Breves en Ciencia y Tecnología

Herramientas bioinformáticas para el análisis de secuencias en el Instituto Nacional de Higiene “Rafael Rangel”

Bioinformatic tools for sequence analysis at Instituto Nacional de Higiene “Rafael Rangel”

Carmen D González L[1], Carla P Castillo D[1], Giovanny Angiolillo R[1], David J Fernández S[1], Ascanio Rojas A[2].

[1]División de Biotecnología y Desarrollo, Gerencia Sectorial de Producción, Instituto Nacional de Higiene “Rafael Rangel”. Caracas, Venezuela. Teléfono: +582122191715. Correo: carmen.gonzalez@inhrr.gob.ve

[2]Centro Nacional de Cálculo Científico de la Universidad de Los Andes (CeCalCULA). Mérida, Venezuela.

Avances en las tecnologías de secuenciación de ácidos nucleicos ha incrementado sustancialmente la cantidad de datos que se generan del estudio de genomas completos o por regiones, datos que deberán ser analizados haciendo usos de herramientas computacionales enfocadas en ciencias biológicas, revalorizando el alcance de la bioinformática. El Instituto Nacional de Salud de Estados Unidos (NIH, por sus siglas en inglés) define a la bioinformática como la investigación, desarrollo o aplicación de herramientas computacionales con el fin de difundir el uso de datos biológicos, médicos, conductuales o de salud, incluyendo aquellas herramientas utilizadas para adquirir, almacenar, organizar, archivar, analizar y visualizar dichos datos(1). Por lo tanto, la bioinformática es esencial para la gestión de datos en la biología moderna.

Uno de los primeros acercamientos a la bioinformática surge en los años 60 con Margaret Dayhoff, quien transformó las secuencias de datos dispersas en la literatura impresa en una colección computarizada. Dayhoff, junto con Richard Eck, publicaron en 1965 una compilación de secuencias de aminoácidos, con el nombre de Atlas of Protein Structure and Sequence, como resultado de recolectar, comparar y analizarlas computacionalmente, cuyo fin era producir conocimiento sobre la estructura, función y evolución de las proteínas (2). Atlas, también los llevó a pensar sobre la mejor forma de manipular secuencias usando computación, lo que conllevó a adoptar la anotación de una letra para aminoácidos y no el código usual de tres letras, con el objetivo de reducir espacio en la memoria del computador (2).

Las bases de datos como las conocemos hoy en día fueron creadas basándose en el modelo de Atlas, tal es el caso de Protein Data Bank (PDB) para estructura de moléculas biológicas y GenBank para secuencias de ADN y proteínas. Estas se han convertido en herramientas esenciales en investigación (2).

HERRAMIENTAS PARA ANÁLISIS DE SECUENCIAS

El análisis de secuencias incluye búsqueda en las bases de datos, alineamiento de secuencias, descubrimiento de motivos y patrones en proteínas, predicción de genes y regiones promotoras, regulación, reconstrucción de relaciones evolutivas, ensamblaje de genomas y su comparación (3). Es por ello que existen diversas bases de datos biológicas y herramientas gratuitas que permiten buscar, visualizar, editar y analizar secuencias de nucleótidos y aminoácidos.

Bases de Datos Biológicas

Una base de datos es un archivo computarizado usado para almacenar y organizar la información en registros estructurados, de manera que pueda ser recuperada fácilmente a través de diferentes criterios de búsqueda. Para recuperar la información de un registro, conocido también como *entry*, el usuario especifica una parte particular de la información, con lo cual se recupera el registro completo de la data (3).

Actualmente, existen numerosas bases de datos biológicas disponibles en la Web. Una de las primeras bases de datos creadas fue Protein Data Bank (PDB; http://www.rcsb.org/pdb), en el año 1971, donde se encuentra depositada la información de estructuras tridimensionales de moléculas biológicas, incluyendo proteínas y ácidos nucleicos, determinadas mediante cristalografía de rayos X o resonancia magnética nuclear. El entendimiento de la forma de estas moléculas permite deducir el rol de las mismas en enfermedades, así como pueden ser utilizadas para el desarrollo de fármacos (4).

Asimismo, en se crea GenBank en el LANL (Los Alamos, NM) (http://www.ncbi.nlm.nih.gov), siendo esta una base de datos pública de secuencias de nucleótidos, bibliografía de apoyo y anotación biológica y a partir de 1992 es mantenida por el National Center for Biotechnology Information (NCBI), el cual forma parte del NIH (5). GenBank es hoy en día una de las colecciones más completas de datos y anotaciones de secuencias nucleotídicas de una gran cantidad de organismos. El contenido incluye ADN, ARN mensajero (ARNm), ADN complementario (ADNc), marcadores de secuencia expresada (EST, por sus siglas en inglés) y datos crudos de secuencias realizadas masivamente (3). Estos son solo dos ejemplos de las numerosas bases de datos biológicas que existen actualmente.

Programas y servicios para alineamiento, edición y análisis de secuencias

El alineamiento de secuencias es uno de los pasos críticos al realizar análisis filogenético, debido a que a partir de ellos se realizan muchas inferencias biológicas,

llegando incluso a la inferencia evolutiva y filogenética. Al comparar secuencias en un alineamiento se pueden identificar patrones de identidad, regiones conservadas o variables. Por ejemplo, la variación entre secuencias puede reflejar los cambios ocurridos por evolución, en la forma de sustituciones, inserciones o deleciones (3). Existe una gran variedad de modelos computaciones para inferir la comparación de secuencias, T-Coffee, Clustal Omega, MUSCLE, entre muchos otros, muchos de los cuales tienen interfaz Web o portales, que permite llevar a cabo el alineamiento remotamente, un ejemplo de ello es el European Bioinformatics Institute (EMBL-EBI; http://www.ebi.ac.uk/), que provee datos y servicios bioinformáticos de manera gratuita con el fin de contribuir en el entendimiento de los procesos biológicos (6).

Bioinformática en el Instituto Nacional de Higiene "Rafael Rangel"

En Venezuela, el Centro Nacional de Cálculo Científico de la Universidad de Los Andes (CeCalCULA) imparte talleres sobre bioinformática, siendo uno de ellos el Taller de Herramientas para Análisis de Secuencias (THAS) desde el año 1998 en la ciudad de Mérida y desde el 2014 se dicta en las instalaciones del Instituto Nacional de Higiene "Rafael Rangel" (INHRR), con el apoyo de la División de Biotecnología y Desarrollo adscrita a la Gerencia Sectorial de Producción, con el objetivo de actualizar conceptos en biología molecular y entrenar a los profesionales e investigadores que hacen vida en la Institución en el empleo de herramientas bioinformáticas para el tratamiento, organización, visualización, análisis e interpretación de secuencias de ácidos nucleicos y proteínas.

El único requisito para realizar el THAS es que los participantes tengan conocimientos básicos en genética, biología molecular y en el uso del computador como herramienta de trabajo. Actualmente, en el INHRR han participado más de 40 profesionales entre biólogos, bioanalistas y farmacéuticos, pertenecientes a las distintas áreas sustantivas de la institución.

El THAS se enfoca en la búsqueda de información y de secuencias en las bases de datos del NCBI; en los servicios para alineamiento múltiples de secuencias a través de MUSCLE y Clustal Omega de EMBL-EBI; en el uso de **BLAST** (Basic Local Alignment Search Tool) para encontrar regiones de similitud entre secuencias y en el manejo de algunos programas disponibles de forma gratuita como el **BioEdit/ Ugene** para la edición, manipulación, ensamblaje y análisis de secuencias de ácidos nucleicos y aminoácidos (7). Asimismo, en el THAS se enseña el diseño de oligos

(*primers)* para una secuencia de interés utilizando el programa **Oligo Explorer/ Primer-Blast**, mediante el cálculo de la temperatura de *melting* (Tm), determinación del contenido de GC, bucles y dímeros de oligos, entre otros. Por último, en el Taller se introduce al participante en el establecimiento de relaciones evolutivas y la construcción de árboles filogenéticos mediante el empleo del programa Molecular Evolutionary Genetics Analysis **(MEGA)** y conceptos básicos de genética de poblaciones con DNAsp.

APLICACIONES DE LA BIOINFORMÁTICA

La bioinformática tiene un papel central en muchas áreas de la investigación en biología, como en genómica, específicamente secuenciación de genomas, mapeo, anotación y comparación de genomas. Es esencial para proteómica, permitiendo el análisis de secuencias de proteínas con el fin de determinar motivos funcionales, para la determinación de estructura de proteínas, interacciones proteína-proteína, entre otras. Asimismo, permite el descubrimiento de marcadores moleculares, como polimorfismos de un solo nucleótido (SNP), así como forma parte de los estudios de evolución y filogenia (8).

Esta versatilidad de la bioinformática ha permitido que hoy en día sea usada para el diseño de nuevos medicamentos y análisis forenses. En el caso del diseño de nuevos medicamentos, los estudios de interacciones proteína-ligando proveen las bases para la identificación de nuevos sitios de acción para medicamentos sintéticos, asimismo, conocer las estructuras tridimensionales de proteínas permite el diseño de moléculas que puedan unirse a un receptor de una proteína blanco con alta especificidad y afinidad (3).

Por otra parte, la bioinformática es de vital importancia en la secuenciación de ADN ayudando a identificar la información de importancia biológica, de manera de tener un mejor entendimiento de los organismos. Por ejemplo, la bioinformática en el campo de la biotecnología de microorganismos se emplea de diferentes formas: analizando computacionalmente la data proveniente de experimentos, secuenciación de genomas, determinación de la función de genes, construcción de árboles filogenéticos, identificación de segmentos que codifican a proteínas, entre otras (9).

Cuenta también con aplicaciones a nivel médico y clínico, ayudando a determinar reacciones adversas de medicamentos en individuos, y podría ser usada en la medicina personalizada, donde se individualiza un tratamiento a partir de la información genética (10).

Es por ello que no solo es necesaria la data proveniente de los experimentos de genómica o proteómica, sino también personas formadas en esta área, capaces de interpretar dicha información. Este es el objetivo primordial de realizar los talleres bioinformática, como el THAS en el INHRR.

REFERENCIAS BIBLIOGRÁFICAS

1. Huerta M, Downing G, Haseltine F, Seto B, Liu Y. NIH working definition of bioinformatics and computational biology. 2000. Disponible en: https://www.bisti.nih.gov/docs/CompuBioDef.pdf. (Consultado 10 de febrero de 2015).
2. Strasser B. Collecting, comparing, and computing sequences: the making of Margaret O. Dayhoff's Atlas of Protein Sequence and Structure, 1954–1965. J Hist Biol. 2010; 43: 623–660.
3. Xiong J. Essential bioinformatics. 1ra ed. Cambridge: Cambridge University Press; 2006.
4. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acid Res. 2000; 28 (1): 235-242.
5. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, et al. GenBank. Nucleic Acid Res. 2015; 43: D30 - D35.
6. Brooksbank C, Bergman MT, Apweiler R, Birney E, Thornton J. The European Bioinformatics Institute's data. Nucleic Acid Res. 2014; 42: D18 - D25.
7. Hall T. BioEdit: a user friendly biological sequence alignment editor and analysis program for Windows 95/98/NT.Nucleic Acids Symp Ser (Oxf). 1999; 41: 95 - 98.
8. Goodman N. Biological data becomes computer literate: new advances. Curr Opin Biotechnol. 2002; 13: 68 - 71.
9. Bansal AK. Bioinformatics in microbial biotechnology – a mini review. Microb Cell Fact. 2005; 4:19.
10. Bayat A. Science, medicine, and the future: bioinformatics. BMJ. 2002; 324: 1018 - 1022.