Upsize your clustering with Clusterize

Erik S. Wright

October 24, 2023

Contents

1 Introduction to supersize clustering

You may have found yourself in a familiar predicament for many bioinformaticians: you have a lot of sequences and you need to downs*ize* before you can get going. You may also theor*ize* that this must be an easy problem to solve —given sequences, output clusters. But what can you util*ize* to solve this problem? This vignette will familiar*ize* you with the `Clusterize` function in the DECIPHER package. Clusterize will revolution*ize* all your clustering needs!

Why `Clusterize`?:

- Scalability - `Clusterize` will linear*ize* the search space so that many sequences can be clustered in a reasonable amount of time.
- Simplicity - Although you can individual*ize* `Clusterize`, the defaults are straightforward and should meet most of your needs.
- Accuracy - `Clusterize` will maxim*ize* your ability to extract biologically meaningful results from your sequences.

This vignette will summar*ize* the use of `Clusterize` to cluster DNA, RNA, or protein sequences.

2 Getting started with Clusterize

To get started we need to load the DECIPHER package, which automatically mobil*ize* a few other required packages.

```
> library(DECIPHER)
```

There's no need to memor*ize* the inputs to `Clusterize`, because its help page can be accessed through:

```
> ? Clusterize
```

3 Optimize your inputs to Clusterize

`Clusterize` requires that you first digit*ize* your sequences by loading them into memory. For the purpose of this vignette, we will capital*ize* on the fact that DECIPHER already includes some built-in sets of sequences.

```
> # specify the path to your file of sequences:
> fas <- "<<path to training FASTA file>>"
> # OR use the example DNA sequences:
> fas <- system.file("extdata",
          "50S_ribosomal_protein_L2.fas",
          package="DECIPHER")
> # read the sequences into memory
> dna <- readDNAStringSet(fas)
> dna
DNAStringSet object of length 317:
      width seq                                               names
  [1]   819 ATGGCTTTAAAAAATTTTAATC...ATTTATTGTAAAAAAAAGAAAA Rickettsia prowaz...
  [2]   822 ATGGGAATACGTAAACTCAAGC...CATCATTGAGAGAAGGAAAAAG Porphyromonas gin...
  [3]   822 ATGGGAATACGTAAACTCAAGC...CATCATTGAGAGAAGGAAAAAG Porphyromonas gin...
  [4]   822 ATGGGAATACGTAAACTCAAGC...CATCATTGAGAGAAGGAAAAAG Porphyromonas gin...
  [5]   819 ATGGCTATCGTTAAATGTAAGC...CATCGTACGTCGTCGTGGTAAA Pasteurella multo...
  ...   ... ...
[313]   819 ATGGCAATTGTTAAATGTAAAC...TATCGTACGTCGCCGTACTAAA Pectobacterium at...
[314]   822 ATGCCTATTCAAAAATGCAAAC...TATTCGCGATCGTCGCGTCAAG Acinetobacter sp....
[315]   864 ATGGGCATTCGCGTTTACCGAC...GGGTCGCGGTGGTCGTCAGTCT Thermosynechococc...
[316]   831 ATGGCACTGAAGACATTCAATC...AAGCCGCCACAAGCGGAAGAAG Bradyrhizobium ja...
[317]   840 ATGGGCATTCGCAAATATCGAC...CAAGACGGCTTCCGGGCGAGGT Gloeobacter viola...
```

The `Clusterize` algorithm will general*ize* to nucleotide or protein sequences, so we must choose which we are going to use. Here, we hypothes*ize* that weaker similarities can be detected between proteins and, therefore, decide to use the translated coding (amino acid) sequences. If you wish to cluster at high similarity, you could also strateg*ize* that nucleotide sequences would be better because there would be more nucleotide than amino acid differences.

```
> aa <- translate(dna)
> aa
AAStringSet object of length 317:
      width seq                                               names
  [1]   273 MALKNFNPITPSLRELVQVDKT...STKGKKTRKNKRTSKFIVKKRK Rickettsia prowaz...
  [2]   274 MGIRKLKPTTPGQRHKVIGAFD...KGLKTRAPKKHSSKYIIERRKK Porphyromonas gin...
  [3]   274 MGIRKLKPTTPGQRHKVIGAFD...KGLKTRAPKKHSSKYIIERRKK Porphyromonas gin...
  [4]   274 MGIRKLKPTTPGQRHKVIGAFD...KGLKTRAPKKHSSKYIIERRKK Porphyromonas gin...
  [5]   273 MAIVKCKPTSAGRRHVVKIVNP...TKGKKTRHNKRTDKFIVRRRGK Pasteurella multo...
  ...   ... ...
[313]   273 MAIVKCKPTSPGRRHVVKVVNP...TKGKKTRSNKRTDKFIVRRRTK Pectobacterium at...
[314]   274 MPIQKCKPTSPGRRFVEKVVHS...KGYKTRTNKRTTKMIIRDRRVK Acinetobacter sp....
[315]   288 MGIRVYRPYTPGVRQKTVSDFA...SDALIVRRRKKSSKRGRGGRQS Thermosynechococc...
[316]   277 MALKTFNPTTPGQRQLVMVDRS...KKTRSNKSTNKFILLSRHKRKK Bradyrhizobium ja...
[317]   280 MGIRKYRPMTPGTRQRSGADFA...RKRRKPSSKFIIRRRKTASGRG Gloeobacter viola...
```

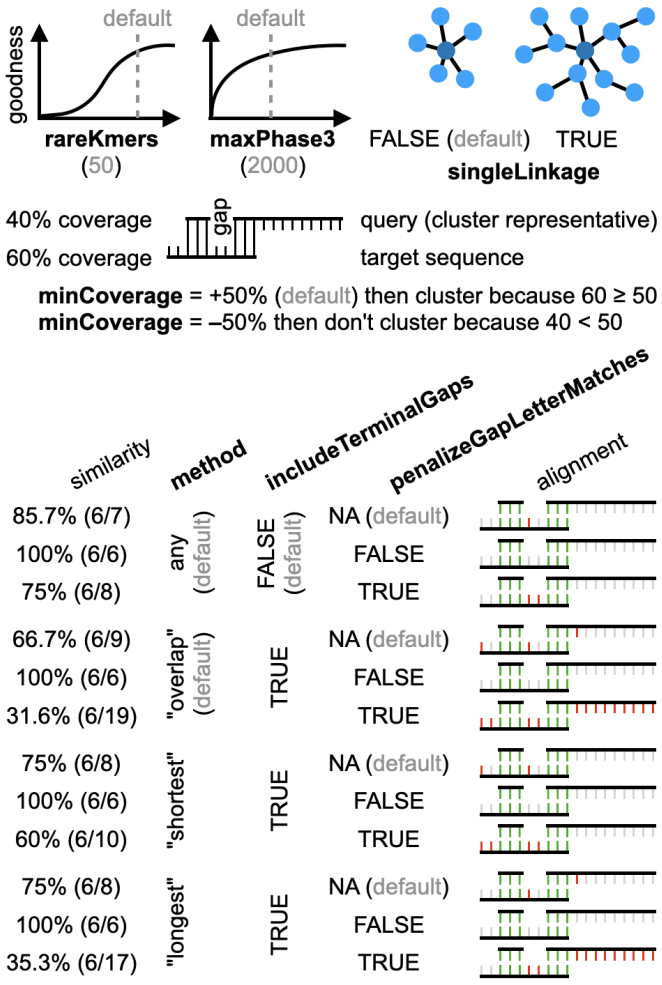

Figure 1: The most important parameters (in **bold**) to custom*ize* your use of `Clusterize`.

```
> seqs <- aa # could also cluster the nucleotides
> length(seqs)
[1] 317
```

Now you can choose how to parameter*ize* the function, with the main arguments being *myXStringSet* and *cutoff*. In this case, we will initial*ize* *cutoff* at `seq(0.5, 0, -0.1)` to cluster sequences from 50% to 100% similarity by 10%'s. It is important to recogn*ize* that *cutoff*s can be provided in *ascending* or *descending* order and, when *descending*, groups at each *cutoff* will be nested within the previous *cutoff*'s groups.

We must also choose whether to custom*ize* the calculation of distance. The defaults will penal*ize* gaps as single events, such that each consecutive set of gaps (i.e., insertion or deletion) is considered equivalent to one mismatch. If you want to standard*ize* the definition of distance to be the same as most other clustering programs then set: *penalizeGapLetterMatches* to `TRUE` (i.e., every gap position is a mismatch), *method* to `"shortest"`, *minCoverage* to `0`, and *includeTerminalGaps* to `TRUE`. It is possible to rational*ize* many different measures of distance – see the `DistanceMatrix` function for more information about alternative distance parameterizations.

We can further personal*ize* the inputs as desired. The main function argument to emphas*ize* is *processors*, which controls whether the function is parallelized on multiple computer threads (if DECIPHER) was built with OpenMP enabled). Setting *processors* to a value greater than `1` will speed up clustering considerably, especially for large s*ize* clustering problems. Once we are ready, it's time to run `Clusterize` and wait for the output to material*ize*!

```
> clusters <- Clusterize(seqs, cutoff=seq(0.5, 0, -0.1), processors=1)
Partitioning sequences by 4-mer similarity:
================================================================================
Time difference of 0.05 secs

Sorting by relatedness within 5 groups:

iteration 47 of up to 47 (100.0% stability)

Time difference of 1.33 secs

Clustering sequences by 4-mer similarity:
================================================================================

Time difference of 0.17 secs

Clusters via relatedness sorting: 100% (0% exclusively)
Clusters via rare 4-mers: 100% (0% exclusively)
Estimated clustering effectiveness: 100%
> class(clusters)
[1] "data.frame"
> colnames(clusters)
[1] "cluster_0_5" "cluster_0_4" "cluster_0_3" "cluster_0_2" "cluster_0_1"
[6] "cluster_0"
> str(clusters)
'data.frame':        317 obs. of  6 variables:
 $ cluster_0_5: int  4 1 1 1 4 4 4 3 3 3 ...
 $ cluster_0_4: int  1 25 25 25 2 2 2 9 9 9 ...
 $ cluster_0_3: int  49 1 1 1 43 43 44 31 31 31 ...
 $ cluster_0_2: int  1 71 71 71 12 12 10 29 29 29 ...
 $ cluster_0_1: int  90 1 1 1 72 72 74 51 51 51 ...
 $ cluster_0  : int  2 102 102 102 24 24 22 46 46 46 ...
> apply(clusters, 2, max) # number of clusters per cutoff
cluster_0_5 cluster_0_4 cluster_0_3 cluster_0_2 cluster_0_1   cluster_0
          4          25          49          71          90         102
> apply(clusters, 2, function(x) which.max(table(x))) # max sizes
cluster_0_5 cluster_0_4 cluster_0_3 cluster_0_2 cluster_0_1   cluster_0
          3          10          28          37          42          55
```

Notice that `Clusterize` will character*ize* the clustering based on how many clustered pairs came from relatedness sorting versus rare k-mers, and `Clusterize` will predict the effectiveness of clustering. Depending on the input sequences, the percentage of clusters originating from relatedness sorting will equal*ize* with the number originating from rare k-mers, but more commonly clusters will originate from one source or the other. The clustering effectiveness formal*ize*s the concept of "inexact" clustering by approximating the fraction of possible sequence pairs

that were correctly clustered together. You can incentiv*ize* a higher clustering effectiveness by increasing *maxPhase3* at the expense of (proportionally) longer run times.

We can now real*ize* our objective of decreasing the number of sequences. Here, we will prioriti*ze* keeping only the longest diverse sequences.

```
> o <- order(clusters[[2]], width(seqs), decreasing=TRUE) # 40% cutoff
> o <- o[!duplicated(clusters[[2]])]
> aa[o]
AAStringSet object of length 25:
     width seq                                               names
 [1]   274 MGIRKLKPTTPGQRHKVIGAFDK...KGLKTRAPKKHSSKYIIERRKK Porphyromonas gin...
 [2]   274 MGIRKLKPTTPGQRHKVIGAFDK...KGLKTRAPKKHSSKYIIERRKK Porphyromonas gin...
 [3]   274 MAVRKLKPTTPGQRHKIIGTFEE...KGLKTRAPKKQSSKYIIERRKK Bacteroides theta...
 [4]   276 MALVKTKPTSPGRRSMVKVVNPD...KGYRTRSNKRTTSMIVQRRHKR Ralstonia solanac...
 [5]   278 MGIRKYKPTTPGRRGSSVADFVE...TRSPKKASNKYIVRRRKTNKKR Streptomyces coel...
 ...   ... ...
[21]   277 MALKHFNPITPGQRQLVIVDRSE...KKTRSNKATDKFIMRSRHQRKK Brucella melitens...
[22]   277 MALKHFNPITPGQRQLVIVDRSE...KKTRSNKATDKFIMRSRHQRKK Brucella sp. NF 2653
[23]   274 MAIVKCKPTSAGRRHVVKVVNAD...TKGYKTRSNKRTDKYIVRRRNK Vibrio cholerae PS15
[24]   274 MAIVKCKPTSAGRRHVVKVVNAD...TKGYKTRSNKRTDKYIVRRRNK Vibrio cholerae H...
[25]   274 MAIVKCKPTSPGRRFVVKVVNQE...PTKGAKTRGNKRTDKMIVRRRK Pseudomonas syrin...
> dna[o]
DNAStringSet object of length 25:
     width seq                                               names
 [1]   822 ATGGGAATACGTAAACTCAAGCC...CATCATTGAGAGAAGGAAAAAG Porphyromonas gin...
 [2]   822 ATGGGAATACGTAAACTCAAGCC...CATCATTGAGAGAAGGAAAAAG Porphyromonas gin...
 [3]   822 ATGGCAGTACGTAAATTAAAGCC...CATTATTGAGAGAAGAAAAAAG Bacteroides theta...
 [4]   828 ATGGCACTCGTCAAGACCAAGCC...CGTGCAACGCCGTCACAAGCGT Ralstonia solanac...
 [5]   834 ATGGGAATCCGCAAGTACAAGCC...CCGCAAGACGAACAAGAAGCGC Streptomyces coel...
 ...   ... ...
[21]   831 ATGGCACTCAAGCATTTTAATCC...TTCGCGCCATCAGCGCAAGAAG Brucella melitens...
[22]   831 ATGGCACTCAAGCATTTTAATCC...TTCGCGCCATCAGCGCAAGAAG Brucella sp. NF 2653
[23]   822 ATGGCTATTGTTAAATGTAAGCC...CATCGTACGTCGTCGTAATAAG Vibrio cholerae PS15
[24]   822 ATGGCTATTGTTAAATGTAAGCC...CATCGTACGTCGTCGTAATAAG Vibrio cholerae H...
[25]   822 ATGGCAATCGTTAAATGCAAACC...AATGATCGTCCGTCGTCGCAAG Pseudomonas syrin...
```

4 Visualize the output of Clusterize

We can scrutin*ize* the clusters by selecting them and looking at their multiple sequence alignment:

```
> t <- table(clusters[[1]]) # select the clusters at a cutoff
> t <- sort(t, decreasing=TRUE)
> head(t)
  3   4   1   2
138 111  55  13
> w <- which(clusters[[1]] == names(t[1]))
> AlignSeqs(seqs[w], verbose=FALSE)
AAStringSet object of length 138:
      width seq                                               names
```

```
  [1]   333 VGIKKYKPTT-NGRRNMTASDF...NKKARSNKLIVRGRRPGKH--- Lactobacillus pla...
  [2]   333 VGIKKYKPTT-NGRRNMTASDF...NKKARSNKLIVRGRRPGKH--- Lactobacillus pla...
  [3]   333 VGIKKYKPTT-NGRRNMTASDF...NKKARSNKLIVRGRRPGKH--- Lactobacillus pla...
  [4]   333 VGIKKYKPTT-NGRRNMTASDF...NKKARSNKLIVRGRRPGKH--- Lactobacillus pla...
  [5]   333 VGIKKYKPTT-NGRRNMTASDF...NKKARSNKLIVRGRRPGKH--- Lactobacillus pla...
  ...   ... ...
[134]   333 MAIKKIISRSNSGIHNATVIDF...NMKKHSTNLIIRNRKGEQY--- Mycoplasma genita...
[135]   333 MAIKKIISRSNSGIHNATVIDF...NMKKHSTNLIIRNRKGEQY--- Mycoplasma genita...
[136]   333 MAIRKLNPTT-NGTRNMSILDY...DNKKSSTKLIIRRRKES---K* Mycoplasma pulmonis
[137]   333 MPVKKIVNRSNSGIHHKISIDY...NNKKSSTQLIIRRRNSK----* Mycoplasma gallis...
[138]   333 MAIKKYKSTT-NGRRNMTTIDY...NTKKTSEKLIVRKRSNK---K* Mycoplasma mycoid...
```

It's possible to util*ize* the `heatmap` function to view the clustering results.

As can be seen in Figure 2, `Clusterize` will organ*ize* its clusters such that each new cluster is within the previous cluster when *cutoff* is provided in descending order. We can also see that sequences from the same species tend to cluster together, which is an alternative way to systemat*ize* sequences without clustering.

```
> aligned_seqs <- AlignSeqs(seqs, verbose=FALSE)
> d <- DistanceMatrix(aligned_seqs, verbose=FALSE)
> tree <- TreeLine(myDistMatrix=d, method="UPGMA", verbose=FALSE)
> heatmap(as.matrix(clusters), scale="column", Colv=NA, Rowv=tree)
```

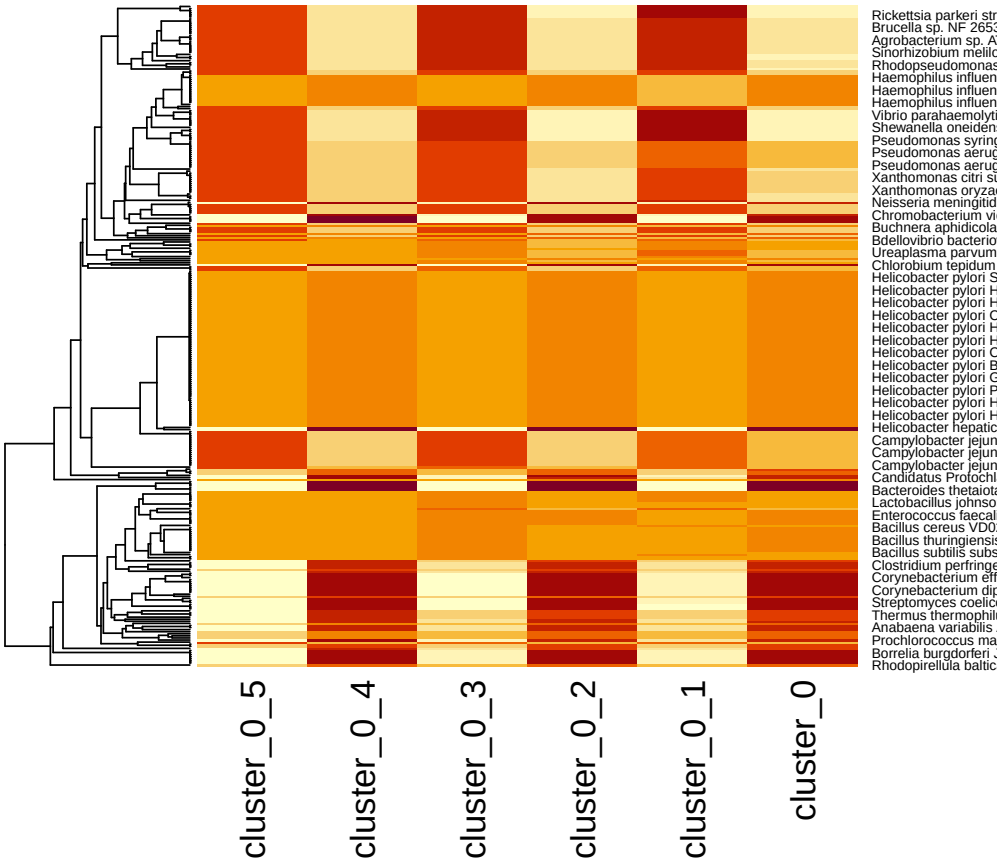

Figure 2: Visualization of the clustering.

5 Specialize clustering for your goals

The most common use of clustering is to categor*ize* sequences into groups sharing similarity above a threshold and pick one representative sequence per group. These settings empitom*ize* this typical user scenario:

```
> c1 <- Clusterize(dna, cutoff=0.2, invertCenters=TRUE, processors=1)
Partitioning sequences by 5-mer similarity:
================================================================================
Time difference of 0.1 secs

Sorting by relatedness within 1 group:

iteration 1 of up to 158 (100.0% stability)

Time difference of 0.29 secs

Clustering sequences by 9-mer similarity:
================================================================================

Time difference of 1.3 secs

Clusters via relatedness sorting: 100% (0% exclusively)
Clusters via rare 5-mers: 100% (0% exclusively)
Estimated clustering effectiveness: 100%
> w <- which(c1 < 0 & !duplicated(c1))
> dna[w] # select cluster representatives (negative cluster numbers)
DNAStringSet object of length 77:
     width seq                                               names
 [1]   819 ATGGCTTTAAAAAATTTTAATCC...ATTTATTGTAAAAAAAAGAAAA Rickettsia prowaz...
 [2]   822 ATGGGAATACGTAAACTCAAGCC...CATCATTGAGAGAAGGAAAAAG Porphyromonas gin...
 [3]   837 GTGGGTATTAAGAAGTATAAACC...TGGTCGCCGTCCAGGCAAACAC Lactobacillus pla...
 [4]   825 ATGCCATTGATGAAGTTCAAACC...CATCGTCCGCGATCGTAGGGGC Xanthomonas vesic...
 [5]   828 ATGGGTATTCGTAATTATCGGCC...GATTGTCCGCCGTCGCACCAAA Synechocystis sp....
 ...   ... ...
[73]   831 ATGGCACTTAAGCAGTTTAATCC...TACGCGTCATCAGCGCAAGAAA Bartonella hensel...
[74]   843 ATGTTTAAGAAATATCGACCTGT...CGTGAAACGTCGAAGGAAGAAG Candidatus Protoc...
[75]   822 ATGCCTATTCAAAAATGCAAACC...TATTCGCGATCGTCGCGTCAAG Acinetobacter sp....
[76]   864 ATGGGCATTCGCGTTTACCGACC...GGGTCGCGGTGGTCGTCAGTCT Thermosynechococc...
[77]   840 ATGGGCATTCGCAAATATCGACC...CAAGACGGCTTCCGGGCGAGGT Gloeobacter viola...
```

By default, `Clusterize` will cluster sequences with linkage to the representative sequence in each group, but it is also possible to tell `Clusterize` to minim*ize* the number of clusters by establishing linkage to any sequence in the cluster (i.e., single-linkage). This is often how we conceptual*ize* natural groupings and, therefore, may better match alternative classification systems such as taxonomy:

```
> c2 <- Clusterize(dna, cutoff=0.2, singleLinkage=TRUE, processors=1)
Partitioning sequences by 5-mer similarity:
================================================================================
Time difference of 0.07 secs

Sorting by relatedness within 1 group:
```

```
iteration 112 of up to 158 (100.0% stability)

Time difference of 19.98 secs

Clustering sequences by 9-mer similarity:
================================================================================

Time difference of 0.42 secs

Clusters via relatedness sorting: 100% (0% exclusively)
Clusters via rare 5-mers: 100% (0% exclusively)
Estimated clustering effectiveness: 100%
> max(abs(c1)) # center-linkage
[1] 77
> max(c2) # single-linkage (fewer clusters, but broader clusters)
[1] 77
```

It is possible to synthes*ize* a plot showing a cross tabulation of taxonomy and cluster number. We may ideal*ize* the clustering as matching taxonomic labels (3), but this is not exactly the case.

```
> genus <- sapply(strsplit(names(dna), " "), `[`, 1)
> t <- table(genus, c2[[1]])
> heatmap(sqrt(t), scale="none", Rowv=NA, col=hcl.colors(100))
```

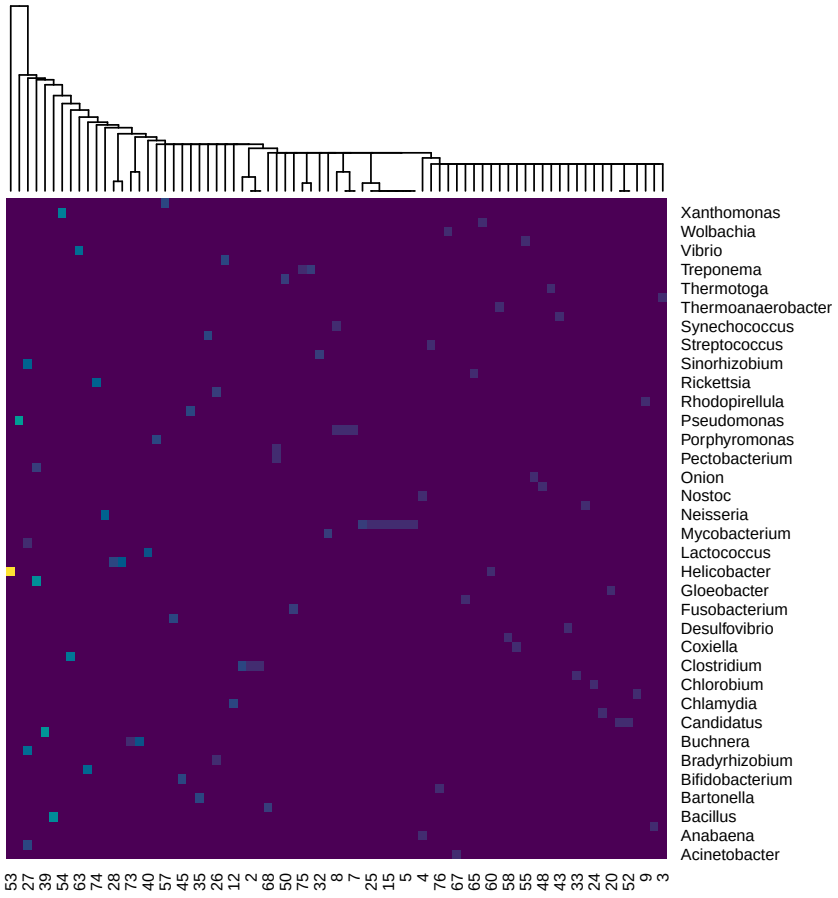

Figure 3: Another visualization of the clustering.

6 Resize to fit within less memory

What should you do if you have more sequences than you can cluster on your mids*ize* computer? If there are far fewer clusters than sequences (e.g., *cutoff* is high) then it is likely possible to res*ize* the clustering problem. This is accomplished by processing the sequences in batches that miniatur*ize* the memory footprint and are at least as large as the final number of clusters. The number of sequences processed per batch is critical to atom*ize* the problem appropriately while limiting redundant computations. Although not ideal from a speed perspective, the results will not jeopard*ize* accuracy relative to as if there was sufficient memory available to process all sequences in one batch.

```
> batchSize <- 2e2 # normally a large number (e.g., 1e6 or 1e7)
> o <- order(width(seqs), decreasing=TRUE) # process largest to smallest
> c3 <- integer(length(seqs)) # cluster numbers
> repeat {
         m <- which(c3 < 0) # existing cluster representatives
         m <- m[!duplicated(c3[m])] # remove redundant sequences
         if (length(m) >= batchSize)
                 stop("batchSize is too small")
         w <- head(c(m, o[c3[o] == 0L]), batchSize)
         if (!any(c3[w] == 0L)) {
                 if (any(c3[-w] == 0L))
                         stop("batchSize is too small")
                 break # done
         }
         m <- m[match(abs(c3[-w]), abs(c3[m]))]
         c3[w] <- Clusterize(seqs[w], cutoff=0.05, invertCenters=TRUE)[[1]]
         c3[-w] <- ifelse(is.na(c3[m]), 0L, abs(c3[m]))
 }
Partitioning sequences by 3-mer similarity:
================================================================================
Time difference of 0.04 secs

Sorting by relatedness within 6 groups:

iteration 1 of up to 25 (100.0% stability)

Time difference of 0.02 secs

Clustering sequences by 4-mer similarity:
================================================================================

Time difference of 0.16 secs

Clusters via relatedness sorting: 100% (0% exclusively)
Clusters via rare 3-mers: 100% (0% exclusively)
Estimated clustering effectiveness: 100%

Partitioning sequences by 3-mer similarity:
================================================================================
Time difference of 0.03 secs

Sorting by relatedness within 4 groups:
```

```
iteration 1 of up to 46 (100.0% stability)

Time difference of 0.03 secs

Clustering sequences by 4-mer similarity:
================================================================================

Time difference of 0.28 secs

Clusters via relatedness sorting: 100% (0% exclusively)
Clusters via rare 3-mers: 100% (0% exclusively)
Estimated clustering effectiveness: 100%
> table(abs(c3)) # cluster sizes
 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
 1  1  1  1  1  2  1  1  1  1  2  1  3  1  1  1  1  7  3  1  1  1  1  3  1  1
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
 1  5  6  3  2  1  3  3  6  1  1  2  7  2  1  1  1  8  3  1 17  3  2  2  2  3
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
75  1  1  1  3  4  1 12  1  1  1 11  3  1  1  1  1  1  1  5  6  3  3  2  1  1
79 80 81 82 83 84 85 86 87 88 89 90 91
 1  1  1 17 13  6  3  1  1  1  1  1  1
```

7 Finalize your use of Clusterize

Notably, `Clusterize` is a stochastic algorithm, meaning it will random*ize* which sequences are selected during pre-sorting. Even though the clusters will typically stabil*ize* with enough iterations, you can set the random number seed (before every run) to guarantee reproducibility of the clusters:

```
> set.seed(123) # initialize the random number generator
> clusters <- Clusterize(seqs, cutoff=0.1, processors=1)
Partitioning sequences by 4-mer similarity:
================================================================================
Time difference of 0.02 secs

Sorting by relatedness within 5 groups:

iteration 1 of up to 47 (100.0% stability)

Time difference of 0.02 secs

Clustering sequences by 4-mer similarity:
================================================================================

Time difference of 0.25 secs

Clusters via relatedness sorting: 100% (0% exclusively)
Clusters via rare 4-mers: 100% (0% exclusively)
Estimated clustering effectiveness: 100%
```

```
> set.seed(NULL) # reset the seed
```

Now you know how to util*ize* `Clusterize` to cluster sequences. To public*ize* your results for others to reproduce, make sure to provide your random number seed and version number:

- R version 4.3.1 (2023-06-16 ucrt), `x86_64-w64-mingw32`
- Running under: `Windows Server 2022 x64 (build 20348)`
- Matrix products: default
- Base packages: base, datasets, grDevices, graphics, methods, parallel, stats, stats4, utils
- Other packages: BiocGenerics 0.48.0, Biostrings 2.70.0, DECIPHER 2.30.0, GenomeInfoDb 1.38.0, IRanges 2.36.0, RSQLite 2.3.1, S4Vectors 0.40.0, XVector 0.42.0
- Loaded via a namespace (and not attached): DBI 1.1.3, GenomeInfoDbData 1.2.11, RCurl 1.98-1.12, bit 4.0.5, bit64 4.0.5, bitops 1.0-7, blob 1.2.4, cachem 1.0.8, cli 3.6.1, compiler 4.3.1, crayon 1.5.2, fastmap 1.1.1, memoise 2.0.1, pkgconfig 2.0.3, rlang 1.1.1, tools 4.3.1, vctrs 0.6.4, zlibbioc 1.48.0