Nama : Muhamad Hilmy Haidar
NIM : G64170030
LKP : 9

Dataset 1:

1. Lakukan ekstraksi fitur menggunakan Amino Acid Composition (AAC) [X]
2. Lakukan analisis clustering pada file “dataset 1” dengan menggunakan agglomerative hierarchical clustering, ke dalam 3 cluster
3. Lakukan clustering menggunakan complete dan centroid linkage, metric yang digunakan adalah Euclidean distance
4. Gambarkan dendogramnya dan analisis hasilnya

Jawaban 1:

Pertama, lakukan ekstraksi fitur ke tiap-tiap file fasta dari dataset :

```
# Fungsi untuk ekstrak fitur
extract_feature <- function(fasta_path) {
  fasta_file <- readFASTA(fasta_path)[[1]]
  extracted <- extractAAC(fasta_file)
  return(extracted)
}
```

Hasil yang diperoleh dari ekstraksi fitur adalah sebagai berikut :

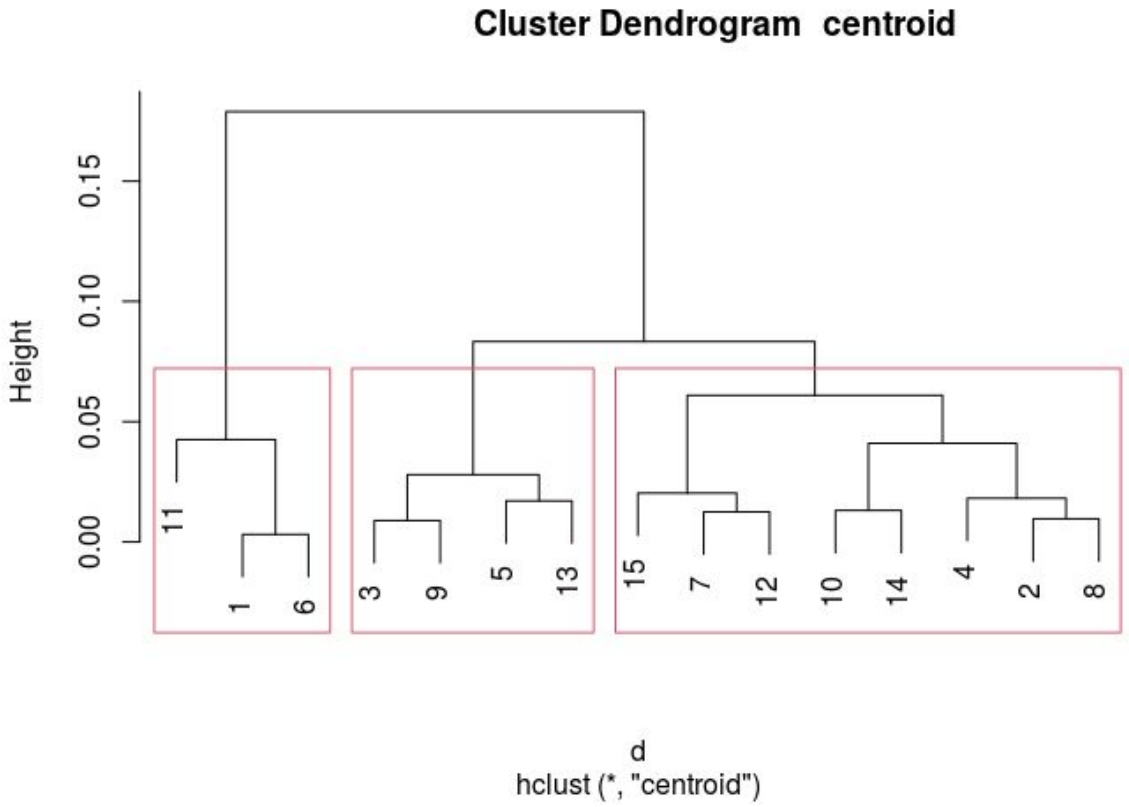
```
> extract_feature("dataset01/A0PJY2.fasta")
          A           R           N           D           C           E
Q           G           H           I           L           K           M
F
0.08631579 0.04000000 0.04421053 0.02315789 0.04210526 0.03157895
0.04210526 0.06315789 0.06105263 0.02105263 0.08631579 0.08210526
0.02315789 0.04842105
          P           S           T           W           Y           V
0.10947368 0.06947368 0.05473684 0.00000000 0.02315789 0.04842105
```

Pasangan ekspresi gen dengan nilai kemunculannya dari data pertama akan digabung dengan data lain sampai membentuk matriks yang barisnya sebanyak jumlah data dan kolom nya sebanyak jumlah ekspresi gen yang berbeda seperti gambar dibawah :

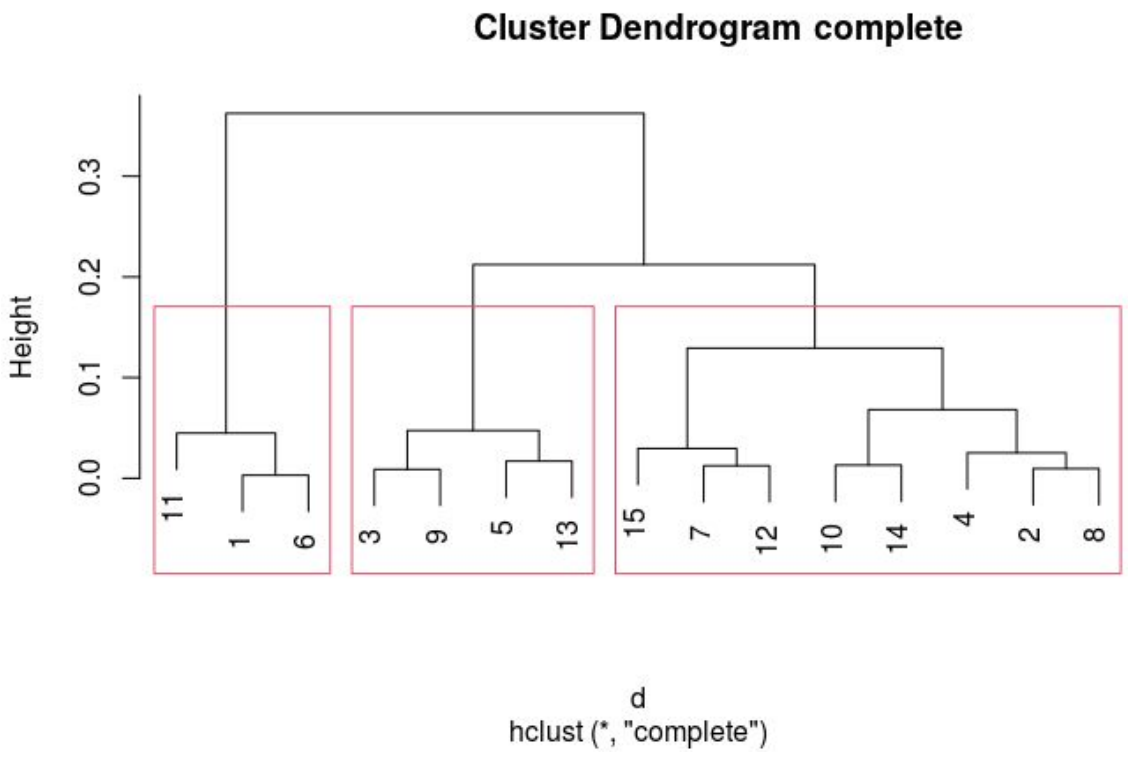
```
             A          R          N          D           ...
1  0.08631579 0.08631579 0.08631579 0.08631579           ...
```

```
2  0.03314917 0.03314917 0.03314917 0.03314917           ...
3  0.02524698 0.02524698 0.02524698 0.02524698           ...
4  0.03881279 0.03881279 0.03881279 0.03881279           ...
5  0.01463415 0.01463415 0.01463415 0.01463415           ...
...           ...           ...           ...             ...
```

Dari matriks di atas, dilakukan clustering dengan 2 metode. Metodenya yaitu complete dan centroid. Dendogram hasil dari metode centroid adalah sebagai berikut :



Sedangkan hasil dendogram untuk linkage complete adalah :



Perbedaan yang terlihat adalah di bagian height nya. Height (jarak data) centroid lebih pendek dibanding dengan metode complete linkage.

Dari analisis lanjutan :

```
> # Analisis cluster dengan centroid linkage
> res_cut_centr = do_clustering(res_mtx, "centroid", "euclidean", 3)
> res_ct_centr = contingency_table(res_cut_centr, res_mtx, c(1:15))
  cluster n
1       1 3
2       2 8
3       3 4

    1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
  1 1 0 0 0 0 1 0 0 0  0  1  0  0  0  0
  2 0 1 0 1 0 0 1 1 0  1  0  1  0  1  1
  3 0 0 1 0 1 0 0 0 1  0  0  0  1  0  0
> res_centr = purity_value(res_ct_centr, res_mtx)
> print(res_centr)
[1] 0.2

> # Analisis cluster dengan complete linkage
```

```
> res_cut_comp = do_clustering(res_mtx, "complete", "euclidean", 3)
> res_ct_comp = contingency_table(res_cut_comp, res_mtx, c(1:15))
  cluster n
1       1 3
2       2 8
3       3 4

    1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
  1 1 0 0 0 0 1 0 0 0  0  1  0  0  0  0
  2 0 1 0 1 0 0 1 1 0  1  0  1  0  1  1
  3 0 0 1 0 1 0 0 0 1  0  0  0  1  0  0
> res_comp = purity_value(res_ct_comp, res_mtx)
> print(res_comp)
[1] 0.2
```

Terlihat bahwa nilai purity antara complete dan centroid sama sama bernilai **0.2**

Dataset 2:

1. Compounds direpresentasikan oleh com_smiles
2. Lakukan ekstraksi fitur menggunakan PaDEL-MACCS
3. Lakukan analisis clustering pada file “dataset 2” dengan menggunakan agglomerative hierarchical clustering ke dalam 3 cluster
4. Lakukan clustering menggunakan complete dan centroid linkage, metric yang digunakan adalah Euclidean distance
5. Gambarkan dendogramnya dan analisis hasilnya

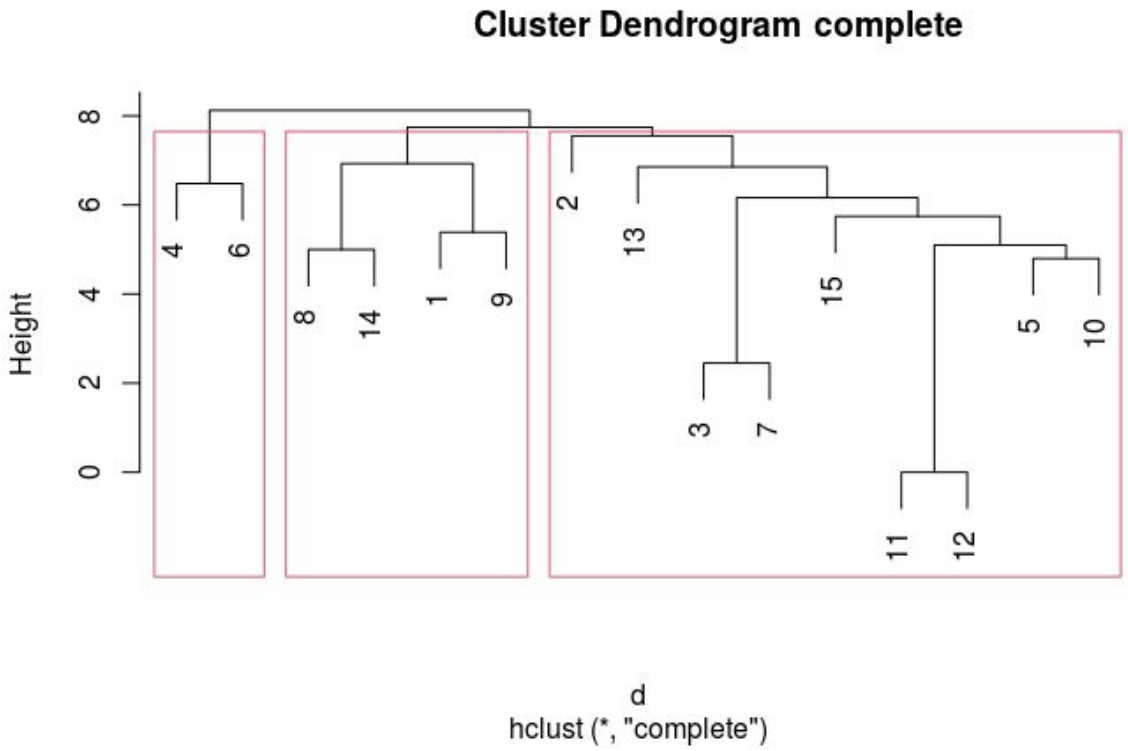
Jawaban 2 :

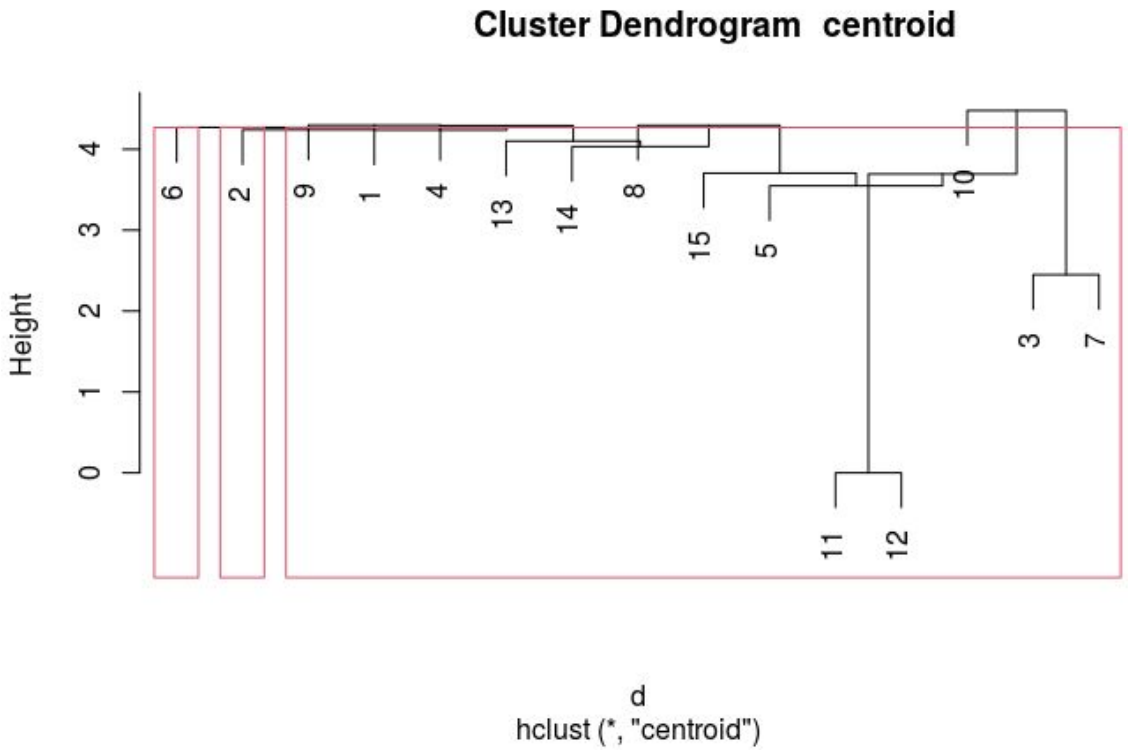
Untuk nomor 2, caranya sangat mirip dengan nomor 1. perbedaanya hanya di tahap menggabungkan file csv (saya menggunakan file csv yang terpisah) :

```
# Fungsi untuk generate dataframe
do_generate_df <- function(arr_files) {
  result <- matrix(nrow = length(arr_files), ncol = 166)
  for(i in seq_along(arr_files)) {
    f <- read.csv(paste(folder_name, arr_files[i], sep = "/"))
    for(j in 2:167) {
      result[i, j-1] = f[1, j]
    }
  }
  return(data.frame(result))
}
```

```
res_mtx = do_generate_df(csv_files)
```

Dendogram hasil clustering adalah berikut ini :





Gambar dendogram terlihat berbeda tetapi hasil perhitungan purity nya ternyata sama :

```
> # Analisis cluster dengan complete linkage
> res_cut_comp = do_clustering(res_mtx, "complete", "euclidean", 3)
> res_ct_comp = contingency_table(res_cut_comp, res_mtx, fasta_files[1:15])
  cluster n
1       1 4
2       2 9
3       3 2

    1.csv 10.csv 11.csv 12.csv 13.csv 14.csv 15.csv 2.csv 3.csv 4.csv 5.csv 6.csv 7.csv 8.csv 9.csv
  1     1      0      0      0      0      0      0     1     1     0     0     0     0     1     0
  2     0      1      1      0      1      0      1     0     0     1     1     1     1     0     1
  3     0      0      0      1      0      1      0     0     0     0     0     0     0     0     0
> res_comp = purity_value(res_ct_comp, res_mtx)
> print(res_comp)
[1] 0.2
>
> # Analisis cluster dengan centroid linkage
> res_cut_centr = do_clustering(res_mtx, "centroid", "euclidean", 3)
> res_ct_centr = contingency_table(res_cut_centr, res_mtx, fasta_files[1:15])
  cluster  n
1       1 13
2       2  1
3       3  1

    1.csv 10.csv 11.csv 12.csv 13.csv 14.csv 15.csv 2.csv 3.csv 4.csv 5.csv 6.csv 7.csv 8.csv 9.csv
  1     1      0      1      1      1      0      1     1     1     1     1     1     1     1     1
  2     0      1      0      0      0      0      0     0     0     0     0     0     0     0     0
  3     0      0      0      0      0      1      0     0     0     0     0     0     0     0     0
> res_centr = purity_value(res_ct_centr, res_mtx)
> print(res_centr)
[1] 0.2
```

Nilai purity yang diperoleh yaitu **0.2**