Implementación de un algoritmo rápido y preciso para determinar grupos ortólogos.

Daniel Tello, Jorge Gomez, Juan Camilo Zuluaga-Monares, Laura Gonzalez, Ricardo Angel, Nicolas Cardozo, Camilo Escobar, Mario Linares-Vasquez, Jorge Duitama

Departamento de Ingeniería de Sistemas y Computación, Universidad de los Andes, Bogotá, Colombia.

Los ortólogos y parálogos son genes que comparten un mismo origen evolutivo y que han divergido debido a procesos de especiación y duplicación respectivamente. La identificación de genes homólogos (ortólogos y parálogos) es la base para llevar a cabo análisis de genómica comparativa, incluyendo la construcción de árboles filogenéticos, alineamiento de genomas, construcción de pangenomas, entre otros. Dada su importancia biológica, se han desarrollado múltiples herramientas para la identificación de relaciones de ortología. Sin embargo, la mayoría de estas herramientas tienen problemas importantes de confiabilidad, eficiencia y usabilidad.

En este trabajo presentamos una nueva funcionalidad para identificación de ortólogos en NGSEP (Next Generation Sequencing Experience Platform). NGSEP recibe como entrada un conjunto de genomas anotados o de genes (nucleótidos o aminoácidos), y calcula de manera eficiente el porcentaje de k-mers compartidos entre todos los posibles pares de secuencias para identificar posibles relaciones de homología. Dos proteínas son consideradas homólogas si comparten un porcentaje de k-mers determinado. Suponiendo que los ortólogos y parálogos comparten un ancestro común, las relaciones iniciales se utilizan para derivar una partición de genes en grupos ortólogos. Para esto, NGSEP ejecuta un algoritmo de identificación de componentes conectados en el grafo inducido por las relaciones identificadas en el paso anterior, luego del cual se ejecuta un proceso de agrupamiento de Markov. Los grupos ortólogos sirven como entrada para identificar bloques de sintenia entre los genomas analizados, y calcular frecuencias de genes pertenecientes a los ortogrupos para construir una matriz de presencia/ausencia que sirve como primera versión del pangenoma de las especies analizadas.

Para evaluar el efecto de los parámetros de este algoritmo en la identificación de ortogrupos y comparar con otras herramientas, se utilizaron 70 grupos ortólogos curados, obtenidos de la base de datos OrthoBench. Se comparó el desempeño del NGSEP con el de Orthofinder y SonicParanoid. El mejor F-score se obtuvo con Orthofinder1 basado en BLAST (92.7), seguido por Orthofinder1 basado en Diamond (92.4). NGSEP obtuvo el tercer lugar (91.5), superando a SonicParanoid (86). NGSEP está a menos de 1% del mejor resultado y su tiempo de ejecución es al menos 10 veces más rápido que Orthofinder, y SonicParanoid. Realizando el mismo análisis con el subconjunto de secuencias correspondiente a mamíferos, NGSEP obtuvo el mejor desempeño, con un F-score del 94.3, superando en casi 10 puntos al segundo mejor resultado obtenido con Orthofinder basado en Diamond (F-score 85). Con respecto a la funcionalidad de alineamiento de genomas, NGSEP permite alinear genomas como los de referencia para humanos y chimpancés en unos pocos minutos y con menos de 16Gb de RAM. Post otra parte, se logró identificar el genoma “core” de e-coli, a partir del alineamiento de 100 genomas descargados de NCBI.

Esperamos que este desarrollo sea muy útil para grupos de investigación en biología evolutiva en una gran cantidad de especies, para llevar a cabo comparación de genomas intra e interespecies de manera eficiente y precisa.