CLASIFICACION AUTOMATICA DE TRANSPOSASAS

JAVIER IGNACIO ROMERO MARTINEZ
INGENIERO EN BIOINFORMÁTICA

RESUMEN

Las secuencias de inserción (IS) son los elementos genéticos móviles más simples que se conocen, de hasta 2 Kpb de extensión. Se caracterizan por tener en sus extremos secuencias repetidas invertidas involucradas en la transposición. Una IS contiene un gen que codifica para una Transposasa, proteína que ocupa la mayor extensión de la IS y es la principal responsable de la transposición. En 1998, Mahillon y Chandler clasificaron manualmente todas las IS que pudieron reunir en la época. Esta clasificación fue revisada en el 2002, que cuenta con cerca de 1000 elementos agrupados en 19 familias. Esta clasificación está públicamente disponible en la base de datos ISFinder. No obstante, este número de secuencias contrasta con el creciente número de Transposasas almacenadas en NCBI, cuyo total suma sobre 120.000 Transposasas no clasificadas en la base de datos no redundante (nr) de proteínas. Las bases de dato de proteínas, incluyendo las Transposasas aún están en etapa exponencial de crecimiento, doblando a cada 2 años. Esto deja en evidencia la necesidad de métodos de clasificación automáticos que permitan la identificación de la función de proteína a partir de su secuencia. El presente trabajo apunta a encontrar un método de clasificación automático que mejor reproduzca la clasificación manual realizada por Mahillon y Chandler. Proceso realizado en 4 importantes secciones: (1)

Preprocesar datos, a partir de archivos fasta de secuencias de Trasnposasas. Se crean matrices de distancias para ser utilizadas como inputs por distintos algoritmos de *clustering*. (2) Desarrollar distintos *clusterings* por medio de los algoritmos: Blastclust, CD-HIT, K-means, MCL, SCPS y UPGMA. (3) Comparar *clusterings* de resultados por medio de Variación de la Información (VI). VI es un criterio de comparación que mide la cantidad de información perdida o ganada al comparar 2 *clusterings*. (4) Identificar el algoritmo que mejor reproduzca la clasificación manual y clasificar todas las proteínas anotadas como Transposasas en nr. A través de VI se identificó SCPS, como el algoritmo que mejor reproduce la clasificación manual. Permitiendo clasificar las Transposasas contenidas en nr, proceso que permite identificar características

propias de cada grupo a las nuevas secuencias clasificadas, por ejemplo: patrones de secuencia y propiedades estructurales. Además, contribuye a mejorar la anotación de genomas y a entender los mecanismos de transposición de estas proteínas y de los elementos móviles en general.

ABSTRACT

The Insertion Sequences (IS) are the simplest mobile genetic elements known. They happen mainly in prokaryotes and are about 2 Kbp in length. The ends of an IS has inverted repeated sequences that are involved in the transposition process. An IS encodes for a protein called Transposase, which takes most of the length of the IS and is the main responsible protein for the transposition. In 1998, Mahillon and Chandler classified all the ISs known in that time. That classification was reviewed in 2002 and counts with around 1000 ISs or Transposases grouped in 19 families. This classification is publicly available in the ISFinder database. This number contrasts with the growing number of Transposases sheltered by NCBI. The total sums over 120.000 unclassified Transposases in the non-redundant Protein database. The protein databases, including Transposases, are still in a exponential growth phase, doubling every two years. This leaves in evidence the need for automatic methods of classification and identification of the protein function from the sequence. This work aims to find an automatic method that best reproduces the classification criteria of Mahillon and Chandler for classification of Transposases. Process carried out in 4 major sections: (1) preprocess of data from fasta files of Trasnposasas sequences. Distance matrices are created to be used as inputs by different clustering algorithms. (2) Develop of different clusterings through algorithms: BLASTCLUST, CD-HIT, K-means, MCL, SCPS and UPGMA. (3) Compare of clusterings of results through Variation of Information (VI). (4) Identify the algorithm that best reproduce the manual sorting and classifying all proteins annotated as transposases in nr. VI identifies SCPS like the algorithm that best reproduces the manual classification of Transposases in nr. Through this process it is possible to identify characteristics of each group to the new classified sequences. For example, sequence patterns and structural properties. On the other hand it improves the genomes annotation and it helps to understand the

transposition mechanisms of proteins and mobile elements in general.