OGG1_Homo_sapiens      1 ----MPARALLPRRMGHRTLASTPALWASIPC--PRSELRLDL----VLPSGQSFRW-----------------------REQSPAHWSGVLAD 61
MBD4_Homo_sapiens      1 MGTTGLESLSLGDRGAAPTVTSSERLVPDPPNDLRKEDVAMELERVGEDEEQMMIKRSSECNPLLQEPIASAQFGATAGTECRKSVPCGWERVVKQ 96

NTHL1_Homo_sapiens    56 K-RPRKAQRLRVAYEGSDSEKGEGAEPLKVPV-------WEPQDWQQQLVNIRAMRNK--------KDAPVDHLGTEHCYDSSAPPKV----RRYQ 131
MUTYH_Homo_sapiens    43 NSQAKPSACDGMIAECPGAPAGLARQPEEVVL----QASVSSYHLFRDVAEVTAFRGS--------LLSWYDQEKRDLPWRRRAEDEMDLDRRAYA 126
OGG1_Homo_sapiens     62 QVWTLTQTEEQLHCTVYRGDKSQASRPTPDEL----EAVRKYFQLDVTLAQLYHHWGS--------VDSHFQEVAQKFQGVRLLRQDP------IE 139
MBD4_Homo_sapiens     97 RLFGKTAGRFDVYFISPQGLKFRSKSSLANYLHKNGETSLKPEDFDFTVLSKRGIKSRYKDCSMAALTSHLQNQSNNSNWNLRTRSKC-----KKD 187

NTHL1_Homo_sapiens   132 VLLSLMLSSQTKDQVTAGAMQRLRA----RGLTVDSILQTDDATLGKLIYP----------VGFWRSKVKYIKQTSAILQQHYGGDIPASV-----208
MUTYH_Homo_sapiens   127 VWVSEVMLQQTQVATVINYYTGWMQ----KWPTLQDLASASLEEVNQLWAG----------LGYY-SRGRRLQEGARKVVEELGGHMPRTAE----203
OGG1_Homo_sapiens    140 CLFSFICSSNNNIARITGMVERLCQAFGPRLIQLDDVTYHGFPSLQALAGPEVEAHLRKLGLGY---RARYVSASARAILEEQGGLAWLQQLRESS 232
MBD4_Homo_sapiens    188 VFMPPSSSSELQESRGLSNFTSTHL----LLKEDEGVDDVNFRKVRKPKGK----------VTIL--KGIPIKKTKKGCRKSCSGFVQSDSKRESV 267

NTHL1_Homo_sapiens   209 -----AELVALPGVGPKMAHLAMAVAWGTV-SGIAVDTHVHRIANR-LRWTKKATKSPEETRAALEEW--LPRELW----HEINGLL-------VG 284
MUTYH_Homo_sapiens   204 -----TLQQLLPGVGRYTAGAIASIAFGQA-TGV-VDGNVARVLCR----VRAIGADPSSTLVSQQLW-GLAQQLV---DPARPGDFNQAA---ME 281
OGG1_Homo_sapiens    233 YEEAHKALCILPGVGTKVADCICLMALDKP-QAVPVDVHMWHIAQRDYSWHPTTSQAKGPSPQTNKELGNFFRSLW--------QPY-------AG 312
MBD4_Homo_sapiens    268 CNKA-DAESEPVAQKSQLDRTVCISDAGACGETLSVTSEENSLVKK-----KERSLSSGSNFCSEQKTSGIINKFCSAKDSEHNEKYEDTFLESEE 357

NTHL1_Homo_sapiens   285 FGQQTCLPVH--------PRCHACLNQALCPAAQGL-------------------------------------------------------------312
MUTYH_Homo_sapiens   282 LGATVCTPQR--------PLCSQCPVESLCRARQRVEQEQLLASGSLSGSPDVEECAPNTGQCHLCLPPSEPWDQTLGVVNFPRKASRKPPREESS 369
OGG1_Homo_sapiens    313 WAQAVLFSAD--------LRQSRHAQEPPAKRRKGSKGPEG-------------------------------------------------------345
MBD4_Homo_sapiens    358 IGTKVEVVERKEHLHTDILKRGSEMDNNCSPTRKDFTGEKIFQEDTIPRT----QIERRKTSLYFSSKYNKEALSPPRRKAFKKWTPPRSPFNLVQ 449

NTHL1_Homo_sapiens     ------------------------------------------------------------------------------------------------
MUTYH_Homo_sapiens   370 ATCVLEQPGALGAQILLVQRPNSGLLAGLWEFPSVTWEPSEQLQRKALLQELQRWAGPLPATHLRHLGEVVHTFSHIKLTYQVYGLALEGQTPVTT 465
OGG1_Homo_sapiens      ------------------------------------------------------------------------------------------------
MBD4_Homo_sapiens    450 ETLFHDPWKLLIATIFLNRTSGKMAIPVLWKF--LEKYPSAEVARTADWRDVSELLKPLGLYDLRAKTIVKFS----------------------520

NTHL1_Homo_sapiens     ------------------------------------------------------------------------------------------------
MUTYH_Homo_sapiens   466 VPPGARWLTQEEFHTAAVSTAMKKVFRVYQGQQPGTCMGSKRSQVSSPCSRKKPRMGQQVLDNFFRSHISTDAHSLNSAAQ                546
OGG1_Homo_sapiens      ------------------------------------------------------------------------------------------------
MBD4_Homo_sapiens    521 ----------DEYLTKQWKYPIELHGIGKYGNDSYRIFCVNEWKQVHPEDHKLNKYHDWLWENHEKLSLS----------                580
```

MAFFT:

```
NTHL1_Homo_sapiens     1 MCSPQESGMTA-----------LSARMLTRSRS------LGPGAGPRGCREEPGPLRREAA--------------AEARKSHSPVKRPRKAQRLRVAYE 69
MUTYH_Homo_sapiens     1 -MTPLVSRLSR-----------LWAIMRKPRAA------VGSGH-----------RKQAAS--------------QEGRQKHA----KNNSQAKPSACD 52
OGG1_Homo_sapiens      1 --MPARALLPRRMGHRTLASTPALWASIPCPRSELRLDLVLPSGQSFRW--------REQSPAHWSGVLADQVWTLTQTEEQLHCTVYRGDKSQASRPTPD 91
MBD4_Homo_sapiens      1 MGTTGLESLS------------LGDRGAAPTVT----------SSERLVPDPPNDLRKEDVA--------------MELERV----------------46

NTHL1_Homo_sapiens    70 GSDSE-----KGEGAEPLKVPVWEPQDWQ---------QQL----VNIRAMR---------NKKDAP-----VDHLGTEHCYDSSAPPKVRRYQVLLSLML 138
MUTYH_Homo_sapiens    53 GMIAECPGAPAGLARQPEEV-VLQASVSS---------YHLFRDVAEVTAFRGSLLSWYDQEKRDLPWRRRAEDEMDLDR----------RAYAVWVSEVM 133
OGG1_Homo_sapiens     92 ELEAV---------RKYFQLDVTLAQLYHHWGSVDSHFGEVAQKFQGVRLLR------------------QDPI----------------ECLFSFIC 146
MBD4_Homo_sapiens     47 GEDEE-------------------------------QMMIKRSSECNPLL------------------QEPIASAQFGATAGTECRKSVPCGWERVV 94

NTHL1_Homo_sapiens   139 SSQTKDQVT------AGAMQRLRARGLTVDSILQTDDATLGKLIYP----------VGFW-----RSKVKYIKQTSAILQQHYGGDI--------------204
MUTYH_Homo_sapiens   134 LQQT--QVATVINYYTGWMQKWP----TLQDLASASLEEVNQLWAG----------LGYY------SRGRRLQEGARKVVEELGGHM--------------198
OGG1_Homo_sapiens    147 SSNN--NIARITGMVERLCQAFGPRLIQLDDVTYHGFPSLQALAGPEVEAHLRKLGLGY--------RARYVSASARAILEEQGG--------------221
MBD4_Homo_sapiens     95 KQRLFGKTA----------GRFDVYFISPQGLKFRSKSSLANYLHKNGETSLKPEDFDFTVLSKRGIKSRYKDCSMAALTSHLQNQSNNSNWNLRTRSKCK 185

NTHL1_Homo_sapiens   205 ------PASVAELVA-------------------------------LPGVGPK---------MAHLAMAVAWGTVSGIAVDTHVHRIANR-----248
MUTYH_Homo_sapiens   199 ------PRTAETLQQL-------------------------------LPGVGRY---------TAGAIASIAFGQATGV-VDGNVARVLCR-----242
OGG1_Homo_sapiens    222 ---------LAWLQQLRESS-------------------YEEAHKALCILPGVGTK--------VADCICLMALDKPQAVPVDVHMWHIAQRDYSWH 282
MBD4_Homo_sapiens    186 KDVFMPPSSSSELQESRGLSNFTSTHLLLKEDEGVDDVNFRKVRKPKGKVTILKGIPIKKTKKGCRKSCSGFVQSDSKRESVCNKADAESEPVAQKSQLDR 286

NTHL1_Homo_sapiens   249 LRWTKKATKSPEETRAALEEW-LPRELWH-----EINGLLVGFGQQTCLPVHPRCHACLNQALCPAAQ-------------------------------310
MUTYH_Homo_sapiens   243 VR---AIGADPSSTLVSQQLWGLAQQLVDPARPGDFNQAAMELGATVCTPQRPLCSQCPVESLCRARQ-------------------------------307
OGG1_Homo_sapiens    283 PTTSQAKGPSPQTNK---ELGNFFRSLWGP-----------------------YAGWAQAVLFSAD-------------------------------322
MBD4_Homo_sapiens    287 TVCISDAGACGETLSVTSEENSLVKK----------KERSLSSGSNFCSEQK---TSGIINKFCSAKDSEHNEKYEDTFLESEEIGTKVEVVERKEHLHTD 374

NTHL1_Homo_sapiens     ------------------------------------------------------------------------------------------------
MUTYH_Homo_sapiens   308 ------------------------------RVEQEQLLASGSLSGSPDVEECAPNTGQCHLCLPPSEPWDQTLGVVNFPRKASRKPPREESSATCVLEQP 377
OGG1_Homo_sapiens      ------------------------------------------------------------------------------------------------
MBD4_Homo_sapiens    375 ILKRGSEMDNNCSPTRKDFTGEKIFQEDTIPRTQIERRKTSLYFSSKYNKEALSPPRRKA-----------------FKKWTPPRSPFNLVQETLFHDPW 457

NTHL1_Homo_sapiens   311 ---------------------------------------------------------------------------GL-----------------------312
MUTYH_Homo_sapiens   378 GALGAQILLVQRPNSGLLAGLWEFPSVTWEPSEQLQRKALLQELQRWAGPLPATHLRHLGEVVHTFSHIKLTYQVYGLALEGQTPVTTVPPGARWLTQEEF 478
OGG1_Homo_sapiens      ------------------------------------------------------------------------------------------------
MBD4_Homo_sapiens    458 KLLIATIFLNRTSGKMAIPVLWKF--LEKYPSAEVARTADWRDVSELLKPLGLYDLRAKTIVKFSDEYL-------------------------524

NTHL1_Homo_sapiens     ------------------------------------------------------------------------------------------------
MUTYH_Homo_sapiens   479 HTAAVSTAMKKVFRVYQGQQPGTCMGSKRSQVSSPCSRKKPRMGQQVLDNFFRSHISTDAHSLNSAAQ-----------                546
OGG1_Homo_sapiens    323 ----------------------LRQSRHAQEPPAKRRKGSKGPEG-------------------------------                345
MBD4_Homo_sapiens    525 ----------------------TKQWKYPIELHGIGKYGNDSYRIFCVNEWKQVHPEDHKLNKYHDWLWENHEKLSLS                580
```

CLUSTAL:

```
NTHL1_Homo_sapiens     1 -------------MCSPQESGMTALSARMLTRSRSLGPGAGPRGCREEPGPLRRREAAAEARKSHSPVKRPRK-----AQRLRVA 67
MUTYH_Homo_sapiens     1 --------------MTPLVSRLSRLWAIMRKPRAAVGS-----GHRKQAASQEGRQKHAKNNSQAKPSACDGM-----IAECPGA 61
OGG1_Homo_sapiens      1 ----MPARALLPRRMGHRTLASTPALWASIPCPRSELRLD----LVLPSGQSFRWREQSPAHWSGVLADQVWTL-----TQTEEQL 73
MBD4_Homo_sapiens      1 MGTTGLESLSLGDRGAAPTVTSSERLVPDPPNDLRKEDVAMELERVGEDEEQMMIKRSSECNPLLQEPIASAQFGATAGTECRKSV 86

NTHL1_Homo_sapiens    68 YEGGSDS-EKGEGAEPLKVPVWEPQD-----WQQQLVNIR-AMRNKKDAPVDHLGTEHCYDSSAPPKVRRYQVLLSLMLSSQTKDQV 146
MUTYH_Homo_sapiens    62 PAGLAR-QPEEVVLQASVSSYHLFR-----DVAEVTAFRGSLLSWYDQEKRDLPWRRRAEDEMDLDRRAYAVWVSEVMLQQTQVAT 141
OGG1_Homo_sapiens     74 HCTVYRGDKSQASRPTPDELEAVRK-----YFQLDVTLAQLYHHWGSVDSHFQEVAQKFQGVRLLRQDPIECLFSFICSSNNNIAR 154
MBD4_Homo_sapiens     87 PCGWERVVKQRLFGKTAGRFDVYFISPQGLKFRSKSSLANYLHKNGETSLKPEDFDFTVLSKRGIKSRYKDCSMAALTSHLQNQSN 172

NTHL1_Homo_sapiens   147 TAGAMQRLRARG------------LTVDSILQTDDATLGKLIYPVGFWRSKVKYIKQTSAILQQHYGG---------DIPASVAEL 211
MUTYH_Homo_sapiens   142 VINYYTGWMQKW------------PTLQDLASASLEEVNQLWAGLGYY-SRGRRLQEGARKVVEELGG---------HMPRTAETL 205
OGG1_Homo_sapiens    155 ITGMVERLCQAFGPRLIQLDDVTYHGFPSLQALAGPEVEAHLRKLGLG-YRARYVSASARAILEEQGGLAWLQQLRESSYEEAHKA 239
MBD4_Homo_sapiens    173 NSNWNLRTRSKCK------KDVFMPPSSSSELQESRGLSNFTSTHLLLKEDEGVDDVNFRKVRKPKGKVTILKGIPIKKTKKGCRK 252

NTHL1_Homo_sapiens   212 VALP-GVGPKMAHLAMAVAWGTVSGIAVDTHVHRIANRLRWT---KKATKSPEETRAALEEWLPRELWHEIN--GLLVGFGQQTCL 291
MUTYH_Homo_sapiens   206 QQLLPGVGRYTAGAIASIAFGQATG-VVDGNVARVLCRVRAIGADPSSTLVSQQLWGLAQQLVDPARPGDFN--QAAMELGATVCT 288
OGG1_Homo_sapiens    240 LCILPGVGTKVADCICLMALDKPQAVPVDVHMWHIAQRDYSWHPTTSQAKGPSPQTNKELGNFFRSLWGPYAGWAQAVLFSADLRQ 325
MBD4_Homo_sapiens    253 SCSGFVQSDSKRESVCNKADAESEPVAQKSQLDRTVCISDAGACGETLSVTSEENSLVKKKERSLSSGSNFCSEQKTSGIINKFCS 338

NTHL1_Homo_sapiens   292 PVHPRCHACLNQALCPAAQGL------------------------------------------------------------------ 312
MUTYH_Homo_sapiens   289 PQRPLCSQCPVESLCRARQRVEQEQLLASGSLSGSPDVEECAPNTGQCHLCLPPSEPWDQTLGVVNFPRKASRKPPREESSATCVL 374
OGG1_Homo_sapiens    326 SRHAQEPPAKRRKGSKGPEG---------------------------------------------------------------- 345
MBD4_Homo_sapiens    339 AKDSEHNEKYEDTFLESEEIGTKVEVVERKEHLHTDILKRGSEMDNNCSPTRKDFTGEKIFQEDTIPRTQIERRKTSLYFSSKYNK 424

NTHL1_Homo_sapiens       ------------------------------------------------------------------------------------
MUTYH_Homo_sapiens   375 EQPGALGAQILLVQRPNSGLLAGLWEFPSVTWEPSEQLQRKALLQELQRWAGPLPATHLRHLGEVVHTFSHIKLTYQVYGLALEGQ 460
OGG1_Homo_sapiens        ------------------------------------------------------------------------------------
MBD4_Homo_sapiens    425 EALSPPRRKAFKKWTPPRSPFN-----LVQETLFHDPWKLLIATIFLNRTSGKMAIPVLWKFLEKYPSAEVARTADWRDVSELLKP 505

NTHL1_Homo_sapiens       ------------------------------------------------------------------------------------
MUTYH_Homo_sapiens   461 TPVTTVPPGARWLTQEEFHTAAVSTAMKKVFRVYQGQQPGTCMGSKRSQVSSPCSRKKPRMGQQVLDNFFRSHISTDAHSLNSAAQ 546
OGG1_Homo_sapiens        ------------------------------------------------------------------------------------
MBD4_Homo_sapiens    506 LGLYDLRAKTIVKFSDEYLTKQWKYPIELHGIGKYGNDSYRIFCVNEWKQVHPEDHKLNKYHDWLWENHEKLSLS----------- 580

NTHL1_Homo_sapiens
MUTYH_Homo_sapiens
OGG1_Homo_sapiens
MBD4_Homo_sapiens
```

33. 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 completely random.

    The HhH motif is well conserved in NTHL1, MUTYH and OGG1. However, ClustalW does not align the initial L and P correctly for NTHL1.

    The [4Fe-4S] cluster motif is fully conserved in NTHL1 and MUTYH, but not nicely aligned above due to the other two sequences that are lacking the motif. Actually, there is no [4Fe-4S] cluster in OGG1 or MBD4. Hence, there is no need to conserve, during evolution, the Cys residues that are complexing the iron-sulphur cluster in the other homologs.

    MBD4 is also aligned to both these motifs with all three programs, but to the wrong part of MBD4.

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

    The Clustal W program seemes to produce the worst alignment, as the HhH motif was not well aligned in NTHL1.

35. Finally, make MUSCLE and MAFFT alignments where you also include the bacterial and vertebrate Nth and MutY sequences that we worked with earlier. Do not

duplicate the human Nth and MutY. Format the alignment as earlier, but sort “by ID”. Include the alignments in your report, but crop the images so that only the region “core region” with the HhH motif is shown. Are any of the programs able to correctly align the HhH motif when all sequences are included? Which important lesson can we learn from this? Which program performed best?

MUSCLE:

```
Bacillus_anthracis_NP_844020          31 FELVIAVALSAQCTDALVNKVTKNLFQK------------YKTPEDYLSVSLEELQQDIRSIGLYRNKAKNIQKLCRMLL 98
Bos_taurus_MutY                      102 YAVWVAEVMLQQTQVATVINYYTRWMQK------------WPTLQDLASASLEEVNQLWAGLGYY-SRGRWLQEGARKVV 168
Bos_taurus_Nth                       123 YQVLLSLMLSSQTKDQVTAGAMQRLRAR------------GLTVDSILQTDDSTLGALIYPVGFWRSKVKYIKQTSAILQ 190
Escherichia_coli_NP_416150            30 FELLIAVLLSAQATDVSVNKATAKLYPV------------ANTPAAMLELGVEGVKTYIKTIGLYNSKAENIIKTCRILL 97
Gallus_gallus_MutY                   166 LAVWVSEIMLQQTQVATVIDYYNRWMQK------------WPTLQALAASLEEVNELWAGLGYY-SRGKRLQEAARKVV 232
Gallus_gallus_Nth                     99 YQVLLSLMLSSQTKDQVTSAAMLRLRQR------------GLTVDSILQMDDATLGQIIYPVGFWRNKVKYIKQTTAILK 166
Homo_sapiens_MutY                    125 YAVWVSEVMLQQTQVATVINYYTGWMQK------------WPTLQDLASASLEEVNQLWAGLGYY-SRGRRLQEGARKVV 191
Homo_sapiens_Nth                     130 YQVLLSLMLSSQTKDQVTAGAMQRLRAR------------GLTVDSILQTDDATLGKLIYPVGFWRSKVKYIKQTSAILQ 197
MBD4_Homo_sapiens                    156 KDCSMAALTSHLQNQSNNSNWNLRTRSK------------CKKDVFMPPSSSSELQESRGLSNF--------TSTHLLL 214
Mus_musculus_MutY                     99 YAVWVSEVMLQQTQVATVIDYYTRWMQK------------WPKLQDLASASLEEVNQLWSGLGYY-SRGRRLQEGARKVV 165
Mus_musculus_Nth                     118 YQVLLSLMLSSQTKDQVTAGAMQRLRAR------------GLTVESILQTDDDTLGRLIYPVGFWRNKVKYIKQTTAILQ 185
Mycobacterium_tuberculosis_NP_218191  41 LELAVATILSAQSTDKRVNLTTPALFAR------------YRTARDYAQADRTELESLIRPTGFYRNKAASLIGLGQALV 108
Neisseria_meningitidis_NP_273578      30 FELLIAVLLSAQATDVGVNKATAKLFPV------------ADTPQAMLDLGLDGVMEYTKTIGLYKTKSKHIMQTCRILL 97
OGG1_Homo_sapiens                    138 IECLFSFICSSNNNIARITGMVERLCQAFGPRLIQLDDVTYHGFPSLQALAGPEVEAHLRKLGLG-YRARYVSASARAIL 216
Streptococcus_pneumoniae_NP_358750    32 FELLVAVMLSAQTTDAAVNKATPGLFVA------------FPTPQAMSVATESEIASHISRLGLYRNKAKFLKKCAQQLL 99
Takifugu_rubripes_MutY               287 YAVWVSEIMLQQTQVATVIDYYNKWMKR------------WPTVQDLATATLEDVNQMWAGLGYY-SRGKRLHEGAQKVV 353
Takifugu_rubripes_Nth                143 FQVLVSLMLSSQTKDQVTSAAMQKLRAH------------GCTVENILATNDETLGQLIYPVGFWRNKVKYLKLTSAMLQ 210
Xenopus_(Silurana)_tropicalis_MutY   83 YAVWVSEVMLQQTQVATVIDYYNKWMKV------------WPTMEDLARSSLEEVNEMWSGLGYY-SRGRRLQEGAKKVV 149
Xenopus_(Silurana)_tropicalis_Nth    115 YQILLSLMLSSQTKDQVTSAAMCRLRQH------------GLTVSRILETDDGTLGKLIYPVGFWKNKVKYIKQTTEILQ 182

Bacillus_anthracis_NP_844020          99 DDYNG----------EVPKDRDELTK-LP------GVGRKTANVVSVAFGIP-AIAVDTHVERVSKR----LAICRWKD 156
Bos_taurus_MutY                      169 EELGG----------HMPRTAETLQQFLP------GVGRYTAGAIASIAFGQAAGV-VDGNVIRVLCR----VRAIGADS 227
Bos_taurus_Nth                       191 QRYDG----------DIPASVAELVA-LP------GVGPKMAHLAMAVAWGTVSGIAVDTHVHRIANR----LRWTKKAT 249
Escherichia_coli_NP_416150            98 EQHNG----------EVPEDRAALEA-LP------GVGRKTANVVLNTAFGWP-TIAVDTHIFRVCNR----TQFAPGKN 155
Gallus_gallus_MutY                   233 SELAG----------RMPRTAEDLQRLLP------GVGRYTAGAIASISFGQATGV-VDGNVIRVLCR----LRCIGADT 291
Gallus_gallus_Nth                    167 QKYGG----------DIPGTVEELVK-LP------GVGPKMAHLAMNIAWNSVSGIAVDTHVHRITNR----LKWVKKET 225
Homo_sapiens_MutY                    192 EELGG----------HMPRTAETLQQLLP------GVGRYTAGAIASIAFGQATGV-VDGNVARVLCR----VRAIGADP 250
Homo_sapiens_Nth                     198 QHYGG----------DIPASVAELVA-LP------GVGPKMAHLAMAVAWGTVSGIAVDTHVHRIANR----LRWTKKAT 256
MBD4_Homo_sapiens                    215 KEDEG-VDDVNFRKVRKPKGKVTILKGIPIKKTKKGCRKSCSGFVQSDSKRESVCNKADAESEPVAQKSQLDRTVCISDA 293
Mus_musculus_MutY                    166 EELGG----------HMPRTAETLQQLLP------GVGRYTAGAIASIAFDQVTGV-VDGNVLRVLCR----VRAIGADP 224
Mus_musculus_Nth                     186 QRYEG----------DIPASVAELVA-LP------GVGPKMAHLAMAVAWGTISGIAVDTHVHRIANR----LRWTKKMT 244
Mycobacterium_tuberculosis_NP_218191 109 ERFGG----------EVPATMDKLVT-LP------GVGRKTANVILGNAFGIP-GITVDTHFGRLVRR----WRWTTAED 166
Neisseria_meningitidis_NP_273578      98 EKYNG----------EVPEDREALES-LP------GVGRKTANVVLNTAFGHP-VMAVDTHIFRVSNR----TKIAPGKD 155
OGG1_Homo_sapiens                    217 EEQGGLAWLQQLRESSYEEAHKALCI-LP------GVGTKVADCICLMALDKPQAVPVDVHMWHIAQR---DYSWHPTTS 286
Streptococcus_pneumoniae_NP_358750   100 DDFDG----------QVPQTREELES-LA------GVGRKTANVVMSVGFGIP-AFAVDTHVERICKH----HDIVKKSA 157
Takifugu_rubripes_MutY               354 SQLQG----------EMPRTVDALLKQLP------GVGRYTAGAIGSIALGQVTGA-VDGNVIRVLCR----LRAIGADC 412
Takifugu_rubripes_Nth                211 KEFGG----------DIPDSVEGLVR-LP------GVGPKMAHLAMDIAWDQVSGIGVDTHVHRISNR----LGWLKKPT 269
Xenopus_(Silurana)_tropicalis_MutY   150 LELGG----------SMPRSADELQKLLP------GVGRYTAGAIASISYGQVTGV-VDGNVIRVLSR----LRCIGADS 208
Xenopus_(Silurana)_tropicalis_Nth    183 EKYGG----------DIPDNVTDLVK-LP------GVGPKMAHLVMDIAWNNVSGIGVDTHVHRISNR----LKWVRKET 241

Bacillus_anthracis_NP_844020         157 SVLEVEKT----LMKKIPMDEWSVTHHR--MIFFGRYHCKAQRPQCEECPLLEVCR---- 206
Bos_taurus_MutY                      228 SSTLVSQHLWSLAQQLVDPARPGDFNQA--AMELGAIVCTPKRPLCSHCPVQNLCRARQR 285
Bos_taurus_Nth                       250 KSPEETRRA---LEEWLPRELWSEINGL--LVGFGQQTCLPIRPRCQACLNRALC----- 299
Escherichia_coli_NP_416150           156 VEQVEEK-----LLKVVPAEFKVDCHHW--LILHGRYTCIARKPRCGSCIIEDLC----- 203
Gallus_gallus_MutY                   292 SSLAVIDCLWDMANTLVDRSRPGDFNQA--LMELGATVCTPKSPLCRECPVKEHCHAWRR 349
Gallus_gallus_Nth                    226 RYPEETRVA---LEDWLPRDLWREINWL--LVGFGQQTCLPVNPRCKECLNQDIC----- 275
Homo_sapiens_MutY                    251 SSTLVSQQLWGLAQQLVDPARPGDFNQA--AMELGATVCTPQRPLCSQCPVESLCRARQR 308
Homo_sapiens_Nth                     257 KSPEETRAA---LEEWLPRELWHEINGL--LVGFGQQTCLPVHPRCHACLNQALC----- 306
MBD4_Homo_sapiens                    294 GACGETLSVTSEENSLVKKKER--------SLSSGSNFCSEQK---TSGIINKFCSAKDS 342
Mus_musculus_MutY                    225 TSTLVSHHLWNLAQQLVDPARPGDFNQA--AMELGATVCTPQRPLCSHCPVQSLCRAYQR 282
Mus_musculus_Nth                     245 KTPEETRKN---LEEWLPRVLWSEVNGL--LVGFGQQICLPVHPRCQACLNKALC----- 294
Mycobacterium_tuberculosis_NP_218191 167 PVKVEQA-----VGELIERKEWTLLSHR--VIFHGRRVCHARRPACGVCVLAKDCPSFG- 218
Neisseria_meningitidis_NP_273578     156 VREVEDK-----LMRFIPKEFLMDAHHW--LILHGRYTCKALKPQCQTCIINDLC----- 203
OGG1_Homo_sapiens                    287 QAKGPSPQTNKELGNFF-RSLWGPYAGWAQAVLFSADLRQSRHAQEPPAKRRKGS----- 340
Streptococcus_pneumoniae_NP_358750   158 TPLEVEKR----VMDILPPEQWLAAHQA--MIYFGRAICHPKNPECDQYP--QLY----- 204
Takifugu_rubripes_MutY               413 TGPTVTEALWSLANTLVDPDRPGDFNQA--LMELGARVCTPKGPLCTRCPIQSHCHSYHK 470
Takifugu_rubripes_Nth                270 KTPEETRKS---LEEWLPRELWSEINWL--LVGFGQQVCLPVSPLCSVCLNQHDCPSAHK 324
Xenopus_(Silurana)_tropicalis_MutY   209 STLAVSDKLWNLANALVDPDRPGDFNQG--MMELGATVCTPKKPLCTACPLQGQCKAYLK 266
Xenopus_(Silurana)_tropicalis_Nth    242 KTPEETRVA---MEDWMPRELWSEINWL--LVGFGQQVCLPVSPRCSECLNKDIC----- 291
```

MAFFT:

```
Bacillus_anthracis_NP_844020          31 FELVIAVALSAQCTDALVNKVTKNLFQK--------------YKTPEDYLSVSLE-ELQQDIRSIGLYRNKA 87
Bos_taurus_MutY                      102 YAVWVAEVMLQQTQVATVINYYTRWMQK--------------WPTLQDLASASLE-EVNQLWAGLGYY-SRG 157
Bos_taurus_Nth                       123 YQVLLSLMLSSQTKDQVTAGAMQRLRAR--------------GLTVDSILQTDDS-TLGALIYPVGFWRSKV 179
Escherichia_coli_NP_416150            30 FELLIAVLLSAQATDVSVNKATAKLYPV--------------ANTPAAMLELGVE-GVKTYIKTIGLYNSKA 86
Gallus_gallus_MutY                   166 LAVWVSEIMLQQTQVATVIDYYNRWMQK--------------WPTLQALAAASLE-EVNELWAGLGYY-SRG 221
Gallus_gallus_Nth                     99 YQVLLSLMLSSQTKDQVTSAAMLRLRQR--------------GLTVDSILQMDDA-TLGQIIYPVGFWRNKV 155
Homo_sapiens_MutY                    125 YAVWVSEVMLQQTQVATVINYYTGWMQK--------------WPTLQDLASASLE-EVNQLWAGLGYY-SRG 180
Homo_sapiens_Nth                     130 YQVLLSLMLSSQTKDQVTAGAMQRLRAR--------------GLTVDSILQTDDA-TLGKLIYPVGFWRSKV 186
MBD4_Homo_sapiens                    457 WKLLIATIFLNRTSGKMAIPVLWKFLEK--------------YPSAEVARTADWR-DVSELLKPLGLYDLRA 513
Mus_musculus_MutY                     99 YAVWVSEVMLQQTQVATVIDYYTRWMQK--------------WPKLQDLASASLE-EVNQLWSGLGYY-SRG 154
Mus_musculus_Nth                     118 YQVLLSLMLSSQTKDQVTAGAMQRLRAR--------------GLTVESILQTDDD-TLGRLIYPVGFWRNKV 174
Mycobacterium_tuberculosis_NP_218191  41 LELAVATILSAQSTDKRVNLTTPALFAR--------------YRTARDYAQADRT-ELESLIRPTGFYRNKA 97
Neisseria_meningitidis_NP_273578      30 FELLIAVLLSAQATDVGVNKATAKLFPV--------------ADTPQAMLDLGLD-GVMEYTKTIGLYKTKS 86
OGG1_Homo_sapiens                    138 IECLFSFICSSNNNIARITGMVERLCQAFGPRLIQLDDVTYHGFPSLQALAGPEVEAHLRKL--GLGY---RA 205
Streptococcus_pneumoniae_NP_358750    32 FELLVAVMLSAQTTDAAVNKATPGLFVA--------------FPTPQ-AMSVATESEIASHISRLGLYRNKA 88
Takifugu_rubripes_MutY               287 YAVWVSEIMLQQTQVATVIDYYNKWMKR--------------WPTVQDLATATLE-DVNQMWAGLGYY-SRG 342
Takifugu_rubripes_Nth                143 FQVLVSLMLSSQTKDQVTSAAMQKLRAH--------------GCTVENILATNDE-TLGQLIYPVGFWRNKV 199
Xenopus_(Silurana)_tropicalis_MutY    83 YAVWVSEVMLQQTQVATVIDYYNKWMKV--------------WPTMEDLARSSLE-EVNEMWSGLGYY-SRG 138
Xenopus_(Silurana)_tropicalis_Nth    115 YQILLSLMLSSQTKDQVTSAAMCRLRQH--------------GLTVSRILETDDG-TLGKLIYPVGFWKNKV 171

Bacillus_anthracis_NP_844020          88 KNIQKLCRMLLDDYNG----------EVPKDRDELTK-LPGVGRKTANVVVSVAFG-IPAIAVDTHVERVSKR 148
Bos_taurus_MutY                      158 RWLQEGARKVVEELGG----------HMPRTAETLQQFLPGVGRYTAGAIASIAFGQAAGV-VDGNVIRVLCR 219
Bos_taurus_Nth                       180 KYIKQTSAILQQRYDG----------DIPASVAELVA-LPGVGPKMAHLAMAVAWGTVSGIAVDTHVHRIANR 241
Escherichia_coli_NP_416150            87 ENIIKTCRILLEQHNG----------EVPEDRAALEA-LPGVGRKTANVVLNTAFG-WPTIAVDTHIFRVCNR 147
Gallus_gallus_MutY                   222 KRLQEAARKVVSELAG----------RMPRTAEDLQRLLPGVGRYTAGAIASISFGQATGV-VDGNVIRVLCR 283
Gallus_gallus_Nth                    156 KYIKQTTAILKQKYGG----------DIPGTVEELVK-LPGVGPKMAHLAMNIAWNSVSGIAVDTHVHRITNR 217
Homo_sapiens_MutY                    181 RRLQEGARKVVEELGG----------HMPRTAETLQQLLPGVGRYTAGAIASIAFGQATGV-VDGNVARVLCR 242
Homo_sapiens_Nth                     187 KYIKQTSAILQQHYGG----------DIPASVAELVA-LPGVGPKMAHLAMAVAWGTVSGIAVDTHVHRIANR 248
MBD4_Homo_sapiens                    514 KTIVKFSDEYLTKQW----------KYPIE-------LHGIGKY--------------GNDSYRIFCVNE 552
Mus_musculus_MutY                    155 RRLQEGARKVVEELGG----------HMPRTAETLQQLLPGVGRYTAGAIASIAFDQVTGV-VDGNVLRVLCR 216
Mus_musculus_Nth                     175 KYIKQTTAILQQRYEG----------DIPASVAELVA-LPGVGPKMAHLAMAVAWGTISGIAVDTHVHRIANR 236
Mycobacterium_tuberculosis_NP_218191  98 ASLIGLGQALVERFGG----------EVPATMDKLVT-LPGVGRKTANVILGNAFG-IPGITVDTHFGRLVRR 158
Neisseria_meningitidis_NP_273578      87 KHIMQTCRILLEKYNG----------EVPEDREALES-LPGVGRKTANVVLNTAFG-HPVMAVDTHIFRVSNR 147
OGG1_Homo_sapiens                    206 RYVSASARAILEEQGGLAWLQQLRESSYEEAHKALCI-LPGVGTKVADCICLMALDKPQAVPVDVHMWHIAQR 277
Streptococcus_pneumoniae_NP_358750    89 KFLKKCAQQLLDDFDG----------QVPQTREELES-LAGVGRKTANVVMSVGFG-IPAFAVDTHVERICKH 149
Takifugu_rubripes_MutY               343 KRLHEGAQKVVSQLQG----------EMPRTVDALLKQLPGVGRYTAGAIGSIALGQVTGA-VDGNVIRVLCR 404
Takifugu_rubripes_Nth                200 KYLKLTSAMLQKEFGG----------DIPDSVEGLVR-LPGVGPKMAHLAMDIAWDQVSGIGVDTHVHRISNR 261
Xenopus_(Silurana)_tropicalis_MutY   139 RRLQEGAKKVVLELGG----------SMPRSADELQKLLPGVGRYTAGAIASISYGQVTGV-VDGNVIRVLSR 200
Xenopus_(Silurana)_tropicalis_Nth    172 KYIKQTTEILQEKYGG----------DIPDNVTDLVK-LPGVGPKMAHLVMDIAWNNVSGIGVDTHVHRISNR 233

Bacillus_anthracis_NP_844020         149 LAICRWK-DSVLEVEKTL----MKKIPMDEWSVTHHRMIFFGRYHCKAQRPQCEECPLLEVCREGKKRMK    213
Bos_taurus_MutY                      220 VRAIGAD-SSSTLVSQHLWSLAQQLVDPARPGDFNQAAMELGAIVCTPKRPLCSHCPVQNLCRARQRVER    288
Bos_taurus_Nth                       242 LRWTKKATKSPEETRRAL----EEWLPRELWSEINGLLVGFGQQTCLPIRPRCQACLNRALCPAARGL--    305
Escherichia_coli_NP_416150           148 TQFAPGK-N-VEQVEEKL----LKVVPAEFKVDCHHWLILHGRYTCIARKPRCGSC-IIEDLCEYKEKVD    210
Gallus_gallus_MutY                   284 LRCIGAD-TSSLAVIDCLWDMANTLVDRSRPGDFNQALMELGATVCTPKSPLCRECPVKEHCHAWRRVEK    352
Gallus_gallus_Nth                    218 LKWVKKETRYPEETRVAL----EDWLPRDLWREINWLLVGFGQQTCLPVNPRCKECLNQDICPAAKRF--    281
Homo_sapiens_MutY                    243 VRAIGAD-PSSTLVSQQLWGLAQQLVDPARPGDFNQAAMELGATVCTPQRPLCSQCPVESLCRARQRVEQ    311
Homo_sapiens_Nth                     249 LRWTKKATKSPEETRAAL----EEWLPRELWHEINGLLVGFGQQTCLPVHPRCHACLNQALCPAAQGL--    312
MBD4_Homo_sapiens                    553 WKQVHPEDHKLNKYHDWLWENHEKLSLS--------------------------------------------    580
Mus_musculus_MutY                    217 VRAIGAD-PTSTLVSHHLWNLAQQLVDPARPGDFNQAAMELGATVCTPQRPLCSHCPVQSLCRAYQRVQR    285
Mus_musculus_Nth                     237 LRWTKKMTKTPEETRKNL----EEWLPRVLWSEVNGLLVGFGQQICLPVHPRCQACLNKALCPAAQDL--    300
Mycobacterium_tuberculosis_NP_218191 159 WRWTTAE-D-PVKVEQAV----GELIERKEWTLLSHRVIFHGRRVCHARRPACGVCVLAKDCPSFGLGPT    222
Neisseria_meningitidis_NP_273578     148 TKIAPGK-D-VREVEDKL----MRFIPKEFLMDAHHWLILHGRYTCKALKPQCQTC-IINDLCEYPAKA-    209
OGG1_Homo_sapiens                    278 ------DYSWHPTTSQAKGPSPQT----------NKELGNFFRSLWGPYAGWAQAVLFSADLRQSRHAQE    331
Streptococcus_pneumoniae_NP_358750   150 HDIVKKS-ATPLEVEKRV----MDILPPEQWLAAHQAMIYFGRAICHPKNPECDQYPQLYDFSNL-----    209
Takifugu_rubripes_MutY               405 LRAIGAD-CTGPTVTEALWSLANTLVDPDRPGDFNQALMELGARVCTPKGPLCTRCPIQSHCHSYHKAHF    473
Takifugu_rubripes_Nth                262 LGWLKKPTKTPEETRKSL----EEWLPRELWSEINWLLVGFGQQVCLPVSPLCSVCLNQHDCPSAHKNSP    327
Xenopus_(Silurana)_tropicalis_MutY   201 LRCIGAD-SSTLAVSDKLWNLANALVDPDRPGDFNQGMMELGATVCTPKKPLCTACPLQGQCKAYLKVIA    269
Xenopus_(Silurana)_tropicalis_Nth    234 LKWVRKETKTPEETRVAM----EDWMPRELWSEINWLLVGFGQQVCLPVSPRCSECLNKDICPGAKKKKP    299
```

MAFFT was able to correctly align the HhH motif of MBD4 with all the others when all sequences are included. MUSCLE did not. MAFFT performed best, but this is no general rule. MUSCLE and T-Coffee are also excellent MSA programs.

Important: **Very often, you get a better alignment of two, or a few sequences, if you align these sequences together with many homologs!**

36. Near the start of the exercise you found that the Nth homolog from *Pantholops hodgsonii*, the Tibetan antelope (identifier XP_005981298) was 55% identical to *E. coli* Nth. No other mammals had Nth-like homologs that were more than roughly 33% identical to *E. coli* Nth. Why is mammalian Nth rather unlike *E. coli* Nth, while Tibetan antelope Nth is quite similar? Do you have a suggestions?
37. Run a blastp search in the full nr database with default settings with Tibetan antelope sequence XP_005981298. What are the top hits? Do you now have any suggestions why mammalian Nth is rather unlike *E. coli* Nth, while Tibetan antelope Nth is quite similar?

The top hit is XP_005981298, itself, then follows Nth from *Phenylobacterium zucineum* (78% identical) and *Caulobacter segnis* (74%) among other sequences. Actually, *all* the 100 top hits are from bacteria, except the Tibetan antelope. If you Google *Phenylobacterium zucineum*, you find that it is a recently identified bacterial species that lives intracellularly in the human leukemia cell line K562. *Caulobacter segnis* is a bacteria that hardly has been studied at all.

There are three "possible" explanations here:

- **Tibetan antelope Nth is evolving and becoming more similar to bacterial Nth, through convergent evolution. This can safely be ruled out! This is *not possible!***
- **Tibetan antelope recently obtained XP_005981298 by a horisontal gene transfer event. The gene has jumped from a bacteria, into the genome of the antelope. This is unlikely, but perhaps not impossible!**
- **The DNA from the Tibetan antelope that was used for sequencing was contaminated by DNA from an unknown, possibly intracellular, bacteria. XP_005981298 is not encoded by the antelope genome at all, but by a bacterial contamination, and this was not spotted during sequencing or sequence processing. Personally, I think this is the most likely explanantion! XP_005981298 is wrongly annotated as a mammalian, antelope protein. It is actually bacterial…**

38. If you have more time, experiment and modify the script for example to
    a. Use H_sapiens, M_musculus, and so on in the headers
    b. Automatically generate 12verts_final.fasta from 12verts.fasta
    c. Or download a few hundred Nth vertebrate homologs from the BLAST results and test the script on this bigger data set. If necessary, modify the script to be more robust

APPENDIX 1:

Bacterial Nth homologs, original sequences

>gi|16129591|ref|NP_416150.1| DNA glycosylase and apyrimidinic (AP) lyase (endonuclease III) [Escherichia coli str. K-12 substr. MG1655]
MNKAKRLEILTRLRENNPHPTTELNFSSPFELLIAVLLSAQATDVSVNKATAKLYPVANTPAAMLELGVE
GVKTYIKTIGLYNSKAENIIKTCRILLEQHNGEVPEDRAALEALPGVGRKTANVVLNTAFGWPTIAVDTH
IFRVCNRTQFAPGKNVEQVEEKLLKVVPAEFKVDCHHWLILHGRYTCIARKPRCGSCIIEDLCEYKEKVD
I
>gi|57117142|ref|NP_218191.2| Probable endonuclease III Nth (DNA-(apurinic or apyrimidinic site)lyase) (AP lyase) (AP endonuclease class I) (endodeoxyribonuclease (apurinic or apyrimidinic)) (deoxyribonuclease (apurinic or apyrimidinic)) [Mycobacterium tuberculosis H37Rv]
MPGRWSAETRLALVRRARRMNRALAQAFPHVYCELDFTTPLELAVATILSAQSTDKRVNLTTPALFARYR
TARDYAQADRTELESLIRPTGFYRNKAASLIGLGQALVERFGGEVPATMDKLVTLPGVGRKTANVILGNA
FGIPGITVDTHFGRLVRRWRWTTAEDPVKVEQAVGELIERKEWTLLSHRVIFHGRRVCHARRPACGVCVL
AKDCPSFGLGPTEPLLAAPLVQGPETDHLLALAGL
>gi|30261643|ref|NP_844020.1| endonuclease III [Bacillus anthracis str. Ames]
MLNKTQIRYCLDTMADMYPEAHCELIHDNPFELVIAVALSAQCTDALVNKVTKNLFQKYKTPEDYLSVSL
EELQQDIRSIGLYRNKAKNIQKLCRMLLDDYNGEVPKDRDELTKLPGVGRKTANVVVSVAFGIPAIAVDT
HVERVSKRLAICRWKDSVLEVEKTLMKKIPMDEWSVTHHRMIFFGRYHCKAQRPQCEECPLLEVCREGKK
RMKGK
>gi|15676439|ref|NP_273578.1| endonuclease III [Neisseria meningitidis MC58]
MNRHIRQEIFERFRAANPHPTTELNFNSPFELLIAVLLSAQATDVGVNKATAKLFPVADTPQAMLDLGLD
GVMEYTKTIGLYKTKSKHIMQTCRILLEKYNGEVPEDREALESLPGVGRKTANVVLNTAFGHPVMAVDTH
IFRVSNRTKIAPGKDVREVEDKLMRFIPKEFLMDAHHWLILHGRYTCKALKPQCQTCIINDLCEYPAKA
>gi|15903200|ref|NP_358750.1| endonuclease III [Streptococcus pneumoniae R6]
MVLSKKRARKVLEEIIALFPDAKPSLDFTNHFELLVAVMLSAQTTDAAVNKATPGLFVAFPTPQAMSVAT
ESEIASHISRLGLYRNKAKFLKKCAQQLLDDFDGQVPQTREELESLAGVGRKTANVVMSVGFGIPAFAVD
THVERICKHHDIVKKSATPLEVEKRVMDILPPEQWLAAHQAMIYFGRAICHPKNPECDQYPQLYDFSNL

APPENDIX 2:

Bacterial Nth homologs, modified headers

```
>Escherichia_coli_NP_416150
MNKAKRLEILTRLRENNPHPTTELNFSSPFELLIAVLLSAQATDVSVNKATAKLYPVANTPAAMLELGVE
GVKTYIKTIGLYNSKAENIIKTCRILLEQHNGEVPEDRAALEALPGVGRKTANVVLNTAFGWPTIAVDTH
IFRVCNRTQFAPGKNVEQVEEKLLKVVPAEFKVDCHHWLILHGRYTCIARKPRCGSCIIEDLCEYKEKVD
I
>Mycobacterium_tuberculosis_NP_218191
MPGRWSAETRLALVRRARRMNRALAQAFPHVYCELDFTTPLELAVATILSAQSTDKRVNLTTPALFARYR
TARDYAQADRTELESLIRPTGFYRNKAASLIGLGQALVERFGGEVPATMDKLVTLPGVGRKTANVILGNA
FGIPGITVDTHFGRLVRRWRWTTAEDPVKVEQAVGELIERKEWTLLSHRVIFHGRRVCHARRPACGVCVL
AKDCPSFGLGPTEPLLAAPLVQGPETDHLLALAGL
>Bacillus_anthracis_NP_844020
MLNKTQIRYCLDTMADMYPEAHCELIHDNPFELVIAVALSAQCTDALVNKVTKNLFQKYKTPEDYLSVSL
EELQQDIRSIGLYRNKAKNIQKLCRMLLDDYNGEVPKDRDELTKLPGVGRKTANVVVSVAFGIPAIAVDT
HVERVSKRLAICRWKDSVLEVEKTLMKKIPMDEWSVTHHRMIFFGRYHCKAQRPQCEECPLLEVCREGKK
RMKGK
>Neisseria_meningitidis_NP_273578
MNRHIRQEIFERFRAANPHPTTELNFNSPFELLIAVLLSAQATDVGVNKATAKLFPVADTPQAMLDLGLD
GVMEYTKTIGLYKTKSKHIMQTCRILLEKYNGEVPEDREALESLPGVGRKTANVVLNTAFGHPVMAVDTH
IFRVSNRTKIAPGKDVREVEDKLMRFIPKEFLMDAHHWLILHGRYTCKALKPQCQTCIINDLCEYPAKA
>Streptococcus_pneumoniae_NP_358750
MVLSKKRARKVLEEIIALFPDAKPSLDFTNHFELLVAVMLSAQTTDAAVNKATPGLFVAFPTPQAMSVAT
ESEIASHISRLGLYRNKAKFLKKCAQQLLDDFDGQVPQTREELESLAGVGRKTANVVMSVGFGIPAFAVD
THVERICKHHDIVKKSATPLEVEKRVMDILPPEQWLAAHQAMIYFGRAICHPKNPECDQYPQLYDFSNL
```

APPENDIX 3:

12 vertebrate homologs, original sequences

```
>gi|4505471|ref|NP_002519.1| endonuclease III-like protein 1 [Homo sapiens]
MCSPQESGMTALSARMLTRSRSLGPGAGPRGCREEPGPLRRREAAAEARKSHSPVKRPRKAQRLRVAYEGSDSEKGEGAE
PLKVPVWEPQDWQQQLVNIRAMRNKKDAPVDHLGTEHCYDSSAPPKVRRYQVLLSLMLSSQTKDQVTAGAMQRLRARGLT
VDSILQTDDATLGKLIYPVGFWRSKVKYIKQTSAILQQHYGGDIPASVAELVALPGVGPKMAHLAMAVAWGTVSGIAVDT
HVHRIANRLRWTKKATKSPEETRAALEEWLPRELWHEINGLLVGFGQQTCLPVHPRCHACLNQALCPAAQGL
>gi|6912520|ref|NP_036354.1| A/G-specific adenine DNA glycosylase isoform 1 [Homo sapiens]
MTPLVSRLSRLWAIMRKPRAAVGSGHRKQAASQEGRQKHAKNNSQAKPSACDGMIAECPGAPAGLARQPEEVVLQASVSS
YHLFRDVAEVTAFRGSLLSWYDQEKRDLPWRRRAEDEMDLDRRAYAVWVSEVMLQQTQVATVINYYTGWMQKWPTLQDLA
SASLEEVNQLWAGLGYYSRGRRLQEGARKVVEELGGHMPRTAETLQQLLPGVGRYTAGAIASIAFGQATGVVDGNVARVL
CRVRAIGADPSSTLVSQQLWGLAQQLVDPARPGDFNQAAMELGATVCTPQRPLCSQCPVESLCRARQRVEQEQLLASGSL
SGSPDVEECAPNTGQCHLCLPPSEPWDQTLGVVNFPRKASRKPPREESSATCVLEQPGALGAQILLVQRPNSGLLAGLWE
FPSVTWEPSEQLQRKALLQELQRWAGPLPATHLRHLGEVVHTFSHIKLTYQVYGLALEGQTPVTTVPPGARWLTQEEFHT
AAVSTAMKKVFRVYQGQQPGTCMGSKRSQVSSPCSRKKPRMGQQVLDNFFRSHISTDAHSLNSAAQ
>gi|227908769|ref|NP_032769.2| endonuclease III-like protein 1 [Mus musculus]
MNSGVRMVTRSRSRATRIASEGCREELAPREAAAEGRKSHRPVRHPRRTQKTHVAYEAANGEEGEDAEPLKVPVWEPQNW
QQQLANIRIMRSKKDAPVDQLGAEHCYDASASPKVRRYQVLLSLMLSSQTKDQVTAGAMQRLRARGLTVESILQTDDDTL
GRLIYPVGFWRNKVKYIKQTTAILQQRYEGDIPASVAELVALPGVGPKMAHLAMAVAWGTISGIAVDTHVHRIANRLRWT
KKMTKTPEETRKNLEEWLPRVLWSEVNGLLVGFGQQICLPVHPRCQACLNKALCPAAQDL
>gi|227330621|ref|NP_573513.2| A/G-specific adenine DNA glycosylase [Mus musculus]
MKKLQASVRSHKKQPANHKRRRTRALSSSQAKPSSLDGLAKQKREELLQASVSPYHLFSDVADVTAFRSNLLSWYDQEKR
DLPWRNLAKEEANSDRRAYAVWVSEVMLQQTQVATVIDYYTRWMQKWPKLQDLASASLEEVNQLWSGLGYYSRGRRLQEG
ARKVVEELGGHMPRTAETLQQLLPGVGRYTAGAIASIAFDQVTGVVDGNVLRVLCRVRAIGADPTSTLVSHHLWNLAQQL
VDPARPGDFNQAAMELGATVCTPQRPLCSHCPVQSLCRAYQRVQRGQLSALPGRPDIEECALNTRQCQLCLTSSSPWDPS
MGVANFPRKASRRPPREEYSATCVVEQPGAIGGPLVLLVQRPDSGLLAGLWEFPSVTLEPSEQHQHKALLQELQRWCGPL
PAIRLQHLGEVIHIFSHIKLTYQVYSLALDQAPASTAPPGARWLTWEEFCNAAVSTAMKKVFRMYEDHRQGTRKGSKRSQ
VCPPSSRKKPSLGQQVLDTFFQRHIPTDKPNSTTQ
>gi|114051958|ref|NP_001039862.1| endonuclease III-like protein 1 [Bos taurus]
MNAAGVRMVVTRARSRGTGASLRRRGEKAAPLRSGEAAAEERKSYSPVKRRRKAQRLSVAYEASEGEGGEGAEHLQAPSW
QPQDWRQQLDNIRTMRSGKDAPVDQLGAEHCFDPSASPKVRRYQVLLSLMLSSQTKDQVTAGAMQRLRARGLTVDSILQT
DDSTLGALIYPVGFWRSKVKYIKQTSAILQQRYDGDIPASVAELVALPGVGPKMAHLAMAVAWGTVSGIAVDTHVHRIAN
RLRWTKKATKSPEETRRALEEWLPRELWSEINGLLVGFGQQTCLPIRPRCQACLNRALCPAARGL
>gi|281485563|ref|NP_001039600.2| A/G-specific adenine DNA glycosylase [Bos taurus]
MKKSRAAVGNRSGRRKQASSQEGKEKCAFGSSQAKPSAPSAGPARQQKALLQASVSPYHLFRDVAEVTALQESLLDWYDR
KKRDLPWRRLVEDEVDLDRRAYAVWVAEVMLQQTQVATVINYYTRWMQKWPTLQDLASASLEEVNQLWAGLGYYSRGRWL
QEGARKVVEELGGHMPRTAETLQQFLPGVGRYTAGAIASIAFGQAAGVVDGNVIRVLCRVRAIGADSSSTLVSQHLWSLA
QQLVDPARPGDFNQAAMELGAIVCTPKRPLCSHCPVQNLCRARQRVEREQLSASQSLPGNCDVEECAPNTGQCPLCAPPT
EPWDQTLGVTNFPRKASRKPPREECSAICVLEQPKALGGAHILLVQRPNSGLLAGLWEFPSVSVNAEASGQHQRAALLQE
LQSWVGPLPDTRLQHLGQVVHTFSHIKMTYQVYSLALEEHTPVTIVPPGARWLTREDFHTAAVSTAMKKVFRMYEGQQPG
TCKGSKRSQVATLSKRKKPSPGQQVLESFFWPHVPTDAPSLNTAAQ
>gi|118601744|ref|NP_001073043.1| endonuclease III-like protein 1 [Gallus gallus]
MCAAAPRGGGRAARRLGAATAGSRVPSAAPRYSRRTRRVPIAYEAEPKPESPGPKWEPENWQQQLERIREMRRHRDAPVD
EMGVDKCYDTSAPPQVMRYQVLLSLMLSSQTKDQVTSAAMLRLRQRGLTVDSILQMDDATLGQIIYPVGFWRNKVKYIKQ
TTAILKQKYGGDIPGTVEELVKLPGVGPKMAHLAMNIAWNSVSGIAVDTHVHRITNRLKWVKKETRYPEETRVALEDWLP
RDLWREINWLLVGFGQQTCLPVNPRCKECLNQDICPAAKRF
>gi|513197809|ref|XP_422433.3| PREDICTED: A/G-specific adenine DNA glycosylase isoform X5
[Gallus gallus]
MGGAAVRARRSVKVRAGGEHVGPGLGSPAIALRTHRRCCDPTPVPVSRQGLPLDHMHCISSVTPSRSMPYVAACSPGMTK
AGGTFPGGRWLQLSWMLTGGRMQLLGLLVASERQVWARERSCGDEDGEEGCWVGFCCSSWNQQHGDRGACCEKWHWHLCT
DPSVLLAVWVSEIMLQQTQVATVIDYYNRWMQKWPTLQALAAASLEEVNELWAGLGYYSRGKRLQEAARKVVSELAGRMP
RTAEDLQRLLPGVGRYTAGAIASISFGQATGVVDGNVIRVLCRLRCIGADTSSLAVIDCLWDMANTLVDRSRPGDFNQAL
MELGATVCTPKSPLCRECPVKEHCHAWRRVEKELASASQKLFGKTTLVPDVEDCGPGGCPLCLPAAEPWDSSLGVTNFPR
KAAKKQPRVEWTATCVLERRGRLGAPEYLIVQRPSSGLLAGLWEFPSLPLAPGLQEEQQKEVLADHLRAWTRQPVQTQSL
CFIGEVVHIFSHIHQTYVVYSLCLDGDVALDAASSPSRWVTEEEFRASAVSTAMKKVLKARETQRGVQSGRAKGSKRKRE
SKLGAAGSTPTGMQLSLRAFLRAQPPP
>gi|113205550|ref|NP_001037884.1| nth endonuclease III-like 1 [Xenopus (Silurana) tropicalis]
MSGSLRPLGRRGRRGVLKAVGGKDQQDGTSKGQVIDDSEDEKPSSPKERSKRRVSVEYEQAASETVAKRPKWQPKNWAQH
LENIRQMRSRRDAPVDQMGAEKCYDQNAAPEVMRYQILLSLMLSSQTKDQVTSAAMCRLRQHGLTVSRILETDDGTLGKL
IYPVGFWKNKVKYIKQTTEILQEKYGGDIPDNVTDLVKLPGVGPKMAHLVMDIAWNNVSGIGVDTHVHRISNRLKWVRKE
TKTPEETRVAMEDWMPRELWSEINWLLVGFGQQVCLPVSPRCSECLNKDICPGAKKKKPR
>gi|118403607|ref|NP_001072831.1| mutY homolog [Xenopus (Silurana) tropicalis]
MPPPRTKTSLGRSAAASGKRKSPKQAFPKREEHVLQSSIYHSFTSQETEIIRDKLLAWYDKSKRDLPWRTMACTEPDLDR
KAYAVWVSEVMLQQTQVATVIDYYNKWMKVWPTMEDLARSSLEEVNEMWSGLGYYSRGRRLQEGAKKVVLELGGSMPRSA
DELQKLLPGVGRYTAGAIASISYGQVTGVVDGNVIRVLSRLRCIGADSSTLAVSDKLWNLANALVDPDRPGDFNQGMMEL
```

GATVCTPKKPLCTACPLQGQCKAYLKVIAEKESAVKTLIKKQASPIAKDVGDIEDCDLGPGLCALCVPTSDPWDSSLGVA
NFPRKSAKKPSRMEQTAICVWEKCGDHGELEYLIVQRPSSGLLAGLWEFPSILLDEKFTEQNRQHSLLGLLQDLSGHAVP
LQKLQYKGEVVHIFSHIHQTYVVYFLSLNTTENCSVKTEETERPLTRWVTKKEFLNSAVPTAMKKIMKLCESHGSSCTAV
NTSKKRKGDLAKVQLPSGRIKTEKGKQQSIQSFFKLATEK

>gi|410917257|ref|XP_003972103.1| PREDICTED: endonuclease III-like protein 1-like [Takifugu rubripes]

MTSHYFAQSRSVVTRRGAQNAAHKPATSLKSKLTIQPEKDDLVSSSAGVKLEEEEAKISGNALKPETDAPTLSSHSRRRR
QLKVEYDKDGSMPQLKTEPWEPPRWKTQLENIRAMRSGRDAPVDNMGADKCHDADAPAHVKRFQVLVSLMLSSQTKDQVT
SAAMQKLRAHGCTVENILATNDETLGQLIYPVGFWRNKVKYLKLTSAMLQKEFGGDIPDSVEGLVRLPGVGPKMAHLAMD
IAWDQVSGIGVDTHVHRISNRLGWLKKPTKTPEETRKSLEEWLPRELWSEINWLLVGFGQQVCLPVSPLCSVCLNQHDCP
SAHKNSPVRRPKFLERSPWIKSPRLFITPGANTFLIR

>gi|410921366|ref|XP_003974154.1| PREDICTED: A/G-specific adenine DNA glycosylase-like [Takifugu rubripes]

MSTQGEPVQVGKVLSFLREWDRGDRSARGRMLSSFLGRSAGRTQGELEYLGEFVGHGGVVTLLEVLTQPQSNEETKAEAL
CLLLAISDAGRKYKELICQSCGAMAAAECLTHSGTGETQESAWMLLESLSHGNPKYEGEIYKGLIGHLTCTSAKAQQFVL
HTLHTLQSKMEIAHHSIVEPLLGVLTSLHPDVQSEVARLIFELRRYDVRPMLLRALCGLGLNVARAPTYPEEESSHASSP
SAYHFFHDAADVALLRSRLLAWYDQEKRELPWRTLALTEPDVNIRTYAVWVSEIMLQQTQVATVIDYYNKWMKRWPTVQD
LATATLEDVNQMWAGLGYYSRGKRLHEGAQKVVSQLQGEMPRTVDALLKQLPGVGRYTAGAIGSIALGQVTGAVDGNVIR
VLCRLRAIGADCTGPTVTEALWSLANTLVDPDRPGDFNQALMELGARVCTPKGPLCTRCPIQSHCHSYHKAHFKQEKNSM
KLMGKLDRKSSALPDIEDCLSSGTCTLCLSEPWDDELGVQNFPRKPAKKPPRAERCLTCVVIRQGEGGEHEFLLTQRPSK
GLLAGLWEFPCINHEEKNAVEEKKVLCAEINRILGTSLTHGLLQYVGEVVHIFSHIHQTYVVHTLRLKDAVSQSENMQWL
TPSALQEAAVSTGVKKIMKLCNSALGQQGAPDGEEKRPKKDRKGQITKRPRLSGANSRSRQLSLSSFFQTVKQDC

APPENDIX 4:

12 vertebrate homologs, modified headers

```
>Homo_sapiens_NP_002519
MCSPQESGMTALSARMLTRSRSLGPGAGPRGCREEPGPLRRREAAAEARKSHSPVKRPRKAQRLRVAYEGSDSEKGEGAE
PLKVPVWEPQDWQQQLVNIRAMRNKKDAPVDHLGTEHCYDSSAPPKVRRYQVLLSLMLSSQTKDQVTAGAMQRLRARGLT
VDSILQTDDATLGKLIYPVGFWRSKVKYIKQTSAILQQHYGGDIPASVAELVALPGVGPKMAHLAMAVAWGTVSGIAVDT
HVHRIANRLRWTKKATKSPEETRAALEEWLPRELWHEINGLLVGFGQQTCLPVHPRCHACLNQALCPAAQGL
>Homo_sapiens_NP_036354
MTPLVSRLSRLWAIMRKPRAAVGSGHRKQAASQEGRQKHAKNNSQAKPSACDGMIAECPGAPAGLARQPEEVVLQASVSS
YHLFRDVAEVTAFRGSLLSWYDQEKRDLPWRRRAEDEMDLDRRAYAVWVSEVMLQQTQVATVINYYTGWMQKWPTLQDLA
SASLEEVNQLWAGLGYYSRGRRLQEGARKVVEELGGHMPRTAETLQQLLPGVGRYTAGAIASIAFGQATGVVDGNVARVL
CRVRAIGADPSSTLVSQQLWGLAQQLVDPARPGDFNQAAMELGATVCTPQRPLCSQCPVESLCRARQRVEQEQLLASGSL
SGSPDVEECAPNTGQCHLCLPPSEPWDQTLGVVNFPRKASRKPPREESSATCVLEQPGALGAQILLVQRPNSGLLAGLWE
FPSVTWEPSEQLQRKALLQELQRWAGPLPATHLRHLGEVVHTFSHIKLTYQVYGLALEGQTPVTTVPPGARWLTQEEFHT
AAVSTAMKKVFRVYQGQQPGTCMGSKRSQVSSPCSRKKPRMGQQVLDNFFRSHISTDAHSLNSAAQ
>Mus_musculus_NP_032769
MNSGVRMVTRSRSRATRIASEGCREELAPREAAAEGRKSHRPVRHPRRTQKTHVAYEAANGEEGEDAEPLKVPVWEPQNW
QQQLANIRIMRSKKDAPVDQLGAEHCYDASASPKVRRYQVLLSLMLSSQTKDQVTAGAMQRLRARGLTVESILQTDDDTL
GRLIYPVGFWRNKVKYIKQTTAILQQRYEGDIPASVAELVALPGVGPKMAHLAMAVAWGTISGIAVDTHVHRIANRLRWT
KKMTKTPEETRKNLEEWLPRVLWSEVNGLLVGFGQQICLPVHPRCQACLNKALCPAAQDL
>Mus_musculus_NP_573513
MKKLQASVRSHKKQPANHKRRRTRALSSSQAKPSSLDGLAKQKREELLQASVSPYHLFSDVADVTAFRSNLLSWYDQEKR
DLPWRNLAKEEANSDRRAYAVWVSEVMLQQTQVATVIDYYTRWMQKWPKLQDLASASLEEVNQLWSGLGYYSRGRRLQEG
ARKVVEELGGHMPRTAETLQQLLPGVGRYTAGAIASIAFDQVTGVVDGNVLRVLCRVRAIGADPTSTLVSHHLWNLAQQL
VDPARPGDFNQAAMELGATVCTPQRPLCSHCPVQSLCRAYQRVQRGQLSALPGRPDIEECALNTRQCQLCLTSSSPWDPS
MGVANFPRKASRRPPREEYSATCVVEQPGAIGGPLVLLVQRPDSGLLAGLWEFPSVTLEPSEQHQHKALLQELQRWCGPL
PAIRLQHLGEVIHIFSHIKLTYQVYSLALDQAPASTAPPGARWLTWEEFCNAAVSTAMKKVFRMYEDHRQGTRKGSKRSQ
VCPPSSRKKPSLGQQVLDTFFQRHIPTDKPNSTTQ
>Bos_taurus_NP_001039862
MNAAGVRMVVTRARSRGTGASLRRRGEKAAPLRSGEAAAEERKSYSPVKRRRKAQRLSVAYEASEGEGGEGAEHLQAPSW
QPQDWRQQLDNIRTMRSGKDAPVDQLGAEHCFDPSASPKVRRYQVLLSLMLSSQTKDQVTAGAMQRLRARGLTVDSILQT
DDSTLGALIYPVGFWRSKVKYIKQTSAILQQRYDGDIPASVAELVALPGVGPKMAHLAMAVAWGTVSGIAVDTHVHRIAN
RLRWTKKATKSPEETRRALEEWLPRELWSEINGLLVGFGQQTCLPIRPRCQACLNRALCPAARGL
>Bos_taurus_NP_001039600
MKKSRAAVGNRSGRRKQASSQEGKEKCAFGSSQAKPSAPSAGPARQQKALLQASVSPYHLFRDVAEVTALQESLLDWYDR
KKRDLPWRRLVEDEVDLDRRAYAVWVAEVMLQQTQVATVINYYTRWMQKWPTLQDLASASLEEVNQLWAGLGYYSRGRWL
QEGARKVVEELGGHMPRTAETLQQFLPGVGRYTAGAIASIAFGQAAGVVDGNVIRVLCRVRAIGADSSSTLVSQHLWSLA
QQLVDPARPGDFNQAAMELGAIVCTPKRPLCSHCPVQNLCRARQRVEREQLSASQSLPGNCDVEECAPNTGQCPLCAPPT
EPWDQTLGVTNFPRKASRKPPREECSAICVLEQPKALGGAHILLVQRPNSGLLAGLWEFPSVSVNAEASGQHQRAALLQE
LQSWVGPLPDTRLQHLGQVVHTFSHIKMTYQVYSLALEEHTPVTIVPPGARWLTREDFHTAAVSTAMKKVFRMYEGQQPG
TCKGSKRSQVATLSKRKKPSPGQQVLESFFWPHVPTDAPSLNTAAQ
>Gallus_gallus_NP_001073043
MCAAAPRGGGRAARRLGAATAGSRVPSAAPRYSRRTRRVPIAYEAEPKPESPGPKWEPENWQQQLERIREMRRHRDAPVD
EMGVDKCYDTSAPPQVMRYQVLLSLMLSSQTKDQVTSAAMLRLRQRGLTVDSILQMDDATLGQIIYPVGFWRNKVKYIKQ
TTAILKQKYGGDIPGTVEELVKLPGVGPKMAHLAMNIAWNSVSGIAVDTHVHRITNRLKWVKKETRYPEETRVALEDWLP
RDLWREINWLLVGFGQQTCLPVNPRCKECLNQDICPAAKRF
>Gallus_gallus_XP_422433
MGGAAVRARRSVKVRAGGEHVGPGLGSPAIALRTHRRCCDPTPVPVSRQGLPLDHMHCISSVTPSRSMPYVAACSPGMTK
AGGTFPGGRWLQLSWMLTGGRMQLLGLLVASERQVWARERSCGDEDGEEGCWVGFCCSSWNQQHGDRGACCEKWHWHLCT
DPSVLLAVWVSEIMLQQTQVATVIDYYNRWMQKWPTLQALAAASLEEVNELWAGLGYYSRGKRLQEAARKVVSELAGRMP
RTAEDLQRLLPGVGRYTAGAIASISFGQATGVVDGNVIRVLCRLRCIGADTSSLAVIDCLWDMANTLVDRSRPGDFNQAL
MELGATVCTPKSPLCRECPVKEHCHAWRRVEKELASASQKLFGKTTLVPDVEDCGPGGCPLCLPAAEPWDSSLGVTNFPR
KAAKKQPRVEWTATCVLERRGRLGAPEYLIVQRPSSGLLAGLWEFPSLPLAPGLQEEQQKEVLADHLRAWTRQPVQTQSL
CFIGEVVHIFSHIHQTYVVYSLCLDGDVALDAASSPSRWVTEEEFRASAVSTAMKKVLKARETQRGVQSGRAKGSKRKRE
SKLGAAGSTPTGMQLSLRAFLRAQPPP
>Xenopus_(Silurana)_tropicalis_NP_001037884
MSGSLRPLGRRGRRGVLKAVGGKDQQDGTSKGQVIDDSEDEKPSSPKERSKRRVSVEYEQAASETVAKRPKWQPKNWAQH
LENIRQMRSRRDAPVDQMGAEKCYDQNAAPEVMRYQILLSLMLSSQTKDQVTSAAMCRLRQHGLTVSRILETDDGTLGKL
IYPVGFWKNKVKYIKQTTEILQEKYGGDIPDNVTDLVKLPGVGPKMAHLVMDIAWNNVSGIGVDTHVHRISNRLKWVRKE
TKTPEETRVAMEDWMPRELWSEINWLLVGFGQQVCLPVSPRCSECLNKDICPGAKKKKPR
>Xenopus_(Silurana)_tropicalis_NP_001072831
MPPPRTKTSLGRSAAASGKRKSPKQAFPKREEHVLQSSIYHSFTSQETEIIRDKLLAWYDKSKRDLPWRTMACTEPDLDR
KAYAVWVSEVMLQQTQVATVIDYYNKWMKVWPTMEDLARSSLEEVNEMWSGLGYYSRGRRLQEGAKKVVLELGGSMPRSA
DELQKLLPGVGRYTAGAIASISYGQVTGVVDGNVIRVLSRLRCIGADSSTLAVSDKLWNLANALVDPDRPGDFNQGMMEL
GATVCTPKKPLCTACPLQGQCKAYLKVIAEKESAVKTLIKKQASPIAKDVGDIEDCDLGPGLCALCVPTSDPWDSSLGVA
```

```
NFPRKSAKKPSRMEQTAICVWEKCGDHGELEYLIVQRPSSGLLAGLWEFPSILLDEKFTEQNRQHSLLGLLQDLSGHAVP
LQKLQYKGEVVHIFSHIHQTYVVYFLSLNTTENCSVKTEETERPLTRWVTKKEFLNSAVPTAMKKIMKLCESHGSSCTAV
NTSKKRKGDLAKVQLPSGRIKTEKGKQQSIQSFFKLATEK
>Takifugu_rubripes_XP_003972103
MTSHYFAQSRSVVTRRGAQNAAHKPATSLKSKLTIQPEKDDLVSSSAGVKLEEEEAKISGNALKPETDAPTLSSHSRRRR
QLKVEYDKDGSMPQLKTEPWEPPRWKTQLENIRAMRSGRDAPVDNMGADKCHDADAPAHVKRFQVLVSLMLSSQTKDQVT
SAAMQKLRAHGCTVENILATNDETLGQLIYPVGFWRNKVKYLKLTSAMLQKEFGGDIPDSVEGLVRLPGVGPKMAHLAMD
IAWDQVSGIGVDTHVHRISNRLGWLKKPTKTPEETRKSLEEWLPRELWSEINWLLVGFGQQVCLPVSPLCSVCLNQHDCP
SAHKNSPVRRPKFLERSPWIKSPRLFITPGANTFLIR
>Takifugu_rubripes_XP_003974154
MSTQGEPVQVGKVLSFLREWDRGDRSARGRMLSSFLGRSAGRTQGELEYLGEFVGHGGVVTLLEVLTQPQSNEETKAEAL
CLLLAISDAGRKYKELICQSCGAMAAAECLTHSGTGETQESAWMLLESLSHGNPKYEGEIYKGLIGHLTCTSAKAQQFVL
HTLHTLQSKMEIAHHSIVEPLLGVLTSLHPDVQSEVARLIFELRRYDVRPMLLRALCGLGLNVARAPTYPEEESSHASSP
SAYHFFHDAADVALLRSRLLAWYDQEKRELPWRTLALTEPDVNIRTYAVWVSEIMLQQTQVATVIDYYNKWMKRWPTVQD
LATATLEDVNQMWAGLGYYSRGKRLHEGAQKVVSQLQGEMPRTVDALLKQLPGVGRYTAGAIGSIALGQVTGAVDGNVIR
VLCRLRAIGADCTGPTVTEALWSLANTLVDPDRPGDFNQALMELGARVCTPKGPLCTRCPIQSHCHSYHKAHFKQEKNSM
KLMGKLDRKSSALPDIEDCLSSGTCTLCLSEPWDDELGVQNFPRKPAKKPPRAERCLTCVVIRQGEGGEHEFLLTQRPSK
GLLAGLWEFPCINHEEKNAVEEKKVLCAEINRILGTSLTHGLLQYVGEVVHIFSHIHQTYVVHTLRLKDAVSQSENMQWL
TPSALQEAAVSTGVKKIMKLCNSALGQQGAPDGEEKRPKKDRKGQITKRPRLSGANSRSRQLSLSSFFQTVKQDC
```

APPENDIX 5:

12 vertebrate homologs, final headers

```
>Homo_sapiens_Nth
MCSPQESGMTALSARMLTRSRSLGPGAGPRGCREEPGPLRRREAAAEARKSHSPVKRPRKAQRLRVAYEGSDSEKGEGAE
PLKVPVWEPQDWQQQLVNIRAMRNKKDAPVDHLGTEHCYDSSAPPKVRRYQVLLSLMLSSQTKDQVTAGAMQRLRARGLT
VDSILQTDDATLGKLIYPVGFWRSKVKYIKQTSAILQQHYGGDIPASVAELVALPGVGPKMAHLAMAVAWGTVSGIAVDT
HVHRIANRLRWTKKATKSPEETRAALEEWLPRELWHEINGLLVGFGQQTCLPVHPRCHACLNQALCPAAQGL
>Homo_sapiens_MutY
MTPLVSRLSRLWAIMRKPRAAVGSGHRKQAASQEGRQKHAKNNSQAKPSACDGMIAECPGAPAGLARQPEEVVLQASVSS
YHLFRDVAEVTAFRGSLLSWYDQEKRDLPWRRRAEDEMDLDRRAYAVWVSEVMLQQTQVATVINYYTGWMQKWPTLQDLA
SASLEEVNQLWAGLGYYSRGRRLQEGARKVVEELGGHMPRTAETLQQLLPGVGRYTAGAIASIAFGQATGVVDGNVARVL
CRVRAIGADPSSTLVSQQLWGLAQQLVDPARPGDFNQAAMELGATVCTPQRPLCSQCPVESLCRARQRVEQEQLLASGSL
SGSPDVEECAPNTGQCHLCLPPSEPWDQTLGVVNFPRKASRKPPREESSATCVLEQPGALGAQILLVQRPNSGLLAGLWE
FPSVTWEPSEQLQRKALLQELQRWAGPLPATHLRHLGEVVHTFSHIKLTYQVYGLALEGQTPVTTVPPGARWLTQEEFHT
AAVSTAMKKVFRVYQGQQPGTCMGSKRSQVSSPCSRKKPRMGQQVLDNFFRSHISTDAHSLNSAAQ
>Mus_musculus_Nth
MNSGVRMVTRSRSRATRIASEGCREELAPREAAAEGRKSHRPVRHPRRTQKTHVAYEAANGEEGEDAEPLKVPVWEPQNW
QQQLANIRIMRSKKDAPVDQLGAEHCYDASASPKVRRYQVLLSLMLSSQTKDQVTAGAMQRLRARGLTVESILQTDDDTL
GRLIYPVGFWRNKVKYIKQTTAILQQRYEGDIPASVAELVALPGVGPKMAHLAMAVAWGTISGIAVDTHVHRIANRLRWT
KKMTKTPEETRKNLEEWLPRVLWSEVNGLLVGFGQQICLPVHPRCQACLNKALCPAAQDL
>Mus_musculus_MutY
MKKLQASVRSHKKQPANHKRRRTRALSSSQAKPSSLDGLAKQKREELLQASVSPYHLFSDVADVTAFRSNLLSWYDQEKR
DLPWRNLAKEEANSDRRAYAVWVSEVMLQQTQVATVIDYYTRWMQKWPKLQDLASASLEEVNQLWSGLGYYSRGRRLQEG
ARKVVEELGGHMPRTAETLQQLLPGVGRYTAGAIASIAFDQVTGVVDGNVLRVLCRVRAIGADPTSTLVSHHLWNLAQQL
VDPARPGDFNQAAMELGATVCTPQRPLCSHCPVQSLCRAYQRVQRGQLSALPGRPDIEECALNTRQCQLCLTSSSPWDPS
MGVANFPRKASRRPPREEYSATCVVEQPGAIGGPLVLLVQRPDSGLLAGLWEFPSVTLEPSEQHQHKALLQELQRWCGPL
PAIRLQHLGEVIHIFSHIKLTYQVYSLALDQAPASTAPPGARWLTWEEFCNAAVSTAMKKVFRMYEDHRQGTRKGSKRSQ
VCPPSSRKKPSLGQQVLDTFFQRHIPTDKPNSTTQ
>Bos_taurus_Nth
MNAAGVRMVVTRARSRGTGASLRRRGEKAAPLRSGEAAAEERKSYSPVKRRRKAQRLSVAYEASEGEGGEGAEHLQAPSW
QPQDWRQQLDNIRTMRSGKDAPVDQLGAEHCFDPSASPKVRRYQVLLSLMLSSQTKDQVTAGAMQRLRARGLTVDSILQT
DDSTLGALIYPVGFWRSKVKYIKQTSAILQQRYDGDIPASVAELVALPGVGPKMAHLAMAVAWGTVSGIAVDTHVHRIAN
RLRWTKKATKSPEETRRALEEWLPRELWSEINGLLVGFGQQTCLPIRPRCQACLNRALCPAARGL
>Bos_taurus_MutY
MKKSRAAVGNRSGRRKQASSQEGKEKCAFGSSQAKPSAPSAGPARQQKALLQASVSPYHLFRDVAEVTALQESLLDWYDR
KKRDLPWRRLVEDEVDLDRRAYAVWVAEVMLQQTQVATVINYYTRWMQKWPTLQDLASASLEEVNQLWAGLGYYSRGRWL
QEGARKVVEELGGHMPRTAETLQQFLPGVGRYTAGAIASIAFGQAAGVVDGNVIRVLCRVRAIGADSSSTLVSQHLWSLA
QQLVDPARPGDFNQAAMELGAIVCTPKRPLCSHCPVQNLCRARQRVEREQLSASQSLPGNCDVEECAPNTGQCPLCAPPT
EPWDQTLGVTNFPRKASRKPPREECSAICVLEQPKALGGAHILLVQRPNSGLLAGLWEFPSVSVNAEASGQHQRAALLQE
LQSWVGPLPDTRLQHLGQVVHTFSHIKMTYQVYSLALEEHTPVTIVPPGARWLTREDFHTAAVSTAMKKVFRMYEGQQPG
TCKGSKRSQVATLSKRKKPSPGQQVLESFFWPHVPTDAPSLNTAAQ
>Gallus_gallus_Nth
MCAAAPRGGGRAARRLGAATAGSRVPSAAPRYSRRTRRVPIAYEAEPKPESPGPKWEPENWQQQLERIREMRRHRDAPVD
EMGVDKCYDTSAPPQVMRYQVLLSLMLSSQTKDQVTSAAMLRLRQRGLTVDSILQMDDATLGQIIYPVGFWRNKVKYIKQ
TTAILKQKYGGDIPGTVEELVKLPGVGPKMAHLAMNIAWNSVSGIAVDTHVHRITNRLKWVKKETRYPEETRVALEDWLP
RDLWREINWLLVGFGQQTCLPVNPRCKECLNQDICPAAKRF
>Gallus_gallus_MutY
MGGAAVRARRSVKVRAGGEHVGPGLGSPAIALRTHRRCCDPTPVPVSRQGLPLDHMHCISSVTPSRSMPYVAACSPGMTK
AGGTFPGGRWLQLSWMLTGGRMQLLGLLVASERQVWARERSCGDEDGEEGCWVGFCCSSWNQQHGDRGACCEKWHWHLCT
DPSVLLAVWVSEIMLQQTQVATVIDYYNRWMQKWPTLQALAAASLEEVNELWAGLGYYSRGKRLQEAARKVVSELAGRMP
RTAEDLQRLLPGVGRYTAGAIASISFGQATGVVDGNVIRVLCRLRCIGADTSSLAVIDCLWDMANTLVDRSRPGDFNQAL
MELGATVCTPKSPLCRECPVKEHCHAWRRVEKELASASQKLFGKTTLVPDVEDCGPGGCPLCLPAAEPWDSSLGVTNFPR
KAAKKQPRVEWTATCVLERRGRLGAPEYLIVQRPSSGLLAGLWEFPSLPLAPGLQEEQQKEVLADHLRAWTRQPVQTQSL
CFIGEVVHIFSHIHQTYVVYSLCLDGDVALDAASSPSRWVTEEEFRASAVSTAMKKVLKARETQRGVQSGRAKGSKRKRE
SKLGAAGSTPTGMQLSLRAFLRAQPPP
>Xenopus_(Silurana)_tropicalis_Nth
MSGSLRPLGRRGRRGVLKAVGGKDQQDGTSKGQVIDDSEDEKPSSPKERSKRRVSVEYEQAASETVAKRPKWQPKNWAQH
LENIRQMRSRRDAPVDQMGAEKCYDQNAAPEVMRYQILLSLMLSSQTKDQVTSAAMCRLRQHGLTVSRILETDDGTLGKL
IYPVGFWKNKVKYIKQTTEILQEKYGGDIPDNVTDLVKLPGVGPKMAHLVMDIAWNNVSGIGVDTHVHRISNRLKWVRKE
TKTPEETRVAMEDWMPRELWSEINWLLVGFGQQVCLPVSPRCSECLNKDICPGAKKKKPR
>Xenopus_(Silurana)_tropicalis_MutY
MPPPRTKTSLGRSAAASGKRKSPKQAFPKREEHVLQSSIYHSFTSQETEIIRDKLLAWYDKSKRDLPWRTMACTEPDLDR
KAYAVWVSEVMLQQTQVATVIDYYNKWMKVWPTMEDLARSSLEEVNEMWSGLGYYSRGRRLQEGAKKVVLELGGSMPRSA
DELQKLLPGVGRYTAGAIASISYGQVTGVVDGNVIRVLSRLRCIGADSSTLAVSDKLWNLANALVDPDRPGDFNQGMMEL
GATVCTPKKPLCTACPLQGQCKAYLKVIAEKESAVKTLIKKQASPIAKDVGDIEDCDLGPGLCALCVPTSDPWDSSLGVA
```

```
NFPRKSAKKPSRMEQTAICVWEKCGDHGELEYLIVQRPSSGLLAGLWEFPSILLDEKFTEQNRQHSLLGLLQDLSGHAVP
LQKLQYKGEVVHIFSHIHQTYVVYFLSLNTTENCSVKTEETERPLTRWVTKKEFLNSAVPTAMKKIMKLCESHGSSCTAV
NTSKKRKGDLAKVQLPSGRIKTEKGKQQSIQSFFKLATEK
>Takifugu_rubripes_Nth
MTSHYFAQSRSVVTRRGAQNAAHKPATSLKSKLTIQPEKDDLVSSSAGVKLEEEEAKISGNALKPETDAPTLSSHSRRRR
QLKVEYDKDGSMPQLKTEPWEPPRWKTQLENIRAMRSGRDAPVDNMGADKCHDADAPAHVKRFQVLVSLMLSSQTKDQVT
SAAMQKLRAHGCTVENILATNDETLGQLIYPVGFWRNKVKYLKLTSAMLQKEFGGDIPDSVEGLVRLPGVGPKMAHLAMD
IAWDQVSGIGVDTHVHRISNRLGWLKKPTKTPEETRKSLEEWLPRELWSEINWLLVGFGQQVCLPVSPLCSVCLNQHDCP
SAHKNSPVRRPKFLERSPWIKSPRLFITPGANTFLIR
>Takifugu_rubripes_MutY
MSTQGEPVQVGKVLSFLREWDRGDRSARGRMLSSFLGRSAGRTQGELEYLGEFVGHGGVVTLLEVLTQPQSNEETKAEAL
CLLLAISDAGRKYKELICQSCGAMAAAECLTHSGTGETQESAWMLLESLSHGNPKYEGEIYKGLIGHLTCTSAKAQQFVL
HTLHTLQSKMEIAHHSIVEPLLGVLTSLHPDVQSEVARLIFELRRYDVRPMLLRALCGLGLNVARAPTYPEEESSHASSP
SAYHFFHDAADVALLRSRLLAWYDQEKRELPWRTLALTEPDVNIRTYAVWVSEIMLQQTQVATVIDYYNKWMKRWPTVQD
LATATLEDVNQMWAGLGYYSRGKRLHEGAQKVVSQLQGEMPRTVDALLKQLPGVGRYTAGAIGSIALGQVTGAVDGNVIR
VLCRLRAIGADCTGPTVTEALWSLANTLVDPDRPGDFNQALMELGARVCTPKGPLCTRCPIQSHCHSYHKAHFKQEKNSM
KLMGKLDRKSSALPDIEDCLSSGTCTLCLSEPWDDELGVQNFPRKPAKKPPRAERCLTCVVIRQGEGGEHEFLLTQRPSK
GLLAGLWEFPCINHEEKNAVEEKKVLCAEINRILGTSLTHGLLQYVGEVVHIFSHIHQTYVVHTLRLKDAVSQSENMQWL
TPSALQEAAVSTGVKKIMKLCNSALGQQGAPDGEEKRPKKDRKGQITKRPRLSGANSRSRQLSLSSFFQTVKQDC
```

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
```

APPENDIX 7:

4 human homologs, modified headers

```
>NTHL1_Homo_sapiens
MCSPQESGMTALSARMLTRSRSLGPGAGPRGCREEPGPLRRREAAAEARKSHSPVKRPRKAQRLRVAYEG
SDSEKGEGAEPLKVPVWEPQDWQQQLVNIRAMRNKKDAPVDHLGTEHCYDSSAPPKVRRYQVLLSLMLSS
QTKDQVTAGAMQRLRARGLTVDSILQTDDATLGKLIYPVGFWRSKVKYIKQTSAILQQHYGGDIPASVAE
LVALPGVGPKMAHLAMAVAWGTVSGIAVDTHVHRIANRLRWTKKATKSPEETRAALEEWLPRELWHEING
LLVGFGQQTCLPVHPRCHACLNQALCPAAQGL
>MUTYH_Homo_sapiens
MTPLVSRLSRLWAIMRKPRAAVGSGHRKQAASQEGRQKHAKNNSQAKPSACDGMIAECPGAPAGLARQPE
EVVLQASVSSYHLFRDVAEVTAFRGSLLSWYDQEKRDLPWRRRAEDEMDLDRRAYAVWVSEVMLQQTQVA
TVINYYTGWMQKWPTLQDLASASLEEVNQLWAGLGYYSRGRRLQEGARKVVEELGGHMPRTAETLQQLLP
GVGRYTAGAIASIAFGQATGVVDGNVARVLCRVRAIGADPSSTLVSQQLWGLAQQLVDPARPGDFNQAAM
ELGATVCTPQRPLCSQCPVESLCRARQRVEQEQLLASGSLSGSPDVEECAPNTGQCHLCLPPSEPWDQTL
GVVNFPRKASRKPPREESSATCVLEQPGALGAQILLVQRPNSGLLAGLWEFPSVTWEPSEQLQRKALLQE
LQRWAGPLPATHLRHLGEVVHTFSHIKLTYQVYGLALEGQTPVTTVPPGARWLTQEEFHTAAVSTAMKKV
FRVYQGQQPGTCMGSKRSQVSSPCSRKKPRMGQQVLDNFFRSHISTDAHSLNSAAQ
>OGG1_Homo_sapiens
MPARALLPRRMGHRTLASTPALWASIPCPRSELRLDLVLPSGQSFRWREQSPAHWSGVLADQVWTLTQTE
EQLHCTVYRGDKSQASRPTPDELEAVRKYFQLDVTLAQLYHHWGSVDSHFQEVAQKFQGVRLLRQDPIEC
LFSFICSSNNNIARITGMVERLCQAFGPRLIQLDDVTYHGFPSLQALAGPEVEAHLRKLGLGYRARYVSA
SARAILEEQGGLAWLQQLRESSYEEAHKALCILPGVGTKVADCICLMALDKPQAVPVDVHMWHIAQRDYS
WHPTTSQAKGPSPQTNKELGNFFRSLWGPYAGWAQAVLFSADLRQSRHAQEPPAKRRKGSKGPEG
>MBD4_Homo_sapiens
MGTTGLESLSLGDRGAAPTVTSSERLVPDPPNDLRKEDVAMELERVGEDEEQMMIKRSSECNPLLQEPIA
SAQFGATAGTECRKSVPCGWERVVKQRLFGKTAGRFDVYFISPQGLKFRSKSSLANYLHKNGETSLKPED
FDFTVLSKRGIKSRYKDCSMAALTSHLQNQSNNSNWNLRTRSKCKKDVFMPPSSSSELQESRGLSNFTST
HLLLKEDEGVDDVNFRKVRKPKGKVTILKGIPIKKTKKGCRKSCSGFVQSDSKRESVCNKADAESEPVAQ
KSQLDRTVCISDAGACGETLSVTSEENSLVKKKERSLSSGSNFCSEQKTSGIINKFCSAKDSEHNEKYED
TFLESEEIGTKVEVVERKEHLHTDILKRGSEMDNNCSPTRKDFTGEKIFQEDTIPRTQIERRKTSLYFSS
KYNKEALSPPRRKAFKKWTPPRSPFNLVQETLFHDPWKLLIATIFLNRTSGKMAIPVLWKFLEKYPSAEV
ARTADWRDVSELLKPLGLYDLRAKTIVKFSDEYLTKQWKYPIELHGIGKYGNDSYRIFCVNEWKQVHPED
HKLNKYHDWLWENHEKLSLS
```