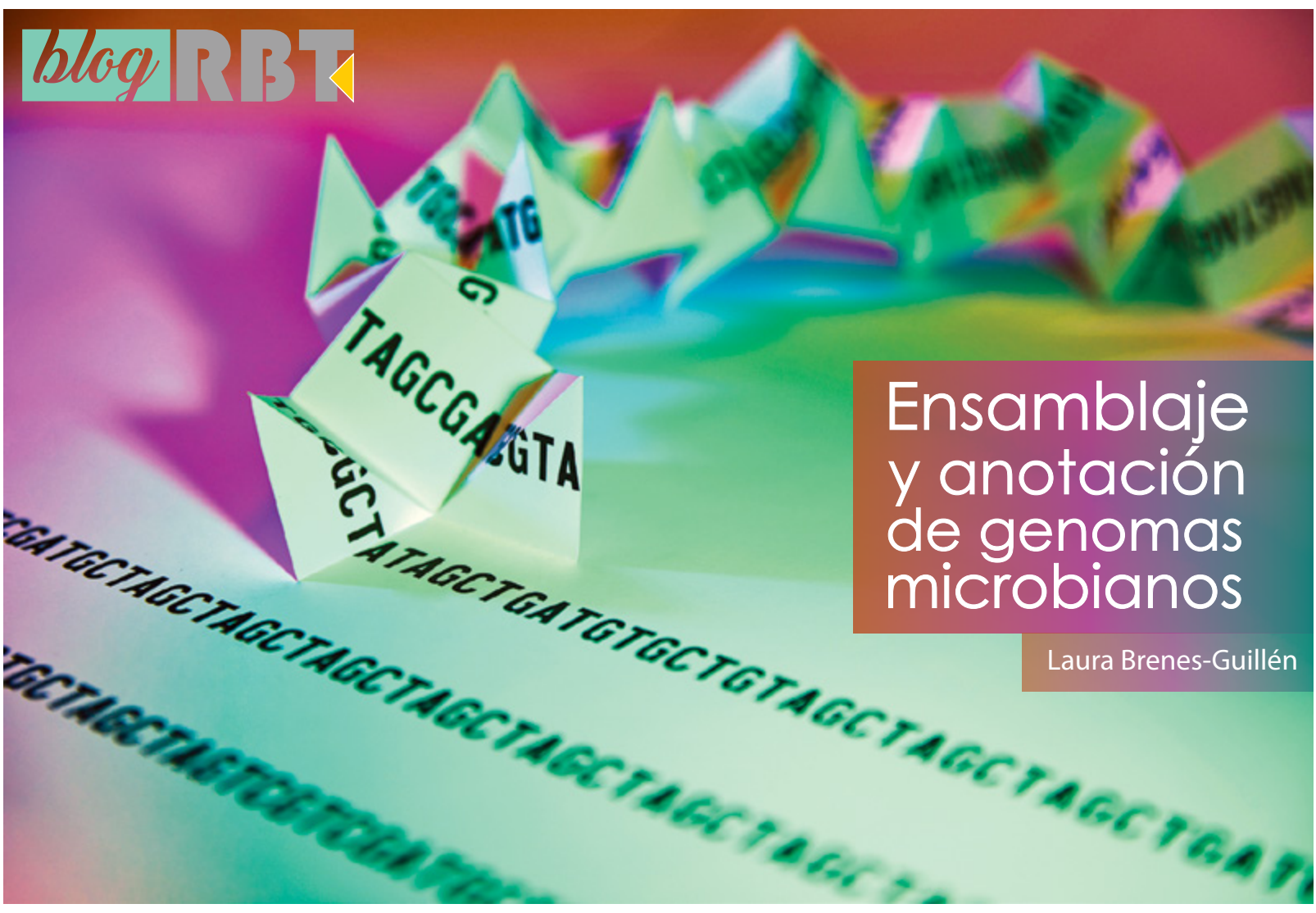
blog RBT

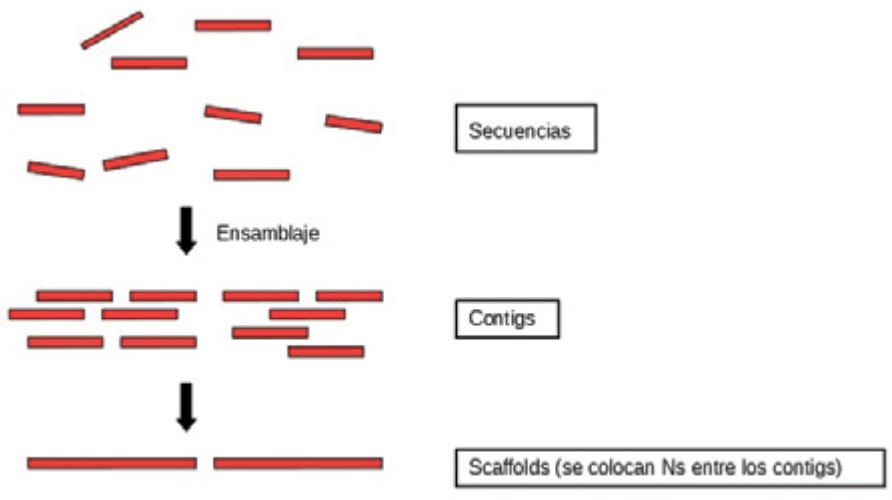


Ensamblaje y anotación de genomas microbianos

Laura Brenes-Guillén

La secuenciación masiva del genoma completo de un microorganismo es una herramienta importante para la identificación y caracterización de organismos conocidos o nuevos registros. Esta herramienta en conjunto con los análisis bioinformáticos permiten la comparación entre diversos individuos ya no solo a nivel de un grupo de genes sino de todo el genoma (al menos todo lo que pueda ser ensamblado), además de la identificación y anotación de genes, lo que permite caracterizar metabólicamente el genoma de un individuo en estudio.

Una vez que se ha **secuenciado el genoma**, ya sea utilizando la plataforma Illumina (secuencias cortas) o la plataforma PacBio (secuencias largas), estas secuencias se deben ensamblar para con ello formar secuencias más largas llamadas "*contigs*", los contigs a su vez se pueden unir y formar "*scaffolds*". Aparte de estas dos plataformas, **existen algunas otras** que también se pueden utilizar.



La comparación de varios algoritmos y ensambladores es de gran importancia para escoger la herramienta que ofrezca los resultados de acuerdo a la pregunta de investigación que se tenga, y que posea los valores de las estadísticas que el investigador anda buscando. Muchas de las diferencias que se pueden obtener entre los ensamblajes puede ser inherente al genoma y a la calidad de los datos[1].

Existen básicamente dos métodos para **ensamblar un genoma**, ya sea utilizando un genoma de referencia o haciendo un ensamblaje *de novo*, por otro lado, también se pueden utilizar ensamblajes mixtos con secuencias cortas y secuencias largas, lo que permite corroborar la información entre ambos tipos de secuenciación. Una vez que se ha obtenido el ensamblaje del genoma se puede hacer la anotación del mismo.

La anotación del ADN de un genoma consiste en la identificación estructural de genes, regiones codificantes y motivos, así como la identificación funcional de esas regiones; consiste en el proceso de identificación y etiquetado de todas las características relevantes en una secuencia del genoma[2]. Existen algunos algoritmos como PRODIGAL[3], los cuales realizan el reconocimiento de genes e identifican el sitio de inicio de la traducción. Existen además otros algoritmos y herramientas como Infernal, RNAmmer, Aragorn y SignalP.

Algunos de los genes previamente identificados no se van a encontrar en las bases de datos, por lo que no se les va a poder asignar una función —estos podrían ser artefactos, genes no descritos o genes nuevos—

Una vez que se han identificado las regiones de interés, se les debe asignar una función, para ello se utilizan herramientas de búsqueda de secuencias homólogas. Básicamente estos buscadores (Blast, HHMER, Diamond) comparan las secuencias del genoma contra bases de datos como **NCBI** (*National Center for Biotechnology Information*) y **COG** (*Cluster of Orthologous Groups*). Una vez que se han alineado e identificado estas secuencias se les asigna una función si es posible. Estas herramientas son útiles para hacer comparaciones metabólicas e identificar genes de interés.

Por otro lado, existen herramientas que utilizan las secuencias de proteínas para buscar enzimas y vías metabólicas en las cuales se encuentran involucradas, tal y como lo es **KEGG** (*Kyoto Encyclopedia of Genes and Genomes*), lo que resulta ventajoso si se quiere hacer análisis y comparaciones de vías metabólicas y genes involucrados. Además, existe desde 1988 el proyecto Ontología Génica (***Gene Ontology***), o GO por sus siglas en inglés, el cual posee un código de información controlado y común entre la comunidad científica que permite la identificación del gen y su función, ya sea molecular, celular o biológica; esta base de información se actualiza constantemente. Es importante saber que algunos de los genes previamente identificados no se van a encontrar en las bases de datos, por lo que no se les va a poder asignar una función —estos podrían ser artefactos, genes no descritos o genes nuevos—.

La secuenciación, ensamblaje y anotación de genomas es un proceso que requiere de la modificación de parámetros hasta lograr los mejores resultados según el criterio del investigador, existen muchas herramientas y tutoriales en línea, sin embargo, estos deben de adaptarse a los datos y considerar los múltiples parámetros que tienen estas herramientas, ya que la modificación de los mismos podría generar efectos en los resultados.

Laura Brenes-Guillén
Centro de Investigación en Biología Celular y Molecular (CIBCM), Universidad de Costa Rica
San José, Costa Rica

Imágenes
Ilustración representativa del genoma. Fotografía de **Darryl Leja y Ernesto Del Aguila III**, *National Human Genome Research Institute* (**CC BY-NC 2.0**)
Representación de secuencias, contigs y scaffolds. Diagrama preparado por Laura Brenes-Guillén

Referencias

[1]Salzberg, S. L., et al. (2012). GAGE: A critical evaluation of genome assemblies and assembly algorithms. ***Genome Research*, 22(3), 557–567**.
[2]Richardson, E. J., & Watson, M. (2013). The automatic annotation of bacterial genomes. ***Briefings in Bioinformatics*, 14(1), 1–12**.
[3]Hyatt, D., et al. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. **BMC Bioinformatics, 11(1), 119**.