

# Sequence searching and sequence alignments – MBV-INF410

---

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) where 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](http://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
IFRVCNRTQFAPGKNVEQVEEKLKVVPAEFKVDCHHWLILHGRYTCIARKPRCGSCIIEDLCEYKEKVD
I
```

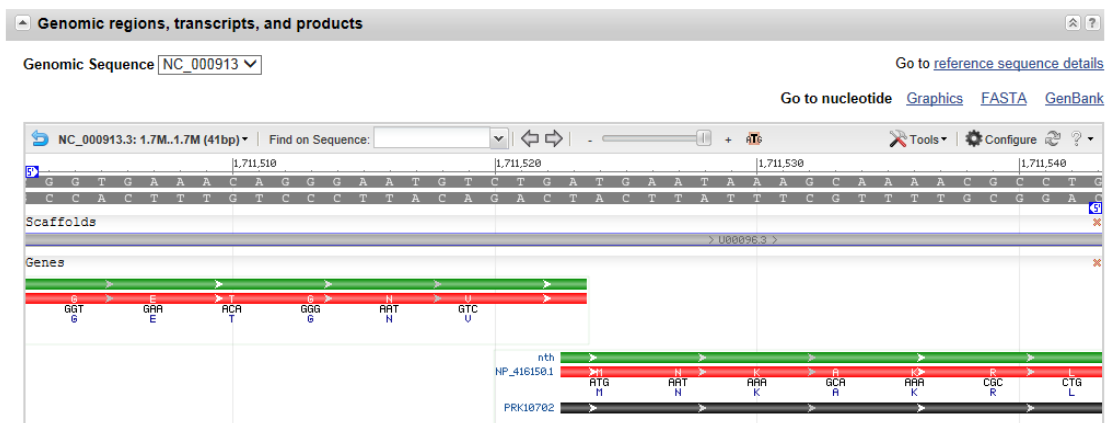
3. Follow the link to the corresponding Gene (in the list of “Related information” at the right hand side). What is the NCBI gene identifier (Gene ID) for the gene? What is the Swiss-Prot identifier for this protein? Which genes are found directly upstream and downstream of *nth*? Are the three genes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

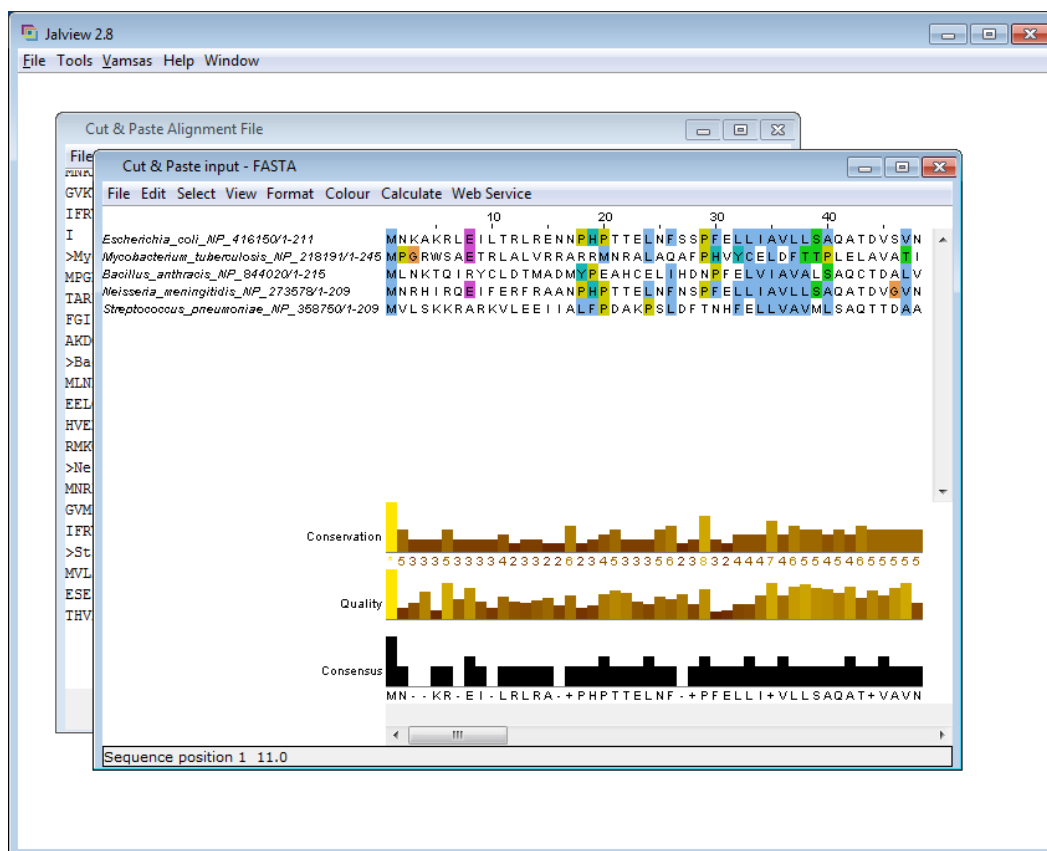
The last amino acid of *rsxE* is Val, encoded by GTC, and the stop codon is TGA. The start codon of *nth* is encoded by ATG. In this case the A of TGA (stop) is the same as the A of ATG (start). The two genes overlap by a single nucleotide, and there are, obviously, no nucleotides between them.

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 anthracis* 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”.

- 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.

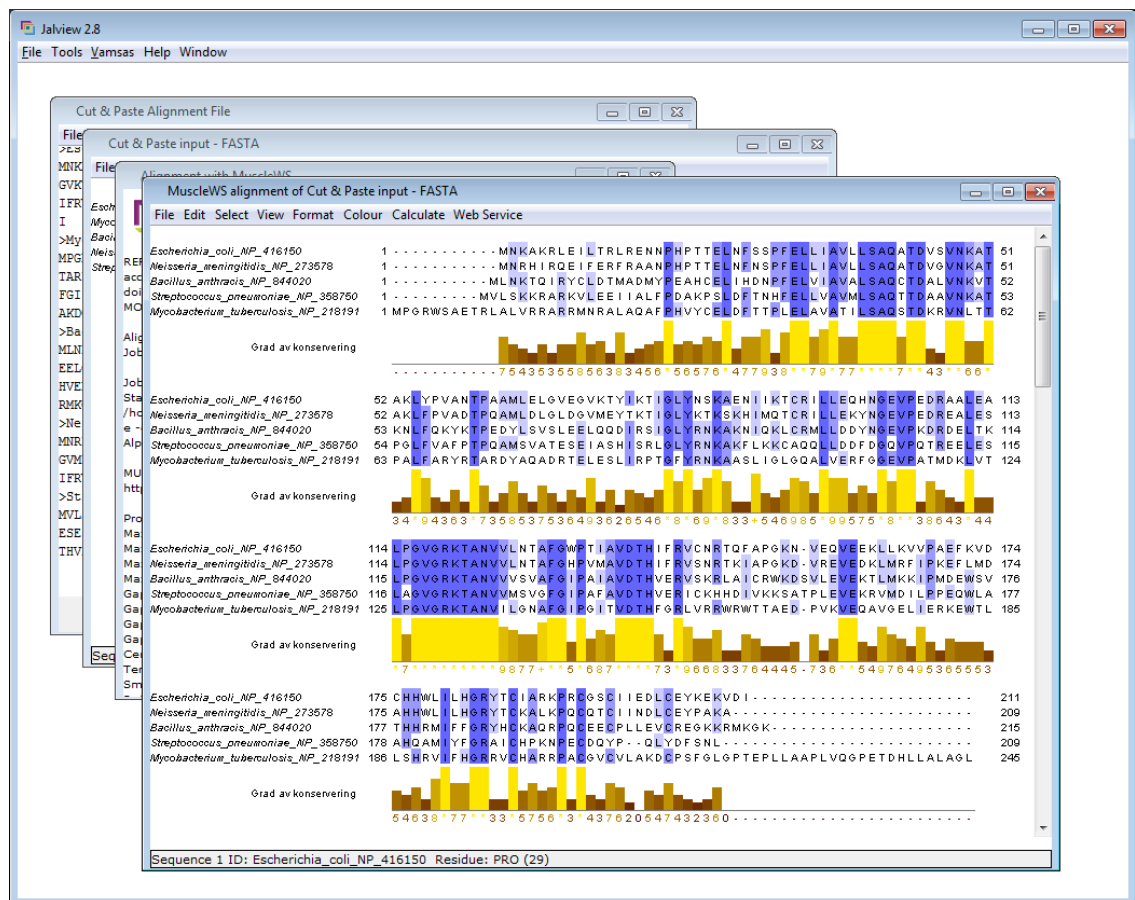


- Use the MUSCLE-algorithm web service (found under “Web Service” → “Alignment”) (with “Muscle with Defaults”) to generate a multiple sequence alignment (MSA). Can you say anything about on which computer the MUSCLE-algorithm is running? Where on the planet? Hint: Look in “Tools” → “Preferences” → “Web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

- 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".

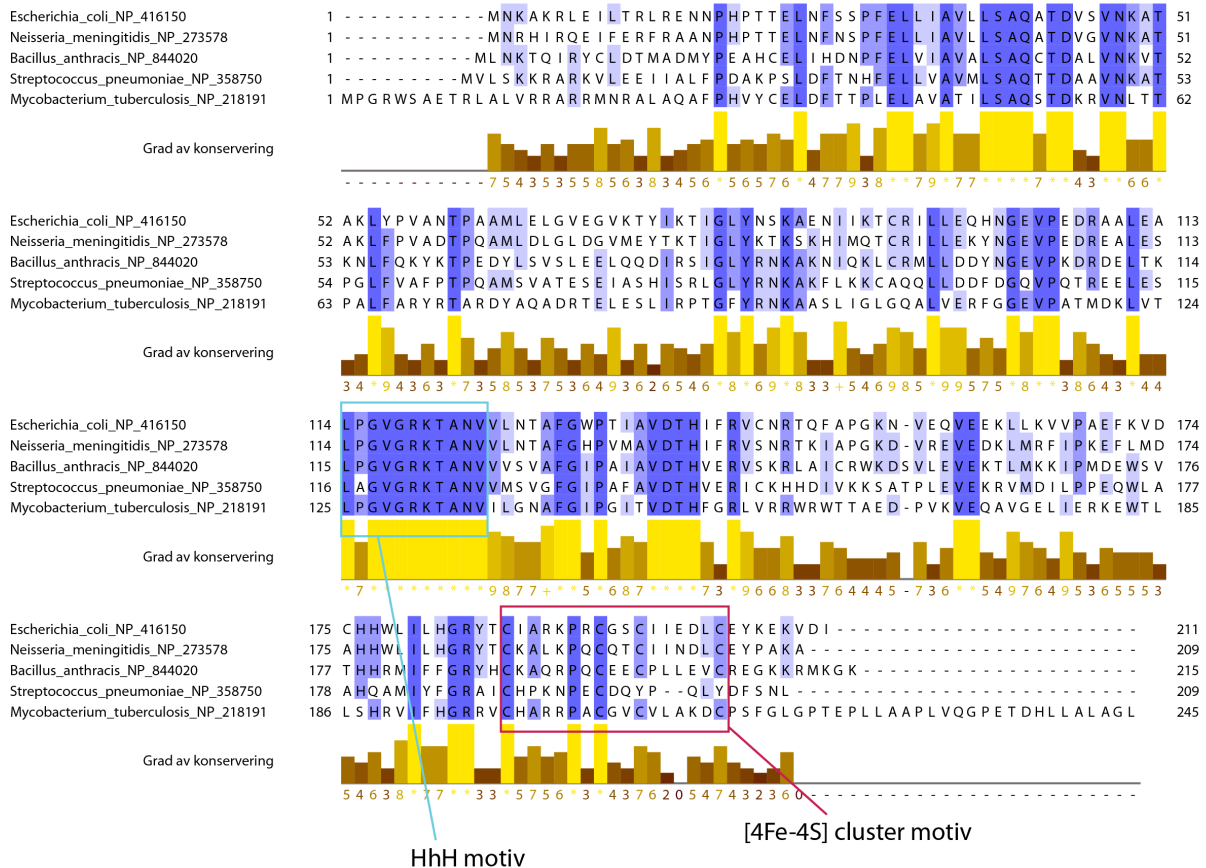


Export the alignment in PNG format, and import it into PowerPoint, Adobe Illustrator, or a similar program in order to add some extra information in the MSA. Indicate the residues involved in the helix-hairpin-helix (HhH) motif (LxGVGxK) and the [4Fe-4S] (iron sulphur) cluster motif (CxxxxxCxxCxxxxxC).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

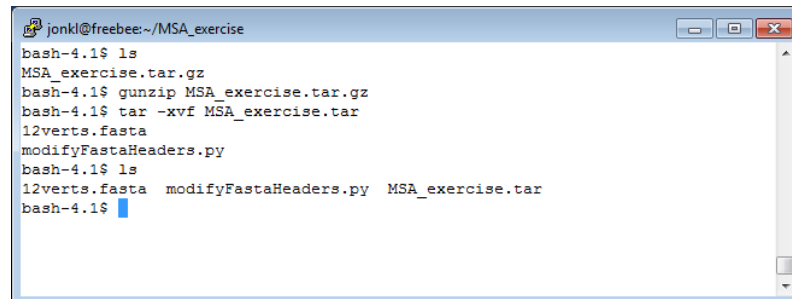
16. We want, as for the bacterial protein sequences, to shorten each sequence title to contain only the species name and RefSeq identifier. We could do this manually, in a text editor, as we did above for the bacterial homologs. However, the task here will be to use a little program or script to do this. 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](http://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](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.

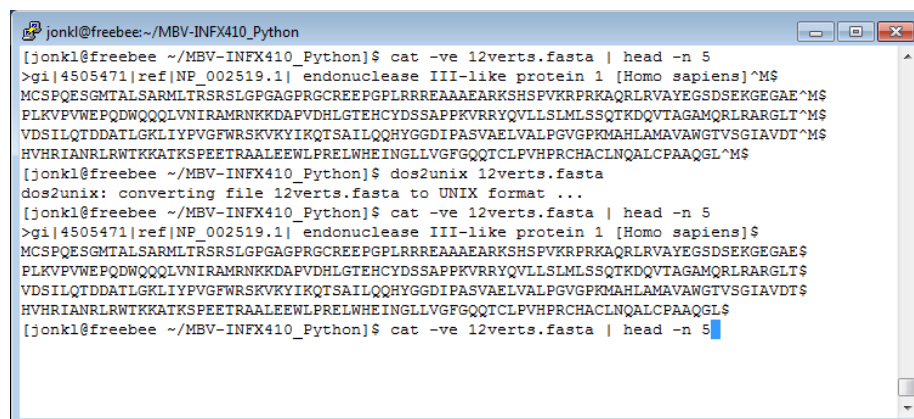


```

jonkl@freebee: ~/MSA_exercise
bash-4.1$ ls
MSA_exercise.tar.gz
bash-4.1$ gunzip MSA_exercise.tar.gz
bash-4.1$ tar -xvf MSA_exercise.tar
12verts.fasta
modifyFastaHeaders.py
bash-4.1$ ls
12verts.fasta  modifyFastaHeaders.py  MSA_exercise.tar
bash-4.1$

```

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`.



```

jonkl@freebee: ~/MBV-INF410_Python
[jonkl@freebee ~/MBV-INF410_Python]$ cat -ve 12verts.fasta | head -n 5
>gi|4505471|ref|NP_002519.1| endonuclease III-like protein 1 [Homo sapiens]^M$
MCSPQESGMTALSARMLTRSRSLGPGAGPRGCREEPGPLRRREAAAEARKSHSPVKRPRKAQRLRVAYEGSDSEKGEAE^M$
PLKVPVWNEPQDWQQQLVNIRAMRNKKDAPVDHLGTEHCYDSSAPPKVRRYQVLLSLMLSSQTKDQVTAGAMQRLRARGLT^M$
VDSILQTDATLGKLIYPVGFWRSKVKYIKQTSAILQQHYGGDIPASVAELVALPGVGPKMAHLAMAVAWGTVSGIAVDT^M$
HVVHRIANRLRWTKKATKSPEETRAALEEWLPRELWHEINGLLVGFQQTCPLVHPRCHACLNQALCPAAQGL^M$
[jonkl@freebee ~/MBV-INF410_Python]$ dos2unix 12verts.fasta
dos2unix: converting file 12verts.fasta to UNIX format ...
[jonkl@freebee ~/MBV-INF410_Python]$ cat -ve 12verts.fasta | head -n 5
>gi|4505471|ref|NP_002519.1| endonuclease III-like protein 1 [Homo sapiens]$
MCSPQESGMTALSARMLTRSRSLGPGAGPRGCREEPGPLRRREAAAEARKSHSPVKRPRKAQRLRVAYEGSDSEKGEAE$
PLKVPVWNEPQDWQQQLVNIRAMRNKKDAPVDHLGTEHCYDSSAPPKVRRYQVLLSLMLSSQTKDQVTAGAMQRLRARGLT$
VDSILQTDATLGKLIYPVGFWRSKVKYIKQTSAILQQHYGGDIPASVAELVALPGVGPKMAHLAMAVAWGTVSGIAVDT$
HVVHRIANRLRWTKKATKSPEETRAALEEWLPRELWHEINGLLVGFQQTCPLVHPRCHACLNQALCPAAQGL$
[jonkl@freebee ~/MBV-INF410_Python]$ cat -ve 12verts.fasta | head -n 5

```

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]
MCSPQESGMTALSARMLTRSRSLGPGAGPRGCREEPGPLRRREAAAEARKSHSPVKRPRKAQRLRVAYEGSDSEKGEAE
PLKVPVWNEPQDWQQQLVNIRAMRNKKDAPVDHLGTEHCYDSSAPPKVRRYQVLLSLMLSSQTKDQVTAGAMQRLRARGLT
...

```

should become

```

>Homo_sapiens_NP_002519

```



```

MCSPPQESGMTALSARMLTRSRSLGPGAGPRGCREEPGLRRREAAAARKSHSPVKRPRKAQRLRVAYEGSDSEKGEAE
PLKVPVWEPQDWQQQLVNIRAMRNKKDAPVDHLGTEHCYDSSAPPKVRRYQVLLSMLSSQTKDQVTAGAMQRLRARGLT
...

```

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

```

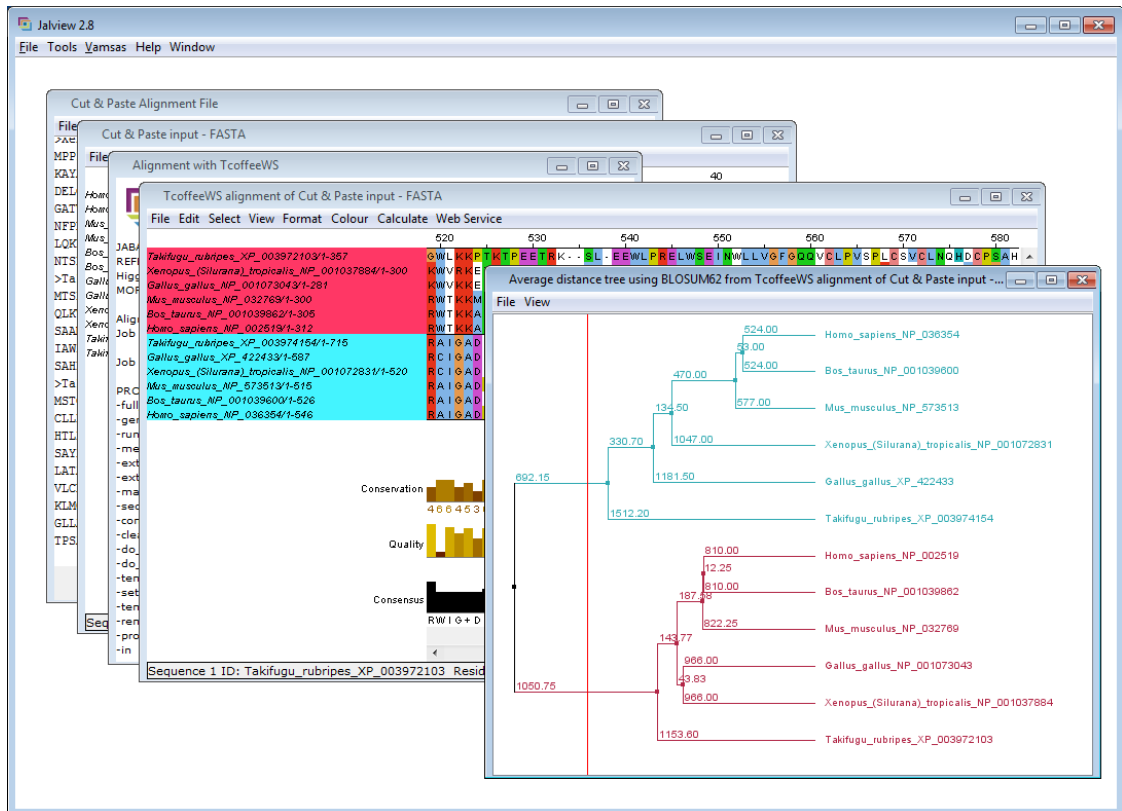
jonkl@freebee: ~/MSA_exercise
[jonkl@freebee ~/MSA_exercise]$ python modifyFastaHeaders.py 12verts.fasta 12verts_new.fasta
Processing NP_002519 from species: Homo_sapiens
Processing NP_036354 from species: Homo_sapiens
Processing NP_032769 from species: Mus_musculus
Processing NP_573513 from species: Mus_musculus
Processing NP_001039862 from species: Bos_taurus
Processing NP_001039600 from species: Bos_taurus
Processing NP_001073043 from species: Gallus_gallus
Processing XP_422433 from species: Gallus_gallus
Processing NP_001037884 from species: Xenopus_(Silurana)_tropicalis
Processing NP_001072831 from species: Xenopus_(Silurana)_tropicalis
Processing XP_003972103 from species: Takifugu_rubripes
Processing XP_003974154 from species: Takifugu_rubripes
[jonkl@freebee ~/MSA_exercise]$ head -n 6 12verts_new.fasta
>Homo_sapiens_NP_002519
MCSPPQESGMTALSARMLTRSRSLGPGAGPRGCREEPGLRRREAAAARKSHSPVKRPRKAQRLRVAYEGSDSEKGEAE
PLKVPVWEPQDWQQQLVNIRAMRNKKDAPVDHLGTEHCYDSSAPPKVRRYQVLLSMLSSQTKDQVTAGAMQRLRARGLT
VDSLLQDDADLGLIYPVGFWRSKVKYIKQTSAILQQHYGGDIPASVAELVALPGVGPMAHLAMAVANGTVSGIAVDT
HVHRIANRLRWTKKATKSPEETRAALEEWLPRELWHEINGLLVFGQQTCLPVHPRCHACLNQALCPAAQGL
>Homo_sapiens_NP_036354
[jonkl@freebee ~/MSA_exercise]$

```

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.



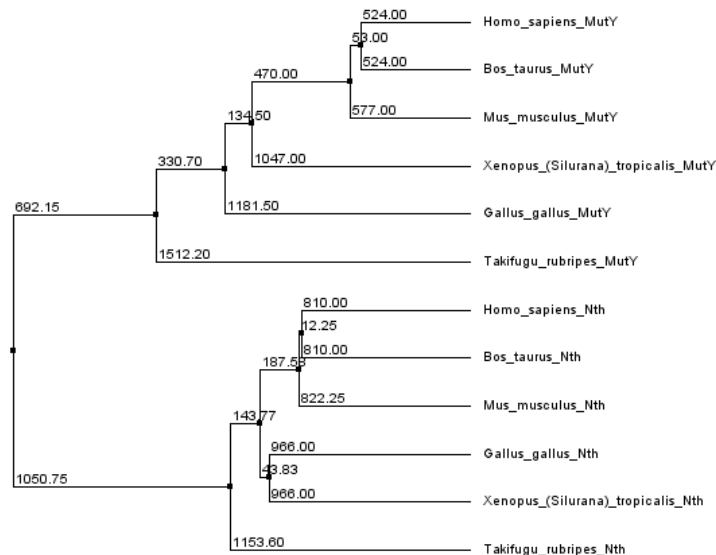
26. Use the terms homologs, paralogs, and orthologs to describe the relationships between these proteins/genes.

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\_MutY). 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?

```

Takifugu_rubripes_Nth      133 -DADAPAHVKRFQVLVSLMLSSTQTKDQVTS AAMQKLR AHGCTVENILATNDET LGQLIYPVGFWRNKVKYLKLT S 206
Xenopus_(Silurana)_tropicalis_Nth 105 -DQNAAPEVMRYQILLSLMLSSTQTKDQVTS AAMCRLRQHGLTVSRILETDDGT LGQLIYPVGFWRNKVKYIKQTT 178
Gallus_gallus_Nth        89 -DTSAPPQVMRYQVLLSLMLSSTQTKDQVTS AAMLRLRQRGLTVDSILQMDDAT LGQLIYPVGFWRNKVKYIKQTT 162
Mus_musculus_Nth         108 -DASASPQVRRYQVLLSLMLSSTQTKDQVTS AAMQRLRARGLTVESILQTDGDT LGRLIYPVGFWRNKVKYIKQTT 181
Bos_taurus_Nth           113 -DPSASPQVRRYQVLLSLMLSSTQTKDQVTS AAMQRLRARGLTVDISILQTDGDT LGRLIYPVGFWRNKVKYIKQTS 186
Homo_sapiens_Nth         120 -DSSASPQVRRYQVLLSLMLSSTQTKDQVTS AAMQRLRARGLTVDISILQTDGDT LGRLIYPVGFWRNKVKYIKQTS 193
Takifugu_rubripes_MutY    281 - - - - -DVNIRITAVVWSEIMLQQTQVATVIDYYNKWMKWPTVQDLATATLEDVNQMMWAGLGYS -RGKRLHEGA 349
Gallus_gallus_MutY       158 LCTDPSV - - - - -LLAVVWSEIMLQQTQVATVIDYYNKWMKWPTLQALAAASLEEVNELWAGLGYS -RGKRLHEGA 228
Xenopus_(Silurana)_tropicalis_MutY 73 -CTEPDLDRKAYAVVWSEVMLQQTQVATVIDYYNKWMKWPTMEDLARSSLEEVNEMWSGLGYS -RGRRLQEGA 145
Mus_musculus_MutY        93 - - - - -NSDRRAYAVVWSEVMLQQTQVATVIDYYTWMQKWPKLQDLASASLEEVNQLWSGLGYS -RGRRLQEGA 161
Bos_taurus_MutY          96 - - - - -DLDRRAYAVVWAEVMLQQTQVATVINYYTWMQKWPTLQDLASASLEEVNQLWAGLGYS -RGRRLQEGA 164
Homo_sapiens_MutY        119 - - - - -DLDRRAYAVVWSEVMLQQTQVATVINYYTGMQKWPTLQDLASASLEEVNQLWAGLGYS -RGRRLQEGA 187

Takifugu_rubripes_Nth      207 AMLQKEFGGDIIPDSVEGLVR -LPGVGPKMAHLAMDIAWDQVSGIGVDTHVHRIINRLGWLKKPTKTPEETR K - - S 278
Xenopus_(Silurana)_tropicalis_Nth 179 EILQEKYGGDIIPDNVTDLVK -LPGVGPKMAHLVMDIAWNNVSGIGVDTHVHRIINRLKWKVKETKTPEETR V - - A 250
Gallus_gallus_Nth        163 AILKQKYGGDIIPGTVEELVK -LPGVGPKMAHLAMNLAWNSVSGIAYDTHVHRIINRLKWKVKETRYPEETR V - - A 234
Mus_musculus_Nth         182 AILQQRYESDIPASVAELVA -LPGVGPKMAHLAMAVAWGTISGIAVDTHVHRIINRLRWTKKMTKTPEETR K - - N 253
Bos_taurus_Nth           187 AILQQRYESDIPASVAELVA -LPGVGPKMAHLAMAVAWGTISGIAVDTHVHRIINRLRWTKKMTKTPEETR R - - A 258
Homo_sapiens_Nth         194 AILQQRYESDIPASVAELVA -LPGVGPKMAHLAMAVAWGTISGIAVDTHVHRIINRLRWTKKMTKTPEETR A - - A 265
Takifugu_rubripes_MutY    350 QKVVSOLQEMPTVDAALLKQLPGVGRYTAGAIGSIALGQVTGA -VDGNVIRVLCRLCIGADTGTPTVEALWS 423
Gallus_gallus_MutY       229 RKVVSLELAGMPRTAEDLQRLPGVGRYTAGAIGSISFGQATGV -VDGNVIRVLCRLCIGADTSLAVIDCLW 302
Xenopus_(Silurana)_tropicalis_MutY 146 KKVVELLGGMPRSADLQKLPGVGRYTAGAIGSISYQVGTGV -VDGNVIRVLCRLCIGADTSLAVSDKLWN 219
Mus_musculus_MutY        162 RKVVEELGGHMPRTAETLQQLPGVGRYTAGAIGSIAFDQVGTGV -VDGNVIRVLCRLCIGADTSLVSHHLWN 235
Bos_taurus_MutY          165 RKVVEELGGHMPRTAETLQQLPGVGRYTAGAIGSIAFGQAAGV -VDGNVIRVLCRLCIGADTSLVSHHLWN 238
Homo_sapiens_MutY        188 RKVVEELGGHMPRTAETLQQLPGVGRYTAGAIGSIAFGQAAGV -VDGNVIRVLCRLCIGADTSLVSHHLWN 261

Takifugu_rubripes_Nth      279 L -EEWLPRELWSEINWLLVGFQQQVCLVSPCLSVCLNQHDGPSAHKN - - - - - 325
Xenopus_(Silurana)_tropicalis_Nth 251 M -EDWMPRELWSEINWLLVGFQQQVCLVSPCLSVCLNQHDGPSAHKN - - - - - 297
Gallus_gallus_Nth        235 L -EDWLPRLDLWREINWLLVGFQQQVCLVSPCLSVCLNQHDGPSAHKN - - - - - 280
Mus_musculus_Nth         254 L -EEWLPRLVLEWSEINWLLVGFQQQVCLVSPCLSVCLNQHDGPSAHKN - - - - - 299
Bos_taurus_Nth           259 L -EEWLPRELWSEINWLLVGFQQQVCLVSPCLSVCLNQHDGPSAHKN - - - - - 304
Homo_sapiens_Nth         266 L -EEWLPRELWSEINWLLVGFQQQVCLVSPCLSVCLNQHDGPSAHKN - - - - - 311
Takifugu_rubripes_MutY    424 LANTLVDPDRPGDFNQALMELGATVCTPKKPLCTAPLQGGCKAYLKVIAEKESAVKTLIKKQASP 488
Gallus_gallus_MutY       303 MANTLVDRSRPGDFNQALMELGATVCTPKKPLCTAPLQGGCKAYLKVIAEKESAVKTLIKKQASP 365
Xenopus_(Silurana)_tropicalis_MutY 220 LANTLVDPDRPGDFNQALMELGATVCTPKKPLCTAPLQGGCKAYLKVIAEKESAVKTLIKKQASP 285
Mus_musculus_MutY        236 LAQQLVDPARPGDFNQAAAMELGATVCTPKKPLCTAPLQGGCKAYLKVIAEKESAVKTLIKKQASP 293
Bos_taurus_MutY          239 LAQQLVDPARPGDFNQAAAMELGATVCTPKKPLCTAPLQGGCKAYLKVIAEKESAVKTLIKKQASP 299
Homo_sapiens_MutY        262 LAQQLVDPARPGDFNQAAAMELGATVCTPKKPLCTAPLQGGCKAYLKVIAEKESAVKTLIKKQASP 322

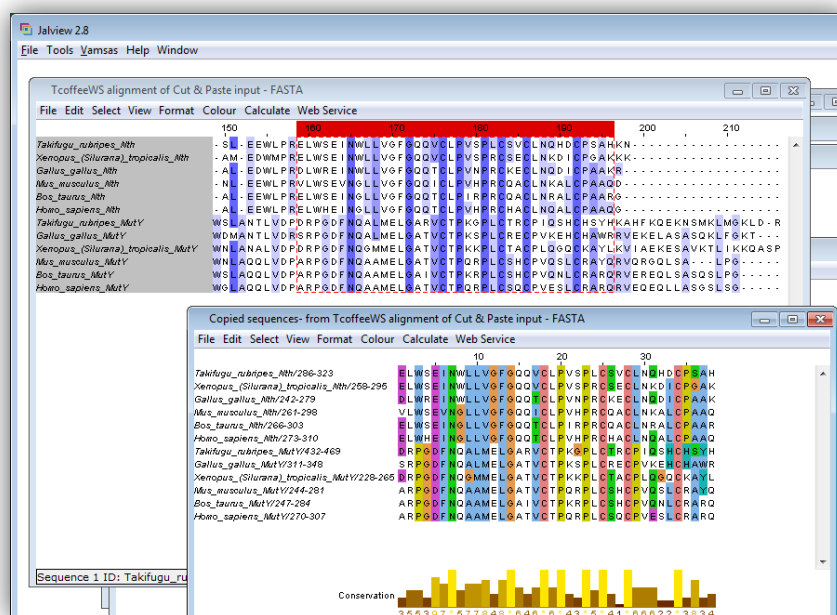
```

HHH motif

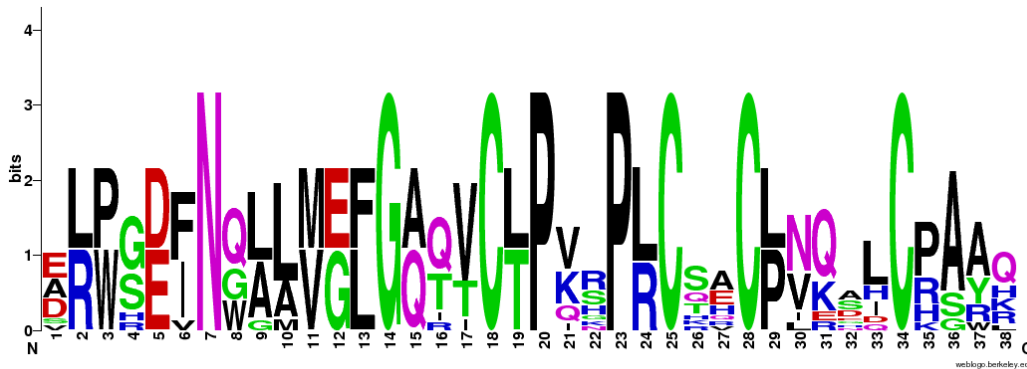
[4Fe-4S] cluster motif

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** of *E. coli* Nth as query, perform an iterative protein PSI-BLAST search against the NCBI Reference protein sequence database (Refseq protein). Limit the search to mammalian sequences, set the maximum 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 ..... MCSFOESOMTA..... SARMLT..... RSRSLGPGAGPRGCREEPGLRR..... REAAAEARKSHSPV55
MUTYH_Homo_sapiens 1 ..... MTLVLSRLSRWAIMRK..... PRAAVGSGH..... RK..... QAAASQEGROKHAKN42
OGG1_Homo_sapiens 1 ..... MPARALLPRRMGHRRLASTPALWASIPCP..... PRSELRLDL..... VLPSSGQSRW..... REQSPAHSQVLAD61
MBD4_Homo_sapiens 1 MGTTGLESLSLGDGGAARTVTSSERLVDPDPNDLRKEDVAMELSEVGEDEEQMMIKRSSECNPLLQEP IASAQFGATAGTEC..... KSVSPGGERVVKQ96

NTHL1_Homo_sapiens 56 K...RPRKAQRLRVAYEGSDSEKGEGAELKVPV..... WEPODWOQQLVNI RAMRNK..... KDAPVDHLGTEHCYDSSAPPKV..... RRYQ131
MUTYH_Homo_sapiens 43 NSQAKPSACDGMIAECGAPAGLAROEVEVL..... QASVSSYHLFRDVAEVTAFRGS..... LLSWYDQEKRDLPWRRRAEDEMOLDRRRAYA126
OGG1_Homo_sapiens 62 QVWTLTQTTEQLHCTVYRGDKSQASRPDPDEL..... EAVRKYFQLDVTLAQLYHHWGS..... VDSHFQEVAAKFQGVRLLRQDP..... IE139
MBD4_Homo_sapiens 97 RLFGKTAGRFDVYFISPOGLKFRSKSSLANYLHKNGETSLKPEDFDFTVLSKRGIKSRYKDCSMAALTSHLQNGSNNSNWNLRTRSKC..... KKD187

NTHL1_Homo_sapiens 132 VLLSLMLSSQTKDQVTAAMORLRA..... RGLTVDSILQTDATLGLIYP..... VGFWRSKVKYIKQTSAILQQHYGGDI PASV..... 208
MUTYH_Homo_sapiens 127 VVWSEVMLQQTQVATVINYYTGWMQ..... KWPTLQDLASASLEEVLNQLWAG..... LGYY..... SRGRRLQEGARKVVEELGGHMRPTAE..... 203
OGG1_Homo_sapiens 140 CLFSFICSSNNNIARITGMVERLCOAFGPRLIQLDDVTYHGFPSLQALAGREVEAHLRKLGLGY..... RARYVSASARAILLEEQGLAWLQQLRESS232
MBD4_Homo_sapiens 188 VMFPSSSSELQESRGLSNFTSTHL..... LLKEDEGVDDVNFVRKVRKPKGK..... VTIL..... KGIPIKTKKGGCRKSCS..... FVQSDSKRESV267

NTHL1_Homo_sapiens 209 ..... AELVALPGVGPKMAHMAVAVGTV..... SGIAVDTHVHRIANR..... LRWTKKATKSPEETRAALEEW..... LPRELW..... HEINOLL..... V6284
MUTYH_Homo_sapiens 204 ..... TLQQLLPQVGRYTAGAIASIAFGQA..... TGV..... VDGNAVRLCR..... VRAIGADPSSTLVSQQLW..... GLAQQLV..... DPARPQDFNQAA..... ME281
OGG1_Homo_sapiens 233 YEEAHKALCILPGVGTKVADICLMLADKP..... QAVPVDVHMWHIAQRDYSWHPTTSSQAKGPSPTQNKELGNFFRSLW..... GPY..... AG312
MBD4_Homo_sapiens 268 CNKA..... DAESPEVAQKSQDLRTVGISDAGACGETLS..... TSEENS LVKK..... KERSLSGGSNFCSEQKTS..... INKFSKAKDSEHNKEDTFFLESESE357

NTHL1_Homo_sapiens 285 FGQQTCLPVRH..... PRCHACLNQALCPAAQGL..... LQPCAGPRGCREEPGLRR..... REAAAEARKSHSPV55
MUTYH_Homo_sapiens 282 LGATVCTPOR..... PLCSQCPVESLQRARORVEEQQLLASGSLSGSPDVEECAPNTGQCHLCLPPEPWDTLQGVNFRPKASRKPREESS309
OGG1_Homo_sapiens 313 WAQAVLFSAD..... LRSRHAQEPAPAKRRKSKGPEG..... YAGWAQAVLFSAD..... 345
MBD4_Homo_sapiens 358 IGTKEVVERKEHLHTDILKRGSEMDNNSCPTRKDFTGKIFQEDTIPRT..... QIERRTSLYFSSKYNKEALSPPRRKA..... KKWTPPRS..... FNLVQETLFHDPW457

NTHL1_Homo_sapiens 370 ATCVLEQPGALGAQILLVQRPNSGLLAGLWEPFSPVTWEPSEQLQRKALLQELQRWAGPLPATHLRHLGEEVVHTFSHIKLTYYQVYGLALEGQTPVTT465
MUTYH_Homo_sapiens 370 ATCVLEQPGALGAQILLVQRPNSGLLAGLWEPFSPVTWEPSEQLQRKALLQELQRWAGPLPATHLRHLGEEVVHTFSHIKLTYYQVYGLALEGQTPVTT465
OGG1_Homo_sapiens 450 ETLFHDWPWKLIIATIFLNRTSGKMAIPVLWKFI..... LEKYPSEAVARTADWRDVSSELLKPLGLYDLRAKTI VKFS..... 520
MBD4_Homo_sapiens 450 ETLFHDWPWKLIIATIFLNRTSGKMAIPVLWKFI..... LEKYPSEAVARTADWRDVSSELLKPLGLYDLRAKTI VKFS..... 520

NTHL1_Homo_sapiens 466 VPPGARWLQTQEEFHTAAVSTAMKKVFRVYQGGQPGTCMGSKRSQVSSPCSRKKPRMGQQQLDNFFFRSHISTDAHSLNSAAQ..... 546
MUTYH_Homo_sapiens 466 VPPGARWLQTQEEFHTAAVSTAMKKVFRVYQGGQPGTCMGSKRSQVSSPCSRKKPRMGQQQLDNFFFRSHISTDAHSLNSAAQ..... 546
OGG1_Homo_sapiens 521 ..... DEYLTKQWKYPIELHGIGKYGNDSYRIFCVNEWQVHPEDHKLNKYHDWLWENHEKLSLS..... 580
MBD4_Homo_sapiens 521 ..... DEYLTKQWKYPIELHGIGKYGNDSYRIFCVNEWQVHPEDHKLNKYHDWLWENHEKLSLS..... 580

```

**MAFFT:**

```

NTHL1_Homo_sapiens 1 MCSFOESOMTA..... SARMLTRSR..... LQPCAGPRGCREEPGLRR..... REAAAEARKSHSPV55
MUTYH_Homo_sapiens 1 MTLVLSRLSRWAIMRK..... PRAAVGSGH..... RK..... QAAASQEGROKHAKN42
OGG1_Homo_sapiens 1 MPARALLPRRMGHRRLASTPALWASIPCP..... PRSELRLDL..... VLPSSGQSRW..... REQSPAHSQVLADQVWTLTQTTEQLHCTVYRGDKSQASRPDP91
MBD4_Homo_sapiens 1 MGTTGLESLSLGDGGAARTVTSSERLVDPDPNDLRKEDVAMELSEVGEDEEQMMIKRSSECNPLLQEP IASAQFGATAGTEC..... KSVSPGGERVVKQ96

NTHL1_Homo_sapiens 70 SDSE..... KGEGAELKVPVWEPODWO..... QQL..... VNI RAMR..... NKDAP..... VDLHGTEHCYDSSAPPKVRRYQVLLSLML138
MUTYH_Homo_sapiens 63 GMIAECPGAPAGLAROEVEVLQASVSS..... YHLFRDVAEVTAFRGSLLSWYDQEKRDLPWRRRAEDEMOLDR..... RAYAVWSEVML133
OGG1_Homo_sapiens 92 ELEAV..... RKYFQLDVTLAQLYHHWGSVDSHFQEVAAKFQGVRLRL..... QDPI..... ECLFIFIC146
MBD4_Homo_sapiens 47 EDEFE..... LQPCAGPRGCREEPGLRR..... REAAAEARKSHSPV55

NTHL1_Homo_sapiens 139 SSQTKDQV..... AGAMQRLRARGLTVDSILQTDATLGLIYP..... VGFWR..... RSKVKYIKQTSAILQQHYGGDI PASV..... 204
MUTYH_Homo_sapiens 134 LQQT..... QVATVINYYTGWMQKWP..... TLQDLASASLEEVLNQLWAG..... LGYY..... SRGRRLQEGARKVVEELGGHMRPTAE..... 203
OGG1_Homo_sapiens 147 SSNN..... NIARITGMVERLCOAFGPRLIQLDDVTYHGFPSLQALAGREVEAHLRKLGLGY..... RARYVSASARAILLEEQGLAWLQQLRESS232
MBD4_Homo_sapiens 95 KQRLFQKTA..... GRFDVYFISPOGLKFRSKSSLANYLHKNGETSLKPEDFDFTVLSKRGIKSRYKDCSMAALTSHLQNGSNNSNWNLRTRSKCK185

NTHL1_Homo_sapiens 205 ..... PASVAELVA..... LPVGPK..... MAHMAVAVGTVSGIAVDTHVHRIANR..... 248
MUTYH_Homo_sapiens 199 ..... PRTAETLQQL..... TAGAIASIAFGQATGV..... VDGNAVRLCR..... VRAIGADPSSTLVSQQLW..... GLAQQLV..... DPARPQDFNQAA..... ME281
OGG1_Homo_sapiens 222 ..... LAWLQQLRESS..... YEEAHKALCILPGVGTKVADICLMLADKP..... QAVPVDVHMWHIAQRDYSWHPTTSSQAKGPSPTQNKELGNFFRSLW..... GPY..... AG312
MBD4_Homo_sapiens 186 KDVFMPSSESSELQESRGLSNFTSTHLLLKEDEGVDDVNFVRKVRKPKGKVTILKGIPIKTKKGGCRKSCS..... FVQSDSKRESV267

NTHL1_Homo_sapiens 249 LRWTKKATKSPEETRAALEEW..... LPRELW..... EINHOLLVGFQQTCLPVRHPRCHACLNQALCPAAQ..... 310
MUTYH_Homo_sapiens 243 VR..... AIGADPSSTLVSQQLWGLAQQLVDPARPGDNQAAMELGATVDTQRLQSCCPVESLQRARO..... 307
OGG1_Homo_sapiens 283 PTTSSQAKGPSPTQNK..... ELGNFFRSLWGP..... YAGWAQAVLFSAD..... 345
MBD4_Homo_sapiens 287 TVCISDAGACGETLSVTSEENS LVKK..... KERSLSGGSNFCSEQKTS..... INKFSKAKDSEHNKEDTFFLESEIEGTVEVVERKEHLHTD374

NTHL1_Homo_sapiens 311 ..... SSQTKDQV..... AGAMQRLRARGLTVDSILQTDATLGLIYP..... VGFWR..... RSKVKYIKQTSAILQQHYGGDI PASV..... 204
MUTYH_Homo_sapiens 378 GALGAQILLVQRPNSGLLAGLWEPFSPVTWEPSEQLQRKALLQELQRWAGPLPATHLRHLGEEVVHTFSHIKLTYYQVYGLALEGQTPVTTVPPGARWLQTQEEF478
OGG1_Homo_sapiens 378 GALGAQILLVQRPNSGLLAGLWEPFSPVTWEPSEQLQRKALLQELQRWAGPLPATHLRHLGEEVVHTFSHIKLTYYQVYGLALEGQTPVTTVPPGARWLQTQEEF478
MBD4_Homo_sapiens 458 KLLIATIFLNRTSGKMAIPVLWKFI..... LEKYPSEAVARTADWRDVSSELLKPLGLYDLRAKTI VKFSDEYL..... 520

NTHL1_Homo_sapiens 479 HTAAVSTAMKKVFRVYQGGQPGTCMGSKRSQVSSPCSRKKPRMGQQQLDNFFFRSHISTDAHSLNSAAQ..... 546
MUTYH_Homo_sapiens 479 HTAAVSTAMKKVFRVYQGGQPGTCMGSKRSQVSSPCSRKKPRMGQQQLDNFFFRSHISTDAHSLNSAAQ..... 546
OGG1_Homo_sapiens 525 ..... LRSRHAQEPAPAKRRKSKGPEG..... YAGWAQAVLFSAD..... 345
MBD4_Homo_sapiens 525 ..... TKQWKYPIELHGIGKYGNDSYRIFCVNEWQVHPEDHKLNKYHDWLWENHEKLSLS..... 580

```

**CLUSTAL:**

```

NTHL1_Homo_sapiens 1 .....MCS PQESGMTALSARMLTRSRSLGPGAGPRGCREEPGLRRREAAAEARKSHS VVKRPRK.....AQLRVA 67
MUTYH_Homo_sapiens 1 .....MTPLVSRSLRSLWAIMRKPRAAVGS.....GHRKQASQEGRKHKAKNNSQAKSACDGM.....IAECPGA 61
OGG1_Homo_sapiens 1 .....MPARALLPRRMGHRTLASTPALWASIPCPRSELRLD.....LVLP SGQSFWRREQSPAHSGLADQVWTL.....TQTEEQ 73
MBD4_Homo_sapiens 1 MGTTLGLESLSLGDGRGAAPTIVTSSERLVPDPNDLRKEDVAMELERVGEDEEQMMIKRSSECNPLLQETIASAQFGATAGTECRKSV 86

NTHL1_Homo_sapiens 68 YEGSDS- EKGEAEPLKVPVWEPQD.....WQQQLVNI R-AMRNKKDAPVDHLGTEHCYDSSAPPKVRYQVLLSLMLSSQTKDQV 146
MUTYH_Homo_sapiens 62 PAGLAR-QPEEVVLQASVSSYHLFR.....DVAEVTAFRGSLLSWYDQEKRDLPWRRRAEDEMDLDRAYAVWVSEVMLQQTQVAT 141
OGG1_Homo_sapiens 74 HCTVYRGDKSQASRPTPDELEAVRK.....YFQLDVTLAQLYHHWGSVDSHFQEVAKFGQVRLRQDPICELFSFICSSNNNIAR 154
MBD4_Homo_sapiens 87 PCGWERVVVKQLRFGKTAGRFDVYFI SPQGLKFRSKSSLANYLHKNGETSLKPEDFDFTVLSKRGIKSRKYKDCSMAALTSHLQNGSN 172

NTHL1_Homo_sapiens 147 TAGAMQRLRARG.....LTVDLSILQTDATL6KLIYPVGFWRSKVKYIKQTSAILQQHYGG.....DIPASVAEL 211
MUTYH_Homo_sapiens 142 VINYYTGWMQKW.....PTLQDLASASLEEVLNQLWAGLGYYSRGRRLQEGARKVVEELGG.....HMPRTAETL 205
OGG1_Homo_sapiens 155 ITGMVERLCCAFGPRLIQLDDVITYHGFPSSLOALAGPEVEAHLRLGLG-YRARYVSASARAILEEQGLAWLQQLRESSYEEAHKA 239
MBD4_Homo_sapiens 173 NSNWNLRTRSKCK.....KDVFMPPSSSELQESRGLSNFTSTHLLKKEDEGVDDVNF RKVRKPKGKVITLKGIPIKKTKKGRK 252

NTHL1_Homo_sapiens 212 VALP- GVGPKMAHLAMAVWGTVSQIAVDTHVRIANRLRWTSKATKSPEETRAALEEWLPRELWHEIN-GLLVGFGQQTGL 291
MUTYH_Homo_sapiens 206 QQLLPQVGRYTAGAIASIAFGQATG-VVDGNVARVLCRVRAIGADPSSSTLVSSQLWGLAQQLVDPARPQDFN-QAAMELGATVCT 288
OGG1_Homo_sapiens 240 LCILPQVGTKVADCICLMALDKPQAVPVDVHMWHIAQRDYSWHPTTSQAKGSPQTNKELGNFFRSLWGPYAGWAQAVLFSADLRQ 325
MBD4_Homo_sapiens 253 SCSGFVQSDSKRESVGNKDAESEPVAKKSQLDRTVCISDAGACGETLSVTSEENSLVKKKERSLSSGSGNFCSEQKTSGLINKFSS 338

NTHL1_Homo_sapiens 292 PVHFRCHACLNQALCPAAQGL.....PQRF LCSQCPVESLCRARQRVQEQLLASGSLSGSPDVEECAPNTGQCHLCLPPSEPWDQTLGVVNFPRKASRKPPREESATCVL 312
MUTYH_Homo_sapiens 289 PQRF LCSQCPVESLCRARQRVQEQLLASGSLSGSPDVEECAPNTGQCHLCLPPSEPWDQTLGVVNFPRKASRKPPREESATCVL 374
OGG1_Homo_sapiens 326 SRHAQEPAPAKRRKGSKGPEG.....AKDSEHNEKYEDTFLESEEIGTKVEVVERKEHLTDILKRGSEMDNNCSPTRKDFTGKIFQEDTIPRTQIERRKTSLYFSKYNK 424
MBD4_Homo_sapiens 339 AKDSEHNEKYEDTFLESEEIGTKVEVVERKEHLTDILKRGSEMDNNCSPTRKDFTGKIFQEDTIPRTQIERRKTSLYFSKYNK 424

NTHL1_Homo_sapiens .....EOPGALGAQILLVQRFNSSGLLAGLWEFPVSVTWEPSEQLQKALLQELORWAGPLPATHRLHLGEVVFTHFSHIKLTQVYVGLALEGQ 460
MUTYH_Homo_sapiens 375 EOPGALGAQILLVQRFNSSGLLAGLWEFPVSVTWEPSEQLQKALLQELORWAGPLPATHRLHLGEVVFTHFSHIKLTQVYVGLALEGQ 460
OGG1_Homo_sapiens .....EALSPRRKAFKKWTPRSPFN.....LVQETLFHDPWKLLIATIFLNRTSGKMAIPVLWKFLKYPYSAEVARTADWRDVSLELKP 505
MBD4_Homo_sapiens 425 EALSPRRKAFKKWTPRSPFN.....LVQETLFHDPWKLLIATIFLNRTSGKMAIPVLWKFLKYPYSAEVARTADWRDVSLELKP 505

NTHL1_Homo_sapiens .....TPVTTPVPGARWLTQEEFHTAAVSTAMKKVFRVYQSQQPCTCMGSKRSQVSSPCSRKPRMGQQVLDNFFRSHISTDAHSLNSAAQ 546
MUTYH_Homo_sapiens 461 TPVTTPVPGARWLTQEEFHTAAVSTAMKKVFRVYQSQQPCTCMGSKRSQVSSPCSRKPRMGQQVLDNFFRSHISTDAHSLNSAAQ 546
OGG1_Homo_sapiens .....LGLYDLRAKTI VKFSD EYLTQKWYPIELHGIGKYGNDSYRIFCVNEWKQVHREDHKLNKYHDWLWENHEKLSLS..... 580
MBD4_Homo_sapiens 506 LGLYDLRAKTI VKFSD EYLTQKWYPIELHGIGKYGNDSYRIFCVNEWKQVHREDHKLNKYHDWLWENHEKLSLS..... 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 seems 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_anthraxis_NP_844020      31 FELVIAVALSAGCTDALVNKVTKNLFQK.....YKTPEDYLSVSLLEELQDDRSILYRNKAKNIQKLCRML198
Bos_taurus_MutY                102 YAVVVAEVMLOQTQVATVINYTRWMQK.....WPTLQDLASASLEEVNQLWAGLGYYSRGRWLQEGARKVY198
Bos_taurus_Nth                 123 YQVLLSLMLSSQTKDOVTAGAMORLRRAR.....GLTVDSILQTDGTLGALIPVGFWRSKVKYIKOTSAILO190
Escherichia_coli_NP_416150     30 FELLIAVLLSAGATDVSNKATAKLYPV.....ANTPAAMLELGVGVKTYIKTIGLYNSKAENIKTCTRILL97
Gallus_gallus_MutY            166 LAVVWSEIMLQQTQVATVIDYNNRWMQK.....WPTLQALAAASLEEVNQLWAGLGYYSRGRWLQEGARKVY232
Gallus_gallus_Nth             99 YQVLLSLMLSSQTKDOVTSAAMLRLRRAR.....GLTVDSILQMDATLGGIIPVGFWRNKVKYIKOTTAILO168
Homo_sapiens_MutY             125 YAVVWSEVMLOQTQVATVINYTGWMQK.....WPTLQDLASASLEEVNQLWAGLGYYSRGRWLQEGARKVY191
Homo_sapiens_Nth             130 YQVLLSLMLSSQTKDOVTAGAMORLRRAR.....GLTVDSILQTDGTLGKLIYPVGFWRSKVKYIKOTSAILO197
MBD4_Homo_sapiens             156 KDCSMAALTSHLQNSNNSNWNLRTRSK.....CKKDVFMPSSSSSELSQESRGLSNF.....TSTHLL1214
Mus_musculus_MutY            99 YAVVWSEVMLOQTQVATVIDYNNRWMQK.....WPKLQDLASASLEEVNQLWAGLGYYSRGRWLQEGARKVY165
Mus_musculus_Nth             118 YQVLLSLMLSSQTKDOVTAGAMORLRRAR.....GLTVESILQTDGTLGRLIPVGFWRNKVKYIKOTTAILO195
Mycobacterium_tuberculosis_NP_218191 41 LELAVATILSAGSTDKRVNLTTPALFAR.....YRTARDYAQADRTLESIRPTGFYRNKAASLIGLGGALV108
Neisseria_meningitidis_NP_273578 30 FELLIAVLLSAGATDVGNKATAKLPV.....ADTPOAMLDLGLDGVMEYTKTIGLYKTSKHHIMQCTRILL97
OGG1_Homo_sapiens            138 IECILSFICSSNNNIARITGMVERLQCAFGRPLIQLDDVTYHGFPSLOALAGPEVEAHLRKLGLG-YRARYVSASARAIL216
Streptococcus_pneumoniae_NP_358750 32 FELLVAVMLSAGTTDAAVNKATPGLFVA.....FPTPOAMSVATESEIASHSRLGLYRNKAKFLKKCAQOOL99
Takifugu_rubripes_MutY        287 YAVVWSEIMLQQTQVATVIDYNNKWMK.....WPTVQDLATATLEDVNMWAGLGYYSRGRWLQEGARKVY353
Takifugu_rubripes_Nth        143 FOVLVSLMLSSQTKDOVTSAAAMKLRAR.....GCTVENILATNDETGLQLIYPVGFWRNKVKYIKLTSAKMLQ210
Xenopus_(Silurana)_tropicalis_MutY 83 YAVVWSEVMLOQTQVATVIDYNNKWMK.....WPTMEDLARSSLEEVNEMWAGLGYYSRGRWLQEGARKVY140
Xenopus_(Silurana)_tropicalis_Nth 115 YQVLLSLMLSSQTKDOVTSAAAMRLRQH.....GLTVSRILETDGTLGKLIYPVGFWRNKVKYIKOTTEILO182

Bacillus_anthraxis_NP_844020      99 DDYNG.....EVFKDRDEITKLP.....GVRKRTANVVVSVAFGIP-AIAVDTHVERVSKR....LAICRWKD156
Bos_taurus_MutY                169 EELGG.....HMPRTAETLQQLLP.....GVRGYTAGAIASIAFGQAAGV.VDGNVIRVLCR....VRAIGADS227
Bos_taurus_Nth                 191 QRYDG.....DIPASVAELVA-LP.....GVRPKMAHLAMAVAWGTISGIAVDTHVHRIANR....LRWTKKAT249
Escherichia_coli_NP_416150     98 EQHNG.....EVEDRAAEAE-LP.....GVRKRTANVVLNTAFGWP-TIAVDTHIFRVGNR....TQFAPGKN155
Gallus_gallus_MutY            233 SELAG.....RMPRTAEDLQRLLP.....GVRGYTAGAIASISFGQATGV.VDGNVIRVLCR....LRCIGADT201
Gallus_gallus_Nth             167 QKYGG.....DIPGTVEELVVK-LP.....GVRPKMAHLAMNIAWNSVSGIAVDTHVHRIANR....LKWVKKET225
Homo_sapiens_MutY             102 EELGG.....HMPRTAETLQQLLP.....GVRGYTAGAIASIAFGQAAGV.VDGNVIRVLCR....VRAIGADP250
Homo_sapiens_Nth             108 QHYGG.....DIPASVAELVA-LP.....GVRPKMAHLAMAVAWGTISGIAVDTHVHRIANR....LRWTKKAT256
MBD4_Homo_sapiens             215 KEDES-VDDVNFKRVRKPKGKVTILKGLPIFKKTKKCKRKSQSGFVQSDSKRESVCNKADAESEPVAAQSKQLDRVTCISDA293
Mus_musculus_MutY            166 EELGG.....HMPRTAETLQQLLP.....GVRGYTAGAIASIAFGQVTVGV.VDGNVIRVLCR....VRAIGADP224
Mus_musculus_Nth             186 QRYEG.....DIPASVAELVA-LP.....GVRPKMAHLAMAVAWGTISGIAVDTHVHRIANR....LRWTKKMT244
Mycobacterium_tuberculosis_NP_218191 109 ERFGG.....EVPATMDKLVLT-LP.....GVRKRTANVILGNAGFIP-GITVDTHFORLVRR....WRWTTAED166
Neisseria_meningitidis_NP_273578 98 EKYNG.....EVEDREALLES-LP.....GVRKRTANVVLNTAFGHP-VMAVDTHIFRVSNR....TKIAPGKD155
OGG1_Homo_sapiens            217 EEQGG-LAWLQQLRESSYEEAHKALCI-LP.....GVTKVADCICLMALDKPAQAVPDVHMWHIAQR....DYSWHPPTS286
Streptococcus_pneumoniae_NP_358750 100 DFDG.....QVPTREELES-LA.....GVRKRTANVVMSVGFIP-AFAVDTHVERICKH....HDIVKKA157
Takifugu_rubripes_MutY        354 SLOLG.....EMPTVDALLKQLP.....GVRGYTAGAIGSIALGQVTGA.VDGNVIRVLCR....LRAIGADC412
Takifugu_rubripes_Nth        211 KEFGG.....DIPDSVEGLVR-LP.....GVRPKMAHLAMDIAWDQVSGIGVDTHVHRIANR....LGWLKKPT269
Xenopus_(Silurana)_tropicalis_MutY 150 LELGG.....SMPSADELQKLLP.....GVRGYTAGAIASISYQVTVGV.VDGNVIRVLCR....LRCIGADS208
Xenopus_(Silurana)_tropicalis_Nth 183 EKYGG.....DIPDNVTDLVK-LP.....GVRPKMAHLVMDIAWNNVSGIGVDTHVHRIANR....LKWVRKET241

Bacillus_anthraxis_NP_844020      157 SVLEVEKT...LMKKIPMDEWVSVTHHR-MIFFGRYHCKAQRQEEPLLEVOR...206
Bos_taurus_MutY                228 SSTLVSQLWLSAQQLVDPARPGDFNOA-AMELGAIVCTPKRPLCSHPVQNLQARQR...285
Bos_taurus_Nth                 250 KSPEETRAA...LEEWLPRELWSEINGL-LVGFQQQTCLRIRPQQAQLNRALQ...299
Escherichia_coli_NP_416150     156 VEQVEEK...LLKVPAEFKVDCHHW-LILHGRYTICIAKPRDGSQIIEDLQ...203
Gallus_gallus_MutY            292 SSLAVIDCLWDMANTLVDRSRPGDFNOA-LMELGATVCTPKRPLCRECPVKEHSHAWRR...349
Gallus_gallus_Nth             226 RYPEETRAVA...LEDWLPRLWREINWL-LVGFQQQTCLRVNPRCKELNODIQ...275
Homo_sapiens_MutY             251 SSTLVSQLWGLAQQLVDPARPGDFNOA-AMELGAIVCTPKRPLCSQCPVESLQARQR...308
Homo_sapiens_Nth             257 KSPEETRAA...LEEWLPRELWSEINGL-LVGFQQQTCLRVHRRHALLNQALQ...306
MBD4_Homo_sapiens             294 GACGETLSVTSEENSLVKKKER...SLSSQSNFCSEQK...TSGLINKFCSAKDS...342
Mus_musculus_MutY            225 SSTLVSHHLWNLAQQLVDPARPGDFNOA-AMELGAIVCTPKRPLCSHPVQSLQRAYQR...282
Mus_musculus_Nth             245 KTPETTRKN...LEEWLPRLVWSEVNGL-LVGFQQQICLRVHRRQQAQLNKALQ...294
Mycobacterium_tuberculosis_NP_218191 167 PVKVEQA...VGELIERKEWTLLSHR-VIFHRRVCHARRADGVCLAKDPSFG...218
Neisseria_meningitidis_NP_273578 156 VREVEDK...LMRFIPKEFLMDAHHW-LILHGRYTCKALKRQQTGIIINDLQ...203
OGG1_Homo_sapiens            287 QAKGPSPTNKELGNFF-RSLWGPYAGWAQAVLFSADLRQSRHAQEPAPAKRRKGS...340
Streptococcus_pneumoniae_NP_358750 158 TPLEVEKR...VMDILPPEQWLAAHQ-MIYFGRALCHFKNFEEDQYP...QLY...204
Takifugu_rubripes_MutY        413 TGPTVTEALWSLANTLVDPDRPGDFNOA-LMELGARVCTPKRPLCTRPPIQSHHSYHK...470
Takifugu_rubripes_Nth        270 KTPETTRKS...LEEWLPRELWSEINWL-LVGFQQQVCLRVSLVCLNQHDPSAHK...324
Xenopus_(Silurana)_tropicalis_MutY 209 STLAVSDKLWNLANALVDPDRPGDFNOG-MMELGATVCTPKRPLCTACPLQGGKAYLK...286
Xenopus_(Silurana)_tropicalis_Nth 242 KTPETTRVA...MEDWMPRELWSEINWL-LVGFQQQVCLRVSEELNKDQ...291

```

## MAFFT:

```

Bacillus_anthraxis_NP_844020      31 FELVIAVALSADCTDALVNKVTKNLFK.....YKPEDYLSVSLEELQQDIRSIGLYRNKA 87
Bos_taurus_MutY                 102 YAVVVAEVMLOQTQVATVINYYTRWMQK.....WPLQDLASASLEEVNQWAGLGYY.SRG 157
Bos_taurus_Nth                  123 YQVLLSLMLSSQTKDQVDTAGAMQRLRAR.....GLTVDSILQTDSD.TLGLIYIPVGFWRNKA 179
Escherichia_coli_NP_416150      30 FELLIAVLLSADATDVSVNKATAKLYPV.....ANTPAAMLELGVEGVKTYIKTIGLYNSKA 86
Gallus_gallus_MutY             166 LAVVWSEIMLOQTQVATVIDYNNRWMQK.....WPLQALAAASLEEVNELWAGLGYY.SRG 221
Gallus_gallus_Nth              99 YQVLLSLMLSSQTKDQVTSAAMLRLRQR.....GLTVDSILQMDDA.TLGLIYIPVGFWRNKA 155
Homo_sapiens_MutY              125 YAVVWSEVMLOQTQVATVINYYTGWMMQK.....WPLQDLASASLEEVNQWAGLGYY.SRG 180
Homo_sapiens_Nth               130 YQVLLSLMLSSQTKDQVDTAGAMQRLRAR.....GLTVDSILQTDSD.TLGLIYIPVGFWRNKA 186
MBD4_Homo_sapiens              457 WKLLIATIFLNRITSGKMAIPVLWKFLEK.....YPSAEVARTADWR.DVSELKPLGLYDLRA 513
Mus_musculus_MutY              99 YAVVWSEVMLOQTQVATVIDYNNRWMQK.....WPKLQDLASASLEEVNQWAGLGYY.SRG 154
Mus_musculus_Nth              118 YQVLLSLMLSSQTKDQVDTAGAMQRLRAR.....GLTVDSILQTDSD.TLGLIYIPVGFWRNKA 174
Mycobacterium_tuberculosis_NP_218191 41 LELAVATILSADSTDKRVNLTTPALFAR.....YRTARDYAQADRT.ELESIRPTGTYRNKA 97
Neisseria_meningitidis_NP_273578 30 FELLIAVLLSADATDVGVNKAATAKLFV.....ADTPQAMLDGLD.GVMEYTKTIGLYKTSK 86
OGG1_Homo_sapiens              138 IECLFDFICSSNNNIARITGMVERLQQAQF6PRLIQLDDVITYHGFPSLQALAGPEVEAHLRL..GLGY...RA 205
Streptococcus_pneumoniae_NP_358750 32 FELLVAVMLSADTTDAAVNKATPQLFVA.....FPTPQ.AMSVATESIASHSIRLGLYRNKA 88
Takifugu_rubripes_MutY         287 YAVVWSEIMLOQTQVATVIDYNNKWMKR.....WPTVQDLATATLE.DVNQWAGLGYY.SRG 342
Takifugu_rubripes_Nth          143 FQVLLSLMLSSQTKDQVTSAAMLRLRAH.....GCTVENILATNDE.TLGLIYIPVGFWRNKA 199
Xenopus_(Silurana)_tropicalis_MutY 83 YAVVWSEVMLOQTQVATVIDYNNKWMKV.....WPTMEDLARSSLE.EVNEWWSGLGY.SRG 138
Xenopus_(Silurana)_tropicalis_Nth 115 YQILLSLMLSSQTKDQVTSAAMLRLRH.....GLTVSRILETDDG.TLGLIYIPVGFWRNKA 171

Bacillus_anthraxis_NP_844020      88 KNIQKLCRMLDDYNG.....EVPKDRDELTK.LPGVGRKTANVVSVAFG.IPAIAVDTHVERVSKR 148
Bos_taurus_MutY                158 RWLQEGARKVVEELGG.....HMPRTAETLQQF.LPGVGRYTAGAIASIAFGQAAAGV.VDGNVIRVLRCR 219
Bos_taurus_Nth                 160 KYIKQTSAILQORYDG.....DIPASVAELVA.LPGVGPMAHMAVAVAGT.VSGIAVDTHVHRIANR 241
Escherichia_coli_NP_416150      87 ENIIKTCRILLEQHNG.....EVPEDRAALEA.LPGVGRKTANVVLTAFG.WPTIAVDTHIFRVCRN 147
Gallus_gallus_MutY             222 KRLQEAARKVSELAG.....RMPRTAEDLQRL.LPGVGRYTAGAIASISFGQATGV.VDGNVIRVLRCR 283
Gallus_gallus_Nth              156 KYIKQTTAILKQKYGG.....DIPGTVEELVK.LPGVGPMAHMAVAVAGT.VSGIAVDTHVHRIANR 242
Homo_sapiens_MutY              181 RRLQEGARKVVEELGG.....HMPRTAETLQQF.LPGVGRYTAGAIASIAFGQATGV.VDGNVIRVLRCR 242
Homo_sapiens_Nth               187 KYIKQTSAILQORYDG.....DIPASVAELVA.LPGVGPMAHMAVAVAGT.VSGIAVDTHVHRIANR 248
MBD4_Homo_sapiens              614 KTIIVKFSDEYLTQW.....KYPFIE.....GNDSYRIFCVNE 552
Mus_musculus_MutY              155 RRLQEGARKVVEELGG.....HMPRTAETLQQF.LPGVGRYTAGAIASIAFGQATGV.VDGNVIRVLRCR 216
Mus_musculus_Nth              175 KYIKQTTAILQORYDG.....DIPASVAELVA.LPGVGPMAHMAVAVAGT.VSGIAVDTHVHRIANR 236
Mycobacterium_tuberculosis_NP_218191 87 KHIQTCRILLEQYNG.....EVPEDRAALEA.LPGVGRKTANVVLTAFG.HPYMAVDTHIFRVSNR 148
Neisseria_meningitidis_NP_273578 88 KHIQTCRILLEQYNG.....EVPEDRAALEA.LPGVGRKTANVVLTAFG.HPYMAVDTHIFRVSNR 148
OGG1_Homo_sapiens              208 RYVSASARAILLEGGGLAWLQQLRESSYEAHAKLCI.LPGVGRKTANVVSVAFG.IPAIAVDTHVERVSKR 277
Streptococcus_pneumoniae_NP_358750 89 KFLKKCAQQLLDDFDG.....QVPTREELES.LPGVGRKTANVVSVAFG.IPAIAVDTHVERVSKR 277
Takifugu_rubripes_MutY         343 KRLHEGAQKVVSQLOG.....EMPRITVDALLKQ.LPGVGRYTAGAIASISFGQATGV.VDGNVIRVLRCR 404
Takifugu_rubripes_Nth          200 KYLKLTSAMLOKEFGG.....DIPDSVEGLVR.LPGVGPMAHMAVAVAGT.VSGIAVDTHVHRIANR 281
Xenopus_(Silurana)_tropicalis_MutY 139 RRLQEGARKVVEELGG.....SMPSADELQKL.LPGVGRYTAGAIASISFGQATGV.VDGNVIRVLRCR 200
Xenopus_(Silurana)_tropicalis_Nth 172 KYIKQTTAILQORYDG.....DIPDNVTDLVK.LPGVGPMAHMAVAVAGT.VSGIAVDTHVHRIANR 233

Bacillus_anthraxis_NP_844020      140 LAICRWK.DSVLEVEKTL....MKKIPMDEWSVTHHRMIFGGRYHKAQRPQDEEPLLEVEGREGKRMK 213
Bos_taurus_MutY                220 VRAIGAD.SSSTLVSHLWLSLAQLVDPARPQDFNQAAMELGATVCTPKRPLCSHPVQNLCRARQVR 288
Bos_taurus_Nth                  242 LRWTKKATKSPEETRAAL....EELWPRELWSEINGLLVGFQQTCLPIRPRQQAALNRLCPAARGL.. 305
Escherichia_coli_NP_416150      148 TQFAPGK.NVEQVEEK....LKVVPAEFKVDCHHWLILHGRYTDIARKPRGSG..IIEDLCYEYKEKVD 210
Gallus_gallus_MutY             284 LRCIGAD.TSSLAVIDCLWMANTLVDRSRPQDFNQAAMELGATVCTPKRPLCSHPVQNLCRARQVR 352
Gallus_gallus_Nth              218 LKWVKKETRYPEETRVALL....EDWLPRLWREINWLVGFQQTCLPVNPRKEKLNQDIPAAKRF.. 281
Homo_sapiens_MutY              243 VRAIGAD.PSSTLVSQQLWGLAQQLVDPARPQDFNQAAMELGATVCTPKRPLCSHPVQNLCRARQVR 311
Homo_sapiens_Nth               249 LRWTKKATKSPEETRAAL....EELWPRELWSEINGLLVGFQQTCLPVNPRKEKLNQDIPAAKRF.. 312
MBD4_Homo_sapiens              553 WKQVHPEDHKLNKYHDWLWENHEKLSLS.....NKELGNFRRSLWGPYAGWAQAVLFADLRQSRHAQE 580
Mus_musculus_MutY              217 VRAIGAD.PTSTLVSHLWLSLAQLVDPARPQDFNQAAMELGATVCTPKRPLCSHPVQNLCRARQVR 285
Mus_musculus_Nth              237 LRWTKKMTKTPEETRKLL....EELWPRELWSEINGLLVGFQQTCLPVNPRKEKLNQDIPAAKRF.. 300
Mycobacterium_tuberculosis_NP_218191 159 WRWTTAE.DPVKVEQAV....GELIERKEWTLSSHRVIFHRRRVCHARRPAQGVVLAKDPSFGLPT 222
Neisseria_meningitidis_NP_273578 148 TKIAPGK.DVREVEDK....MRFIPEKLMADAHHWLILHGRYTDIARKPRGSG..IINDLCYPAKA.. 209
OGG1_Homo_sapiens              278 .....DYSWHPTTSQAKGSPQTD.....NKELGNFRRSLWGPYAGWAQAVLFADLRQSRHAQE 331
Streptococcus_pneumoniae_NP_358750 150 HDIVKKS.ATPLEVEKRV....MDLPPQWLAHQAMIFYGRAIDHPKNPEQDQVQLYDFSNL.... 209
Takifugu_rubripes_MutY         405 LRAIGAD.CTGPTVTEALWLSANTLVDPDRPQDFNQAAMELGATVCTPKRPLCSHPVQNLCRARQVR 473
Takifugu_rubripes_Nth          262 LGWLKKPTKTPEETRKSL....EELWPRELWSEINWLVGFQQTCLPVNPRKEKLNQDIPAAKRF.. 327
Xenopus_(Silurana)_tropicalis_MutY 201 LRCIGAD.SSTLAVSDKLWNLANALVDPDRPQDFNQAAMELGATVCTPKRPLCSHPVQNLCRARQVR 269
Xenopus_(Silurana)_tropicalis_Nth 234 LKWVKKETKTPEETRVAM....EDWMPRELWSEINWLVGFQQTCLPVNPRKEKLNQDIPAAKRF.. 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 horizontal 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 explanation! 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]
MNKAKRLEILTRLRENNPHPTTELNFSSPFELLIIVLLSAQATDVSVNKATAKLYPVANTPAAMLELGVE
GVKTYIKTIGLYNSKAENIIKTCRILLEQHNGEVPEDRAALEALPGVGRKTANVVLNTAFGWPTIAVDTH
IFRVCNRTQFAPGKNVEQVEEKLKVVPAEFKVDCHHWLILHGRYTCTARKPRCGSCIIEDLCEYKEKVD
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]
MPGRWSAETRLALVRRARRMNRALAQAFFPHVYCELDFTTPLELAVATILSAQSTDKRVNLTPALFARYR
TARDYAQADRTELESIRPTGFYRNKAASLIGLGQALVERFGGEVPATMDKLVTLPGVGRKTANVILGNA
FGIPGITVDTHFGRLVRRWRWTTAEDPVKVEQAVGELIERKEWTLTSHRVIFHGRRVCHARRPACGVCVL
AKDCPSFGLGPTEPLLAAPLVQGPETDHLALAGL
>gi|30261643|ref|NP_844020.1| endonuclease III [Bacillus anthracis str. Ames]
MLNKQTQIRYCLDTMADMYPEAHCELIHDNPFELVIAVALSAQCTDALVNKVTKNLFQKYKTPEDYLSVSL
EELQQDIRSIGLYRNKAKNIQKLCRMLDDYNGEVPKDRDELTKLPGVGRKTANVVVSVAFGIPAIAVDT
HVERVSKRLAICRWKDSVLEVEKTLMKKIPMDEWSVTHHRMIFFGRYHCKAQRPCCEECPLLEVCREGKK
RMKGK
>gi|15676439|ref|NP_273578.1| endonuclease III [Neisseria meningitidis MC58]
MNRHIRQEIFERFRAANPHPTTELNFNSPFELLIIVLLSAQATDVGVNKATAKLFVADTPQAMLDLGLD
GVMYTKTIGLYKTSKHIMQTCRILLEKYNGEVPEDREALESLPGVGRKTANVVLNTAFGHPVMAVDTH
IFRVSNRTKIAPGKDVREVEDKLMRFIPKEFLMDAHHWLILHGRYTCKALKPQCQTCIINDLCEYPAKA
>gi|15903200|ref|NP_358750.1| endonuclease III [Streptococcus pneumoniae R6]
MVLSKKRARKVLEEIIALFPDAKPSLDFTNHFELLVAVMLSAQTDAAVNKATPGLFVAFPTPQAMSVAT
ESEIASHISRLGLYRNKAKFLKKCAQQLLDDFDGQVPQTREELESLAGVGRKTANVMSVGFIPAFVD
THVERICKHHDIVKKSATPLEVEKRVMDILPPEQWLAHQAMIYFGRAICHKPNPECDQYPQLYDFSNL

```

**APPENDIX 2:****Bacterial Nth homologs, modified headers**

```

>Escherichia_coli_NP_416150
MNKAKRLEILTRLRENNPHPTTELNFSSPFELLIIVLLSAQATDVSVNKATAKLYPVANTPAAMLELGVE
GVKTYIKTIGLYNSKAENI IKTCRILLEQHNGEVPEDRAALEALPGVGRKTANVVLNTAFGWPTIAVDTH
IFRVCNRTQFAPGKNVEQVEEKLLKVVPAEFKVDCHHWLILHGRTYTCIARKPRCGSCIIEDLCEYKEKVD
I
>Mycobacterium_tuberculosis_NP_218191
MPGRWSAETRLALVRRARRMNRALAQAFPHVYCELDFTTPELAVATILSAQSTDKRVNLTPALFARYR
TARDYAQADRTELESIRPTGFYRNKAASLIGLQALVERFGGEVPATMDKLVTLPGVGRKTANVILGNA
FGIPGITVDTHFGRLVRRWRWTTAEDPVKVEQAVGELIERKEWTLLSHRVIFHGRRVCHARRPACGVCVL
AKDCPSFGLGPTPELLAAPLVQGPETDHLALAGL
>Bacillus_anthraxis_NP_844020
MLNKTQIRYCLDTMADMYPEAHCELIHDNPFELVIAVALSAQCTDALVNKVTKNLFQKYKTPEDYLSVSL
EELQQDIRSIGLYRNKAKNIQKLCRMLLDDYNGEVPKDRDELTKLPGVGRKTANVVVSVAFGIPAIAVD
HVERVSKRLAICRWKDSVLEVEKTLMKKIPMDEWSVTHHRMIFFGRYHCKAQR PQCEECPLLEVCREGKK
RMKGK
>Neisseria_meningitidis_NP_273578
MNRHIRQEIFERFRAANPHPTTELNFNSPFELLIIVLLSAQATDVGVNKATAKLFVADTPQAMLDLGLD
GVMEYTKTIGLYKTSKHIMQTCRILLEKYNGEVPEDREALESLPGVGRKTANVVLNTAFGHPVMAVDTH
IFRVSNRTKIAPGKDVREVEDKLMRFIPKEFLMDAHHWLILHGRTYCKALKPQCQT CIINDLCEYPAKA
>Streptococcus_pneumoniae_NP_358750
MVLSSKKRARKVLEEIIALFPDAKPSLDFTNHFELLVAVMLSAQTDDAAVNKATPGLFVAFPTPQAMSVAT
ESEIASHISRLGLYRNKAKFLKKCAQQLDDDFDGQVPQTREELES LAGVGRKTANVMSVGFIPAFVD
THVERICKHHDIVKKSATPLEVEKRVMDILPPEQWLAHQAMIYFGRAICHPKNPECDQYPQLYDFSNL

```

## APPENDIX 3:

## 12 vertebrate homologs, original sequences

```

>gi|4505471|ref|NP_002519.1| endonuclease III-like protein 1 [Homo sapiens]
MCSPEQESGMTALSARMLTRSRSLGPGAGPRGCREEPGLRRREAAAEARKSHSPVKRPRKAQRLRVAYEGSDSEKGEAE
PLKVPVWEPQDWQQQLVNIRAMRNKKDAPVDHLGTEHCYDSSAPPKVRRYQVLLSMLSSQTKDQVTAGAMQRLRARGLT
VDSILQTDATLTKGLIYPVGFWRSKVKYIKQTSAILQQHYGGDIPASVAELVALPGVGPKMAHLAMAVAWGTVSGIAVD
HVRIRANRLRWTKKATKSPEETRAALEEWLPRELWHEINGLLVGFQQTCPLVHPRCHACLNQALCPAAQGL
>gi|6912520|ref|NP_036354.1| A/G-specific adenine DNA glycosylase isoform 1 [Homo sapiens]
MTPLVSRSLRLWAIMRKPRAAVSGSHRKAASQEGRQKHAKNNSQAKPSACDGMIAECPGAPAGLARQPEEVVLQASVSS
YHLFRDVAEVTAFRGSLLSWYDQEKRDLPWRRRAEDEMDLDRRAYAVWVSEVMLQQTQVATVINYYTGWMQKWPTLQDLA
SASLEEVNQLWAGLGYYSRGRRLQEGARKVVEELGGHMPRTAETLQQLLPGVGGRYTAGAIAISAFQATGVVDGNVARVL
CRVRAIGADPSSSTLVSSQLWGLAQQLVDPARPGDFNQAAAMELGATVCTPQRPLCSQCPVESLCRARQRVEQEQLLASGSL
SGSPDVEECAPNTGQCHLCLPPSEPWDQTLGVVNFPRKASRKPPREESATCVLEQPGALGAQILLVQRPNSGLLAGLWE
FPSVTWEPSEQLRKALLQELQWAGPLPATHRLHGLVGVHTFSHIKLTQVYGLALEGQTPVTTVPPGARWLTQEEFHT
AAVSTAMKKVFRVYQGGQPGTCMGSKRSQVSSPCSRKKPRMGQVLDNFFRSHISTDAHSLNSAAQ
>gi|227908769|ref|NP_032769.2| endonuclease III-like protein 1 [Mus musculus]
MNSGVRMVTRSRSRATRIASEGCREELAPREAAAEGRKSHRPVRHPRRTQKTHVAYEAANGEEGEDAEPLKVPVWEPQNW
QQQLANIRIMRSKKMDAPVDQLGAEHYDASAPKVVRRYQVLLSMLSSQTKDQVTAGAMQRLRARGLTVESILQTDDDL
GRLIYPVGFWRNKVKYIKQTTAILQQRYEGDIPASVAELVALPGVGPKMAHLAMAVAWGTISGIAVDTHVRIANRLRW
TKMTKTPEETRNLEEWLPRLVWSEVNGLLVGFQQICPLVHPRCQACLNKALCPAAQDL
>gi|227330621|ref|NP_573513.2| A/G-specific adenine DNA glycosylase [Mus musculus]
MKKLQASVRSCHKQPANHKRRRTRALSSSQAKPSSLDGLAKQKREELLQASVSPYHLFSDVADVTAFRSNLLSWYDQEK
DLPWRNLAKKEANSRRAYAVWVSEVMLQQTQVATVIDYYTRWMQKWPKLQDLASASLEEVNQLWVSGLGYYSRGRRLQ
ARKVVEELGGHMPRTAETLQQLLPGVGGRYTAGAIAISAFDQVTGVVDGNVLRVLCRVRAIGADPTSTLVSHHLWNLQQL
VDPARPGDFNQAAAMELGATVCTPQRPLCSHCPVQSLCRAYQVRVQRLSALPGRPDIEECALNTRQCLCTSSSPWDP
MGVANFPRKASRRPPREYSATCVVEQPGAIGGPLVLLVQRPDSGLLAGLWEPFVSTLEPSEQHQHKALLQELQWCGPL
PAIRLQHLGEVIHIFSHIKLTQVYSLALDQAPASTAPPGARWLTWEEFCNAAVSTAMKKVFRMYEDHRGQTRKGSKR
VCPSSRRKKPSLQGVLDTFQRIPTDKPNSTQ
>gi|114051958|ref|NP_001039862.1| endonuclease III-like protein 1 [Bos taurus]
MNAAGVRMVVTRARSRGTGASLRRRGEKAAPLRSGEAAAEERKSYSPVKRRRKAQRLSVAYEASEGEGGEGAEHLQAPSW
QPQDWRLQDLNIRTMRSKGDAPVDQLGAEHCFDPSASPKVRRYQVLLSMLSSQTKDQVTAGAMQRLRARGLTVDSILQ
TVDSTLQALIIYPVGFWRNKVKYIKQTSAILQQRYDGDIPASVAELVALPGVGPKMAHLAMAVAWGTVSGIAVDTHVRIAN
RLRWTKKATKSPEETRAALEEWLPRELWSEINGLLVGFQQTCPLIRPRCQACLNALCPAARGL
>gi|281485563|ref|NP_001039600.2| A/G-specific adenine DNA glycosylase [Bos taurus]
MKKSRAAVGNRSRRKQASSQEGKEKCAFSSQAKPSAPSAGPARQQKALLQASVSPYHLFRDVAEVTALQESLLDWYDR
KKRDLPWRRLVEDEVLDLRRAYAVWVAEVMQQTQVATVINYYTRWMQKWPTLQDLASASLEEVNQLWAGLGYYSRGRWL
QEGARKVVEELGGHMPRTAETLQQFLPGVGGRYTAGAIAISAFGQAAGVVDGNVIRVLCRVRAIGADSSSTLVSHLWLA
QQLVDPARPGDFNQAAAMELGATVCTPKRPLCSHCPVQNLRCRARQRVEREQLSASQSLPGNCDVEECAPNTGQCP
EPWDQTLGVNTNFRKASRKPPREECSAICVLEQPKALGAHILLVQRPNSGLLAGLWEPFVSVVNAEASGQHQAALLQE
LQSWVGLPDTRLQHLGVVHTFHSIKMTYQVYSLALEEHTPTVTPPGARWLTREDFHTAAVSTAMKKVFRMYEGQQPG
TCKGSKRSQVATLSKRKKPSPGQVLESFFWPHVPTDAPSLNTAAQ
>gi|118601744|ref|NP_001073043.1| endonuclease III-like protein 1 [Gallus gallus]
MCAAPRGGGRAARRLGAATAGSRVPSAAPRYSRRTTRVPIAYEAEKPKESPGRPKEPENWQQQLERIREMRHRDAPVD
EMGVDKCYDTSAPQVMRYQVLLSMLSSQTKDQVTSAMRLRQRGLTVDSILQMDATLGQIIYPVGFWRNKVKYIKQ
TTAILKQKYGDPGTVEELVKLPGVGPKMAHLAMNIWNSVSGIAVDTHVHRIITNRLKWVKETRYPEETRALEDWLP
RDLWREINWLLVGFQQTCPLVNPRCKECLNQDIPCAAKRF
>gi|513197809|ref|XP_422433.3| PREDICTED: A/G-specific adenine DNA glycosylase isoform X5
[Gallus gallus]
MGGAAVRARRSVKVRAGGEHVGPGLGSPAIALRTHRRCCDPTVPVSRQGLPLDHMHCISSVTPSRSMPIYAACSPGMTK
AGGTFPGGRWLQLSWMLTGGRMQLGLLVASERQVWARERSCGDEDEGEGCWVGFCCSSWNQHGDRGACCEKWHWHLCT
DPSVLLAVVWSEIMLQQTQVATVIDYNNRWMQKWPTLQALAAASLEEVNELWAGLGYYSRGKRLQEAARKVVSELAGRMP
RTAEDLQRLLPVGGRYTAGAIAISIFGQATGVVDGNVIRVLCRLRCIGADTSSLAVIDCLWDMANTLVDRSRPGDFNQAL
MELGATVCTPKSPLCRECPVKECHAWRRVEKELASASQKLFGKTTLPDVEDCGPGGCLPLPAAEPWDSSLGVTNFP
KAAKQPRVEWTATCVLERRGRLAGAPEYLIVQRPSSGLLAGLWEPFSLPLAPGLQEEQQKEVLADHLRAWTRQPVTQSL
CFIGEVVHIFSHIHQTYVVYSLCLDGDVALDAASSPSRWVTEEFRAVSTAMKKVLKARETQRGVQSGRAKGSKRKRE
SKLGAAGSTPTGMQLSLRAFLRAQPPP
>gi|113205550|ref|NP_001037884.1| nth endonuclease III-like 1 [Xenopus (Silurana) tropicalis]
MSGSLRPLGRRGRGVLKAVGGKDQDGTSGKQVIDDSEDEKPSPKERSKRRVSVEYEQAASETVAKRPKWQPKNWAQH
LENIRQMRSRDAPVDQGAEEKCYDQNAAEVVMRYQIILLMLSSQTKDQVTSAMCRLRQHGLTVSRILETDDGTGLGKL
IYPVGFWNKVKYIKQTEILQEKYGGDIPDNVTDLVKLPVGVPKMAHLVMDIAWNNVSGIGVDTHVHRIISNRLKWVRKE
TKTPEETRVAMEDWMPRELWSEINWLLVGFQQVCLPVSRCSECLNKDICPGAKKKKPR
>gi|118403607|ref|NP_001072831.1| mutY homolog [Xenopus (Silurana) tropicalis]
MPPPRTKTSLGRSAAASGKRKSPKQAFPKREEHVLQSSYHSFTSQETEIIRDKLLAWYDKSRDLPWRMTACTEPDLDR
KAYAVWVSEVMLQQTQVATVIDYNNKMKVWPTMEDLARSSLEEVNEMWVSGLGYYSRGRRLQEGAKKVLELGGSMPSRA
DELQKLLPGVGGRYTAGAIAISYQVTVVDGNVIRVLSRLRCIGADSSTLAVSDKLWNLANALVDPDRPGDFNQGMME

```

GATVCTPKKPLCTACPLQGQCKAYLKVIAEKESAVKTLIKKQASPIAKDVGDIEDCDLGPGLCALCVPTSDPDWSSSLGVA  
NFRKSAKKPSRMEQTAICVWEKCGDHGELEYLIVQRPSSGLLAGLWEFPSILLDEKFTQNRQHSLLGLLQDLSGHAVP  
LQKLQYKGEVVHIFSHIHQTYVVYFLSLNTTENC SVKTEETERPLTRWVTKKEFLNSAVPTAMKKIMKLCESHGSSCTAV  
NTSKRRKGD LAKVQLPSGRIKTEKGKQSQSIQSFFKLATEK  
>gi|410917257|ref|XP\_003972103.1| PREDICTED: endonuclease III-like protein 1-like [Takifugu  
rubripes]  
MTSHYFAQSRVSVTRRGAQNAAHKPATSLKSKLTIQPEKDDLVSSSAGVKLEEEEA KISGNALKPETDAPTLSSH SRRRR  
QLKVEYDKDGSMPLKTEPWEPPRWKTQLENIRAMRSGRDAPVDNMGADKCHDADAPAHVKRFQVLVSLMLSSQTKDQVT  
SAAMQKLRAGHGTVENILATNDETLGQLIYPVGFWRNKVYKLKLT SAMLQKEFGGDIPDSVEGLVRLPGVGPMAHLAMD  
IAWDQVSGIGVDTHVHRISNRLGWLKKPTKTPEETRKSLEEWLPRELWSEINWLLVGFQGVCLPVSPLCSVCLNQHDPC  
SAHKNSPVRRPKFLERSPWIKSPRLFITPGANTFLIR  
>gi|410921366|ref|XP\_003974154.1| PREDICTED: A/G-specific adenine DNA glycosylase-like  
[Takifugu rubripes]  
MSTQGEVPQVGKVL SFLREWD RGDRSARGRMLSSFLGRSAGRTQGELEYLGEFVGHGGVVTLLEVLTQPPQSNEETKAEAL  
CLLLAISDAGRKYKELICQSCGAMAAAECLTHSGTGETQESAWMLLESLSHGNPKYEGE IYKGLIGHLTCTSAKAQQFVL  
HTLHTLQSKMEIAHHSIVEPLLGVLTSLHPDVQSEVARLIFELRRYDVRPMLLRALCGLGLNVARAPTYPEEESHASSP  
SAYHFFHDAADVALLRSRLAWYDQEKRELPWRTLALTEPDVNIRTYAVVWSEIMLQQTQVATVIDYYNKWMMRWPTVQD  
LATATLEDVNQMWAGLGYYSRGKRLHEGAQKVVSQ LQGEMPRTVDALLKQLPGVG RYTAGAIGSIALGQVTGAVDGNVIR  
VLCRLRAIGADCTGPTVTEALWSLANTLVDPDRPGDFNQALMELGARVCTPKGPLCTRCP IQSHCHSYHKAHFKQEKN SM  
KLMGKLD RKSSALPDIEDCLSSGTCTLCLEP WDELGVQNFPRKPAKKPPRAERCLTCVVIRQGE GGEHEFLLTQRPSK  
GLLAGLWEFPCINHEEKNAVEEKKVLC AEINRILGTS LTHGLLQYVGEVVHIFSHIHQTYVVHTLRLKDAVSQSENMQWL  
TPSALQEAAVSTGVKKIMKLCNSALGQQGAPDGEKRPKKDRKGQITKRPRLSGANSRSRQLSLSSFFQTVKQDC



**APPENDIX 4:****12 vertebrate homologs, modified headers**

```

>Homo_sapiens_NP_002519
MCSPEQESGMTALSARMLTRSRSLGPGAGPRGCREEPGLRRREAAAEARKSHSPVKRPRKAQRLRVAYEGSDSEKGEAE
PLKVPVWEPQDWQQQLVNIRAMRNKKDAPVDHLGTEHCYDSSAPKVVRYQVLLSMLSSQTKDQVTAGAMQRLRARGLT
VDSILQTDATLTKGLIYPVGFWRSKVKYIKQTSAILQQHYGGDIPASVAELVALPGVGPKMAHLAMAVAWGTVSGIAVD
THVHRIANRLRWTKKATKSPEETRAALEEWLPRELWHEINGLLVGFQQQTCLPVHPRCHACLNQALCPAAQGL

>Homo_sapiens_NP_036354
MTPLVSRSLRWAIMRKPRAAVSGSHRKQAASQEGRQKHAKNNSQAKPSACDGMIAECPGAPAGLARQPEEVVLQASVSS
YHLFRDVAEVTAFRGSLLSWYDQEKRDLPWRRRAEDEMDLDRRAYAVVWSEVMLQQTQVATVINYYTGWMQKWPTLQDLA
SASLEEVNQLWAGLGYYSRGRRLQEGARKVVEELGGHMPRTAETLQQLLPGVGGRYTAGAIAISAFGQATGVVDGNVAVL
CRVRAIGADPSSSTLVSSQQLWGLAQQLVDPARPGDFNQAAAMELGATVCTPQRPLCSQCPVESLCRARQRVEQEQLLASGSL
SGSPDVEECAPNTGQCCHLCLPSEPQDWQTLGVVNFPRKASRKPPREESATCVLEQPGALGAQILLVQRPNSGLLAGLWE
FPSVTWEPSEQLRKALLQELQRWAGPLPATHRLHGLVGVHTFHSIKLTYQVYGLALEGQTPVTVPVPGARWLTQEEFHT
AAVSTAMKKVFRVYQQQPGTCMGSKRSQVSSPCSRKKPRMGQQVLDNFRSHISTDAHSLNSAAQ

>Mus_musculus_NP_032769
MNSGVRMVRTRSRSRATRIASEGCEELAPREAAAEGRKSHRPVHRPRRTQKTHVAYEAANGEEGEDAEPLKVPVWEPQNW
QQQLANIRIMRSKKDAPVDQLGAEHCYDASAPKVVRYQVLLSMLSSQTKDQVTAGAMQRLRARGLTVESILQTDGDTL
GRLIYPVGFWRNKVKYIKQTTAILQQRYEGDIPASVAELVALPGVGPKMAHLAMAVAWGTISGIAVDTHVHRIANRLRW
TKMTKTPEETRNLEEWLPRVLWSEVNGLLVGFQQQICLPVHPRCQACLNKALCPAAQDL

>Mus_musculus_NP_573513
MKKLQASVRSRSHKQPANHKKRRRTRALSSSQAKPSSLDGLAKQKREELLQASVSPYHLFSDVADVTAFRSNLLSWYDQEK
RDLPRNLAKKEANSRRAYAVVWSEVMLQQTQVATVIDYYTRWMQKWPKLQDLASASLEEVNQLWAGLGYYSRGRRLQEG
ARKVVEELGGHMPRTAETLQQLLPGVGGRYTAGAIAISAFDQVTGVVDGNVLRVLCRVRAIGADPTSTLVSHHLWNLQQL
VDPARPGDFNQAAAMELGATVCTPQRPLCSHCPVQSLCRAYQVRVQRGQLSALPGRPDIEECALNTRQCQLCLTSSSPWDP
SGVANFPRKASRRPPREESATCVVEQPGAIGGPLVLLVQRPDSGLLAGLWEPSPVTLEPSEQHQHKALLQELQRCWGP
PAIRLQHLGEVIHIFSHIKLTYQVYSLALDQAPASTAPPGARWLTWEEFCNAAVSTAMKKVFRMYEDHRGQTRKGSKR
SQVCPSSRKKPSLGQQVLDTFQRIPTDKPNSTQ

>Bos_taurus_NP_001039862
MNAAGVRMVRTARSRGTGASLRRRGEKAAPLRSGEAAAEERKSYSVPKRRRKAQRLSVAYEASEGEGGEGAEHLQAPSW
QPQDWRQQLDNIRTMRSKGDAPVDQLGAEHCFDPSASPKVVRYQVLLSMLSSQTKDQVTAGAMQRLRARGLTVD
SILQTDALGALIIYPVGFWRNKVKYIKQTSAILQQRYDGDIPASVAELVALPGVGPKMAHLAMAVAWGTVSGIAVD
THVHRIANRLRWTKKATKSPEETRAALEEWLPRELWSEINGLLVGFQQQTCLPIRPRCQACLNALCPAARGL

>Bos_taurus_NP_001039600
MKKSRAAVGNRSRGRKQASSQEGKEKCAFSSQAKPSAPSAGPARQQKALLQASVSPYHLFRDVAEVTALQESL
LDWYDRKKRDLPRWRLVEDEVLDLRRAYAVVWSEVMLQQTQVATVINYYTRWMQKWPTLQDLASASLEEVNQLWAGL
GYYSRGRWLQEGARKVVEELGGHMPRTAETLQQFLPGVGGRYTAGAIAISAFGQAAGVVDGNVIRVLCRVRAIGAD
SSSTLVSQHLWLSLQQQLVDPARPGDFNQAAAMELGAICTPKRPLCSHCPVQNLCLRARQRVEREQLSASQSLPGN
CDVEECAPNTGQCPLCAPPTEPWDQTLGVNTNFRKASRKPPREECSAICVLEQPKALGAHILLVQRPNSGLLAGLW
EFPSPSVNNAEASGQHQRAALLQELQSWVGPLPDTRLQHLGQVHTFHSIKMITYQVYSLALEHTPTVTPVPGARW
LTREDFTAAVSTAMKKVFRMYEGQQPGTCGSKRSQVATLSKRKKPSPGQQVLESFFWPHVPTDAPSLNTAAQ

>Gallus_gallus_NP_001073043
MCAAPRGGGGAARRLGAATAGSRVPSAAPRYSRRTTRVPIAYEAEKPKESPGRPKEPENWQQQLERIREMRHRDAP
VD EMGVDKCYDTSAPQVMRYQVLLSMLSSQTKDQVTSAMLRRLRQGLTVDSILQMDATLGQIIYPVGFWRNKVKYIK
QTTAILKQKYGGDIPGTVEELVKLPGVGPKMAHLAMNIWNSVSGIAVDTHVHRIITNRLKWVKETRYPEETRALE
DWLPRDLWREINWLLVGFQQQTCLPVNPRCKECLNQDPCPAAKRF

>Gallus_gallus_XP_422433
MGGAAVRARRSVKVRAGGEHVGPGLGSPAIALRTHRRCCDPTVPVSRQGLPLDHMHCISSVTPSRSMYPVAACSP
GMTKAGGTFFPGRWLQLSWMLTGGRMQLGLLVASERQVWARERSCGDEGEEGCWVGFCSSWNQQHGDGACCEKWH
HHLCTDPSVLLAVVWSEIMLQQTQVATVIDYYNRWMQKWPTLQALAAASLEEVNELWAGLGYYSRGKRLQEAARKV
VSELAGRMPRTAEDLQRLLPVGGRYTAGAIAISISFGQATGVVDGNVIRVLCRLCIGADTSSLAVIDCLWDMANTL
VDRSRPGDFNQALMELGATVCTPKSPLCRECPVKECHAWRRVEKELASASQKLFKTTLPDVEDCGPGGCPLCLP
AAEPWDSSLGVTNFRPKAAKKQPRVEWTATCVLERRGRLAGAPEYLIVQRPSSGLLAGLWEPFSLPLAPGLQEEQ
QKEVLADHLRAWTRQPVQTQSLCFI GEVVHIFSHIHQTYVVYSLCLDGDVALDAASSPSRWVTEEEFRASAVSTAM
KKVLKARETQRGVQSGRAKGSKRKRESKLGAAGSTPTGMQLSLRAFLRAQPPP

>Xenopus_(Silurana)_tropicalis_NP_001037884
MSGSLRPLGRGRGVKAVGGKDQDGTSGKQVIDDSEDEKPSPKERSKRRVSVEYEQAASETVAKRPKWQPKNWAQH
LENIRQMRSRRDAPVDQMAEKCYDQNAAPVEMRYQIILLMLSSQTKDQVTSAMCRLRQHGLTVSRIETDDGT
LGKLIYPVGFWRNKVKYIKQTEILQEKYGGDIPDNVTDLVKLPVGVPKMAHLVMDIAWNNVSGIGVDTHVHRI
SNRLKWVRKETKTPEETRVAMEDWMPRELWSEINWLLVGFQQQVCLPVSPRCSECLNKDPCGAKKKKPR

>Xenopus_(Silurana)_tropicalis_NP_001072831
MPPTRTKSLGRSAAASGKRKSPKQAFPKREEHLQSSIIYHSFTSQTEIIRDKLLAWYDKSKRDLPWRTMACTE
PDLDRKAYAVVWSEVMLQQTQVATVIDYYNKWMKVWPTMEDLARSSLEEVNEMWWSGLGYYSRGRRLQEGAKK
VVLELGGMPRSADLQKLLPGVGGRYTAGAIAISISYGQVTGVVDGNVIRVLSRLRCIGADSSSTLAVSDKLWN
LANALVDPDRPGDFNQGMMELGATVCTPKKPLCTACPLGQCKAYLKVIAEKESAVKTLIKQASPIAKDVGDI
EDCDLGPGLCALCVPTSDPWDSSLGVA

```

NFPRKSAKKPSRMEQTAICVWEKCGDHGELEYLIVQRPSSGLLAGLWEFFSILLDEKFTEQNRQHSLLGLLQDLSGHAVP  
LQKLQYKGEVVHIFSHIHQTYVYVFLSLNTTENC SVKTEETERPLTRWVTKKEFLNSAVPTAMKKIMKLCESHGSSCTAV  
NTSKKRKGD LAKVQLPSGRIKTEKKGQSQSFFKLATEK  
>Takifugu\_rubripes\_XP\_003972103  
MTSHYFAQSRSVVTRRG AQNAAHKPATSLKSKLTIQPEKDDLVS SAGVKLEEEEEAKISGNALKPETDAPTLSSH SRRRR  
QLKVEYDKDGSM PQLKTEPWEPPRWKTQLENIRAMRSGRDAPVDNMGADKCHDADAPAHVKRFQVLVSLMLSSQTKDQVT  
SAAMQKLRAHGCTVENILATNDET LGQLIYPVGFWRNKVYLKLT SAMLQKEFGGDI PDSVEGLVRLPGVGPKMAHLAMD  
IAWDQVSGIGVDTHVHRISNRLGWLKKPTKTPEETRKSLEEWLPRELWSEINWLLVGFGQQVCLPVSP LCSVCLNQHD CP  
SAHKNSPVRRPKFLERSPWIKSPRLFITPGANTFLIR  
>Takifugu\_rubripes\_XP\_003974154  
MSTQGE PVQVGKVL SFLREWDRGDRSARGRMLSSFLGRSAGRTQGELEYLGEFVGHGGVVTLLLEVLTQ PQSNEETKAEAL  
CLLLAISDAGRKYKELICQSCGAMAAA ECLTHSGTGETQESAWMLLESLSHG NPKYEGEIYKGLIGHLTCTSAKAQQFVL  
HTLHTLQSKMEIAHHSIVEPLLGVLTSLHPDVQSEVARLIFELRRYDVRPMLLRALCGLGNVARAPTYPEEESHASSP  
SAYHFFHDAADVALLRSRL LAWYDQEKREL PWRTLALTEPDVNIRTYAVVWSEIMLQQTQVATVIDYYNKMMKRWPTVQD  
LATATLEDVNQMWAGLGYYSRGKRLHEGAQKVVSQ LQGEMPRTVDALLKQLPGVGGRYTAGAIGSIALGQVTGAVDGNVIR  
VLCRLRAIGADCTGPTVTEALWSLANTLVDPDRPGDFN QALMELGARVCTPKGPLCTRCP IQSHCHSYHKAHFKQEKN SM  
KLMGKLD RKSSALPDIEDCLSSGTCTLCLSEPWDELGVQNFPRKPAKKPPRAERCLTCVVIRQGE GEGEHEFLTQRPSK  
GLLAGLWEFFCINHEEKNAVEEKKVLC AEINRILGTS LTHGLLQYVGEVVHIFSHIHQTYV VHTLR LKDAVSQSENMQWL  
TPSALQEAAVSTGVKKIMKLCNSALGQQGAPDGE EKRPKKDRKGQITKRPRLSGANSRSRQLSLSSFFQTVKQDC

## 12 vertebrate homologs, final headers

25

NFPRKSAKKPSRMEQTAICVWEKCGDHGELEYLIVQRPSSGLLAGLWEFFPSILLDEKFTEQNRQHSLLGLLQDLSGHAVP  
LQKLQYKGEVVHIFSHIHQTYVYVFLSLNTTENCSCVKTEETERPLTRWVTKKEFLNSAVPTAMKKIMKLCESHGSSCTAV  
NTSKKRKGD LAKVQLPSGRIKTEKGKQSQSIQSFFKLATEK  
>Takifugu\_rubripes\_Nth  
MTSHYFAQSRSVVTRRGQAHAHKPATSLKSKLTIQPEKDDLVS SAGVKLEEEEEAKISGNALKPETDAPTLSSH SRRRR  
QLKVEYDKDGSMPLKTEPWEPPRWKTQLENIRAMRSGRDAPVDNMGADKCHDADAPAHVKRFQVLVSLMLSSQTKDQVT  
SAAMQKLRAHGCTVENILATNDET LGQLIYPVGFWRNKVYLKLT SAMLQKEFGGDI PDSVEGLVRLPGVGPKMAHLAMD  
IAWDQVSGIGVDTHVHRISNRLGWLKKPTKTPEETRKSLEEWLPRELWSEINWLLVGFGQQVCLPVSP LCSVCLNQHD CP  
SAHKNSPVRP KFLERSPWIKSPRLFITPGANTFLIR  
>Takifugu\_rubripes\_MutY  
MSTQGE PVQVGK VLSFLREWD RGDRSARGRMLSSFLGRSAGRTQGELEYLGEFVGHGGVVTLLLEVLTQPQSNEETKAEAL  
CLLLAISDAGRKYKELICQSCGAMAAA ECLTHSGTG ETQESAWMLLESLSHG NPKYEGEIYKGLIGHLTCTSAKAQQFVL  
HTLHTLQSKMEIAHHSIVEPLLGVLTSLHPDVQSEVARLIFELRRYDVRPMLLRALCGLGNVARAPTYPEEESHASSP  
SAYHFFHDAADVALLRSRL LAWYDQEKREL PWRTLALTEPDVNIRTYAVWVSEIMLQQTQVATVIDYYNKMMKRWPTVQD  
LATATLEDVNQMMWAGLGYYSRGRLHEGAQKVVSQ LQGEMPR TVDALLKQLPGVG RYTAGAIGSIALGQVTGAVDGNVIR  
VLCRLRAIGADCTGPTVTEALWSLANTLVDPDRPGDFN QALMELGARVCTPKGPLCTRCP IQSHCHSYHKAHFKQEKN SM  
KLMGKLD RKSSALPDIEDCLSSGTCTLCLSEP WDELGVQNFPRKPAKKPPRAERCLTCVVIRQEGEGEHEFLLTQRPSK  
GLLAGLWEFFCINHEEKNAVEEKKVLC AEINRILGTS LTHGLLQYVGEVVHIFSHIHQTYVVHTLR LKDAVSQSENMQWL  
TPSALQEAAVSTGVKKIMKLCNSALGQQGAPDGE EKRPKKDRKGQITKRPRLSGANSRSRQLSLSSFFQT VVKQDC

**APPENDIX 6:****4 human homologs, original headers**

```

>gi|4505471|ref|NP_002519.1| endonuclease III-like protein 1 [Homo sapiens]
MCSPQESGMTALSARMLTRSRSLGPGAGPRGCREEPGLRRREAAAEARKSHSPVKRPRKAQRLRVAYEG
SDSEKGEAEPLKVPVWEPQDWQQQLVNIRAMRNKKDAPVDHLGTEHCYDSSAPPKVRRYQVLLSMLSS
QTKDQVTAGAMQRLRLRAGLTVDSTILQTDATLGKLIYPVGFWRSKVYIKQTSAILQQHYGGDIPASVAE
LVALPGVGPKMAHLAMAVAWGTVSGIAVDTHVHRIANRLRWTKKATKSPETRAALEEWLPRELWHEING
LLVGFGQQTCLPVHPRCHACLNQALCPAAQGL
>gi|6912520|ref|NP_036354.1| A/G-specific adenine DNA glycosylase isoform 1 [Homo sapiens]
MTPLVSRSLRLWAIMRKPRAAVGSGRKQAASQEGRQKHAKNNSQAKPSACDGMIAECPGAPAGLARQPE
EVLVQASVSSYHLFRDVAEVTAFRGSLLSWYDQEKRDLPWRRRAEDEMDLRRAYAVWVSEVMLQQTQVA
TVINYTTGWMQWPTLQDLASASLEEVNQLWAGLGYYSRGRRLQEGARKVVEELGGHMPRTAETLQQLLP
GVGRYTAGAIIASIAFGQATGVVDGNVARVLCRVRAIGADPSSTLVSQQLWGLAQQLVDPARPGDFNQAM
ELGATVCTPQRPLCSQCQVESLCRARQVEQEQLLASGSLSGSPDVEECAPNTGQCHLCLPPSEFPWDQTL
GVVNFPRKASRKPPREESATCVLEQPGALGAQILLVQRPNSSGLLAGLWEFFSVTWEPSEQLQRKALLQE
LQRWAGPLPATHRLHLEGEVHTFSHIKLTYYQVYGLALEGQTPVTVPPGARWLTQEEFHTAAVSTAMKKV
FRVYQGGQPGTCMGSKRQVSSPCSRKKPRMGQVLDNFFRSHISTDAHSLNSAAQ
>gi|4505495|ref|NP_002533.1| N-glycosylase/DNA lyase isoform 1a [Homo sapiens]
MPARALLPRRMGHRTLASTPALWASIPCPRSELRLDLVLPSSGQSFRWREQSPAHWSGVLADQVWTLTQTE
EQLHCTVYRGDKSQASRPTPDELEAVRKYFQLDVTLAQLYHHWGSVDSHFQEVAKQKFGVRLLRQDPIEC
LFSFICSSNNNIARITGMVERLCQAFGPRLIQLDDVTYHGFPSLQALAGPEVEAHLRKLGLGYRARYVSA
SARAILEEQGGLAWLQQLRESSYEEAHKALCILPGVGTQVADICLMALDKPQAVPVDVHMHIAQRDYS
WHPTTSQAKGPPSPQTNKELGNFFRSLWGPYAGWAQAVLFSADLRQSRHAQEPAPAKRRKSGSGPEG
>gi|4505121|ref|NP_003916.1| methyl-CpG-binding domain protein 4 [Homo sapiens]
MGTTGLESLSLGDRAAPTPTSSERLVPDPPNDLRKEDVAMELERVGEDEEQMMIKRSSECNPLLQEPPIA
SAQFGATAGTECRKSVPCGWERVVKQRLFGKTAGRFVDYFISPGQLKFRSKSSLANYLHKNGETSLKPED
FDFTVLSKRGIKSRYKDCSMAALTSHLQNSNNSNWNLRTRSKCKKDVFMPPSSSSSELQESRGLSNFTST
HLLKKEDEGVDDVNFVRKPKGKVTILKGIPIKKTGCRKSCSGFVQSDSKRESVCNKADAEESEPVAQ
KSQLDRTVCISDAGACGETLSVTSEENSLVKKKERSLSSGSNFCSEQKTSGIINKFCSAKDSEHNEKYED
TFLESEEIGTKVEVVERKEHLHTDILKRGSEMDNNSPTRKDFTEGKIFQEDTIPRTQIERRKTSLYFSS
KYNKEALSPRRKAFKKWTPPRSPFNLVQETLFHDPWKLIIATIFLNRTSGKMAIPVLWKFLKYPSPAEV
ARTADWRDVSELLKPLGLYDLRAKTIVKFSDEYLTQWKYPIELHGIGKYGNDSYRIFCVNEWKQVHPED
HKLNKYHDLWLENHEKLSLS

```

**APPENDIX 7:****4 human homologs, modified headers**

```

>NTHL1_Homo_sapiens
MCSPQESGMTALSARMLTRSRSLGPGAGPRGCREEPGLRRREAAAEARKSHSPVKRPRKAQRLRVAYEG
SDSEKGEAEPLKVPVWEPQDWQQQLVNIRAMRNKKDAPVDHLGTEHCYDSSAPPKVRRYQVLLSMLSS
QTKDQVTAGAMQRLRLRAGLTVDLSILQTDATLGKLIYPVGFWRSKVKYIKQTSAILQQHYGGDIPASVAE
LVALPGVGPKMAHLAMAVAWGTVSGIAVDTHVHRIANRLRWTKKATKSPETRAALEEWLPRELWHEING
LLVGFGQQTCPLVHPRCHACLNQALCPAAQGL
>MUTYH_Homo_sapiens
MTPLVSRLSRLWAIMRKPRAAVSGSHRKQAASQEGRQKHAKNNSQAKPSACDGMIAECPGAPAGLARQPE
EVVLQASVSSYHLFRDVAEVTAFRGSLLSWYDQEKRDLPWRRRAEDEMDLDRRAYAVVWSEVMLQQTQVA
TVINYTTGWMQKWPTLQDLASASLEEVLWAGLGYYSRGRRLQEGARKVVEELGGHMPRTAETLQQLLP
GVGRYTAGAIIASIAFGQATGVVDGNVARVLCRVRAIGADPSSTLVSQQLWGLAQQLVDPARPGDFNQAM
ELGATVCTPQRLCSQCPEVSLCRARQVEQEQLLASGSLSGSPDVEECAPNTGQCHLCLPPSEFPWDQTL
GVVNFPRKASRPPREESATCVLEQPGALGAQILLVQRPNSGLLAGLWEFFPSVTWEPSEQLQRKALLQE
LQRWAGPLPATHRLHLEVVHTFSHIKLTYYQVYGLALEGQTPVTVPPGARWLTQEEFHTAAVSTAMKKV
FRVYQGGQPGTCMGSKRSQVSSPCSRKKPRMGQQVLDNFFRSHISTDAHSLNSAAQ
>OGG1_Homo_sapiens
MPARALLPRRMGHRTLASTPALWASIPCPRSELRLDLVLPSSGQSFWRREQSPAHWSGVLADQVWTLTQTE
EQLHCTVYRGDKSQASRPTPDELEAVRKYFQLDVTLAQLYHHWGSVDSHFQEVAKKFQGVRLLRQDPIEC
LFSFICSSNNNIARITGMVERLCQAFGPRLIQLDDVTYHGFPSLQALAGPEVEAHLRKLGLGYRARYVSA
SARAILEEQGGLAWLQQLRESSYEEAHKALCILPGVGTKVADCICLMALDKPQAVPVDVHMHIAQRDYS
WHPTTSQAKGPPSPQTNKELGNFRSLWGPYAGWAQAVLFSADLRQSRHAQEPAPAKRRKSGKGPGE
>MBD4_Homo_sapiens
MGTTGLESLSLGDGRGAAPTVTSSERLVPDPNDLRKEDVAMELERVGEDEEQMMIKRSSECNPLLQEPPIA
SAQFGATAGTECRKSVPCGWERVVKQRLFGKTAGRFDVYFISPGQLKFRSKSSLANYLHKNGETSLKPED
FDFTVLSKRGIKSRYKDCSMAALTSHLQNSNNSNWNLRTRSKCKKDVFMPPSSSELQESRGLSNFTST
HLLKKEDEGVDDVNFVRKVRPKGKVTILKGIPIKKTGCRKSCSGFVQSDSKRESVCNKADAESPEVAQ
KSQLDRTVCISDAGACGETLSVTSEENSLVKKKERSLSSGSNFCSEQKTSGIINKFCSAKDSEHNEKYED
TFLESEEIGTKVEVVERKEHLHTDILKRGSEMDNNSPTRKDFTEKIFQEDTIPRTQIERRKTSLYFSS
KYNKEALSPRRKAFKKWTPPRSPFNLVQETLFHDPWKLIIATIFLNRTSGKMAIPVLWKFLKYPSPAEV
ARTADWRDVSELKPLGLYDLRAKTIVKFSDEYLTQWKYPIELHGIGKYGNDSYRIFCVNEWKQVHPED
HKLNKYHDLWENHEKLSLS

```