ANÁLISIS DE EXPRESIONES DE GENES DE TUMORES CANCERÍGENOS MEDIANTE TÉCNICAS DE CLUSTERING

José Antonio Pozo Núñez

Técnicas Inteligentes en Bioinformática

Máster Universitario en Lógica, Computación e Inteligencia Artificial

Índice

1. Introducción

En los últimos tiempos se han producido importantes avances tecnológicos que han afectado a prácticamente a todos los sectores existentes, generando un gran impacto y desarrollo en dichos ámbitos. Dentro de estos, el campo de la Biología ha evolucionado y se ha beneficiado tremendamente gracias al progreso informático que hemos vivido en las últimas décadas.

La ingente cantidad de datos que genera la Biología necesita de técnicas para almacenar, organizar y manejar toda esa información, por supuesto de manera digital. Por tanto, se puede definir la Bioinformática como el resultado de combinar la Biología con la Tecnología de la Información y de la Computación. Es una nueva área de la ciencia que utiliza métodos computacionales para responder a cuestiones biológicas.

La Genómica, campo de la Biología Molecular, entendida como la disciplina encargada del estudio integral de los genomas utilizando tecnologías avanzadas de alto rendimiento, está revolucionando la investigación biomédica, o lo que es lo mismo, nuestra forma de investigar las causas de las enfermedades humanas. La medicina personalizada va a revolucionar la salud y jugará un papel dominante en el futuro de la terapia del cáncer. Los análisis a nivel genómico de los tumores se convertirán en una práctica rutinaria en la clínica y podrá ser posible identificar a los pacientes que se beneficien de un tratamiento en base a su perfil molecular.

En este contexto, existen multitud de técnicas y herramientas para el análisis de esa gran cantidad de información generada por las distintas investigaciones de tumores cancerígenos. Dentro de estas técnicas existe el llamado Clustering, que no es más que un procedimiento de agrupación en base a un criterio. Se lo considera una técnica de aprendizaje no supervisado puesto que busca encontrar relaciones entre variables descriptivas pero no la que guardan con respecto a una variable objetivo.

Este proyecto viene motivado por dos razones principales, primeramente, por el propósito de profundizar en el ámbito de las técnicas Clustering, el lenguaje R y la herramienta Weka, y como segunda finalidad, la de aplicar esos conocimientos en un caso práctico como el que nos ocupa. Utilizaremos estás técnicas y herramientas para analizar una gran cantidad de información sobre genes cancerígenos.

2. Proyecto Pan-Cáncer

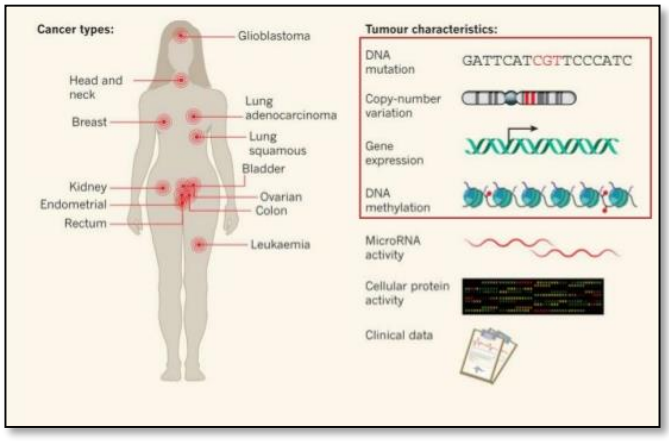
El proyecto Pan-Cáncer fue planteado por la red de investigadores integrados en el proyecto de El Atlas del Genoma del Cáncer (TCGA). Con esta iniciativa, se pretende observar alteraciones comunes entre diferentes linajes de tumores con el fin de diseñar terapias efectivas en un tipo de cáncer y poder extenderlas a otros perfiles tumorales similares. La red de investigadores integrados en el proyecto de El Atlas del Genoma del Cáncer propone este proyecto como una iniciativa coordinada cuyos objetivos principales serían los siguientes: identificar y analizar aberraciones en el genoma tumoral y el fenotipo que definan distintos linajes de cáncer, así como identificar aberraciones trascendentes en linajes tumorales concretos.

Se escogieron 12 tipos tumorales, ya analizados de forma individual en el proyecto de El Atlas del Genoma del Cáncer, y se procedió a caracterizar su genoma y su epigenética para identificar rutas biológicas comunes y elementos regulatorios activados o desactivados.

Para ello, se procede de la siguiente forma. Se comienza obteniendo muestras de los tumores a estudiar desde diferentes plataformas de tejidos. Posteriormente, se procede a purificar DNA, RNA y proteínas para, después, mandar las preparaciones a centros de secuenciación y de caracterización que realicen un perfil molecular. Finalmente, estos datos son depositados en el centro de coordinación de datos de TCGA y se acaban interpretando.

Dentro del campo de la genética, un concepto muy importante es el de la expresión génica, que no es más que el proceso por el cual las instrucciones genéticas son utilizadas de sintetizar productos del gen. Estos productos son generalmente las proteínas

En este proyecto utilizaremos mediciones de las expresiones de genes de distintos tumores para realizar el análisis Clustering.



Proceso de generación de datos por la TCGA y el proyecto Pan-Cáncer

3. Clustering

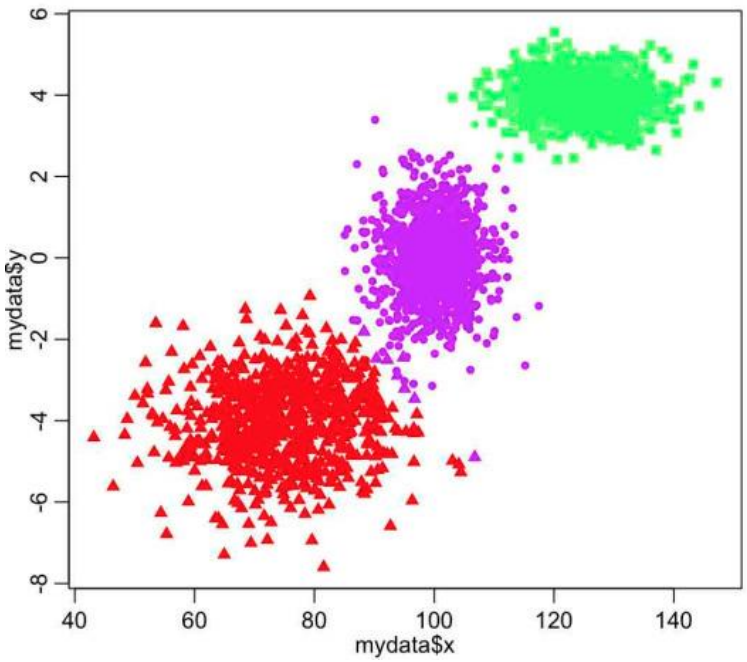
El Clustering es una tarea que consiste en agrupar un conjunto de objetos (no clasificados) en subconjuntos de objetos llamados Clusters. Cada Cluster está formado por una colección de objetos que son similares entre sí, pero que son distintos respecto a los objetos de otros Clusters.

En el ámbito del aprendizaje automático, el Clustering se enmarca dentro del aprendizaje no supervisado, es decir, que para esta técnica solo disponemos de un conjunto de datos de entrada, sobre los que debemos obtener información sobre la estructura del dominio de salida, que es una información de la cual no se dispone.

Existen dos grandes técnicas para el agrupamiento de casos:

- Agrupamiento jerárquico, que puede ser aglomerativo o divisivo.
- Agrupamiento no jerárquico, en los que el número de grupos se determina de antemano y las observaciones se van asignando a los grupos en función de su cercanía. Por ejemplo, el método de k-mean.

En este proyecto, utilizaremos el agrupamiento jerárquico y no jerárquico para analizar un conjunto de datos que contiene información acerca de las expresiones de genes cancerígenos de un conjunto de personas enfermas con cáncer.



Ejemplo de Clustering

4. Dataset

El dataset utilizado en este proyecto contiene información sobre las expresiones de genes cancerígenos de un conjunto de personas que padece distintos tipos de cáncer. Las características de dicho conjunto de datos son las siguientes:

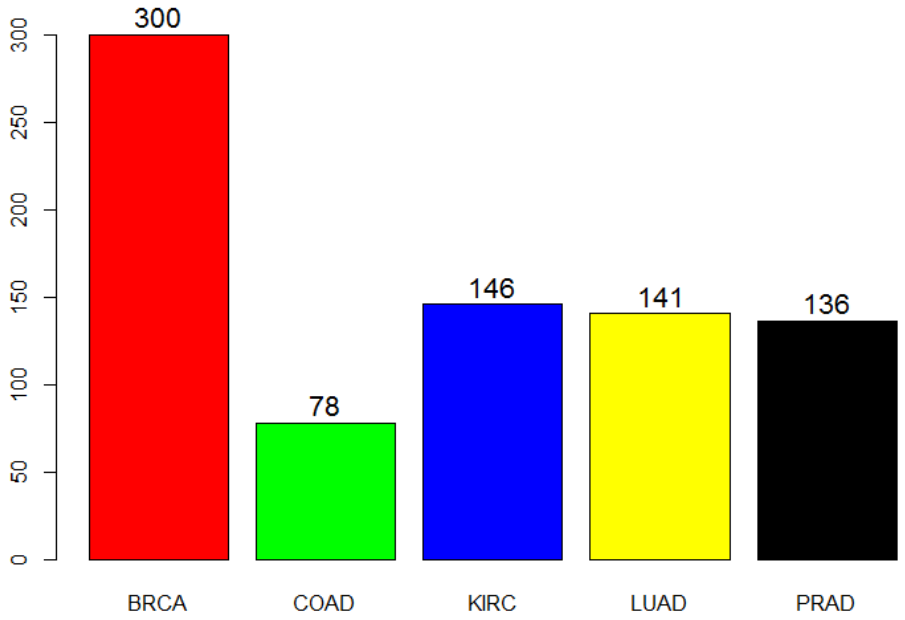
- 801 ejemplos, cada ejemplo corresponde a una persona distinta enferma de cáncer.

- 20531 atributos genéticos, no son más que mediciones de la expresión génica de distintos genes.

- 1 atributo con la clase de tumor, usado para validar el modelo, no para la construcción del mismo, las distintas clases, y por tanto, los distintos tipos de cáncer con los que vamos a tratar son:

    - BRCA: Cáncer de mama.
    - COAD: Cáncer de colon.
    - KIRC: Cáncer de células renales.
    - LUAD: Cáncer de pulmón.
    - PRAD: Cáncer de próstata.

La distribución de frecuencias de cada clase de cáncer es la siguiente:



Distribución de frecuencias de cada clase

5. Proceso de Clustering

Para la implementación de todo el proceso he utilizado el lenguaje R, el IDE RStudio, y la implementación por defecto de los distintos algoritmos de agrupamiento que contiene R.

5.1. Preprocesamiento

Primeramente, realizamos un preprocesado previo a la construcción del modelo. Comenzamos comprobando y eliminando los ejemplos que contienen valores vacíos, para ello ejecutamos lo siguiente:

```
> #Comprobamos que no haya valores vacíos, si es así, los informamos con la media de la columna
> genes.cancer.dataset.omit <- na.omit(genes.cancer.dataset)
> porcentaje.valores.vacios <- (1-(nrow(genes.cancer.dataset.omit)/nrow(genes.cancer.dataset)))
> print(paste("Porcentaje de valores vacíos:",porcentaje.valores.vacios))
[1] "Porcentaje de valores vacíos: 0"
```

Se observa que el porcentaje de ejemplos con valores vacíos es cero, por tanto, no se ha eliminado ningún ejemplo. A continuación, comprobamos y eliminamos los atributos (genes) cuyos valores sean todos 0 para todos los ejemplos, para ello ejecutamos lo siguiente:

```
> #Eliminamos los Genes (atributos) que tengan todos sus valores a cero
> genes.valores.cero <- obtener_genes_a_cero(genes.cancer.dataset)
> genes.cancer.dataset <- eliminar_atributos(genes.cancer.dataset,genes.valores.cero)
> print(paste("Numero de atributos después de la eliminación:",length(genes.cancer.dataset)))
[1] "Numero de atributos después de la eliminación: 20264"
```

Tras este paso, se eliminan un total de 267 atributos (genes), el siguiente paso es realizar un normalizado, para ellos pintamos la siguiente tabla:

```
> #Comprobamos el rango de varios atributo
> apply(genes.cancer.dataset[25:30], 2, range)
      gene_26   gene_27  gene_28  gene_29  gene_30  gene_31
[1,]  0.00000  6.200934  0.00000  2.20389 0.000000 0.000000
[2,] 13.26481 11.537946 10.25141 13.10509 9.962827 9.520505
```

Comprobamos que el rango de los distintos atributos son similares, además como todos los atributos miden lo mismo, por lo tanto, no es necesario un normalizado previo. Como última medida, eliminamos los valores outliers, pero al intentarlo, el sistema devuelve el siguiente error:

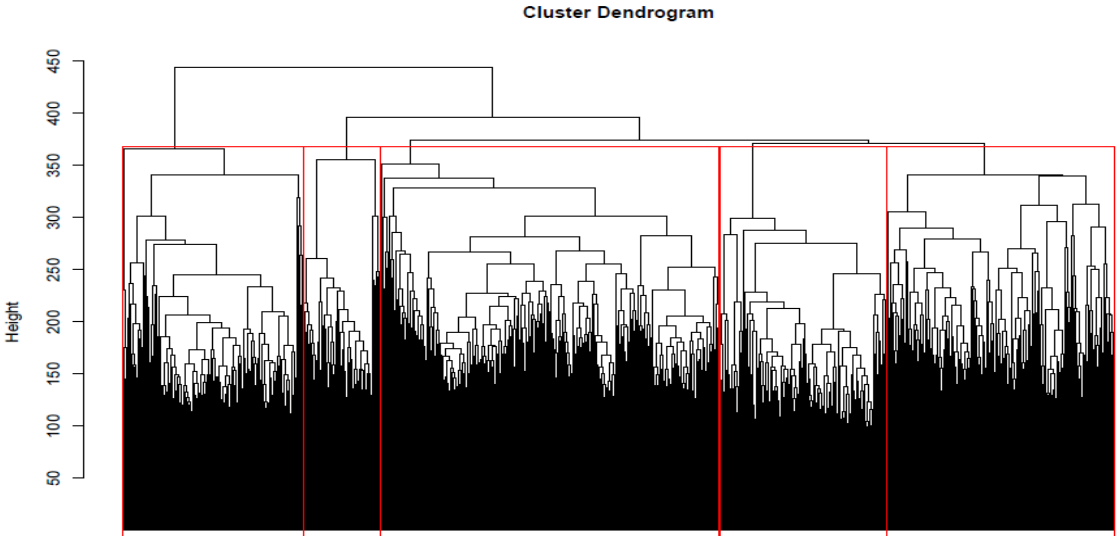
```
> mvOutlier(genes.cancer.dataset)
Error in covMcd(data, alpha = alpha) : n <= p -- you can't be serious!
```

5.2. Construcción y análisis del modelo

Para la construcción y análisis del modelo, comenzamos ejecutamos distintas configuraciones del algoritmo jerárquico para comprobar si se diferencian correctamente los 5 grupos.

- Distancia Euclídea y Enlace Completo

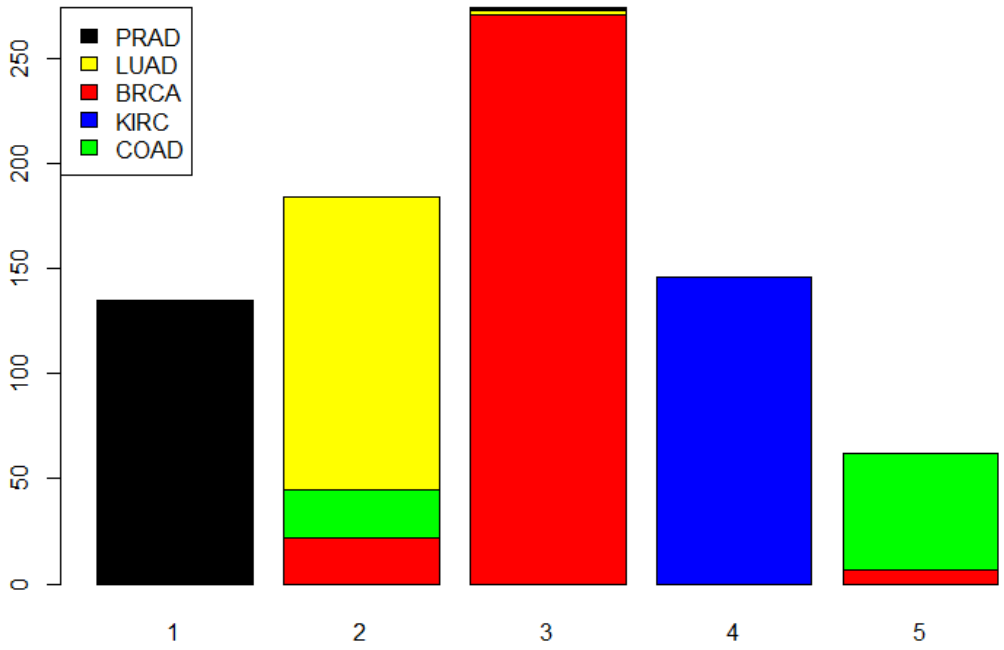
```
> #Calculamos la matriz de distancia, utilizando la distancia euclidea
> genes.cancer.dist<-dist(genes.cancer.dataset,method = "euclidean")
> #Construimos el modelo de clustering jerárquico, con distancia completa
> genes.cancer.modelo.hclust<-hclust(genes.cancer.dist,method = "complete")
> #Pintamos el árbol jerárquico
> plot(genes.cancer.modelo.hclust, hang = -1, cex = 0.6)
> #Dividimos el árbol en 5 clústers, como el número de clases
> rect.hclust(genes.cancer.modelo.hclust, k = 5, border = "red")
```



Árbol Jerárquico

Se aprecia cómo se divide correctamente en los 5 grupos estudiados.

```
> barplot(table(genes.cancer.dataset.groups$Class, genes.cancer.dataset.groups$groups)
,col = c("red","green","blue","yellow","black"))

[1] "Porcentaje de elementos bien agrupados con la distancias Euclidea y Completa:
 0.931335830212235"
```

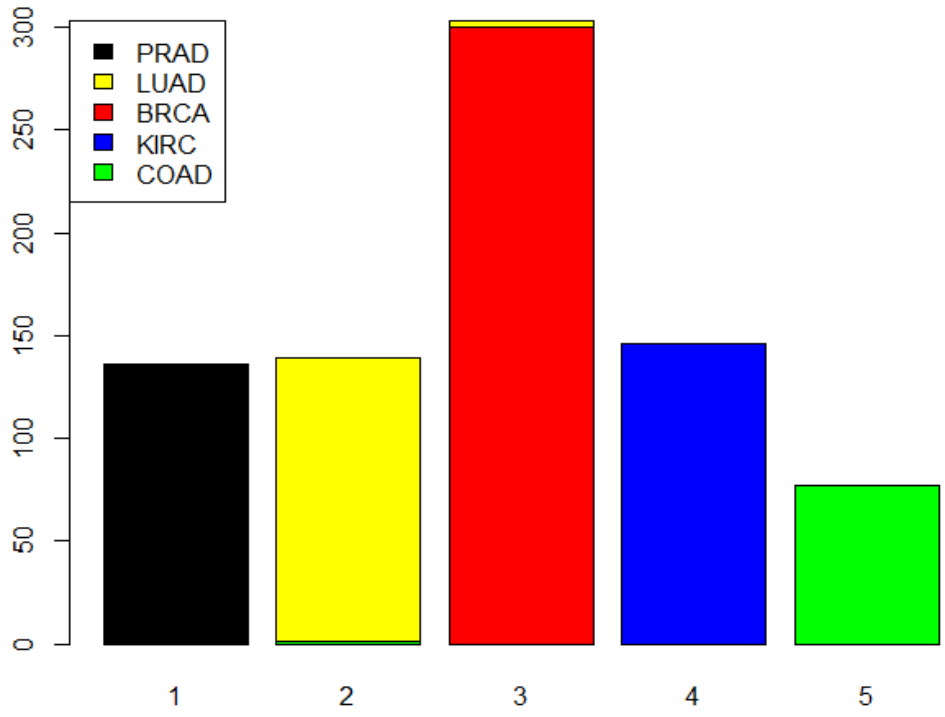


Distancia Euclídea y Enlace Completo

Si examinamos la gráfica anterior, parece que en términos generales, el algoritmo ha agrupado correctamente la mayoría de ejemplos, pero ya comenzamos a constatar que los grupos LUAD, COAD y BRCA son más similares entres ellos que con los demás grupos.

- Distancia Euclídea y Enlace War.D

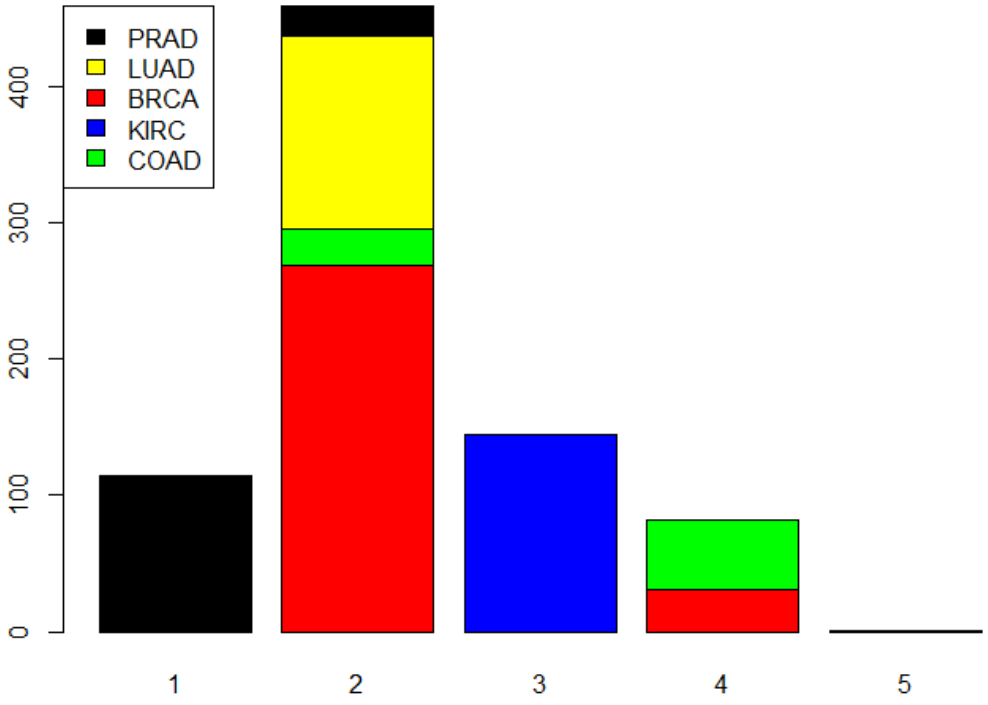
```
[1] "Porcentaje de elementos bien agrupados con la distancias Euclidea y Ward.D:
 0.995006242197253"
```



Distancia Euclídea y Enlace War.D

- Distancia Manhattan y Enlace Completo

```
[1] "Porcentaje de elementos bien agrupados con la distancias Manhattan y
Completa: 0.722846441947566"
```

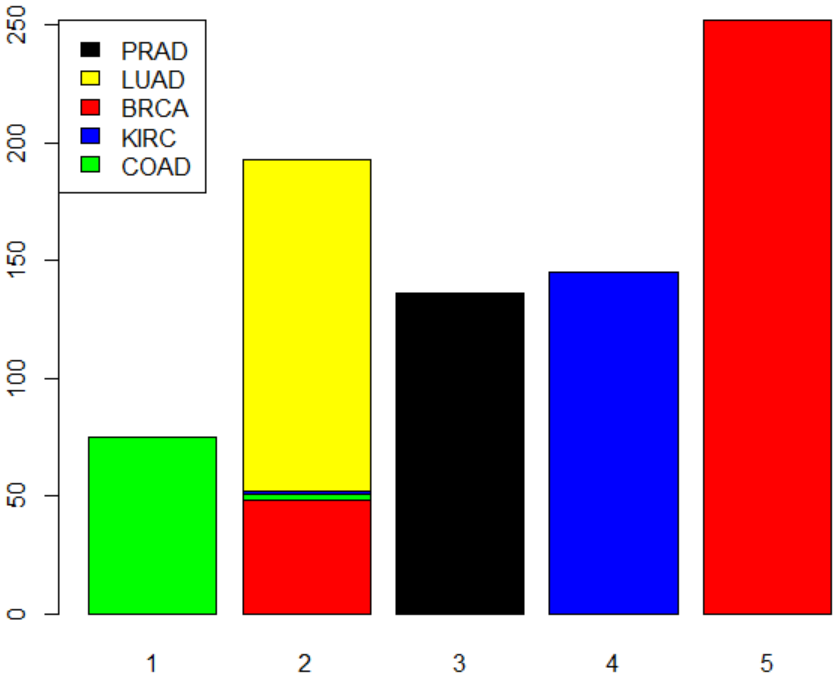


Distancia Manhattan y Enlace Completo

Parece que la configuración del algoritmo jerárquico que obtiene mejores resultados es la de distancia Euclídea y enlace War.D y realmente es algo que tiene bastante sentido ya que la distancia euclídea se usa cuando cada dimensión mide propiedades similares y la Manhattan en caso contrario.

Después de constatar con el algoritmo jerárquico que los grupos se dividen correctamente, ahora utilizamos el algoritmo K-medias, con k=5, para observar cómo se dividen los clusters y analizar los resultados.

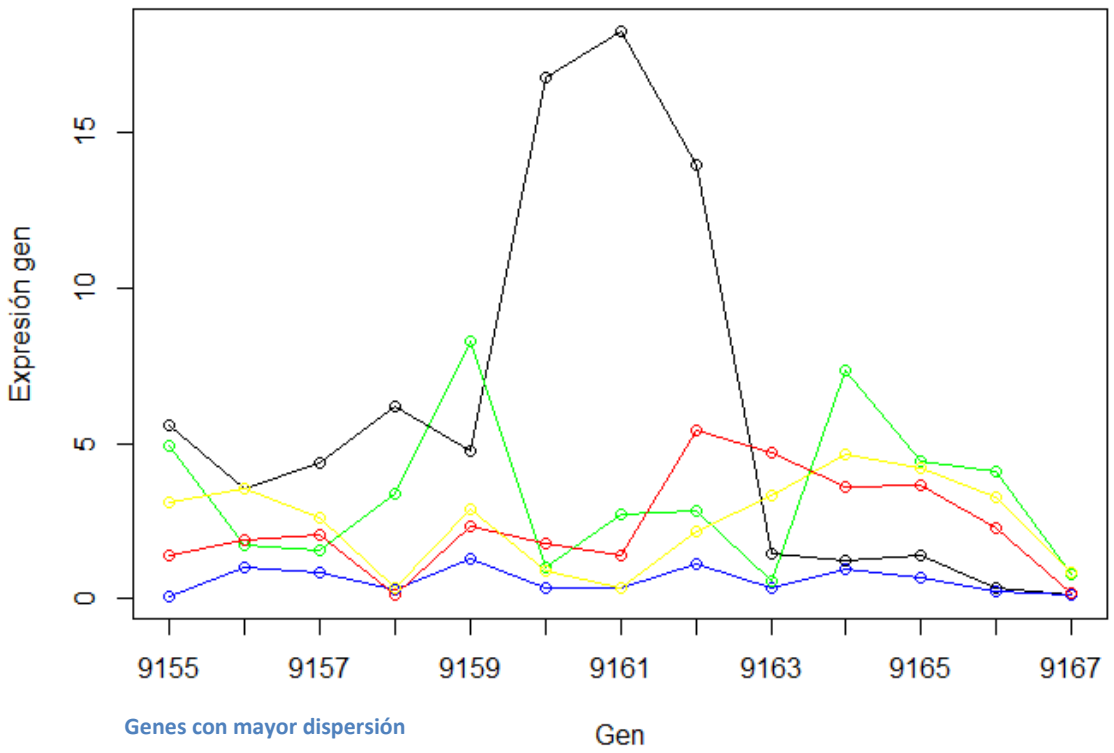
```
> #Construimos el modelo kmeans
> genes.cancer.modelo.kmeans <- kmeans(genes.cancer.dataset,5)
> table(genes.cancer.modelo.kmeans$cluster, as.vector(t(genes.cancer.dataset.class)))

    BRCA COAD KIRC LUAD PRAD
  1    0   75    0    0    0
  2   48    3    1  141    0
  3    0    0    0    0  136
  4    0    0  145    0    0
  5  252    0    0    0    0

[1] "Porcentaje de elementos bien agrupados con k-medias: 0.935081148564295"
```



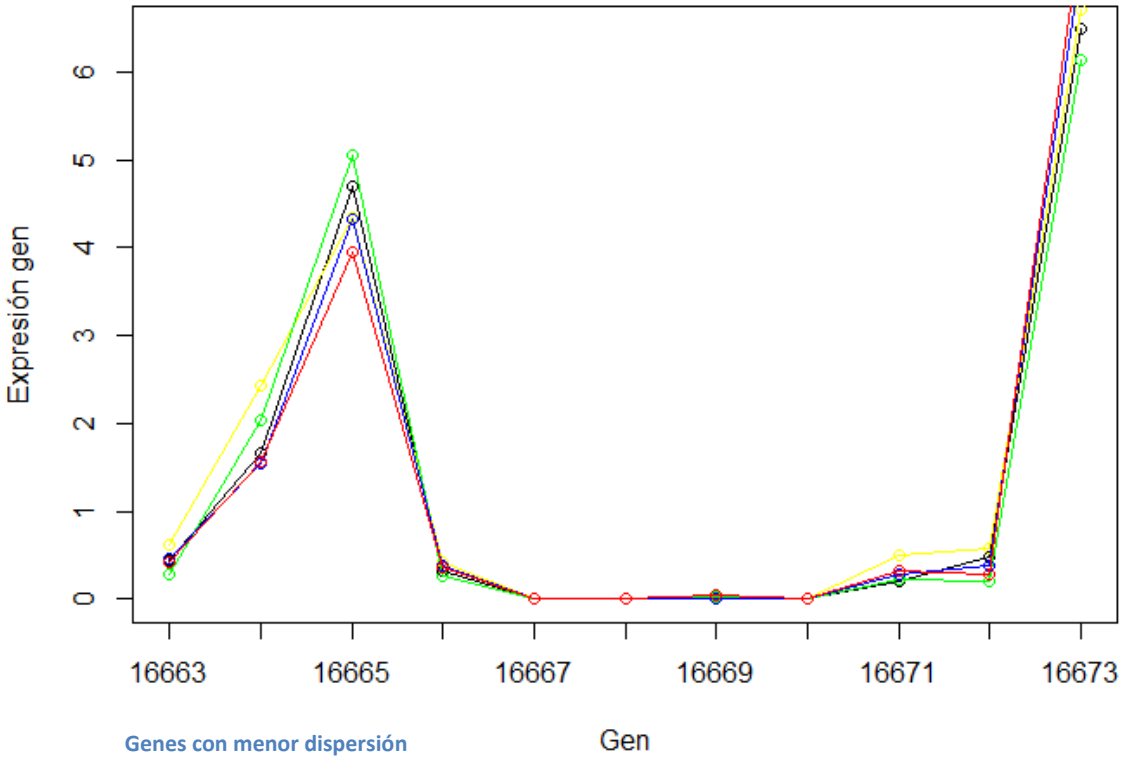
K-medias, 5 clusters

De nuevo, esta vez, mediante el algoritmo K-medias, comprobamos que los grupos LUAD, COAD y BRCA mantienen cierta similitud, y que los grupos PRAD y BRCA son distintos a todos los demás.

Ahora, después de haber ejecutado el algoritmo K-medias y obtener los clusters, mediante una medida de dispersión, como la varianza, obtenemos aquellos atributos (genes), cuyos centroides tienen mayor y mejor dispersión.



Genes con mayor dispersión



Genes con menor dispersión

6. Weka

En este apartado utilizamos la herramienta Weka para generar distintos modelos de Clustering a partir del dataset anteriormente comentado mediante el algoritmo K-medias, más específicamente, ejecutamos dicho algoritmo con 5 y 3 clusters respectivamente y analizamos su resultado.

- 5 Clusters:

```
                        Classes to Clusters:

                          0   1   2   3   4  <-- assigned to cluster
Cluster 0 <-- LUAD        0   0 136   0   0 | PRAD
Cluster 1 <-- KIRC      141   0   0   0   0 | LUAD
Cluster 2 <-- PRAD       41   0   0   0 259 | BRCA
Cluster 3 <-- COAD        1 145   0   0   0 | KIRC
Cluster 4 <-- BRCA        4   0   0  74   0 | COAD

Incorrectly clustered instances :       46.0      5.7428 %
```

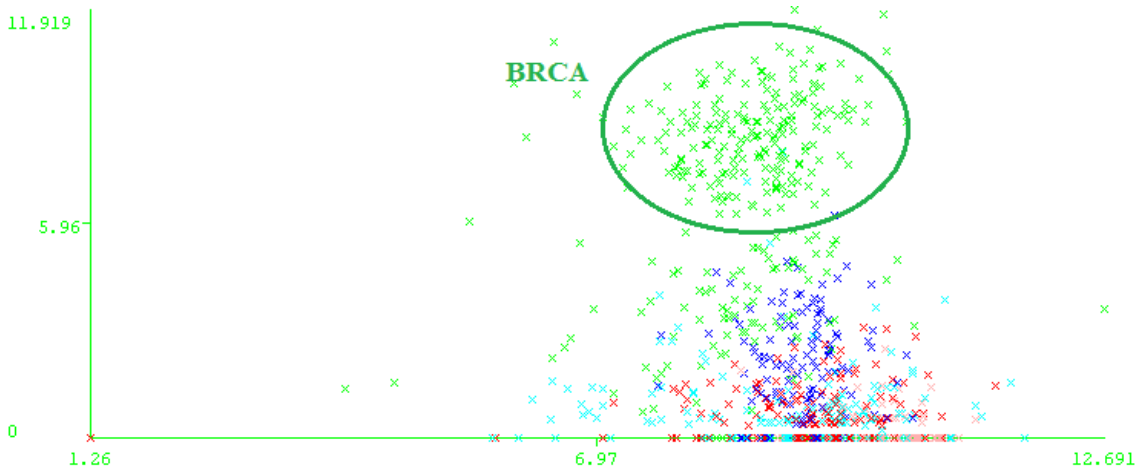
Se comprueba visiblemente que las agrupaciones coinciden en un alto porcentaje con las clasificaciones, apenas hay un 5% de ejemplos no agrupados correctamente.

- 3 Clusters:

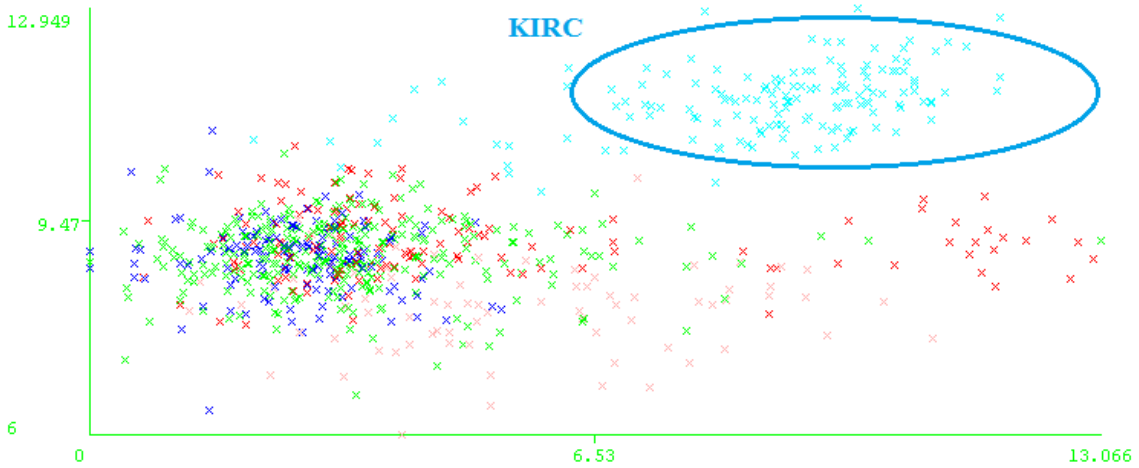
```
                        Classes to Clusters:

                          0   1   2  <-- assigned to cluster
                          0   0 136 | PRAD
                        141   0   0 | LUAD
Cluster 0 <-- BRCA      300   0   0 | BRCA
Cluster 1 <-- KIRC        1 145   0 | KIRC
Cluster 2 <-- PRAD       78   0   0 | COAD

Incorrectly clustered instances :       220.0    27.4657 %
```

Parece que una vez más se hace patente que existen 3 grandes grupos diferenciados, PRAD, KIRC y, por otro lado, todos los demás. Es evidente que el alto porcentaje de ejemplos no agrupados correctamente se debe a que hemos agrupado ejemplos de 5 distintas clases en tan solo 3 clusters.

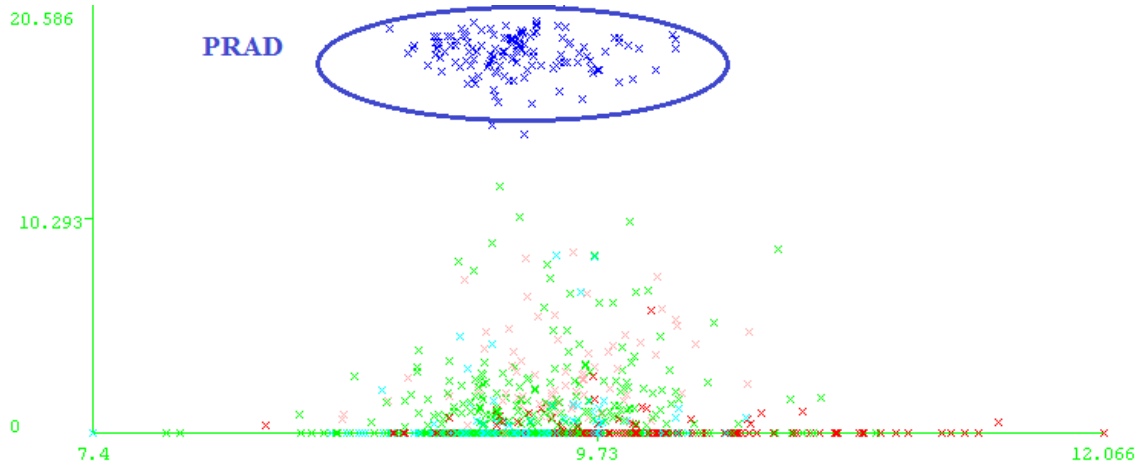
A continuación, se muestran distintas gráficas donde se evidencian grandes diferencias entre los distintos clusters en base a ciertos genes (atributos).



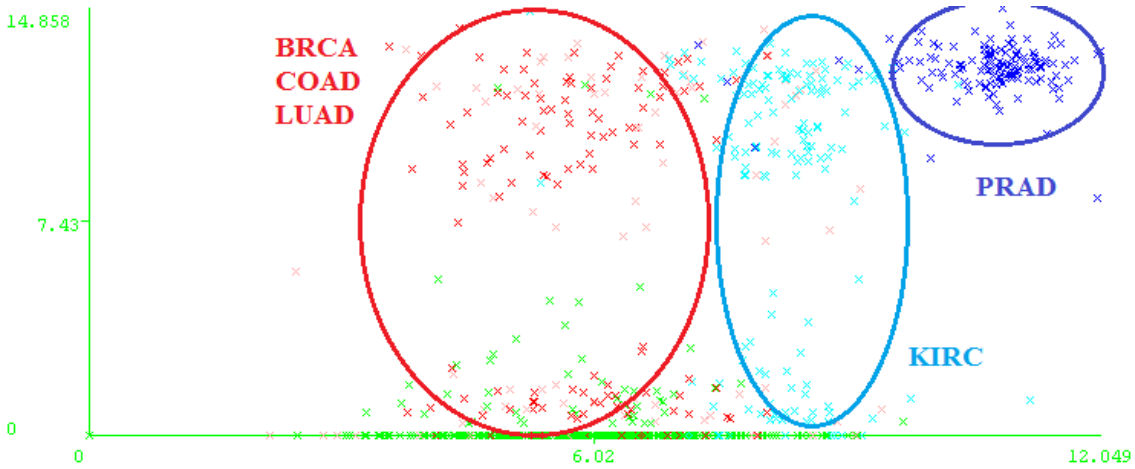
Gen 6530 - Diferencia entre clase BRCA y todas las demás



Gen 18178 - Diferencia entre clase KIRC y todas las demás



Gen 9176 - Diferencia entre clase PRAD y todas las demás



Gen 15301 - Diferencia entre las clases PRAD, KIRC y todas las demás

7. Conclusiones

La importancia de avanzar en el conocimiento a nivel genético de los distintos tipos de cáncer es vital para la lucha contra esta enfermedad. Tras este pequeño estudio que hemos realizado, podemos afirmar que existen ciertas diferencias y similitudes a nivel génico entre los distintos tipos de tumores estudiados. Como hemos ido comprobando en el análisis de los resultados de este proyecto, parece que el cáncer de mama, de colón y de pulmón tienen más similitudes a nivel genético en comparación con los restantes, por el contrario, el cáncer de próstata se diferencia considerablemente de todos los demás, igualmente ocurre con el cáncer de células renales.

Toda esta información supone un conocimiento muy valioso para los expertos en el dominio del problema. La envergadura de los proyectos Pan-Cáncer y Atlas del Genoma del Cáncer hace necesario sinergias entre expertos en genética y medicina, e ingenieros y científicos de ramas más técnicas, que colaboren y den soporte a la consecución de resultados que ayuden a acabar por fin con esta enfermedad.

A nivel personal, podemos reseñar especialmente la dificultad y complejidad de trabajar con un dataset tan sumamente grande, el simple hecho de intentar abrir el fichero en cualquier aplicación suponía varios minutos, además de dejar el ordenador ciertamente ralentizado. Esto implica lo complicado de, por ejemplo, realizar una visualización por encima de los datos.

Además de esto, y aunque R no es excesivamente complejo, sí que me he encontrado con la traba de aprender y tratar con un lenguaje desconocido hasta entonces para mí, afortunadamente, existe abundantes recursos en internet que ha ayudado a salvar ese problema.

Bibliografía

- UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq
- Proyecto Pan Cáncer: https://es.wikipedia.org/wiki/Proyecto_Pan-Cáncer
- Algoritmo de Agrupamiento: https://es.wikipedia.org/wiki/Algoritmo_de_agrupamiento
- Clustering: https://jarroba.com/que-es-el-clustering
- Universidad del País Vasco: http://www.ehu.eus
- Sociedad Española de Bioquímica y Biología Molecular: https://www.sebbm.es