

Ejercicio 5

A) Halle la distancia genética de Prevosti entre variedades

Para calcular esta distancia usamos la siguiente sentencia

```
matriz_distancia <- dist(datos, method = "manhattan", diag = TRUE) / 33
```

Mostramos la matriz resultante de manera parcial, ya que es muy grande como para mostrarla de manera completa en una pagina.

	A1	A2	A3	A4	A5	A6	A7	A8	F1	F2
A1	0.00	0.27	0.30	0.33	0.30	0.39	0.39	0.33	0.55	0.47
A2	0.27	0.00	0.33	0.30	0.33	0.30	0.30	0.24	0.45	0.33
A3	0.30	0.33	0.00	0.03	0.18	0.15	0.15	0.09	0.42	0.41
A4	0.33	0.30	0.03	0.00	0.15	0.12	0.12	0.06	0.39	0.38
A5	0.30	0.33	0.18	0.15	0.00	0.09	0.09	0.15	0.42	0.41
A6	0.39	0.30	0.15	0.12	0.09	0.00	0.00	0.06	0.39	0.38
A7	0.39	0.30	0.15	0.12	0.09	0.00	0.00	0.06	0.39	0.38
A8	0.33	0.24	0.09	0.06	0.15	0.06	0.06	0.00	0.45	0.38
F1	0.55	0.45	0.42	0.39	0.42	0.39	0.39	0.45	0.00	0.18
F2	0.47	0.33	0.41	0.38	0.41	0.38	0.38	0.38	0.18	0.00

Pero lo que si podemos observar son los pares mas similares y los mas distintos.

Variedad 1	Variedad 2	Distancia
A6	A7	0
F3	F4	0
H3	H4	0
H5	H6	0
H7	H8	0
H7	H15	0
H8	H15	0

Variedad 1	Variedad 2	Distancia
F5	H7	0.59
F5	H8	0.59
F5	H11	0.59
F5	H15	0.59
A1	F5	0.60

Donde vemos que hay 7 pares de variedades que presentan valores identicos para sus variables moleculares, y que la distancia maxima entre pares es 0.6.

B) Podría aplicar el coeficiente de similitud SM ? Porque ?

Si, pero no lo hacemos porque perderiamos informacion, ya que las bandas presentan mas de 2 valores posibles. Categorizar los valores observados en solamente dos categorias implicaria una perdida de informacion.

C) Realice un Análisis de Coordenadas Principales para encontrar la configuración de las variedades de pepino en función de esta caracterización molecular. Encuentra asociaciones en función del tipo de pepino?

```
coordenadas_principales <- cmdscale(sqrt(matriz_distancia), eig = TRUE)
```

En la Figura 1 podemos ver que los pepinos del tipo **F** suelen encontrarse en el primer cuadrante, los del tipo **P** en el segundo cuadrante, los de tipo **A** en el tercero, y los de tipo **H** en el cuarto. Sin embargo esta ordenación es un tanto imprecisa, ya que por ejemplo, hay pepinos de los tipos **F** y **A** en el segundo cuadrante, así como pepinos del tipo **H** en el tercero. Si no estuvieran los colores que indican los tipos de pepinos, probablemente obtendríamos agrupamientos que estuvieran compuestos en su mayoría un único tipo de pepino, pero que también incluirían pepinos de otros tipos.

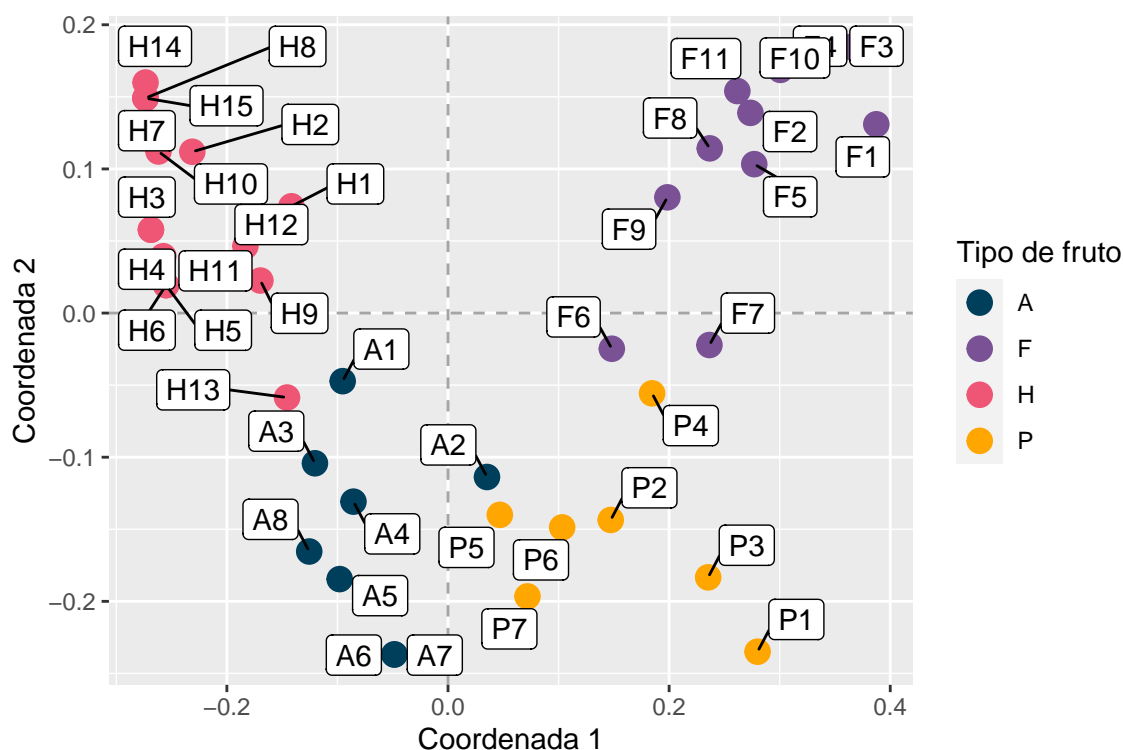


Figure 1: Caracterización molecular de las variedades de pepino en el plano principal.

D) Encuentre el dendrograma ultramétrico con ligamiento UPGMA

```
cluster_molecular <- hclust(matriz_distancia, method = "average")
```

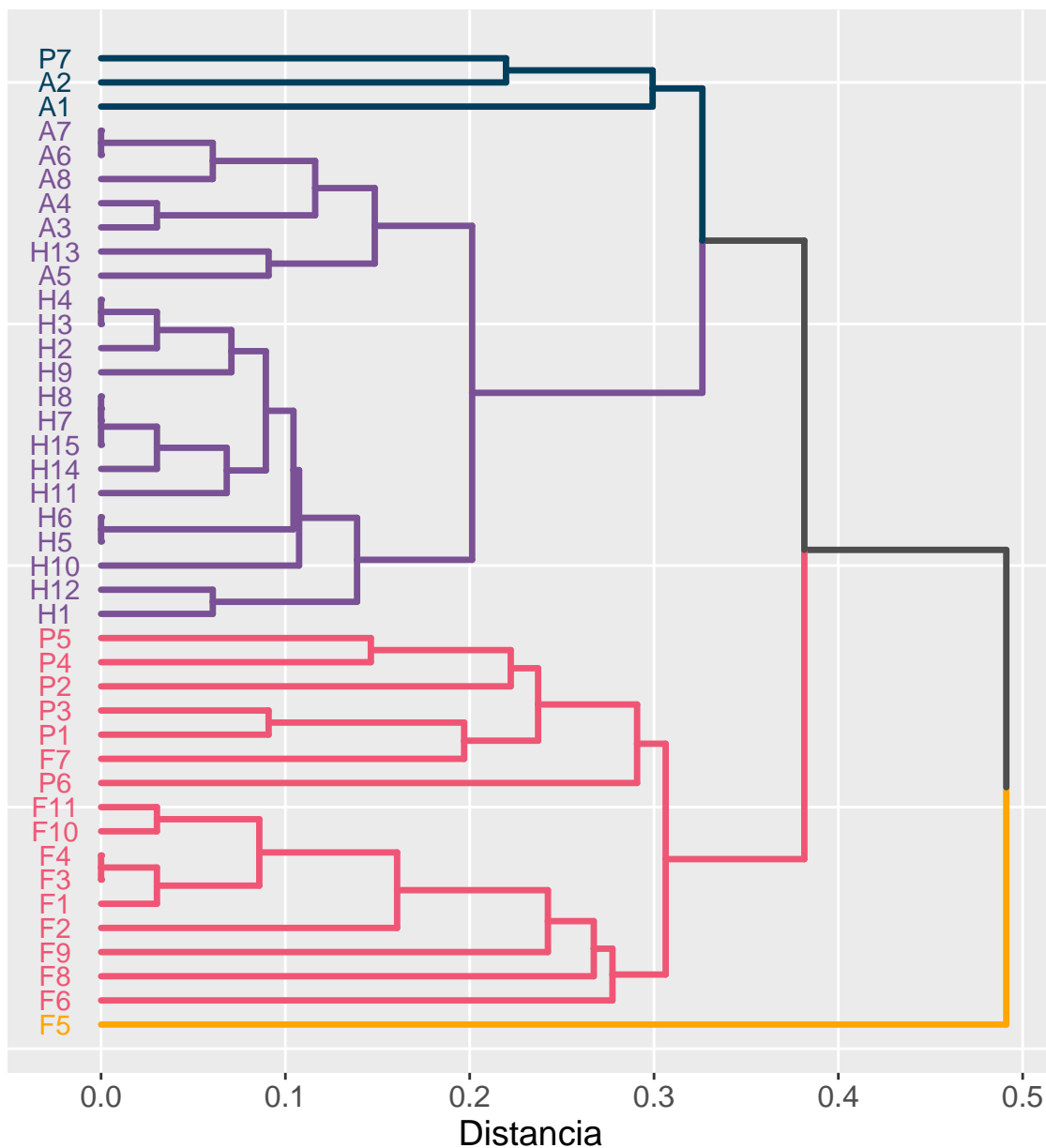


Figure 2: Dendrograma Ultrametrico con ligamiento UPGMA.

Resulta interesante comparar los agrupamientos representados en la Figura 2 con los representados en la Figura 2 del Ejercicio 2. En ambos casos sucede que hay una mezcla de variedades pertenecientes a los tipos alfa-beta y holandes. Sin embargo, en el agrupamiento de este ejercicio se observa que las variedades **A1** y **A2** no pertenecen al mismo cluster que el resto de las alfa-betas, sino que se unen con una variedad del tipo pepinillo, **P7**. Esto se corresponde con la cercanía entre estas tres variedades en el plano principal.

En este dendrograma, así como sucede con el dendrograma del Ejercicio 2, también se observa una variedad que se diferencia sustancialmente del resto de las variedades, la variedad **F5**. De solo haber mirado la Figura 1, hubieramos interpretado que esta variedad es similar al resto de los pepinillos franceses, cuando el dendrograma no sugiere lo mismo.

E) Mida a través de su matriz cofenética la concordancia con la matriz de distancias que le dio origen

La concordancia entre la matriz de distancias cofenética y la matriz de distancia original es igual a 0.872, lo que habla de una alta similaridad entre las mismas. En la Figura 3 se puede ver la asociación positiva entre las dos medidas de distancia. La dispersión en la nube de puntos aumenta a medida que la distancia es mayor, lo que significa que las dos medidas de distancia tienden a diferir más cuando la distancia entre los tipos de pepinos es mayor.

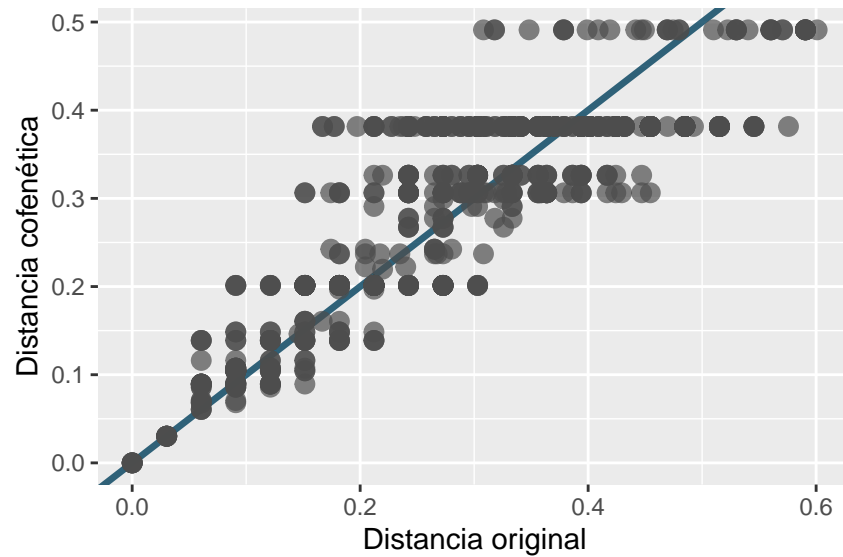


Figure 3: Grafico de dispersion entre distancia original y distancia cofenética a partir de d endograma Ultrametrico con ligamiento UPGMA. La linea azul representa a la recta identidad.

F) Halle el dendrograma aditivo Neighbor Joining

```
rapds_nj <- nj(matriz_distancia)
```

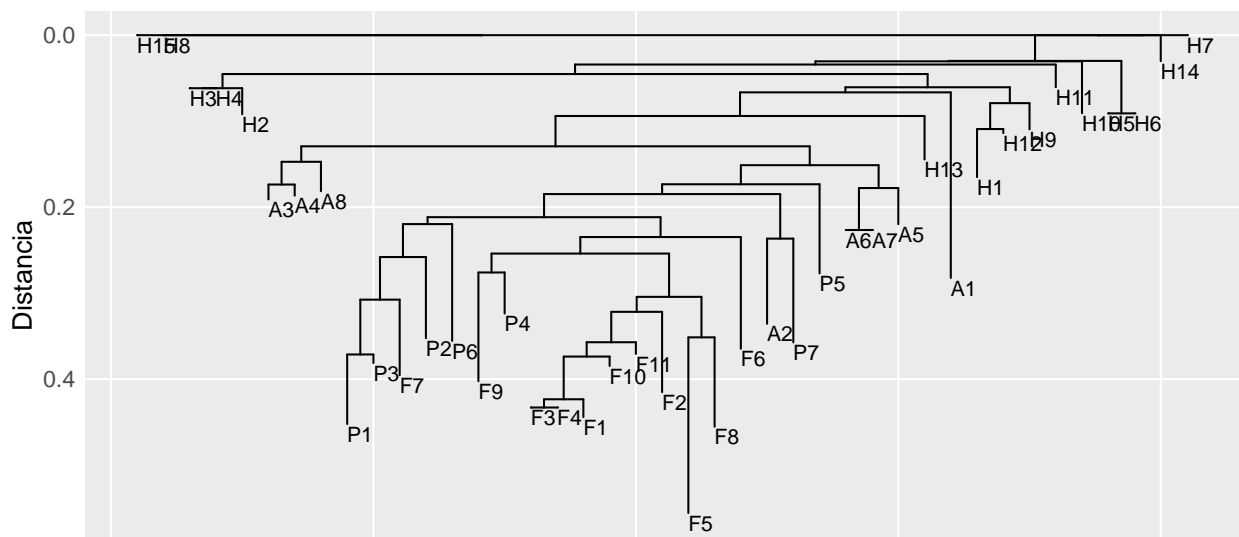


Figure 4: Representacion jerarquizada del arbol aditivo Neighbor-Joining

G) Mida su concordancia con matriz de distancia original

```
distancia_cofenetica_nj <- as.dist(cophenetic(rapds_nj), diag = TRUE, upper = FALSE)
concordancia <- cor(distancia_cofenetica_nj, matriz_distancia)
```

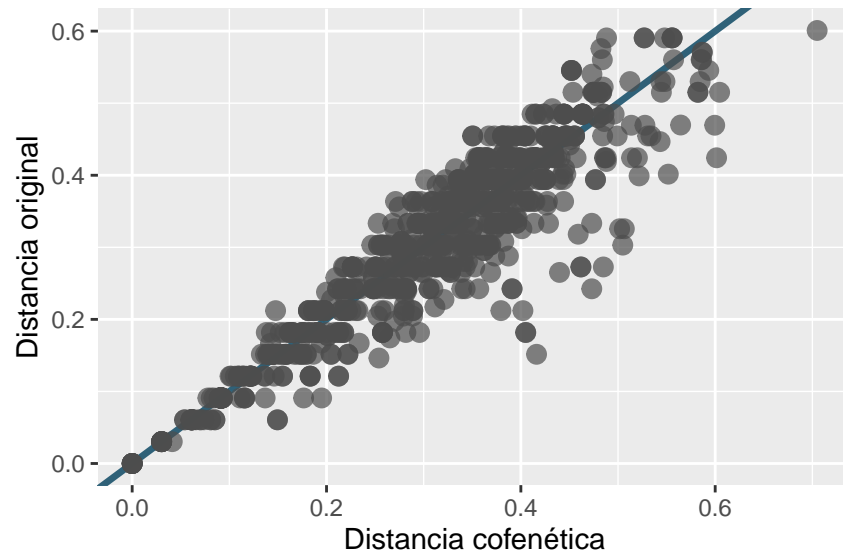


Figure 5: Grafico de dispersion entre distancia original y distancia cofenética a partir del dendograma aditivo. La linea azul representa a la recta identidad.

La concordancia entre la matriz de distancias cofenética construida a partir del dendograma aditivo Neighbor Joining y la matriz de distancia original es igual a 0.926, por lo que este arbol aditivo es el que mejor representa las distancias originales entre las variedades, y se corresponde con lo mencionado en clase de que en general estos arboles aditivos dan mejor.

H) Relacione ambos dendogramas y saque conclusiones

```
concordancia <- cor(distancia_cofenetica_nj, distancia_cofenetica_upgma)
```

La concordancia entre ambos dendogramas es 0.872. Es decir, en ambos casos se preserva altamente el ordenamiento entre variedades pero no de manera perfecta. Por ejemplo, si miramos la Figura 4, vemos que la variedad **P7** se uniría primero a la variedad **A2**, pero no así a la variedad **A1**, como si sucede en la Figura 2.