post02: Sequence Alignment Tools in R

Hannah Spinner

11/28/2017

Introduction & Background

Being a Genetics major at UC Berkeley means many things. I have rigorous coursework, impressive peers, and ground-breaking research all at my fingertips. I work in the Doudna Lab and the Savage Lab on engineering CRISPR proteins. CRISPR[1] is a cutting-edge molecular biology technique that was partly discovered here by Jennifer Doudna. These CRISPR-associated proteins (Cas proteins) are RNA-guided DNA endonucleases meaning that they cut DNA as a specified location. Do you want this one 'lil section of the vast human genome to be cut precisely and reliably? Well boy oh boy, do we have the tool for you! Just pop the Cas protein into the cell with a small piece of RNA that can heterodimerize with your DNA of interest and *pow pow* your DNA is cleaved!

To make these proteins and RNAs in a way that is easy to put into cells, we have to do a lot of cloning and a lot of sequence alignments[2]. Aligning sequences of DNA allows us to see if certain pieces of DNA have been change in the way we want (or in ways we didn't expect). Additionally, it can be used to see levels of similarity between proteins. For instance, there has been a lot of news about the CRISPR patent battle[3] between Berkeley and MIT. This patent[4] includes some statements about similarity between other CRISPR proteins and Cas9. For example, if I changed one amino acid residue on Cas9, I would not be able to get a new patent on it becuase it would still be 99% of the original Cas9. Sequence alignment technologies are hugely important for disputes like this and it is a vital process to modern molecular biology. While this aligned could potentially be with packages like stringr, it is much better to use packages designed specifically for this purpose for the alignment and use stringr to search within the alignment.

Motivation

The motivation for this post is to align amino acid and DNA sequences to look for similarities and differences between unique sequences.

Examples

Getting a Basic Alignment

First, we can look at a basic example dataset that comes with the msa package. We load the file and then we read the string of amino acids. This yields a table with the amino acid sequence and the name of the organism from which that protein is derived. Note: This is simply reading the sequence and nothing has been aligned yet.

```
Example1file <- system.file("examples", "exampleAA.fasta", package="msa")
Example1 <- readAAStringSet(Example1file)
Example1
```

```
##   A AAStringSet instance of length 9
##     width seq                                          names
## [1]   452 MSTAVLENPGLGRKLSDFGQE...LADSINSEIGILCSALQKIK PH4H_Homo_sapiens
## [2]   453 MAAVVLENGVLSRKLSDFGQE...ADSINSEVGILCNALQKIKS PH4H_Rattus_norve...
## [3]   453 MAAVVLENGVLSRKLSDFGQE...ADSINSEVGILCHALQKIKS PH4H_Mus_musculus
## [4]   297 MNDRADFVVPDITTRKNVGLS...PDDLVLNAGDRQGWADTEDV PH4H_Chromobacter...
## [5]   262 MKTTQYVARQPDDNGFIHYPE...VHEAMRLGLHAPLFPPKQAA PH4H_Pseudomonas_...
## [6]   451 MSALVLESRALGRKLSDFGQE...LADSISSEVEILCSALQKLK PH4H_Bos_taurus
## [7]   313 MAIATPTSAAPTPAPAGFTGT...DGDAVLNAGTREGWADTADI PH4H_Ralstonia_so...
## [8]   294 MSGDGLSNGPPPGARPDWTID...TRGTQAYATAGGRLAGAAAG PH4H_Caulobacter_...
## [9]   275 MSVAEYARDCAAQGLRGDYSV...DFEAIVARRKDQKALDPATV PH4H_Rhizobium_loti
```

To visualize the alignment we just use the msa function.

```
Example1Alignment <- msa(Example1)
```

```
## use default substitution matrix
```

```
Example1Alignment
```

```
## CLUSTAL 2.1
##
## Call:
##    msa(Example1)
##
## MsaAAMultipleAlignment with 9 rows and 456 columns
##     aln                                         names
## [1] MAAVVLENGVLSRKLSDFGQETS...KILADSINSEVGILCNALQKIKS PH4H_Rattus_norve...
## [2] MAAVVLENGVLSRKLSDFGQETS...KILADSINSEVGILCHALQKIKS PH4H_Mus_musculus
## [3] MSTAVLENPGLGRKLSDFGQETS...KILADSINSEIGILCSALQKIK- PH4H_Homo_sapiens
## [4] MSALVLESRALGRKLSDFGQETS...KILADSISSEVEILCSALQKLK- PH4H_Bos_taurus
## [5] -----------------------...AGDRQGWADTEDV---------- PH4H_Chromobacter...
## [6] -----------------------...AGTREGWADTADI---------- PH4H_Ralstonia_so...
## [7] -----------------------...RGT-QAYATAGGRLAGAAAG--- PH4H_Caulobacter_...
## [8] -----------------------...----------------------- PH4H_Pseudomonas_...
## [9] -----------------------...----------------------- PH4H_Rhizobium_loti
## Con -----------------------...????????????IL??A???--- Consensus
```

The output of this is the beginning and end of the sequences under the "aln" section accompanied by the name of each sequence in the "names" section. This is somewhat similar to the head() function for a data frame in that it will give the reader a gist of what is going on with the sequence alignment, but not the full story. Additionally, there is a new row called the “Consensus” sequence. This sequence includes all of the amino acids that more than 50% of the species share. This threshold can be changed in the settings of the line of code. This sequence is important because it can allow you to detect common motifs throughout the alignment.

Getting the Whole Sequence

As you can see from this simple example, though, the entire sequence is not shown. In some cases this is fine, but in others we would like to see the entire amino acid sequence alignment and investigate the middle regions. To remedy that, we can show the complete output with the ‘show’ setting.

```
print(Example1Alignment, show = 'complete')
```

```
##
## MsaAAMultipleAlignment with 9 rows and 456 columns
##     aln (1..49)                                       names
## [1] MAAVVLENGVLSRKLSDFGQETSYIEDNSNQNGAISLIFSLKEEVGALA PH4H_Rattus_norve...
## [2] MAAVVLENGVLSRKLSDFGQETSYIEDNSNQNGAVSLIFSLKEEVGALA PH4H_Mus_musculus
## [3] MSTAVLENPGLGRKLSDFGQETSYIEDNCNQNGAISLIFSLKEEVGALA PH4H_Homo_sapiens
## [4] MSALVLESRALGRKLSDFGQETSYIEGNSDQN-AVSLIFSLKEEVGALA PH4H_Bos_taurus
## [5] ------------------------------------------------- PH4H_Chromobacter...
## [6] ------------------------------------------------- PH4H_Ralstonia_so...
## [7] ------------------------------------------------- PH4H_Caulobacter_...
## [8] ------------------------------------------------- PH4H_Pseudomonas_...
## [9] ------------------------------------------------- PH4H_Rhizobium_loti
## Con ------------------------------------------------- Consensus
##
##     aln (50..98)                                      names
## [1] KVLRLFEENDINLTHIESRPSRLNKDEYEFFTYLDKRTKPVLGSIIKSL PH4H_Rattus_norve...
## [2] KVLRLFEENEINLTHIESRPSRLNKDEYEFFTYLDKRSKPVLGSIIKSL PH4H_Mus_musculus
## [3] KVLRLFEENDVNLTHIESRPSRLKKDEYEFFTHLDKRSLPALTNIIKIL PH4H_Homo_sapiens
## [4] RVLRLFEENDINLTHIESRPSRLRKDEYEFFTNLDQRSVPALANIIKIL PH4H_Bos_taurus
## [5] ------------------------------------------------- PH4H_Chromobacter...
## [6] ------------------------------------------------- PH4H_Ralstonia_so...
## [7] ------------------------------------------------- PH4H_Caulobacter_...
## [8] ------------------------------------------------- PH4H_Pseudomonas_...
## [9] ------------------------------------------------- PH4H_Rhizobium_loti
## Con ------------------------------------------------- Consensus
##
##     aln (99..147)                                     names
## [1] RNDIGATVHELSRDKEKNTVPWFPRTIQELDRFANQILSYGAELDADHP PH4H_Rattus_norve...
## [2] RNDIGATVHELSRDKEKNTVPWFPRTIQELDRFANQILSYGAELDADHP PH4H_Mus_musculus
## [3] RHDIGATVHELSRDKKKDTVPWFPRTIQELDRFANQILSYGAELDADHP PH4H_Homo_sapiens
## [4] RHDIGATVHELSRDKKKDTVPWFPRTIQELDNFANQVLSYGAELDADHP PH4H_Bos_taurus
## [5] ----------------------------------MNDRADFVVPD---- PH4H_Chromobacter...
## [6] -----------------------------MAIATPTSAAPTPAPAGFTG PH4H_Ralstonia_so...
## [7] ----------------------------------MSG------------ PH4H_Caulobacter_...
## [8] ------------------------------------------------- PH4H_Pseudomonas_...
## [9] ----------------------------------MSVAEYAR------- PH4H_Rhizobium_loti
## Con -----------------------------??????????Y????D???? Consensus
##
##     aln (148..196)                                    names
## [1] GFKDPVYRARRKQFADIAYNYRHGQPIPRVEYTEEEKQTWGTVFRTLKA PH4H_Rattus_norve...
## [2] GFKDPVYRARRKQFADIAYNYRHGQPIPRVEYTEEERKTWGTVFRTLKA PH4H_Mus_musculus
## [3] GFKDPVYRARRKQFADIAYNYRHGQPIPRVEYMEEEKKTWGTVFKTLKS PH4H_Homo_sapiens
## [4] GFKDPVYRARRKQFADIAYNYRHGQPIPRVEYTEEEKKTWGTVFRTLKS PH4H_Bos_taurus
## [5] -ITTRKNVGLSHDAN------DFTLPQPLDRYSAEDHATWATLYQRQCK PH4H_Chromobacter...
## [6] TLTDKLREQFAEGLDGQTLRPDFTMEQPVHRYTAADHATWRTLYDRQEA PH4H_Ralstonia_so...
## [7] ---DGLSNGPPPGAR-----PDWTIDQGWETYTQAEHDVWITLYERQTD PH4H_Caulobacter_...
## [8] ---MKTTQYVARQPD----------DNGFIHYPETEHQVWNTLITRQLK PH4H_Pseudomonas_...
## [9] ---DCAAQGLRGDYS--VCRADFTVAQDYD-YSDEEQAVWRTLCDRQTK PH4H_Rhizobium_loti
## Con ???D??????R?Q????????????P?P???YTEEE??TW?TL??RQ?? Consensus
##
##     aln (197..245)                                    names
## [1] LYKTHACYEHNHIFPLLEKYCGFREDNIPQLEDVSQFLQTCTGFRLRPV PH4H_Rattus_norve...
## [2] LYKTHACYEHNHIFPLLEKYCGFREDNIPQLEDVSQFLQTCTGFRLRPV PH4H_Mus_musculus
## [3] LYKTHACYEYNHIFPLLEKYCGFHEDNIPQLEDVSQFLQTCTGFRLRPV PH4H_Homo_sapiens
## [4] LYKTHACYEHNHIFPLLEKYCGFREDNIPQLEEVSQFLQSCTGFRLRPV PH4H_Bos_taurus
## [5] LLPGRACDEFMEGL----ERLEVDADRVPDFNKLNQKLMAATGWKIVAV PH4H_Chromobacter...
## [6] LLPGRACDEFLQGL----STLGMSREGVPSFDRLNETLMRATGWQIVAV PH4H_Ralstonia_so...
## [7] MLHGRACDEFMRGL----DALDLHRSGIPDFARINEELKRLTGWTVVAV PH4H_Caulobacter_...
## [8] VIEGRACQEYLDGI----EQLGLPHERIPQLDEINRVLQATTGWRVARV PH4H_Pseudomonas_...
## [9] LTRKLAHHSYLDGV----EKLGL-LDRIPDFEDVSTKLRKLTGWEIIAV PH4H_Rhizobium_loti
## Con L????AC?E???G?----??LG???D?IPQLE?VSQ?LQ??TGWR???V Consensus
##
##     aln (246..294)                                    names
## [1] AGLLSSRDFLGGLAFRVFHCTQYIRHGSKPMYTPEPDICHELLGHVPLF PH4H_Rattus_norve...
## [2] AGLLSSRDFLGGLAFRVFHCTQYIRHGSKPMYTPEPDICHELLGHVPLF PH4H_Mus_musculus
## [3] AGLLSSRDFLGGLAFRVFHCTQYIRHGSKPMYTPEPDICHELLGHVPLF PH4H_Homo_sapiens
## [4] AGLLSSRDFLGGLAFRVFHCTQYIRHGSKPMYTPEPDICHELLGHVPLF PH4H_Bos_taurus
## [5] PGLIPDDVFFEHLANRRFPVTWWLREPHQLDYLQEPDVFHDLFGHVPLL PH4H_Chromobacter...
## [6] PGLVPDEVFFEHLANRRFPASWWMRRPDQLDYLQEPDGFHDIFGHVPLL PH4H_Ralstonia_so...
## [7] PGLVPDDVFFDHLANRRFPAGQFIRKPHELDYLQEPDIFHDVFGHVPML PH4H_Caulobacter_...
## [8] PALIPFQTFFELLASQQFPVATFIRTPEELDYLQEPDIFHEIFGHCPLL PH4H_Pseudomonas_...
## [9] PGLIPAAPFFDHLANRRFPVTNWLRTRQELDYIVEPDMFHDFFGHVPVL PH4H_Rhizobium_loti
## Con PGL?P???FF??LA?R?FP?TQ?IR????LDY??EPDIFHELFGHVPLL Consensus
```

```
##
##      aln (295..343)                                    names
## [1] SDRSFAQFSQEIG-LASLGAPDEYIEKLATIYWFTVEFGLCKEG-DSIK PH4H_Rattus_norve...
## [2] SDRSFAQFSQEIG-LASLGAPDEYIEKLATIYWFTVEFGLCKEG-DSIK PH4H_Mus_musculus
## [3] SDRSFAQFSQEIG-LASLGAPDEYIEKLATIYWFTVEFGLCKQG-DSIK PH4H_Homo_sapiens
## [4] SDRSFAQFSQEIG-LASLGAPDEYIEKLATIYWFTVEFGLCKQG-DSIK PH4H_Bos_taurus
## [5] INPVFADYLEAYGKGGVKAKALGALPMLARLYWYTVEFGLINTP-AGMR PH4H_Chromobacter...
## [6] INPVFADYMQAYGQGGLKAARLGALDMLARLYWYTVEFGLIRTP-AGLR PH4H_Ralstonia_so...
## [7] TDPVFADYMQAYGEGGRRALGLGRLANLARLYWYTVEFGLMNTP-AGLR PH4H_Caulobacter_...
## [8] TNPWFAEFTHTYGKLGLKASKE-ERVFLARLYWMTIEFGLVETD-QGKR PH4H_Pseudomonas_...
## [9] SQPVFADFMQMYGKKAGDIIALGGDEMITRLYWYTAEYGLVQEAGQPLK PH4H_Rhizobium_loti
## Con SDP?FA?F?Q?YG?LA???A?????E?LARLYW?TVEFGL????-???K Consensus
##
##      aln (344..392)                                    names
## [1] AYGAGLLSSFGELQYCLSD-KPKLLPLELEKTACQEYSVTEFQPLYYVA PH4H_Rattus_norve...
## [2] AYGAGLLSSFGELQYCLSD-KPKLLPLELEKTACQEYTVTEFQPLYYVA PH4H_Mus_musculus
## [3] AYGAGLLSSFGELQYCLSE-KPKLLPLELEKTAIQNYTVTEFQPLYYVA PH4H_Homo_sapiens
## [4] AYGAGLLSSFGELQYCLSD-KPKLLPLELEKTAVQEYTITEFQPLYYVA PH4H_Bos_taurus
## [5] IYGAGILSSKSESIYCLDSASPNRVGFDLMRIMNTRYRIDTFQKTYFVI PH4H_Chromobacter...
## [6] IYGAGIVSSKSESVYALDSASPNRIGFDVHRIMRTRYRIDTFQKTYFVI PH4H_Ralstonia_so...
## [7] IYGAGIVSSRTESIFALDDPSPNRIGFDLERVMRTLYRIDDFQQVYFVI PH4H_Caulobacter_...
## [8] IYGGGILSSPKETVYSLSD-EPLHQAFNPLEAMRTPYRIDILQPLYFVL PH4H_Pseudomonas_...
## [9] AFGAGLMSSFTELQFAVEGKDAHHVPFDLETVMRTGYEIDKFQRAYFVL PH4H_Rhizobium_loti
## Con AYGAGLLSSF?ELQYCLSD-?P???PF?LE??M?T?Y?ID?FQPLYFV? Consensus
##
##      aln (393..441)                                    names
## [1] ESFSDAKEKVRTFAATIPRPFSVRYDPYTQRVEVLDNTQQLKILADSIN PH4H_Rattus_norve...
## [2] ESFNDAKEKVRTFAATIPRPFSVRYDPYTQRVEVLDNTQQLKILADSIN PH4H_Mus_musculus
## [3] ESFNDAKEKVRNFAATIPRPFSVRYDPYTQRIEVLDNTQQLKILADSIN PH4H_Homo_sapiens
## [4] ESFNDAKEKVRNFAATIPRPFSVHYDPYTQRIEVLDNTQQLKILADSIS PH4H_Bos_taurus
## [5] DSFKQLFDATA-PDFAPLYLQLADAQPWGAGDVAPDDLVLNAGDRQGWA PH4H_Chromobacter...
## [6] DSFEQLFDATR-PDFTPLYEALGTLPTFGAGDVVDGDAVLNAGTREGWA PH4H_Ralstonia_so...
## [7] DSIQTLQEVTL-RDFGAIYERLASVSDIGVAEIVPGDAVLTRGT-QAYA PH4H_Caulobacter_...
## [8] PDLKRLFQLAQ-EDIMALVHEAMRLG-LHAPLFPPKQAA---------- PH4H_Pseudomonas_...
## [9] PSFDALRDAFQTADFEAIVARRKDQKALDPATV---------------- PH4H_Rhizobium_loti
## Con ?SF??L?E??R??D?T??????????P??????V?D????????????? Consensus
##
##      aln (442..456)  names
## [1] SEVGILCNALQKIKS PH4H_Rattus_norve...
## [2] SEVGILCHALQKIKS PH4H_Mus_musculus
## [3] SEIGILCSALQKIK- PH4H_Homo_sapiens
## [4] SEVEILCSALQKLK- PH4H_Bos_taurus
## [5] DTEDV---------- PH4H_Chromobacter...
## [6] DTADI---------- PH4H_Ralstonia_so...
## [7] TAGGRLAGAAAG--- PH4H_Caulobacter_...
## [8] --------------- PH4H_Pseudomonas_...
## [9] --------------- PH4H_Rhizobium_loti
## Con ????IL??A???--- Consensus
```

The benefit of visualizing the data like this is that you can see it all at once. This shows all of the organisms and all of their protein sequences lined up. This is much more powerful information than simply the beginning and end of a sequence. With this, you can find specific sequences within your data set and see if they have changed.

Searching Within the Sequence

Searching for Intentional Mutations

For example, say I wanted to create a single amino acid substitution at the at the 101-111th position of Rattus, Mus, Homo, and Bos. I could first use the grep function to search within that conserved region of the orginal sequencing using the grep function within the stringr package.

```
grep("DIGATVHELSRDK", Example1Alignment)
```

```
## [1] 1 2 3 4
```

This tells us the that pattern is in the first, second, third, and fourth (Rattus, Mus, Homo, and Bos) sequences. Now, if I were to make my mutation in the lab and re-run this program with the new sequences, I could see if any of them have a new pattern here. For instance, if I wanted to change the first alanine (‘A’) residue to a leucine (‘L’), I would use the pattern “DIG**L**TVHELSRDK” instead of the original “DIG**A**TVHELSRDK.” Becuase I haven’t actually made that amino acid change, I cannot search for it inside of these sequences.

Searching for Specific Short Sequences

Similarly, say that alanine (‘A’), arginine (‘R’) and leucine (‘L’) is an important region for a twist in an alpha helix or beta pleated sheet. If you wanted to see which organisms out of your population could produce this sequence, you could use grepexpr functoin to search the sequence alignment for that.

```
position_of_ARL = gregexpr('ARL', Example1Alignment)
position_of_ARL
```

```
## [[1]]
## [1] -1
## attr(,"match.length")
## [1] -1
## attr(,"useBytes")
## [1] TRUE
##
## [[2]]
## [1] -1
## attr(,"match.length")
## [1] -1
## attr(,"useBytes")
## [1] TRUE
##
## [[3]]
## [1] -1
## attr(,"match.length")
## [1] -1
## attr(,"useBytes")
## [1] TRUE
##
## [[4]]
## [1] -1
## attr(,"match.length")
## [1] -1
## attr(,"useBytes")
## [1] TRUE
##
## [[5]]
## [1] 323
## attr(,"match.length")
## [1] 3
## attr(,"useBytes")
## [1] TRUE
##
## [[6]]
## [1] 314 323
## attr(,"match.length")
## [1] 3 3
## attr(,"useBytes")
## [1] TRUE
##
## [[7]]
## [1] 323
## attr(,"match.length")
## [1] 3
## attr(,"useBytes")
## [1] TRUE
##
## [[8]]
## [1] 323
## attr(,"match.length")
## [1] 3
## attr(,"useBytes")
## [1] TRUE
##
## [[9]]
## [1] -1
## attr(,"match.length")
## [1] -1
## attr(,"useBytes")
## [1] TRUE
```

This tells you that sequences 1, 2, 3, 4, and 9 do not have it. But, sequences 5, 6, 7, and 8 all have it at the 323rd position. This example also emphasizes the importance of aligning the sequences before searching within them. For instance, let's see what happens if we search through unalingned sequences.

```
position_of_unaligned_ARL = gregexpr('ARL', Example1)
position_of_unaligned_ARL
```

```
## [[1]]
## [1] -1
## attr(,"match.length")
## [1] -1
## attr(,"useBytes")
## [1] TRUE
##
## [[2]]
## [1] -1
## attr(,"match.length")
## [1] -1
## attr(,"useBytes")
## [1] TRUE
##
## [[3]]
## [1] -1
## attr(,"match.length")
## [1] -1
## attr(,"useBytes")
## [1] TRUE
##
## [[4]]
## [1] 176
## attr(,"match.length")
## [1] 3
## attr(,"useBytes")
## [1] TRUE
##
## [[5]]
## [1] 158
## attr(,"match.length")
## [1] 3
## attr(,"useBytes")
## [1] TRUE
##
## [[6]]
## [1] -1
## attr(,"match.length")
## [1] -1
## attr(,"useBytes")
## [1] TRUE
##
## [[7]]
## [1] 183 192
## attr(,"match.length")
## [1] 3 3
## attr(,"useBytes")
## [1] TRUE
##
## [[8]]
## [1] 167
## attr(,"match.length")
## [1] 3
## attr(,"useBytes")
## [1] TRUE
##
## [[9]]
## [1] -1
## attr(,"match.length")
## [1] -1
## attr(,"useBytes")
## [1] TRUE
```

Note: it might seem odd that all of a sudden sequence number 4 now contains “ARG” where it didn’t before, but remember that the unaligned sequences are in a different order than the aligned ones.

Here we can see that when the sequences aren’t aligned, it appears as if ARG is at all different loci in each sequence. This is because each protein is of slightly difference length, so before they are aligned by functional units, the positions will be slightly off. If the scientist did, however, want the sequence position independent of other sequences, this would be useful.

Just for Fun

You can also look for fun words within the sequence, like a word search. Let’s see if my name appears anywhere in the sequence!

```
position_of_HANNAH = gregexpr("HANNAH", Example1Alignment)
position_of_HANNAH
```

```
## [[1]]
## [1] -1
## attr(,"match.length")
## [1] -1
## attr(,"useBytes")
## [1] TRUE
##
## [[2]]
## [1] -1
## attr(,"match.length")
## [1] -1
## attr(,"useBytes")
## [1] TRUE
##
## [[3]]
## [1] -1
## attr(,"match.length")
## [1] -1
## attr(,"useBytes")
## [1] TRUE
##
## [[4]]
## [1] -1
## attr(,"match.length")
## [1] -1
## attr(,"useBytes")
## [1] TRUE
##
## [[5]]
## [1] -1
## attr(,"match.length")
## [1] -1
## attr(,"useBytes")
## [1] TRUE
##
## [[6]]
## [1] -1
## attr(,"match.length")
## [1] -1
## attr(,"useBytes")
## [1] TRUE
##
## [[7]]
## [1] -1
## attr(,"match.length")
## [1] -1
## attr(,"useBytes")
## [1] TRUE
##
## [[8]]
## [1] -1
## attr(,"match.length")
## [1] -1
## attr(,"useBytes")
## [1] TRUE
##
## [[9]]
## [1] -1
## attr(,"match.length")
## [1] -1
## attr(,"useBytes")
## [1] TRUE
```

Dang, foiled again!

Getting Sequence Alignments from DNA sequences

CRISPR

So far we have looked at aligning amino acid sequences from the package database. However, what if we wanted to compare DNA sequences relavent to my research? Below are sequences of Cas9[5] and a newly discovered CRISPR protein CasX[6]. If we want to compare these sequences, we can use the DECIPHER package[7] to translate these sequences to amino acids and then align them.

```
Cas9_sequence = "ATGCGTTATAAGATTGGCCTGGACATCGGTATTACCTCTGTTGGTTGGGCAGTCATGAACCTGGATATCCCTCGTATCGAAGATCTGGGCGTGCGCA
TTTTTGACCGTGCGGAGAACCCGCAGACCGGTGAATCTCTGGCTCTGCCGCGTCGTCTGGCACGTAGCGCACGCCGCCGCCTGCGTCGTCGTAAACACCGTCTGGAGCGTATTC
GTCGCCTGGTTATTCGTGAAGGCATCCTGACGAAAGAAGAACTGGATAAACTGTTCGAAGAAAAACACGAGATCGACGTATGGCAGCTGCGTGTAGAAGCCCTGGACCGTAAGC
TGAACAACGACGAACTGGCGCGTGTCCTGCTGCATCTGGCAAAGCGTCGTGGCTTCAAATCTAACCGTAAATCTGAACGCTCCAATAAAGAGAACTCCACTATGCTGAAACATA
TTGAGGAGAACCGTGCAATTCTGTCTAGCTACCGTACCGTGGGCGAAATGATTGTTAAAGACCCGAAATTCGCACTGCATAAGCGTAACAAAGGCGAAAACTACACCAACACCA
TTGCACGCGATGACCTGGAACGTGAAATCCGTCTGATTTTCTCCAAACAGCGCGAATTCGGCAACATGTCTTGCACCGAAGAATTCGAAAACGAATATATTACCATTTGGGCAT
CTCAGCGTCCGGTGGCGTCTAAAGATGATATCGAAAAAAAAGTAGGCTTTTGTACTTTCGAACCGAAGGAAAAACGTGCGCCGAAAGCCACCTATACCTTCCAGTCTTTTATCG
CGTGGGAACATATCAACAAACTGCGTCTGATTTCTCCGTCTGGCGCCCGCGGCCTGACCGACGAAGAACGTCGTCTGCTGTATGAACAAGCATTCCAGAAAAACAAAATTACCT
ACCACGATATTCGTACCCTGCTGCATCTGCCGGACGACACCTACTTCAAGGGCATCGTTTACGATCGCGGTGAATCTCGTAAGCAGAACGAAAACATTCGTTTCCTGGAACTGG
ATGCATACCACCAGATCCGTAAAGCTGTAGATAAAGTTTACGGCAAGGGTAAATCCAGCAGCTTCCTGCCGATCGACTTTGATACCTTCGGTTACGCGCTGACCCTGTTTAAAG
ACGATGCGGATATCCACTCTTACCTGCGCAACGAGTACGAACAGAACGGCAAACGTATGCCTAACCTGGCTAACAAAGTTTACGATAACGAGCTGATTGAAGAACTGCTGAACC
TGTCCTTCACTAAATTCGGTCACCTGTCTCTGAAAGCTCTGCGTTCCATCCTGCCGTATATGGAACAGGGTGAAGTCTACTCCTCCGCTTGTGAACGTGCAGGCTACACCTTCA
CCGGTCCGAAAAAGAAGCAAAAACTATGCTGCTGCCGAACATCCCGCCGATTGCGAACCCTGTAGTAATGCGTGCACTGACCCAGGCGCGCAAAGTAGTCAACGCGATCATCA
AAAAGTACGGCAGCCCGGTTTCCATCCATATCGAACTGGCGCGCGACCTGAGCCAGACTTTTGACGAGCGTCGTAAAACTAAAAAGGAACAGGATGAAAACCGTAAAAAAAACG
AAACCGCGATCCGCCAGCTGATGGAATACGGTCTGACTCTGAACCCTACTGGTCACGATATTGTGAAGTTCAAGCTGTGGTCTGAACAGAACGGTCGCTGTGCTTACTCTCTGC
AGCCGATCGAGATCGAACGTCTGCTGGAGCCAGGTTACGTTGAAGTAGATCATGTGATCCCGTACTCCCGCTCTCTGGATGATTCTTATACCAACAAAGTTCTGGTTCTGACTC
GCGAAAACCGTGAGAAAGGCAACCGCATCCCAGCTGAATATCTGGGTGTTGGCACTGAGCGTTGGCAACAGTTCGAAACCTTCGTCCTGACCAATAAACAGTTCTCTAAAAAGA
AACGTGACCGTCTGCTGCGTCTGCACTACGATGAAAACGAAGAGACTGAATTCAAAAACCGTAACCTGAACGATACTCGCTACATCAGCCGCTTCTTCGCAAACTTCATTCGTG
AACACCTGAAATTTGCGGAATCCGACGATAAACAGAAAGTTTATACCGTAAACGGCCGTGTTACCGCCCACCTGCGTTCTCGCTGGGAGTTCAACAAGAACCGTGAGGAAAGCG
ATCTGCACCACGCTGTTGACGCCGTTATTGTGGCGTGCACCACCCCAAGCGATATCGCTAAGGTGACCGCATTCTACCAGCGTCGTGAGCAGAACAAGGAACTGGCCAAAAAAA
CCGAACCGCATTTTCCGCAGCCGTGGCCGCACTTCGCGGACGAACTGCGTGCTCGTCTGTCCAAACATCCTAAAGAAAGCATCAAAGCTCTGAACCTGGGTAACTACGATGACC
AAAAACTGGAATCTCTGCAGCCGGTGTTTGTCAGCCGTATGCCGAAACGTTCTGTTACTGGCGCTGCGCACCAGGAAACGCTGCGCCGTTACGTGGGCATCGACGAACGCTCCG
GTAAAATCCAGACCGTAGTAAAAACCAAACTGTCCGAGATTAAACTGGATGCATCCGGCCACTTCCCAATGTACGGTAAAGAATCCGATCCACGCACTTATGAAGCCATCCGCC
AGCGTCTGCTGGAGCATAACAACGACCCGAAAAAGGCATTCCAGGAGCCTCTGTACAAACCGAAAAAAAACGGCGAACCGGGCCCGGTAATCCGTACTGTAAAAATTATCGACA
CGAAAAACCAGGTGATCCCTCTGAACGATGGTAAAACCGTGGCCTACAATTCCAACATCGTTCGCGTGGACGTGTTCGAAAAAGATGGTAAATACTACTGTGTACCGGTGTATA
CCATGGACATCATGAAAGGCATTCTGCCGAACAAAGCGATTGAACCGAACAAGCCGTACTCTGAATGGAAAGAAATGACCGAAGATTACACGTTTCGTTTCAGCCTGTATCCGA
ACGACCTGATCCGCATCGAACTGCCGCGTGAAAAAACCGTTAAAACCGCTGCAGGCGAAGAAATTAACGTGAAAGACGTGTTCGTTTACTATAAAACGATCGACTCCGCAAACG
GCGGCCTGGAACTGATTTCTCACGACCACCGTTTCTCTCTGCGTGGCGTTGGCTCTCGCACCCTGAAACGTTTCGAGAAATATCAAGTTGATGTTCTGGGTAACATCTATAAAG
TGCGTGGCGAGAAACGTGTCGGTCTGGCGTCCTCCGCACACAGCAAACCTGGCAAAACCATTCGTCCACTGCAATCTACTCGTGACTAA"

CasX1_sequence = "ATGGAGAAGCGTATTAACAAGATTCGTAAGAAGTTATCGGCTGATAATGCCACAAAACCTGTAAGCCGCTCAGGACCTATGAAAACGCTTTTAGTT
CGTGTTATGACAGATGATCTGAAAAAACGTCTTGAGAAGCGTCGTAAAAAGCCGGAGGTCATGCCTCAAGTCATTTCGAATAATGCCGCCAATAATTTACGTATGCTTCTTGAT
GACTATACAAAGATGAAAGAAGCAATCCTGCAAGTCTACTGGCAAGAGTTCAAAGATGATCATGTAGGTTTGATGTGTAAGTTTGCACAGCCCGCTAGTAAAAAGATCGACCAA
AACAAGTTAAAGCCAGAGATGGACGAGAAAGGGAATCTTACCACTGCCGGGTTCGCTTGCTCTCAATGCGGCCAACCGCTGTTCGTATATAAGCTGGAGCAAGTTTCAGAAAAG
GGTAAGGCATACACAAATTATTTCGGACGTTGCAACGTCGCTGAGCACGAGAAGCTTATTCTTTTAGCCCAGCTGAAACCAGAGAAGGACTCTGATGAAGCGGTTACCTACTCC
TTGGGAAAGTTCGGACAGCGCGCTCTTGACTTTTATTCCATCCACGTGACAAAGGAAAGTACACACCCCGTCAAACCACTGGCCCAAATTGCAGGGAACCGCTACGCGAGCGGG
CCAGTGGGAAAAGCATTGTCAGACGCTTGTATGGGCACGATTGCTTCCTTTCTTAGTAAATACCAGGATATTATCATTGAACATCAAAAGGTAGTTAAGGGGAACCAGAAGCGT
CTGGAAAGCTTACGTGAACTTGCAGGGAAAGAGAACCTTGAGTACCCATCGGTCACATTGCCTCCGCAGCCGCACACAAAGGAGGGTGTCGATGCTTATAACGAGGTGATTGCA
CGCGTACGCATGTGGGTAAATTTGAATCTGTGGCAGAAGTTGAAACTGTCGCGTGATGACGCCAAACCTCTTTTACGCCTGAAGGGCTTCCCCTCATTTCCAGTCGTAGAGCGC
CGTGAAAATGAAGTAGACTGGTGGAATACTATCAACGAAGTAAAAAAGCTGATTGACGCGAAGCGCGACATGGACGCGTGTTTTGGAGCGGCGTTACTGCTGAAAAGCGCAAC
ACCATCCTTGAAGGATACAACTATTTACCCAACGAGAACGATCATAAAAAGCGCGAGGGCAGTTTAGAGAATCCGAAGAAGCCGGCCAAACGTCAGTTTGGCGACCTTTTATTA
TACCTGGAAAAGAAATATGCGGGTGACTGGGGCAAAGTGTTCGATGAAGCCTGGGAGCGCATTGACAAGAAAATCGCTGGTCTGACGTCGCACATTGAGCGCGAGGAAGCTCGT
AACGCAGAAGACGCCCAATCCAAAGCGGTCCTGACAGACTGGCTGCGCGCTAAGGCGTCCTTTGTTCTTGAGCGTTTAAAAGAGATGGACGAAAAGGAATTCTATGCCTGTGAG
ATTCAATTACAAAAGTGGTATGGTGACTTACGCGGGAACCCCTTTGCTGTAGAGGCTGAGAATCGTGTAGTGGACATTTCAGGTTTCAGCATCGGATCGGATGGACATTCTATC
CAATATCGTAACCTTCTTGCATGGAAATATTTGGAGAATGGTAAGCGCGAGTTCTATCTGCTTATGAACTACGGGAAAAAAGGACGTATTCGCTTCACCGATGGTACAGACATT
AAAAAAAGTGGCAAATGGCAAGGGTTGTTATATGGGGGTGGCAAAGCAAAGGTTATCGATTTAACGTTTGATCCCGATGATGAGCAGCTGATTATCTTGCCTTTAGCTTTCGGA
ACCCGCCAGGGCCGTGAATTTATCTGGAACGACTTGCTTTCTCTTGAGACAGGACTTATCAAACTTGCGAATGGACGCGTGATCGAAAAGACTATTTACAATAAAAAGATCGGT
CGCGATGAGCCGGCCCTGTTTGTCGCCTTAACATTCGAGCGTCGCGAAGTTGTTGATCCGTCAAATATCAAGCCGGTAAACTTAATCGGGGTCGATCGTGGTGAAAATATCCCG
GCAGTAATCGCTCTTACAGACCCAGAAGGCTGCCCCTTACCCGAGTTCAAGGACTCATCCGGCGGTCCAACAGACATCCTTCGCATTGGCGAAGGATATAAAGAGAAACAGCGT
GCAATCCAGGCAGCAAAAGAAGTGGAACAACGTCGCGCCGGAGGCTACAGCCGCAAATTTGCGTCAAAATCGCGTAACTTGGCAGATGATATGGTCCGCAACAGCGCCCGTGAT
TTGTTTTACCACGCGGTGACACATGACGCTGTTTTAGTTTTTGAGAACTTGTCACGTGGGTTCGGACGTCAGGGAAAACGCACCTTCATGACTGAGCGCCAATACACAAAAATG
GAAGACTGGTTGACGGCTAAACTGGCTTACGAGGGGCTTACTAGTAAAACGTATTTGTCGAAAACTCTTGCTCAGTACACCTCTAAAACTTGTTCTAACTGCGGGTTCACGATT
ACCACAGCGGACTACGATGGCATGCTGGTACGCCTTAAAAAGACGAGCGACGGGTGGGCAACGACATTGAACAATAAAGAACTTAAGGCCGAGGGACAAATCACCTACTATAAT
CGTTATAAGCGCCAGACGGTAGAAAAAGAGTTAAGCGCCGAACTTGACCGCCTGAGCGAGGAGAGTGGGAACAACGACATTTCAAAATGGACTAAGGGGCGCCGTGACGAGGCT
TTGTTTTTGCTGAAGAAACGTTTCTCCCATCGTCCCGTACAAGAGCAGTTCGTTTGCTTAGATTGTGGCCATGAAGTTCATGCTGATGAACAGGCTGCTTTAAATATCGCGCGT
TCCTGGTTGTTTCTTAATAGTAACTCAACTGAGTTCAAATCATATAAGTCAGGGAAACAGCCGTTTGTGGGAGCATGGCAGGCTTTTTATAAACGCCGCCTGAAGGAGGTCTGG
AAGCCGAATGCT"

Cas9_dna <- DNAStringSet(x = Cas9_sequence)
CasX_dna <- DNAStringSet(x = CasX1_sequence)

DNA <- readDNAStringSet("DNA_Sequences")

Alignment <- AlignSeqs(DNA) #aligns the DNA sequence
```

```
## Determining distance matrix based on shared 9-mers:
## 
  |
  |                                                                 |   0%
  |
  |=================================================================| 100%
## 
## Time difference of 0 secs
## 
## Clustering into groups by similarity:
## 
  |
  |                                                                 |   0%
  |
  |=================================================================| 100%
## 
## Time difference of 0.01 secs
## 
## Aligning Sequences:
## 
  |
  |                                                                 |   0%
  |
  |=================================================================| 100%
## 
## Time difference of 0.49 secs
## 
## Determining distance matrix based on alignment:
## 
  |
  |                                                                 |   0%
  |
  |=================================================================| 100%
## 
## Time difference of 0 secs
## 
## Reclustering into groups by similarity:
## 
  |
  |                                                                 |   0%
  |
  |=================================================================| 100%
## 
## Time difference of 0 secs
## 
## Realigning Sequences:
## 
  |
  |                                                                 |   0%
  |
  |=================================================================| 100%
## 
## Time difference of 0 secs
```

```
Protein_Alignment <- AlignTranslation(DNA) #aligns the protein sequence
```

```
## Determining distance matrix based on shared 5-mers:
## 
  |                                                                       
  |                                                                 |   0%
  |                                                                       
  |=================================================================| 100%
## 
## Time difference of 0 secs
## 
## Clustering into groups by similarity:
## 
  |                                                                       
  |                                                                 |   0%
  |                                                                       
  |=================================================================| 100%
## 
## Time difference of 0 secs
## 
## Aligning Sequences:
## 
  |                                                                       
  |                                                                 |   0%
  |                                                                       
  |=================================================================| 100%
## 
## Time difference of 0.1 secs
## 
## Determining distance matrix based on alignment:
## 
  |                                                                       
  |                                                                 |   0%
  |                                                                       
  |=================================================================| 100%
## 
## Time difference of 0 secs
## 
## Reclustering into groups by similarity:
## 
  |                                                                       
  |                                                                 |   0%
  |                                                                       
  |=================================================================| 100%
## 
## Time difference of 0 secs
## 
## Realigning Sequences:
## 
  |                                                                       
  |                                                                 |   0%
  |                                                                       
  |=================================================================| 100%
## 
## Time difference of 0 secs
```

```
BrowseSeqs(Protein_Alignment, highlight=1) #visualize!
```

The BrowseSeqs function will open up a tab in your web browser and you can visualize the sequence alignment! You won’t be able to see this in the HTML file, sadly. Similar to the msa package, there is a consensus sequence. However, with DECIPHER you can easily see what is similar and different based on the highlighting that the program does for you! It might seem shocking that such functionally similar proteins are so different sequence-wise, but it’s not actually that surprising. CasX is notorious for being a wacky protein that doesn’t follow typical rules whereas Cas9 is somewhat ordinary.

Ribosomal RNA

We can also do a similar example with data about 50S subunit of bacterial ribosomes. This is very important to determine the identity of bacteria and understand how their translational machinery works.

```
ribosome <- system.file("extdata", "50S_ribosomal_protein_L2.fas", package="DECIPHER")
ribosome_dna <- readDNAStringSet(ribosome)
ribosome_dna
```

```
##   A DNAStringSet instance of length 317
##       width seq                                          names
##   [1]   819 ATGGCTTTAAAAAATTTTAA...TATTGTAAAAAAAAGAAAA Rickettsia prowaz...
##   [2]   822 ATGGGAATACGTAAACTCAA...CATTGAGAGAAGGAAAAAG Porphyromonas gin...
##   [3]   822 ATGGGAATACGTAAACTCAA...CATTGAGAGAAGGAAAAAG Porphyromonas gin...
##   [4]   822 ATGGGAATACGTAAACTCAA...CATTGAGAGAAGGAAAAAG Porphyromonas gin...
##   [5]   819 ATGGCTATCGTTAAATGTAA...CGTACGTCGTCGTGGTAAA Pasteurella multo...
##   ...   ... ...
## [313]   819 ATGGCAATTGTTAAATGTAA...CGTACGTCGCCGTACTAAA Pectobacterium at...
## [314]   822 ATGCCTATTCAAAAATGCAA...TCGCGATCGTCGCGTCAAG Acinetobacter sp....
## [315]   864 ATGGGCATTCGCGTTTACCG...TCGCGGTGGTCGTCAGTCT Thermosynechococc...
## [316]   831 ATGGCACTGAAGACATTCAA...CCGCCACAAGCGGAAGAAG Bradyrhizobium ja...
## [317]   840 ATGGGCATTCGCAAATATCG...GACGGCTTCCGGGCGAGGT Gloeobacter viola...
```

```r
ribosome_aa <- AlignTranslation(ribosome_dna, type="AAStringSet")
```

```
## Determining distance matrix based on shared 4-mers:
## 
  |                                                                       
  |                                                                 |   0%
  |                                                                       
  |=                                                                |   1%
  |                                                                       
  |=                                                                |   2%
  |                                                                       
  |==                                                               |   3%
  |                                                                       
  |===                                                              |   4%
  |                                                                       
  |===                                                              |   5%
  |                                                                       
  |====                                                             |   6%
  |                                                                       
  |=====                                                            |   7%
  |                                                                       
  |=====                                                            |   8%
  |                                                                       
  |======                                                           |   9%
  |                                                                       
  |======                                                           |  10%
  |                                                                       
  |=======                                                          |  11%
  |                                                                       
  |========                                                         |  12%
  |                                                                       
  |========                                                         |  13%
  |                                                                       
  |=========                                                        |  14%
  |                                                                       
  |==========                                                       |  15%
  |                                                                       
  |==========                                                       |  16%
  |                                                                       
  |===========                                                      |  17%
  |                                                                       
  |============                                                     |  18%
  |                                                                       
  |============                                                     |  19%
  |                                                                       
  |=============                                                    |  20%
  |                                                                       
  |==============                                                   |  21%
  |                                                                       
  |==============                                                   |  22%
  |                                                                       
  |===============                                                  |  23%
  |                                                                       
  |================                                                 |  24%
  |                                                                       
  |================                                                 |  25%
  |                                                                       
  |=================                                                |  26%
  |                                                                       
  |==================                                               |  27%
  |                                                                       
  |==================                                               |  28%
  |                                                                       
  |===================                                              |  29%
  |                                                                       
  |====================                                             |  30%
  |                                                                       
  |====================                                             |  31%
  |                                                                       
  |=====================                                            |  32%
  |
  |=====================                                            |  33%
  |
  |======================                                           |  34%
  |
  |=======================                                          |  35%
  |
  |=======================                                          |  36%
  |
  |========================                                         |  37%
  |
  |=========================                                        |  38%
  |
  |=========================                                        |  39%
  |
  |==========================                                       |  40%
  |
  |===========================                                      |  41%
  |
  |===========================                                      |  42%
  |
  |============================                                     |  43%
  |
  |=============================                                    |  44%
  |
  |=============================                                    |  45%
  |
  |==============================                                   |  46%
  |
  |===============================                                  |  47%
  |
  |===============================                                  |  48%
  |
  |================================                                 |  49%
  |
  |================================                                 |  50%
  |
  |=================================                                |  51%
  |
  |==================================                               |  52%
  |
  |==================================                               |  53%
  |
  |===================================                              |  54%
  |
  |====================================                             |  55%
  |
  |====================================                             |  56%
  |
  |=====================================                            |  57%
  |
  |======================================                           |  58%
  |
  |======================================                           |  59%
  |
  |=======================================                          |  60%
  |
  |========================================                         |  61%
  |
  |========================================                         |  62%
  |
  |=========================================                        |  63%
  |
  |==========================================                       |  64%
  |
  |==========================================                       |  65%
  |
  |===========================================                      |  66%
  |
  |============================================                     |  67%
  |
  |============================================                     |  68%
  |
  |=============================================                    |  69%
  |
  |==============================================                   |  70%
  |
  |==============================================                   |  71%
  |
  |===============================================                  |  72%
  |
  |===============================================                  |  73%
  |
  |================================================                 |  74%
  |
  |
  |=================================================                |  75%
  |
  |=================================================                |  76%
  |
  |==================================================               |  77%
  |
  |===================================================              |  78%
  |
  |===================================================              |  79%
  |
  |====================================================             |  80%
  |
  |=====================================================            |  81%
  |
  |=====================================================            |  82%
  |
  |======================================================           |  83%
  |
  |=======================================================          |  84%
  |
  |=======================================================          |  85%
  |
  |========================================================         |  86%
  |
  |=========================================================        |  87%
  |
  |=========================================================        |  88%
  |
  |==========================================================       |  89%
  |
  |==========================================================       |  90%
  |
  |===========================================================      |  91%
  |
  |============================================================     |  92%
  |
  |============================================================     |  93%
  |
  |=============================================================    |  94%
  |
  |==============================================================   |  95%
  |
  |==============================================================   |  96%
  |
  |===============================================================  |  97%
  |
  |================================================================ |  98%
  |
  |================================================================ |  99%
  |
  |=================================================================| 100%
##
## Time difference of 1.07 secs
##
## Clustering into groups by similarity:
##
  |
  |                                                                 |   0%
  |
  |=                                                                |   1%
  |
  |=                                                                |   2%
  |
  |==                                                               |   3%
  |
  |===                                                              |   4%
  |
  |===                                                              |   5%
  |
  |====                                                             |   6%
  |
  |=====                                                            |   7%
  |
  |=====                                                            |   8%
  |
  |======                                                           |   9%
  |
  |======                                                           |  10%
  |
  |=======                                                          |  11%
  |
  |========                                                         |  12%
  |
  |========                                                         |  13%
  |
|=========                                                        |  14%
|
|==========                                                       |  15%
|
|==========                                                       |  16%
|
|===========                                                      |  17%
|
|============                                                     |  18%
|
|============                                                     |  19%
|
|=============                                                    |  20%
|
|==============                                                   |  21%
|
|==============                                                   |  22%
|
|===============                                                  |  23%
|
|================                                                 |  24%
|
|================                                                 |  25%
|
|=================                                                |  26%
|
|==================                                               |  27%
|
|==================                                               |  28%
|
|===================                                              |  29%
|
|====================                                             |  30%
|
|====================                                             |  31%
|
|=====================                                            |  32%
|
|=====================                                            |  33%
|
|======================                                           |  34%
|
|=======================                                          |  35%
|
|=======================                                          |  36%
|
|========================                                         |  37%
|
|=========================                                        |  38%
|
|=========================                                        |  39%
|
|==========================                                       |  40%
|
|===========================                                      |  41%
|
|===========================                                      |  42%
|
|============================                                     |  43%
|
|=============================                                    |  44%
|
|=============================                                    |  45%
|
|==============================                                   |  46%
|
|===============================                                  |  47%
|
|===============================                                  |  48%
|
|================================                                 |  49%
|
|================================                                 |  50%
|
|=================================                                |  51%
|
|==================================                               |  52%
|
|==================================                               |  53%
|
|===================================                              |  54%
|
|====================================                             |  55%
|
|====================================                             |  56%
```

```
|
|=====================================                            |  57%
|
|======================================                           |  58%
|
|======================================                           |  59%
|
|=======================================                          |  60%
|
|========================================                         |  61%
|
|========================================                         |  62%
|
|=========================================                        |  63%
|
|==========================================                       |  64%
|
|==========================================                       |  65%
|
|===========================================                      |  66%
|
|============================================                     |  67%
|
|============================================                     |  68%
|
|=============================================                    |  69%
|
|==============================================                   |  70%
|
|==============================================                   |  71%
|
|===============================================                  |  72%
|
|===============================================                  |  73%
|
|================================================                 |  74%
|
|=================================================                |  75%
|
|=================================================                |  76%
|
|==================================================               |  77%
|
|===================================================              |  78%
|
|===================================================              |  79%
|
|====================================================             |  80%
|
|=====================================================            |  81%
|
|=====================================================            |  82%
|
|======================================================           |  83%
|
|=======================================================          |  84%
|
|=======================================================          |  85%
|
|========================================================         |  86%
|
|=========================================================        |  87%
|
|=========================================================        |  88%
|
|==========================================================       |  89%
|
|==========================================================       |  90%
|
|===========================================================      |  91%
|
|============================================================     |  92%
|
|============================================================     |  93%
|
|=============================================================    |  94%
|
|==============================================================   |  95%
|
|==============================================================   |  96%
|
|===============================================================  |  97%
|
|================================================================ |  98%
|
  |================================================================ |  99%
  |
  |=================================================================| 100%
##
## Time difference of 0.28 secs
##
## Aligning Sequences:
##
  |
  |                                                                 |   0%
  |
  |=                                                                |   1%
  |
  |=                                                                |   2%
  |
  |==                                                               |   3%
  |
  |===                                                              |   4%
  |
  |===                                                              |   5%
  |
  |====                                                             |   6%
  |
  |=====                                                            |   7%
  |
  |=====                                                            |   8%
  |
  |======                                                           |   9%
  |
  |======                                                           |  10%
  |
  |=======                                                          |  11%
  |
  |========                                                         |  12%
  |
  |========                                                         |  13%
  |
  |=========                                                        |  14%
  |
  |==========                                                       |  15%
  |
  |==========                                                       |  16%
  |
  |===========                                                      |  17%
  |
  |============                                                     |  18%
  |
  |============                                                     |  19%
  |
  |=============                                                    |  20%
  |
  |==============                                                   |  21%
  |
  |==============                                                   |  22%
  |
  |===============                                                  |  23%
  |
  |================                                                 |  24%
  |
  |================                                                 |  25%
  |
  |=================                                                |  26%
  |
  |==================                                               |  27%
  |
  |==================                                               |  28%
  |
  |===================                                              |  29%
  |
  |====================                                             |  30%
  |
  |====================                                             |  31%
  |
  |=====================                                            |  32%
  |
  |=====================                                            |  33%
  |
  |======================                                           |  34%
  |
  |=======================                                          |  35%
  |
  |=======================                                          |  36%
  |
  |========================                                         |  37%
  |
  |=========================                                        |  38%
```

```
|=========================                                        |  38%
|
|=========================                                        |  39%
|
|==========================                                       |  40%
|
|===========================                                      |  41%
|
|===========================                                      |  42%
|
|============================                                     |  43%
|
|=============================                                    |  44%
|
|=============================                                    |  45%
|
|==============================                                   |  46%
|
|===============================                                  |  47%
|
|===============================                                  |  48%
|
|================================                                 |  49%
|
|================================                                 |  50%
|
|=================================                                |  51%
|
|==================================                               |  52%
|
|==================================                               |  53%
|
|===================================                              |  54%
|
|====================================                             |  55%
|
|====================================                             |  56%
|
|=====================================                            |  57%
|
|======================================                           |  58%
|
|======================================                           |  59%
|
|=======================================                          |  60%
|
|========================================                         |  61%
|
|========================================                         |  62%
|
|=========================================                        |  63%
|
|==========================================                       |  64%
|
|==========================================                       |  65%
|
|===========================================                      |  66%
|
|============================================                     |  67%
|
|============================================                     |  68%
|
|=============================================                    |  69%
|
|==============================================                   |  70%
|
|==============================================                   |  71%
|
|===============================================                  |  72%
|
|===============================================                  |  73%
|
|================================================                 |  74%
|
|=================================================                |  75%
|
|=================================================                |  76%
|
|==================================================               |  77%
|
|===================================================              |  78%
|
|===================================================              |  79%
|
|====================================================             |  80%
|
  |
  |=====================================================            |  81%
  |
  |=====================================================            |  82%
  |
  |======================================================           |  83%
  |
  |=======================================================          |  84%
  |
  |=======================================================          |  85%
  |
  |========================================================         |  86%
  |
  |=========================================================        |  87%
  |
  |=========================================================        |  88%
  |
  |==========================================================       |  89%
  |
  |==========================================================       |  90%
  |
  |===========================================================      |  91%
  |
  |============================================================     |  92%
  |
  |============================================================     |  93%
  |
  |=============================================================    |  94%
  |
  |==============================================================   |  95%
  |
  |==============================================================   |  96%
  |
  |===============================================================  |  97%
  |
  |================================================================ |  98%
  |
  |================================================================ |  99%
  |
  |=================================================================| 100%
##
## Time difference of 1.55 secs
##
## Determining distance matrix based on alignment:
##
  |
  |                                                                 |   0%
  |
  |=                                                                |   1%
  |
  |=                                                                |   2%
  |
  |==                                                               |   3%
  |
  |===                                                              |   4%
  |
  |===                                                              |   5%
  |
  |====                                                             |   6%
  |
  |=====                                                            |   7%
  |
  |=====                                                            |   8%
  |
  |======                                                           |   9%
  |
  |======                                                           |  10%
  |
  |=======                                                          |  11%
  |
  |========                                                         |  12%
  |
  |========                                                         |  13%
  |
  |=========                                                        |  14%
  |
  |==========                                                       |  15%
  |
  |==========                                                       |  16%
  |
  |===========                                                      |  17%
  |
  |============                                                     |  18%
  |
  |============                                                     |  19%
  |
|=============                                                    |  20%
|
|==============                                                   |  21%
|
|==============                                                   |  22%
|
|===============                                                  |  23%
|
|================                                                 |  24%
|
|================                                                 |  25%
|
|=================                                                |  26%
|
|==================                                               |  27%
|
|==================                                               |  28%
|
|===================                                              |  29%
|
|====================                                             |  30%
|
|====================                                             |  31%
|
|=====================                                            |  32%
|
|=====================                                            |  33%
|
|======================                                           |  34%
|
|=======================                                          |  35%
|
|=======================                                          |  36%
|
|========================                                         |  37%
|
|=========================                                        |  38%
|
|=========================                                        |  39%
|
|==========================                                       |  40%
|
|===========================                                      |  41%
|
|===========================                                      |  42%
|
|============================                                     |  43%
|
|=============================                                    |  44%
|
|=============================                                    |  45%
|
|==============================                                   |  46%
|
|===============================                                  |  47%
|
|===============================                                  |  48%
|
|================================                                 |  49%
|
|================================                                 |  50%
|
|=================================                                |  51%
|
|==================================                               |  52%
|
|==================================                               |  53%
|
|===================================                              |  54%
|
|====================================                             |  55%
|
|====================================                             |  56%
|
|=====================================                            |  57%
|
|======================================                           |  58%
|
|======================================                           |  59%
|
|=======================================                          |  60%
|
|========================================                         |  61%
|
|========================================                         |  62%
```

```
  |
  |=========================================                        |  63%
  |
  |==========================================                       |  64%
  |
  |==========================================                       |  65%
  |
  |===========================================                      |  66%
  |
  |============================================                     |  67%
  |
  |============================================                     |  68%
  |
  |=============================================                    |  69%
  |
  |==============================================                   |  70%
  |
  |==============================================                   |  71%
  |
  |===============================================                  |  72%
  |
  |===============================================                  |  73%
  |
  |================================================                 |  74%
  |
  |=================================================                |  75%
  |
  |=================================================                |  76%
  |
  |==================================================               |  77%
  |
  |===================================================              |  78%
  |
  |===================================================              |  79%
  |
  |====================================================             |  80%
  |
  |=====================================================            |  81%
  |
  |=====================================================            |  82%
  |
  |======================================================           |  83%
  |
  |=======================================================          |  84%
  |
  |=======================================================          |  85%
  |
  |========================================================         |  86%
  |
  |=========================================================        |  87%
  |
  |=========================================================        |  88%
  |
  |==========================================================       |  89%
  |
  |==========================================================       |  90%
  |
  |===========================================================      |  91%
  |
  |============================================================     |  92%
  |
  |============================================================     |  93%
  |
  |=============================================================    |  94%
  |
  |==============================================================   |  95%
  |
  |==============================================================   |  96%
  |
  |===============================================================  |  97%
  |
  |================================================================ |  98%
  |
  |================================================================ |  99%
  |
  |=================================================================| 100%
##
## Time difference of 0.16 secs
##
## Reclustering into groups by similarity:
##
  |
  |                                                                 |   0%
  |
  |=                                                                |   1%
  |
|
|=                                                                |   2%
|
|==                                                               |   3%
|
|===                                                              |   4%
|
|===                                                              |   5%
|
|====                                                             |   6%
|
|=====                                                            |   7%
|
|=====                                                            |   8%
|
|======                                                           |   9%
|
|======                                                           |  10%
|
|=======                                                          |  11%
|
|========                                                         |  12%
|
|========                                                         |  13%
|
|=========                                                        |  14%
|
|==========                                                       |  15%
|
|==========                                                       |  16%
|
|===========                                                      |  17%
|
|============                                                     |  18%
|
|============                                                     |  19%
|
|=============                                                    |  20%
|
|==============                                                   |  21%
|
|==============                                                   |  22%
|
|===============                                                  |  23%
|
|================                                                 |  24%
|
|================                                                 |  25%
|
|=================                                                |  26%
|
|==================                                               |  27%
|
|==================                                               |  28%
|
|===================                                              |  29%
|
|====================                                             |  30%
|
|====================                                             |  31%
|
|=====================                                            |  32%
|
|=====================                                            |  33%
|
|======================                                           |  34%
|
|=======================                                          |  35%
|
|=======================                                          |  36%
|
|========================                                         |  37%
|
|=========================                                        |  38%
|
|=========================                                        |  39%
|
|==========================                                       |  40%
|
|===========================                                      |  41%
|
|===========================                                      |  42%
|
|============================                                     |  43%
|
|=============================                                    |  44%
```

```
|=============================                                    |  44%
|
|=============================                                    |  45%
|
|==============================                                   |  46%
|
|===============================                                  |  47%
|
|===============================                                  |  48%
|
|================================                                 |  49%
|
|================================                                 |  50%
|
|=================================                                |  51%
|
|==================================                               |  52%
|
|==================================                               |  53%
|
|===================================                              |  54%
|
|====================================                             |  55%
|
|====================================                             |  56%
|
|=====================================                            |  57%
|
|======================================                           |  58%
|
|======================================                           |  59%
|
|=======================================                          |  60%
|
|========================================                         |  61%
|
|========================================                         |  62%
|
|=========================================                        |  63%
|
|==========================================                       |  64%
|
|==========================================                       |  65%
|
|===========================================                      |  66%
|
|============================================                     |  67%
|
|============================================                     |  68%
|
|=============================================                    |  69%
|
|==============================================                   |  70%
|
|==============================================                   |  71%
|
|===============================================                  |  72%
|
|===============================================                  |  73%
|
|================================================                 |  74%
|
|=================================================                |  75%
|
|=================================================                |  76%
|
|==================================================               |  77%
|
|===================================================              |  78%
|
|===================================================              |  79%
|
|====================================================             |  80%
|
|=====================================================            |  81%
|
|=====================================================            |  82%
|
|======================================================           |  83%
|
|=======================================================          |  84%
|
|=======================================================          |  85%
|
|========================================================         |  86%
|
  |
  |=========================================================        |  87%
  |
  |=========================================================        |  88%
  |
  |==========================================================       |  89%
  |
  |==========================================================       |  90%
  |
  |===========================================================      |  91%
  |
  |============================================================     |  92%
  |
  |============================================================     |  93%
  |
  |=============================================================    |  94%
  |
  |==============================================================   |  95%
  |
  |==============================================================   |  96%
  |
  |===============================================================  |  97%
  |
  |================================================================ |  98%
  |
  |================================================================ |  99%
  |
  |=================================================================| 100%
##
## Time difference of 0.25 secs
##
## Realigning Sequences:
##
  |
  |                                                                 |   0%
  |
  |=                                                                |   1%
  |
  |=                                                                |   2%
  |
  |==                                                               |   3%
  |
  |===                                                              |   4%
  |
  |===                                                              |   5%
  |
  |====                                                             |   6%
  |
  |=====                                                            |   7%
  |
  |=====                                                            |   8%
  |
  |======                                                           |   9%
  |
  |======                                                           |  10%
  |
  |=======                                                          |  11%
  |
  |========                                                         |  12%
  |
  |========                                                         |  13%
  |
  |=========                                                        |  14%
  |
  |==========                                                       |  15%
  |
  |==========                                                       |  16%
  |
  |===========                                                      |  17%
  |
  |============                                                     |  18%
  |
  |============                                                     |  19%
  |
  |=============                                                    |  20%
  |
  |==============                                                   |  21%
  |
  |==============                                                   |  22%
  |
  |===============                                                  |  23%
  |
  |================                                                 |  24%
  |
  |================                                                 |  25%
  |
  |=================                                                |  26%
  |
  |==================                                               |  27%
  |
  |==================                                               |  28%
  |
  |===================                                              |  29%
  |
  |====================                                             |  30%
  |
  |====================                                             |  31%
  |
  |=====================                                            |  32%
  |
  |=====================                                            |  33%
  |
  |======================                                           |  34%
  |
  |=======================                                          |  35%
  |
  |=======================                                          |  36%
  |
  |========================                                         |  37%
  |
  |=========================                                        |  38%
  |
  |=========================                                        |  39%
  |
  |==========================                                       |  40%
  |
  |===========================                                      |  41%
  |
  |===========================                                      |  42%
  |
  |============================                                     |  43%
  |
  |=============================                                    |  44%
  |
  |=============================                                    |  45%
  |
  |==============================                                   |  46%
  |
  |===============================                                  |  47%
  |
  |===============================                                  |  48%
  |
  |================================                                 |  49%
  |
  |================================                                 |  50%
  |
  |=================================                                |  51%
  |
  |==================================                               |  52%
  |
  |==================================                               |  53%
  |
  |===================================                              |  54%
  |
  |====================================                             |  55%
  |
  |====================================                             |  56%
  |
  |=====================================                            |  57%
  |
  |======================================                           |  58%
  |
  |=================================================================| 100%
##
## Time difference of 0.8 secs
```

```
BrowseSeqs(ribosome_aa, highlight=1)
```

That alignment is much more impressive because it’s a much larger dataset! If you run this code yourself and see the browser visualization, there are some very conserved motifs throughout the genomes. These would be great places to use stringr to search for bacteria who have a slightly different sequence.

Influenza

To see how similar sequences are to each other, you can also use the “FindSynteny” function to get a table of similarity. Here we can look at the different glycoprotein attachments (Hemagglutinin = “H” and Neuraminidase = “N”) on the surface of viral particles that cause the flu. It’s interesting to see how they compare to each other because diffferent combinations of these glycoprotein decorations cause slightly different forms of the flu.

```
flu_sequences <- system.file("extdata", "Influenza.sqlite", package="DECIPHER")
synteny <- FindSynteny(flu_sequences, minScore=50, verbose=FALSE)
synteny
```

```
##           H9N2     H5N1     H2N2     H7N9     H1N1
## H9N2   8 seqs 53% hits 34% hits 48% hits 34% hits
## H5N1 7 blocks   8 seqs 30% hits 47% hits 44% hits
## H2N2 7 blocks 8 blocks   8 seqs 29% hits 35% hits
## H7N9 7 blocks 6 blocks 8 blocks   8 seqs 32% hits
## H1N1 6 blocks 8 blocks 6 blocks 6 blocks   8 seqs
```

Discusion & Conclusions

In conclusion, msa and DECIPHER are both great tools to combine with stringr to manipulate specific kinds of strings and pull as much meaning out as possible. Here I have talked about how to:

- align protein sequences using msa
- read the output of the alignment
- visualize parts vs. entire sequence alignment using msa
- search for sequences within an alignment using stringr
- align character strings of DNA sequence using DECIPHER
- visualize those alignments in a browser using DECIPHER
- examine different kinds of DNA using DECIPHER
- create a similarity table using DECIPHER

Take home message: Sequence alignments have huge implications for modern biology and aren't trivial. Packages like msa and DECIPHER are important and useful for aligning sequences of DNA and proteins.

References

[1] http://sitn.hms.harvard.edu/flash/2014/crispr-a-game-changing-genetic-engineering-technique/

[2] https://en.wikipedia.org/wiki/Sequence_alignment

[3] http://www.sciencemag.org/news/2017/07/ding-ding-ding-crispr-patent-fight-enters-next-round

[4] https://patents.google.com/patent/US8945839B2/en?q=cas9

[5] https://www.ncbi.nlm.nih.gov/pubmed/24476820

[6] https://www.ncbi.nlm.nih.gov/pubmed/28005056

[7] https://bioconductor.org/packages/devel/bioc/vignettes/DECIPHER/inst/doc/ArtOfAlignmentInR.pdf