Metodología eficiente para obtener cliques de proteínas mediante los mejores aciertos bidireccionales+

Eunice Esther Ponce de León Sentí
Javier Eduardo Reyes Gallegos
Luis Daniel Cuellar Garrido
Eduardo Mauricio Martin Álvarez Tostado
Elva Díaz Díaz
Aurora Torres Soto
María Dolores Torres Soto
Juan José Martínez Guerra

Introducción

El agrupamiento de entidades mediante algún criterio adecuado con métodos de aprendizaje no supervisado ha sido uno de los mecanismos metodológicos utilizados para encontrar similitudes entre esas entidades y poder caracterizar mediante el hallazgo de agrupamientos las características de las entidades que quedan en el mismo grupo respecto a las características de las entidades que quedan en grupos diferentes. La realización de agrupamientos de proteínas en familias ha sido uno de los objetivos más importantes en la Biología computacional desde que establece el primer paso para ganar en información cuando se estudia una o varias proteínas desconocidas, pero se tiene bas-

tante información de otras. Existen muchos grupos de proteínas que han sido bien estudiados (NCBI CDD), por lo que al identificar una o varias proteínas desconocidas se les puede asociar un conjunto de propiedades funcionales, según el grupo al que se predice pertenecen.

En la literatura se pueden identificar tres tipos de criterios para clasificar las proteínas, uno es en función de las familias a las que pertenecen, otro criterio son los dominios que contienen y finalmente las características de secuencia que poseen. En la web se encuentran varias bases de datos donde se clasifican las proteínas y que utilizan estos criterios de forma individual o combinada. CDD es una de las bases de datos más reconocidas para anotar familias de proteínas (Marchler-Bauer *et al.*, 2015) y con la cual hacemos la validación de los resultados de la metodología que aquí se presenta.

La investigación que nos ocupa desarrolla una metodología eficiente para encontrar cliques de proteínas bajo la relación binaria que se puede definir entre pares de ellas usando el concepto de mejores aciertos bidireccionales (BBH) (Altschul *et al.*, 1999) y un mecanismo heurístico que utiliza la topología del árbol filogenético de los organismos en estudio. Como organismos de interés para el estudio de sus proteínas se tomaron 74 proteomas completos de hongos bajados de la página web NCBI.

Materiales y método

En un modelo basado en grafos para la búsqueda de todos los grupos de proteínas que se forman con las proteínas en estudio, los vértices son las proteínas, las aristas representan una relación binaria de la similitud entre ellas, y la búsqueda de grupos de proteínas relacionadas entre sí, puede plantearse como el problema de encontrar todos los cliques maximales del grafo así definido. Un clique de un grafo es un subgrafo donde cualquier par de vértices del subgrafo están unidos por una arista. Un clique maximal es aquel clique tal que no existe otro vértice del grafo que no pertenece al clique, que tenga aristas con todos los vértices del clique. Es decir, es el clique más grande que se pueda formar con esos vértices. Encontrar todos los cliques maximales de un grafo es un problema NP – completo (Garey y Johnson, 1979), es decir, de los más difíciles de resolver computacionalmente de manera exacta, y para ello es necesario utilizar un mecanismo heurístico (Pearl, 1984) que ayude a suavizar

la complejidad del problema, de tal forma que el tiempo de ejecución del algoritmo para obtener soluciones cercanas al óptimo no sea impráctico cuando crece el número de vértices del grafo. El óptimo en este problema es encontrar todos los cliques maximales que tienen el grafo antes definido. Encontrar una solución cercana al óptimo es encontrar un número suficientemente grande de cliques maximales del grafo.

El concepto fundamental del cual partimos en esta metodología es el concepto de Mejor Acierto Bidireccional (MAB), este consiste en un par de proteínas (X, Y) de organismos diferentes que tienen una similaridad reconocida entre ellas (se usa el puntaje [Gertz, 2005] de similaridad BLASTp [NCBI BLAST]), y no existe ninguna proteína del organismo donde está la proteína Y, que tenga una mayor similaridad con X que la similaridad que Y tiene con X. Y viceversa: no existe otra proteína del organismo de donde es X que sea más similar a Y que lo que X fue a Y. En este sentido, es la bidireccionalidad del concepto de MAB que lo hace una relación binaria simétrica por definición.

El concepto de MAB entre un par de proteínas nos permite establecer la relación binaria para las aristas del grafo formado por vértices que representan las proteínas.

Resultados y discusión

Para abordar el problema de encontrar todos los cliques maximales de un grafo utilizando un mecanismo heurístico que se basa en la topología del árbol filogenético de los organismos en estudio, se ha enriquecido el método inicialmente reportado en Ponce de León *et al.* (2017) y se añadió el paso 7 para tratar la búsqueda de cliques de proteínas interespecies. Los pasos de la metodología son como sigue:

1. Una vez definidos los organismos de interés, se bajan sus proteomas de la base de datos del NCBI. Supongamos que tenemos N organismos de interés.
2. Se ejecutan comandos del BLASTP para todo par de organismos en estudio, intercambiando organismo "query" con organismo "Data

Base" (NCBI BLAST). En este punto no se le pone restricciones al puntaje de similaridad del BLASTp.

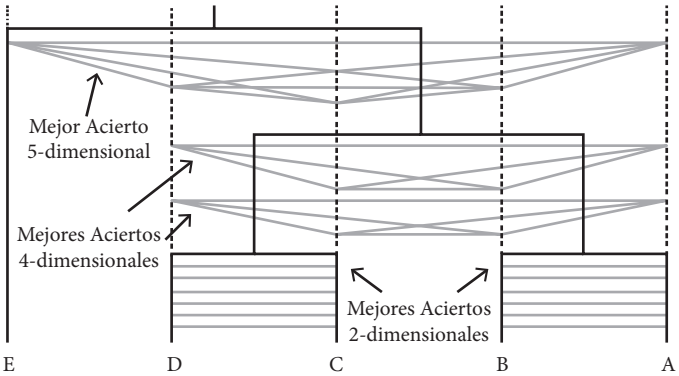
3. Del conjunto de archivos resultantes del BLAST en el paso 2, en total N(N-1) archivos, se buscan aquellos pares de proteínas de organismos diferentes que cumplan con el concepto de ser MAB. La cantidad de archivos resultantes de este paso es N(N-1)/2 y corresponde a los MAB que hay entre todo par de organismos. Cada archivo contiene los pares de proteínas que resultaron ser MAB entre un par de organismos.
4. Elegir un puntaje de similaridad entre proteínas para filtrar los N(N-1)/2 archivos conteniendo sólo aquellos MAB que sí cumplieron con el puntaje. La cantidad de archivos resultante es la misma que en el paso 3, sólo que son diferentes. Los archivos resultantes del paso 3 se guardan por si se desea variar el puntaje de similaridad entre proteínas para otros experimentos con mayor o menor exigencia en la similaridad entre proteínas en los MAB.
5. Se procede al conteo de los MAB para todo par de organismos y se calcula una distancia como la definida en Ponce de León *et al.* (2017) o la definida en Snel *et al.* (1999). Con esta distancia se construye una matriz de distancias entre pares de organismos.
6. Con la matriz de distancias como entrada a un método de construcción de árboles filogenéticos basado en distancias se obtiene un árbol filogenético de organismos, el cual se guarda en un formato Newick. Las hojas de este árbol son los organismos de interés.
7. Una vez obtenido el árbol filogenético de organismos, éste se va a utilizar como mecanismo heurístico para la búsqueda de grupos de proteínas que estén dos a dos relacionadas por ser MAB. Se lee la cadena Newick que contiene codificado el árbol filogenético de organismos (hojas). Para los subárboles formados por dos hojas ya se tienen calculados los cliques de a dos proteínas en el paso 4 con el puntaje de similaridad elegido para el estudio. La cadena Newick del árbol filogenético se va recorriendo desde los subárboles más pequeños que son de dos hojas, luego los subárboles de tres hojas y así sucesivamente. A partir de tres hojas definimos el concepto de Mejor Acierto k dimensional que para el caso de tres hojas, k=3:

> Un conjunto de proteínas interespecies se dice que forma un Mejor Acierto K-dimensional (MAK) si todo par de proteínas del conjunto es un MAB.

Para conformar los cliques de a tres proteínas interespecies se toman los MAB correspondientes a los dos organismos (hojas) que se unen primero, y se busca si en el organismo que representa la tercera hoja hay alguna proteína que tiene MAB con las proteínas que forman MAB de los dos organismos antes mencionados. Las proteínas que sean MAB de las dos hojas, pero que no sean MAB con proteínas de la tercera hoja, no pasan a formar MAK, donde k=3. Los grupos de tres proteínas que sí son MAK pasan al nivel jerárquico superior según la topología del árbol filogenético. Para conformar cliques de a cuatro proteínas interespecies se buscan los MAB de los dos pares de organismos que se unen primero y luego se buscan MAK, donde k=4. Así se sigue el proceso de construcción de grupos de proteínas MAK donde k crece desde 2 y en la medida que se acerca a la raíz del árbol, tal como se aprecia en la Figura 1.

Para la implementación de la metodología propuesta se utilizaron 74 proteomas de hongos y se aplicaron los pasos del 1 al 6. Para el paso 7 se implementa un algoritmo recursivo aprovechando la estructura jerárquica del árbol filogenético (Reyes, 2019). Se procesa la cadena Newick para dividir la búsqueda de los cliques, por el subárbol izquierdo y por el subárbol derecho. La naturaleza recursiva de la estructura de datos "árbol" permite la utilización de una función recursiva para recorrer todos los árboles anidados y divididos en subárboles izquierdo y derecho del árbol hasta llegar a las hojas. Esto permite la construcción de los cliques o grupos de proteínas, tal como se explica en el punto 7 de la metodología.

Figura 1. Cliques de proteínas obtenidos utilizando la topología del árbol filogenético de los organismos. Los cliques más grandes contienen proteínas de mayor número de organismos y podría decirse que son más ubicuas y los cliques más pequeños contienen proteínas de un número menor de organismos.

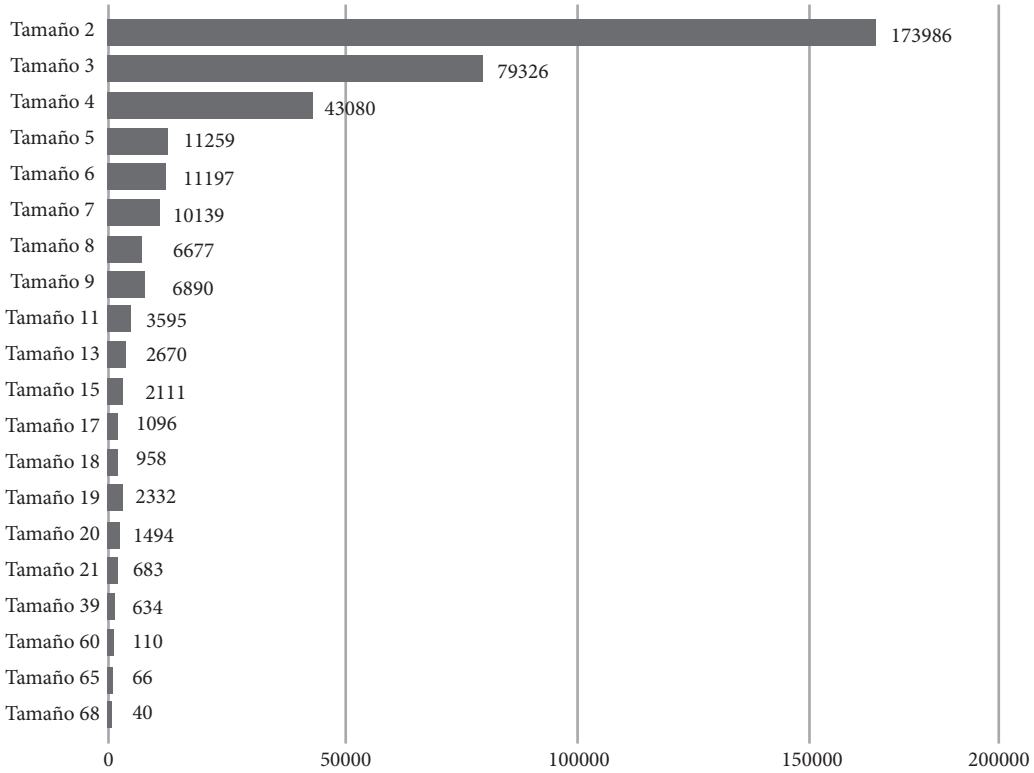


Como resultado de la aplicación completa de la metodología se obtuvieron 73 archivos de cliques de proteínas de los 74 hongos. Cada fila de un archivo contiene un clique de proteínas interespecies de tamaño k, donde todas las proteínas dos a dos han sido MAB o, dicho de otra forma, ese clique es un MAK. Se observa como tendencia que, en la medida en que se buscan cliques de proteínas en los niveles más altos del árbol filogenético, existen menos cliques de proteínas que cumplen con ser un MAK, tal como se observa en la Figura 2.

Para validar los resultados de la metodología se tomaron los 40 cliques de proteínas de hongos de tamaño 68, que son los cliques de proteínas más grandes que se obtuvieron. Se ingresaron a la base de datos CDD (NCBI CDD) para verificar que las proteínas de cada clique pertenecieran a la misma familia según los motivos y dominios registrados en esa base de datos. Se utilizó la herramienta de CDD, "Batch CD-search" para buscar aciertos de dominios y con ella obtener únicamente superfamilias. La salida de "Batch CD-search" es un archivo con los aciertos que tuvo cada una de las proteínas del clique introducido y contiene palabras como "Query", que es la proteína para la cual se buscan aciertos de dominios, "Hit type" es el tipo de acierto que se encontró,

en este caso se buscan superfamilias, “From” y “To” indican de dónde a dónde se encuentra el acierto en la proteína, “E-value” es la medida para indicar qué tan significativo es el acierto, cuanto más bajo sea este valor más confiable es. Finalmente, “Short name” es el nombre de la superfamilia con la que se tuvo el acierto.

Figura 2. Conteo de cuántos cliques de proteínas interespecies hay para cada tamaño (número de proteínas interespecies en el clique).



Para 39 cliques CDD encontró que las 68 proteínas dieron acierto con una superfamilia que se encuentra en la base de datos. Sólo en un clique sucedió que el mayor valor de acierto fue de 20 proteínas para una superfamilia de la

base de datos, teniendo de esta forma una clasificación correcta de los cliques obtenidos al aplicar la metodología de 97%. Un extracto de la clasificación de los cliques se puede ver en la Tabla 1.

Tabla 1. Extracto de superfamilias con las que tuvo aciertos las proteínas de los cliques encontrados por la metodología propuesta.

No. de Clique	Superfamilia	No. Aciertos
1	Pyrophosphatase superfamily	68
1	Atrophin-1	1
2	WD40	68
2	Abhydrolase	1
2	Pex14_N	1
3	P-loop_NTP	68
3	SIMIBI	68
3	ArsA	68
3	PRK02534	3
3	PRK06225	1
4	Brix	68
4	Rilp-LIKE	1
4	TOLa_FULL	1

Conclusiones

Se propone y comprueba la efectividad de la metodología aquí reportada que utiliza la información filogenética de los organismos en estudio como mecanismo de reducción de la dimensión de la complejidad del problema combinatorio subyacente. El método es determinístico y corre en tiempo polinomial porque la topología del árbol filogenético dicta los cliques que se

deben construir. Obtiene como resultados cliques de proteínas con un alto porcentaje de coincidencias con las familias reportadas en las bases de datos de CDD para dominios conservados, el cual es un recurso para la anotación de unidades funcionales de proteínas y se utiliza para conocer el porcentaje de buena clasificación de la metodología propuesta.

Referencias

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403-410.

Garey, M.R., Johnson, D.S. (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness (Series of Books in the Mathematical Sciences)*. W. H. Freeman.

Gertz, E.M. (2005). BLAST Scoring Parameters. Recuperado de at:ftp://ftp.ncbi.nlm.nih.gov/blast/documents/developer/scoring.pdf

Marchler-Bauer, A., Derbyshire, M.K., Gonzales, N.R., Lu, S., Chitsaz, F., Geer, L. Y., Geer, R.C., He, J., Gwadz, M., Hurwitz, D.I., Lanczycki, C.J., Lu, F., Marchler, G.H., Song, J.S., Thanki, N., Wang, Z., Yamashita, R.A., Zhang, D., Zheng, C., & Bryant, S.H. (2015). CDD: NCBI's conserved domain database. *Nucleic Acids Research*, *43*, D222-D226.

NCBI BLAST. *Basic Local Alignment Search Tool*. Recuperado de https://blast.ncbi.nlm.nih.gov/Blast.cgi

NCBI. *CDD Conserved Domain Database*. Recuperado de https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml

NCBI National Center for Biotechnology Information. Búsqueda de organismos de interés. Recuperado de https://www.ncbi.nlm.nih.gov/genome/browse#!/

Pearl, J. (1984). *Heuristics: intelligent search strategies for computer problem solving*.

Ponce-de-León-Sentí, E., Díaz, E., Guardado-Muro, H., Cuellar-Garrido, D., Martínez-Guerra, J.J.,Torres-Soto, A., Torres-Soto, M.D., Hernandez-Aguirre, A. (2017). A distance measure for building phylogenetic trees: a first approach. *Research in Computing Science*, *139*, 149-162.

Reyes Gallegos, J.E. (2019). *Algoritmo eficiente para la agrupación de proteínas en familias basado en mejores aciertos bidireccionales y el árbol filogenético*. Universidad Autónoma de Aguascalientes, Aguascalientes, México.

Snel, B., Bork, P., Huynen, M.A. (1999). Genome phylogeny based on gene content. *Nat Genet*, *21*(1), 108-110.