R Notebook

Code

Rose Abdelmalak

SCIENTIFIC QUESTION: What mutations and alternate splicing differences in DNA sequence exist between four different nucleotide sequence variants of the gene FRMD4A found amongst homo sapians, and how does that impact the homology in the structures of the protein?

BACKGROUND: My gene of interest is the FERM Domain Containing 4A (FRMD4A) gene. This gene FERM domain-containing protein that regulates epithelial cell polarity. It is associatd with adherens junctions and actin cable formation which occurs in epithelial cell polarity (FERM domain-containing protein that regulates epithelial cell polarity). Previous research has shown a correlation between miss regulation of the FRMD4A gene, in particularly >80% up regulation, and oral squamous cell carcinoma (Zheng et al 2016). The study done by Zheng et al revealed that Silencing FRMD4A gene reduced the proliferation of CAL27 cells and led to cell cycle arrest in the G1 phase, as well as significantly suppressing the migration and invasion capacity of CAL27 cells. Furthermore, As an inspiring dentist I am interested in oral squamous cell carcinoma and ways to prevent it. By understanding which variants of the FRMD4A gene have mutated protein structures that alter the function of the FRMD4A gene we can target those genes for therapy and treat for oral squamous cell carcinoma from very early stages (Reva et al 2011). DNA sequence of the four FRMD4A variants are extracted from the ncbi nucleotides database. Variant 1: https://www.ncbi.nlm.nih.gov/nuccore/NM_018027.5 (https://www.ncbi.nlm.nih.gov/nuccore/NM_018027.5) Variant 2: https://www.ncbi.nlm.nih.gov/nuccore/NM_001318336.2 (https://www.ncbi.nlm.nih.gov/nuccore/NM_001318336.2) Variant 3: https://www.ncbi.nlm.nih.gov/nuccore/NM_001318337.2 (https://www.ncbi.nlm.nih.gov/nuccore/NM_001318337.2) Variant 4: https://www.ncbi.nlm.nih.gov/nuccore/NM_001318338 (https://www.ncbi.nlm.nih.gov/nuccore/NM_001318338). The protein structures used for the homology modeling were extracted from SWISS protein and PDB. https://swissmodel.expasy.org/repository/uniprot/Q9P2Q2 (https://swissmodel.expasy.org/repository/uniprot/Q9P2Q2)

“FRMD4A FERM Domain Containing 4A [Homo Sapiens (Human)] - Gene - NCBI.” Nih.Gov, https://www.ncbi.nlm.nih.gov/gene/55691 (https://www.ncbi.nlm.nih.gov/gene/55691). Accessed 5 June 2022.

Reva, Boris, et al. “Predicting the Functional Impact of Protein Mutations: Application to Cancer Genomics.” Nucleic Acids Research, vol. 39, no. 17, 2011, p. e118, doi:10.1093/nar/gkr407 (doi:10.1093/nar/gkr407).

Zheng, Xianghuai, et al. “FRMD4A: A Potential Therapeutic Target for the Treatment of Tongue Squamous Cell Carcinoma.” International Journal of Molecular Medicine, vol. 38, no. 5, 2016, pp. 1443–1449, doi:10.3892/ijmm.2016.2745 (doi:10.3892/ijmm.2016.2745).

SCIENTIFIC HYPOTHESIS: If the sequences of the four FRMD4A gene variants are altered through mutations and/or alternate splicing then they will exhibit differences in their protein structures.

ANALYSIS: First, Multiple sequence alignment (msa) was performed on the DNA sequence from 4 variants of the FRMD4A gene which were extracted from the ncbi nucleotides database as fasta files. This data was visualized using pretty print to show the aligned sequences in a publication worthy figure. Second, Homology modeling was performed on the protein structures of the four variants of the FRMD4A gene. In order to do so I extracted the amino acid sequences of each of the variants as FASTA proteins. I then used those amino acid sequences to model the structures of the variants using the SWISS protein database.The PDB database was used in order to find and extract a template structure to model my variants on. In order to visualize the homology between these

protein structures I created a heat map to see how each protein model varied from the other in terms of their modes and rmsip values which take into acount confirmation/ movement fluctuations. Finally, I used pymol to view my structures but displayed them in the rnootbook using the NGLVieweR package.

Hide

```
library(Biostrings)
library(msa)
library(bio3d)
```

```
Attaching package: ‘bio3d’

The following object is masked from ‘package:seqLogo’:

    consensus

The following object is masked from ‘package:Biostrings’:

    mask

The following object is masked from ‘package:IRanges’:

    trim
```

Hide

```
library(NGLVieweR)
```

```
Registered S3 method overwritten by 'htmlwidgets':
  method           from
  print.htmlwidget tools:rstudio
```

- Biostrings is used for reading long strings of data. In this case biostrings is helpful in reading and maninpulating the nucelotide sequences by reading in the amino acid sequences, which then can be converted into codons, it can also count the length of sequences and along with many other functions.
- msa is used for multiple sequence alignment and includes a function called “msa pretty print” that cleans up the alignment and displays it as a publication worthy alignment. I used this package to compare the sequences of my four FRMD4A gene variants.
- bio3d is a package that allows us to analyze the structures of proteins by giving us the ability to read, write and process biomolecular structure, sequence and dynamics trajectory data. I used this package to perform homology modeling and compare the four protein structures associated with my FRMD4A gene variants.
- NGLVieweR is a package that allows us to code interactive 3D visualizations of molecular structures. I used this package to display the four protein structures associated with my four FRMD4A gene variants.

MULTIPLE SEQUENCE ALIGNMENT

Multiple Sequence Alignment often abbreviated as msa is a bioinformatics method in which DNA sequences are aligned and compared. Msa can be helpful in identifying which sequences are conserved and where mutations or alternate splicing may have occured. I used this bioinformatics method in my project in order to align the nucleotide sequence from the four FRMD4A gene variants that I obtained from ncbi as fasta files (linked above). msa then revealed the aligned sequences and placed a - in place of missing or deleted nucleotides, a ? when the sequence could not be identified and simply the nucleotide (A, T, C or G) when appropriate. By utalizing MSA I am able to determine which sequences were conserved and where there were variations in the sequences.

Hide

```
# read in the fasta file with the DNA sequence of the 4 FRMD4A variants using readAAStri
ngSet and assign to FRMD4A_unaligned because we have yet to perform MSA
system.file("tex", "texshade.sty", package="msa")
```

```
[1] "/Library/Frameworks/R.framework/Versions/4.1/Resources/library/msa/tex/texshade.st
y"
```

Hide

```
FRMD4A_unaligned <- readAAStringSet("FRMD4A.fasta")

#code check to ensure that all four variant sequences were imported
length(FRMD4A_unaligned)
```

```
[1] 4
```

Hide

```
#perform multiple sequence alignment using the function msa() and assign to FRMDA_MSA
FRMD4A_MSA <- msa(FRMD4A_unaligned)
```

```
use default substitution matrix
```

Hide

```
#print the full msa results using show= "complete"
print(FRMD4A_MSA, show="complete")
```

```
MsaAAMultipleAlignment with 4 rows and 3219 columns
    aln (1..108)
names
[1] ATGGTGGTT------CAGGCTGCAGTGGCTCCGAATAG-------ATCCCAAAGACTTTTACTGAAAATTCCTTATGGATCTCT
GAGAAGG----------------- lcl|NC_000010.11_...
[2] ------------------------------------------------------------------------------------
------------------------ lcl|NC_000010.11_...
[3] ATGGCAGTG------CAG----CTGGTGCCCGACTCAG-------CTCTCGG------------------CCTGCTG-------
------------------------ lcl|NC_000010.11_...
[4] ATGGCTGCTGGGCTGCTTGGCTCTGAGGACCCACCGTGGAGTTGGAACCTGACTTGTCGGGCGCTGAGGACCTGCCAAGTGAGG
AACATTCGAGTTCTGCAGCTGCTG lcl|NC_000010.11_...
Con ATGGC?GTT------CAG?---CTG?GGCCCCAC??AG-------ATCCCGA----?----?----?---CCTGCTG--?----
-?-?-------------------- Consensus

    aln (109..216)
names
[1] ----------------------CGCAGCGTTGAAAGGATGACGGAGGGCCGCCGATGTCAAGTACATCTTCTTGATGACAGGAAG
CTGGAACTCCTAGTACAGCCCAAG lcl|NC_000010.11_...
[2] ------------------------------------------------------------------------------------
------------------------ lcl|NC_000010.11_...
[3] --------------------------------ATGATGACGGAGGGCCGCCGATGTCAAGTACATCTTCTTGATGACAGGAAG
CTGGAACTCCTAGTACAGCCCAAG lcl|NC_000010.11_...
[4] CTAAAACCATGGTGCATCTCCAGGGCCGTCTATCAGATGACGGAGGGCCGCCGATGTCAAGTACATCTTCTTGATGACAGGAAG
CTGGAACTCCTAGTACAGCCCAAG lcl|NC_000010.11_...
Con ----------------------?---???--?-A?GATGACGGAGGGCCGCCGATGTCAAGTACATCTTCTTGATGACAGGAAG
CTGGAACTCCTAGTACAGCCCAAG Consensus

    aln (217..324)
names
[1] CTGTTGGCCAAGGAGCTTCTTGACCTTGTGGCTTCTCACTTCAATCTGAAGGAAAAGGAGTACTTTGGAATAGCATTCACAGAT
GAAACGGGACACTTAAACTGGCTT lcl|NC_000010.11_...
[2] ------------------------------------------------------------------------------------
------------------------ lcl|NC_000010.11_...
[3] CTGTTGGCCAAGGAGCTTCTTGACCTTGTGGCTTCTCACTTCAATCTGAAGGAAAAGGAGTACTTTGGAATAGCATTCACAGAT
GAAACGGGACACTTAAACTGGCTT lcl|NC_000010.11_...
[4] CTGTTGGCCAAGGAGCTTCTTGACCTTGTGGCTTCTCACTTCAATCTGAAGGAAAAGGAGTACTTTGGAATAGCATTCACAGAT
GAAACGGGACACTTAAACTGGCTT lcl|NC_000010.11_...
Con CTGTTGGCCAAGGAGCTTCTTGACCTTGTGGCTTCTCACTTCAATCTGAAGGAAAAGGAGTACTTTGGAATAGCATTCACAGAT
GAAACGGGACACTTAAACTGGCTT Consensus

    aln (325..432)
names
[1] CAGCTAGATCGAAGAGTATTGGAACATGACTTCCCTAAAAAGTCAGGACCCGTGGTTTTATACTTTTGTGTCAGGTTCTATATA
GAAAGCATTTCATACCTGAAGGAT lcl|NC_000010.11_...
[2] ------------------------------------------------------------------------------------
------------------------ lcl|NC_000010.11_...
[3] CAGCTAGATCGAAGAGTATTGGAACATGACTTCCCTAAAAAGTCAGGACCCGTGGTTTTATACTTTTGTGTCAGGTTCTATATA
GAAAGCATTTCATACCTGAAGGAT lcl|NC_000010.11_...
[4] CAGCTAGATCGAAGAGTATTGGAACATGACTTCCCTAAAAAGTCAGGACCCGTGGTTTTATACTTTTGTGTCAGGTTCTATATA
GAAAGCATTTCATACCTGAAGGAT lcl|NC_000010.11_...
Con CAGCTAGATCGAAGAGTATTGGAACATGACTTCCCTAAAAAGTCAGGACCCGTGGTTTTATACTTTTGTGTCAGGTTCTATATA
```

```
GAAAGCATTTCATACCTGAAGGAT Consensus

    aln (433..540)
names
[1] AATGCTACCATTGAGCTTTTCTTTCTGAACGCGAAGTCCTGCATCTACAAGGAGCTTATTGACGTTGACAGCGAAGTGGTGTTT
GAATTAGCTTCCTATATTTTACAG lcl|NC_000010.11_...
[2] ------------------------------------------------------------------------------------
------------------------ lcl|NC_000010.11_...
[3] AATGCTACCATTGAGCTTTTCTTTCTGAACGCGAAGTCCTGCATCTACAAGGAGCTTATTGACGTTGACAGCGAAGTGGTGTTT
GAATTAGCTTCCTATATTTTACAG lcl|NC_000010.11_...
[4] AATGCTACCATTGAGCTTTTCTTTCTGAACGCGAAGTCCTGCATCTACAAGGAGCTTATTGACGTTGACAGCGAAGTGGTGTTT
GAATTAGCTTCCTATATTTTACAG lcl|NC_000010.11_...
Con AATGCTACCATTGAGCTTTTCTTTCTGAACGCGAAGTCCTGCATCTACAAGGAGCTTATTGACGTTGACAGCGAAGTGGTGTTT
GAATTAGCTTCCTATATTTTACAG Consensus

    aln (541..648)
names
[1] GAGGCAAAGGGAGATTTTTCTAGCAATGAAGTTGTGAGGAGTGACTTGAAGAAGCTGCCAGCCCTTCCCACCCAAGCCCTGAAG
GAGCACCCTTCCCTGGCCTACTGT lcl|NC_000010.11_...
[2] ------------------------------------------------------------------------------------
------------------------ lcl|NC_000010.11_...
[3] GAGGCAAAGGGAGATTTTTCTAGCAATGAAGTTGTGAGGAGTGACTTGAAGAAGCTGCCAGCCCTTCCCACCCAAGCCCTGAAG
GAGCACCCTTCCCTGGCCTACTGT lcl|NC_000010.11_...
[4] GAGGCAAAGGGAGATTTTTCTAGCAATGAAGTTGTGAGGAGTGACTTGAAGAAGCTGCCAGCCCTTCCCACCCAAGCCCTGAAG
GAGCACCCTTCCCTGGCCTACTGT lcl|NC_000010.11_...
Con GAGGCAAAGGGAGATTTTTCTAGCAATGAAGTTGTGAGGAGTGACTTGAAGAAGCTGCCAGCCCTTCCCACCCAAGCCCTGAAG
GAGCACCCTTCCCTGGCCTACTGT Consensus

    aln (649..756)
names
[1] GAAGACAGAGTCATTGAGCACTACAAGAAACTGAACGGTCAGACAAGAGGTCAAGCAATCGTAAACTACATGAGCATCGTGGAG
TCTCTCCCAACCTACGGGGTTCAC lcl|NC_000010.11_...
[2] ------------------------------------------------------------------------------------
------------------------ lcl|NC_000010.11_...
[3] GAAGACAGAGTCATTGAGCACTACAAGAAACTGAACGGTCAGACAAGAGGTCAAGCAATCGTAAACTACATGAGCATCGTGGAG
TCTCTCCCAACCTACGGGGTTCAC lcl|NC_000010.11_...
[4] GAAGACAGAGTCATTGAGCACTACAAGAAACTGAACGGTCAGACAAGAGGTCAAGCAATCGTAAACTACATGAGCATCGTGGAG
TCTCTCCCAACCTACGGGGTTCAC lcl|NC_000010.11_...
Con GAAGACAGAGTCATTGAGCACTACAAGAAACTGAACGGTCAGACAAGAGGTCAAGCAATCGTAAACTACATGAGCATCGTGGAG
TCTCTCCCAACCTACGGGGTTCAC Consensus

    aln (757..864)
names
[1] TATTATGCAGTGAAGGACAAGCAGGGCATACCATGGTGGCTGGGCCTGAGCTACAAAGGGATCTTCCAGTATGACTACCATGAT
AAAGTGAAGCCAAGAAAGATATTC lcl|NC_000010.11_...
[2] ------------------------------------------------------------------------------------
------------------------ lcl|NC_000010.11_...
[3] TATTATGCAGTGAAGGACAAGCAGGGCATACCATGGTGGCTGGGCCTGAGCTACAAAGGGATCTTCCAGTATGACTACCATGAT
AAAGTGAAGCCAAGAAAGATATTC lcl|NC_000010.11_...
[4] TATTATGCAGTGAAGGACAAGCAGGGCATACCATGGTGGCTGGGCCTGAGCTACAAAGGGATCTTCCAGTATGACTACCATGAT
AAAGTGAAGCCAAGAAAGATATTC lcl|NC_000010.11_...
Con TATTATGCAGTGAAGGACAAGCAGGGCATACCATGGTGGCTGGGCCTGAGCTACAAAGGGATCTTCCAGTATGACTACCATGAT
```

```
AAAGTGAAGCCAAGAAAGATATTC Consensus

    aln (865..972)
names
[1] CAATGGAGACAGTTGGAAAACCTGTACTTCAGAGAAAAGAAGTTTTCCGTGGAAGTTCATGACCCACGCAGGGCTTCAGTGACA
AGGAGGACGTTTGGGCACAGCGGC lcl|NC_000010.11_...
[2] ----------------------------------------------------------------------------------
------------------------ lcl|NC_000010.11_...
[3] CAATGGAGACAGTTGGAAAACCTGTACTTCAGAGAAAAGAAGTTTTCCGTGGAAGTTCATGACCCACGCAGGGCTTCAGTGACA
AGGAGGACGTTTGGGCACAGCGGC lcl|NC_000010.11_...
[4] CAATGGAGACAGTTGGAAAACCTGTACTTCAGAGAAAAGAAGTTTTCCGTGGAAGTTCATGACCCACGCAGGGCTTCAGTGACA
AGGAGGACGTTTGGGCACAGCGGC lcl|NC_000010.11_...
Con CAATGGAGACAGTTGGAAAACCTGTACTTCAGAGAAAAGAAGTTTTCCGTGGAAGTTCATGACCCACGCAGGGCTTCAGTGACA
AGGAGGACGTTTGGGCACAGCGGC Consensus

    aln (973..1080)
names
[1] ATTGCAGTGCACACGTGGTATGCATGTCCGGCATTGATCAAGTCCATCTGGGCTATGGCCATAAGCCAACACCAGTTCTATCTG
GACAGAAAGCAGAGTAAGTCCAAA lcl|NC_000010.11_...
[2] ----------------------------------------------------ATGGCCATAAGCCAACACCAGTTCTATCTG
GACAGAAAGCAGAGTAAGTCCAAA lcl|NC_000010.11_...
[3] ATTGCAGTGCACACGTGGTATGCATGTCCGGCATTGATCAAGTCCATCTGGGCTATGGCCATAAGCCAACACCAGTTCTATCTG
GACAGAAAGCAGAGTAAGTCCAAA lcl|NC_000010.11_...
[4] ATTGCAGTGCACACGTGGTATGCATGTCCGGCATTGATCAAGTCCATCTGGGCTATGGCCATAAGCCAACACCAGTTCTATCTG
GACAGAAAGCAGAGTAAGTCCAAA lcl|NC_000010.11_...
Con ATTGCAGTGCACACGTGGTATGCATGTCCGGCATTGATCAAGTCCATCTGGGCTATGGCCATAAGCCAACACCAGTTCTATCTG
GACAGAAAGCAGAGTAAGTCCAAA Consensus

    aln (1081..1188)
names
[1] ATCCATGCAGCACGCAGCCTGAGTGAGATCGCCATCGACCTGACCGAGACGGGGACGCTGAAGACCTCGAAGCTGGCCAACATG
GGTAGCAAGGGGAAGATCATCAGC lcl|NC_000010.11_...
[2] ATCCATGCAGCACGCAGCCTGAGTGAGATCGCCATCGACCTGACCGAGACGGGGACGCTGAAGACCTCGAAGCTGGCCAACATG
GGTAGCAAGGGGAAGATCATCAGC lcl|NC_000010.11_...
[3] ATCCATGCAGCACGCAGCCTGAGTGAGATCGCCATCGACCTGACCGAGACGGGGACGCTGAAGACCTCGAAGCTGGCCAACATG
GGTAGCAAGGGGAAGATCATCAGC lcl|NC_000010.11_...
[4] ATCCATGCAGCACGCAGCCTGAGTGAGATCGCCATCGACCTGACCGAGACGGGGACGCTGAAGACCTCGAAGCTGGCCAACATG
GGTAGCAAGGGGAAGATCATCAGC lcl|NC_000010.11_...
Con ATCCATGCAGCACGCAGCCTGAGTGAGATCGCCATCGACCTGACCGAGACGGGGACGCTGAAGACCTCGAAGCTGGCCAACATG
GGTAGCAAGGGGAAGATCATCAGC Consensus

    aln (1189..1296)
names
[1] GGCAGCAGCGGCAGCCTGCTGTCTTCAGGTTCTCAGGAATCAGATAGCTCGCAGTCGGCCAAGAAGGACATGCTGGCTGCCTTG
AAGTCCAGGCAGGAAGCTCTGGAG lcl|NC_000010.11_...
[2] GGCAGCAGCGGCAGCCTGCTGTCTTCAGGTTCTCAGGAATCAGATAGCTCGCAGTCGGCCAAGAAGGACATGCTGGCTGCCTTG
AAGTCCAGGCAGGAAGCTCTGGAG lcl|NC_000010.11_...
[3] GGCAGCAGCGGCAGCCTGCTGTCTTCAGGTTCTCAGGAATCAGATAGCTCGCAGTCGGCCAAGAAGGACATGCTGGCTGCCTTG
AAGTCCAGGCAGGAAGCTCTGGAG lcl|NC_000010.11_...
[4] GGCAGCAGCGGCAGCCTGCTGTCTTCAGGTTCTCAGGAATCAGATAGCTCGCAGTCGGCCAAGAAGGACATGCTGGCTGCCTTG
AAGTCCAGGCAGGAAGCTCTGGAG lcl|NC_000010.11_...
Con GGCAGCAGCGGCAGCCTGCTGTCTTCAGGTTCTCAGGAATCAGATAGCTCGCAGTCGGCCAAGAAGGACATGCTGGCTGCCTTG
```

```
AAGTCCAGGCAGGAAGCTCTGGAG Consensus

    aln (1297..1404)
names
[1] GAAACCCTGCGTCAGAGGCTGGAGGAACTGAAGAAGCTGTGTCTCCGAGAAGCTGAGCTCACGGGCAAGCTGCCAGTAGAATAT
CCCCTGGATCCAGGGGAGGAACCA lcl|NC_000010.11_...
[2] GAAACCCTGCGTCAGAGGCTGGAGGAACTGAAGAAGCTGTGTCTCCGAGAAGCTGAGCTCACGGGCAAGCTGCCAGTAGAATAT
CCCCTGGATCCAGGGGAGGAACCA lcl|NC_000010.11_...
[3] GAAACCCTGCGTCAGAGGCTGGAGGAACTGAAGAAGCTGTGTCTCCGAGAAGCTGAGCTCACGGGCAAGCTGCCAGTAGAATAT
CCCCTGGATCCAGGGGAGGAACCA lcl|NC_000010.11_...
[4] GAAACCCTGCGTCAGAGGCTGGAGGAACTGAAGAAGCTGTGTCTCCGAGAAGCTGAGCTCACGGGCAAGCTGCCAGTAGAATAT
CCCCTGGATCCAGGGGAGGAACCA lcl|NC_000010.11_...
Con GAAACCCTGCGTCAGAGGCTGGAGGAACTGAAGAAGCTGTGTCTCCGAGAAGCTGAGCTCACGGGCAAGCTGCCAGTAGAATAT
CCCCTGGATCCAGGGGAGGAACCA Consensus

    aln (1405..1512)
names
[1] CCCATTGTTCGGAGAAGAATAGGAACAGCCTTCAAACTGGATGAACAGAAAATCCTGCCCAAAGGAGAGGAAGCTGAGCTGGAA
CGCCTGGAACGAGAGTTTGCCATT lcl|NC_000010.11_...
[2] CCCATTGTTCGGAGAAGAATAGGAACAGCCTTCAAACTGGATGAACAGAAAATCCTGCCCAAAGGAGAGGAAGCTGAGCTGGAA
CGCCTGGAACGAGAGTTTGCCATT lcl|NC_000010.11_...
[3] CCCATTGTTCGGAGAAGAATAGGAACAGCCTTCAAACTGGATGAACAGAAAATCCTGCCCAAAGGAGAGGAAGCTGAGCTGGAA
CGCCTGGAACGAGAGTTTGCCATT lcl|NC_000010.11_...
[4] CCCATTGTTCGGAGAAGAATAGGAACAGCCTTCAAACTGGATGAACAGAAAATCCTGCCCAAAGGAGAGGAAGCTGAGCTGGAA
CGCCTGGAACGAGAGTTTGCCATT lcl|NC_000010.11_...
Con CCCATTGTTCGGAGAAGAATAGGAACAGCCTTCAAACTGGATGAACAGAAAATCCTGCCCAAAGGAGAGGAAGCTGAGCTGGAA
CGCCTGGAACGAGAGTTTGCCATT Consensus

    aln (1513..1620)
names
[1] CAGTCCCAGATTACGGAGGCCGCCCGCCGCCTAGCCAGTGACCCCAACGTCAGCAAAAAACTGAAGAAACAAAGGAAAACCTCG
TATCTGAATGCACTGAAGAAACTG lcl|NC_000010.11_...
[2] CAGTCCCAGATTACGGAGGCCGCCCGCCGCCTAGCCAGTGACCCCAACGTCAGCAAAAAACTGAAGAAACAAAGGAAAACCTCG
TATCTGAATGCACTGAAGAAACTG lcl|NC_000010.11_...
[3] CAGTCCCAGATTACGGAGGCCGCCCGCCGCCTAGCCAGTGACCCCAACGTCAGCAAAAAACTGAAGAAACAAAGGAAAACCTCG
TATCTGAATGCACTGAAGAAACTG lcl|NC_000010.11_...
[4] CAGTCCCAGATTACGGAGGCCGCCCGCCGCCTAGCCAGTGACCCCAACGTCAGCAAAAAACTGAAGAAACAAAGGAAAACCTCG
TATCTGAATGCACTGAAGAAACTG lcl|NC_000010.11_...
Con CAGTCCCAGATTACGGAGGCCGCCCGCCGCCTAGCCAGTGACCCCAACGTCAGCAAAAAACTGAAGAAACAAAGGAAAACCTCG
TATCTGAATGCACTGAAGAAACTG Consensus

    aln (1621..1728)
names
[1] CAGGAGATTGAAAATGCAATCAATGAGAACCGCATCAAGTCTGGGAAGAAACCCACCCAGAGGGCTTCGCTGATCATAGACGAT
GGAAACATTGCCAGTGAAGACAGC lcl|NC_000010.11_...
[2] CAGGAGATTGAAAATGCAATCAATGAGAACCGCATCAAGTCTGGGAAGAAACCCACCCAGAGGGCTTCGCTGATCATAGACGAT
GGAAACATTGCCAGTGAAGACAGC lcl|NC_000010.11_...
[3] CAGGAGATTGAAAATGCAATCAATGAGAACCGCATCAAGTCTGGGAAGAAACCCACCCAGAGGGCTTCGCTGATCATAGACGAT
GGAAACATTGCCAGTGAAGACAGC lcl|NC_000010.11_...
[4] CAGGAGATTGAAAATGCAATCAATGAGAACCGCATCAAGTCTGGGAAGAAACCCACCCAGAGGGCTTCGCTGATCATAGACGAT
GGAAACATTGCCAGTGAAGACAGC lcl|NC_000010.11_...
Con CAGGAGATTGAAAATGCAATCAATGAGAACCGCATCAAGTCTGGGAAGAAACCCACCCAGAGGGCTTCGCTGATCATAGACGAT
```

```
GGAAACATTGCCAGTGAAGACAGC Consensus

    aln (1729..1836)
names
[1] TCCCTCTCAGATGCCCTTGTTCTTGAGGATGAAGACTCTCAGGTTACCAGCACAATATCCCCCCTACATTCTCCTCACAAGGGA
CTCCCTCCTCGGCCACCGTCGCAC lcl|NC_000010.11_...
[2] TCCCTCTCAGATGCCCTTGTTCTTGAGGATGAAGACTCTCAGGTTACCAGCACAATATCCCCCCTACATTCTCCTCACAAGGGA
CTCCCTCCTCGGCCACCGTCGCAC lcl|NC_000010.11_...
[3] TCCCTCTCAGATGCCCTTGTTCTTGAGGATGAAGACTCTCAGGTTACCAGCACAATATCCCCCCTACATTCTCCTCACAAGGGA
CTCCCTCCTCGGCCACCGTCGCAC lcl|NC_000010.11_...
[4] TCCCTCTCAGATGCCCTTGTTCTTGAGGATGAAGACTCTCAGGTTACCAGCACAATATCCCCCCTACATTCTCCTCACAAGGGA
CTCCCTCCTCGGCCACCGTCGCAC lcl|NC_000010.11_...
Con TCCCTCTCAGATGCCCTTGTTCTTGAGGATGAAGACTCTCAGGTTACCAGCACAATATCCCCCCTACATTCTCCTCACAAGGGA
CTCCCTCCTCGGCCACCGTCGCAC Consensus

    aln (1837..1944)
names
[1] AACAGGCCTCCTCCTCCCCAGTCCCTGGAGGGACTCCGACAGATGCACTATCACCGCAACGACTATGACAAGTCACCCATCAAG
CCCAAAATGTGGAGTGAGTCCTCT lcl|NC_000010.11_...
[2] AACAGGCCTCCTCCTCCCCAGTCCCTGGAGGGACTCCGACAGATGCACTATCACCGCAACGACTATGACAAGTCACCCATCAAG
CCCAAAATGTGGAGTGAGTCCTCT lcl|NC_000010.11_...
[3] AACAGGCCTCCTCCTCCCCAGTCCCTGGAGGGACTCCGACAGATGCACTATCACCGCAACGACTATGACAAGTCACCCATCAAG
CCCAAAATGTGGAGTGAGTCCTCT lcl|NC_000010.11_...
[4] AACAGGCCTCCTCCTCCCCAGTCCCTGGAGGGACTCCGACAGATGCACTATCACCGCAACGACTATGACAAGTCACCCATCAAG
CCCAAAATGTGGAGTGAGTCCTCT lcl|NC_000010.11_...
Con AACAGGCCTCCTCCTCCCCAGTCCCTGGAGGGACTCCGACAGATGCACTATCACCGCAACGACTATGACAAGTCACCCATCAAG
CCCAAAATGTGGAGTGAGTCCTCT Consensus

    aln (1945..2052)
names
[1] TTAGATGAACCCTATGAGAAGGTCAAGAAGCGCTCCTCTCACAGCCATTCCAGCAGCCACAAGCGCTTCCCCAGCACAGGAAGC
TGTGCGGAAGCCGGCGGAGGAAGC lcl|NC_000010.11_...
[2] TTAGATGAACCCTATGAGAAGGTCAAGAAGCGCTCCTCTCACAGCCATTCCAGCAGCCACAAGCGCTTCCCCAGCACAGGAAGC
TGTGCGGAAGCCGGCGGAGGAAGC lcl|NC_000010.11_...
[3] TTAGATGAACCCTATGAGAAGGTCAAGAAGCGCTCCTCTCACAGCCATTCCAGCAGCCACAAGCGCTTCCCCAGCACAGGAAGC
TGTGCGGAAGCCGGCGGAGGAAGC lcl|NC_000010.11_...
[4] TTAGATGAACCCTATGAGAAGGTCAAGAAGCGCTCCTCTCACAGCCATTCCAGCAGCCACAAGCGCTTCCCCAGCACAGGAAGC
TGTGCGGAAGCCGGCGGAGGAAGC lcl|NC_000010.11_...
Con TTAGATGAACCCTATGAGAAGGTCAAGAAGCGCTCCTCTCACAGCCATTCCAGCAGCCACAAGCGCTTCCCCAGCACAGGAAGC
TGTGCGGAAGCCGGCGGAGGAAGC Consensus

    aln (2053..2160)
names
[1] AACTCCTTGCAGAACAGCCCCATCCGCGGCCTCCCGCACTGGAACTCCCAGTCCAGCATGCCGTCCACGCCAGACCTGCGGGTC
CGGAGTCCCCACTACGTCCATTCC lcl|NC_000010.11_...
[2] AACTCCTTGCAGAACAGCCCCATCCGCGGCCTCCCGCACTGGAACTCCCAGTCCAGCATGCCGTCCACGCCAGACCTGCGGGTC
CGGAGTCCCCACTACGTCCATTCC lcl|NC_000010.11_...
[3] AACTCCTTGCAGAACAGCCCCATCCGCGGCCTCCCGCACTGGAACTCCCAGTCCAGCATGCCGTCCACGCCAGACCTGCGGGTC
CGGAGTCCCCACTACGTCCATTCC lcl|NC_000010.11_...
[4] AACTCCTTGCAGAACAGCCCCATCCGCGGCCTCCCGCACTGGAACTCCCAGTCCAGCATGCCGTCCACGCCAGACCTGCGGGTC
CGGAGTCCCCACTACGTCCATTCC lcl|NC_000010.11_...
Con AACTCCTTGCAGAACAGCCCCATCCGCGGCCTCCCGCACTGGAACTCCCAGTCCAGCATGCCGTCCACGCCAGACCTGCGGGTC
```

```
CGGAGTCCCCACTACGTCCATTCC Consensus

    aln (2161..2268)
names
[1] ACGAGGTCGGTGGACATCAGCCCCACCCGACTGCACAGCCTCGCACTGCACTTTAGGCACCGGAGCTCCAGCCTGGAGTCCCAG
GGCAAGCTCCTGGGCTCGGAAAAC lcl|NC_000010.11_...
[2] ACGAGGTCGGTGGACATCAGCCCCACCCGACTGCACAGCCTCGCACTGCACTTTAGGCACCGGAGCTCCAGCCTGGAGTCCCAG
GGCAAGCTCCTGGGCTCGGAAAAC lcl|NC_000010.11_...
[3] ACGAGGTCGGTGGACATCAGCCCCACCCGACTGCACAGCCTCGCACTGCACTTTAGGCACCGGAGCTCCAGCCTGGAGTCCCAG
GGCAAGCTCCTGGGCTCGGAAAAC lcl|NC_000010.11_...
[4] ACGAGGTCGGTGGACATCAGCCCCACCCGACTGCACAGCCTCGCACTGCACTTTAGGCACCGGAGCTCCAGCCTGGAGTCCCAG
GGCAAGCTCCTGGGCTCGGAAAAC lcl|NC_000010.11_...
Con ACGAGGTCGGTGGACATCAGCCCCACCCGACTGCACAGCCTCGCACTGCACTTTAGGCACCGGAGCTCCAGCCTGGAGTCCCAG
GGCAAGCTCCTGGGCTCGGAAAAC Consensus

    aln (2269..2376)
names
[1] GACACCGGGAGCCCCGACTTCTACACCCCGCGGACTCGTAGCAGCAACGGCTCAGACCCCATGGACGACTGCTCGTCGTGCACC
AGCCACTCGAGCTCGGAGCACTAC lcl|NC_000010.11_...
[2] GACACCGGGAGCCCCGACTTCTACACCCCGCGGACTCGTAGCAGCAACGGCTCAGACCCCATGGACGACTGCTCGTCGTGCACC
AGCCACTCGAGCTCGGAGCACTAC lcl|NC_000010.11_...
[3] GACACCGGGAGCCCCGACTTCTACACCCCGCGGACTCGTAGCAGCAACGGCTCAGACCCCATGGACGACTGCTCGTCGTGCACC
AGCCACTCGAGCTCGGAGCACTAC lcl|NC_000010.11_...
[4] GACACCGGGAGCCCCGACTTCTACACCCCGCGGACTCGTAGCAGCAACGGCTCAGACCCCATGGACGACTGCTCGTCGTGCACC
AGCCACTCGAGCTCGGAGCACTAC lcl|NC_000010.11_...
Con GACACCGGGAGCCCCGACTTCTACACCCCGCGGACTCGTAGCAGCAACGGCTCAGACCCCATGGACGACTGCTCGTCGTGCACC
AGCCACTCGAGCTCGGAGCACTAC Consensus

    aln (2377..2484)
names
[1] TACCCGGCGCAGATGAACGCCAACTACTCCACGCTGGCCGAGGACTCGCCGTCCAAGGCGCGCCAGAGGCAGAGGCAGCGGCAG
CGGGCGGCGGGCGCACTGGGCTCA lcl|NC_000010.11_...
[2] TACCCGGCGCAGATGAACGCCAACTACTCCACGCTGGCCGAGGACTCGCCGTCCAAGGCGCGCCAGAGGCAGAGGCAGCGGCAG
CGGGCGGCGGGCGCACTGGGCTCA lcl|NC_000010.11_...
[3] TACCCGGCGCAGATGAACGCCAACTACTCCACGCTGGCCGAGGACTCGCCGTCCAAGGCGCGCCAGAGGCAGAGGCAGCGGCAG
CGGGCGGCGGGCGCACTGGGCTCA lcl|NC_000010.11_...
[4] TACCCGGCGCAGATGAACGCCAACTACTCCACGCTGGCCGAGGACTCGCCGTCCAAGGCGCGCCAGAGGCAGAGGCAGCGGCAG
CGGGCGGCGGGCGCACTGGGCTCA lcl|NC_000010.11_...
Con TACCCGGCGCAGATGAACGCCAACTACTCCACGCTGGCCGAGGACTCGCCGTCCAAGGCGCGCCAGAGGCAGAGGCAGCGGCAG
CGGGCGGCGGGCGCACTGGGCTCA Consensus

    aln (2485..2592)
names
[1] GCCAGCTCGGGCAGCATGCCCAACCTGGCGGCGCGCGGGGGTGCGGGGGGCGCGGGGGGCGCGGGGGGCGGTGTGTACCTGCAC
AGCCAGAGCCAGCCCAGCTCGCAG lcl|NC_000010.11_...
[2] GCCAGCTCGGGCAGCATGCCCAACCTGGCGGCGCGCGGGGGTGCGGGGGGCGCGGGGGGCGCGGGGGGCGGTGTGTACCTGCAC
AGCCAGAGCCAGCCCAGCTCGCAG lcl|NC_000010.11_...
[3] GCCAGCTCGGGCAGCATGCCCAACCTGGCGGCGCGCGGGGGTGCGGGGGGCGCGGGGGGCGCGGGGGGCGGTGTGTACCTGCAC
AGCCAGAGCCAGCCCAGCTCGCAG lcl|NC_000010.11_...
[4] GCCAGCTCGGGCAGCATGCCCAACCTGGCGGCGCGCGGGGGTGCGGGGGGCGCGGGGGGCGCGGGGGGCGGTGTGTACCTGCAC
AGCCAGAGCCAGCCCAGCTCGCAG lcl|NC_000010.11_...
Con GCCAGCTCGGGCAGCATGCCCAACCTGGCGGCGCGCGGGGGTGCGGGGGGCGCGGGGGGCGCGGGGGGCGGTGTGTACCTGCAC
```

```
AGCCAGAGCCAGCCCAGCTCGCAG Consensus

    aln (2593..2700)
names
[1] TACCGCATCAAGGAGTACCCGCTGTACATCGAGGGCGGCGCCACGCCCGTGGTGGTGCGCAGCCTGGAGAGCGACCAGGAGGGC
CACTACAGCGTCAAGGCTCAGTTC lcl|NC_000010.11_...
[2] TACCGCATCAAGGAGTACCCGCTGTACATCGAGGGCGGCGCCACGCCCGTGGTGGTGCGCAGCCTGGAGAGCGACCAGGAGGGC
CACTACAGCGTCAAGGCTCAGTTC lcl|NC_000010.11_...
[3] TACCGCATCAAGGAGTACCCGCTGTACATCGAGGGCGGCGCCACGCCCGTGGTGGTGCGCAGCCTGGAGAGCGACCAGGAGGGC
CACTACAGCGTCAAGGCTCAGTTC lcl|NC_000010.11_...
[4] TACCGCATCAAGGAGTACCCGCTGTACATCGAGGGCGGCGCCACGCCCGTGGTGGTGCGCAGCCTGGAGAGCGACCAGGAGGGC
CACTACAGCGTCAAGGCTCAGTTC lcl|NC_000010.11_...
Con TACCGCATCAAGGAGTACCCGCTGTACATCGAGGGCGGCGCCACGCCCGTGGTGGTGCGCAGCCTGGAGAGCGACCAGGAGGGC
CACTACAGCGTCAAGGCTCAGTTC Consensus

    aln (2701..2808)
names
[1] AAGACGTCCAACTCCTACACGGCGGGCGGCCTGTTCAAGGAGAGCTGGCGCGGCGGCGGCGGCGACGAGGGCGACACGGGCCGC
CTGACGCCGTCGCGATCGCAGATC lcl|NC_000010.11_...
[2] AAGACGTCCAACTCCTACACGGCGGGCGGCCTGTTCAAGGAGAGCTGGCGCGGCGGCGGCGGCGACGAGGGCGACACGGGCCGC
CTGACGCCGTCGCGATCGCAGATC lcl|NC_000010.11_...
[3] AAGACGTCCAACTCCTACACGGCGGGCGGCCTGTTCAAGGAGAGCTGGCGCGGCGGCGGCGGCGACGAGGGCGACACGGGCCGC
CTGACGCCGTCGCGATCGCAGATC lcl|NC_000010.11_...
[4] AAGACGTCCAACTCCTACACGGCGGGCGGCCTGTTCAAGGAGAGCTGGCGCGGCGGCGGCGGCGACGAGGGCGACACGGGCCGC
CTGACGCCGTCGCGATCGCAGATC lcl|NC_000010.11_...
Con AAGACGTCCAACTCCTACACGGCGGGCGGCCTGTTCAAGGAGAGCTGGCGCGGCGGCGGCGGCGACGAGGGCGACACGGGCCGC
CTGACGCCGTCGCGATCGCAGATC Consensus

    aln (2809..2916)
names
[1] CTGCGGACTCCGTCGCTGGGCCGCGAGGGCGCCCACGACAAGGGCGCGGGCCGTGCCGCCGTCTCAGACGAGCTGCGCCAGTGG
TACCAGCGTTCCACCGCCTCGCAC lcl|NC_000010.11_...
[2] CTGCGGACTCCGTCGCTGGGCCGCGAGGGCGCCCACGACAAGGGCGCGGGCCGTGCCGCCGTCTCAGACGAGCTGCGCCAGTGG
TACCAGCGTTCCACCGCCTCGCAC lcl|NC_000010.11_...
[3] CTGCGGACTCCGTCGCTGGGCCGCGAGGGCGCCCACGACAAGGGCGCGGGCCGTGCCGCCGTCTCAGACGAGCTGCGCCAGTGG
TACCAGCGTTCCACCGCCTCGCAC lcl|NC_000010.11_...
[4] CTGCGGACTCCGTCGCTGGGCCGCGAGGGCGCCCACGACAAGGGCGCGGGCCGTGCCGCCGTCTCAGACGAGCTGCGCCAGTGG
TACCAGCGTTCCACCGCCTCGCAC lcl|NC_000010.11_...
Con CTGCGGACTCCGTCGCTGGGCCGCGAGGGCGCCCACGACAAGGGCGCGGGCCGTGCCGCCGTCTCAGACGAGCTGCGCCAGTGG
TACCAGCGTTCCACCGCCTCGCAC Consensus

    aln (2917..3024)
names
[1] AAGGAGCACAGCCGCCTGTCGCACACCAGCTCCACCTCCTCGGACAGCGGCTCGCAGTACAGCACCTCCTCCCAGAGCACCTTC
GTGGCGCACAGCAGGGTCACCAGG lcl|NC_000010.11_...
[2] AAGGAGCACAGCCGCCTGTCGCACACCAGCTCCACCTCCTCGGACAGCGGCTCGCAGTACAGCACCTCCTCCCAGAGCACCTTC
GTGGCGCACAGCAGGGTCACCAGG lcl|NC_000010.11_...
[3] AAGGAGCACAGCCGCCTGTCGCACACCAGCTCCACCTCCTCGGACAGCGGCTCGCAGTACAGCACCTCCTCCCAGAGCACCTTC
GTGGCGCACAGCAGGGTCACCAGG lcl|NC_000010.11_...
[4] AAGGAGCACAGCCGCCTGTCGCACACCAGCTCCACCTCCTCGGACAGCGGCTCGCAGTACAGCACCTCCTCCCAGAGCACCTTC
GTGGCGCACAGCAGGGTCACCAGG lcl|NC_000010.11_...
Con AAGGAGCACAGCCGCCTGTCGCACACCAGCTCCACCTCCTCGGACAGCGGCTCGCAGTACAGCACCTCCTCCCAGAGCACCTTC
```

```
GTGGCGCACAGCAGGGTCACCAGG Consensus

    aln (3025..3132)
names
[1] ATGCCCCAGATGTGCAAGGCCACGTCAGCTGCCTTACCTCAAAGCCAGAGAAGCTCGACACCGTCAAGTGAAATTGGAGCCACC
CCCCCAAGCAGCCCCCACCACATC lcl|NC_000010.11_...
[2] ATGCCCCAGATGTGCAAGGCCACGTCAGCTGCCTTACCTCAAAGCCAGAGAAGCTCGACACCGTCAAGTGAAATTGGAGCCACC
CCCCCAAGCAGCCCCCACCACATC lcl|NC_000010.11_...
[3] ATGCCCCAGATGTGCAAGGCCACGTCAGCTGCCTTACCTCAAAGCCAGAGAAGCTCGACACCGTCAAGTGAAATTGGAGCCACC
CCCCCAAGCAGCCCCCACCACATC lcl|NC_000010.11_...
[4] ATGCCCCAGATGTGCAAGGCCACGTCAGCTGCCTTACCTCAAAGCCAGAGAAGCTCGACACCGTCAAGTGAAATTGGAGCCACC
CCCCCAAGCAGCCCCCACCACATC lcl|NC_000010.11_...
Con ATGCCCCAGATGTGCAAGGCCACGTCAGCTGCCTTACCTCAAAGCCAGAGAAGCTCGACACCGTCAAGTGAAATTGGAGCCACC
CCCCCAAGCAGCCCCCACCACATC Consensus

    aln (3133..3219)
names
[1] CTAACCTGGCAGACTGGAGAAGCAACAGAAAACTCACCCATTCTGGATGGGTCTGAGTCTCCACCTCACCAAAGTACTGATGAA
TAG lcl|NC_000010.11_...
[2] CTAACCTGGCAGACTGGAGAAGCAACAGAAAACTCACCCATTCTGGATGGGTCTGAGTCTCCACCTCACCAAAGTACTGATGAA
TAG lcl|NC_000010.11_...
[3] CTAACCTGGCAGACTGGAGAAGCAACAGAAAACTCACCCATTCTGGATGGGTCTGAGTCTCCACCTCACCAAAGTACTGATGAA
TAG lcl|NC_000010.11_...
[4] CTAACCTGGCAGACTGGAGAAGCAACAGAAAACTCACCCATTCTGGATGGGTCTGAGTCTCCACCTCACCAAAGTACTGATGAA
TAG lcl|NC_000010.11_...
Con CTAACCTGGCAGACTGGAGAAGCAACAGAAAACTCACCCATTCTGGATGGGTCTGAGTCTCCACCTCACCAAAGTACTGATGAA
TAG Consensus
```

Hide

```
#read in the sequences of the 4 FRMD4A variants seperatly
V1 <- readAAStringSet("Variant_1.fasta")
V2 <- readAAStringSet("Variant_2.fasta")
V3 <- readAAStringSet("Variant_3.fasta")
V4 <- readAAStringSet("Variant_4.fasta")

# created a function that calculates the number of nucleotides in each of the sequences
 of the four variants nd assigned it to the variable "length_variants"
length_variants <- function(a) {
   result <- nchar(a)
   print(result)
}

#create a vector with all of the variants combined
variants <- c(V1, V2, V3, V4)

#run the function
length_variants(variants)
```

```
[1] 3120 3168 3219 2193
```

Hide

```
# created a second function that determines the difference between the sequence length i
n comparison with variant 1 and assigned it to the variable "length_difference"
length_difference <- function(a, b) {
   result <- nchar(a) - nchar(b)
   print(result)
}

#perofmed the function on all of the variants in comparison with Variant 1
length_difference(variants, V1)
```

```
[1]    0   48   99 -927
```

MSA PRETTY PRINT

msaPrettyPrint is the data analysis and visualization method that I chose to display my msa results. This analysis method highlights where all of the aligned sequences were conserved in blue. It also displays the sequence logo results at the top of the aligned sequences in order to display which nucleotides were conserved amongst all of the aligned sequences.

Hide

```
#use the msa package function "msaPrettyPrint()" to clean up the DNA sequence into a pub
lication worthy sequence alignment.
msaPrettyPrint(FRMD4A_MSA, output="tex", showNames="left", showLogo="top",
logoColors="rasmol", shadingMode="functional", shadingModeArg="structure",
showLegend=FALSE, askForOverwrite=FALSE)
```

HOMOLOGY MODELING

Homology modeling is a bioinformatics method that compares different protein structures and contrasts their similarities and differences. Homology modeling was performed on four pdb files that were extracted from the SWISS model database. The four pdb files contain the protein structure of the four variants of the FRMD4A gene aligned above.

Hide

```
#create a vector with the names of the 4 protein structures of the 4 variants and assign
to the variable ids
ids <- c("1e5w","3u8z", "5xq0", "4cgk")
```

Hide

```
#obrain the pdb files for the ids using the function get.pdb and assign to raw.files
raw.files <- get.pdb(ids, rm.alt=TRUE)
```

```
Warning in get.pdb(ids, rm.alt = TRUE) :
  ./1e5w.pdb  exists. Skipping download
Warning in get.pdb(ids, rm.alt = TRUE) :
  ./3u8z.pdb  exists. Skipping download
Warning in get.pdb(ids, rm.alt = TRUE) :
  ./5xq0.pdb  exists. Skipping download
Warning in get.pdb(ids, rm.alt = TRUE) :
  ./4cgk.pdb  exists. Skipping download
```

Hide

```
#code check to ensure we read in the correct pdb files
head(raw.files)
```

```
[1] "./1e5w.pdb" "./3u8z.pdb" "./5xq0.pdb" "./4cgk.pdb"
```

Hide

```
#split the four pdb files with their corresponding protein PDB code using the function p
dbsplit() and asign to splitpdbfiles
splitpdbfiles1 <- pdbsplit(raw.files, ids)
```

```
  |
  |
|   0%
  |
  |================================
|  25%
  |
  |=====================================
|  31%
  |
  |============================================
|  38%
  |
  |=================================================
|  44%
  |
  |============================================================
|  50%
  |
  |=======================================================================
|  62%
  |
  |================================================================================
========                                  |  75%
  |
  |================================================================================
=======================                   |  88%
  |
  |================================================================================
======================================|  100%
```

Hide

```
splitpdbfiles <- (splitpdbfiles1[c(1,2,6,8)])
```

Hide

```
#code check to ensure I only included structure A for each pdb
print(splitpdbfiles)
```

```
[1] "split_chain/1e5w_A.pdb" "split_chain/3u8z_A.pdb" "split_chain/5xq0_A.pdb" "split_ch
ain/4cgk_A.pdb"
```

Hide

```
#use the function pdbaln to align the protein structures for analysis. Assign to the var
iable pdbs. Use exefile = "msa" to avoid having to install muscle.
pdbs <- pdbaln(splitpdbfiles, exefile='msa')
```

```
Reading PDB files:
split_chain/1e5w_A.pdb
split_chain/3u8z_A.pdb
split_chain/5xq0_A.pdb
split_chain/4cgk_A.pdb
....

Extracting sequences

pdb/seq: 1   name: split_chain/1e5w_A.pdb
pdb/seq: 2   name: split_chain/3u8z_A.pdb
pdb/seq: 3   name: split_chain/5xq0_A.pdb
pdb/seq: 4   name: split_chain/4cgk_A.pdb
```

Hide

```
#code check to check the sequence idenity of the pdb files
summary(c(seqidentity(pdbs)))
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1270  0.1385  0.1690  0.4231  0.7345  1.0000
```

Hide

```
#perform normal mode analysis on the pdb structures
modes2 <- nma.pdbs(pdbs, fit = TRUE, outpath= NULL, chain = "A")
```

```
Warning in nma.pdbs(pdbs, fit = TRUE, outpath = NULL, chain = "A") :
  3u8z_A.pdb, 5xq0_A.pdb might have missing residue(s) in structure:
   Fluctuations at neighboring positions may be affected.
```

```
Details of Scheduled Calculation:
  ... 4 input structures
  ... storing 675 eigenvectors for each structure
  ... dimension of x$U.subspace: ( 681x675x4 )
  ... coordinate superposition prior to NM calculation
  ... aligned eigenvectors (gap containing positions removed)
  ... estimated memory usage of final 'eNMA' object: 14.1 Mb


  |
  |
|   0%
  |
  |==============================
|  25%
  |
  |============================================================
|  50%
  |
  |============================================================================================
========                                  |  75%
  |
  |============================================================================================
=======================================| 100%
```

Hide

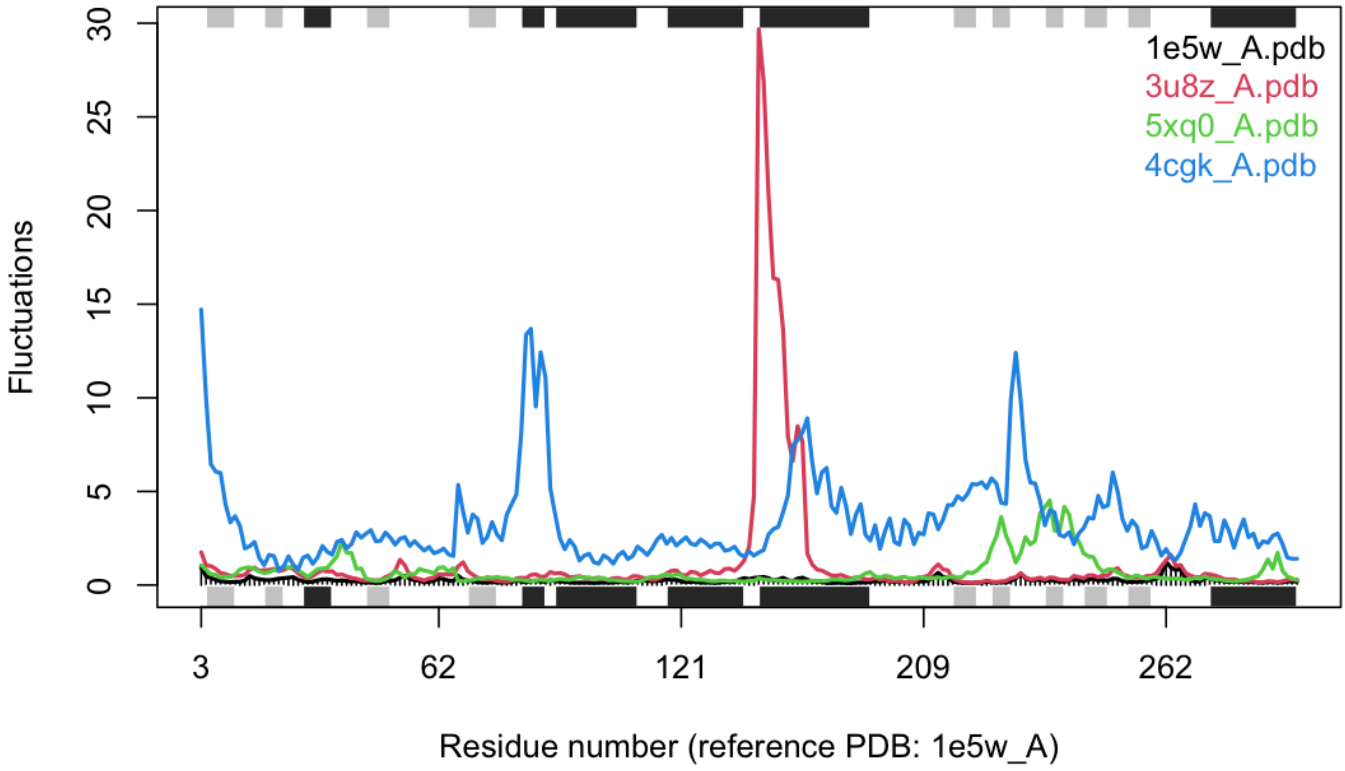
```
print(modes2)
```

```
Call:
  nma.pdbs(pdbs = pdbs, fit = TRUE, outpath = NULL, chain = "A")

Class:
  enma

Number of structures:
  4

Attributes stored:
  - Root mean square inner product (RMSIP)
  - Aligned atomic fluctuations
  - Aligned eigenvectors (gaps removed)
  - Dimensions of x$U.subspace: 681x675x4

Coordinates were aligned prior to NMA calculations

+ attr: fluctuations, rmsip, U.subspace, L, full.nma, xyz,
        call
```

Hide

```
#plot the fluctuation data between the pdb files
plot(modes2, pdbs, type="h")
```

```
Extracting SSE from pdbs$sse attribute
```



HEATMAP Heatmaps are a way to display the magnitude of data and is a great way to visualize the level of variations using colors and shades. In this case I am using the heatmap in order to display the magnitude of similarity between the modes of my four protein structures using RMISP values. The darker the color, the hgher the magnitude of similarity between the protein structures. The reason that modes are helpful is because they account for fluctations and take into acount the most common conformations of the proteins.

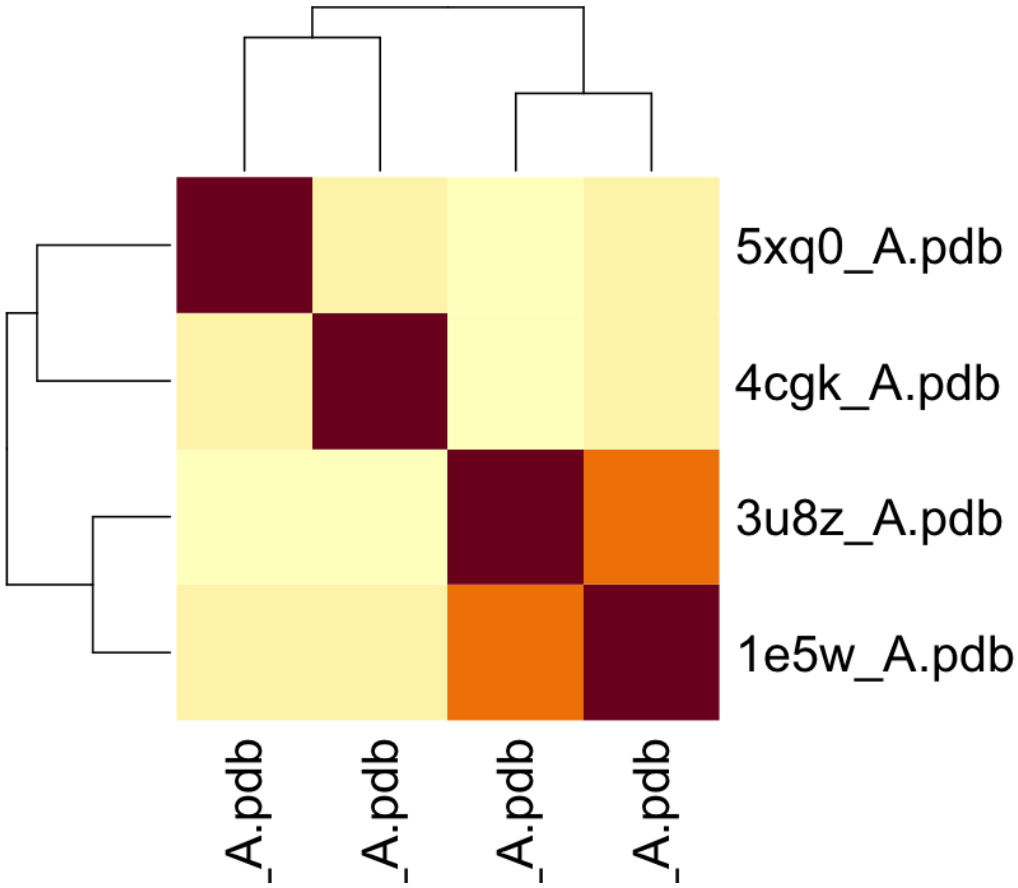
Hide

```
#Root mean square inner product
#RMSIP is a measure for the similarity between two set of modes obtained from principal
 component or normal modes analysis.
modes2$rmsip
```

```
           1e5w_A.pdb 3u8z_A.pdb 5xq0_A.pdb 4cgk_A.pdb
1e5w_A.pdb     1.0000     0.6569     0.3828     0.3794
3u8z_A.pdb     0.6569     1.0000     0.2741     0.2932
5xq0_A.pdb     0.3828     0.2741     1.0000     0.3854
4cgk_A.pdb     0.3794     0.2932     0.3854     1.0000
```

Hide

```
#creating a heat map in order to plot the RSMIP values of the four protein structures th
at were calculated above
#The darker the color of the square the higher the RSMIP value
heatmap(modes2$rmsip, symm = TRUE)
```

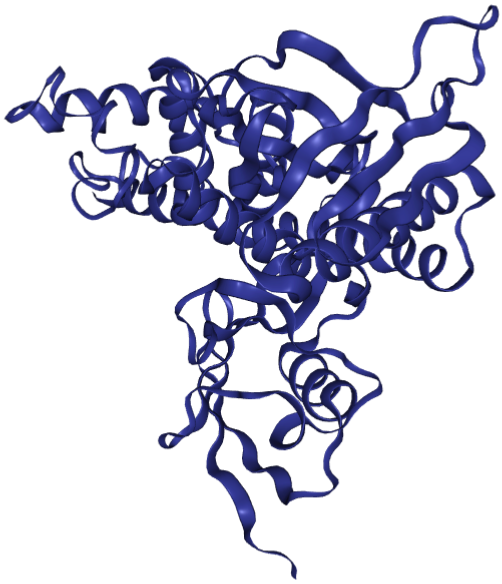


3D PROTEIN STRUCTURES

I am choosing to display my data by coding interactive 3D models for my four protein structures. This is a great way to visualize the protein structures and be able to move them around into different conformations. The structures were extracted using the 4 letter codes associated with each protein from the SWISS-Model database.

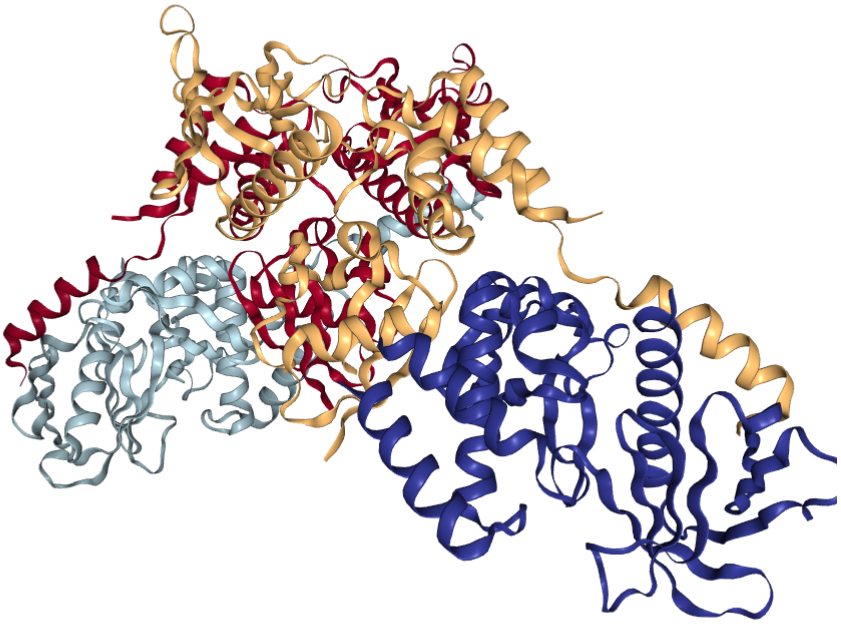
Hide

```
#using the function NGLVieweR we can extract and display protein #1 - 1w5w
NGLVieweR("1e5w") %>%
addRepresentation("cartoon")
```



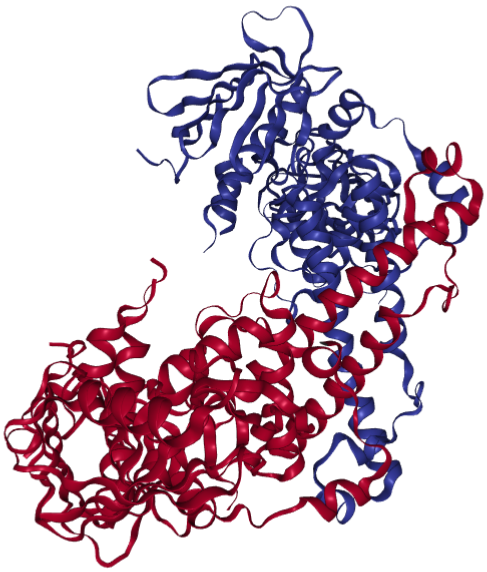
Hide

```
#using the function NGLVieweR we can extract and display protein #2 - 3u8z
NGLVieweR("3u8z") %>%
addRepresentation("cartoon")
```



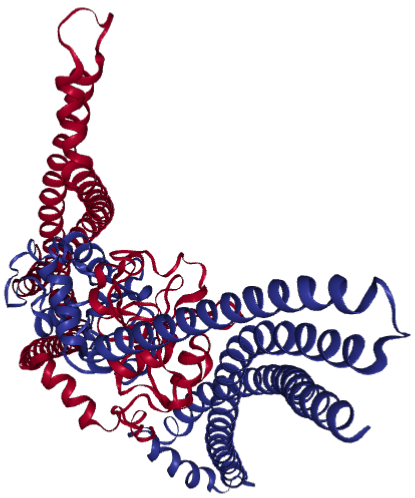
Hide

```
#using the function NGLVieweR we can extract and display protein #3 - 5xq0
NGLVieweR("5xq0") %>%
addRepresentation("cartoon")
```



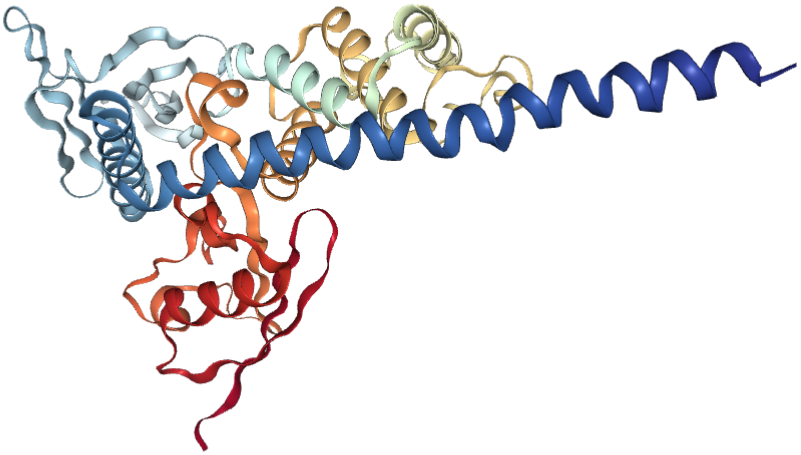
Hide

```
#using the function NGLVieweR we can extract and display protein #4 - 4cgk
NGLVieweR("4cgk") %>%
addRepresentation("cartoon")
```



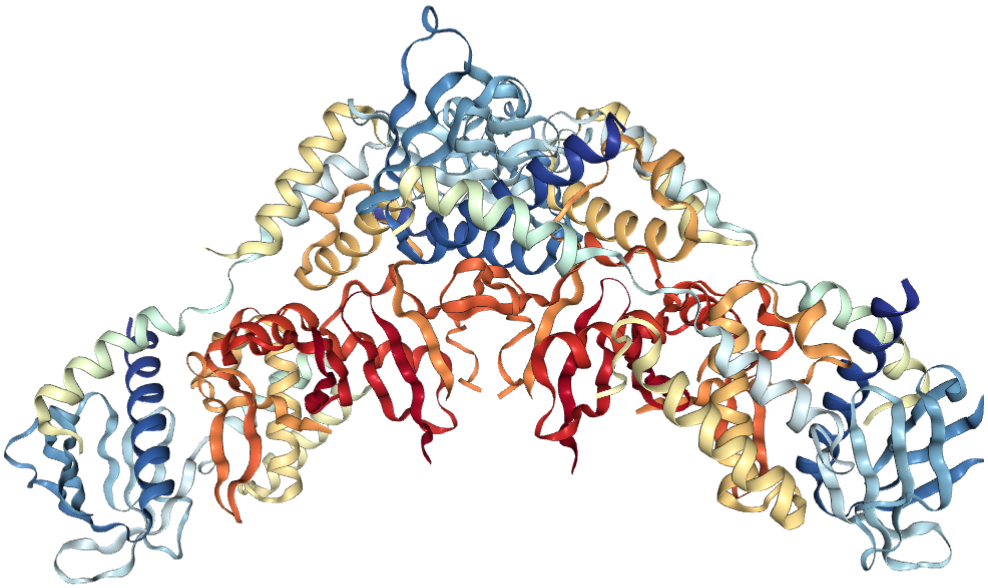
Hide

```
#using setSpin we can set the protein structure to continuously spin
NGLVieweR("1e5w") %>%
  stageParameters(backgroundColor = "white", zoomSpeed = 1) %>%
  addRepresentation("cartoon",
    param = list(name = "cartoon", colorScheme = "residueindex")
  ) %>%
  setSpin()
```



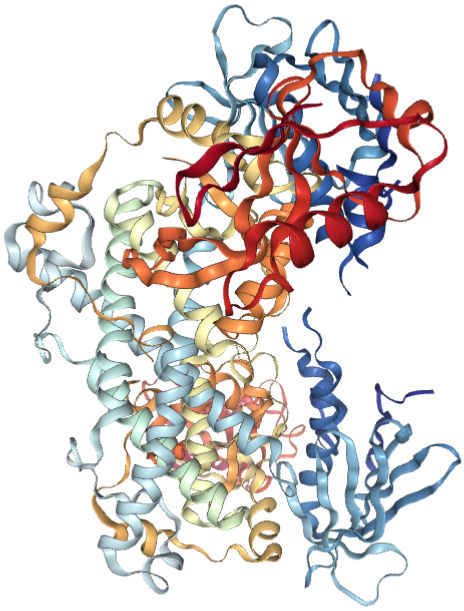
Hide

```
NGLVieweR("3u8z") %>%
  stageParameters(backgroundColor = "white", zoomSpeed = 1) %>%
  addRepresentation("cartoon",
    param = list(name = "cartoon", colorScheme = "residueindex")
  ) %>%
  setSpin()
```



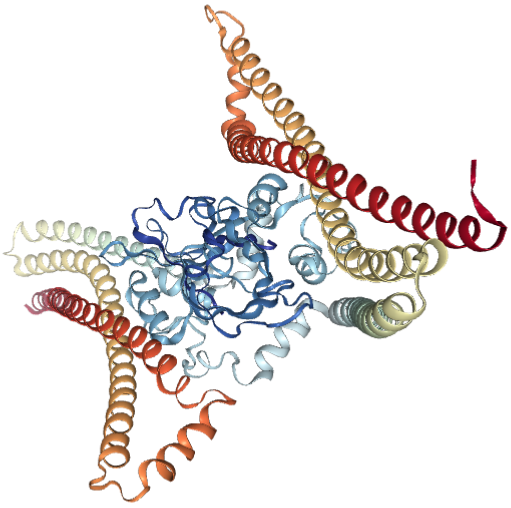
Hide

```
NGLVieweR("5xq0") %>%
  stageParameters(backgroundColor = "white", zoomSpeed = 1) %>%
  addRepresentation("cartoon",
    param = list(name = "cartoon", colorScheme = "residueindex")
  ) %>%
  setSpin()
```



Hide

```
NGLVieweR("4cgk") %>%
  stageParameters(backgroundColor = "white", zoomSpeed = 1) %>%
  addRepresentation("cartoon",
    param = list(name = "cartoon", colorScheme = "residueindex")
  ) %>%
  setSpin()
```



ANALYSIS OF DATA

The four sequences used in the multiple sequence alignment are four different isotope variants from the FRMD4A gene that were derived from the ncbi nuceliotide database. When multiple sequence alignment was performed the alignment revealed that the most amount of variation occurred at the beginning of the alignment and thus likely affective the enhancer and promoter regions of the FRMD4A gene. when compared to variant 1, variants 2 and 3 were longer sequences meanwhile variant 4 was significantly shorter. The sequence from Variant 2 did not begin until about 1kbp into the msa.

In order to test the effect of the different DNA sequences on the protein structure, Homology modeling was performed using may bio3d methods and functions. First the protein structures were read in as pdb files extracted from the SWISS model database, they structures were aligned and the mode analysis was performed. I specifically looked at the Root mean square inner product often abbreviated as RMSIP which yielded me a chart with scores from 1 - 0 with 1 being most identical and 0 being the least identical when it comes to mode and the structure conformations/ movement. The heatmap displays the results of the mode and RSMIP results well and shows that struture 1e5w (variant 1) which I used as the target protein was most similar to structure 3u8z (variant 2) their rismp score was 0.65 which is an excellent-fair score according to the journal of molecular graphics and modeling (David & Jacobs, 2011). When the structures of variants 3 and 4 were compared with variant 1 they yielded rmsip scores of 0.3828 and 0.3794 which are below the “fair” thershold determined above.The results displayed in the heat map can also be validated by the fluctuation plot which shows structure 3u8z (variant 2) to only have 1 main fluctuation when compared with structure 4cgk (variant 4) which has about 6 main fluctation peaks.

Overall, my hypothesis was correct that the mutations and alternate splicing that differentiated the four variants of the FRMD4A gene had an impact on the structure of the proteins which they translated as the protein structures were signifigantly different form one another due to their nucelotide sequence differences.

David, C. C., & Jacobs, D. J. (2011). Characterizing protein motions from structure. Journal of Molecular Graphics & Modelling, 31, 41–56. https://doi.org/10.1016/j.jmgm.2011.08.004 (https://doi.org/10.1016/j.jmgm.2011.08.004)