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) 13-24

TAB

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

Análisis filogenético de Proteínas: Una herramienta básica para la Medicina Evolutiva o Filomedicina.

Phylogenetic analysis of Proteins: A basic tool for Evolutionary Medicine or Phylomedicine

Riveros-Rosas, Héctor*[1]; Julián-Sánchez, Adriana[1] y Carrillo-Campos, Javier[1].

1. Departamento de Bioquímica, Facultad de Medicina, Universidad Nacional Autónoma de México.

*Correspondencia. Piso 8, Torre de Investigación, Depto. Bioquimica, Fac. Medicina, Universidad Nacional Autonoma de Mexico. Ciudad Universitaria, Coyoacán, CDMX, 04510. Tel. +52(55)56232284; hriveros@unam.mx

Resumen

Actualmente, como resultado de la implementación de técnicas de secuenciación masiva, tenemos acceso público a la información de miles de genomas completos y más de 400 millones de secuencias de genes y proteínas. Sin embargo, en menos del 1% de esas proteínas, contamos con información que vaya más allá de un marco de lectura abierto que determine su secuencia primaria. Por esta razón ha sido imperativo implementar métodos de análisis de secuencias y uno de ellos en particular es el análisis filogenético de secuencias que ha resultado muy exitoso; permitiendo, por ejemplo: discernir entre genes ortólogos y parálogos, estimar tiempos de divergencia, reconstruir proteínas ancestrales, o incluso determinar residuos importantes para la selección natural e identificar mutaciones claves que producen enfermedades o permiten que una proteína adquiera nuevas funciones. Su trascendencia actual va más allá de la Biología Evolutiva, y se ha convertido en una de las herramientas básicas de la Medicina moderna, conformando lo que ahora denominamos como Medicina Evolutiva, Medicina Darwiniana o simplemente Filomedicina. El objetivo de este capítulo es proporcionar algunos lineamientos generales que permitan a los estudiantes que por primera vez realizan un análisis filogenético de proteínas, obtener resultados útiles y confiables.

Abstract

Today, as a result of the implementation of massive sequencing techniques, we have public access to the information contained in thousands of complete genomes and more than 400 million gene and protein sequences. However, in less than 1% of these proteins, we have information that goes beyond an open reading frame that determines their primary sequence. For this reason, it has been imperative to implement sequence analysis methods and one of them in particular, phylogenetic sequence analysis, has been particularly successful: allowing, for example, to discern between orthologous and paralogous genes, estimate divergence times, reconstruct ancestral proteins, or even more identify important residues for natural selection or determine key mutations that cause diseases or allow a protein to acquire new functions. Its current importance goes beyond Evolutionary Biology, and has become one of the basic tools of modern Medicine, forming what we now call Evolutionary Medicine, Darwinian Medicine or just Philomedicine. The goal of this chapter is to provide some general guidelines that will enable students who are performing protein phylogenetic analysis for the first time to obtain useful and reliable results.

Palabras claves: Uniprot, BLASTP, alineamiento de secuencias múltiple, análisis de secuencias, Medicina Darwiniana.

Keywords: Uniprot, BLASTP, multiple sequence alignment, sequence analysis, Darwinian Medicine.

Introducción

A partir del desarrollo de las técnicas de secuenciación masiva, tenemos acceso público a miles de genomas completos y millones de secuencias de genes y proteínas. Esta abundancia de información ha permitido la aplicación de nuevas técnicas para responder preguntas fundamentales en Medicina, que nos ayudan a poner en contexto las características que observamos en nuestras células, tejidos, órganos, etc., tanto en condiciones normales como patológicas. De hecho, nuevas aplicaciones de la Biología Evolutiva a los problemas médicos, son cada vez más numerosas, lo que ha dado pie al desarrollo de una nueva rama del conocimiento médico: la Medicina Evolutiva o Darwiniana. Actualmente, PubMed registra casi 2,000 publicaciones que incluyen los términos "*Evolutionary Medicine*" o "*Darwinian Medicine*". Una de las primeras publicaciones en esta área, fue por ejemplo, el análisis evolutivo entre los genes y la resistencia a los antibióticos (1), y la primera publicación que incluyó en su título el término "*Darwinian Medicine*", se publicó apenas en 1991 (2).

La Medicina Evolutiva o Darwiniana (e.g. (3-7), o simplemente Filomedicina (*Phylomedicine*) (8), coloca a los humanos en un ámbito biológico, tratando de interpretar nuestras características, fortalezas y limitaciones en un contexto evolutivo, interpretándonos nosotros (los humanos) como el resultado de un largo y complejo proceso evolutivo, que ha dejado sus huellas tanto en el genoma como en nuestra anatomía y fisiología, en donde muchas de nuestras enfermedades se pueden entender y analizar, (y tal vez tratar) de manera más adecuada tomando en cuenta los principios básicos de la Bioquímica y la Biología Evolutiva, que comprenden tanto las características y propiedades de nuestros genes y proteínas que codifican, como el análisis filogenético de sus secuencias.

El análisis filogenético de nuestros genes y proteínas, nos permite identificar los sitios conservados a lo largo de la evolución, y que son críticos para su función o estructura, y por ende, nos permite identificar también con precisión, las mutaciones responsables de patologías, y cómo la selección natural puede actuar sobre ellas. El poder elucidar la historia evolutiva de nuestros genes, nos permite ahora plantear preguntas y/o hipótesis que antes no era posible.

Todo esto no ha pasado desapercibido para los expertos en Educación Médica. Así, la Medicina Evolutiva, se ha consolidado en los últimos 25 años, al grado que la *American Association of Medical Colleges* y el *Howard Hughes Medical Institute*, proponen que los estudiantes de Medicina deben adquirir los fundamentos básicos de la Biología Evolutiva, a efecto de aplicarlos en su práctica médica, como una más de las competencias que deben adquirir los estudiantes en sus estudios formales de licenciatura (9). Esta misma propuesta, ha sido formulada de manera independiente, por muchos otros autores en diferentes Escuelas de Medicina (e.g., (10-14)). Así, los estudiantes de Medicina, además de poseer una formación básica en matemáticas, física, química, bioquímica, biología celular, fisiología y método científico, deben tener también una formación básica en Biología Evolutiva. Esto no quiere decir, que los conocimientos que adquieran en matemáticas, física, química o alguna de las otras competencias básicas, tengan aplicación directa en su práctica médica, pero si evidentemente, les permite a los médicos durante su práctica clínica, evaluar mejor la información diagnóstica disponible, y tomar mejores decisiones médicas.

La perspectiva evolucionista también permite tomar ventaja de la gran cantidad de secuencias de genes y proteínas disponible. Desafortunadamente, dentro del universo de proteínas reportadas, en muy pocos casos se cuenta con información más allá de la predicción de una secuencia de aminoácidos. Al día de hoy por ejemplo, de los poco más de 230 millones de secuencias incluidas en el Universal *Protein Resource (UniProt) del European Bioinformatics Instituto* (EBI) (https://www.uniprot.org/) sólo en el 0.08% de los casos se cuenta con evidencia experimental que demuestre su existencia, en el 0.6% sólo se cuenta con evidencia experimental de que existe un RNAm que puede dirigir su síntesis; en el 31.67% se sabe que corresponden a probables proteínas reales porque son homólogas a alguna otra proteína ya conocida, pero en dos terceras partes de las secuencias de UniProt, lo único con que se cuenta, es con un marco de lectura abierto (*open reading frame*) de una secuencia de DNA (Tabla 1). Esto no hace más que subrayar la necesidad de desarrollar estrategias que nos permitan obtener

Comité Editorial: Apellido Paterno1 Apellido Materno1 N1.; Apellido Paterno2 Apellido Materno2 N2. y Apellido Paterno3 Apellido Materno3 N3. *Publicado por el Departamento de Bioquímica de la Facultad de Medicina; UNAM.*

información a partir del análisis de secuencias primarias de proteínas, y que constituyen la abrumadora mayoría.

Tabla 1. Proporción de secuencias de proteínas depositadas en la base de datos Uniprot/TrEMBLE distribuidas de acuerdo con la evidencia experimental disponible sobre su existencia.

Evidencia experimental que demuestra la existencia de proteínas	Número de secuencias	Porcentaje
Evidencia a nivel de proteínas	179,881	0.08%
Evidencia a nivel de transcritos (RNAm)	1,376,788	0.60%
Evidencia inferida a partir de homología	72,941,987	31.67%
Evidencia sólo por predicción (marco de lectura abierto)	155,829,992	67.66%
Universo total de proteínas secuenciadas	230,328,648	100.00%

Datos obtenidos a partir de las estadísticas publicadas en UniProt: https://www.ebi.ac.uk/uniprot/TrEMBLstats (19 de enero de 2022).

Estos datos proporcionan la razón de ser del presente capítulo, cuya intención es proporcionar al lector que se aproxima por primera vez al empleo del análisis filogenético de proteínas, algunas guías generales que le permitan obtener información útil y confiable a partir del análisis de secuencias. Para ello, podemos dividir el trabajo en tres etapas principales: *i)* obtención de secuencias; *ii)* elaboración de alineamientos, y *iii)* construcción de árboles filogenéticos.

I. Obtención de secuencias

Todo análisis *in silico* de proteínas debe comenzar por recuperar la(s) secuencia(s) de aminoácidos de la(s) proteína(s) de nuestro interés. En un principio (hace un par de décadas), el problema fundamental radicaba en tratar de recuperar de las bases de datos públicas todas las posibles secuencias homólogas a nuestra proteína de interés. Sólo por dar un ejemplo, nosotros publicamos en 2003 un análisis de todas las proteínas que formaban parte de la superfamilia de alcohol deshidrogenasas dependientes de Zinc (Zn-ADHs) (15). En aquel entonces recuperamos 656 secuencias distintas de Zn-ADHs y que constituían el total de las secuencias homólogas disponibles en ese momento. Si el día de hoy quisiéramos repetir el trabajo y recuperar todas las secuencias disponibles de Zn-ADHs disponibles en las bases de datos, recuperaríamos casi 120,000 secuencias (un número inmanejable para fines prácticos).

Este problema se repite ahora prácticamente con cualquier familia de proteínas. Para dar una idea del incremento podemos referirnos a la figura 1 que muestra el número de secuencias disponibles en la base de datos Uniprot/TrEMBLE desde el año 2000 a la fecha. En ella puede observarse que el número de secuencias disponibles se ha incrementado de forma prácticamente exponencial a lo largo de los años. Por esta razón es muy importante tener claro de antemano qué tipo de secuencias homólogas estamos interesados en recuperar, y debemos tener presenteque el tamaño de las diferentes familias de proteínas es realmente muy variable.

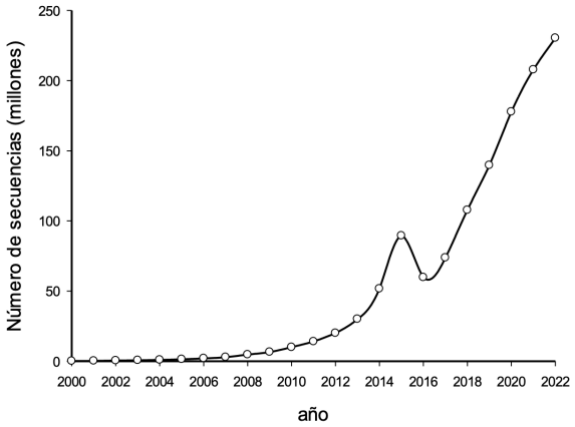

Figura 1. Número de secuencias depositadas en la base de datos Uniprot/TrEMBLE a lo largo del tiempo. Nota: entre los años 2015-2016 se observa una notable disminución en el número de secuencias depositadas en la base de datos. Esto fue resultado de un cambio en los criterios de inclusión de secuencias en la base de datos con el objetivo de eliminar secuencias idénticas que eran resultado de la inclusión de proteomas bacterianos redundantes (para más información ver https://insideuniprot.blogspot.com/2015/05/uniprot-knowledgebase-just-got-smaller.html).

Para realizar una búsqueda de secuencias homólogas, se requiere únicamente contar con la secuencia primaria de una proteína, o en su defecto, con el número de acceso asociado a una secuencia de aminoácidos de una proteína. Para facilitar el análisis de la secuencia de aminoácidos de las proteínas, cada aminoácido se sustituye por una letra, de manera que la estructura primaria de las proteínas es representada por una secuencia de letras. Esta forma de representar la estructura de las proteínas, con base en la asignación de letras establecida por Margaret Dayhoff (16), se le denominó formato *fasta*, y fue implementado para que el programa FASTA, desarrollado por Lipman & Pearson, pudiera realizar

Comité Editorial: Apellido Paterno1 Apellido Materno1 N1.; Apellido Paterno2 Apellido Materno2 N2. y Apellido Paterno3 Apellido Materno3 N3. *Publicado por el Departamento de Bioquímica de la Facultad de Medicina; UNAM.*

análisis de similitud de secuencias (17). En el formato fasta, la primera línea indicada con el carácter ">" corresponde siempre al nombre de la secuencia. Todos los caracteres indicados después de la primera línea corresponden a los aminoácidos que componen la secuencia primaria de la proteína. Aquí, es conveniente siempre indicar en el nombre de la secuencia su número de acceso, para que si alguien quiere repetir/revisar los análisis realizados pueda hacerlo (Figura 2).

```
>WP_012714300.1|GCF_000022385|Sulfolobus_islandicus_LS215|478¶
MKSYQGLADKWIKGSGEEYLDINPADKDHVLAKIRLYTKDDVKEAINKAVAKFDEWSRTP¶
APKRGSILLKAGELMEQEAQEFALLMTLEEGKTLKDSMFEVTRSYNLLKFYGALAFKISG¶
KTLPSADPNTRIFTVKEPLGVVALITPWNFPLSIPVWKLAPALAAGNTAVIKPATKTPLM¶
VAKLVEVLSKAGLPEGVVNLVVGKGSEVGDTIVSDDNIAAVSFTGSTEVGKRIYKLVGNK¶
NRMTRIQLELGGKNALYVDKSADLTLAAELAIRGGFGLTGQSCTATSRLIINKDVYTQFK¶
QRLLERVKKWRVGPGTEDVDMGPVVDEGQFKKDLEYIEYGKNVGAKLIYGGNIIPGKGYF¶
LEPTIFEGVTSDMRLFKEEIFGPVLSVTEAKDLDEAIRLVNAVDYGHTAGIVASDIKAIN¶
EFVSRVEAGVIKVNKPTVGLELQAPFGGFKNSGATTWKEMGEDALEFYLKEKTVYEGW¶
¶
>P28469|ADH1A_MACMU|Macaca_mulatta|375¶
MSTAGKVIKCKAAVLWEVMKPFSIEDVEVAPPKAYEVRIKMVTVGICGTDDHVVSGTMVT¶
PLPVILGHEAAGIVESVGEGVTTVEPGDKVIPLALPQCGKCRICKTPERNYCLKNDVSNP¶
RGTLQDGTSRFTCRGKPIHHFLGVSTFSQYTVVDENAVAKIDAASPMEKVCLIGCGFSTG¶
YGSAVKVAKVTPGSTCAVFGLGGVGLSAVMGCKAAGAARIIAVDINKDKFAKAKELGATE¶
CINPQDYKKPIQEVLKEMTDGGVDFSFEVIGRLDTMMASLLCCHEACGTSVIVGVPPDSQ¶
NLSINPMLLLTGRTWKGAVYGGFKSKEDIPKLVADFMAKKFSLDALITHVLPFEKINEGF¶
DLLRSGKSIRTILTF¶
¶
```

Figura 2. Ejemplo de dos secuencias primarias de proteínas en formato fasta, que es el requerido por los programas de búsqueda de secuencias homólogas. El nombre de la secuencia del primer ejemplo incluye el número de acceso: WP_012714300.1 (del GenBank); la clave del genoma: GCF_000022385; el nombre de la especie: Sulfolobus_islandicus_LS215; y el número de aminoácidos de la proteína: 478. En el segundo ejemplo, el número de acceso es: P28469 (de UniProt); el nombre de la proteína es: ADH1_MACMU (asignado por UniProt); el nombre de la especie es Macaca_mulatta y su longitud en aminoácidos es: 375. En UniProt es importante no confundir el nombre de la proteína con el número de acceso, porque este último SI es permanente, mientras que el nombre de la proteína NO, y puede ser modificado (Elisabeth Gasteiger, comunicación personal). Se recomienda no utilizar espacios en el nombre porque algunos programas cortan el nombre de la proteína al encontrar el primer espacio en blanco.

Una de las herramientas más populares y útiles para recuperar secuencias homólogas de proteínas es BLASTP, que corresponde al acrónimo en inglés de *Basic Local Alignment Search Tool* (18). Esta herramienta identifica, muy rápidamente, a través de fragmentos muy cortos de proteínas (2 a 6 aminoácidos) secuencias idénticas. Una vez identificadas las proteínas con fragmentos idénticos, el algoritmo revisa si la similitud entre las secuencias puede extenderse antes y/o después del fragmento previamente identificado para finalmente proporcionar un listado de las secuencias homólogas identificadas, ordenándolas de acuerdo a su significancia estadística medida a través del valor estadístico E (E-value). Este último valor depende principalmente de dos parámetros, la identidad o similitud entre las secuencias a comparar, y del número de aminoácidos que pueden alinearse entre ellas. Si el valor estadístico E al comparar dos secuencias nos arroja números menores a 10^{-5} (E-values < 0.00001) podemos tener confianza en que la similitud entre ellas es estadísticamente significativa, pero si el valor estadístico E presenta valores en el rango 10^{-5} < E-value < 0.001 deben ser tomados con ciertas reservas porque la similitud observada pudiera ser simplemente resultado del azar. La similitud entre secuencias con E-values > 0.001 es casi con certeza, resultado del azar.

BLASTP puede ser utilizado con facilidad a través de la WEB, en dos de las bases de datos más utilizadas para la búsqueda de proteínas: el GenBank del *National Center for Biotechnology Information* (NCBI) ((19); https://blast.ncbi.nlm.nih.gov/Blast.cgi); y Uniprot ((20); https://www.uniprot.org/blast/). Tanto en el NCBI-GenBank como en UniProt, las búsquedas de secuencias de proteínas pueden realizarse consultando distintas bases de datos. Las más importantes están listadas en la Tabla 2 para el caso del GenBank y la Tabla 3 para el caso de UniProt. Así, uno puede darse una idea del número de

Comité Editorial: Apellido Paterno1 Apellido Materno1 N1.; Apellido Paterno2 Apellido Materno2 N2. y Apellido Paterno3 Apellido Materno3 N3. *Publicado por el Departamento de Bioquímica de la Facultad de Medicina; UNAM.*

secuencias que se podrían recuperar al emplear cada una de las distintas bases de datos. Qué base de datos consultar dependerá de la(s) pregunta(s) específica(s) que se quiere(n) resolver.

Tabla 2. Número de secuencias de proteínas depositadas en las diferentes bases de datos contenidas en el GenBank del NCBI.

Base de datos: GenBank (NCBI)	Número de secuencias de proteínas
Secuencias de proteínas no redundantes *[Non-redundant protein sequences (nr)]*	464,314,559
Proteínas de referencia seleccionadas *[RefSeq Select proteins (refseq select)]*	26,294,491
Proteínas de referencia *[Reference Proteins]*	217,955,956
Secuencias de proteínas no redundantes agrupadas(1) *[Clustered no redundant (nr_clustered)]*	212,789,408
Organismos modelo *[Model organisms (landmark)]*	440,691
Base de datos del Swiss-Prot (Uniprot) *[Swiss-Prot database]*	478,866
Banco de datos de estructura de proteínas *[Protein Data Bank (pdb)]*	136,675

Secuencias depositadas en el National Center for Biotechnology Information al 21 de marzo de 2022. https://blast.ncbi.nlm.nih.gov/Blast.cgi
(1) Las secuencias son agrupadas cuando presentan un 90% de identidad o más y conservan al menos un 90% de su longitud de secuencia.

Tabla 3. Número de secuencias de proteínas depositadas en las diferentes bases de datos contenidas en UniProt (EBI).

Base de datos UniProt	Número de secuencias de proteínas
Uniparc	485,953,918
Uniprotkb (Swiss-Prot + TrEMBLE)	230,936,395
Uniprotkb (reference proteomes + Swiss-Prot)	64,368,998
Uniprotkb (reference proteomes)	64,126,468
Uniprotkb (Swiss-Prot)	607,747
Uniprotkb (Protein Data Bank)	68,320
Sequence clusters	
Uniref100	297,827,854
uniref90	144,113,457
Uniref50	51,333,317

Secuencias depositadas en UniProt al 21 de marzo de 2022. La base de datos UniRef100 combina secuencias idénticas y fragmentos de proteínas de un organismo específico en un registro único. UniRef90 y UniRef 50 se construyen agrupando las secuencias de UniRef100 a un nivel de identidad del 90 o 50% respectivamente.

Existen también otras herramientas de búsqueda de secuencias homólogas de proteínas. DIAMOND por ejemplo (21,22), es una herramienta cuya ventaja principal es ser más rápida que BLASTP al momento de realizar la búsqueda de secuencias homólogas. HMMER por su parte (https://www.ebi.ac.uk/Tools/hmmer/), es una herramienta que a través del empleo de perfiles de modelos ocultos de Markov (profile hidden Markov models (HMM)), es capaz de identificar con mayor sensibilidad que BLASTP, la existencia de homología entre secuencias (23).

Comité Editorial: Apellido Paterno1 Apellido Materno1 N1.; Apellido Paterno2 Apellido Materno2 N2. y Apellido Paterno3 Apellido Materno3 N3. *Publicado por el Departamento de Bioquímica de la Facultad de Medicina; UNAM.*

Finalmente, no está de más mencionar que si en un genoma reportado como completamente secuenciado en las bases de datos, no podemos recuperar con BLASTP la proteína de nuestro interés, esto no puede ser considerado como una prueba inequívoca de que la proteína en cuestión está ausente, porque los métodos automatizados de predicción de genes, con relativa frecuencia no identifican correctamente los marcos de lectura abierta que conforman un gen (falsos negativos). Si es importante tener certeza sobre la presencia/ausencia de un gen determinado, en ese caso puede utilizarse la herramienta TBLASTN, que busca a partir de la secuencia primaria de una proteína, secuencias de nucleótidos que una vez traducidas (con los seis posibles marcos de lectura), pudieran codificar la proteína de nuestro interés. Sólo si después de buscar con TBLASTN no podemos recuperar una secuencia de cDNA que pudiera codificar nuestra proteína de interés (utilizando como punto de partida la secuencia homóloga más cercana), entonces se puede concluir que el gen que codifica esa proteína está ausente.

II. Elaboración de alineamientos

Una vez que tenemos un conjunto de secuencias homólogas, se puede emplear alguno de los programas que generan alineamientos múltiples de proteínas, aunque primero debe prestarse atención a lo siguiente:

1. Los programas que generan alineamientos de proteínas, alinean cualquier secuencia que uno les proporcione, sean homólogas o no. Esto es importante tenerlo presente porque en algunos casos se recuperan secuencias de proteínas de las bases de datos mediante palabras clave (usando el nombre de la enzima por ejemplo). Una estrategia de este tipo no es recomendable porque se corre el riesgo de recuperar proteínas no homólogas que pertenezcan a familias diferentes (proteínas análogas).

2. Con cierta frecuencia los genes que codifican para una proteína dada, aunque hayan sido recuperados por medio de BLASTP, presentan además de la secuencia de interés, dominios adicionales extra que usualmente son resultado de la fusión de dos o más genes (proteínas múltidominio). Si las proteínas que usted va a alinear tienen secuencias de tamaños mucho mayores al esperado, es probable que sus secuencias tengan dominios adicionales.

Así, es importante tomar en cuenta que los programas de alineamiento no pueden discriminar entre segmentos homólogos y no homólogos. Si no se tiene el debido cuidado con las secuencias, se obtendrá un alineamiento de baja calidad, y en consecuencia árboles filogenéticos que probablemente no tengan ninguna utilidad o que incluso, induzcan a conclusiones erróneas. De esta manera, antes de alinear las secuencias, es importante remover todos los segmentos no homólogos que pudieran estar presentes en algunas proteínas.

Actualmente existen distintos programas capaces de generar alineamientos múltiples de proteínas. La Tabla 4 lista algunos de los más populares. Todos los programas de alineamiento múltiple basados en la comparación de secuencias, emplean una matriz de similitud de aminoácidos y generan un árbol guía con base en alineamientos pareados. Una vez identificadas las secuencias más similares entre sí, el resto de las secuencias se van añadiendo de manera progresiva, optimizando el alineamiento mediante la adición de huecos o "gaps" dentro de las secuencias. Para no añadir un exceso de "*gaps*"·los programas penalizan el alineamiento cada vez que se abre y/o extiende un "*gap*" dentro una secuencia, y puede considerarse que los "*gaps*" son el equivalente biológico a mutaciones que ocurrieron por inserción o deleción. En casi todos los programas de alineamiento, la penalización por apertura y extensión de "*gaps*" puede modificarse a criterio del investigador.

Clustal W fue uno de los primeros programas especializado en realizar alineamientos múltiples confiables (24) y esa tal vez es una de las razones por la que aún sigue siendo popular. De hecho, los artículos que describen la primera versión, tanto de Clustal W (24) como de Clustal X (25), se encuentran entre los 100 artículos más citados en ciencias (26). Actualmente se considera que tanto MUSCLE, como MAFFT y T-Coffee generan alineamientos múltiples más precisos (27), sin embargo, tienen el inconveniente de añadir mayor cantidad de "gaps" dentro de los alineamientos, lo cual también tiene sus inconvenientes, tal y como se comenta más adelante.

Algunas instituciones como por ejemplo, el *European Molecular Biology Laboratory* del *European Bioinformatics Institute* (EMBL-EBI) permiten construir alineamientos múltiples en línea, empleando cualquiera de los programas listados en la Tabla 4, e inclusive, con algunos otros programas más (https://www.ebi.ac.uk/Tools/msa/).

Comité Editorial: Apellido Paterno1 Apellido Materno1 N1.; Apellido Paterno2 Apellido Materno2 N2. y Apellido Paterno3 Apellido Materno3 N3. *Publicado por el Departamento de Bioquímica de la Facultad de Medicina; UNAM.*

Tabla 4. Algunos Programas representativos (populares) para la elaboración de alineamientos múltiples de secuencias

Programa	Sitio web	Referencia
Clustal (W. X y Omega)	http://www.ebi.ac.uk/Tools/clustalw/ http://www.clustal.org/	(28,29)
MUSCLE	https://drive5.com/muscle5/manual/topics.html	(30,31)
MAFFT	https://mafft.cbrc.jp/alignment/server/large.html https://mafft.cbrc.jp/alignment/software/	(32,33)
T-Coffee	http://www.tcoffee.org/ https://www.tcoffee.org/Projects/tcoffee/index.html#DOWNLOAD	(34-36)

Una vez que se ha generado el alineamiento múltiple, es importante considerar que algunas secuencias pueden generar "ruido" y demeritar la calidad del alineamiento. En especial se debe tener cuidado en los siguientes casos:

1. Las secuencias que codifican fragmentos de proteínas suelen ser problemáticas, ya que con frecuencia los programas de alineamiento al tratar de alinear los fragmentos, los distribuyen sobre toda la extensión de las proteínas completas, lo que resulta en errores de alineamiento.

2. Los métodos automatizados de predicción de genes no son infalibles, y con cierta frecuencia las bases de datos incluyen predicciones de secuencias de aminoácidos incorrectas, sobre todo en el caso de genomas eucariontes, y casi siempre por problemas en la identificación de los límites de intrones-exones, aunque los problemas en la secuenciación o ensamblado de los contigs del cDNA pueden también ser una causa.

De esta manera, si los fragmentos de proteínas no están correctamente alineados, es muy probable que, al construir los árboles filogenéticos, estas secuencias terminen formando ramas anómalas: que resultan ser largas y en posiciones inesperadas dentro del árbol. Por otra parte, todas aquellas secuencias que en el alineamiento se observen como mal alineadas, antes de concluir que se trata de secuencias divergentes, es conveniente revisar si la predicción es correcta, sobre todo si hay secuencias bien alineadas de especies filogenéticamente cercanas. En este caso, se puede recuperar de las bases de datos la secuencia de nucleótidos de donde se realizó la predicción automatizada original de la proteína, y utilizar algún programa especializado para revisar dicha predicción. Para el caso de proteínas de eucariontes, puede emplearse por ejemplo FGENESH+ (http://www.softberry.com/), un programa que realiza predicciones de genes con base en la secuencia de una proteína homóloga cercana (37).

Finalmente, es importante considerar que algunos programas de alineamiento insertan una cantidad considerable de "*gaps*" para optimizar el alineamiento múltiple entre las secuencias; de hecho, matemáticamente un alineamiento puede optimizarse mientras mayor es el número de "gaps" que se incluyen. El problema de esto último, es que conforme se añaden "gaps" se pierde información desde el punto de vista biológico y el número de posiciones útiles para generar los árboles filogenéticos se reduce conforme se añaden "gaps" (38). Los alineamientos incluidos en algunas bases de datos, como por ejemplo Pfam (https://pfam.xfam.org/) (39), presentan tal cantidad de "gaps", que aunque son útiles para muchos fines, no es adecuado emplearlos para realizar análisis filogenéticos.

No existe hasta el momento, un consenso sobre cuál debe ser el mejor criterio para medir la calidad de un alineamiento múltiple de secuencias (40). Sin embargo, si se cuenta con información estructural de alguna de las secuencias del alineamiento, debe ocurrir entonces que la mayoría de los "gaps" del alineamiento se ubiquen preferentemente en las "asas" de la estructura tridimensional, o en los extremos amino y carboxilo, y muy pocos "*gaps*" dentro de las estructuras secundarias de la(s) proteína(s). De hecho, una alternativa es realizar primero un alineamiento estructural que sirva de guía para alinear el resto de las secuencias. Un ejemplo de un programa capaz de generar un alineamiento múltiple a partir de estructuras tridimensionales de proteínas es VAST (41) (https://www.ncbi.nlm.nih.gov/Structure/VAST/vastsearch.html).

Comité Editorial: Apellido Paterno1 Apellido Materno1 N1.; Apellido Paterno2 Apellido Materno2 N2. y Apellido Paterno3 Apellido Materno3 N3. *Publicado por el Departamento de Bioquímica de la Facultad de Medicina; UNAM.*

III. Construcción de árboles filogenéticos

Una vez que ya se han obtenido y alineado las secuencias de proteínas, el siguiente paso es construir el árbol filogenético a partir de estas. El árbol filogenético nos permite inferir relaciones evolutivas entre las secuencias que se han utilizado para construir el árbol. A partir de las relaciones evolutivas podemos conocer distintos aspectos de las secuencias de proteínas, analizar por ejemplo, la ortología y paralogía, estimar tiempos de divergencia, reconstruir proteínas ancestrales, buscar residuos importantes para la selección natural e identificar mutaciones claves (42).

Existen diversos métodos para obtener un árbol filogenético, que pueden estar basados en distancias, como el del vecino más próximo (*Neighbour-joining*), o basados en caracteres como el de máxima parsimonia (*parsimony*), Máxima Verosimilitud (Maximum Likelihood) o Bayesiano (*Bayesian*). Un aspecto importante que considerar para la construcción de árboles filogenéticos es el modelo de sustitución. Este modelo nos permite estimar la distancia evolutiva entre las secuencias y en los métodos Bayesianos y de Máxima Verosimilitud nos permite calcular la probabilidad de cambio a lo largo de las ramas del árbol filogenético. Los modelos de sustitución utilizan un método de reemplazo de cada aminoácido tomando en cuenta las propiedades biológicas, químicas y físicas de cada uno de ellos, a través de una matriz que resume la tasa de sustitución de cada aminoácido considerando las propiedades mencionadas previamente. Así, por ejemplo, es bastante frecuente un cambio entre una arginina a una lisina mientras que un cambio entre una arginina a un aspartato no lo es tanto.

Algunos modelos de sustitución comúnmente utilizados para construir árboles filogenéticos de secuencias de proteínas son el propuesto por Margaret O. Dayhoff (43), el JTT de Jones, Taylor & Thornton (44) y el WAG de Whelan & Goldman (45); de hecho, el desarrollo de más modelos de sustitución ha sido constante (46). Cada alineamiento debe ser analizado para determinar (empíricamente) cuál modelo de sustitución será el mejor para dicho alineamiento, y poder así, construir el árbol filogenético; una selección arbitraria (sin justificación), del modelo de sustitución puede generar árboles filogenéticos cuyo análisis podría carecer de validez (47). Un ejemplo de un programa capaz de realizar el análisis filogenético de secuencias, y determinar el mejor modelo de sustitución de aminoácidos es MEGA, desarrollado por el grupo de Sudhir Kumar y colaboradores (48,49).

Método del Vecino más Próximo (Neighbour-Joining; NJ)

Este método propuesto por Saitou y Nei (50), es quizás el más popular y más utilizado, su principal ventaja es que es considerablemente más rápido que el resto de los métodos. El método en un inicio estima la distancia evolutiva entre las secuencias, es decir, estima los cambios que han ocurrido entre las secuencias; para este proceso se debe de tener en cuenta que los cambios entre las secuencias no siempre son un reflejo preciso de las distancias evolutivas, un aminoácido puede mutar varias veces y hacer creer que en realidad han sido pocas, provocando una cercanía artificial entre secuencias, el conocimiento de las secuencias será determinante para saber si este método puede ser utilizado.

Máxima Parsimonia (*Maximum Parsimony*)

El principio de máxima parsimonia en Biología aplica de la misma forma que en otras ciencias; la explicación más sencilla que se ajusta a los datos es la que se debe de elegir. En filogenia, este método construye la historia de las secuencias en árboles considerando el mínimo número de cambios. Así, en cada rama del árbol se evalúan las mutaciones necesarias para poder explicar las secuencias alineadas, cada rama recibe un puntaje y este se puede entender como el número mínimo de mutaciones que puede dar lugar a los datos (51).

Al utilizar la máxima parsimonia para obtener el árbol filogenético, es necesario considerar que el resultado será más confiable en tanto la búsqueda de secuencias haya sido exhaustiva, esto permitirá inferir correctamente las relaciones evolutivas; de otra forma podemos presenciar eventos de atracción de ramas largas, en donde erróneamente podemos relacionar secuencias que han evolucionado rápidamente y que de manera correcta deberían de estar separadas dentro del árbol (52,53). Dado que el método de máxima parsimonia busca el camino más sencillo, se omite hacer una búsqueda completa de todos los cambios dentro de las secuencias que puedan explicar los datos; existen otros métodos como Máxima verosimilitud que sí exploran todos los escenarios posibles.

Máxima Verosimilitud (*Maximum-Likelihood*; ML)

Cuando estamos tratando de inferir las relaciones evolutivas a partir de un alineamiento debemos de

Comité Editorial: Apellido Paterno1 Apellido Materno1 N1.; Apellido Paterno2 Apellido Materno2 N2. y Apellido Paterno3 Apellido Materno3 N3. *Publicado por el Departamento de Bioquímica de la Facultad de Medicina; UNAM.*

poder utilizar un método que considere los múltiples eventos de mutación que pudieron haber tenido lugar en las secuencias, ML nos permite conseguir esto. En ML, múltiples árboles son generados y aquel que tiene la mayor verosimilitud es el que se prefiere. ML es uno de los métodos más robustos y que más se prefieren para la construcción de árboles filogenéticos, tienen un respaldo estadístico fuerte, se considera que, con una buena selección de secuencias y un buen alineamiento, ML será capaz de proporcionar en la totalidad de los casos el árbol con la topología correcta (54).

El principal inconveniente del método de ML es el poder de cómputo requerido para poder obtener el árbol filogenético, un alineamiento con un número pequeño de secuencia será fácil de procesar con ML. Sin embargo, alineamientos con un número grande de secuencias requerirán un tiempo considerable para ser procesados con ML a fin de obtener el árbol filogenético.

Bayesiano (*Bayesian*)

Este método es relativamente nuevo, al igual que ML tiene un sustento estadístico firme y en este método se considera la probabilidad posterior, esto es, aquella probabilidad de que un evento ocurra (una mutación en un aminoácido) tomando en consideración toda la información del alineamiento y las secuencias.

El método Bayesiano permite utilizar modelos complejos de evolución para explicar los datos de las secuencias alineadas, esto en principio no es posible con ML y en caso de ser posible, toma considerablemente más tiempo en comparación con el método Bayesiano. En resumen, los métodos de ML y Bayesiano deben ser la primera elección si el poder de cómputo lo permite.

Análisis estadístico por Bootstrap

En cualquiera de los métodos previamente mencionados el producto del método es un árbol filogenético que permite estimar las relaciones evolutivas, la duda que inmediatamente surge es ¿qué tan buen soporte estadístico tiene este árbol? Para contestar esta pregunta comúnmente se utiliza la técnica de Bootstrap (55). Con esta técnica la matriz de datos se reutiliza mediante muestreos para generar distintos árboles filogenéticos (usualmente mediante repeticiones del orden de 500 veces). Cada nuevo árbol puede compartir topología con el anterior en todo el árbol o en algunas regiones, aquellas regiones que no se repiten tan seguido se denominan de bajo soporte y se puede decir que no se resuelven de manera adecuada, de tal forma que el análisis de esas regiones debe ser hacerse con cautela. Así, en cada rama se registra la proporción de repeticiones en donde se obtuvo la misma topología. Valores de bootstrap cuyo porcentaje de repeticiones es >70% se consideran con buen soporte estadístico.

Corolario

El análisis filogenético de proteínas, empleado correctamente, es una herramienta muy poderosa para estudiar la estructura, función y evolución de familias de proteínas. Para ello, es muy importante seleccionar cuidadosamente las secuencias de proteínas que se utilizaran para generar los alineamientos, revisar que en los alineamientos obtenidos los motivos y residuos clave estén correctamente identificados y alineados. Determinar, previo a la construcción de los árboles filogenéticos, cuál es el mejor modelo de sustitución de aminoácidos, para finalmente obtener los mejores árboles posibles de acuerdo con la capacidad de cómputo disponible (Figura 3). Desafortunadamente, los métodos que generan alineamientos no cuentan con una metodología estandarizada para medir la calidad o confianza estadística de los mismos, y si bien la confianza estadística de los árboles filogenéticos puede evaluarse por el método de bootstrap o probabilidad posterior en el caso de los árboles bayesianos, su utilidad siempre está supeditada a la calidad de los alineamientos, y dado que los programas de alineamiento siempre son capaces de generar uno (no importando si las secuencias incluidas son homólogas o no), y que los programas de análisis filogenético también siempre son capaces de generar un árbol (no importando la calidad del alineamiento), no es extraño encontrar en la literatura resultados filogenéticos que no aportan nada más allá de una simple figura extra, a partir de la cuál es casi imposible extraer alguna conclusión útil y/o confiable. Dado que hoy en día contamos con una enorme cantidad de información en forma de secuencias de proteínas (230 millones), es imperativo analizarlas correctamente para aprovechar al máximo la información que pueden aportarnos.

Comité Editorial: Apellido Paterno1 Apellido Materno1 N1.; Apellido Paterno2 Apellido Materno2 N2. y Apellido Paterno3 Apellido Materno3 N3. *Publicado por el Departamento de Bioquímica de la Facultad de Medicina; UNAM.*

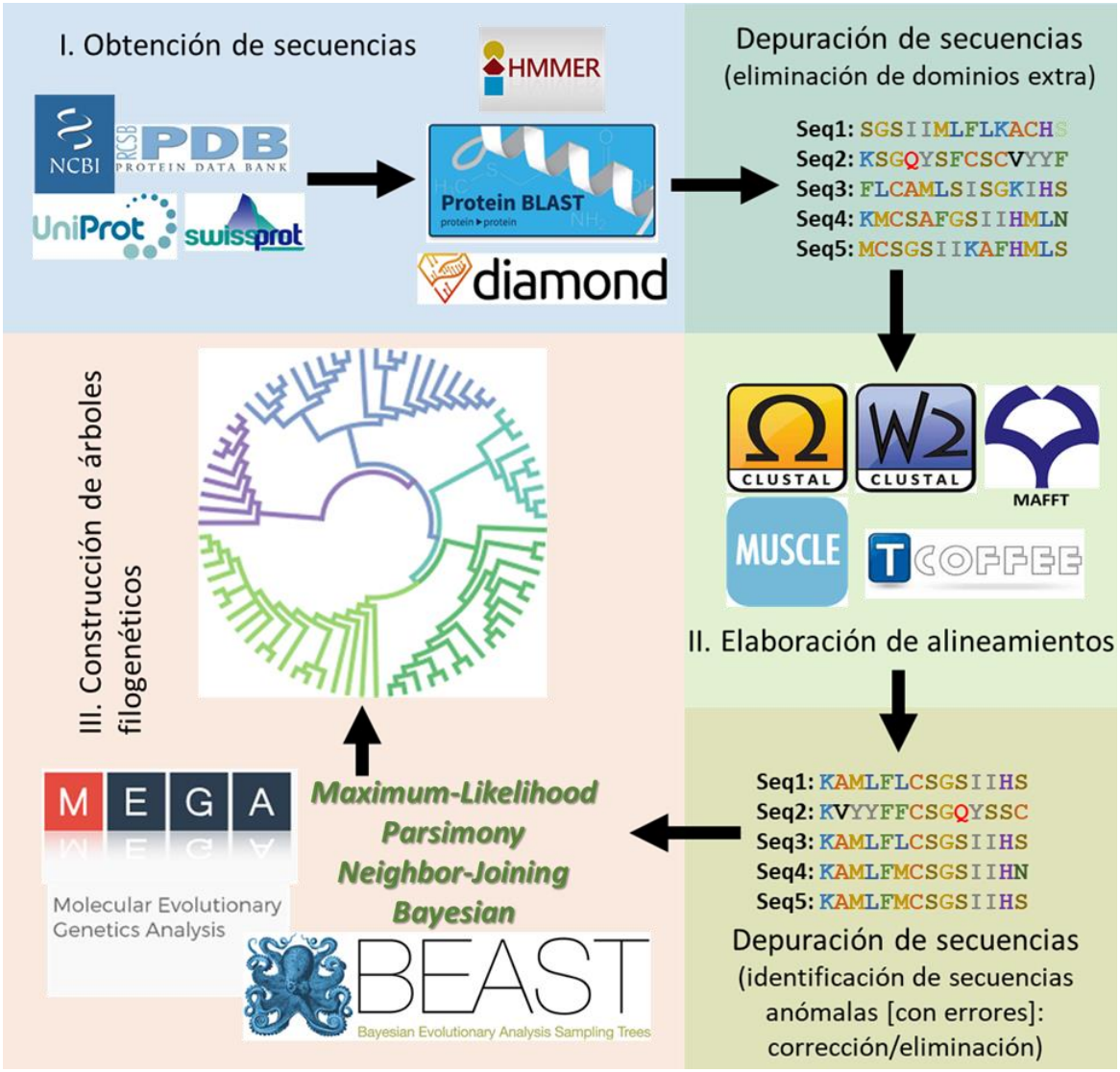

Figura 3. Esquema que muestra las etapas que conforman un análisis filogenético de proteínas: I) obtención de las secuencias (usualmente a partir de bases de datos públicas como el NCBI-GeneBank y Uniprot; II) elaboración de alineamientos múltiples, verificando que las secuencias sólo presentan el dominio de interés (eliminación de dominios extra no homólogos); y III) construcción de árboles filogenéticos (identificando y eliminando o corrigiendo aquellas secuencias con posiciones anómalas en el árbol).

Agradecimientos

Se agradece el apoyo del Programa UNAM-PAPIIT IN219022 (HRR) y del Programa de becas Posdoctorales en la UNAM (JCC).

Referencias

1. Davies, J., and Gray, G. (1984) Evolutionary relationships among genes for antibiotic resistance. Ciba Foundation symposium 102, 219-232
2. Williams, G. C., and Nesse, R. M. (1991) The dawn of Darwinian medicine. The Quarterly review of biology 66, 1-22
3. Andersson, D. I., Balaban, N. Q., Baquero, F., Courvalin, P., Glaser, P., Gophna, U., Kishony, R., Molin, S., and Tonjum, T. (2020) Antibiotic resistance: turning evolutionary principles into clinical reality. FEMS microbiology reviews 44, 171-188
4. Moltzau Anderson, J., and Horn, F. (2020) (Re-) Defining evolutionary medicine. Ecology and evolution 10, 10930-10936
5. Perlman, R. L. (2013) Evolution and medicine. Perspectives in biology and medicine 56, 167-183
6. Perry, G. H. (2021) Evolutionary medicine. eLife 10
7. Stearns, S. C. (2020) Frontiers in Molecular Evolutionary Medicine. Journal of molecular evolution 88, 3-11

Comité Editorial: Apellido Paterno1 Apellido Materno1 N1.; Apellido Paterno2 Apellido Materno2 N2. y Apellido Paterno3 Apellido Materno3 N3. *Publicado por el Departamento de Bioquímica de la Facultad de Medicina; UNAM.*

8. Kumar, S., Dudley, J. T., Filipski, A., and Liu, L. (2011) Phylomedicine: an evolutionary telescope to explore and diagnose the universe of disease mutations. Trends in genetics : TIG 27, 377-386
9. Nesse, R. M., Bergstrom, C. T., Ellison, P. T., Flier, J. S., Gluckman, P., Govindaraju, D. R., Niethammer, D., Omenn, G. S., Perlman, R. L., Schwartz, M. D., Thomas, M. G., Stearns, S. C., and Valle, D. (2010) Evolution in health and medicine Sackler colloquium: Making evolutionary biology a basic science for medicine. Proceedings of the National Academy of Sciences of the United States of America 107 Suppl 1, 1800-1807
10. Antolin, M. F., Jenkins, K. P., Bergstrom, C. T., Crespi, B. J., De, S., Hancock, A., Hanley, K. A., Meagher, T. R., Moreno-Estrada, A., Nesse, R. M., Omenn, G. S., and Stearns, S. C. (2012) Evolution and medicine in undergraduate education: a prescription for all biology students. Evolution; international journal of organic evolution 66, 1991-2006
11. Graves, J. L., Jr., Reiber, C., Thanukos, A., Hurtado, M., and Wolpaw, T. (2016) Evolutionary Science as a Method to Facilitate Higher Level Thinking and Reasoning in Medical Training. Evolution, medicine, and public health 2016, 358-368
12. Harris, E. E., and Malyango, A. A. (2005) Evolutionary explanations in medical and health profession courses: are you answering your students' "why" questions? BMC medical education 5, 16
13. Nesse, R. M., and Stearns, S. C. (2008) The great opportunity: Evolutionary applications to medicine and public health. Evolutionary applications 1, 28-48
14. Palanza, P., and Parmigiani, S. (2016) Why human evolution should be a basic science for medicine and psychology students. Journal of anthropological sciences = Rivista di antropologia : JASS 94, 183-192
15. Riveros-Rosas, H., Julian-Sanchez, A., Villalobos-Molina, R., Pardo, J. P., and Pina, E. (2003) Diversity, taxonomy and evolution of medium-chain dehydrogenase/reductase superfamily. European journal of biochemistry 270, 3309-3334
16. Dayhoff, M. O., Eck, R. V., Chang, M. A., and Sochard, M. R. (1965) Atlas of Protein Sequence and Structure, National Biomedical Research Foundation, Silver Spring, Maryland
17. Lipman, D. J., and Pearson, W. R. (1985) Rapid and sensitive protein similarity searches. Science 227, 1435-1441
18. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. Journal of molecular biology 215, 403-410
19. Cock, P. J., Chilton, J. M., Gruning, B., Johnson, J. E., and Soranzo, N. (2015) NCBI BLAST+ integrated into Galaxy. GigaScience 4, 39
20. Pundir, S., Martin, M. J., O'Donovan, C., and UniProt, C. (2016) UniProt Tools. Current protocols in bioinformatics 53, 1 29 21-21 29 15
21. Buchfink, B., Reuter, K., and Drost, H. G. (2021) Sensitive protein alignments at tree-of-life scale using DIAMOND. Nature methods 18, 366-368
22. Buchfink, B., Xie, C., and Huson, D. H. (2015) Fast and sensitive protein alignment using DIAMOND. Nature methods 12, 59-60
23. Potter, S. C., Luciani, A., Eddy, S. R., Park, Y., Lopez, R., and Finn, R. D. (2018) HMMER web server: 2018 update. Nucleic acids research 46, W200-W204
24. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic acids research 22, 4673-4680
25. Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic acids research 25, 4876-4882
26. Van Noorden, R., Maher, B., and Nuzzo, R. (2014) The top 100 papers. Nature 514, 550-553
27. Edgar, R. C., and Batzoglou, S. (2006) Multiple sequence alignment. Current opinion in structural biology 16, 368-373
28. Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., and Higgins, D. G. (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23, 2947-2948
29. Sievers, F., and Higgins, D. G. (2021) The Clustal Omega Multiple Alignment Package. Methods in molecular biology 2231, 3-16
30. Edgar, R. C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC bioinformatics 5, 113
31. Edgar, R. C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research 32, 1792-1797
32. Katoh, K., Rozewicki, J., and Yamada, K. D. (2019) MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Briefings in bioinformatics 20, 1160-1166
33. Katoh, K., and Standley, D. M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution 30, 772-780
34. Di Tommaso, P., Moretti, S., Xenarios, I., Orobitg, M., Montanyola, A., Chang, J. M., Taly, J. F., and Notredame, C. (2011) T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. Nucleic acids research 39, W13-17
35. Garriga, E., Di Tommaso, P., Magis, C., Erb, I., Mansouri, L., Baltzis, A., Floden, E., and Notredame, C. (2021) Multiple Sequence Alignment Computation Using the T-Coffee Regressive Algorithm Implementation. Methods in molecular biology 2231, 89-97
36. Notredame, C., Higgins, D. G., and Heringa, J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. Journal of molecular biology 302, 205-217
37. Solovyev, V., Kosarev, P., Seledsov, I., and Vorobyev, D. (2006) Automatic annotation of eukaryotic genes, pseudogenes and promoters. Genome biology 7 Suppl 1, S10 11-12
38. Hartmann, S., and Vision, T. J. (2008) Using ESTs for phylogenomics: can one accurately infer a phylogenetic tree from a gappy alignment? BMC evolutionary biology 8, 95
39. Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., and Bateman, A. (2021) Pfam: The protein families database in 2021. Nucleic acids research 49, D412-D419
40. Wong, T. K. F., Kalyaanamoorthy, S., Meusemann, K., Yeates, D. K., Misof, B., and Jermiin, L. S. (2020) A minimum reporting standard for multiple sequence alignments. NAR genomics and bioinformatics 2, lqaa024
41. Gibrat, J. F., Madej, T., and Bryant, S. H. (1996) Surprising similarities in structure comparison. Current opinion in structural biology 6, 377-385
42. Holder, M., and Lewis, P. O. (2003) Phylogeny estimation: traditional and Bayesian approaches. Nature reviews. Genetics 4, 275-284
43. Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C. (1978) A model of evolutionary change in proteins. in Atlas of Protein Sequence and Structure (Dayhoff, M. O. ed.), National

Comité Editorial: Apellido Paterno1 Apellido Materno1 N1.; Apellido Paterno2 Apellido Materno2 N2. y Apellido Paterno3 Apellido Materno3 N3. *Publicado por el Departamento de Bioquímica de la Facultad de Medicina; UNAM.*

Biomedical Research Foundation, Washington, DC. pp 345-352
44. Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992) The rapid generation of mutation data matrices from protein sequences. Computer applications in the biosciences : CABIOS 8, 275-282
45. Whelan, S., and Goldman, N. (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Molecular biology and evolution 18, 691-699
46. Arenas, M. (2015) Trends in substitution models of molecular evolution. Frontiers in genetics 6, 319
47. Kelchner, S. A., and Thomas, M. A. (2007) Model use in phylogenetics: nine key questions. Trends in ecology & evolution 22, 87-94
48. Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018) MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Molecular biology and evolution 35, 1547-1549
49. Tamura, K., Stecher, G., and Kumar, S. (2021) MEGA11: Molecular Evolutionary Genetics Analysis Version 11. Molecular biology and evolution 38, 3022-3027
50. Saitou, N., and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular biology and evolution 4, 406-425
51. Swofford, D. L. (2002) PAUP: Phylogenetic Analysis Using Parsimony (and Other Methods), Version 4.0 Beta 10., Sinauer Associates., Sunderland, MA. USA
52. Felsenstein, J. (1978) Cases in which Parsimony or Compatibility Methods Will be Positively Misleading. Systematic Zoology 27, 401-410
53. Felsenstein, J. (1996) Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. Methods in enzymology 266, 418-427
54. Whelan, S., Lio, P., and Goldman, N. (2001) Molecular phylogenetics: state-of-the-art methods for looking into the past. Trends in genetics : TIG 17, 262-272
55. Efron, B. (1979) Bootstrap Methods: Another Look at the Jackknife. Ann. Statist. 7, 1-26.

DR. HECTOR RIVERO-ROSAS
ORCID: 0000-0003-0194-2537

Biólogo egresado de la Facultad de Ciencias, UNAM. Realizó sus estudios de Maestría y Doctorado en Ciencias Biomédicas (Bioquímica) en la Facultad de Medicina, UNAM, bajo la dirección del Dr. Enrique Piña Garza. Realizó una estancia sabática en la Wilfrid Laurier University, en Waterloo, Ontario en el laboratorio del Dr. Gabriel Moreno-Hagelsieb.

Es miembro del Sistema Nacional de Investigadores, nivel 2, y revisor ad hoc de más de una veintena de revistas arbitradas. Es autor de 43 trabajos de investigación original publicados en revistas internacionales, dos trabajos de divulgación internacional, así como autor de un libro sobre Método Científico (Ed. Trillas), editor de seis libros nacionales, incluyendo las tres últimas ediciones de la Bioquímica de Laguna (Ed. Manual Moderno), autor de cinco capítulos de libros internacionales, y cuatro nacionales, además de 16 trabajos de docencia/difusión. Algunos de sus trabajos han sido reseñados en *The Scientist* y *ASBMB Today*. Recibió con sus colegas en 2008, el Premio “Dr. Maximiliano Ruíz Castañeda”, otorgado por la Academia Nacional de Medicina.

Sus principales líneas de investigación giran alrededor del análisis filogenético de proteínas, utilizando como paradigma las enzimas responsables de la oxidación del etanol. Sus publicaciones han alcanzado ya las 1,500 citas (excluyendo autocitas).

Actualmente es Profesor Titular B de TC, Departamento de Bioquímica, Facultad de Medicina, UNAM.

Comité Editorial: Apellido Paterno1 Apellido Materno1 N1.; Apellido Paterno2 Apellido Materno2 N2. y Apellido Paterno3 Apellido Materno3 N3. *Publicado por el Departamento de Bioquímica de la Facultad de Medicina; UNAM.*