Answers to exercises in MBV-INF4410 - Sequence alignment and searching

Exercise 1

a)

The E.coli Nth protein was found by searching the NCBI protein database for (“Escherichia coli” AND “Endonuclease III”) using NCBI Entrez, and filtering for RefSeq proteins and for the K-12 MG1655 strain.

b)

Here is the entry for the E.coli Nth protein in FASTA format:

```
>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
```

c)

The Nth protein sequences were found by searching NCBI Entrez with the gi numbers. Here are the entries for the other Nth proteins:

```
>gi|57117142|ref|NP_218191.2| endonuclease III [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
```

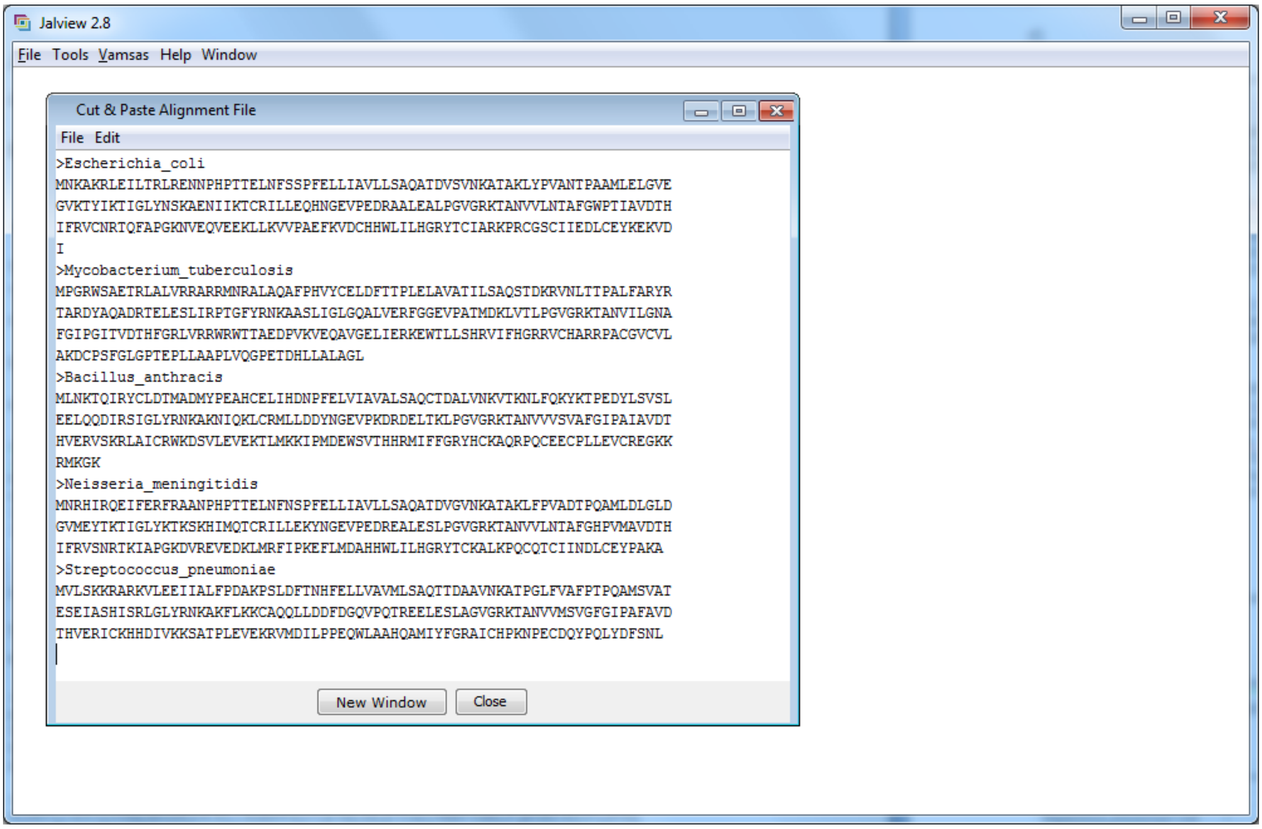
d)

Here are all five entries for Nth proteins with shortened descriptions:

```
>Escherichia_coli
MNKAKRLEILTRLRENNPHPTTELNFSSPFELLIAVLLSAQATDVSVNKATAKLYPVANTPAAMLELGVE
GVKTYIKTIGLYNSKAENIIKTCRILLEQHNGEVPEDRAALEALPGVGRKTANVVLNTAFGWPTIAVDTH
IFRVCNRTQFAPGKNVEQVEEKLLKVVPAEFKVDCHHWLILHGRYTCIARKPRCGSCIIEDLCEYKEKVD
I
>Mycobacterium_tuberculosis
MPGRWSAETRLALVRRARRMNRALAQAFPHVYCELDFTTPLELAVATILSAQSTDKRVNLTTPALFARYR
TARDYAQADRTELESLIRPTGFYRNKAASLIGLGQALVERFGGEVPATMDKLVTLPGVGRKTANVILGNA
FGIPGITVDTHFGRLVRRWRWTTAEDPVKVEQAVGELIERKEWTLLSHRVIFHGRRVCHARRPACGVCVL
AKDCPSFGLGPTEPLLAAPLVQGPETDHLLALAGL
>Bacillus_anthracis
MLNKTQIRYCLDTMADMYPEAHCELIHDNPFELVIAVALSAQCTDALVNKVTKNLFQKYKTPEDYLSVSL
EELQQDIRSIGLYRNKAKNIQKLCRMLLDDYNGEVPKDRDELTKLPGVGRKTANVVVSVAFGIPAIAVDT
HVERVSKRLAICRWKDSVLEVEKTLMKKIPMDEWSVTHHRMIFFGRYHCKAQRPQCEECPLLEVCREGKK
RMKGK
>Neisseria_meningitidis
MNRHIRQEIFERFRAANPHPTTELNFNSPFELLIAVLLSAQATDVGVNKATAKLFPVADTPQAMLDLGLD
GVMEYTKTIGLYKTKSKHIMQTCRILLEKYNGEVPEDREALESLPGVGRKTANVVLNTAFGHPVMAVDTH
IFRVSNRTKIAPGKDVREVEDKLMRFIPKEFLMDAHHWLILHGRYTCKALKPQCQTCIINDLCEYPAKA
>Streptococcus_pneumoniae
MVLSKKRARKVLEEIIALFPDAKPSLDFTNHFELLVAVMLSAQTTDAAVNKATPGLFVAFPTPQAMSVAT
ESEIASHISRLGLYRNKAKFLKKCAQQLLDDFDGQVPQTREELESLAGVGRKTANVVMSVGFGIPAFAVD
THVERICKHHDIVKKSATPLEVEKRVMDILPPEQWLAAHQAMIYFGRAICHPKNPECDQYPQLYDFSNL
```

e)

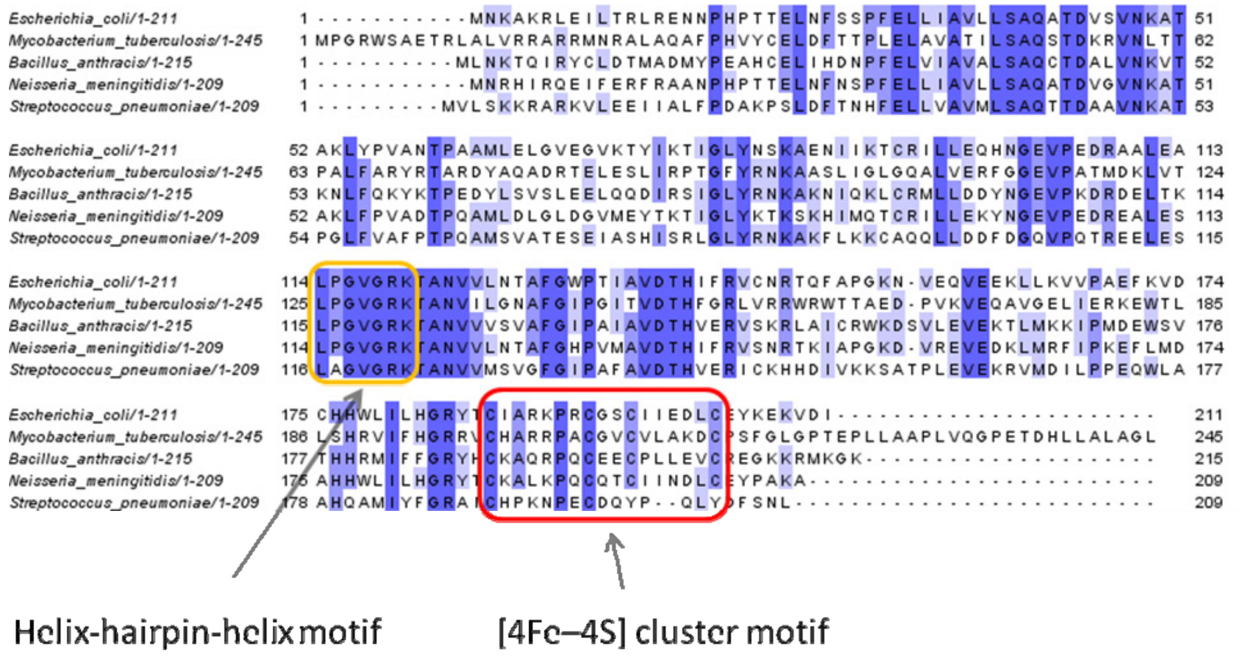
Below is a screenshot from Jalview with the input sequences



```
>Escherichia_coli
MNKAKRLEILTRLRENNPHPTTELNFSSPFELLIAVLLSAQATDVSVNKATAKLYPVANTPAAMLELGVE
GVKTYIKTIGLYNSKAENIIKTCRILLEQHNGEVPEDRAALEALPGVGRKTANVVLNTAFGWPTIAVDTH
IFRVCNRTQFAPGKNVEQVEEKLLKVVPAEFKVDCHHWLILHGRYTCIARKPRCGSCIIEDLCEYKEKVD
I
>Mycobacterium_tuberculosis
MPGRWSAETRLALVRRARRMNRALAQAFPHVYCELDFTTPLELAVATILSAQSTDKRVNLTTPALFARYR
TARDYAQADRTELESLIRPTGFYRNKAASLIGLGQALVERFGGEVPATMDKLVTLPGVGRKTANVILGNA
FGIPGITVDTHFGRLVRRWRWTTAEDPVKVEQAVGELIERKEWTLLSHRVIFHGRRVCHARRPACGVCVL
AKDCPSFGLGPTEPLLAAPLVQGPETDHLLALAGL
>Bacillus_anthracis
MLNKTQIRYCLDTMADMYPEAHCELIHDNPFELVIAVALSAQCTDALVNKVTKNLFQKYKTPEDYLSVSL
EELQQDIRSIGLYRNKAKNIQKLCRMLLDDYNGEVPKDRDELTKLPGVGRKTANVVVSVAFGIPAIAVDT
HVERVSKRLAICRWKDSVLEVEKTLMKKIPMDEWSVTHHRMIFFGRYHCKAQRPQCEECPLLEVCREGKK
RMKGK
>Neisseria_meningitidis
MNRHIRQEIFERFRAANPHPTTELNFNSPFELLIAVLLSAQATDVGVNKATAKLFPVADTPQAMLDLGLD
GVMEYTKTIGLYKTKSKHIMQTCRILLEKYNGEVPEDREALESLPGVGRKTANVVLNTAFGHPVMAVDTH
IFRVSNRTKIAPGKDVREVEDKLMRFIPKEFLMDAHHWLILHGRYTCKALKPQCQTCIINDLCEYPAKA
>Streptococcus_pneumoniae
MVLSKKRARKVLEEIIALFPDAKPSLDFTNHFELLVAVMLSAQTTDAAVNKATPGLFVAFPTPQAMSVAT
ESEIASHISRLGLYRNKAKFLKKCAQQLLDDFDGQVPQTREELESLAGVGRKTANVVMSVGFGIPAFAVD
THVERICKHHDIVKKSATPLEVEKRVMDILPPEQWLAAHQAMIYFGRAICHPKNPECDQYPQLYDFSNL
```

f)

Below is the multiple alignment of the bacterial Nth sequences as produced by MUSCLE. The Helix-hairpin-helix (HhH) and the [4Fe-4S] cluster motif are indicated. The HhH motif is fully conserved in all species, while the [4Fe-4S] cluster is conserved in all but the Streptococcus species, which lack the two last cysteines.



```
Escherichia_coli/1-211                1 -----------MNKAKRLEILTRLRENNPHPTTELNFSSPFELLIAVLLSAQATDVSVNKAT 51
Mycobacterium_tuberculosis/1-245      1 MPGRWSAETRLALVRRARRMNRALAQAFPHVYCELDFTTPLELAVATILSAQSTDKRVNLTT 62
Bacillus_anthracis/1-215              1 ----------MLNKTQIRYCLDTMADMYPEAHCELIHDNPFELVIAVALSAQCTDALVNKVT 52
Neisseria_meningitidis/1-209          1 -----------MNRHIRQEIFERFRAANPHPTTELNFNSPFELLIAVLLSAQATDVGVNKAT 51
Streptococcus_pneumoniae/1-209        1 ---------MVLSKKRARKVLEEIIALFPDAKPSLDFTNHFELLVAVMLSAQTTDAAVNKAT 53

Escherichia_coli/1-211               52 AKLYPVANTPAAMLELGVEGVKTYIKTIGLYNSKAENIIKTCRILLEQHNGEVPEDRAALEA 113
Mycobacterium_tuberculosis/1-245     63 PALFARYRTARDYAQADRTELESLIRPTGFYRNKAASLIGLGQALVERFGGEVPATMDKLVT 124
Bacillus_anthracis/1-215             53 KNLFQKYKTPEDYLSVSLEELQQDIRSIGLYRNKAKNIQKLCRMLLDDYNGEVPKDRDELTK 114
Neisseria_meningitidis/1-209         52 AKLFPVADTPQAMLDLGLDGVMEYTKTIGLYKTKSKHIMQTCRILLEKYNGEVPEDREALES 113
Streptococcus_pneumoniae/1-209       54 PGLFVAFPTPQAMSVATESEIASHISRLGLYRNKAKFLKKCAQQLLDDFDGQVPQTREELES 115

Escherichia_coli/1-211              114 LPGVGRKTANVVLNTAFGWPTIAVDTHIFRVCNRTQFAPGKN-VEQVEEKLLKVVPAEFKVD 174
Mycobacterium_tuberculosis/1-245    125 LPGVGRKTANVILGNAFGIPGITVDTHFGRLVRRWRWTTAED-PVKVEQAVGELIERKEWTL 185
Bacillus_anthracis/1-215            115 LPGVGRKTANVVVSVAFGIPAIAVDTHVERVSKRLAICRWKDSVLEVEKTLMKKIPMDEWSV 176
Neisseria_meningitidis/1-209        114 LPGVGRKTANVVLNTAFGHPVMAVDTHIFRVSNRTKIAPGKD-VREVEDKLMRFIPKEFLMD 174
Streptococcus_pneumoniae/1-209      116 LAGVGRKTANVVMSVGFGIPAFAVDTHVERICKHHDIVKKSATPLEVEKRVMDILPPEQWLA 177

Escherichia_coli/1-211              175 CHHWLILHGRYTCIARKPRCGSCIIEDLCEYKEKVDI-----------------------   211
Mycobacterium_tuberculosis/1-245    186 LSHRVIFHGRRVCHARRPACGVCVLAKDCPSFGLGPTEPLLAAPLVQGPETDHLLALAGL   245
Bacillus_anthracis/1-215            177 THHRMIFFGRYHCKAQRPQCEECPLLEVCREGKKRMKGK---------------------   215
Neisseria_meningitidis/1-209        176 AHHWLILHGRYTCKALKPQCQTCIINDLCEYPAKA-------------------------   209
Streptococcus_pneumoniae/1-209      178 AHQAMIYFGRAICHPKNPECDQYP--QLYDFSNL--------------------------   209
```

Exercise 2

a)

Vertebrate sequences in Refseq were searched with BLAST using *E.coli* Nth as a query.

b)

Here are the entries for the Nth and MutY proteins in the selected organisms identified using BLAST using *E. coli* Nth as the query:

```
>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
```

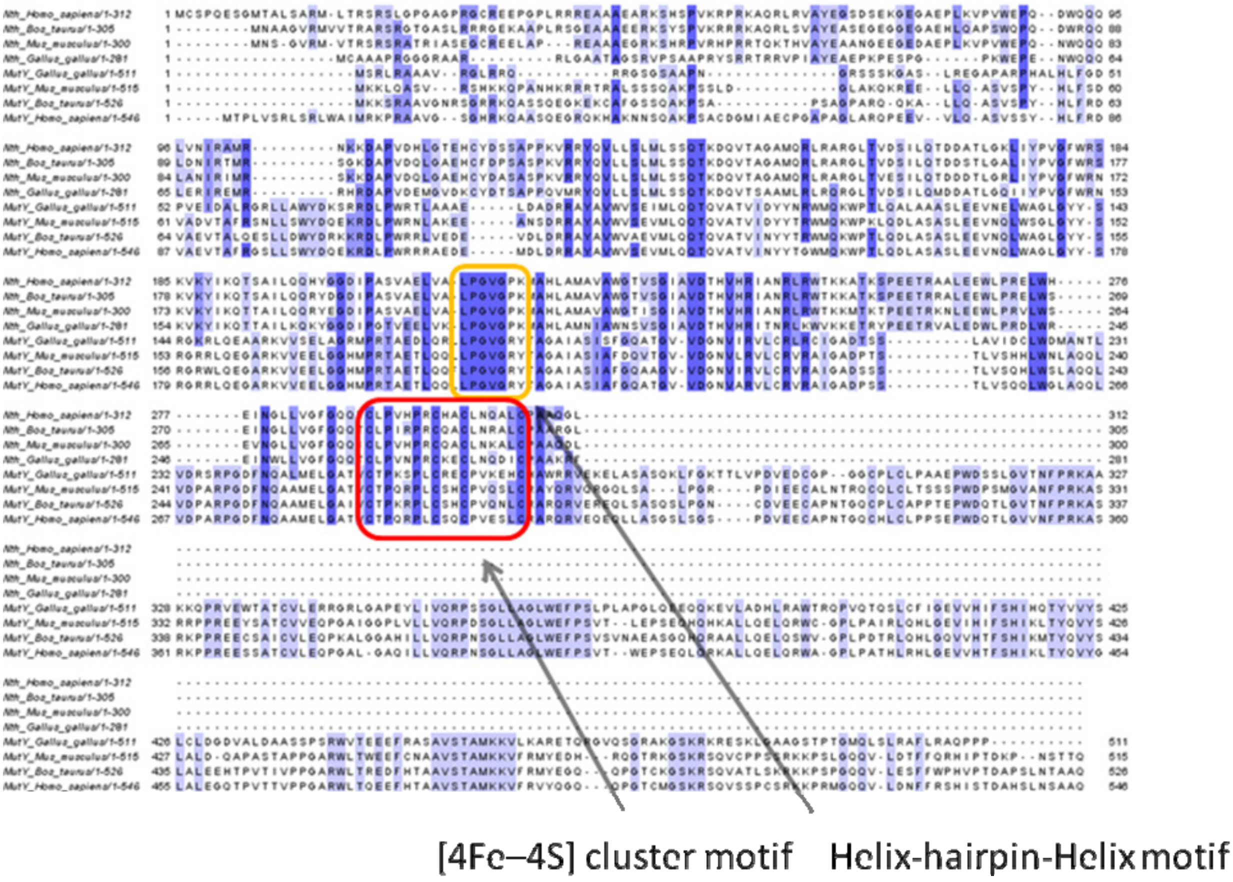
```
>gi|227908769|ref|NP_032769.2| endonuclease III-like protein 1 [Mus musculus]
MNSGVRMVTRSRSRATRIASEGCREELAPREAAAEGRKSHRPVRHPRRTQKTHVAYEAANGEEGEDAEPL
KVPVWEPQNWQQQLANIRIMRSKKDAPVDQLGAEHCYDASASPKVRRYQVLLSLMLSSQTKDQVTAGAMQ
RLRARGLTVESILQTDDDTLGRLIYPVGFWRNKVKYIKQTTAILQQRYEGDIPASVAELVALPGVGPKMA
HLAMAVAWGTISGIAVDTHVHRIANRLRWTKKMTKTPEETRKNLEEWLPRVLWSEVNGLLVGFGQQICLP
VHPRCQACLNKALCPAAQDL
>gi|227330621|ref|NP_573513.2| A/G-specific adenine DNA glycosylase [Mus musculus]
MKKLQASVRSHKKQPANHKRRRTRALSSSQAKPSSLDGLAKQKREELLQASVSPYHLFSDVADVTAFRSN
LLSWYDQEKRDLPWRNLAKEEANSDRRAYAVWVSEVMLQQTQVATVIDYYTRWMQKWPKLQDLASASLEE
VNQLWSGLGYYSRGRRLQEGARKVVEELGGHMPRTAETLQQLLPGVGRYTAGAIASIAFDQVTGVVDGNV
LRVLCRVRAIGADPTSTLVSHHLWNLAQQLVDPARPGDFNQAAMELGATVCTPQRPLCSHCPVQSLCRAY
QRVQRGQLSALPGRPDIEECALNTRQCQLCLTSSSPWDPSMGVANFPRKASRRPPREEYSATCVVEQPGA
IGGPLVLLVQRPDSGLLAGLWEFPSVTLEPSEQHQHKALLQELQRWCGPLPAIRLQHLGEVIHIFSHIKL
TYQVYSLALDQAPASTAPPGARWLTWEEFCNAAVSTAMKKVFRMYEDHRQGTRKGSKRSQVCPPSSRKKP
SLGQQVLDTFFQRHIPTDKPNSTTQ
>gi|114051958|ref|NP_001039862.1| endonuclease III-like protein 1 [Bos taurus]
MNAAGVRMVVTRARSRGTGASLRRRGEKAAPLRSGEAAAEERKSYSPVKRRRKAQRLSVAYEASEGEGGE
GAEHLQAPSWQPQDWRQQLDNIRTMRSGKDAPVDQLGAEHCFDPSASPKVRRYQVLLSLMLSSQTKDQVT
AGAMQRLRARGLTVDSILQTDDSTLGALIYPVGFWRSKVKYIKQTSAILQQRYDGDIPASVAELVALPGV
GPKMAHLAMAVAWGTVSGIAVDTHVHRIANRLRWTKKATKSPEETRRALEEWLPRELWSEINGLLVGFGQ
QTCLPIRPRCQACLNRALCPAARGL
>gi|281485563|ref|NP_001039600.2| A/G-specific adenine DNA glycosylase [Bos taurus]
MKKSRAAVGNRSGRRKQASSQEGKEKCAFGSSQAKPSAPSAGPARQQKALLQASVSPYHLFRDVAEVTAL
QESLLDWYDRKKRDLPWRRLVEDEVDLDRRAYAVWVAEVMLQQTQVATVINYYTRWMQKWPTLQDLASAS
LEEVNQLWAGLGYYSRGRWLQEGARKVVEELGGHMPRTAETLQQFLPGVGRYTAGAIASIAFGQAAGVVD
GNVIRVLCRVRAIGADSSSTLVSQHLWSLAQQLVDPARPGDFNQAAMELGAIVCTPKRPLCSHCPVQNLC
RARQRVEREQLSASQSLPGNCDVEECAPNTGQCPLCAPPTEPWDQTLGVTNFPRKASRKPPREECSAICV
LEQPKALGGAHILLVQRPNSGLLAGLWEFPSVSVNAEASGQHQRAALLQELQSWVGPLPDTRLQHLGQVV
HTFSHIKMTYQVYSLALEEHTPVTIVPPGARWLTREDFHTAAVSTAMKKVFRMYEGQQPGTCKGSKRSQV
ATLSKRKKPSPGQQVLESFFWPHVPTDAPSLNTAAQ
>gi|118601744|ref|NP_001073043.1| endonuclease III-like protein 1 [Gallus gallus]
MCAAAPRGGGRAARRLGAATAGSRVPSAAPRYSRRTRRVPIAYEAEPKPESPGPKWEPENWQQQLERIRE
MRRHRDAPVDEMGVDKCYDTSAPPQVMRYQVLLSLMLSSQTKDQVTSAAMLRLRQRGLTVDSILQMDDAT
LGQIIYPVGFWRNKVKYIKQTTAILKQKYGGDIPGTVEELVKLPGVGPKMAHLAMNIAWNSVSGIAVDTH
VHRITNRLKWVKKETRYPEETRVALEDWLPRDLWREINWLLVGFGQQTCLPVNPRCKECLNQDICPAAKR
F
>gi|118094461|ref|XP_422433.2| PREDICTED: A/G-specific adenine DNA glycosylase
[Gallus gallus]
MSRLRAAAVRGLRRQRRGSGSAAPNGRSSSKGASLREGAPARPHALHLFGDPVEIDALRGRLLAWYDKSR
RDLPWRTLAAAELDADRRAYAVWVSEIMLQQTQVATVIDYYNRWMQKWPTLQALAAASLEEVNELWAGLG
YYSRGKRLQEAARKVVSELAGRMPRTAEDLQRLLPGVGRYTAGAIASISFGQATGVVDGNVIRVLCRLRC
IGADTSSLAVIDCLWDMANTLVDRSRPGDFNQALMELGATVCTPKSPLCRECPVKEHCHAWRRVEKELAS
ASQKLFGKTTLVPDVEDCGPGGCPLCLPAAEPWDSSLGVTNFPRKAAKKQPRVEWTATCVLERRGRLGAP
EYLIVQRPSSGLLAGLWEFPSLPLAPGLQEEQQKEVLADHLRAWTRQPVQTQSLCFIGEVVHIFSHIHQT
YVVYSLCLDGDVALDAASSPSRWVTEEEFRASAVSTAMKKVLKARETQRGVQSGRAKGSKRKRESKLGAA
GSTPTGMQLSLRAFLRAQPPP
>gi|113205550|ref|NP_001037884.1| nth endonuclease III-like 1 [Xenopus (Silurana)
tropicalis]
MSGSLRPLGRRGRRGVLKAVGGKDQQDGTSKGQVIDDSEDEKPSSPKERSKRRVSVEYEQAASETVAKRP
KWQPKNWAQHLENIRQMRSRRDAPVDQMGAEKCYDQNAAPEVMRYQILLSLMLSSQTKDQVTSAAMCRLR
QHGLTVSRILETDDGTLGKLIYPVGFWKNKVKYIKQTTEILQEKYGGDIPDNVTDLVKLPGVGPKMAHLV
MDIAWNNVSGIGVDTHVHRISNRLKWVRKETKTPEETRVAMEDWMPRELWSEINWLLVGFGQQVCLPVSP
RCSECLNKDICPGAKKKKPR
>gi|118403607|ref|NP_001072831.1| mutY homolog [Xenopus (Silurana) tropicalis]
MPPPRTKTSLGRSAAASGKRKSPKQAFPKREEHVLQSSIYHSFTSQETEIIRDKLLAWYDKSKRDLPWRT
MACTEPDLDRKAYAVWVSEVMLQQTQVATVIDYYNKWMKVWPTMEDLARSSLEEVNEMWSGLGYYSRGRR
LQEGAKKVVLELGGSMPRSADELQKLLPGVGRYTAGAIASISYGQVTGVVDGNVIRVLSRLRCIGADSST
LAVSDKLWNLANALVDPDRPGDFNQGMMELGATVCTPKKPLCTACPLQGQCKAYLKVIAEKESAVKTLIK
KQASPIAKDVGDIEDCDLGPGLCALCVPTSDPWDSSLGVANFPRKSAKKPSRMEQTAICVWEKCGDHGEL
EYLIVQRPSSGLLAGLWEFPSILLDEKFTEQNRQHSLLGLLQDLSGHAVPLQKLQYKGEVVHIFSHIHQT
YVVYFLSLNTTENCSVKTEETERPLTRWVTKKEFLNSAVPTAMKKIMKLCESHGSSCTAVNTSKKRKGDL
AKVQLPSGRIKTEKGKQQSIQSFFKLATEK
```

Here are the entries for the Nth and MutY proteins in selected organisms, with short titles:

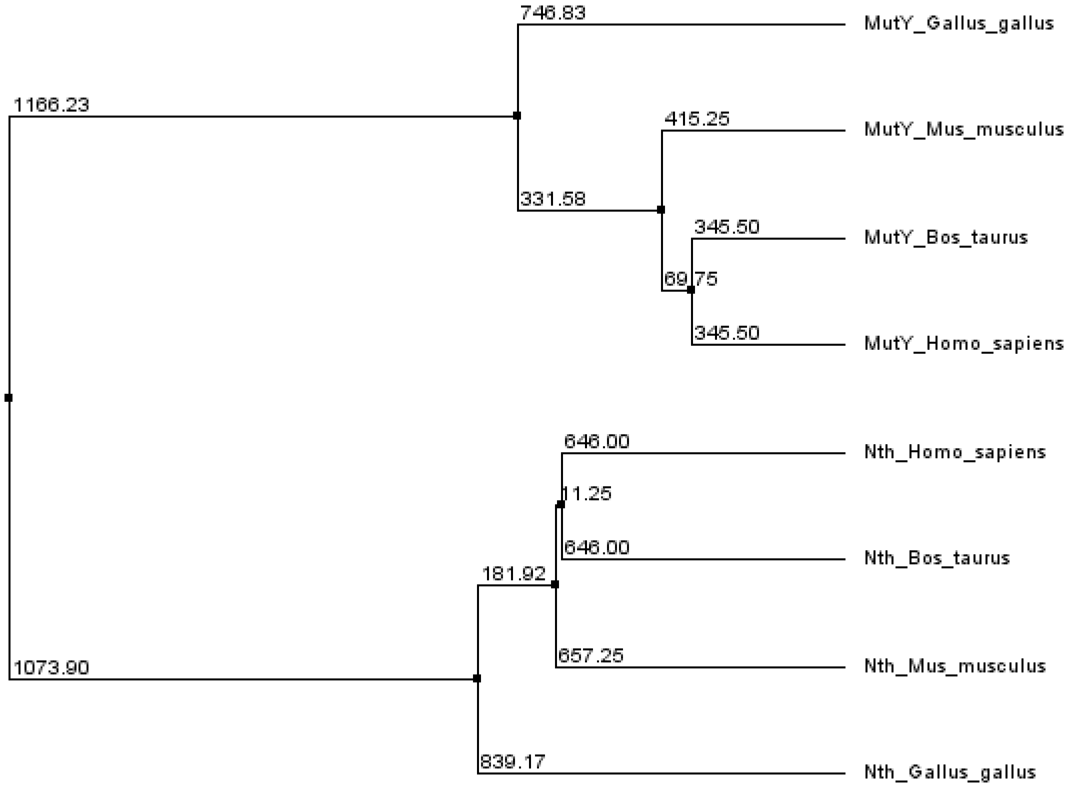
```
>Nth_Homo_sapiens
MCSPQESGMTALSARMLTRSRSLGPGAGPRGCREEPGPLRRREAAAEARKSHSPVKRPRKAQRLRVAYEG
SDSEKGEGAEPLKVPVWEPQDWQQQLVNIRAMRNKKDAPVDHLGTEHCYDSSAPPKVRRYQVLLSLMLSS
QTKDQVTAGAMQRLRARGLTVDSILQTDDATLGKLIYPVGFWRSKVKYIKQTSAILQQHYGGDIPASVAE
LVALPGVGPKMAHLAMAVAWGTVSGIAVDTHVHRIANRLRWTKKATKSPEETRAALEEWLPRELWHEING
LLVGFGQQTCLPVHPRCHACLNQALCPAAQGL
>MutY_Homo_sapiens
MTPLVSRLSRLWAIMRKPRAAVGSGHRKQAASQEGRQKHAKNNSQAKPSACDGMIAECPGAPAGLARQPE
EVVLQASVSSYHLFRDVAEVTAFRGSLLSWYDQEKRDLPWRRRAEDEMDLDRRAYAVWVSEVMLQQTQVA
TVINYYTGWMQKWPTLQDLASASLEEVNQLWAGLGYYSRGRRLQEGARKVVEELGGHMPRTAETLQQLLP
GVGRYTAGAIASIAFGQATGVVDGNVARVLCRVRAIGADPSSTLVSQQLWGLAQQLVDPARPGDFNQAAM
ELGATVCTPQRPLCSQCPVESLCRARQRVEQEQLLASGSLSGSPDVEECAPNTGQCHLCLPPSEPWDQTL
GVVNFPRKASRKPPREESSATCVLEQPGALGAQILLVQRPNSGLLAGLWEFPSVTWEPSEQLQRKALLQE
LQRWAGPLPATHLRHLGEVVHTFSHIKLTYQVYGLALEGQTPVTTVPPGARWLTQEEFHTAAVSTAMKKV
FRVYQGQQPGTCMGSKRSQVSSPCSRKKPRMGQQVLDNFFRSHISTDAHSLNSAAQ
>Nth_Mus_musculus
MNSGVRMVTRSRSRATRIASEGCREELAPREAAAEGRKSHRPVRHPRRTQKTHVAYEAANGEEGEDAEPL
KVPVWEPQNWQQQLANIRIMRSKKDAPVDQLGAEHCYDASASPKVRRYQVLLSLMLSSQTKDQVTAGAMQ
RLRARGLTVESILQTDDDTLGRLIYPVGFWRNKVKYIKQTTAILQQRYEGDIPASVAELVALPGVGPKMA
HLAMAVAWGTISGIAVDTHVHRIANRLRWTKKMTKTPEETRKNLEEWLPRVLWSEVNGLLVGFGQQICLP
VHPRCQACLNKALCPAAQDL
>MutY_Mus_musculus
MKKLQASVRSHKKQPANHKRRRTRALSSSQAKPSSLDGLAKQKREELLQASVSPYHLFSDVADVTAFRSN
LLSWYDQEKRDLPWRNLAKEEANSDRRAYAVWVSEVMLQQTQVATVIDYYTRWMQKWPKLQDLASASLEE
VNQLWSGLGYYSRGRRLQEGARKVVEELGGHMPRTAETLQQLLPGVGRYTAGAIASIAFDQVTGVVDGNV
LRVLCRVRAIGADPTSTLVSHHLWNLAQQLVDPARPGDFNQAAMELGATVCTPQRPLCSHCPVQSLCRAY
QRVQRGQLSALPGRPDIEECALNTRQCQLCLTSSSPWDPSMGVANFPRKASRRPPREEYSATCVVEQPGA
IGGPLVLLVQRPDSGLLAGLWEFPSVTLEPSEQHQHKALLQELQRWCGPLPAIRLQHLGEVIHIFSHIKL
TYQVYSLALDQAPASTAPPGARWLTWEEFCNAAVSTAMKKVFRMYEDHRQGTRKGSKRSQVCPPSSRKKP
SLGQQVLDTFFQRHIPTDKPNSTTQ
>Nth_Bos_taurus
MNAAGVRMVVTRARSRGTGASLRRRGEKAAPLRSGEAAAEERKSYSPVKRRRKAQRLSVAYEASEGEGGE
GAEHLQAPSWQPQDWRQQLDNIRTMRSGKDAPVDQLGAEHCFDPSASPKVRRYQVLLSLMLSSQTKDQVT
AGAMQRLRARGLTVDSILQTDDSTLGALIYPVGFWRSKVKYIKQTSAILQQRYDGDIPASVAELVALPGV
GPKMAHLAMAVAWGTVSGIAVDTHVHRIANRLRWTKKATKSPEETRRALEEWLPRELWSEINGLLVGFGQ
QTCLPIRPRCQACLNRALCPAARGL
>MutY_Bos_taurus
MKKSRAAVGNRSGRRKQASSQEGKEKCAFGSSQAKPSAPSAGPARQQKALLQASVSPYHLFRDVAEVTAL
QESLLDWYDRKKRDLPWRRLVEDEVDLDRRAYAVWVAEVMLQQTQVATVINYYTRWMQKWPTLQDLASAS
LEEVNQLWAGLGYYSRGRWLQEGARKVVEELGGHMPRTAETLQQFLPGVGRYTAGAIASIAFGQAAGVVD
GNVIRVLCRVRAIGADSSSTLVSQHLWSLAQQLVDPARPGDFNQAAMELGAIVCTPKRPLCSHCPVQNLC
RARQRVEREQLSASQSLPGNCDVEECAPNTGQCPLCAPPTEPWDQTLGVTNFPRKASRKPPREECSAICV
LEQPKALGGAHILLVQRPNSGLLAGLWEFPSVSVNAEASGQHQRAALLQELQSWVGPLPDTRLQHLGQVV
HTFSHIKMTYQVYSLALEEHTPVTIVPPGARWLTREDFHTAAVSTAMKKVFRMYEGQQPGTCKGSKRSQV
ATLSKRKKPSPGQQVLESFFWPHVPTDAPSLNTAAQ
>Nth_Gallus_gallus
MCAAAPRGGGRAARRLGAATAGSRVPSAAPRYSRRTRRVPIAYEAEPKPESPGPKWEPENWQQQLERIRE
MRRHRDAPVDEMGVDKCYDTSAPPQVMRYQVLLSLMLSSQTKDQVTSAAMLRLRQRGLTVDSILQMDDAT
LGQIIYPVGFWRNKVKYIKQTTAILKQKYGGDIPGTVEELVKLPGVGPKMAHLAMNIAWNSVSGIAVDTH
VHRITNRLKWVKKETRYPEETRVALEDWLPRDLWREINWLLVGFGQQTCLPVNPRCKECLNQDICPAAKR
F
>MutY_Gallus_gallus
MSRLRAAAVRGLRRQRRGSGSAAPNGRSSSKGASLREGAPARPHALHLFGDPVEIDALRGRLLAWYDKSR
RDLPWRTLAAAELDADRRAYAVWVSEIMLQQTQVATVIDYYNRWMQKWPTLQALAAASLEEVNELWAGLG
YYSRGKRLQEAARKVVSELAGRMPRTAEDLQRLLPGVGRYTAGAIASISFGQATGVVDGNVIRVLCRLRC
IGADTSSLAVIDCLWDMANTLVDRSRPGDFNQALMELGATVCTPKSPLCRECPVKEHCHAWRRVEKELAS
ASQKLFGKTTLVPDVEDCGPGGCPLCLPAAEPWDSSLGVTNFPRKAAKKQPRVEWTATCVLERRGRLGAP
EYLIVQRPSSGLLAGLWEFPSLPLAPGLQEEQQKEVLADHLRAWTRQPVQTQSLCFIGEVVHIFSHIHQT
YVVYSLCLDGDVALDAASSPSRWVTEEEFRASAVSTAMKKVLKARETQRGVQSGRAKGSKRKRESKLGAA
GSTPTGMQLSLRAFLRAQPPP
```

c)

Below is a multiple alignment of the vertebrate Nth and MutY homologs, as produced by MAFFT. The Helix-hairpin-Helix motif and the [4Fe-4S] cluster motif are indicated. The [4Fe-4S] cluster is fully conserved in all sequences, as well as the Helix-hairpin-helix motif, except for the final K in the HhH-motif which is not conserved in the MutY sequences.



d)



Based on the phylogenetic tree above, it is clear that human Nth and chicken Nth are more similar than human Nth and human MutY, because Nth and MutY forms seperate lineages near the root of the tree.

Exercise 3

a)

Here are the human NTHL1, MUTYH, OGG1 and MBD4 proteins found by the PSI-BLAST search:

```
>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
```

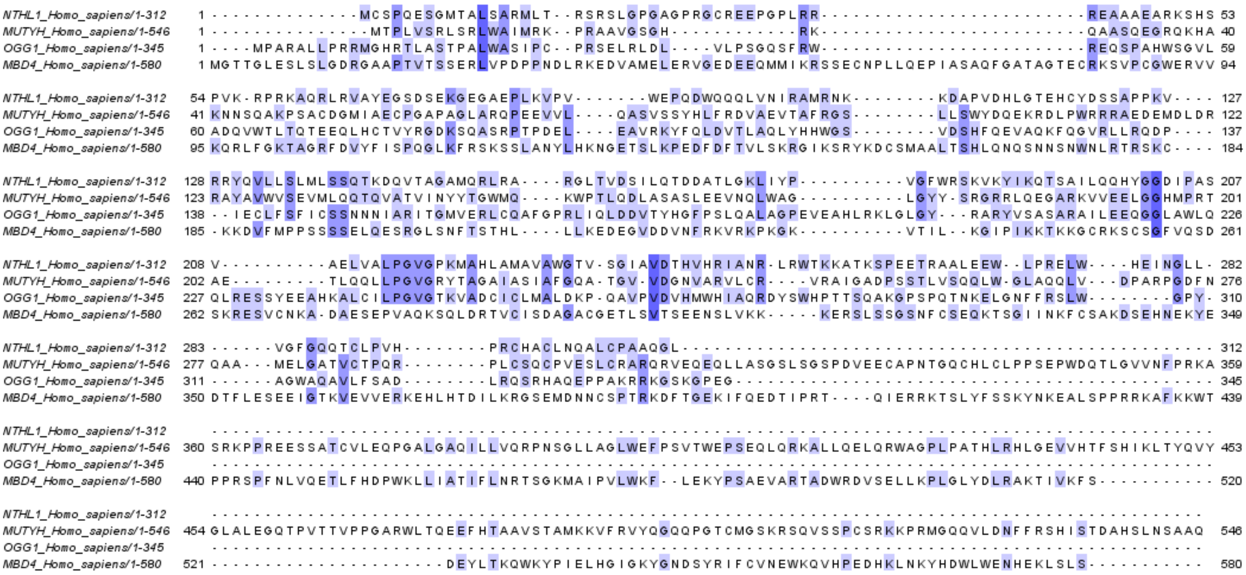
```
>gi|4505495|ref|NP_002533.1| N-glycosylase/DNA lyase isoform 1a [Homo sapiens]
MPARALLPRRMGHRTLASTPALWASIPCPRSELRLDLVLPSGQSFRWREQSPAHWSGVLADQVWTLTQTE
EQLHCTVYRGDKSQASRPTPDELEAVRKYFQLDVTLAQLYHHWGSVDSHFQEVAQKFQGVRLLRQDPIEC
LFSFICSSNNNIARITGMVERLCQAFGPRLIQLDDVTYHGFPSLQALAGPEVEAHLRKLGLGYRARYVSA
SARAILEEQGGLAWLQQLRESSYEEAHKALCILPGVGTKVADCICLMALDKPQAVPVDVHMWHIAQRDYS
WHPTTSQAKGPSPQTNKELGNFFRSLWGPYAGWAQAVLFSADLRQSRHAQEPPAKRRKGSKGPEG

>gi|4505121|ref|NP_003916.1| methyl-CpG-binding domain protein 4 [Homo sapiens]
MGTTGLESLSLGDRGAAPTVTSSERLVPDPPNDLRKEDVAMELERVGEDEEQMMIKRSSECNPLLQEPIA
SAQFGATAGTECRKSVPCGWERVVKQRLFGKTAGRFDVYFISPQGLKFRSKSSLANYLHKNGETSLKPED
FDFTVLSKRGIKSRYKDCSMAALTSHLQNQSNNSNWNLRTRSKCKKDVFMPPSSSSELQESRGLSNFTST
HLLLKEDEGVDDVNFRKVRKPKGKVTILKGIPIKKTKKGCRKSCSGFVQSDSKRESVCNKADAESEPVAQ
KSQLDRTVCISDAGACGETLSVTSEENSLVKKKERSLSSGSNFCSEQKTSGIINKFCSAKDSEHNEKYED
TFLESEEIGTKVEVVERKEHLHTDILKRGSEMDNNCSPTRKDFTGEKIFQEDTIPRTQIERRKTSLYFSS
KYNKEALSPPRRKAFKKWTPPRSPFNLVQETLFHDPWKLLIATIFLNRTSGKMAIPVLWKFLEKYPSAEV
ARTADWRDVSELLKPLGLYDLRAKTIVKFSDEYLTKQWKYPIELHGIGKYGNDSYRIFCVNEWKQVHPED
HKLNKYHDWLWENHEKLSLS
```

Here are the same sequences with short descriptions:

```
>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
```

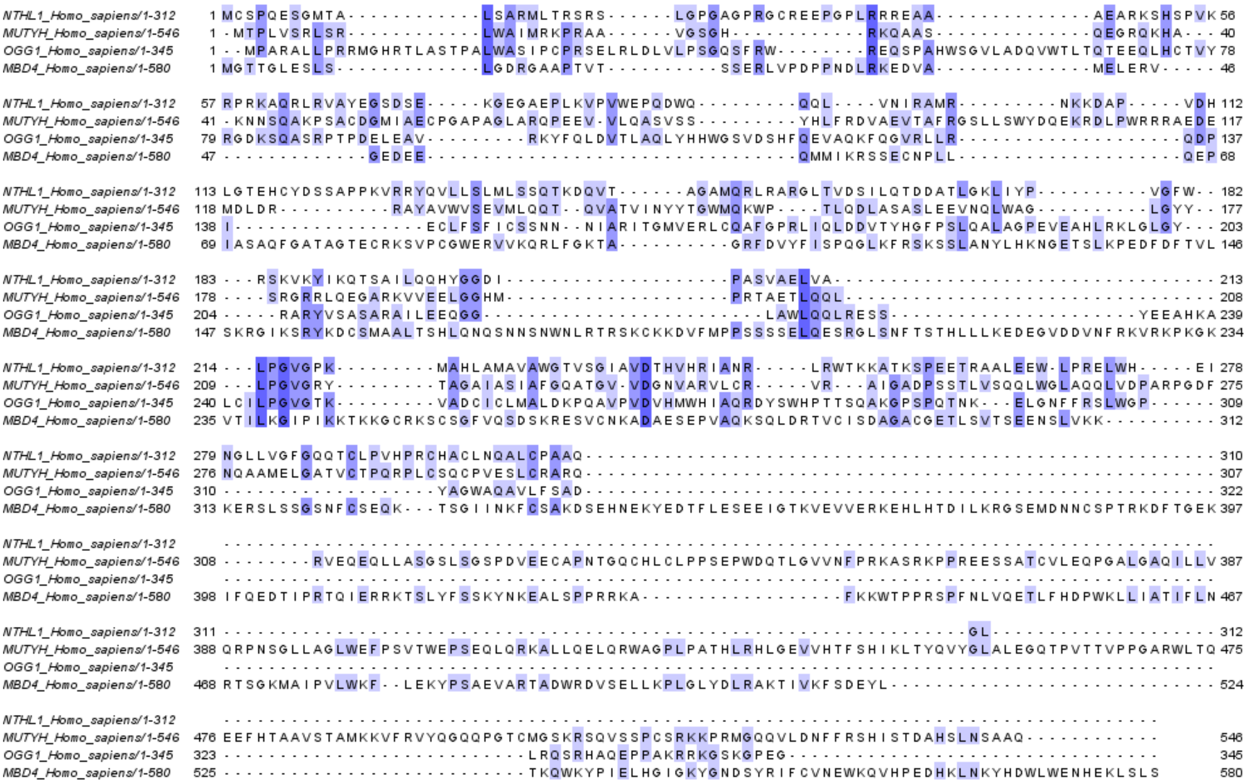
b)

MUSCLE alignment of the human proteins:



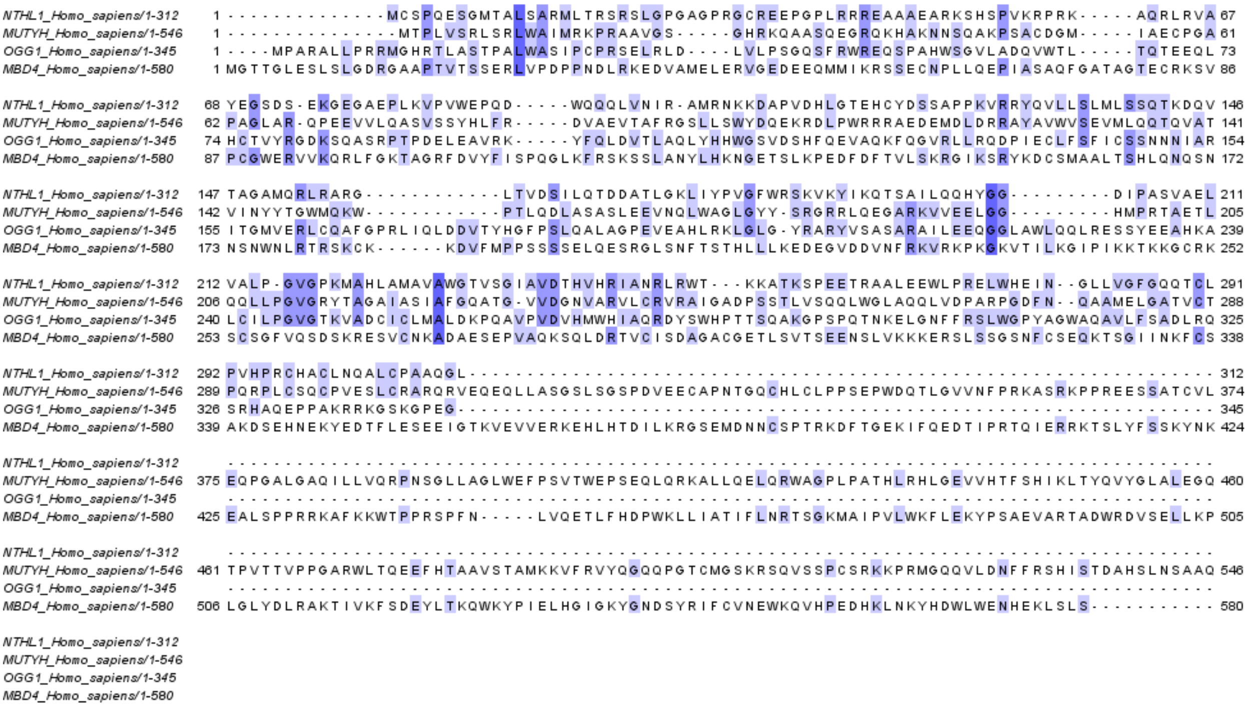
```
NTHL1_Homo_sapiens/1-312    1 --------------MCSPQESGMTALSARMLT--RSRSLGPGAGPRGCREEPGPLRR-------------------------REAAAEARKSHS 53
MUTYH_Homo_sapiens/1-546    1 ---------------MTPLVSRLSRLWAIMRK--PRAAVGSGH------------RK-------------------------QAASQEGRQKHA 40
OGG1_Homo_sapiens/1-345     1 ----MPARALLPRRMGHRTLASTPALWASIPC--PRSELRLDL----VLPSGQSFRW-------------------------REQSPAHWSGVL 59
MBD4_Homo_sapiens/1-580     1 MGTTGLESLSLGDRGAAPTVTSSERLVPDPPNDLRKEDVAMELERVGEDEEQMMIKRSSECNPLLQEPIASAQFGATAGTECRKSVPCGWERVV 94

NTHL1_Homo_sapiens/1-312   54 PVK-RPRKAQRLRVAYEGSDSEKGEGAEPLKVPV-------WEPQDWQQQLVNIRAMRNK--------KDAPVDHLGTEHCYDSSAPPKV---- 127
MUTYH_Homo_sapiens/1-546   41 KNNSQAKPSACDGMIAECPGAPAGLARQPEEVVL----QASVSSYHLFRDVAEVTAFRGS--------LLSWYDQEKRDLPWRRRAEDEMDLDR 122
OGG1_Homo_sapiens/1-345    60 ADQVWTLTQTEEQLHCTVYRGDKSQASRPTPDEL----EAVRKYFQLDVTLAQLYHHWGS--------VDSHFQEVAQKFQGVRLLRQDP---- 137
MBD4_Homo_sapiens/1-580    95 KQRLFGKTAGRFDVYFISPQGLKFRSKSSLANYLHKNGETSLKPEDFDFTVLSKRGIKSRYKDCSMAALTSHLQNQSNNSNWNLRTRSKC---- 184

NTHL1_Homo_sapiens/1-312  128 RRYQVLLSLMLSSQTKDQVTAGAMQRLRA----RGLTVDSILQTDDATLGKLIYP----------VGFWRSKVKYIKQTSAILQQHYGGDIPAS 207
MUTYH_Homo_sapiens/1-546  123 RAYAVWVSEVMLQQTQVATVINYYTGWMQ----KWPTLQDLASASLEEVNQLWAG----------LGYY-SRGRRLQEGARKVVEELGGHMPRT 201
OGG1_Homo_sapiens/1-345   138 --IECLFSFICSSNNNIARITGMVERLCQAFGPRLIQLDDVTYHGFPSLQALAGPEVEAHLRKLGLGY---RARYVSASARAILEEQGGLAWLQ 226
MBD4_Homo_sapiens/1-580   185 -KKDVFMPPSSSSELQESRGLSNFTSTHL----LLKEDEGVDDVNFRKVRKPKGK----------VTIL--KGIPIKKTKKGCRKSCSGFVQSD 261

NTHL1_Homo_sapiens/1-312  208 V----------AELVALPGVGPKMAHLAMAVAWGTV-SGIAVDTHVHRIANR-LRWTKKATKSPEETRAALEEW--LPRELW----HEINGLL- 282
MUTYH_Homo_sapiens/1-546  202 AE---------TLQQLLPGVGRYTAGAIASIAFGQA-TGV-VDGNVARVLCR----VRAIGADPSSTLVSQQLW-GLAQQLV---DPARPGDFN 276
OGG1_Homo_sapiens/1-345   227 QLRESSYEEAHKALCILPGVGTKVADCICLMALDKP-QAVPVDVHMWHIAQRDYSWHPTTSQAKGPSPQTNKELGNFFRSLW--------GPY- 310
MBD4_Homo_sapiens/1-580   262 SKRESVCNKA-DAESEPVAQKSQLDRTVCISDAGACGETLSVTSEENSLVKK-----KERSLSSGSNFCSEQKTSGIINKFCSAKDSEHNEKYE 349

NTHL1_Homo_sapiens/1-312  283 ------VGFGQQTCLPVH--------PRCHACLNQALCPAAQGL-------------------------------------------------- 312
MUTYH_Homo_sapiens/1-546  277 QAA---MELGATVCTPQR--------PLCSQCPVESLCRARQRVEQEQLLASGSLSGSPDVEECAPNTGQCHLCLPPSEPWDQTLGVVNFPRKA 359
OGG1_Homo_sapiens/1-345   311 ------AGWAQAVLFSAD--------LRQSRHAQEPPAKRRKGSKGPEG--------------------------------------------- 345
MBD4_Homo_sapiens/1-580   350 DTFLESEEIGTKVEVVERKEHLHTDILKRGSEMDNNCSPTRKDFTGEKIFQEDTIPRT----QIERRKTSLYFSSKYNKEALSPPRRKAFKKWT 439

NTHL1_Homo_sapiens/1-312      ----------------------------------------------------------------------------------------------
MUTYH_Homo_sapiens/1-546  360 SRKPPREESSATCVLEQPGALGAQILLVQRPNSGLLAGLWEFPSVTWEPSEQLQRKALLQELQRWAGPLPATHLRHLGEVVHTFSHIKLTYQVY 453
OGG1_Homo_sapiens/1-345       ----------------------------------------------------------------------------------------------
MBD4_Homo_sapiens/1-580   440 PPRSPFNLVQETLFHDPWKLLIATIFLNRTSGKMAIPVLWKF--LEKYPSAEVARTADWRDVSELLKPLGLYDLRAKTIVKFS----------- 520

NTHL1_Homo_sapiens/1-312      ---------------------------------------------------------------------------------------------
MUTYH_Homo_sapiens/1-546  454 GLALEGQTPVTTVPPGARWLTQEEFHTAAVSTAMKKVFRVYQGQQPGTCMGSKRSQVSSPCSRKKPRMGQQVLDNFFRSHISTDAHSLNSAAQ 546
OGG1_Homo_sapiens/1-345       ---------------------------------------------------------------------------------------------
MBD4_Homo_sapiens/1-580   521 ----------------------DEYLTKQWKYPIELHGIGKYGNDSYRIFCVNEWKQVHPEDHKLNKYHDWLWENHEKLSLS----------- 580
```

MAFFT alignment:



```
NTHL1_Homo_sapiens/1-312    1 MCSPQESGMTA------------LSARMLTRSRS------LGPGAGPRGCREEPGPLRRREAA--------------AEARKSHSPVK 56
MUTYH_Homo_sapiens/1-546    1 -MTPLVSRLSR------------LWAIMRKPRAA------VGSGH------------RKQAAS--------------QEGRQKHA--- 40
OGG1_Homo_sapiens/1-345     1 --MPARALLPRRMGHRTLASTPALWASIPCPRSELRLDLVLPSGQSFRW--------REQSPAHWSGVLADQVWTLTQTEEQLHCTVY 78
MBD4_Homo_sapiens/1-580     1 MGTTGLESLS-------------LGDRGAAPTVT----------SSERLVPDPPNDLRKEDVA--------------MELERV----- 46

NTHL1_Homo_sapiens/1-312   57 RPRKAQRLRVAYEGSDSE-----KGEGAEPLKVPVWEPQDWQ---------QQL----VNIRAMR---------NKKDAP-----VDH 112
MUTYH_Homo_sapiens/1-546   41 -KNNSQAKPSACDGMIAECPGAPAGLARQPEEV-VLQASVSS---------YHLFRDVAEVTAFRGSLLSWYDQEKRDLPWRRRAEDE 117
OGG1_Homo_sapiens/1-345    79 RGDKSQASRPTPDELEAV---------RKYFQLDVTLAQLYHHWGSVDSHFQEVAQKFQGVRLLR--------------------QDP 137
MBD4_Homo_sapiens/1-580    47 -------------GEDEE---------------------------------QMMIKRSSECNPLL--------------------QEP 68

NTHL1_Homo_sapiens/1-312  113 LGTEHCYDSSAPPKVRRYQVLLSLMLSSQTKDQVT------AGAMQRLRARGLTVDSILQTDDATLGKLIYP----------VGFW-- 182
MUTYH_Homo_sapiens/1-546  118 MDLDR----------RAYAVWVSEVMLQQT--QVATVINYYTGWMQKWP----TLQDLASASLEEVNQLWAG----------LGYY-- 177
OGG1_Homo_sapiens/1-345   138 I-----------------ECLFSFICSSNN--NIARITGMVERLCQAFGPRLIQLDDVTYHGFPSLQALAGPEVEAHLRKLGLGY--- 203
MBD4_Homo_sapiens/1-580    69 IASAQFGATAGTECRKSVPCGWERVVKQRLFGKTA----------GRFDVYFISPQGLKFRSKSSLANYLHKNGETSLKPEDFDFTVL 146

NTHL1_Homo_sapiens/1-312  183 ---RSKVKYIKQTSAILQQHYGGDI--------------------PASVAELVA---------------------------------- 213
MUTYH_Homo_sapiens/1-546  178 ----SRGRRLQEGARKVVEELGGHM--------------------PRTAETLQQL--------------------------------- 208
OGG1_Homo_sapiens/1-345   204 -----RARYVSASARAILEEQGG-------------------------LAWLQQLRESS----------------------YEEAHKA 239
MBD4_Homo_sapiens/1-580   147 SKRGIKSRYKDCSMAALTSHLQNQSNNSNWNLRTRSKCKKDVFMPPSSSSELQESRGLSNFTSTHLLLKEDEGVDDVNFRKVRKPKGK 234

NTHL1_Homo_sapiens/1-312  214 ---LPGVGPK---------MAHLAMAVAWGTVSGIAVDTHVHRIANR-----LRWTKKATKSPEETRAALEEW-LPRELWH-----EI 278
MUTYH_Homo_sapiens/1-546  209 ---LPGVGRY---------TAGAIASIAFGQATGV-VDGNVARVLCR-----VR---AIGADPSSTLVSQQLWGLAQQLVDPARPGDF 275
OGG1_Homo_sapiens/1-345   240 LCILPGVGTK---------VADCICLMALDKPQAVPVDVHMWHIAQRDYSWHPTTSQAKGPSPQTNK---ELGNFFRSLWGP------ 309
MBD4_Homo_sapiens/1-580   235 VTILKGIPIKKTKKGCRKSCSGFVQSDSKRESVCNKADAESEPVAQKSQLDRTVCISDAGACGETLSVTSEENSLVKK---------- 312

NTHL1_Homo_sapiens/1-312  279 NGLLVGFGQQTCLPVHPRCHACLNQALCPAAQ-------------------------------------------------------- 310
MUTYH_Homo_sapiens/1-546  276 NQAAMELGATVCTPQRPLCSQCPVESLCRARQ-------------------------------------------------------- 307
OGG1_Homo_sapiens/1-345   310 -------------------YAGWAQAVLFSAD-------------------------------------------------------- 322
MBD4_Homo_sapiens/1-580   313 KERSLSSGSNFCSEQK---TSGIINKFCSAKDSEHNEKYEDTFLESEEIGTKVEVVERKEHLHTDILKRGSEMDNNCSPTRKDFTGEK 397

NTHL1_Homo_sapiens/1-312      ----------------------------------------------------------------------------------------
MUTYH_Homo_sapiens/1-546  308 --------RVEQEQLLASGSLSGSPDVEECAPNTGQCHLCLPPSEPWDQTLGVVNFPRKASRKPPREESSATCVLEQPGALGAQILLV 387
OGG1_Homo_sapiens/1-345       ----------------------------------------------------------------------------------------
MBD4_Homo_sapiens/1-580   398 IFQEDTIPRTQIERRKTSLYFSSKYNKEALSPPRRKA------------------FKKWTPPRSPFNLVQETLFHDPWKLLIATIFLN 467

NTHL1_Homo_sapiens/1-312  311 ------------------------------------------------------------------GL-------------------- 312
MUTYH_Homo_sapiens/1-546  388 QRPNSGLLAGLWEFPSVTWEPSEQLQRKALLQELQRWAGPLPATHLRHLGEVVHTFSHIKLTYQVYGLALEGQTPVTTVPPGARWLTQ 475
OGG1_Homo_sapiens/1-345       ----------------------------------------------------------------------------------------
MBD4_Homo_sapiens/1-580   468 RTSGKMAIPVLWKF--LEKYPSAEVARTADWRDVSELLKPLGLYDLRAKTIVKFSDEYL----------------------------- 524

NTHL1_Homo_sapiens/1-312      -----------------------------------------------------------------------------------
MUTYH_Homo_sapiens/1-546  476 EEFHTAAVSTAMKKVFRVYQGQQPGTCMGSKRSQVSSPCSRKKPRMGQQVLDNFFRSHISTDAHSLNSAAQ------------ 546
OGG1_Homo_sapiens/1-345   323 ---------------------------LRQSRHAQEPPAKRRKGSKGPEG--------------------------------- 345
MBD4_Homo_sapiens/1-580   525 ---------------------------TKQWKYPIELHGIGKYGNDSYRIFCVNEWKQVHPEDHKLNKYHDWLWENHEKLSLS 580
```

Clustal W alignment:



```
NTHL1_Homo_sapiens/1-312    1 --------------MCSPQESGMTALSARMLTRSRSLGPGAGPRGCREEPGPLRRREAAAEARKSHSPVKRPRK-----AQRLRVA 67
MUTYH_Homo_sapiens/1-546    1 ---------------MTPLVSRLSRLWAIMRKPRAAVGS-----GHRKQAASQEGRQKHAKNNSQAKPSACDGM-----IAECPGA 61
OGG1_Homo_sapiens/1-345     1 ----MPARALLPRRMGHRTLASTPALWASIPCPRSELRLD----LVLPSGQSFRWREQSPAHWSGVLADQVWTL-----TQTEEQL 73
MBD4_Homo_sapiens/1-580     1 MGTTGLESLSLGDRGAAPTVTSSERLVPDPPNDLRKEDVAMELERVGEDEEQMMIKRSSECNPLLQEPIASAQFGATAGTECRKSV 86

NTHL1_Homo_sapiens/1-312   68 YEGSDS-EKGEGAEPLKVPVWEPQD-----WQQQLVNIR-AMRNKKDAPVDHLGTEHCYDSSAPPKVRRYQVLLSLMLSSQTKDQV 146
MUTYH_Homo_sapiens/1-546   62 PAGLAR-QPEEVVLQASVSSYHLFR-----DVAEVTAFRGSLLSWYDQEKRDLPWRRRAEDEMDLDRRAYAVWVSEVMLQQTQVAT 141
OGG1_Homo_sapiens/1-345    74 HCTVYRGDKSQASRPTPDELEAVRK-----YFQLDVTLAQLYHHWGSVDSHFQEVAQKFQGVRLLRQDPIECLFSFICSSNNNIAR 154
MBD4_Homo_sapiens/1-580    87 PCGWERVVKQRLFGKTAGRFDVYFISPQGLKFRSKSSLANYLHKNGETSLKPEDFDFTVLSKRGIKSRYKDCSMAALTSHLQNQSN 172

NTHL1_Homo_sapiens/1-312  147 TAGAMQRLRARG------------LTVDSILQTDDATLGKLIYPVGFWRSKVKYIKQTSAILQQHYGG---------DIPASVAEL 211
MUTYH_Homo_sapiens/1-546  142 VINYYTGWMQKW------------PTLQDLASASLEEVNQLWAGLGYY-SRGRRLQEGARKVVEELGG---------HMPRTAETL 205
OGG1_Homo_sapiens/1-345   155 ITGMVERLCQAFGPRLIQLDDVTYHGFPSLQALAGPEVEAHLRKLGLG-YRARYVSASARAILEEQGGLAWLQQLRESSYEEAHKA 239
MBD4_Homo_sapiens/1-580   173 NSNWNLRTRSKCK------KDVFMPPSSSSELQESRGLSNFTSTHLLLKEDEGVDDVNFRKVRKPKGKVTILKGIPIKKTKKGCRK 252

NTHL1_Homo_sapiens/1-312  212 VALP-GVGPKMAHLAMAVAWGTVSGIAVDTHVHRIANRLRWT---KKATKSPEETRAALEEWLPRELWHEIN--GLLVGFGQQTCL 291
MUTYH_Homo_sapiens/1-546  206 QQLLPGVGRYTAGAIASIAFGQATG-VVDGNVARVLCRVRAIGADPSSTLVSQQLWGLAQQLVDPARPGDFN--QAAMELGATVCT 288
OGG1_Homo_sapiens/1-345   240 LCILPGVGTKVADCICLMALDKPQAVPVDVHMWHIAQRDYSWHPTTSQAKGPSPQTNKELGNFFRSLWGPYAGWAQAVLFSADLRQ 325
MBD4_Homo_sapiens/1-580   253 SCSGFVQSDSKRESVCNKADAESEPVAQKSQLDRTVCISDAGACGETLSVTSEENSLVKKKERSLSSGSNFCSEQKTSGIINKFCS 338

NTHL1_Homo_sapiens/1-312  292 PVHPRCHACLNQALCPAAQGL----------------------------------------------------------------- 312
MUTYH_Homo_sapiens/1-546  289 PQRPLCSQCPVESLCRARQRVEQEQLLASGSLSGSPDVEECAPNTGQCHLCLPPSEPWDQTLGVVNFPRKASRKPPREESSATCVL 374
OGG1_Homo_sapiens/1-345   326 SRHAQEPPAKRRKGSKGPEG------------------------------------------------------------------ 345
MBD4_Homo_sapiens/1-580   339 AKDSEHNEKYEDTFLESEEIGTKVEVVERKEHLHTDILKRGSEMDNNCSPTRKDFTGEKIFQEDTIPRTQIERRKTSLYFSSKYNK 424

NTHL1_Homo_sapiens/1-312      --------------------------------------------------------------------------------------
MUTYH_Homo_sapiens/1-546  375 EQPGALGAQILLVQRPNSGLLAGLWEFPSVTWEPSEQLQRKALLQELQRWAGPLPATHLRHLGEVVHTFSHIKLTYQVYGLALEGQ 460
OGG1_Homo_sapiens/1-345       --------------------------------------------------------------------------------------
MBD4_Homo_sapiens/1-580   425 EALSPPRRKAFKKWTPPRSPFN-----LVQETLFHDPWKLLIATIFLNRTSGKMAIPVLWKFLEKYPSAEVARTADWRDVSELLKP 505

NTHL1_Homo_sapiens/1-312      --------------------------------------------------------------------------------------
MUTYH_Homo_sapiens/1-546  461 TPVTTVPPGARWLTQEEFHTAAVSTAMKKVFRVYQGQQPGTCMGSKRSQVSSPCSRKKPRMGQQVLDNFFRSHISTDAHSLNSAAQ 546
OGG1_Homo_sapiens/1-345       --------------------------------------------------------------------------------------
MBD4_Homo_sapiens/1-580   506 LGLYDLRAKTIVKFSDEYLTKQWKYPIELHGIGKYGNDSYRIFCVNEWKQVHPEDHKLNKYHDWLWENHEKLSLS----------- 580

NTHL1_Homo_sapiens/1-312
MUTYH_Homo_sapiens/1-546
OGG1_Homo_sapiens/1-345
MBD4_Homo_sapiens/1-580
```

c)

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

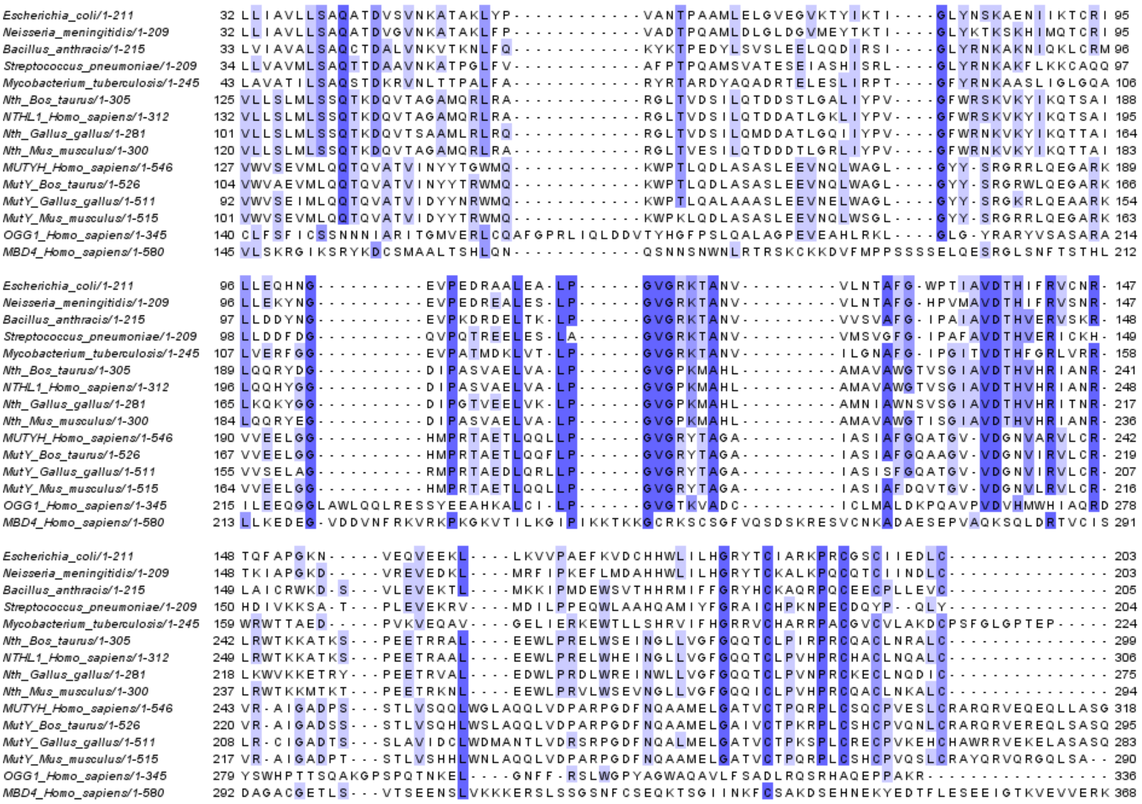
The [4Fe-4S] cluster motif is fully conserved in NTHL1 and MUTYH.

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

d)

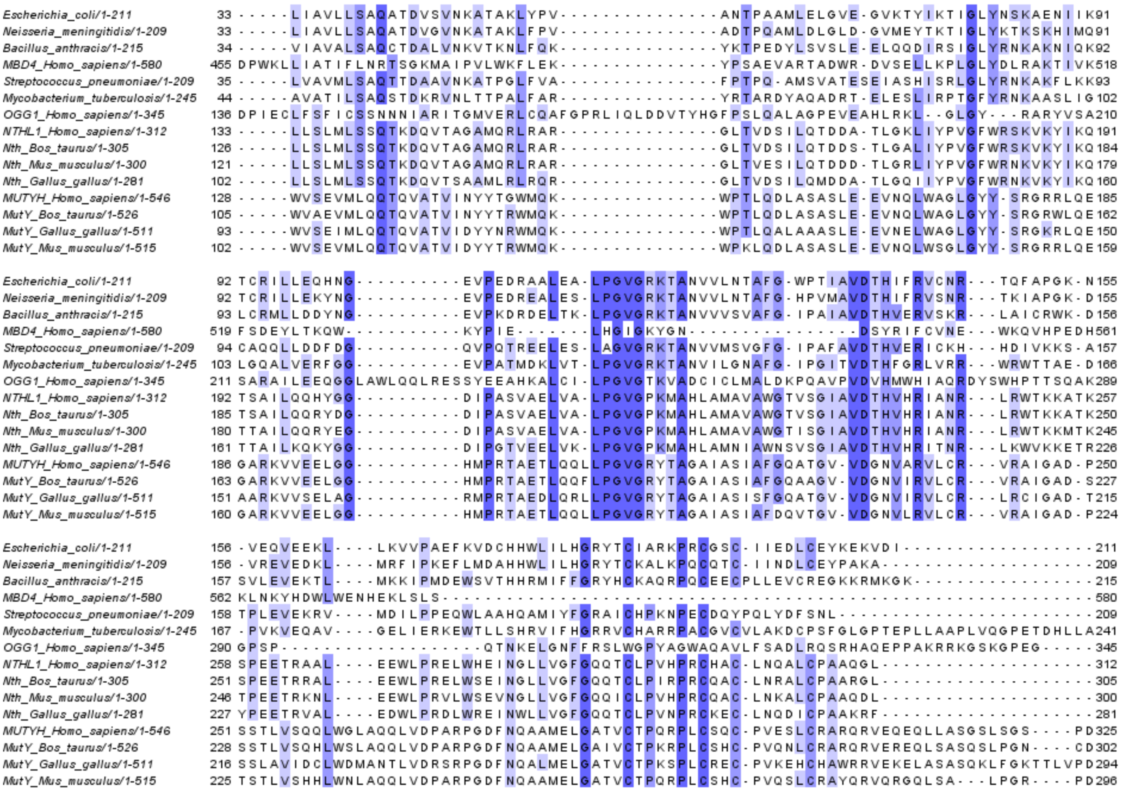
The Clustal W program seemed to produce the worst alignment, as the HhH motif was not well aligned in NTHL1.

e) Here is the MUSCLE alignment with all sequences:



```
Escherichia_coli/1-211                 32 LLIAVLLSAQATDVSVNKATAKLYP------------VANTPAAMLELGVEGVKTYIKTI----GLYNSKAENIIKTCRI 95
Neisseria_meningitidis/1-209           32 LLIAVLLSAQATDVGVNKATAKLFP------------VADTPQAMLDLGLDGVMEYTKTI----GLYKTKSKHIMQTCRI 95
Bacillus_anthracis/1-215               33 LVIAVALSAQCTDALVNKVTKNLFQ------------KYKTPEDYLSVSLEELQQDIRSI----GLYRNKAKNIQKLCRM 96
Streptococcus_pneumoniae/1-209         34 LLVAVMLSAQTTDAAVNKATPGLFV------------AFPTPQAMSVATESEIASHISRL----GLYRNKAKFLKKCAQQ 97
Mycobacterium_tuberculosis/1-245       43 LAVATILSAQSTDKRVNLTTPALFA------------RYRTARDYAQADRTELESLIRPT----GFYRNKAASLIGLGQA 106
Nth_Bos_taurus/1-305                  125 VLLSLMLSSQTKDQVTAGAMQRLRA------------RGLTVDSILQTDDSTLGALIYPV----GFWRSKVKYIKQTSAI 188
NTHL1_Homo_sapiens/1-312              132 VLLSLMLSSQTKDQVTAGAMQRLRA------------RGLTVDSILQTDDATLGKLIYPV----GFWRSKVKYIKQTSAI 195
Nth_Gallus_gallus/1-281               101 VLLSLMLSSQTKDQVTSAAMLRLRQ------------RGLTVDSILQMDDATLGQIIYPV----GFWRNKVKYIKQTTAI 164
Nth_Mus_musculus/1-300                120 VLLSLMLSSQTKDQVTAGAMQRLRA------------RGLTVESILQTDDDTLGRLIYPV----GFWRNKVKYIKQTTAI 183
MUTYH_Homo_sapiens/1-546              127 VWVSEVMLQQTQVATVINYYTGWMQ------------KWPTLQDLASASLEEVNQLWAGL----GYY-SRGRRLQEGARK 189
MutY_Bos_taurus/1-526                 104 VWVAEVMLQQTQVATVINYYTRWMQ------------KWPTLQDLASASLEEVNQLWAGL----GYY-SRGRWLQEGARK 166
MutY_Gallus_gallus/1-511               92 VWVSEIMLQQTQVATVIDYYNRWMQ------------KWPTLQALAAASLEEVNELWAGL----GYY-SRGKRLQEAARK 154
MutY_Mus_musculus/1-515               101 VWVSEVMLQQTQVATVIDYYTRWMQ------------KWPKLQDLASASLEEVNQLWSGL----GYY-SRGRRLQEGARK 163
OGG1_Homo_sapiens/1-345               140 CLFSFICSSNNNIARITGMVERLCQAFGPRLIQLDDVTYHGFPSLQALAGPEVEAHLRKL----GLG-YRARYVSASARA 214
MBD4_Homo_sapiens/1-580               145 VLSKRGIKSRYKDCSMAALTSHLQN------------QSNNSNWNLRTRSKCKKDVFMPPSSSSELQESRGLSNFTSTHL 212

Escherichia_coli/1-211                 96 LLEQHNG----------EVPEDRAALEA-LP------GVGRKTANV---------VLNTAFG-WPTIAVDTHIFRVCNR- 147
Neisseria_meningitidis/1-209           96 LLEKYNG----------EVPEDREALES-LP------GVGRKTANV---------VLNTAFG-HPVMAVDTHIFRVSNR- 147
Bacillus_anthracis/1-215               97 LLDDYNG----------EVPKDRDELTK-LP------GVGRKTANV---------VVSVAFG-IPAIAVDTHVERVSKR- 148
Streptococcus_pneumoniae/1-209         98 LLDDFDG----------QVPQTREELES-LA------GVGRKTANV---------VMSVGFG-IPAFAVDTHVERICKH- 149
Mycobacterium_tuberculosis/1-245      107 LVERFGG----------EVPATMDKLVT-LP------GVGRKTANV---------ILGNAFG-IPGITVDTHFGRLVRR- 158
Nth_Bos_taurus/1-305                  189 LQQRYDG----------DIPASVAELVA-LP------GVGPKMAHL---------AMAVAWGTVSGIAVDTHVHRIANR- 241
NTHL1_Homo_sapiens/1-312              196 LQQHYGG----------DIPASVAELVA-LP------GVGPKMAHL---------AMAVAWGTVSGIAVDTHVHRIANR- 248
Nth_Gallus_gallus/1-281               165 LKQKYGG----------DIPGTVEELVK-LP------GVGPKMAHL---------AMNIAWNSVSGIAVDTHVHRITNR- 217
Nth_Mus_musculus/1-300                184 LQQRYEG----------DIPASVAELVA-LP------GVGPKMAHL---------AMAVAWGTISGIAVDTHVHRIANR- 236
MUTYH_Homo_sapiens/1-546              190 VVEELGG----------HMPRTAETLQQLLP------GVGRYTAGA---------IASIAFGQATGV-VDGNVARVLCR- 242
MutY_Bos_taurus/1-526                 167 VVEELGG----------HMPRTAETLQQFLP------GVGRYTAGA---------IASIAFGQAAGV-VDGNVIRVLCR- 219
MutY_Gallus_gallus/1-511              155 VVSELAG----------RMPRTAEDLQRLLP------GVGRYTAGA---------IASISFGQATGV-VDGNVIRVLCR- 207
MutY_Mus_musculus/1-515               164 VVEELGG----------HMPRTAETLQQLLP------GVGRYTAGA---------IASIAFDQVTGV-VDGNVLRVLCR- 216
OGG1_Homo_sapiens/1-345               215 ILEEQGGLAWLQQLRESSYEEAHKALCI-LP------GVGTKVADC---------ICLMALDKPQAVPVDVHMWHIAQRD 278
MBD4_Homo_sapiens/1-580               213 LLKEDEG-VDDVNFRKVRKPKGKVTILKGIPIKKTKKGCRKSCSGFVQSDSKRESVCNKADAESEPVAQKSQLDRTVCIS 291

Escherichia_coli/1-211                148 TQFAPGKN-----VEQVEEKL----LKVVPAEFKVDCHHWLILHGRYTCIARKPRCGSCIIEDLC--------------- 203
Neisseria_meningitidis/1-209          148 TKIAPGKD-----VREVEDKL----MRFIPKEFLMDAHHWLILHGRYTCKALKPQCQTCIINDLC--------------- 203
Bacillus_anthracis/1-215              149 LAICRWKD-S---VLEVEKTL----MKKIPMDEWSVTHHRMIFFGRYHCKAQRPQCEECPLLEVC--------------- 205
Streptococcus_pneumoniae/1-209        150 HDIVKKSA-T---PLEVEKRV----MDILPPEQWLAAHQAMIYFGRAICHPKNPECDQYP--QLY--------------- 204
Mycobacterium_tuberculosis/1-245      159 WRWTTAED-----PVKVEQAV----GELIERKEWTLLSHRVIFHGRRVCHARRPACGVCVLAKDCPSFGLGPTEP----- 224
Nth_Bos_taurus/1-305                  242 LRWTKKATKS---PEETRRAL----EEWLPRELWSEINGLLVGFGQQTCLPIRPRCQACLNRALC--------------- 299
NTHL1_Homo_sapiens/1-312              249 LRWTKKATKS---PEETRAAL----EEWLPRELWHEINGLLVGFGQQTCLPVHPRCHACLNQALC--------------- 306
Nth_Gallus_gallus/1-281               218 LKWVKKETRY---PEETRVAL----EDWLPRDLWREINWLLVGFGQQTCLPVNPRCKECLNQDIC--------------- 275
Nth_Mus_musculus/1-300                237 LRWTKKMTKT---PEETRKNL----EEWLPRVLWSEVNGLLVGFGQQICLPVHPRCQACLNKALC--------------- 294
MUTYH_Homo_sapiens/1-546              243 VR-AIGADPS---STLVSQQLWGLAQQLVDPARPGDFNQAAMELGATVCTPQRPLCSQCPVESLCRARQRVEQEQLLASG 318
MutY_Bos_taurus/1-526                 220 VR-AIGADSS---STLVSQHLWSLAQQLVDPARPGDFNQAAMELGAIVCTPKRPLCSHCPVQNLCRARQRVEREQLSASQ 295
MutY_Gallus_gallus/1-511              208 LR-CIGADTS---SLAVIDCLWDMANTLVDRSRPGDFNQALMELGATVCTPKSPLCRECPVKEHCHAWRRVEKELASASQ 283
MutY_Mus_musculus/1-515               217 VR-AIGADPT---STLVSHHLWNLAQQLVDPARPGDFNQAAMELGATVCTPQRPLCSHCPVQSLCRAYQRVQRGQLSA-- 290
OGG1_Homo_sapiens/1-345               279 YSWHPTTSQAKGPSPQTNKEL----GNFF-RSLWGPYAGWAQAVLFSADLRQSRHAQEPPAKR----------------- 336
MBD4_Homo_sapiens/1-580               292 DAGACGETLS---VTSEENSLVKKKERSLSSGSNFCSEQKTSGIINKFCSAKDSEHNEKYEDTFLESEEIGTKVEVVERK 368
```

And here is the MAFFT alignment with all sequences:



```
Escherichia_coli/1-211                 33 -----LIAVLLSAQATDVSVNKATAKLYPV---------------ANTPAAMLELGVE-GVKTYIKTIGLYNSKAENIIK91
Neisseria_meningitidis/1-209           33 -----LIAVLLSAQATDVGVNKATAKLFPV---------------ADTPQAMLDLGLD-GVMEYTKTIGLYKTKSKHIMQ91
Bacillus_anthracis/1-215               34 -----VIAVALSAQCTDALVNKVTKNLFQK---------------YKTPEDYLSVSLE-ELQQDIRSIGLYRNKAKNIQK92
MBD4_Homo_sapiens/1-580               455 DPWKLLIATIFLNRTSGKMAIPVLWKFLEK---------------YPSAEVARTADWR-DVSELLKPLGLYDLRAKTIVK518
Streptococcus_pneumoniae/1-209         35 -----LVAVMLSAQTTDAAVNKATPGLFVA---------------FPTPQ-AMSVATESEIASHISRLGLYRNKAKFLKK93
Mycobacterium_tuberculosis/1-245       44 -----AVATILSAQSTDKRVNLTTPALFAR---------------YRTARDYAQADRT-ELESLIRPTGFYRNKAASLIG102
OGG1_Homo_sapiens/1-345               136 DPIECLFSFICSSNNNIARITGMVERLCQAFGPRLIQLDDVTYHGFPSLQALAGPEVEAHLRKL--GLGY---RARYVSA210
NTHL1_Homo_sapiens/1-312              133 -----LLSLMLSSQTKDQVTAGAMQRLRAR---------------GLTVDSILQTDDA-TLGKLIYPVGFWRSKVKYIKQ191
Nth_Bos_taurus/1-305                  126 -----LLSLMLSSQTKDQVTAGAMQRLRAR---------------GLTVDSILQTDDS-TLGALIYPVGFWRSKVKYIKQ184
Nth_Mus_musculus/1-300                121 -----LLSLMLSSQTKDQVTAGAMQRLRAR---------------GLTVESILQTDDD-TLGRLIYPVGFWRNKVKYIKQ179
Nth_Gallus_gallus/1-281               102 -----LLSLMLSSQTKDQVTSAAMLRLRQR---------------GLTVDSILQMDDA-TLGQIIYPVGFWRNKVKYIKQ160
MUTYH_Homo_sapiens/1-546              128 -----WVSEVMLQQTQVATVINYYTGWMQK---------------WPTLQDLASASLE-EVNQLWAGLGYY-SRGRRLQE185
MutY_Bos_taurus/1-526                 105 -----WVAEVMLQQTQVATVINYYTRWMQK---------------WPTLQDLASASLE-EVNQLWAGLGYY-SRGRWLQE162
MutY_Gallus_gallus/1-511               93 -----WVSEIMLQQTQVATVIDYYNRWMQK---------------WPTLQALAAASLE-EVNELWAGLGYY-SRGKRLQE150
MutY_Mus_musculus/1-515               102 -----WVSEVMLQQTQVATVIDYYTRWMQK---------------WPKLQDLASASLE-EVNQLWSGLGYY-SRGRRLQE159

Escherichia_coli/1-211                 92 TCRILLEQHNG----------EVPEDRAALEA-LPGVGRKTANVVLNTAFG-WPTIAVDTHIFRVCNR---TQFAPGK-N155
Neisseria_meningitidis/1-209           92 TCRILLEKYNG----------EVPEDREALES-LPGVGRKTANVVLNTAFG-HPVMAVDTHIFRVSNR---TKIAPGK-D155
Bacillus_anthracis/1-215               93 LCRMLLDDYNG----------EVPKDRDELTK-LPGVGRKTANVVVSVAFG-IPAIAVDTHVERVSKR---LAICRWK-D156
MBD4_Homo_sapiens/1-580               519 FSDEYLTKQW-----------KYPIE-------LHGIGKYGN----------------DSYRIFCVNE---WKQVHPEDH561
Streptococcus_pneumoniae/1-209         94 CAQQLLDDFDG----------QVPQTREELES-LAGVGRKTANVVMSVGFG-IPAFAVDTHVERICKH---HDIVKKS-A157
Mycobacterium_tuberculosis/1-245      103 LGQALVERFGG----------EVPATMDKLVT-LPGVGRKTANVILGNAFG-IPGITVDTHFGRLVRR---WRWTTAE-D166
OGG1_Homo_sapiens/1-345               211 SARAILEEQGGLAWLQQLRESSYEEAHKALCI-LPGVGTKVADCICLMALDKPQAVPVDVHMWHIAQRDYSWHPTTSQAK289
NTHL1_Homo_sapiens/1-312              192 TSAILQQHYGG----------DIPASVAELVA-LPGVGPKMAHLAMAVAWGTVSGIAVDTHVHRIANR---LRWTKKATK257
Nth_Bos_taurus/1-305                  185 TSAILQQRYDG----------DIPASVAELVA-LPGVGPKMAHLAMAVAWGTVSGIAVDTHVHRIANR---LRWTKKATK250
Nth_Mus_musculus/1-300                180 TTAILQQRYEG----------DIPASVAELVA-LPGVGPKMAHLAMAVAWGTISGIAVDTHVHRIANR---LRWTKKMTK245
Nth_Gallus_gallus/1-281               161 TTAILKQKYGG----------DIPGTVEELVK-LPGVGPKMAHLAMNIAWNSVSGIAVDTHVHRITNR---LKWVKKETR226
MUTYH_Homo_sapiens/1-546              186 GARKVVEELGG----------HMPRTAETLQQLLPGVGRYTAGAIASIAFGQATGV-VDGNVARVLCR---VRAIGAD-P250
MutY_Bos_taurus/1-526                 163 GARKVVEELGG----------HMPRTAETLQQFLPGVGRYTAGAIASIAFGQAAGV-VDGNVIRVLCR---VRAIGAD-S227
MutY_Gallus_gallus/1-511              151 AARKVVSELAG----------RMPRTAEDLQRLLPGVGRYTAGAIASISFGQATGV-VDGNVIRVLCR---LRCIGAD-T215
MutY_Mus_musculus/1-515               160 GARKVVEELGG----------HMPRTAETLQQLLPGVGRYTAGAIASIAFDQVTGV-VDGNVLRVLCR---VRAIGAD-P224

Escherichia_coli/1-211                156 -VEQVEEKL----LKVVPAEFKVDCHHWLILHGRYTCIARKPRCGSC-IIEDLCEYKEKVDI------------------211
Neisseria_meningitidis/1-209          156 -VREVEDKL----MRFIPKEFLMDAHHWLILHGRYTCKALKPQCQTC-IINDLCEYPAKA--------------------209
Bacillus_anthracis/1-215              157 SVLEVEKTL----MKKIPMDEWSVTHHRMIFFGRYHCKAQRPQCEECPLLEVCREGKKRMKGK-----------------215
MBD4_Homo_sapiens/1-580               562 KLNKYHDWLWENHEKLSLS-------------------------------------------------------------580
Streptococcus_pneumoniae/1-209        158 TPLEVEKRV----MDILPPEQWLAAHQAMIYFGRAICHPKNPECDQYPQLYDFSNL------------------------209
Mycobacterium_tuberculosis/1-245      167 -PVKVEQAV----GELIERKEWTLLSHRVIFHGRRVCHARRPACGVCVLAKDCPSFGLGPTEPLLAAPLVQGPETDHLLA241
OGG1_Homo_sapiens/1-345               290 GPSP-------------------QTNKELGNFFRSLWGPYAGWAQAVLFSADLRQSRHAQEPPAKRRKGSKGPEG-----345
NTHL1_Homo_sapiens/1-312              258 SPEETRAAL----EEWLPRELWHEINGLLVGFGQQTCLPVHPRCHAC-LNQALCPAAQGL--------------------312
Nth_Bos_taurus/1-305                  251 SPEETRRAL----EEWLPRELWSEINGLLVGFGQQTCLPIRPRCQAC-LNRALCPAARGL--------------------305
Nth_Mus_musculus/1-300                246 TPEETRKNL----EEWLPRVLWSEVNGLLVGFGQQICLPVHPRCQAC-LNKALCPAAQDL--------------------300
Nth_Gallus_gallus/1-281               227 YPEETRVAL----EDWLPRDLWREINWLLVGFGQQTCLPVNPRCKEC-LNQDICPAAKRF--------------------281
MUTYH_Homo_sapiens/1-546              251 SSTLVSQQLWGLAQQLVDPARPGDFNQAAMELGATVCTPQRPLCSQC-PVESLCRARQRVEQEQLLASGSLSGS----PD325
MutY_Bos_taurus/1-526                 228 SSTLVSQHLWSLAQQLVDPARPGDFNQAAMELGAIVCTPKRPLCSHC-PVQNLCRARQRVEREQLSASQSLPGN----CD302
MutY_Gallus_gallus/1-511              216 SSLAVIDCLWDMANTLVDRSRPGDFNQALMELGATVCTPKSPLCREC-PVKEHCHAWRRVEKELASASQKLFGKTTLVPD294
MutY_Mus_musculus/1-515               225 TSTLVSHHLWNLAQQLVDPARPGDFNQAAMELGATVCTPQRPLCSHC-PVQSLCRAYQRVQRGQLSA---LPGR----PD296
```

MAFFT was able to correctly align the HhH motif of MBD4 when all sequences are included, but not MUSCLE. MAFFT performed best.