Intro to Bioinformatics, Group Project – Milestone C

Julie Garcia, Katie Jenike, Joseph Sparks, Jane Ryu

April 1, 2017

[5] Generate a multiple sequence alignment with your novel protein, your original query protein, and a group of other members of this family. A typical number of proteins to use in a multiple sequence alignment is a minimum of 5 or 10 and a maximum 30, although the exact number is up to you.

A blastp search was performed with our original query protein sequence, NP_001317287.1, to find a group of other related proteins. From the blast results page, the link "taxonomy reports" was selected. The protein with the highest score from the western lowland gorilla (XP_018866767), the pygmy chimpanzee (XP_008973836), and the green monkey (XP_007978727) were selected to use in the MSA along with the novel protein sequence and the original query sequence. Each of these sequences make up the methylenetetrahydrofolate reductase protein for their particular sequence. Their amino acid sequences were obtained from NCBI's Protein database.

The following was used as a query in Clustal Omega:

Original query sequence:

>Human

MDHRKARVLPAGHYCPSLGIWASQVGSVRSSVPPSISRNPAMVNEARGNSSLNPCLEGSASSGSESSK
DSSRCSTPGLDPERHERLREKMRRRLESGDKWFSLEFFPPRTAEGAVNLISRFDRMAAGGPLYIDVTWH
PAGDPGSDKETSSMMIASTAVNYCGLETILHMTCCRQRLEEITGHLHKAKQLGLKNIMALRGDPIGDQWE
EEEGGFNYAVDLVKHIRSEFGDYFDICVAGYPKGHPEAGSFEADLKHLKEKVSAGADFIITQLFFEADTFF
RF VKACTDMGITCPIVPGIFPIQGYHSLRQLVKLSKLEVPQEIKDVIEPIKDNDAAIRNYGIELAVSLCQEL
LASGLVPGLHFYTLNREMATTEVLKRLGMWTEDPRRPLPWALSAHPKRREEDVRPIFWASRPKSYIYRT
QEWDEFPNGRWGNSSSPAFGELKDYYLFYLKSKSPKEELLKMWGEELTSEESVFEVFVLYLSGEPNRN
GHKVTCLPWNDEPLAAETSLLKEELLRVNRQGILTINSQPNINGKPSSDPIVGWGPSGGYVFQKAYLEFFT
SRETAEALLQVLKKYELRVNYHLVNVKGENITNAPELQPNAVTWGIFPGREIIQPTVVDPVSFMFWKDEAF
A LWIERWGKLYEEESPSRTIIQYIHDNYFLVNLVDNDFPLDNCLWQVVEDTLELLNRPTQNARETEAP

>Novel

VRAGADLCITDVFYDTNAYAKFIKECREAGIARTFPIVPGILPIHSFKSFEGIVDHLGINVPASIREAIEPIKED
DAAMQEYGISLAESMCLELLNSGLAQGMYFYTFNLEYSVRHLLEERLKVTPKSQLPWRPSANPKRIEEDV
RPIFWANRPKSYLIRTESWNEFPSGRWGSAVESASFSELKDSTLFARETFFERDDIKKKAWGEAPQTRE
EVFEVFAGFVEGRVQFLPWCEESLHLETSVIRDKLVQV

>Gorilla

MDHRKARVFPAGHYCPSLGIWASQVGSVRSSVPPSISRNPAMVNEARGNSSLNPCLEGSASSGSESSK
DSSRCSTPGLDPERHERLREKMRRRLESGDKWFSLEFFPPRTAEGAVNLISRFDRMAAGGPLYIDVTWH
PAGDPGSDKETSSMMIASTAVNYCGLETILHMTCCRQRLEEITGHLHKAKQLGLKNIMALRGDPIGDQWE
EEEGGFNYAVDLVKHIRSEFGDYFDICVAGYPKGHPEAGSFEADLKHLKEKVSAGADFIITQLFFEADTFF
RFVKACTDMGITCPIVPGIFPIQGYHSLRQLVKLSKLEVPQEIKDVIEPIKDNDAAIRNYGIELAVSLCQELLA
SGLVPGLHFYTLNREMATTEVLKRLGMWTEDPRRPLPWALSAHPKRREEDVRPIFWASRPKSYIYRTQE

WDEFPNGRWGNSSSPAFGELKDYYLFYLKSKSPKEELLKMWGEELTSEESVFEVFVLYLSGEPNRNGH
KVTCLPWNDEPLAAETSLLKEELLRVNRQGILTINSQPNINGKPSSDPIVGWGPSGGYVFQKAYLEFFTSR
ETAEALLQVLKKYELRVNYHLVNVKGENITNAPELQPNAVTWGIFPGREIIQPTVVDPVSFMFWKDEAFAL
WIERWGKLYEEESPSRTIIQYIHDNYFLVNLVDNDFPLDNCLWQVVEDTLELLNRPTQNARETEAP

>Pygmy_Chimp

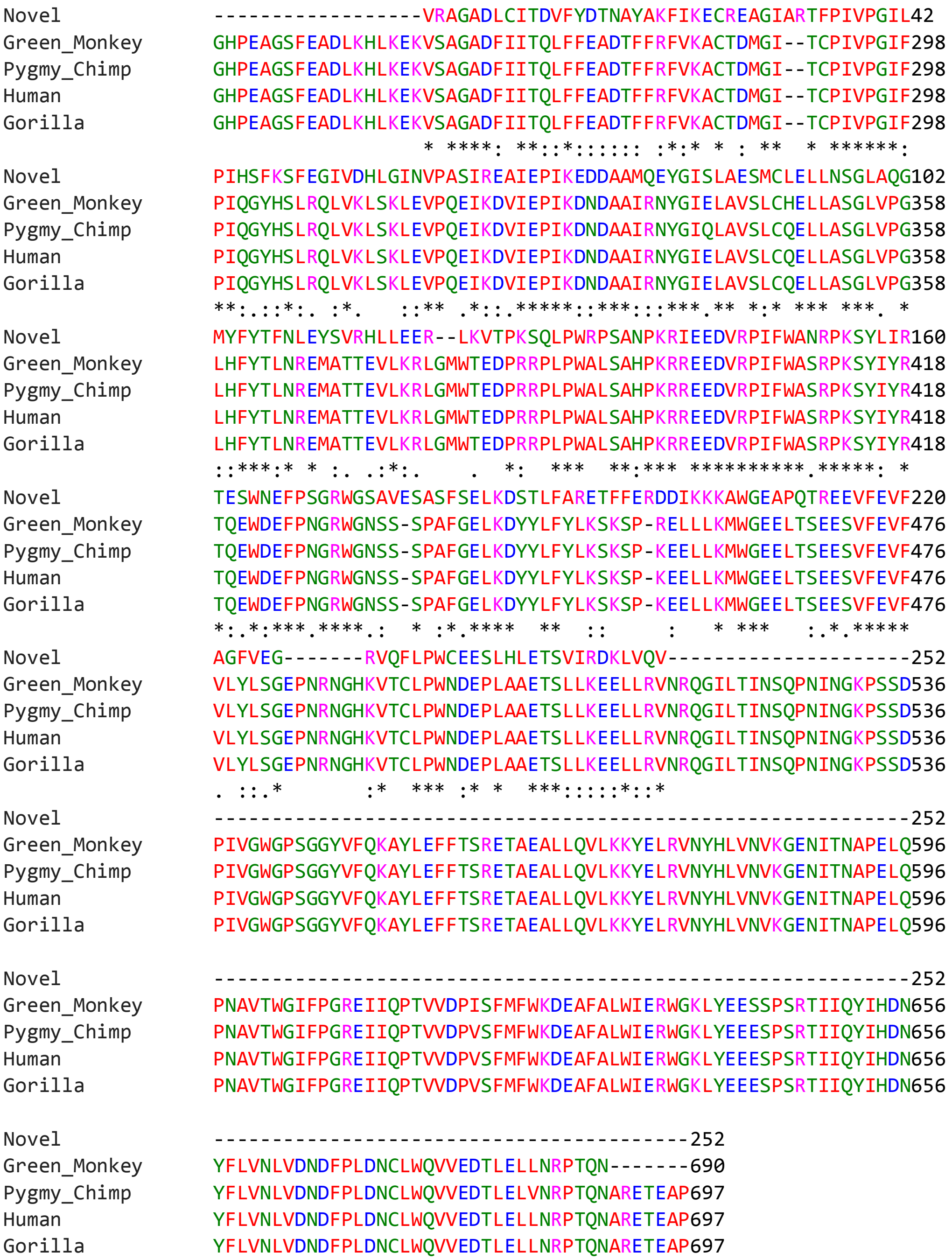
MDHRKARVLPAGHYCPSLGIWASQVGSVRSSVPPSISRNPAMVNEARGNSSLNPCLEGSASSGSESSK
DSSRCSTPGLDPERHERLREKMRRRLESGDKWFSLEFFPPRTAEGAVNLISRFDRMAAGGPLYIDVTWH
PAGDPGSDKETSSMMIASTAVNYCGLETILHMTCCRQRLEEITGHLHKAKQLGLKNIMALRGDPIGDQWE
EEEGGFNYAVDLVKHIRSEFGDYFDICVAGYPKGHPEAGSFEADLKHLKEKVSAGADFIITQLFFEADTFF
RFVKACTDMGITCPIVPGIFPIQGYHSLRQLVKLSKLEVPQEIKDVIEPIKDNDAAIRNYGIQLAVSLCQELLA
SGLVPGLHFYTLNREMATTEVLKRLGMWTEDPRRPLPWALSAHPKRREEDVRPIFWASRPKSYIYRTQE
WDEFPNGRWGNSSSPAFGELKDYYLFYLKSKSPKEELLKMWGEELTSEESVFEVFVLYLSGEPNRNGH
KVTCLPWNDEPLAAETSLLKEELLRVNRQGILTINSQPNINGKPSSDPIVGWGPSGGYVFQKAYLEFFTSR
ETAEALLQVLKKYELRVNYHLVNVKGENITNAPELQPNAVTWGIFPGREIIQPTVVDPVSFMFWKDEAFAL
WIERWGKLYEEESPSRTIIQYIHDNYFLVNLVDNDFPLDNCLWQVVEDTLELVNRPTQNARETEAP

>Green_Monkey

MDHRKARVLPAGHYCPSLGIWASQAGSVRFSVPPSISRNLAMVNEARGNGSLSPCLEGSASSSSESSKD
SSRCSTPGLDPERHERLRDKMRRRMESGDKWFSLEFFPPRTAEGAVNLISRFDRMAAGGPLFIDVTWH
PAGDPGSDKETSSMMIASTAVNYCGLETILHMTCCCQRLEEITGHLHKAKQLGLKNIMALRGDPIGDQWE
EEEGGFNYAVDLVKHIRNEFGDYFDLCVAGYPKGHPEAGSFEADLKHLKEKVSAGADFIITQLFFEADTFF
RFVKACTDMGITCPIVPGIFPIQGYHSLRQLVKLSKLEVPQEIKDVIEPIKDNDAAIRNYGIELAVSLCHELLA
SGLVPGLHFYTLNREMATTEVLKRLGMWTEDPRRPLPWALSAHPKRREEDVRPIFWASRPKSYIYRTQE
WDEFPNGRWGNSSSPAFGELKDYYLFYLKSKSPRELLLKMWGEELTSEESVFEVFVLYLSGEPNRNGH
KVTCLPWNDEPLAAETSLLKEELLRVNRQGILTINSQPNINGKPSSDPIVGWGPSGGYVFQKAYLEFFTSR
ETAEALLQVLKKYELRVNYHLVNVKGENITNAPELQPNAVTWGIFPGREIIQPTVVDPISFMFWKDEAFAL
WIERWGKLYEESSPSRTIIQYIHDNYFLVNLVDNDFPLDNCLWQVVEDTLELLNRPTQN

MSA Results:

```
Novel             ------------------------------------------------------------0
Green_Monkey      MDHRKARVLPAGHYCPSLGIWASQAGSVRFSVPPSISRNLAMVNEARGNGSLSPCLEGSA60
Pygmy_Chimp       MDHRKARVLPAGHYCPSLGIWASQVGSVRSSVPPSISRNPAMVNEARGNSSLNPCLEGSA60
Human             MDHRKARVLPAGHYCPSLGIWASQVGSVRSSVPPSISRNPAMVNEARGNSSLNPCLEGSA60
Gorilla           MDHRKARVFPAGHYCPSLGIWASQVGSVRSSVPPSISRNPAMVNEARGNSSLNPCLEGSA60

Novel             ------------------------------------------------------------0
Green_Monkey      SSSSESSKDSSRCSTPGLDPERHERLRDKMRRRMESGDKWFSLEFFPPRTAEGAVNLISR120
Pygmy_Chimp       SSGSESSKDSSRCSTPGLDPERHERLREKMRRRLESGDKWFSLEFFPPRTAEGAVNLISR120
Human             SSGSESSKDSSRCSTPGLDPERHERLREKMRRRLESGDKWFSLEFFPPRTAEGAVNLISR120
Gorilla           SSGSESSKDSSRCSTPGLDPERHERLREKMRRRLESGDKWFSLEFFPPRTAEGAVNLISR120

Novel             ------------------------------------------------------------0
Green_Monkey      FDRMAAGGPLFIDVTWHPAGDPGSDKETSSMMIASTAVNYCGLETILHMTCCCQRLEEIT180
Pygmy_Chimp       FDRMAAGGPLYIDVTWHPAGDPGSDKETSSMMIASTAVNYCGLETILHMTCCRQRLEEIT180
Human             FDRMAAGGPLYIDVTWHPAGDPGSDKETSSMMIASTAVNYCGLETILHMTCCRQRLEEIT180
Gorilla           FDRMAAGGPLYIDVTWHPAGDPGSDKETSSMMIASTAVNYCGLETILHMTCCRQRLEEIT180

Novel             ------------------------------------------------------------0
Green_Monkey      GHLHKAKQLGLKNIMALRGDPIGDQWEEEEGGFNYAVDLVKHIRNEFGDYFDLCVAGYPK240
Pygmy_Chimp       GHLHKAKQLGLKNIMALRGDPIGDQWEEEEGGFNYAVDLVKHIRSEFGDYFDICVAGYPK240
Human             GHLHKAKQLGLKNIMALRGDPIGDQWEEEEGGFNYAVDLVKHIRSEFGDYFDICVAGYPK240
Gorilla           GHLHKAKQLGLKNIMALRGDPIGDQWEEEEGGFNYAVDLVKHIRSEFGDYFDICVAGYPK240
```



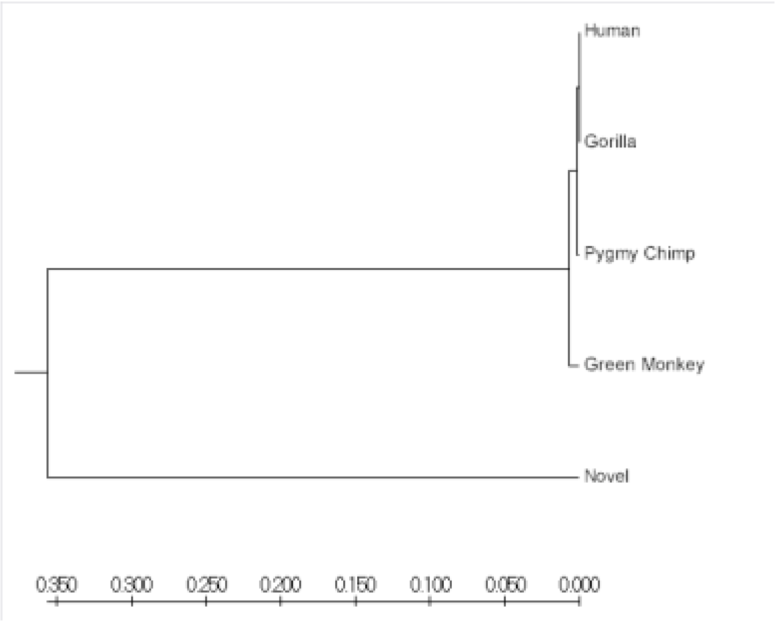
```
Novel             ------------------VRAGADLCITDVFYDTNAYAKFIKECREAGIARTFPIVPGIL42
Green_Monkey      GHPEAGSFEADLKHLKEKVSAGADFIITQLFFEADTFFRFVKACTDMGI--TCPIVPGIF298
Pygmy_Chimp       GHPEAGSFEADLKHLKEKVSAGADFIITQLFFEADTFFRFVKACTDMGI--TCPIVPGIF298
Human             GHPEAGSFEADLKHLKEKVSAGADFIITQLFFEADTFFRFVKACTDMGI--TCPIVPGIF298
Gorilla           GHPEAGSFEADLKHLKEKVSAGADFIITQLFFEADTFFRFVKACTDMGI--TCPIVPGIF298
                                    * ****: **::*:::::: :*:* * : **  * ******:

Novel             PIHSFKSFEGIVDHLGINVPASIREAIEPIKEDDAAMQEYGISLAESMCLELLNSGLAQG102
Green_Monkey      PIQGYHSLRQLVKLSKLEVPQEIKDVIEPIKDNDAAIRNYGIELAVSLCHELLASGLVPG358
Pygmy_Chimp       PIQGYHSLRQLVKLSKLEVPQEIKDVIEPIKDNDAAIRNYGIQLAVSLCQELLASGLVPG358
Human             PIQGYHSLRQLVKLSKLEVPQEIKDVIEPIKDNDAAIRNYGIELAVSLCQELLASGLVPG358
Gorilla           PIQGYHSLRQLVKLSKLEVPQEIKDVIEPIKDNDAAIRNYGIELAVSLCQELLASGLVPG358
                  **:.::*:. :*.   ::** .*::.*****::***:::***.** *:* *** ***. *

Novel             MYFYTFNLEYSVRHLLEER--LKVTPKSQLPWRPSANPKRIEEDVRPIFWANRPKSYLIR160
Green_Monkey      LHFYTLNREMATTEVLKRLGMWTEDPRRPLPWALSAHPKRREEDVRPIFWASRPKSYIYR418
Pygmy_Chimp       LHFYTLNREMATTEVLKRLGMWTEDPRRPLPWALSAHPKRREEDVRPIFWASRPKSYIYR418
Human             LHFYTLNREMATTEVLKRLGMWTEDPRRPLPWALSAHPKRREEDVRPIFWASRPKSYIYR418
Gorilla           LHFYTLNREMATTEVLKRLGMWTEDPRRPLPWALSAHPKRREEDVRPIFWASRPKSYIYR418
                  ::***:* * :. .:*:.    .  *:  ***  **:*** **********.*****: *

Novel             TESWNEFPSGRWGSAVESASFSELKDSTLFARETFFERDDIKKKAWGEAPQTREEVFEVF220
Green_Monkey      TQEWDEFPNGRWGNSS-SPAFGELKDYYLFYLKSKSP-RELLLKMWGEELTSEESVFEVF476
Pygmy_Chimp       TQEWDEFPNGRWGNSS-SPAFGELKDYYLFYLKSKSP-KEELLKMWGEELTSEESVFEVF476
Human             TQEWDEFPNGRWGNSS-SPAFGELKDYYLFYLKSKSP-KEELLKMWGEELTSEESVFEVF476
Gorilla           TQEWDEFPNGRWGNSS-SPAFGELKDYYLFYLKSKSP-KEELLKMWGEELTSEESVFEVF476
                  *:.*:***.****.:  * :*.****  **  ::     :   * ***   :.*.*****

Novel             AGFVEG-------RVQFLPWCEESLHLETSVIRDKLVQV---------------------252
Green_Monkey      VLYLSGEPNRNGHKVTCLPWNDEPLAAETSLLKEELLRVNRQGILTINSQPNINGKPSSD536
Pygmy_Chimp       VLYLSGEPNRNGHKVTCLPWNDEPLAAETSLLKEELLRVNRQGILTINSQPNINGKPSSD536
Human             VLYLSGEPNRNGHKVTCLPWNDEPLAAETSLLKEELLRVNRQGILTINSQPNINGKPSSD536
Gorilla           VLYLSGEPNRNGHKVTCLPWNDEPLAAETSLLKEELLRVNRQGILTINSQPNINGKPSSD536
                  . ::.*       :*  *** :* *  ***:::::*::*

Novel             ------------------------------------------------------------252
Green_Monkey      PIVGWGPSGGYVFQKAYLEFFTSRETAEALLQVLKKYELRVNYHLVNVKGENITNAPELQ596
Pygmy_Chimp       PIVGWGPSGGYVFQKAYLEFFTSRETAEALLQVLKKYELRVNYHLVNVKGENITNAPELQ596
Human             PIVGWGPSGGYVFQKAYLEFFTSRETAEALLQVLKKYELRVNYHLVNVKGENITNAPELQ596
Gorilla           PIVGWGPSGGYVFQKAYLEFFTSRETAEALLQVLKKYELRVNYHLVNVKGENITNAPELQ596

Novel             ------------------------------------------------------------252
Green_Monkey      PNAVTWGIFPGREIIQPTVVDPISFMFWKDEAFALWIERWGKLYEESSPSRTIIQYIHDN656
Pygmy_Chimp       PNAVTWGIFPGREIIQPTVVDPVSFMFWKDEAFALWIERWGKLYEEESPSRTIIQYIHDN656
Human             PNAVTWGIFPGREIIQPTVVDPVSFMFWKDEAFALWIERWGKLYEEESPSRTIIQYIHDN656
Gorilla           PNAVTWGIFPGREIIQPTVVDPVSFMFWKDEAFALWIERWGKLYEEESPSRTIIQYIHDN656

Novel             ------------------------------------------252
Green_Monkey      YFLVNLVDNDFPLDNCLWQVVEDTLELLNRPTQN-------690
Pygmy_Chimp       YFLVNLVDNDFPLDNCLWQVVEDTLELVNRPTQNARETEAP697
Human             YFLVNLVDNDFPLDNCLWQVVEDTLELLNRPTQNARETEAP697
Gorilla           YFLVNLVDNDFPLDNCLWQVVEDTLELLNRPTQNARETEAP697
```

The MSA shows high homology between the Human, Pygmy_Chimp, Gorilla, and Green_Monkey, with an exact match between the first four. This is to be expected, because they are all in the MTHFR family of proteins that are highly conserved across species. Our novel protein shows a lower, but still significant homology with high similarity of ~50% in the middle region of the protein sequences of these other species. This aligns with the expected identity of 46% between the original query (MTHR_HUMAN) and our novel protein.

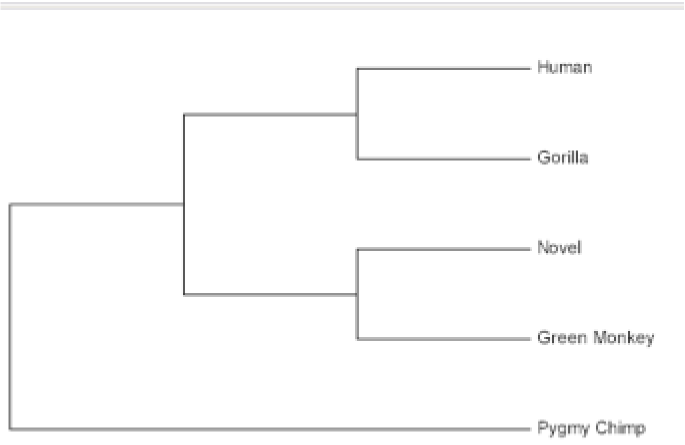
[6] Create a phylogenetic tree, using either a parsimony or distance-based approach. Bootstrapping and tree rooting are optional. Use any program such as MEGA3, PAUP, or Phylip.

We tested two different phylogenetic tree methods using Mega7. The first tree (directly below) is a UPGMA tree with the Poisson model, while the second tree uses maximum parsimony. The maximum parsimony method reduces branch length by minimizing the number of mutations, while UPGMA is a quick bottom-up clustering algorithm. As you can see both trees show a similar trend, with high homology between the four species selected and a lower homology for our novel protein.

UPGMA tree:



Maximum parsimony tree:


Human
Gorilla
Novel
Green Monkey
Pygmy Chimp