Comité Editorial: González Andrade, M.; Hernández Alcántara, G.; Martínez González, J.J.; Meraz Cruz, N.; Ramírez Silva, L.H. y Vilchis Landeros, M.M.
© ISSN-0188-137X

Facultad de Medicina

Versión electrónica en http://bq.facmed.unam.mx/tab

MENSAJE BIOQUÍMICO

Mens. Bioquím. 46 (2022) 1-12

TAB

Memoria del XLIX Taller de Actualización Bioquímica, Facultad de Medicina; UNAM

Una introducción a la bioinformática: avances en la biología y ciencias de la salud.

An introduction to bioinformatics: advances in biology and health sciences.

Portillo Bobadilla, Tobías[1]*; Pérez Hernández, Bertha[2]; Pérez Hernández, Valentín[3] y Hernández Guzmán, Mario[4].

1. Red de Apoyo a la Investigación (RAI), Coordinación de la Investigación Científica, Universidad Nacional Autónoma de México - Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán.
2. Red de Biodiversidad y Sistemática, Instituto de Ecología A.C.
3. Instituto Tecnológico de Tuxtla Gutierrez.
4. Laboratorio de Ecología del Suelo, Cinvestav, Instituto Politécnico Nacional.

*Correspondencia: RAI, Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán Edificio de Radio-Oncología, 2o piso Vasco de Quiroga 15, Belisario Domínguez Sección XVI, Tlalpan, C.P. 14080, CDMX, México.
Tel. +52(55)54870900 ext. 6327, tobias@cic.unam.mx

Resumen

Se presenta una introducción sobre la bioinformática, sus orígenes y campo de aplicación mostrando las herramientas computacionales necesarias para el análisis de secuencias. Se revisa el estado del arte de los repositorios internacionales de datos biológicos, así como de las tecnologías de alto rendimiento, con su aplicación en la medicina y ciencias de la salud. Por ejemplo, se aborda el NCBI, PubMed, RefSeq, PDB, InterPro, entre otras bases de datos, algunas aplicaciones en la biomedicina y se presenta un protocolo para el análisis de datos de secuenciación masiva (microbiota usando el gen 16S rRNA); se detalla el flujo de trabajo, software y requerimientos mínimos de cómputo. Se menciona el alcance de la bioinformática en la interpretación de los resultados, formulación de hipótesis y en general su relevancia para la investigación en la biología y ciencias biomédicas.

Palabras claves: bioinformática, secuenciación masiva NGS, QIIME, bases de datos, biomedicina.

Abstract

An introduction to bioinformatics, its origins and field of application is presented, showing the computational tools for the analysis of sequences. The state of the art of international databases or repositories of biological data and high-performance technologies, with their application in medicine and health sciences, is reviewed. For example, the NCBI, PubMed, RefSeq, PDB, InterPro, and others. Some applications in biomedicine are addressed and a protocol for the analysis of NGS sequencing data (microbiota using the 16S rRNA gene) is presented, detailing the workflow, software, and minimum computing requirements. The scope of bioinformatics in the interpretation of results, formulation of hypotheses and in general its relevance for research in biology and biomedical sciences is mentioned.

Keywords: bioinformatics, next generation sequencing NGS, QIIME, databases, biomedicine.

Introducción

Definición y orígenes de la bioinformática

La bioinformática es una disciplina científica que desarrolla software, bases de datos, algoritmos y métodos computacionales, que son incorporados en sistemas, flujos de trabajo y diversas estrategias de investigación con el objeto de estudiar y comprender los sistemas biológicos (1). La bioinformática ha tenido un impacto relevante en las ciencias biológicas, en particular en la investigación biomédica o medicina de precisión, mejorando el diagnóstico o la práctica clínica (2, 3).

Los inicios de la bioinformática se dan a partir del estudio de las proteínas y de los ácidos nucleicos (4, 5) (ver Tabla 1). Primero se secuenciaron las proteínas y prácticamente dos décadas después el ácido desoxirribonucleico o ADN. En 1949 Pehr Edman introdujo el método para la secuenciación de péptidos utilizando la degradación con proteasas y un marcaje con isótopos radiactivos (6). Así, en 1953 Fred Sanger, Tuppy y Thomson lograron secuenciar la proteína de la insulina, que consta de una cadena de 20 y otra de 30 residuos de aminoácidos (7, 8). Pero se deben recordar tres momentos: a) los experimentos de Frederick Griffith sobre el principio transformador, b) los de Avery, MacLeod y McCarty en 1944 (9), c) el trabajo de Hershey y Chase en 1952 (10). En su conjunto estos trabajos nos permitieron entender que es justo en el ADN, y no en otras moléculas, en donde se almacena la herencia.

En 1953 Watson y Crick descifraron la estructura química del ADN a partir de las imágenes de cristalografía de rayos-X obtenidas por Rosalind Franklin (11, 12), lo que permitió saber la forma en que se almacena, replica y hereda la información genética. Con ello se postuló el dogma central de la biología molecular y se descubrió el código genético universal (Crick 1957, 1958, 1961, 1970). A partir de estas bases, de 1950 a 1970 se desarrolla tanto la biología molecular como la bioinformática. Margaret Dayhoff fue de las pioneras enfocando sus conocimientos de computación a la biología, bioquímica y medicina. Se sabe que se enfrentó a una visión de género desfavorable para la mujer, pues se les veía haciendo tareas secretariales o repetitivas (13). Dayhoff hizo grandes aportes sobre el estudio de las secuencias de proteínas, el cambio evolutivo a nivel molecular y elaboró los primeros programas de cómputo para analizar las secuencias. Los resultados experimentales de diversos grupos de investigación los integró en su Atlas of Protein Sequence and Structure. También, creó las matrices de sustitución (Dayhoff o PAM) utilizadas en los alineamientos de secuencias, así como el código de una letra de los aminoácidos en aras de simplificar el cómputo. El atlas, resultó ser la primera base de datos de secuencias, éste se almacenó en tarjetas perforadas que se ingresaban a la computadora y permitiendo ejecutar los programas que ella misma desarrolló (5, 14). Con el tiempo, el atlas pasó a cintas magnéticas y luego a su versión on line en los inicios de internet en 1978. Sin duda Dayhoff es considerada la madre de la bioinformática.

Campo de acción del bioinformático

A veces la bioinformática no es totalmente entendida (13, 15) y resulta complejo establecer una línea propia de investigación, con objetivos, tiempos y financiamiento adecuados. No pocas veces la bioinformática es vista sólo como una herramienta, servicio (16) u oficina de asesoramiento, lo que la deja totalmente supeditada a las preguntas o valoración de otros laboratorios o departamentos. Se debe considerar el hecho de que la bioinformática es una actividad de origen y naturaleza interdisciplinaria (17). El bagaje de un bioinformático es heterogéneo, con formación de biólogo, médico, informático, actuario, químico, físico, matemático u otro. Por otro lado, las estrategias y métodos empleados cambian al ritmo de los mismos avances científicos y tecnológicos. El bioinformático no tiene una técnica estandarizada, método o pipeline ya establecido que pueda aplicarse, sin una revisión crítica o modificaciones para una correcta interpretación de los resultados, pues el conjunto de datos y el contexto de la investigación es distinto en cada caso (18). Por otro lado, nace la ciencia de datos y queda por ver si esta vertiente absorberá o fortalecerá a la bioinformática, ampliando sus horizontes y complejidades (ciencia de datos, big data, medicina de precisión o bioinformática) (19, 20). En México se han identificado fortalezas y oportunidades pero hay retos importantes que hace falta resolver (21).

Habilidades y herramientas del bioinformático

Además de poseer conocimientos de biología, biología molecular y de biomedicina, se han descrito habilidades y recomendaciones para el bioinformático (22–24). Entre las herramientas más versátiles que posee está Linux, pues este sistema operativo facilita automatizar tareas repetitivas, explorar múltiples archivos, binarios o de texto simple, con miles de millones de secuencias, usando sólo la línea de comandos. Algunas distribuciones Linux son RedHat, Debian, Ubuntu, Fedora o Slackware.

Comité Editorial: González Andrade, M.; Hernández Alcántara, G.; Martínez González, J.J.; Meraz Cruz, N.; Ramírez Silva, L.H. y Vilchis Landeros, M.M.
Publicado por el Departamento de Bioquímica de la Facultad de Medicina; UNAM.

Tabla 1. Algunas aportaciones relevantes que dieron origen a la Bioinformática.

Año	Aportaciones en biología, genética y medicina	Autores
1928	Descubrimiento del *principio transformador* en los neumococos como causante de enfermedades.	Frederick Griffith
1944	Se descubrió que el *principio transformador* (hereditario) que permite conferir virulencia en los neumococos está compuesto de ADN y no de ARN o de proteína.	Oswald T. Avery, Colin M. MacLeod, and Maclyn McCarty
1949	Un método para determinar la secuencia de los residuos de aminoácidos en las proteínas (Degradación de Edman)	Edman Pehr
1952	El ADN y no la proteína es la molécula que almacena la información hereditaria. Experimentos con fagos T2 y bacterias para discriminar más allá de cualquier duda, usando marcaje radiactivo para el ADN (fósforo P-32) y las proteínas (azufre S-35).	Alfred Hershey y Martha Chase
1951 - 1953	Secuenciación de la primera proteína: la insulina. Primeras predicciones de la estructura de las proteínas (hélices y hojas)	Fred Sanger; Pauling and Corey
1953	Descubrimiento de la estructura del ADN	Watson y Crick
1958	La mioglobina es la primera estructura tridimensional de una proteína determinada por cristalografía de rayos-X	Kendrew JC, Bodo G, Dintzis HM, et al.
1955, 1961, 1965	Desciframiento del código genético.	Severo Ochoa, Marshall W. Nirenberg, Har Gobind Khorana
1963 - 1965	Comparación de secuencias de proteínas de la hemoglobina o "*paleogenética*". En 1965 se establece la hipótesis evolutiva del reloj molecular. En 1962 se determinó el tiempo de divergencia de especies basándose en la idea del reloj molecular y en el registro paleontológico para su calibración.	Linus Pauling, Emile Zuckerkandl
1965	Atlas de la Secuencia y Estructura de Proteínas	Dayhoff, M.O., Eck, Richard V., Chang, Marie A. y Sochard, Minnie R.
1967	Secuenciador de proteínas	Edman Pehr
1972	Aislamiento y amplificación de genes cortando con enzimas de restricción e insertando ADN en bacterias transformadas.	Jackson, Symons and Berg
1977	Origen de los métodos de secuenciación de ADN. Primero basado en la reacción de Maxam-Gilbert y posteriormente en polimerasas usando dinucleótidos modificados.	Maxam AM, Gilbert W. A; Sanger F, Nicklen S, Coulson
1987	Invención de la técnica de la PCR	Kary Mullis

Comité Editorial: González Andrade, M.; Hernández Alcántara, G.; Martínez González, J.J.; Meraz Cruz, N.; Ramírez Silva, L.H. y Vilchis Landeros, M.M.
Publicado por el Departamento de Bioquímica de la Facultad de Medicina; UNAM.

Tabla 1. Algunas aportaciones relevantes que dieron origen a la Bioinformática. Continuación.

Año	Aportaciones en computación y estadística	Referencia
1962	COMPROTEIN: un programa de cómputo para determinar la estructura primaria de una proteína, evaluado con éxito en una computadora IBM 7090	Dayhoff, Margaret O y Ledley Robert S.
1970	Se desarrolló el primer algoritmo de programación dinámica para la alineaciones pareada de secuencias de proteínas	Needleman SB y Wunsch CD
1974, 1975	Intel 8080 microprocesadores con circuitos integrados y el sistema Altair 8800.	Roberts E, Yates W
1977	Surgimiento de las primeras computadoras personales como la *Commodore* PET, Apple II y la Tandy TRS-80.	Commodore Business Machines, Inc.; Wozniak, S.
1978	Primer modelo de sustitución de aminoácidos, publicado en el Atlas de la Secuencia y Estructura de Proteínas, basado en 1572 mutaciones aceptadas (PAM) con árboles filogenéticos de 71 familias de proteínas con más del 85% de identidad.	Dayhoff, Schwartz and Orcutt
1979	El primer software dedicado a analizar la secuenciación de Sanger, incluyendo un alfabeto para codificar los caracteres inciertos en las secuencias.	Rodger Staden
1981	Método de máxima verosimilitud para inferir árboles filogenéticos a partir de secuencias de ADN. Filogenias moleculares.	Joseph Felsenstein
1984	La primera colección de software paquete CGC con 33 comandos en línea, dedicado al análisis de secuencias e implementado en la computadora DEC VAX-11.	Devereux J, Haeberli P, Smithies O.
1987	Primera aproximación al problema del alineamiento múltiple de secuencias. Se crea el lenguaje de programación PERL.	Da-Fei Feng and Russell F. Doolitle; Larry Wall
1988	CLUSTAL: paquete de cómputo para realizar alineamientos múltiples en una computadora personal.	Higgins DG, Sharp PM
1996	Métodos bayesianos para inferir las filogenias moleculares.	Rannala B, Yang Z.
1996, 1999	SWISS-PROT y el TrEMBL que incluye las traducciones de las secuencias codificantes o CDS provenientes del EMBL.	Hermjakob H, Fleischmann W, Apweiler Rolf; Bairoch, A.
2000	EMBOSS: the European molecular biology open software suite. Utilidades de software libre para el análisis.	Rice P, Longden I, Bleasby A.
2002	The Bioperl Toolkit: módulos en Perl para las ciencias de la vida, para el acceso, procesamiento y análisis de secuencias.	Stajich JE, Block D, Boulez K, et al.
2005	Secuenciación masiva, tecnología de pirosecuenciación 454.	Margulies M, Egholm M, Altman WE, et al.

Otras características importantes de Linux son sus sistemas de archivos, jerárquicos, con una raíz inicial, el uso de permisos de usuario y de un entorno o ambiente de trabajo llamado shell. Esta es la clásica pantalla (blanca o negra) en la que se escriben los comandos o instrucciones de texto. Bash, tcsh y zsh, son ejemplos del shell y se puede programar en todos ellos. Identificar el PATH de los programas, las rutas de trabajo, modificar un código o script, hacer una instalación e incluso compilar programas, son tareas rutinarias del bioinformático en Linux. Otra herramienta que se integra muy bien a las rutinas

Comité Editorial: González Andrade, M.; Hernández Alcántara, G.; Martínez González, J.J.; Meraz Cruz, N.; Ramírez Silva, L.H. y Vilchis Landeros, M.M.
Publicado por el Departamento de Bioquímica de la Facultad de Medicina; UNAM.

bioinformáticas es Github. Un administrador de versiones que permite llevar el control de las modificaciones que realizamos a nuestro código, así como visualizar, documentar y publicar (online) el trabajo que se realiza. También se utilizan otros lenguajes de programación. AWK por ejemplo, permite procesar textos manipulando columnas en textos (tablas). Perl, Python y Java son lenguajes de programación de alto nivel muy populares entre los bioinformáticos, o bien C o C++. En cuanto a bases de datos se dispone de MySQL o PostgreSQL, pero también existen otros paradigmas noSQL como MongoDB y Cassandra, entre otros.

Bioconductor, R y RStudio (una IDE para escribir y ejecutar código en R) son ampliamente recomendados. Los paquetes de R se encuentran en repositorios públicos, son implementados y distribuidos por la comunidad, lo que favorece la implementación de otros paquetes. En R podemos manipular y explorar datos, realizar pruebas estadísticas y presentar resultados mediante gráficos de calidad profesional. Además, se facilita la automatización y reproducibilidad de los análisis. Phyloseq, edgeR, DESeq, ShortRead, metagenomeSeq son algunos paquetes en R para análisis bioinformáticos.

El software Anaconda o miniConda y los contenedores Docker, Kubernetes o Mesos permiten crear ambientes de trabajo que facilitan la creación de código y flujos de trabajo. Anaconda, por ejemplo, está orientada a Python y R, es de distribución libre y cuenta con herramientas de bioinformática, ciencia de datos, inteligencia artificial, big data, análisis predictivo, cómputo científico, entre otras. Los contenedores como Docker permiten incluir versiones específicas de software (o el núcleo de otros sistemas operativos) para ejecutar un flujo de trabajo. El usuario destino no requiere instalar bibliotecas de código o un nuevo sistema operativo o distribución Linux. De esta forma, cuando un flujo de trabajo es complejo, se evitan errores técnicos en la compilación o ejecución de los programas que son difíciles de resolver para alguien que no es informático. Así, se construye el contenedor o dockerfile, se incluye la imagen del ambiente de trabajo y se puede ejecutar en cualquier otra computadora. Es como tener una mochila para ir a acampar con todo lo necesario, sin tener que improvisar o prepararla uno mismo.

Finalmente, también se dispone de servidores en la nube para el análisis de datos, lo que simplifica la infraestructura y los recursos requeridos en la administración de servidores. Las mismas empresas de secuenciación ofrecen servicios de análisis en sus plataformas web. Otras grandes compañías como Google (Cloud Life Sciences, antes Google Genomics) y Amazon (AWS) también están ofreciendo servicios genómicos o de big data.

Bases de datos biológicas y de información científica

Se dispone de grandes bases de datos sobre la literatura científica e información biológica. Por ejemplo, PubMed, la Biblioteca Pública del gobierno de los Estados Unidos de Norteamérica, comprende más de 33 millones de citas de literatura biomédica provenientes de MEDLINE, revistas y libros en línea. MEDLINE, el componente principal de PubMed, es una base de datos bibliográfica de la Biblioteca Nacional de Medicina (NLM) dependiente de los Institutos Nacionales de Salud (NIH) de los Estados Unidos de Norteamérica. En ella se tienen más de 28 millones de referencias de publicaciones que datan de 1966 a la fecha. Las áreas que abarca son ciencias de la vida, biomedicina y salud, incluyendo investigación básica, clínica, salud pública, política en salud, actividades educativas, y temas de biología, ciencias ambientales, biofísica y química. Además, incluye más de 5200 revistas, a través de criterios de selección definidos por un comité de selección. Estos registros son indexados con palabras de temas médicos (MeSH) y metadatos sobre el fondeo, la genética o química. PubMed Central (PMC) es un repositorio de texto completo de las publicaciones científicas en ciencias biológicas y biomedicina.

Por otro lado, se tienen las bases de datos biológicas primarias y derivadas. Las primarias resultan de la secuenciación de genes o genomas, de la determinación de la estructura de las proteínas y de experimentos de expresión. El GenBank del NCBI aloja las secuencias de genes y comparte diariamente la información obtenida por otras dos bases de datos de ácidos nucleicos, el DNA Data Bank of Japan (DDBJ) y el European Nucleotide Archive (ENA) del EMBL-EBI. Estas tres bases de datos conforman la Colaboración Internacional de Bases de Datos de secuencias de nucleótidos (INSDC) con el objetivo de facilitar su acceso y actualización a nivel global. En su sitio web se describe la Feature Table o definiciones que son reglas para la anotación de los genes. Por ejemplo, las palabras utilizadas para identificar un CDS o secuencia codificante, un origen de la replicación rep_origin, un sitio de unión a proteína protein_bind, o un RNA de transferencia tRNA. Estas palabras o keys son la forma de anotar el significado biológico que se esconde per se en las secuencias de nucleótidos. La estructura 3D de las

Comité Editorial: González Andrade, M.; Hernández Alcántara, G.; Martínez González, J.J.;
Meraz Cruz, N.; Ramírez Silva, L.H. y Vilchis Landeros, M.M.
Publicado por el Departamento de Bioquímica de la Facultad de Medicina; UNAM.

proteínas se encuentra en la base de datos (primaria) Protein Data Bank (PDB). Esta información es obtenida a partir de métodos experimentales por resonancia magnética nuclear o cristalografía de rayos X y contiene las coordenadas atómicas de estas moléculas. Se puede afirmar que para todo enfoque o estudio existe alguna base de datos (p. ej. algunas de uso frecuente son *GenBank, SNPs, WGS, PDB, RDP, Silva, ArrayExpress*).

Las bases de datos derivadas son las que utilizan las bases de datos primarias, agregan valor con nueva información y están elaboradas por terceros. Es a través de software, algoritmos computacionales y del trabajo de curadores que se construyen estas bases de datos. Una base de datos es curada cuando existe un equipo técnico y científico que la revisa cuidadosamente. Por ejemplo, el RefSeq incluye secuencias curadas que eliminan redundancia y son utilizadas como referencia para otros estudios. Otro ejemplo, son las bases de datos de familias y dominios de proteínas (*Conserved Domain o pfam*). Además, todas están relacionadas de alguna u otra forma, comparten y cruzan información a través del internet «en la nube». Otro ejemplo, el KEGG (Enciclopedia de Genes y Genomas de la Universidad de Kyoto en Japón) alberga las rutas metabólicas de los genomas.

Para la clasificación en familias o dominios y el análisis funcional de las proteínas es útil InterPro que se vincula a otras bases de datos para ofrecer una descripción más completa de las proteínas. En su sitio web podemos realizar búsquedas mediante secuencia, texto o palabras clave, o incluso a través de la arquitectura de dominios. La búsqueda incluye los resultados de otras bases de datos relacionadas tales como CATCH, Pfam, HAMAP, Panther, Prosite, que son bastante conocidas. Las bases de datos en general disponen de páginas o sitios web con herramientas para la consulta y el análisis. La revista *Nucleic Acid Research* dedica anualmente en enero un número especial con nuevas bases de datos y la revisión o actualización de las publicadas en números anteriores (25).

Tecnologías de alto rendimiento

Las tecnologías de secuenciación masiva o de siguiente generación, por sus siglas en inglés: NGS next generation sequencing, son un conjunto de tecnologías de alto rendimiento que permiten obtener secuencias de nucleótidos en poco tiempo y a gran escala. Las principales son: Roche 454 o pirosecuenciación, secuenciación Illumina (NextSeq, HiSeq y MiSeq), SOLiD de *Applied Biosystems*, *Ion Torrent* (Proton, S5, Chef, PGM), Helicos, BGISeq-500 que se basa en la síntesis de sondas de anclaje, y las más recientes PacBio (SMRT) y *Oxford Nanopore* (MinION, GridION, PromethION). Para llevar a cabo la secuenciación masiva usando las tecnologías de Roche, Illumina, *Ion Torrent* y *SOLID* se debe amplificar la molécula de interés, es decir, se requiere de un paso que es la amplificación. Esto significa que se hacen muchas copias del ADN, para posteriormente realizar la reacción de secuenciación por síntesis o ligación y así obtener lecturas (secuencias) que son de tamaño pequeño, esto es 100, 150, 200, 300 o 400 nucleótidos en promedio. Las tecnologías Oxford Nanopore o PacBio, por el contrario, permiten obtener secuencias de mucho mayor tamaño, del orden de unos miles o decenas de miles de nucleótidos. Esto es relevante desde el punto de vista de la bioinformática, por ejemplo para lograr ensamblar genomas complejos, o resolver regiones difíciles. Por otro lado, la secuenciación se da en tiempo real conforme avanza la reacción de secuenciación a partir de una sóla molécula. Así, en un sólo experimento se puede obtener el genoma o transcriptoma de una o muchas células individualmente. De estos experimentos obtenemos los datos crudos 'raw data' y se deben analizar a través de un flujo de trabajo o pipeline. Es natural que a partir de estos adelantos tecnológicos exista un crecimiento exponencial de la información y un aumento en la demanda por estrategias bioinformáticas para acceder, analizar e interpretar los datos y obtener información, modelar o hacer predicciones.

Aplicaciones en biología y biomedicina

Las tecnologías de secuenciación han ayudado a la evaluación del cáncer hereditario, permitiendo realizar perfiles genéticos partiendo de cantidades pequeñas de tejido tumoral, inclusive de células individuales. También contribuyen a entender los mecanismos de resistencia a los medicamentos lo que conlleva a una mejora en el tratamiento. Otro gran aporte de estas aplicaciones son los paneles de genes que permiten analizar mutaciones específicas para un tipo de cáncer en particular u otras enfermedades, con un enfoque dirigido con aplicación en la clínica. Estas tecnologías contribuyen a personalizar los tratamientos basándose en la información genética y biomarcadores farmacogenéticos. Se mejora el diagnóstico de las enfermedades y se busca un tratamiento integral y dirigido al paciente en lo que se conoce como medicina de precisión. Otra de las aplicaciones, es la vigilancia de patógenos y detección de genes de resistencia a antibióticos. Esto es a través de la identificación de la diversidad de

Comité Editorial: González Andrade, M.; Hernández Alcántara, G.; Martínez González, J.J.; Meraz Cruz, N.; Ramírez Silva, L.H. y Vilchis Landeros, M.M.
Publicado por el Departamento de Bioquímica de la Facultad de Medicina; UNAM.

bacterias en una muestra (secuenciación microbioma o del gen del ARN ribosomal 16S) o mediante la detección de genes de virulencia y de resistencia a patógenos. Se hace también secuenciación de genoma completo de nuevos organismos. Los experimentos de transcriptoma o RNA-Seq permiten explorar la regulación o expresión génica, y la secuenciación ChIP-Seq, DNaseI-seq, BS-seq o sRNA-Seq obtener información epigenómica (26).

Protocolos de análisis bioinformático: flujos de trabajo

Un flujo de trabajo en bioinformática es una secuencia de pasos para el análisis de los datos, implica revisar y preparar los datos, correr software especializado y proceder con las entradas y salidas de los programas. Esto es así porque siempre se utiliza más de un programa para el análisis. Generalmente, los flujos de trabajo se encuentran bien documentados, en manuales, Github o sitios web. Un pipeline bioinformático hace también referencia a las instrucciones que se ejecutan paso a paso como parte del análisis primario, secundario o terciario. El análisis primario empieza con la secuenciación y termina con la evaluación de la calidad y limpieza de las secuencias. Le sigue el análisis secundario que es un pipeline específico para cada tipo de experimento o aplicación. Por ejemplo, es diferente determinar las variantes genéticas de interés médico en exomas de pacientes a obtener el genoma completo de un individuo. Y por último, el análisis terciario, es la interpretación y evaluación de esos resultados, realizado no solo por el bioinformático sino por todo el grupo de investigadores y especialistas en el tema. Podemos decir que cada tipo de estudio o análisis en particular tiene su propio flujo de trabajo, por lo que siempre se remite a las guías y/o referencias de uso de los programas y al flujo de trabajo que definen los investigadores. Para ello, se suelen utilizar conjuntos de datos reducidos y guías de instrucciones que copiamos y pegamos en nuestras propias consolas de trabajo.

Ejemplo de flujo de trabajo
(https://github.com/MaryoHg/portillo_etal)

Materiales

Equipo: Una laptop o computadora de escritorio, con mínimo 8 GB de memoria RAM (se recomiendan 16GB o más, dependiendo de la cantidad de datos). Software: Suite QIIME 1.9 (o imagen VirtualBox), RStudio o Bioconductor.

Procedimiento para el análisis de la microbiota

Este procedimiento está enfocado en muestras preparadas en bibliotecas con índices duales (Illumina). En las cuales los índices o barcodes están dentro de las secuencias en ambos extremos. Se emplea el pipeline bioinformático QIIME 1.9 (*Quantitative Insights Into Microbial Ecology*) (27, 28). Se pueden usar los datos del BioProject del NCBI con número de acceso PRJNA545497 (29).

1.- Verificar que el mapping file tenga el formato requerido por QIIME 1.9. Este es un formato de texto plano delimitado por tabulaciones que debe incluir la descripción de cada muestra (metadatos). Es obligatorio que el archivo mapping_file.tsv contenga las siguientes columnas con los nombres específicos: #SampleID, BarcodeSequence, LinkerPrimerSequence, Sample, Description (obligatorio como última columna):

#SampleID	BarcodeSeq	LinkerPrimerSeq	Sample	Description

Se usa la instrucción **validate_mapping_file.py** escribiendo en línea de comandos lo siguiente:

```
$ validate_mapping_file.py -m mapping_file.tsv -o mapping_verificado/
```

Explicación: -m (abreviado) --mapping_fp (completo) indica la ruta del archivo; -o --output_dir indica el directorio en donde se guardan los archivos generados. En el directorio mapping_verificado/ se genera un archivo HTML que muestra los errores que pudiera existir en el mapping file.

2.- Extraer los índices (barcodes). Los índices están unidos a las secuencias y requieren ser extraídos. En secuencias pareadas (*paired end*), existe una secuencia *forward* y otra *reverse*. La instrucción genera un archivo llamado barcodes.fastq que contiene los índices de ambas secuencias. Este archivo se usará en el paso 3.

```
$ extract_barcodes.py -f forward.fastq -r reverse.fastq -c barcode_paired_end -l 8 -L 8 -o barcodes/
```

Explicación: -f, --fastq1 indica la ruta del archivo forward.fastq; -r, --fastq2 la ruta del reverse.fastq; -c, --input_type el tipo de secuencia de entrada, por ejemplo: *barcode_paired_end*; -l, --bc1_len la longitud que tiene el índice en pares de bases del

Comité Editorial: González Andrade, M.; Hernández Alcántara, G.; Martínez González, J.J.; Meraz Cruz, N.; Ramírez Silva, L.H. y Vilchis Landeros, M.M.
Publicado por el Departamento de Bioquímica de la Facultad de Medicina; UNAM.

archivo *forward*, por ejemplo: 8; -L, --bc2_len la longitud del índice en pares de bases del archivo *reverse*, por ejemplo: 8; -o, --output_dir es la ruta de salida. Ejemplo: barcodes/

3.- Unir las secuencias forward y reverse en un solo archivo fastq. Se generarán los archivos fastqjoin.join.fastq y fastqjoin.join_barcodes.fastq, que se usarán en el paso siguiente.

```
$ join_paired_ends.py -f forward.fastq -r reverse.fastq -b barcodes/barcodes.fastq -j 100 -p 10 -o join/
```

Explicación: -f --forward_reads_fp indica la ruta del archivo forward.fastq; -r --reverse_reads_fp la ruta del archivo reverse.fastq; -b, --index_reads_fp la ruta del archivo barcodes.fastq generado en el paso anterior; -j --min_overlap es el número de pares de bases que se sobrelapan en la unión. Ejemplo: 100; -p, --perc_max_diff es el porcentaje (%) de diferencias permitidas en la región de sobrelape; -o, --output_dir es la salida para los archivos generados. Ejemplo: join/

4.- Separar las secuencias por nombre de la muestra que se especifica el mapping file (#sampleID). El índice se elimina de las secuencias y se lleva a cabo el filtrado por calidad Phred. La instrucción genera un archivo llamado seqs.fna que contiene las secuencias filtradas con una calidad mínima de Q=25.

```
$ split_libraries_fastq.py -i join/fastqjoin.join.fastq -b join/fastqjoin.join_barcodes.fastq -m mapping_file.tsv --max_barcode_errors 2 --barcode_type 16 -q 25 -o split/
```

Explicación: -i --sequence_read_fps es la ruta del archivo fastqjoin.join.fastq; -b --barcode_read_fps ruta del archivo fastqjoin.join_barcodes.fastq; -m --mapping_fps ruta del archivo mapping_file.tsv; --max_barcode_errors número máximo de errores en los índices; --barcode_type tipo de índice empleado, por ejemplo: 16; -q --phred_quality_threshold calidad mínima Phred que tendrán las secuencias, por ejemplo 25; -o --output_dir es la ruta de salida. Ejemplo: split/

5.- Descartar las secuencias quiméricas. Las secuencias quiméricas son artefactos generados principalmente en el proceso de amplificación por PCR, siendo errores, es necesario eliminarlas. Primero se identifican las quimeras obteniendo un archivo chimeras.txt que contiene la lista de muestras con quimeras:

```
$ identify_chimeric_seqs.py -i split/seqs.fna -r gg_13_8_otus/rep_set/97_otus.fasta -m usearch61 --threads 4 -o quimeras/
```

Explicación: -i --input_fasta_fp ruta del archivo seqs.fna; -r --reference_seqs_fp directorio de las secuencias de referencia greengenes database_13.8, se puede usar ***print_qiime_config.py -t*** para ubicar 97_otus.fasta; -m --chimera_detection_method método para la identificación de quimeras, por ejemplo: usearch61; --threads el número de hilos de ejecución empleados para la identificación de las quimeras, 4; -o, --output_fp es la ruta de salida. Ejemplo: quimeras/. Posteriormente, las secuencias identificadas se filtran del archivo seqs.fna:

```
$ filter_fasta.py -f split/seqs.fna -s quimeras/chimeras.txt -n -o non_chim_seqs.fna
```

Explicación: -f --input_fasta_fp ruta del archivo seqs.fna; -s --seq_id_fp ruta del archivo chimeras.txt generado en el paso anterior; -n --negate indica que todas las secuencias identificadas como quimeras se eliminen del archivo seqs.fna; -o --output_fasta_fp nombre del archivo de salida con las secuencias filtradas. Ejemplo: non_chim_seqs.fna. Este archivo contiene las secuencias sin quimeras.

6.- Agrupar las secuencias en unidades taxonómicas operativas (OTU) y asignación taxonómica. Los OTUs se forman en función del porcentaje de similitud de las secuencias, generalmente se emplea el 97% de similitud para la agrupación de las secuencias. QIIME permite agrupar las secuencias con tres enfoques: de novo, referencia cerrada y referencia abierta (ver más detalles en Navas-Molinas et al. 2013). En este flujo de trabajo se emplea la referencia abierta que permite el uso de una base de datos de referencia y compara las secuencias con ella. El comando ***pick_open_reference_otus.py*** realiza la agrupación en OTUs, construye un árbol filogenético y realiza la asignación taxonómica de las secuencias. Es necesario crear un archivo de texto con el nombre parametros.txt que contenga la siguiente línea: "enable_rev_strand_match: True". Este parámetro permite la anotación taxonómica de las secuencias en orientación reversa.

Comité Editorial: González Andrade, M.; Hernández Alcántara, G.; Martínez González, J.J.; Meraz Cruz, N.; Ramírez Silva, L.H. y Vilchis Landeros, M.M.
Publicado por el Departamento de Bioquímica de la Facultad de Medicina; UNAM.

```
$pick_open_reference_otus.py -i non_chim_seqs.fna -p parametros.txt -s 0.1 -m usearch61 -r 97_otus.fasta -a -O 2 -o otus -v
```

Explicación: -i --input_fps ruta del archivo non_chim_seqs.fna creado en el paso anterior; -p --parameter_fp ruta del archivo parametros.txt; -s --percent_subsample porcentaje de falla de las secuencias para incluir en la agrupación de novo, por ejemplo: 0.1; -m --otu_picking_method método de agrupamiento de OTUs, por ejemplo: usearch61; -r --reference_fp son las secuencias de referencia, por defecto, gg_13_8_otus/rep_set/97_otus.fasta; -a --parallel permite que el proceso se realice en paralelo; -O --jobs_to_start es el número de trabajos a realizar en paralelo, por ejemplo: 2; -o --output_dir directorio de salida. Ejemplo: otus/; -v --verbose muestra en la pantalla los pasos que se están realizando mientras se ejecuta el comando.

Los resultados que se obtienen son dos archivos: otu_table_mc2_w_tax_no_pynast_failures.biom (tabla en formato BIOM que contiene las secuencias agrupadas en OTUs y con su asignación taxonómica) y rep_set.tre (árbol filogenético creado con las secuencias representativas). El archivo BIOM se puede emplear para realizar diversos análisis (diversidad alfa y beta, análisis estadísticos, gráficos) en R, Rstudio, Origin, etc.

7.- En QIIME 1.9 se pueden realizar diversos gráficos y análisis para presentar los resultados. Por ejemplo, usando ***core_diversity_analyses.py*** se ejecutan otros scripts ***alpha_rarefaction.py, beta_diversity_through_plots.py, summarize_taxa _through_plots.py, make_distance_boxplots.py, compare_alpha_diversity.py*** y ***group _significance .py***. Se obtienen gráficos de barras con las abundancias relativas de las secuencias, análisis de componentes principales (PCoA), gráficos de cajas y bigotes. Además, se realizan los análisis de diversidad alfa Observed_species, Chao1, Shannon, Simpson, entre otros. Esta instrucción requiere que indiquemos la profundidad media de nuestras secuencias para el análisis de rarefacción (usado para los análisis de diversidad alfa). Para conocer la profundidad a partir de nuestra tabla BIOM escribimos: ***biom summarize-table*** (https://biom-format.org/documentation).

```
$ biom summarize-table -i out_table_mc2_w_tax _no_pynast_failures.biom -o otu_table_summary.txt
```

Explicación: -i ruta del archivo BIOM creado en el paso anterior; -o archivo de salida. Ejemplo: otu_table_summary.txt.

El archivo otu_table_summary.txt contiene el número de muestras, nombre de las muestras, media, mediana y otros datos obtenidos de la tabla BIOM. Se recomienda emplear el valor de la media de las muestras para el análisis de diversidad. Para especificar los diferentes métodos empleados para el análisis de diversidad es opcional crear el archivo parametros_core.txt con lo siguiente:

```
beta_diversity:metrics bray_curtis, euclidean, unweighted_unifrac,weighted_unifrac
alpha_diversity:metrics observed_species,chao1
```

```
$core_diversity_analyses.py --recover_from_ failure i otu_table_mc2_w_tax_no_pynast_ failures. biom -m mapping_file.txt -t rep_set.tre -e 900 -p parametros_core.txt -a -O 4 -v -o diversity/
```

Explicación: -i --input_biom_fp ruta del archivo BIOM; -m --mapping_fp ruta del archivo mapping_file.txt; -t --tree_fp ruta del árbol filogenético rep_set.tre creado en el paso 6; -e --sampling_depth profundidad de las muestras, calculado con ***biom summarize-table***; -p --parameter_fp ruta del archivo parametros_core.txt; -a --parallel indica que el proceso se realice en paralelo; -O --jobs_to_start número de trabajos a realizar en paralelo, por ejemplo: 4; -o --output_dir directorio de salida de los archivos generados. Ejemplo: diversity/. Se recomienda el uso de la opción --recover_from_failure para recuperar el avance del trabajo tras cierto fallo computacional o de suministro de energía. Los resultados en formato HTML incluyen gráficas y documentos de texto (Figura 1).

Conclusiones

La bioinformática permite almacenar, procesar, y analizar la información biológica, ordenarla y generar nuevo conocimiento. Es una herramienta y disciplina en la intersección entre áreas muy diversas, como son la estadística, las ciencias de la computación y la biología. La bioinformática es interdisciplinaria por lo que se nutre de muchas otras disciplinas relacionadas. La bioinformática tiene un papel relevante en los avances de la biología y la biomedicina, hace uso de protocolo de análisis y flujos de trabajo empleando diversas herramientas computacionales. Los resultados obtenidos con la bioinformática se deben interpretar de forma crítica,

Comité Editorial: González Andrade, M.; Hernández Alcántara, G.; Martínez González, J.J.; Meraz Cruz, N.; Ramírez Silva, L.H. y Vilchis Landeros, M.M.
Publicado por el Departamento de Bioquímica de la Facultad de Medicina; UNAM.

el análisis no puede ser una caja negra en donde sólo se ingresan datos y se obtienen resultados. El bioinformático puede participar en todo el proceso de la investigación, desde la recolección de la información, planteamiento de los objetivos y planeación del experimento, además del procesamiento, análisis e interpretación.

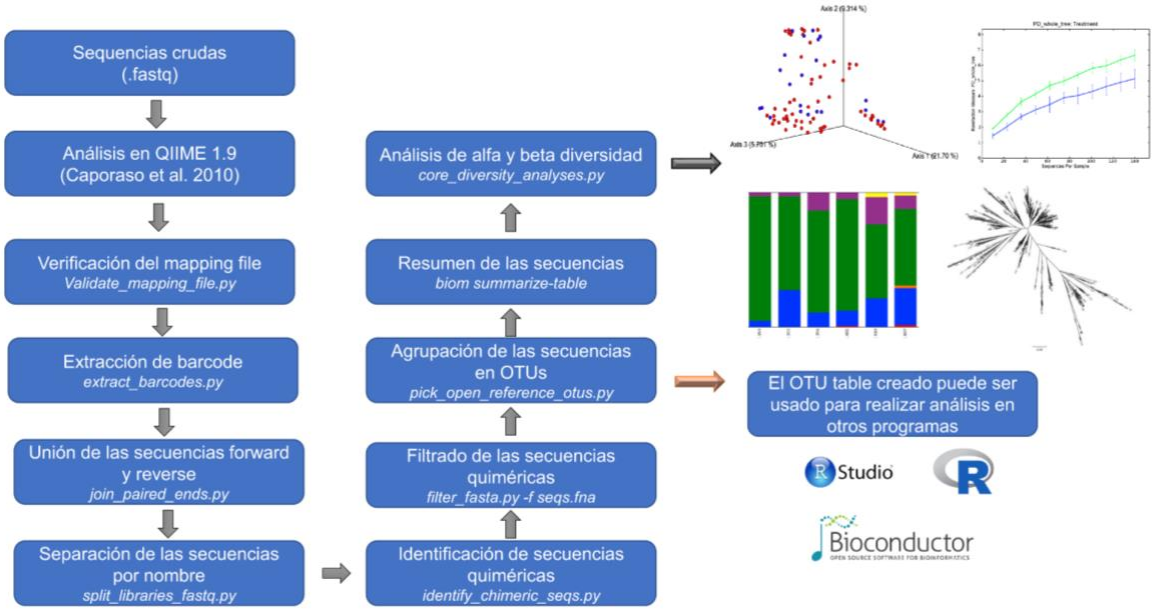

Figura 1. Pipeline del análisis de la microbiota en QIIME 1.9. El tiempo estimado para realizar los análisis dependen del número de secuencias y de la capacidad computacional. Para 3x106 secuencias pareadas y empleando una portátil de 4 hilos, 16 GB de RAM y 2.4 GHz de frecuencia, las estimaciones de tiempo son: 15 minutos para validación del mapping file, extracción de barcode, unión de las secuencias *forward* y *reverse*, separación de las secuencias por nombre. Siendo los procesos que consumen mayor tiempo: identificación y filtro de secuencias quiméricas (15-25 min), agrupación de las secuencias en unidades taxonómicas operacionales (40-80 min) y análisis de alfa y beta diversidad (60-120 min). Opcionalmente se puede emplear comandos específicos (según se requiera) incluidos dentro del script core_diversity_analyses.py para sólo calcular índices o hacer ciertos gráficos y reducir tiempos de cómputo: alpha_diversity.py que calcula los índices tradicionales de diversidad alfa ó alpha_rarefaction.py que nos arroja los gráficos de rarefacción. La OTU table se puede convertir a un archivo de texto tabular (tsv ó csv) y este puede ser empleado para realizar análisis y gráficos en programas externos a QIIME 1.9. Siendo R y Rstudio los más populares para realizar diversos análisis (heatmap, barplots, PCA, PCoA, diagramas de Sankey, perMANOVA, ANOVA).

Referencias

1. Luscombe, N. M., Greenbaum, D., and Gerstein, M. (2001) What is Bioinformatics? A Proposed Definition and Overview of the Field. Methods Inf Med. 40, 346–358
2. Gómez-López, G., Dopazo, J., Cigudosa, J. C., Valencia, A., and Al-Shahrour, F. (2019) Precision medicine needs pioneering clinical bioinformaticians. Briefings in Bioinformatics. 20, 752–766
3. Servant, N., Roméjon, J., Gestraud, P., La Rosa, P., Lucotte, G., Lair, S., Bernard, V., Zeitouni, B., Coffin, F., Jules-Clément, G., Yvon, F., Lermine, A., Poullet, P., Liva, S., Pook, S., Popova, T., Barette, C., Prud'homme, F., Dick, J.-G., Kamal, M., Le Tourneau, C., Barillot, E., and Hupé, P. (2014) Bioinformatics for precision medicine in oncology: principles and application to the SHIVA clinical trial. Front. Genet. 10.3389/fgene.2014.00152
4. Gauthier, J., Vincent, A. T., Charette, S. J., and Derome, N. (2019) A brief history of bioinformatics. Briefings in Bioinformatics. 20, 1981–1996
5. Hagen, J. B. (2000) The origins of bioinformatics. Nat Rev Genet. 1, 231–236
6. Fruton, J. S. (2009) An episode in the history of protein chemistry: Pehr Edman's method for the sequential degradation of peptides. International Journal of Peptide and Protein Research. 39, 189–194
7. Sanger, F., and Tuppy, H. (1951) The amino-acid sequence in the phenylalanyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates. Biochemical Journal. 49, 463–481
8. Sanger, F., and Thompson, E. O. P. (1953) The amino-acid sequence in the glycyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates. Biochemical Journal. 53, 353–366
9. Avery, O. T., MacLeod, C. M., and McCarty, M. (1944) Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Journal of Experimental Medicine. 79, 137–158
10. Hershey, A. D., and Chase, M. (1952) Independent functions of viral protein and nucleic acid in growth of bacteriophage. Journal of General Physiology. 36, 39–56
11. Watson, J. D., and Crick, F. H. C. (1953) Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. Nature. 171, 737–738
12. Stasiak, A. (2003) The first lady of DNA. EMBO Rep. 4, 14–14
13. Strasser, B. J. (2010) Collecting, Comparing, and Computing Sequences: The Making of Margaret O. Dayhoff's Atlas of

Comité Editorial: González Andrade, M.; Hernández Alcántara, G.; Martínez González, J.J.; Meraz Cruz, N.; Ramírez Silva, L.H. y Vilchis Landeros, M.M.
Publicado por el Departamento de Bioquímica de la Facultad de Medicina; UNAM.

Protein Sequence and Structure, 1954–1965. J Hist Biol. 43, 623–660
14. Dayhoff, M. O., and Ledley, R. S. (1962) Comprotein: a computer program to aid primary protein structure determination. in Proceedings of the December 4-6, 1962, fall joint computer conference on - AFIPS '62 (Fall), pp. 262–274, ACM Press, Philadelphia, Pennsylvania, 10.1145/1461518.1461546
15. 15. Bartlett, A., Penders, B., and Lewis, J. (2017) Bioinformatics: indispensable, yet hidden in plain sight? BMC Bioinformatics. 18, 311
16. Chang, J. T., Volk, D. E., Gorenstein, D. G., Steffen, D., and Bernstam, E. V. (2019) Bioinformatics service center projects go beyond service. Journal of Biomedical Informatics. 94, 103192
17. Bartlett, A., Lewis, J., and Williams, M. L. (2016) Generations of interdisciplinarity in bioinformatics. New Genetics and Society. 35, 186–209
18. Chang, J. (2015) Core services: Reward bioinformaticians. Nature. 520, 151–152
19. Hulsen, T., Jamuar, S. S., Moody, A. R., Karnes, J. H., Varga, O., Hedensted, S., Spreafico, R., Hafler, D. A., and McKinney, E. F. (2019) From Big Data to Precision Medicine. Front. Med. 6, 34
20. Parker, M. S., Burgess, A. E., and Bourne, P. E. (2021) Ten simple rules for starting (and sustaining) an academic data science initiative. PLoS Comput Biol. 17, e1008628
21. Armenta-Medina, D., Díaz de León-Castañeda, C., y Valderrama-Blanco, B. (2020) Bioinformatics in Mexico: A diagnostic from the academic perspective and recommendations for a public policy. PLoS ONE. 15, e0243531
22. Corpas, M., Fatumo, S., and Schneider, R. (2012) How Not to Be a Bioinformatician. Source Code Biol Med. 7, 3
23. Smith, D. R. (2015) Broadening the definition of a bioinformatician. Front. Genet. 10.3389/fgene.2015.00258
24. Vincent, A. T., and Charette, S. J. (2015) Who qualifies to be a bioinformatician? Front. Genet. 10.3389/fgene.2015.00164
25. Rigden, D. J., and Fernández, X. M. (2022) The 2022 Nucleic Acids Research database issue and the online molecular biology database collection. Nucleic Acids Research. 50, D1–D10
26. Hsu, F.-M., Gohain, M., Chang, P., Lu, J.-H., and Chen, P.-Y. (2018) Bioinformatics of Epigenomic Data Generated From Next-Generation Sequencing. in Epigenetics in Human Disease, pp. 65–106, Elsevier, 10.1016/B978-0-12-812215-0.00004-2
27. Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., Reeder, J., Sevinsky, J. R., Turnbaugh, P. J., Walters, W. A., Widmann, J., Yatsunenko, T., Zaneveld, J., and Knight, R. (2010) QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 7, 335–336
28. Navas-Molina, J. A., Peralta-Sánchez, J. M., González, A., McMurdie, P. J., Vázquez-Baeza, Y., Xu, Z., Ursell, L. K., Lauber, C., Zhou, H., Song, S. J., Huntley, J., Ackermann, G. L., Berg-Lyons, D., Holmes, S., Caporaso, J. G., and Knight, R. (2013) Advancing Our Understanding of the Human Microbiome Using QIIME. in Methods in Enzymology, pp. 371–444, Elsevier, 531, 371–444
29. Hernández-Guzmán, M., Pérez-Hernández, V., Navarro-Noya, Y. E., Luna-Guido M. L., Verhulst N., Govaerts, B. y Dendooven L. (2022) Application of ammonium to a N limited arable soil enriches a succession of bacteria typically found in the rhizosphere. Scientific Reports. 12, 4110.

M. en C . TOBÍAS PORTILLO BOBADILLA
ORCID: 0000-0002-3448-7959

Biólogo egresado de la Facultad de Ciencias de la UNAM, realizó un diplomado en Desarrollo e Implementación de Sistemas con Software Libre en Linux en la Dirección General de Cómputo Académico (DGSCA, UNAM). Es autor del software educativo multimedia Interacciones macromoleculares ver. 1.0 y 1.6, que fue elaborado como tesis de licenciatura en el Departamento de Programas Audiovisuales de la Facultad de Química (DePA), Instituto de Investigaciones Biomédicas, Facultad de Ciencias y Academia de San Carlos de la UNAM en su versión 2.0. Obtuvo mención Honorífica y Segundo Lugar en el Concurso Latinoamericano a la Mejor Aplicación multimedia en 2001. En el Instituto de Ecología de la UNAM realizó estudios de posgrado sobre la dinámica evolutiva de los genomas de las enterobacterias. Posteriormente, fue miembro del Consorcio Genoma *Taenia solium*. La Maestría en Ciencias Biológicas la realizó en la Unidad de Medicina Experimental de la Facultad de Medicina en la UNAM, estudiando el desarrollo de la microbiota intestinal en la comunidad de Xoxocotla, Morelos.

En docencia ha impartido pláticas de divulgación sobre la estructura de las proteínas en la Universidad Autónoma del Estado de Morelos UAEM, Facultad de Ciencias, FES Iztacala de la UNAM y en Universum. Es profesor de asignatura nivel B en la Facultad de Ciencias de la UNAM.

Actualmente es Técnico Académico Asociado C de Tiempo Completo, a cargo de los servicios de bioinformática en la Unidad de Bioinformática, Bioestadística y Biología Computacional de la Red de Apoyo a la Investigación (RAI), Coordinación de la Investigación Científica (CIC) UNAM e Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán (INCMNSZ). Proporciona asesorías especializadas en bioinformática, genómica, secuenciación masiva y es responsable de la administración de servidores y del diseño del sitio web de

Comité Editorial: González Andrade, M.; Hernández Alcántara, G.; Martínez González, J.J.; Meraz Cruz, N.; Ramírez Silva, L.H. y Vilchis Landeros, M.M.
Publicado por el Departamento de Bioquímica de la Facultad de Medicina; UNAM.

la Red de Apoyo a la Investigación. Coautor en diversos artículos de investigación y de divulgación de la ciencia.

Comité Editorial: González Andrade, M.; Hernández Alcántara, G.; Martínez González, J.J.; Meraz Cruz, N.; Ramírez Silva, L.H. y Vilchis Landeros, M.M.
Publicado por el Departamento de Bioquímica de la Facultad de Medicina; UNAM.