Imagen: "mtHN9DI" (www.wallpapercave.com)

ENLAZANDO QUÍMICA Y BIOLOGÍA: SECUENCIAS DE PROTEÍNAS

Linking chemistry and biology: protein sequences

*Por/ by Roberto Laos y/ and Steven A. Benner**

En los últimos veinte años el número de genomas completos que han sido secuenciados y depositados en bancos de datos ha crecido dramáticamente. Esta abundancia de información de secuencias ha servido de base para la creación de una disciplina llamada paleogenética. En este artículo, sin ahondar en algoritmos complejos, presentamos algunos conceptos clave para comprender cómo las proteínas han evolucionado con el tiempo. Luego ilustraremos como la paleogenética es utilizada en biotecnología. Estos ejemplos resaltan la conexión entre la química y la biología, dos disciplinas que quizás veinte años atrás parecían ser mucho más distintas que lo que parecen ser hoy.

Palabras clave: Bioinformática, proteínas ancestrales, biología sintética, ingeniería de proteínas.

In the last twenty years, the number of complete genomes that have been sequenced and deposited in data banks has grown dramatically. This abundance in sequence information has supported the creation of the discipline known as paleogenetics. In this article, without going into complex algorithms, we present some key concepts for understanding how proteins have evolved in time. We then illustrate how paleogenetic analysis can be used in biotechnology. These examples highlight the connection between chemistry and biology, two disciplines that twenty years ago seemed to be more different than what they seem to be today.

Keywords: Bioinformatics, ancestral proteins, synthetic biology, protein engineering.

A pesar de que la química orgánica y la biología se han desarrollado por separado durante mucho tiempo, los avances en ambas disciplinas están borrando las divisiones entre estas, y fuerzan a los químicos a aprender acerca de biología evolutiva y a los biólogos a aprender más de síntesis orgánica. Hace no mucho tiempo, la química era una disciplina muy diferente a la que conocemos hoy en día, como lo recuerda Frank Westheimer[1], quien es reconocido como uno de los químicos pioneros en combinar biología, síntesis orgánica y bioinformatica: cuando él recibió su doctorado en el año 1935, los instrumentos como UV, IR, RMN no habían sido desarrollados aun; tampoco la cromatografía de papel o de columna. Un estudiante de química en estos días solo se imaginaría un laboratorio de química sin estos instrumentos en un museo.

Sin embargo, el rápido avance científico hizo que el panorama cambiase rápido y en pocos años ya estarían disponibles los RMN. De hecho, la primera persona en tomar un espectro de resonancia magnética nuclear de una proteína[2] en 1957 (Figura 1), el químico Martin Saunders, es hoy en día el profesor más longevo de la Universidad de Yale en los Estados Unidos. Pareciera que el salto del laboratorio "de museo" al laboratorio moderno ocurrió hace aproximadamente 60 años.

**Los autores son doctores en Química e investigadores de la Fundación para la Evolución Molecular Aplicada (Foundation for Applied Molecular Evolution, FAME). (💻) Contacto: 13709 Progress Blvd. No. 7, Alachua, FL 32615, Estados Unidos de América. (Email: rlaos@ffame.org)*

1. Westheimer, F. H., Musings. *Journal of Biological Chemistry* **2003**, *278* (14), 11729-11730. (💻)

2. Saunders, M.; Wishnia, A.; Kirkwood, J. G., The nuclear magnetic resonance spectrum of ribonuclease. *Journal of the American Chemical Society* **1957**, *79* (12), 3289-3290. (💻)

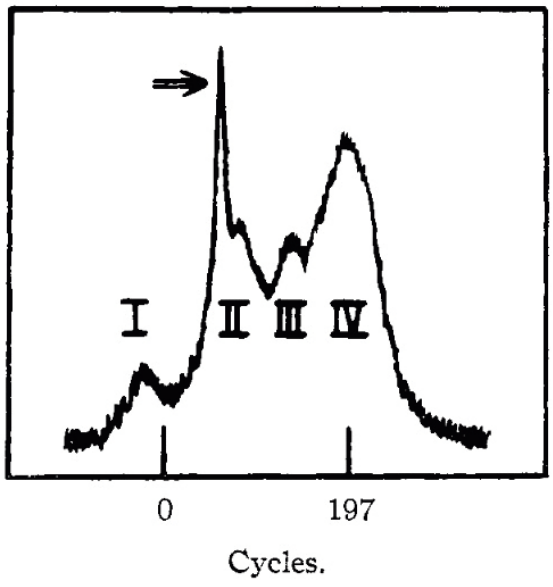

Figura 1. Espectro de resonancia magnética nuclear de una ribonucleasa. Reproducido con permiso de Martin Saunders, Arnold Wishnia, John G. Kirkwood The Nuclear Magnetic Resonance Spectrum of Ribonuclease. *J. Am. Chem. Soc.*, **1957**, *79* (12), pp 3289–3290. Copyright (1957) American Chemical Society.

Secuencias de ADN y síntesis de ADN

La información genética de todas las formas vivientes en la Tierra fluye entre tres biomoléculas: ADN, ARN y proteínas. El ADN es donde se almacena la información genética (excepto por algunos virus que utilizan ARN). El ADN es transcrito como ARN y este es traducido como proteína. Las proteínas cumplen diversas funciones, algunas proveen estructura en un organismo, mientras que otras catalizan reacciones y se les llama enzimas.

La vasta información genética a la que podemos acceder en la actualidad proviene de secuenciar el ADN de diferentes organismos. Las secuencias son depositadas en bancos de datos y de estas secuencias de ADN se pueden inferir las secuencias de proteínas.

Las moléculas de ADN también pueden ser sintetizadas por procedimientos automáticos. Tanto los precios para obtener las secuencias como para sintetizar ADN se han reducido drásticamente. Esta disponibilidad de grandes cantidades de datos, combinada con precios accesibles, ha hecho posible que muchos proyectos que previamente eran considerados demasiado costosos sean ahora viables.

Bioinformática: ¿Cómo estudiar secuencias de proteínas en una computadora?

Las proteínas son cadenas de aminoácidos unidos covalentemente. Los genes (ADN) codifican veinte aminoácidos y cada uno de ellos tiene una abreviatura convencional de tres letras (*Ala* para alanina, *Cys* para cisteína, por ejemplo) y otra abreviatura de una letra (A y C para los dos aminoácidos mencionados); los bioinformáticos generalmente usan la abreviatura de una sola letra. Estas cadenas de letras pueden ser comparadas con otras cadenas de letras de proteínas homólogas que comparten un ancestro común.

Para ilustrar lo que veremos con secuencias de proteínas, tomemos en consideración una analogía para la reconstrucción de palabras que significan “nieve” en diversos idiomas indoeuropeos **(Figura 2)**. La reconstrucción muestra las similitudes entre palabras en idiomas que están relacionados entre sí por tener un ancestro común. Este análisis puede llevar a descubrir palabras que se usaban hace miles de años en antiguas civilizaciones.

¿Qué podemos aprender de esta reconstrucción? La reconstrucción de un lenguaje nos puede decir mucho acerca de una civilización que existió miles de años atrás. Obviamente, vivían en un lugar donde había nieve. Un análisis adicional de los lenguajes indoeuropeos muestra que el lenguaje ancestral no tenía palabras para oro o plata. Sin embargo, tenían palabras para vehículos con ruedas que eran usados para transportar granos. Una reconstrucción similar con secuencias de proteínas puede mostrarnos cómo la dieta de los animales ha cambiado con el tiempo o cómo la capacidad para metabolizar ciertas sustancias ha cambiado con el clima del planeta y, con ello, la temperatura óptima para el funcionamiento de las proteínas.

Las proteínas, una vez puestas en cierto formato para ser usadas por un programa de computadora, pueden ser comparadas entre sí. Inevitablemente, uno encontrará palabras o mensajes escondidos en estas secuencias de 20 letras, ver **Figura 3**.

“*Nada en biología tiene sentido excepto a la luz de la evolución*”[3]

El gen (ADN) que codifica la secuencia de una proteína en particular puede sufrir mutaciones al azar, lo que algunas veces se traduce como un cambio en la secuencia de la proteína. Algunos de estos cambios le confieren una ventaja al organismo y son retenidos por las generaciones futuras. Otros cambios son letales y desaparecen rápidamente de la población. También ocurren cambios neutrales que no confieren ni una ventaja ni una desventaja y estos pueden ser retenidos o no.

Durante 3500 millones de años, desde que la vida ha dejado algún rastro en la Tierra, los genes del último ancestro común se han copiado y pasado de generación en generación. Además, han ocurrido eventos evolutivos en los que se han formado especies diferentes. De esta manera, se crean familias y subfamilias de la misma proteína.

3. Dobzhans.T, Nothing in biology makes sense except in light of evolution. *American Biology Teacher* **1973**, *35* (3), 125-129.

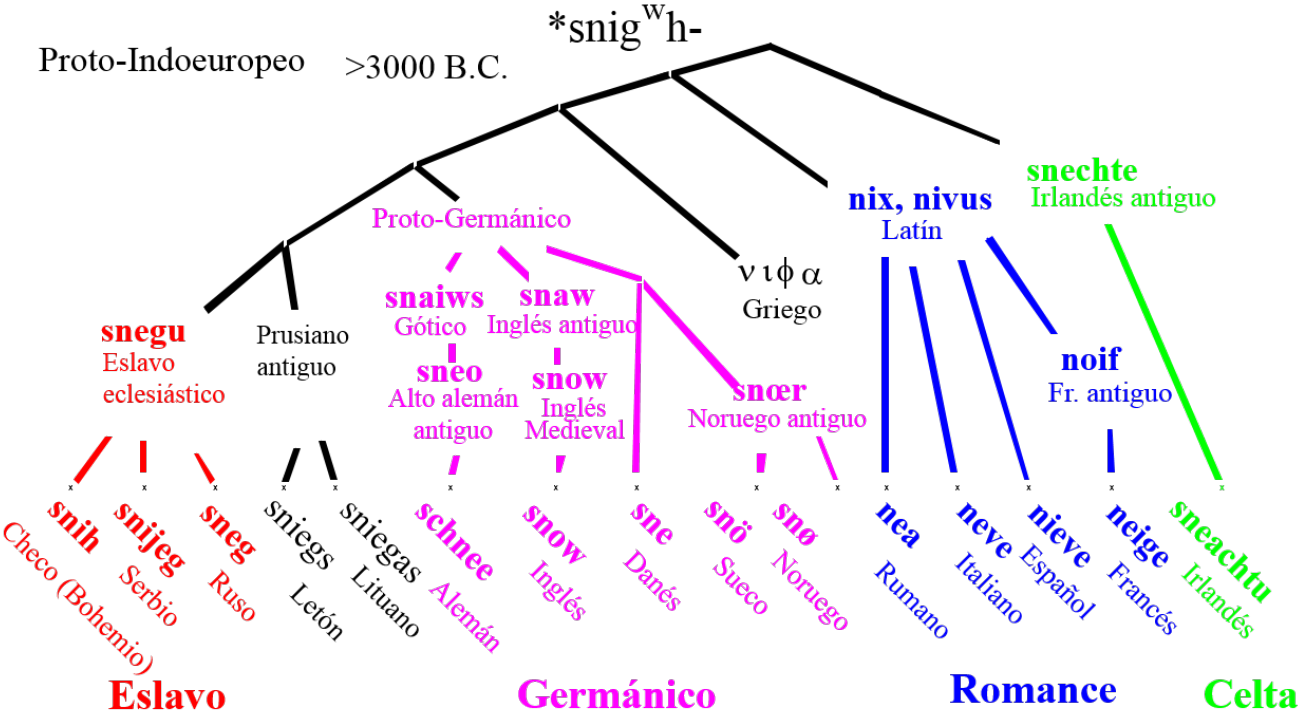

Figura 2. La mayoría de lenguajes europeos descienden de un lenguaje ancestral llamado protoindoeuropeo. Todos los lenguajes modernos tienen una palabra para nieve que es homóloga en todos los lenguajes. Esto quiere decir que las palabras para nieve en los idiomas modernos descienden de una palabra con el mismo significado en el lenguaje ancestral.

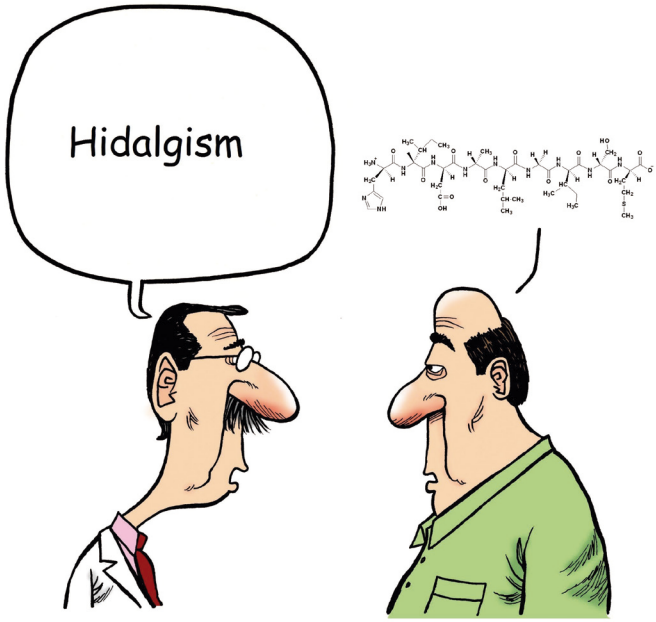

Figura 3. Un químico y un biólogo conversando acerca de proteínas. La secuencia de aminoácidos HIDALGISM[8] fue encontrada en el año 1993 al mismo tiempo en el diccionario no abreviado de Oxford y en el banco de datos de proteínas SwissProt de aquel entonces[9]. Este fue uno de los primeros ejercicios de búsqueda simultánea en dos bases de datos. La palabra "*hidalgism*" (así como la palabra "*ensilsits*") fueron en aquel momento las palabras en inglés más largas que se encontraron en esa búsqueda. Ilustración adaptada de *Life, the Universe and the Scientific Method* por Steven A. Benner © Steven Benner, 2009.

Para comprender la biología necesitamos entender la evolución de las biomoléculas. La historia evolutiva combina tres descripciones. Primero, un árbol evolutivo muestra la relación de las proteínas dentro de una familia. En segundo lugar, un alineamiento de las secuencias de proteínas muestra cómo los aminoácidos de una proteína están relacionados con los aminoácidos de otras proteínas. Finalmente, las secuencias de proteínas ancestrales pueden ser inferidas y son representadas por los puntos que unen a las ramas del árbol filogenético. El árbol, el alineamiento y las secuencias reconstruidas son interdependientes.

Comparando secuencias de proteínas Reconstrucción de proteínas de organismos extintos

Así como en la analogía mostrada en la **Figura 2**, si conocemos las secuencias de proteínas homologas, podemos reconstruir la secuencia de una proteína ancestral de un ancestro común, ya extinto. A continuación se muestran las secuencias (parciales) de la enzima alcohol deshidrogenasa de humano y de caballo*.

```
     5    10    15
EGFDL LRSGK SIRTI LT  caballo
EGFDL LHSGK SIRTI LM  humano
```

* Si usted desea hacer esto en su computadora utilice la fuente Courier New.

Es evidente que esta es la mejor manera de alinear estas dos proteínas: tanto sus similitudes como sus diferencias son evidentes. Podemos añadir más secuencias de alcohol deshidrogenasa de otras especies, para producir un alineamiento múltiple. A continuación se muestra el alineamiento múltiple que incluye, además, la secuencia de una planta (*Arabidopsis thaliana*).

```
    5    10    15
EGFDL LRSGK SIRTI LT  caballo
EGFDL LHSGK SIRTI LM  humano
KAFDY MLKGE SIRCI IT  planta
^ **^ ^  *^ *** * ^
```

Debajo de cada columna está indicado si los aminoácidos son idénticos (*) o si son diferentes pero con cambios conservativos (^). Un cambio conservativo quiere decir que aunque los aminoácidos no sean los mismos, son similares estructuralmente. Por ejemplo, leucina (L) e isoleucina (I) son similares estructuralmente; el ácido aspártico (D) y el ácido glutámico (E) tienen ambos un ácido carboxílico **(Figura 4)**.

La secuencia de la proteína que viene de la planta tiene más diferencias con las dos proteínas de los animales que las secuencias de los dos animales entre sí. Sigue siendo evidente que las secuencias son homólogas. Existen tres posibilidades por las cuales dos proteínas son similares: (1) por accidente: las similitudes aparecieron por azar; (2) por tener un ancestro en común: es posible que ambas enzimas sean descendientes de una enzima ancestral, es decir, son similares por la misma razón por la cual dos hermanos son parecidos. Las diferencias entre ambas representan evolución divergente. (3) Ambas enzimas catalizan la oxidación de alcohol a aldehído y es posible que una proteína que cataliza esta reacción deba tener esta secuencia en particular. Esto implicaría que las dos secuencias aparecieron por evolución convergente.

Veamos ahora un ejemplo para la reconstrucción de una proteína ancestral, definida como la secuencia con la probabilidad más alta de ser la secuencia original de la cual han derivado las secuencias modernas. El caso más sencillo es cuando ambos ancestros tienen el mismo aminoácido en cierta posición. En ese caso se infiere que el ancestro tuvo el mismo aminoácido ya que los aminoácidos en las dos secuencias pueden obtenerse sin reemplazo alguno.

```
5    10    15
EGFDL LRSGK SIRTI LT  caballo
EGFDL LHSGK SIRTI LM  humano
EGFDL L?SGK SIRTI L?  ancestro
```

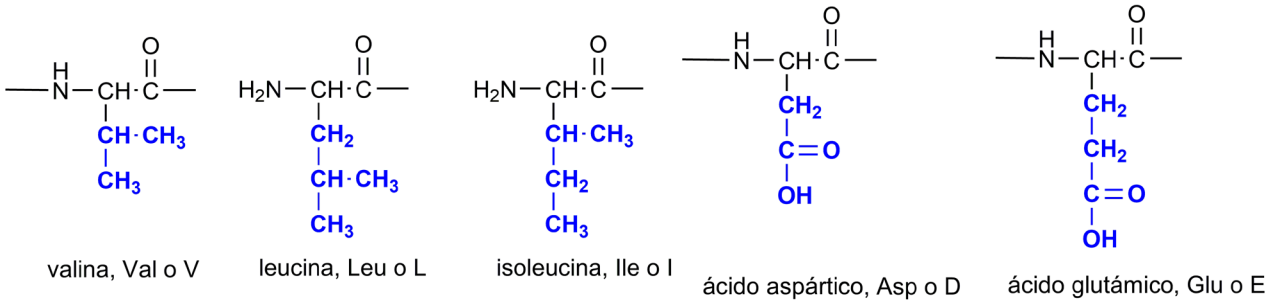

Figura 4. Aminoácidos con cadenas laterales similares (en azul). De izquierda a derecha: valina, leucina e isoleucina tienen cadenas laterales apolares; tanto el ácido aspártico como el ácido glutámico tienen s un ácido carboxílico, un grupo hidrofílico, como cadena lateral. Cuando un aminoácido es reemplazado por otro de estructura similar el cambio se considera conservativo.

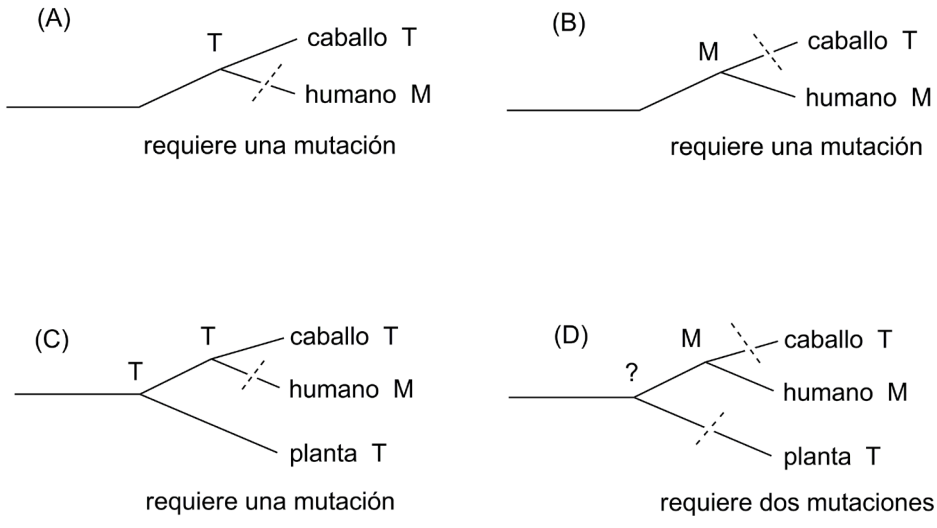

Figura 5. Reconstrucción de proteínas ancestrales en la posición número 17 del alineamiento mostrado en el texto. (a) y (b) muestran dos reconstrucciones posibles para el ancestro común para humano y caballos en la posición 17. Ambas tienen igual probabilidad (el ancestro podría tener una T o M en esta posición) por lo que las dos reconstrucciones requieren una sola mutación indicada por la línea punteada. (c) Al introducir la secuencia correspondiente a la planta, la reconstrucción para el ancestro común en caballos y humanos seria T. Esta reconstrucción requiere una sola mutación. (c) y (d) muestran otras dos posibles reconstrucciones pero incluye información adicional (el aminoácido correspondiente a la planta). Si la reconstrucción del ancestro común para humano y caballo fuese M, entonces se requerirían dos mutaciones indicadas por las líneas punteadas. La reconstrucción mostrada en (c) es mejor que la mostrada en (d).

Consideremos ahora la posición 17 en el alineamiento mostrado arriba. La enzima humana tiene una M (metionina) en esa posición; la enzima de caballo tiene una T (treonina). Si tuviésemos solamente esta información, el ancestro tendría la misma posibilidad de tener M o T en esta posición. Ambas reconstrucciones requerirían tener una sola mutación para explicar las secuencias modernas **(ver Figura 5A y 5B)**. Sin embargo, si añadimos la secuencia de la planta (la cual tiene una T en esta posición) entonces la reconstrucción nos da un árbol filogenético que solo requiere de una mutación, mientras que la reconstrucción con M nos da un árbol que requeriría dos mutaciones **(ver Figura 5C y 5D)**.

Podemos extender este ejemplo introduciendo la secuencia correspondiente a la deshidrogenasa de levadura: en este caso, el alineamiento múltiple sería el mostrado debajo.

```
    5    10    15
EGFDL LRSGK SIRTI LT  caballo
EGFDL LHSGK SIRTI LM  humano
KAFDY MLKGE SIRCI IT  planta
EIYEK MEKGQIVGRYV VD  levadura
```

Entonces, podemos tratar de reconstruir el ancestro en la posición 8. Nos encontramos con dos posibles reconstrucciones, una de ellas es más probable que la otra, como se muestra en la **Figura 6**.

La reconstrucción de proteínas de organismos extintos ha creado una disciplina llamada paleogenética. Nuestro grupo de investigación ha hecho una serie de estudios con proteínas de organismos extintos. Una vez que las secuencias de interés han sido reconstruidas (como proteínas) sintetizamos una secuencia de ADN que codifica la proteína de interés, después se inserta en un vector y se expresa en una bacteria, generalmente en *E. coli*. La proteína purificada obtenida puede ser manipulada y estudiada en el laboratorio[4].

Hemos visto de manera simplificada como un alineamiento múltiple de secuencias puede ser usado para inferir las secuencias de proteínas ya extintas. Examinando fragmentos pequeños y comparando no más de cuatro especies, podemos decidir si un alineamiento es mejor que otro o si un árbol es mejor que otro. Sin embargo, la mayoría de familias de proteínas pueden contener cientos de proteínas. Es entonces cuando las computadoras se hacen necesarias: el alineamiento, el árbol y la mejor reconstrucción de las proteínas ancestrales deben ser inferidos por algoritmos implementados por programas de computadora. Estos programas le asignan un puntaje a cada alineamiento y a cada árbol filogenético. La pionera en establecer un método para asignar puntajes a los alineamientos múltiples es Margaret Dayhoff[5]. El lector interesado en más detalles en este campo puede consultar su trabajo. Otro buen punto de partida es el libro "Of URFS and ORFS" de Russell Doolittle.

Patrones de evolución

Los árboles filogenéticos y los alineamientos múltiples de secuencias de proteínas pueden mostrar patrones de interés. Algunos patrones de especial interés son mostrados en la **Figura 7**.

El patrón conocido como "*heterotachy*" (del griego, *heteros*: diferente y *tachy*: velocidad) ha sido sumamente útil en nuestras investigaciones**. Al introducir cambios en las

4. (a) Gaucher, E. A.; Thomson, J. M.; Burgan, M. F.; Benner, S. A., Experimental paleogenomics as a tool to analyze protein function and predict environmental temperature during early life. *Astrobiology* **2002**, *2* (4), 501; (🖳) (b) Sassi, S. O.; Benner, S. A., The resurrection of ribonucleases from mammals: from ecology to medicine. 2007; p 208-224; (c) Gaucher, E. A.; Govindarajan, S.; Ganesh, O. K., Palaeotemperature trend for Precambrian life inferred from resurrected proteins. *Nature* **2008**, *451* (7179), 704-U2; (d) Carrigan, M. A.; Uryasev, O.; Frye, C. B.; Eckman, B. L.; Myers, C. R.; Hurley, T. D.; Benner, S. A., Hominids adapted to metabolize ethanol long before human-directed fermentation. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112* (2), 458-463.

5. Dayhoff, M. O., Origin and evolution of protein superfamilies. *Federation Proceedings* **1976**, *35* (10), 2132-2138.

** Laos *et al.*, **2016**, por publicarse.

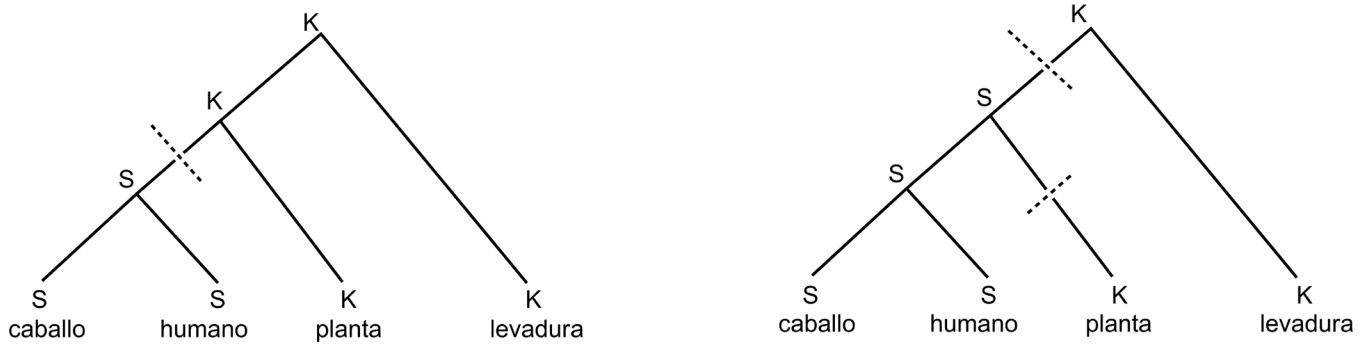

Figura 6. Dos árboles filogenéticos muestran dos posibles reconstrucciones. El árbol de la derecha requiere de dos mutaciones mientras el de la izquierda requiere de una sola mutación (indicadas por la línea punteada). El árbol de la izquierda es una mejor reconstrucción por ser más probable.

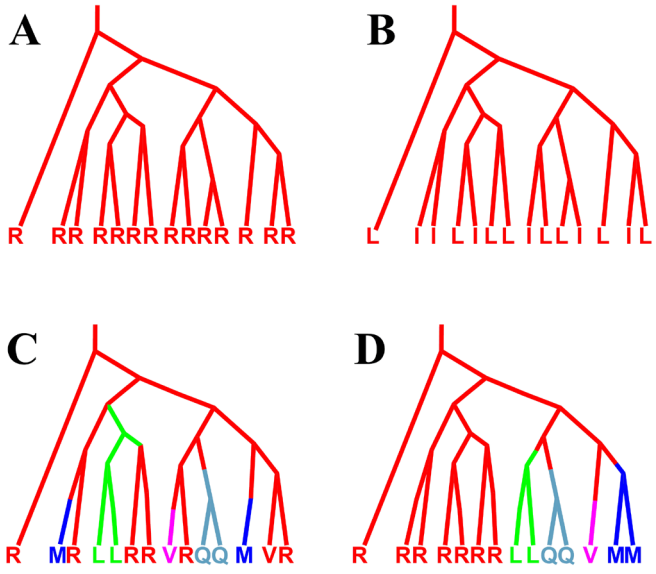

Figura 7. Diferentes patrones de evolución. Esta figura muestra cuatro árboles filogenéticos en los que las letras indican el aminoácido de cada proteína en una posición específica. (**A**) El aminoácido es conservado: en este caso una arginina (R) estáen esta posición en todas las especies presentes en el árbol. Los antepasados que probaron un aminoácido diferente en esta posición están muertos. (**B**) Similar al caso en (A), el sitio es conservado, tiene una leucina (L) o una isoleucina (I) que son muy similares. (**C**) El sitio es altamente variable, este sitio es probablemente irrelevante, pues tolera un amplio rango de aminoácidos. (D) Una subfamilia tiene el sitio altamente conservado mientras que otra tiene alta variabilidad en este sitio. Esto es consecuencia de que dos subfamilias sean expuestas a diferentes presiones evolutivas. Este patrón es llamado *heterotachy*.

secuencias de proteínas en sitios que muestran este patrón de *heterotachy* hemos sido capaces de producir variantes de la enzima ADN polimerasa, que acepta nuevas "letras" de un alfabeto genético artificial llamado AEGIS (*Artificially Expanded Genetic Information System*)[6], desarrollado en nuestros laboratorios. La ventaja de introducir cambios en un lugar que muestra *heterotachy* es que la naturaleza ya ha puesto a prueba estos sitios. Estos son lo suficientemente importantes para influenciar la actividad de la enzima, de modo que podemos notar que el sitio esta conservado en una subfamilia pero puede ser cambiado sin perder la actividad, como se aprecia al mostrar variabilidad en otra subfamilia.

Los ingenieros de proteínas (aquellos que tienen como misión modificar una enzima de modo que tenga una nueva actividad o una actividad mejorada) generalmente se encuentran frente a una tarea colosal dado el tamaño del "*espacio de secuencia*[7]***". Con 20 aminoácidos disponibles para cada posición hay 20^{100} posibilidades para modificar una proteína de solo 100 aminoácidos. Este número es gigantesco y es imposible probar cada una de las posibles variantes. La guía evolutiva puede ayudar a superar este problema.

Visión final

Las diferencias entre la química y la biología vienen de los tiempos en que los químicos sintetizaban pequeñas biomoléculas para confirmar su estructura mientras que los biólogos estudiaban sistemas biológicos retirando partes del sistema e infiriendo un entendimiento de dicho sistema a partir de la perturbación del mismo. Hoy en día, la biología sintética, un buen ejemplo de la intersección de la química y la biología, tiene una meta más alta, la cual es incluir y/o reemplazar biomoléculas naturales por sus análogos artificiales. El reciente interés por la biología sintética puede ser ilustrado por la creación, en 2012, de una nueva revista por parte de la Sociedad Americana de Química (ACS) dedicada a esta área.

En este artículo hemos tocado algunos temas de manera simplificada buscando orientar a los jóvenes estudiantes de química a ver la biología desde el punto de vista evolutivo. El mensaje para el estudiante de química es: "*nada en biología tiene sentido excepto a la luz de la evolución*[3]".

Agradecimientos

Este proyecto y su publicación han sido posibles gracias al apoyo de la Templeton World Charity Foundation, Inc. 0092/AB57, la Defense Threat Reduction Agency (HDTRA1-13-1-0004), y por NASA: NNX14AK37G y NNX15AF46G. Las opiniones expresadas en esta publicación son las de los autores y no necesariamente reflejan las opiniones de la Templeton World Charity Foundation, Inc., o de la National Aeronautics and Space Administration, o del Department of Defense. R. Laos también desea agradecer a Eliana Esparza por las sugerencias ofrecidas mientras se preparaba este artículo.

6. Geyer, C. R.; Battersby, T. R.; Benner, S. A., Nucleobase pairing in Watson-Crick-like genetic expanded information systems. *Structure* **2003**, *11* (12), 1485-1498.

7. Smith, J. M., Natural selection and concept of a protein space. *Nature* **1970**, *225* (5232), 563-564.

*** Del ingles: protein sequence space.

8. Gonnet, G. H.; Benner, S. A., A word in your protein. *Nature* **1993**, *361* (6408), 121. (🖳)

9. Bairoch, A.; Boeckmann, B., The Swiss-prot protein-sequence data-bank. *Nucleic Acids Research* **1992**, *20*, 2019-2022.

BIBLIOGRAFÍA ESENCIAL

Doolittle, R: "*Of URFS and ORFS. A Primer on How to Analyze Derived Amino Acid Sequences*", University Science Books: Mill Valley, 1986.