ProyectoClustering_EduardoArrieta

Eduardo Alejandro Arrieta Donato

El presente escrito presenta la resolución del proyecto correspondiente al módulo de Clustering de la materia de Bioinformática del cuarto semestre de la Licenciatura en Ciencias Genómicas de la UNAM.

Todos los datos biológicos y código en R usado se encuentran en un repositorio GitHub indicado en la sección de discusión.

Introducción

El proyecto consiste en la ejecución de métodos de clustering para formar árboles de relación entre proteínas sin la necesidad de realizar una alineamiento múltiple previó de sus secuencias de aminoácidos. Se nos proporcionó una fasta con las secuencias de aminoácidos de 100 proteínas ABC (ATP-binding cassette) de organismos, con estas, se establecieron distancias por disimilitud de los valores *bitscores* resultantes de hacer un BLAST contra ellas mismas y por último se implementaron los métodos de clustering jerárquico de unión mínima (single), máxima (complete), media (average) y por Método de variación mínima de Ward (ward.D2).

Es importante aclarar que los métodos de clustering son una herramienta de exploración que se debe realizar una vez conocido y explorado los datos; estos métodos desde un conjunto de números estimas relaciones de cercanía los objetos estudiados, que matemáticamente esté correcta la agrupación, puede que no represente una verdad biológica.

La utilidad del uso de estos métodos recae en la capacidad del usuario de poder interpretar los resultados y poderles dar una explicación a partir de su conocimiento previo, por lo que que siempre se debe ser precavido con el uso de clustering.

Discusión

Las proteínas etiquetas con ABC son una familia de transportadores activos de un gran número de proteínas transmembrana diversas. Transportan compuestos a través de membranas contra gradientes de concentración con hidrólisis del ATP. Estos sustratos incluyen aminoácidos, lípidos, iones inorgánicos, péptidos, sacáridos, péptidos para la presentación de antígenos, metales, fármacos y proteínas. La presencia de estos transportadores en diferentes seres vivos nos indica una importancia evolutiva.

Para el inicio del proyecto se realizó un BLAST de las 100 proteínas recibidas contra ellas mismas, se usó la siguiente línea de comandos en los servidores de la licenciatura:

```
blastp -query ABC.faa -subject ABC.faa -outfmt 7 -evalue 100 -max_hsps 1 -use_sw_tback > ABCvABC_blast.out
```

Las proteínas ya se encontraban etiquetadas con ABC1, ABC2 y ABC3 para su pronta identificación.

Tras una inspección rápida con bash se detectaron la ausencia de resultados del BLAST, de 10 000 hits, el programa devolvió 9 992, esos 8 faltantes fueron omitidos por mostrar una E-value fuera de los rangos permitidos, al no ser significativo la relación de esas proteínas se llenaron con 0s los espacios.

De los resultados obtenidos solo nos importa el bitscore para obtener la relación de las proteínas, por lo que se generó una matriz pareada de 100 por 100 (cada proteína contra cada proteína), la matriz es cuadrada pues no es lo mismo un resultado de A a B que de B a A. Se esperar que los valores la diagonal de la matriz sean los más altos pues son los lugares donde una protepina se encuentra con si misma.

La disparidad de valores no permite una correcta relación, por lo que se llevó a cabo una normalización obteniendo una disimilitud.

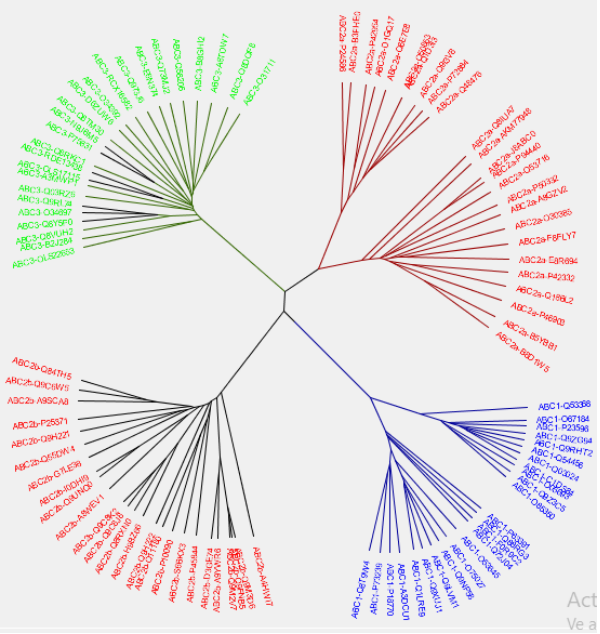
$$d_{i,j} = 1 - \frac{b_{i,j}}{max(b_{x,y})}$$

Donde b es el bitscore entre dos proteínas.

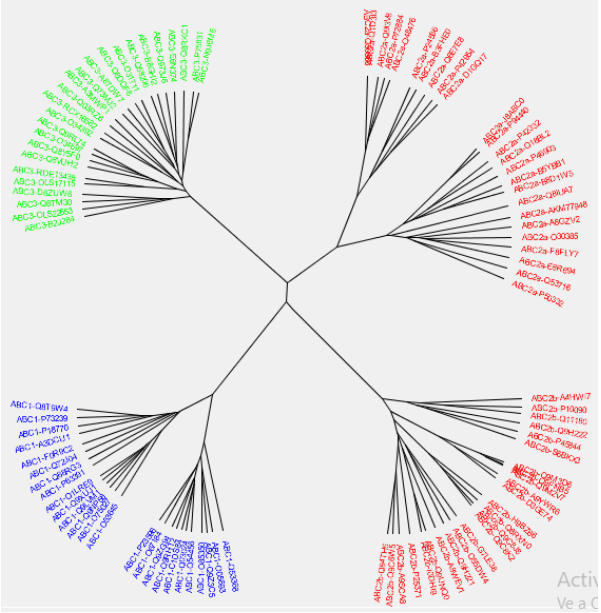
Los métodos fueron llamados y ejecutados desde un script en R, donde se guardaron los árboles de salida en archivos.

Todos los archivos mencionados se encuntran disponibles en https://github.com/eduardo-arrieta/ClusteringProject

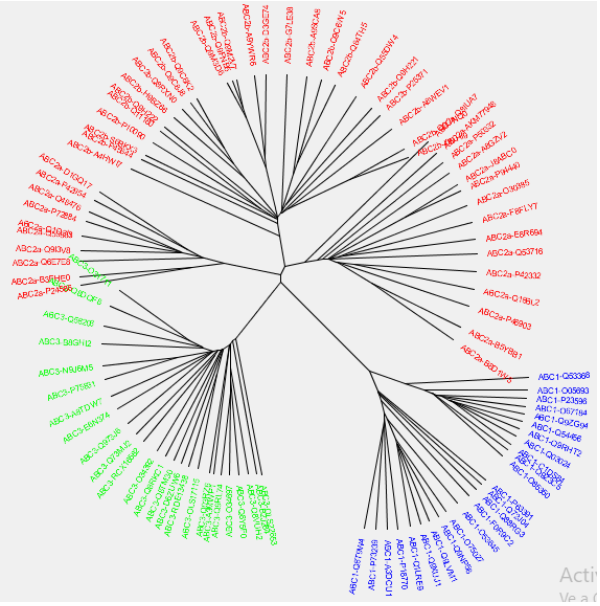
Posteriormente se visualizaron con el programa `figtree`.



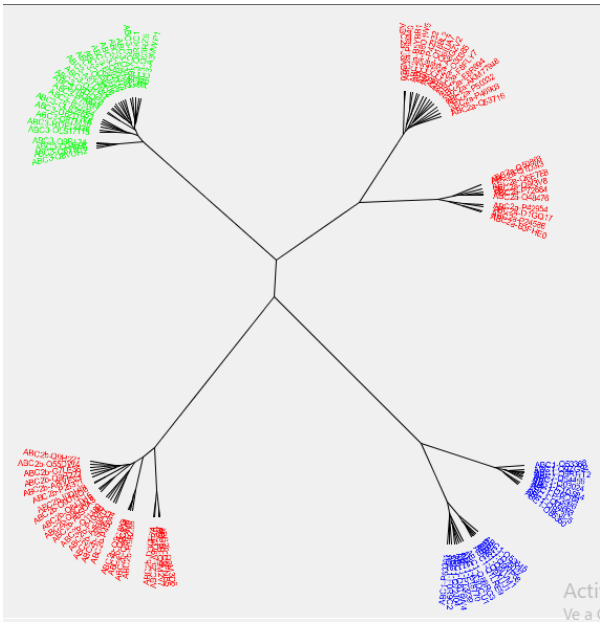
Average, agglomerative coefficient = 0.68



Complete, agglomerative coefficient = 0.75



Single, agglomerative coefficient = 0.58



ward.D2, agglomerative coefficient = 0.93

La simbología de colores es la misma para los cuatro árboles, de los azules son para las ABC1, los rojos ABC2 y los verdes para los ABC3.

Satisfactoriamente vemos que las etiquetas si guardan relación haciendo que los colores se agrupen dentro de una misma rama principal (el coloreo se hizo mediante búsqueda del término ABC#), tres de cuatro métodos (Average, complete y ward) podemos apreciar ciertos

patrones. Los ABC1 (azules) se diversifican en dos subramas, los ABC2 (rojos) se dividen en dos ramas principales y una de estas se divide en dos subramas, mientras los ABC3 (verdes) se aglomeran juntos en una sola rama. Sin embargo, el Single, presenta 3 ramas de ABC2 cercanas, contrastando con los resultados de los demás método, su agglomerative coefficient fue el menor de todos indicando que se encontró una mayor separación de los datos. Mientras, el árbol con ward tuvo el mayor agglomerative coefficient con un valor de **0.93**, claramente se puede ver lo densas que están las ramas. Si nos guiamos con el etiquetado taxonómico de las proteínas, los árboles Complete y Single son los que presentan mayor fidelidad a la cercanía de colores.

Considero que la dispersión del árbol Single nos brinda más información sobre las relaciones entre proteínas, mientras el árbol average no presenta grandes divisiones entre las proteínas, por lo que lo coloco como el meno informativo. El árbol de ward por su alta aglomeración, no permite ver que hay distinciones entre las mismas subramas, sin embargo, no es congruente con el etiquetado.

Como se dijo en clase, una mejor precisión filogenética se logra a partir de un análisis directo de las secuencias, no obstante, este proceso ha demostrado obtener resultados muy similares permitiéndonos hacer exploraciones muy rápidas.

Referencias

REACTOME. (2009). ABC-family proteins mediated transport. https://reactome.org/content/detail/R-HSA-382556