

RECONSTRUYENDO PROTEÍNAS ANCESTRALES MEDIANTE BIOLOGÍA COMPUTACIONAL

Miguel Arenas Busto

*Profesor Titular Universidad e Investigador principal. Dpto. de Bioquímica, Genética e Inmunología,
y CINBIO. Universidad de Vigo*

Ugo Bastolla Bufalini

Científico titular. Centro de Biología Molecular Severo Ochoa (CBMSO). CSIC

RESUMEN

Las proteínas constituyen componentes esenciales de la vida y, al igual que el material genético, están sujetas a procesos evolutivos de cambio. Como consecuencia, las proteínas existentes hace millones de años pueden diferir ampliamente de sus homólogas actuales, presentando propiedades estructurales y funcionales que se han modificado con el paso del tiempo. En las últimas décadas, la reconstrucción de proteínas ancestrales ha despertado un creciente interés debido a sus aplicaciones en biotecnología y biomedicina, así como a satisfacer nuestra curiosidad sobre los organismos ancestrales. Una etapa fundamental de este proceso es la inferencia computacional de secuencias proteicas ancestrales a partir de proteínas actuales. En este artículo se explora el uso de la biología computacional para inferir proteínas ancestrales, con sus limitaciones y desafíos, y sus principales aplicaciones.

1. INTRODUCCIÓN

Las proteínas pertenecientes a organismos que habitaron el planeta hace millones de años pueden presentar propiedades diferentes a sus proteínas homólogas observadas en organismos actuales debido a procesos evolutivos derivados de selección natural en entornos cambiantes. Estas diferentes propiedades, que presentan potenciales aplicaciones en diferentes áreas, han motivado la reconstrucción de proteínas ancestrales. De esta forma, la inferencia de proteínas ancestrales y su posterior síntesis en laboratorio para analizar sus propiedades es, hoy en día, una aplicación fascinante de la biología evolutiva (Chang y Donoghue 2000; Liberles 2007; Merkl y Sterner 2016a).

Los fundamentos del estudio de proteínas ancestrales se basan en la comparación de secuencias de proteínas de organismos actuales para inferir las secuencias ancestrales homólogas a las proteínas actuales, y la síntesis de las proteínas ancestrales en el laboratorio. Esta estrategia fue originalmente diseñada por Pauling y Zuckerkandl (1963), y las primeras reconstrucciones de proteínas ancestrales se centraron en lisozimas (Malcom, y col. 1990) y ribonucleasas (Stackhouse, y col. 1990; Jermann, y col. 1995). Estos estudios iniciales mostraron que las proteínas ancestrales no solo son útiles para adquirir conocimiento sobre el origen y la evolución de la vida, sino también para resolver problemas del mundo actual. Ejemplos prácticos incluyen la reconstrucción de paleoenzimas pertenecientes a diversos organismos extintos, incluyendo organismos ancestrales con estilo de vida termófilo (Miyazaki, y col. 2001; Gaucher, y col. 2003; Iwabata, y col. 2005), y diversas proteínas tales como tiorredoxinas (Perez-Jimenez, y col. 2011), betalactamasas (Risso, y col. 2013), chaperoninas (Severino, y col. 2025), endonucleasas (Alonso-Lerma, y col. 2023), Rubisco (Shih, y col. 2016) y alcohol deshidrogenasas

(Thomson, y col. 2005), entre otras (Merkl y Sterner 2016a). Algunas de estas enzimas ancestrales resucitadas se han utilizado en procesos industriales (Thomson, y col. 2005; Yamashiro, y col. 2010; Alcalde 2015) debido a su elevada termoestabilidad (Trudeau, y col. 2016).

Además, proteínas ancestrales de virus también han sido reconstruidas (dentro en una escala temporal mucho más corta) y aplicadas en biomedicina para desarrollar vacunas. El caso más conocido es el de las vacunas basadas en secuencias centralizadas para el virus de la inmunodeficiencia humana (VIH) (Gaschen, y col. 2002; Gao, y col. 2003; Doria-Rose, y col. 2005; Kothe, y col. 2006). La idea es que las proteínas ancestrales del virus, reconstruidas a partir de una población de virus, presenten propiedades inmunogénicas a la vez que minimicen la distancia genética con sus variantes descendientes, proporcionando así protección frente a tales variantes descendientes. Sin embargo, en la práctica, estas vacunas fueron solo parcialmente efectivas (Kothe, y col. 2006; Kothe, y col. 2007) debido a varias causas que incluyen la extremadamente elevada tasa de evolución del VIH y errores metodológicos en la reconstrucción de las proteínas ancestrales (Arenas y Posada 2010a), estos factores se detallan más adelante. Además, la reconstrucción de proteínas ancestrales ha sido usada para comprender mejor la evolución molecular, la aparición de resistencia y el desarrollo de terapias (Wilson, y col. 2015; Zakas, y col. 2017). Por lo tanto, la reconstrucción de proteínas ancestrales no solo permite comprender las consecuencias de los procesos evolutivos (Schmitt, y col. 2007), incluyendo cómo evolucionan nuevas funciones, como se adaptan las proteínas a cambios ambientales y que características tenían los organismos ancestrales, sino que también nos proporciona aplicaciones de utilidad práctica.

Antes de reconstruir y caracterizar las proteínas ancestrales en el laboratorio (Gaucher 2007), éstas tienen que ser previamente diseñadas de forma computacional. Concretamente, las secuencias de las proteínas ancestrales se deben inferir usando modelos evolutivos implementados en herramientas computacionales. Por supuesto, estas inferencias incluyen errores derivados del uso de modelos y métodos de evolución molecular, que como todos los modelos y métodos contienen sesgos y limitaciones cuando son comparados con los procesos evolutivos reales. Sin embargo, el uso de modelos apropiados permite inferir secuencias ancestrales más realistas que cuando éstos son ignorados, permitiendo identificar, en la medida de lo posible, las propiedades biológicas de las proteínas ancestrales que existieron en la realidad. A continuación, describimos las metodologías actuales para reconstruir computacionalmente proteínas ancestrales, incluyendo avances y aspectos que se pueden mejorar.

2. PRINCIPALES METODOLOGÍAS COMPUTACIONALES PARA LA RECONSTRUCCIÓN DE PROTEÍNAS ANCESTRALES

Considerando que las proteínas evolucionan a lo largo del tiempo, del pasado hacia el presente, se han desarrollado metodologías estadísticas para la reconstrucción de secuencias de proteínas ancestrales que esencialmente recorren el camino evolutivo de forma temporalmente inversa (Figura 1).

Los métodos de reconstrucción de secuencias ancestrales (ASR, del inglés *ancestral sequence reconstruction*) requieren un alineamiento de secuencias genéticas (ADN o proteínas, secuencias modernas indicadas en verde en el ejemplo de la Figura 1) y una historia evolutiva, generalmente un árbol filogenético que incluye nodos internos (donde las ramas divergen del pasado hacia el presente) enraizado con una secuencia que conecta con el nodo raíz (en el ejemplo de la Figura 1, sería la secuencia “ACRSPH”).

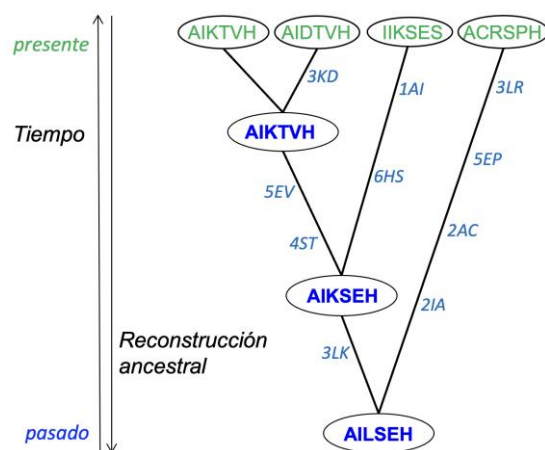


Figura 1. Historia evolutiva de una muestra de secuencias aminoacídicas. Para un alineamiento de cuatro secuencias homólogas (con un ancestro común) de proteínas observadas en el presente (representadas en verde, en este ejemplo cada secuencia contiene seis amino ácidos), se presenta su historia evolutiva y la secuencia de cada nodo ancestral (representada en azul), así como los eventos de reemplazamiento de amino ácidos ocurridos a lo largo del tiempo en las diferentes ramas de la historia evolutiva (representados en azul más claro e indicando posición del reemplazamiento y amino ácidos involucrados). Mientras que el proceso evolutivo real tuvo lugar del pasado hacia el presente, la reconstrucción de las secuencias ancestrales puede entenderse del presente hacia el pasado.

Para realizar la reconstrucción de proteínas ancestrales se debe llevar a cabo las siguientes etapas: (1) selección de un conjunto de secuencias actuales que representen la diversidad observada de la proteína de interés, (2) obtener el alineamiento múltiple de las secuencias, (3) inferencia de un árbol filogenético enraizado a partir de estas secuencias, y (4) realizar la ASR. Los métodos para realizar la ASR, que describimos brevemente a continuación (véase Liberles 2007; Joy, y col. 2016 para detalles teóricos), reconstruyen las secuencias de los nodos ancestrales (las secuencias indicadas en azul en el ejemplo de la Figura 1). Estos métodos suelen implementarse en herramientas computacionales, que también indicamos más adelante.

Se han desarrollado tres métodos estadísticos principales para realizar la ASR: máxima parsimonia (MP), que modela la evolución hacia el pasado mediante el número mínimo de eventos de sustitución; máxima verosimilitud (ML, del inglés *maximum likelihood*), que modela la evolución considerando los eventos de sustitución de amino ácidos más probables de acuerdo con un modelo de sustitución; y la inferencia bayesiana (BI, del inglés *Bayesian inference*), que amplía el método de máxima verosimilitud incorporando incertidumbre asociada a ciertos parámetros evolutivos. Numerosos estudios basados en datos simulados y en datos reales han mostrado que el método de ML proporciona reconstrucciones ancestrales más exactas que el método de MP (Bull, y col. 1993; Zhang y Nei 1997; Randall, y col. 2016), mientras que el método de BI ha mejorado los resultados obtenidos por el método de ML (Williams, y col. 2006; Randall, y col. 2016).

2.1. Reconstrucción de secuencias ancestrales basada en el método de máxima parsimonia

El método de MP es uno de los primeros métodos de ASR desarrollados (Fitch 1971; Sankoff 1975) y se basa en minimizar el número de sustituciones a lo largo de un árbol filogenético dado. Para proteínas, la reconstrucción de alguno de los veinte aminoácidos en cada nodo ancestral se realiza en dos etapas. En una primera etapa, para una posición dada, los aminoácidos observados en las secuencias terminales (o más modernas) se utilizan para inferir los aminoácidos de sus nodos ancestrales. De esta forma, los aminoácidos asignados a un nodo interno provienen de la evaluación de los aminoácidos presentes en sus nodos descendientes, por lo que un nodo interno puede quedar asociado a uno o varios aminoácidos, dependiendo de si los nodos descendientes presentan aminoácidos idénticos o diferentes, respectivamente. En una segunda etapa, el árbol filogenético se recorre en la dirección contraria, del pasado hacia el presente, con el fin de determinar los aminoácidos de cada nodo ancestral.

En este proceso, se resuelven los nodos donde se había asignado más de un aminoácido considerando que los aminoácidos asignados a un nodo ancestral se priorizan para sus nodos descendientes. Sin embargo, todavía podrían encontrarse situaciones donde un nodo ancestral pueda presentar varios aminoácidos posibles, es decir aminoácidos ancestrales igualmente parsimoniosos. Algunos métodos de MP asumen que todos los aminoácidos tienen la misma probabilidad de ocurrir (Fitch 1971), mientras que otros permiten asignar pesos diferentes a los distintos aminoácidos (Sankoff 1975). Es posible obtener todas las secuencias más parsimoniosas para un nodo ancestral, pero esto presenta el problema de como seleccionar una única secuencia entre varias igualmente plausibles.

Para solucionar estas situaciones, algunos métodos incorporaron criterios adicionales, como favorecer cambios de aminoácido que ocurran en determinados periodos evolutivos (por ejemplo, los más recientes) (Swofford y Maddison 1987). A pesar de que los métodos de MP fueron inicialmente muy usados, en la actualidad se han generalmente abandonado debido a que presentan importantes limitaciones. Por ejemplo, estos métodos ignoran las longitudes de las ramas del árbol, asumiendo que solo ocurre un cambio de aminoácido por posición y rama (mientras que en la realidad pueden existir varios cambios en una rama como se muestra en el ejemplo presentado en la Figura 1), e introducen un sesgo hacia los aminoácidos más frecuentes. Los métodos de ASR basados en probabilidades, como ML y BI, que se desarrollaron posteriormente a los métodos de MP, condujeron a ASR de mayor exactitud (Eyre-Walker 1998; Yang y Nielsen 2000; Holder y Lewis 2003).

2.2. Reconstrucción de secuencias ancestrales basada en los métodos de máxima verosimilitud e inferencia Bayesiana

Los métodos de ASR basados en ML y BI calculan la probabilidad de las posibles secuencias para cada nodo ancestral considerando las secuencias observadas en el presente y un modelo de sustitución de aminoácidos (Arenas 2015; Ferreira, y col. 2026). Es decir, consisten en métodos estocásticos que describen la probabilidad de que un aminoácido ancestral sea sustituido por otro en función del tiempo transcurrido hasta ese evento y la naturaleza de los aminoácidos involucrados. Los modelos de sustitución usados para la inferencia filogenética asumen que cada posición evoluciona independientemente del resto de posiciones de la proteína, lo cual no es cierto, pero es una aproximación actualmente necesaria para que se pueda calcular la probabilidad. El elemento fundamental de un modelo de sustitución es una matriz de tasas de sustitución entre pares de aminoácidos, a la cual se denomina “modelo de sustitución”. Este marco computacional se utiliza para inferir la topología del árbol filogenético y las longitudes (duraciones temporales) de sus ramas, y los errores y sesgos derivados son una fuente importante de errores de la inferencia de secuencias ancestrales, cuya naturaleza y duración dependen de la inferencia filogenética. En este artículo, no nos centramos en la inferencia del árbol filogenético, sino en los métodos para inferir secuencias ancestrales para un árbol filogenético dado.

La incorporación de modelos de sustitución realistas, como los que incorporan variación del patrón evolutivo entre posiciones de la proteína (Galtier 2001; Pupko, y col. 2002; Arenas, y col. 2015) e incluso entre ramas del árbol filogenético (Galtier 2001), han sido clave para obtener secuencias ancestrales reconstruidas con elevada exactitud. No obstante, todavía existe espacio de mejora en el desarrollo y aplicación de modelos de sustitución de aminoácidos, como la consideración de sustitución asimétrica entre aminoácidos (por ejemplo, para una posición dada en la proteína, la tasa de cambio de alanina a serina sería diferente a la tasa de cambio de serina a alanina). El uso de modelos asimétricos requiere la resolución de ecuaciones matemáticas complejas que generalmente se han evitado en estudios sobre evolución molecular (Felsenstein 2004; Yang 2006), a pesar de que estos modelos puedan proporcionar reconstrucciones más realistas (Sianga-Mete, y col. 2025; Tinh, y col. 2025).

La función de verosimilitud usada en los métodos ML y BI consiste en una suma de probabilidades que considera todos los aminoácidos posibles en cada secuencia de nodo ancestral, dadas las secuencias observadas en el presente, la topología y longitudes de rama del árbol filogenético y el modelo de sustitución utilizado. Para una descripción matemática sobre las funciones de verosimilitud

se remite al lector a libros tradicionalmente usados en el área (Felsenstein 2004; Yang 2006, 2014). Brevemente, para el método de ML se han desarrollado dos estrategias principales. La estrategia “conjunta” (*joint*), que estima simultáneamente los aminoácidos de las secuencias de todos los nodos ancestrales buscando la combinación con mayor verosimilitud global (Pupko, y col. 2000). Por el contrario, la estrategia “marginal” identifica el aminoácido más probable para cada nodo ancestral de manera individual (Yang, y col. 1995).

Aunque ambas estrategias pueden teóricamente producir reconstrucciones diferentes, en la práctica suelen generar secuencias ancestrales muy similares (Arenas y Posada 2010b; Arenas, y col. 2017). Seguidamente, a menudo se escogen las secuencias ancestrales con mayor verosimilitud, ignorando generalmente las menos probables. Por último, el método de ML suele asumir valores fijos para ciertos parámetros evolutivos como la tasa de evolución, lo que puede producir error cuando existe incertidumbre en ellos. En cambio, el método de BI permite considerar esta incertidumbre especificando distribuciones para cada parámetro en vez de valores fijos. Los métodos de BI permiten estimar las probabilidades posteriores de los aminoácidos ancestrales considerando conjuntamente incertidumbre en los parámetros del árbol filogenético y del modelo de sustitución, es decir, consideran el ajuste estadístico de todos los posibles árboles y modelos de sustitución con las secuencias del presente (Huelsenbeck y Bollback 2001). Sin embargo, estos métodos presentan elevados costes computacionales, especialmente cuando se analizan muchas secuencias, ya que el número de árboles posibles crece más que exponencialmente con el número de secuencias.

2.3. Herramientas computacionales frecuentemente usadas para la reconstrucción de proteínas ancestrales

Las metodologías de ASR previamente descritas (MP, ML y BI) se han implementado en herramientas computacionales habitualmente usadas por los investigadores. Varias herramientas permiten reconstruir secuencias de proteínas ancestrales usando MP, donde podemos destacar *MEGA* (<https://www.megasoftware.net/>) (Kumar, y col. 2018) y *Mesquite* (<https://www.mesquiteproject.org/>) (Maddison y Maddison 2019) debido a su popularidad derivada de la implementación de un interfaz gráfico que facilita el uso a investigadores no familiarizados con la línea de comando de Linux. Con respecto a la metodología de ML, podemos destacar múltiples herramientas incluyendo *MEGA*, *Hyphy* (<http://www.hyphy.org/>) (Kosakovsky Pond, y col. 2005), *FastML* (<https://fastml.tau.ac.il/>) (Moshe y Pupko 2019), *PhyML* (<https://github.com/stephaneguindon/phyml>) (Guindon, y col. 2010), *RAXML-NG* (<https://codeberg.org/amkozlov/raxml-ng>) (Kozlov, y col. 2019), *PAML* (<https://github.com/abacus-gene/paml>) (Yang 2007) y *ProtASR* (<https://github.com/MiguelArenas/protasr>) (Arenas, y col. 2017; Arenas y Bastolla 2020), esta última fue desarrollada por los autores del presente artículo. Estas herramientas presentan diferentes características.

Algunas presentan una interfaz gráfica de uso amigable para no expertos (por ejemplo, *MEGA* y *FastML*), otras han sido optimizadas para analizar rápidamente grandes cantidades de secuencias mediante el uso de supercomputadores o clústeres computacionales (por ejemplo, *RAXML-NG*), y otras implementan modelos de sustitución avanzados que consideran información evolutiva a partir de la estructura de las proteínas (por ejemplo, *ProtASR*). Es el usuario el que debe escoger la herramienta considerando los datos a analizar y los recursos computacionales disponibles. Con respecto a la ASR de proteínas basada en BI, destacamos las herramientas *MrBayes* (<http://nbisweden.github.io/MrBayes/index.html>) (Huelsenbeck y Ronquist 2001), *PhyloBayes* (<http://www.atgc-montpellier.fr/phylobayes/>) (Lartillot, y col. 2009) y *BEAST* (<http://www.beast2.org/>) (Bouckaert, y col. 2019). Estas tres herramientas son muy populares en biología evolutiva, especialmente *BEAST* debido a que incluye un práctico interfaz gráfico y a que implementa características interesantes como violaciones del reloj molecular (es decir, que la tasa de evolución pueda variar en el tiempo) y muestras recogidas a diferentes tiempos. La ASR con estas herramientas requiere más tiempo que con las herramientas de ML y MP, pero en ciertos casos (por ejemplo, en datos con heterogeneidad del proceso evolutivo en el tiempo) la espera puede valer la pena.

3. MECANISMOS BIOLÓGICOS QUE PUEDEN SESGAR LA RECONSTRUCCIÓN DE SECUENCIAS DE PROTEÍNAS ANCESTRALES

Algunos mecanismos biológicos pueden afectar a la inferencia de las secuencias ancestrales. Dado que las proteínas reconstruidas deberían ser lo más realistas posible para presentar propiedades biológicas, tales como estabilidad y actividad, próximas a las de las proteínas ancestrales reales, es importante tener en cuenta dichos mecanismos en la inferencia computacional.

3.1. Influencia de la diversidad proteica sobre la reconstrucción de secuencias ancestrales

La exactitud de las secuencias ancestrales reconstruidas suele depender del grado de diversidad genética observado en las secuencias del presente. Por ejemplo, un estudio basado en simulaciones mostró que el error de reconstrucción de secuencias aminoacídicas ancestrales se incrementa aproximadamente del 1% al 5% al duplicarse la tasa de sustitución aminoacídica (Arenas y Posada 2010b). Este efecto fue descrito como una consecuencia del uso de modelos de sustitución, el cual es requerido por los métodos más avanzados de reconstrucción de secuencias ancestrales (ML y BI). Cualquier modelo constituye una simplificación de la realidad y, por lo tanto, incorpora cierto grado de error. El error derivado del modelo de sustitución puede acumularse a medida que aumenta la profundidad temporal de la reconstrucción. En consecuencia, la inferencia de secuencias muy antiguas, como paleoenzimas existentes hace miles de millones de años (Merkl y Sterner 2016b), donde el modelo debe aplicarse en mayor medida, puede conllevar a errores considerables. Además, el error de reconstrucción también puede variar según el tipo de molécula estudiada. Moléculas con un mayor número de estados posibles (por ejemplo, los 64 codones en la reconstrucción de secuencias de codones o los 20 aminoácidos en la reconstrucción de proteínas, frente a los 4 nucleótidos en la reconstrucción de ADN) pueden mostrar mayor error en la predicción (Arenas y Posada 2010b).

3.2. Sesgos causados por el modelo de sustitución asumido

Relacionado con el apartado anterior, se sabe que la exactitud de las reconstrucciones ancestrales puede variar entre modelos de sustitución (Lemmon y Moriarty 2004; Del Amparo y Arenas 2022, 2023). Por ejemplo, reconstrucciones ancestrales que consideran variación de la tasa de sustitución entre sitios presentan una mayor exactitud, especialmente en secuencias altamente divergentes (Pupko, y col. 2002) que se encuentran frecuentemente en la naturaleza (Chang, y col. 2002; Thornton, y col. 2003). Asimismo, los modelos de sustitución que incorporan información estructural de las proteínas (véanse los estudios de revisión Liberles, y col. 2012; Echave y Wilke 2017; Bastolla y Arenas 2019; Chisholm, y col. 2024; Ferreira, y col. 2026) permitiendo reconstruir proteínas ancestrales más realistas, en términos de estabilidad y actividad, que aquellos modelos que ignoran las restricciones estructurales (Bordner y Mittelman 2014; Arenas, y col. 2017; Echave 2019; Arenas y Bastolla 2020; Ferreira, y col. 2024). Sin embargo, los modelos de sustitución que consideran restricciones estructurales todavía no están ampliamente implantados en el área debido a la complejidad de incorporar sus algoritmos en funciones tradicionales de verosimilitud filogenética, así como a su mayor coste computacional. Como una excepción, los autores del presente artículo han desarrollado la herramienta computacional *ProtASR* (presentada previamente) que implementa un modelo de sustitución aminoacídica que considera información estructural de la proteína (Bastolla, y col. 2006; Arenas, y col. 2015) y puede reconstruir proteínas ancestrales con estabilidad de plegamiento más realista que la obtenida mediante los modelos empíricos tradicionalmente usados (Arenas, y col. 2017; Arenas y Bastolla 2020).

3.3. Sesgos causados por errores en el alineamiento de secuencias

El alineamiento múltiple de secuencias es probablemente una de las mayores fuentes de error en los análisis de evolución molecular (Wong, y col. 2008; Hanson-Smith, y col. 2010). Los métodos de ASR asumen que las posiciones homólogas del alineamiento múltiple de secuencias presentan un ancestro común. Sin embargo, los errores de alineamiento pueden violar esta asunción al asignar una incorrecta homología entre posiciones de diferentes secuencias. De esta forma, los aminoácidos

observados en una posición del alineamiento pueden no corresponder a una misma posición ancestral, introduciendo sesgos en la inferencia de secuencias ancestrales. Este sesgo es más habitual en secuencias con elevada divergencia, donde la identificación de posiciones homólogas resulta más complicada. Estos errores conducirían a la estimación errónea de tasas de sustitución (árbol filogenético con incorrectas longitudes de rama) y a la inferencia incorrecta de estados ancestrales (Hanson-Smith, y col. 2010), al comparar posiciones no homólogas como si lo fuesen.

Por lo tanto, es fundamental asegurarse de disponer de un alineamiento de secuencias apropiado. Otra complicación es la reconstrucción de estados ambiguos (más de un aminoácido detectado para una posición dada) y de inserciones y deleciones (indeles). Las ambigüedades en el alineamiento de secuencias suelen eliminarse antes del análisis, siendo así ignoradas. Además, los modelos de sustitución tradicionalmente usados para inferencia evolutiva no suelen incorporar indeles o se consideran como un estado molecular adicional. Se requiere más investigación en el desarrollo de modelos y métodos de ASR que consideren la incertidumbre en la calidad del alineamiento de secuencias y en el tratamiento de las ambigüedades e indeles.

3.4. Sesgos causados por la recombinación genética

La recombinación genética, donde se intercambia material genético entre dos moléculas de ADN, es una fuerza evolutiva fundamental presente en numerosos organismos (Modesti y Kanaar 2001; Perez-Losada, y col. 2015) y debe tenerse en cuenta en los análisis evolutivos. Por ejemplo, la reconstrucción de árboles filogenéticos, que es requerida para la reconstrucción de secuencias ancestrales, puede verse afectada por la presencia de recombinación, generando árboles con topologías y longitudes de rama incorrectas (Schierup y Hein 2000). Esto se debe a que los fragmentos recombinantes que se intercambian pueden haber evolucionado bajo historias evolutivas diferentes [de hecho, numerosos métodos de detección de recombinación se basan en esta incongruencia filogenética (Martin, y col. 2011; Arenas 2021)].

Como consecuencia, el uso de un árbol filogenético incorrecto derivado de recombinación puede conducir a la inferencia errónea de secuencias ancestrales (Arenas y Posada 2010b). Así, cuanto mayor es la tasa de recombinación (en presencia de suficiente diversidad genética para observar el efecto de la recombinación), mayor es el error en las secuencias ancestrales reconstruidas (Arenas y Posada 2010b). Por lo tanto, cuando existe recombinación, ésta debería considerarse en la metodología de ASR.

Una estrategia para realizar ASR teniendo en cuenta la recombinación consiste en reconstruir las secuencias ancestrales de forma independiente para cada fragmento recombinante, utilizando el árbol filogenético inferido para el fragmento bajo estudio y, posteriormente, concatenar los fragmentos de secuencias ancestrales (Arenas y Posada 2010b). Este proceso requiere: (1) Identificar los puntos de corte recombinación a lo largo de las secuencias [puede realizarse con herramientas computacionales tales como *RDP* (Martin, y col. 2021) y *LDhat* (McVean, y col. 2004), entre otras]; (2) Inferir un árbol filogenético para cada fragmento recombinante (puede realizarse con herramientas computacionales tales como *PhyML* y *RAxML-NG*, entre otros); (3) Reconstruir las secuencias ancestrales de cada fragmento recombinante usando su correspondiente árbol filogenético. Este procedimiento de tres etapas se ha implementado en la herramienta computacional *HyPhy*.

Seguidamente, si las historias evolutivas de los fragmentos recombinantes comparten nodos ancestrales, los fragmentos de secuencia reconstruidos para dichos nodos pueden concatenarse siguiendo el orden definido por los puntos de corte de recombinación. Sin embargo, los fragmentos correspondientes a nodos ancestrales que no son compartidos entre los distintos árboles de fragmentos recombinantes no pueden combinarse ya que pertenecen a tiempos y trayectorias evolutivas diferentes. Considerando la asunción de que todas las secuencias observadas provienen de un ancestro común, más o menos alejado en el tiempo, tuvo que existir un ancestro común más reciente a todos los fragmentos recombinantes, el cual se denomina habitualmente como “gran ancestro común más reciente” (*grand most recent common ancestor*, GMRCA) (Griffiths y Marjoram 1997; Arenas 2013). El GMRCA

contendría material genético de todos los fragmentos recombinantes, por lo que podría ser usado para obtener una secuencia ancestral completa a todos los fragmentos recombinantes (Arenas y Posada 2010b). No obstante, el desarrollo de métodos de ASR que permitan inferir directamente la secuencia del GMRCA constituye una línea de investigación pendiente en el área debido a su complejidad.

4. PERSPECTIVAS SOBRE EL FUTURO DE LA RECONSTRUCCIÓN COMPUTACIONAL DE PROTEÍNAS ANCESTRALES

La reconstrucción de proteínas ancestrales constituye una de las aplicaciones más interesantes de la biología evolutiva para comprender las propiedades biológicas de proteínas pertenecientes a organismos del pasado. En la actualidad, ya se han reconstruido cientos de proteínas ancestrales (Carletti, y col. 2020) y se mantiene como un tópico relevante de investigación (Alonso-Lerma, y col. 2023; Chisuga, y col. 2025; Severino, y col. 2025). Sin embargo, la predicción computacional de proteínas ancestrales con propiedades realistas, que no sean artefactos derivados de asunciones en el proceso metodológico (como las que se han presentado previamente), es esencial. Hemos encontrado numerosos estudios de ASR de proteínas donde se utilizan modelos de sustitución demasiado simplistas o donde se ignoran aspectos fundamentales como encontrar homología apropiada entre secuencias. No obstante, se están realizando esfuerzos para mejorar la exactitud de los modelos y métodos de ASR, tales como el desarrollo de modelos de sustitución avanzados, la consideración de indeles y el uso de redes filogenéticas de recombinación. Todavía existe un amplio margen de mejora. Por ejemplo, sería interesante incorporar presión selectiva variable en el tiempo. En este sentido, una aproximación consiste en realizar una reconstrucción ancestral más realista usando directamente la red ancestral de selección (*ancestral selection graph*, ASG) (Neuhauser y Krone 1997). Estos desarrollos metodológicos presentan importantes desafíos teóricos y computacionales para el área, pero podrían contribuir de forma significativa a mejorar la exactitud de las proteínas ancestrales reconstruidas. También es posible que en el futuro la inteligencia artificial pueda mejorar las metodologías actuales de ASR. Por el momento parece que podría ayudar en algunas etapas de la ASR o en particularidades tales como la consideración de epistasis en los modelos de evolución (De Leonardis, y col. 2025). En total, el futuro de la reconstrucción de secuencias de proteínas ancestrales pasa por la integración de procesos evolutivos cada vez más realistas dentro de los modelos y métodos de inferencia. Estos avances nos permitirán comprender mejor la evolución molecular y las características de los organismos del pasado a nivel molecular.

FINANCIACIÓN Y APOYOS: *La realización de este artículo ha sido posible mediante los Proyectos PID2023-151032NB-C21 y PID2023-151032NB-C22 financiados por MCIU/AEI/10.13039/501100011033 y por FEDER, UE.*

5. REFERENCIAS BIBLIOGRÁFICAS

- Alcalde M. 2015. Engineering the ligninolytic enzyme consortium. *Trends in Biotechnology* 33:155-162.
- Alonso-Lerma B, Jabalera Y, Samperio S, Morin M, Fernandez A, Hille LT, Silverstein RA, Quesada-Ganuza A, Reifs A, Fernandez-Penalver S, y col. 2023. Evolution of CRISPR-associated endonucleases as inferred from resurrected proteins. *Nat Microbiol* 8:77-90.
- Arenas M. 2021. Computational Analysis of Recombination in Viral Nucleotide Sequences. In: Bamford D, Zuckerman M, editors. *Encyclopedia of Virology*: Academic Press (Elsevier). p. In press.
- Arenas M. 2013. The Importance and Application of the Ancestral Recombination Graph. *Front Genet* 4:206.
- Arenas M. 2015. Trends in substitution models of molecular evolution. *Front Genet* 6:319.
- Arenas M, Bastolla U. 2020. ProtASR2: Ancestral reconstruction of protein sequences accounting for folding stability. *Methods Ecol Evol* 11:248-257.
- Arenas M, Posada D. 2010a. Computational Design of Centralized HIV-1 Genes. *Curr HIV Res* 8:613-621.

- Arenas M, Posada D. 2010b. The effect of recombination on the reconstruction of ancestral sequences. *Genetics* 184:1133-1139.
- Arenas M, Sanchez-Cobos A, Bastolla U. 2015. Maximum likelihood phylogenetic inference with selection on protein folding stability. *Molecular Biology and Evolution* 32:2195-2207.
- Arenas M, Weber CC, Liberles DA, Bastolla U. 2017. ProtASR: An Evolutionary Framework for Ancestral Protein Reconstruction with Selection on Folding Stability. *Systematic Biology* 66:1054-1064.
- Bastolla U, Arenas M. 2019. The Influence of Protein Stability on Sequence Evolution: Applications to Phylogenetic Inference. In: Sikosek T, editor. *Computational Methods in Protein Evolution*. New York, NY: Springer New York. p. 215-231.
- Bastolla U, Porto M, Roman HE, Vendruscolo M. 2006. A protein evolution model with independent sites that reproduces site-specific amino acid distributions from the Protein Data Bank. *BMC Evolutionary Biology* 6:43.
- Bordner AJ, Mittelman HD. 2014. A new formulation of protein evolutionary models that account for structural constraints. *Molecular Biology and Evolution* 31:736-749.
- Bouckaert R, Vaughan TG, Barido-Sottani J, Duchene S, Fourment M, Gavryushkina A, Heled J, Jones G, Kuhnert D, De Maio N, y col. 2019. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Comput Biol* 15:e1006650.
- Bull JJ, Cunningham CW, Molineux IJ, Badgett MR, Hillis DM. 1993. Experimental molecular evolution of bacteriophage T7. *Evolution* 47:993-1007.
- Carletti MS, Monzon AM, Garcia-Rios E, Benitez G, Hirsh L, Fornasari MS, Parisi G. 2020. Revenant: a database of resurrected proteins. Database 2020.
- Chang BS, Donoghue MJ. 2000. Recreating ancestral proteins. *Trends in Ecology & Evolution* 15:109-114.
- Chang BS, Jonsson K, Kazmi MA, Donoghue MJ, Sakmar TP. 2002. Recreating a functional ancestral archosaur visual pigment. *Molecular Biology and Evolution* 19:1483-1489.
- Chisholm LO, Orlandi KN, Phillips SR, Shavlik MJ, Harms MJ. 2024. Ancestral Reconstruction and the Evolution of Protein Energy Landscapes. *Annu Rev Biophys* 53:127-146.
- Chisuga T, Takinami S, Liao Z, Karasawa M, Adachi N, Kawasaki M, Moriya T, Senda T, Terada T, Kudo F, y col. 2025. Ancestral sequence reconstruction as a tool for structural analysis of modular polyketide synthases. *Nat Commun* 16:6847.
- De Leonardis M, Pagnani A, Barrat-Charlaix P. 2025. Reconstruction of Ancestral Protein Sequences Using Autoregressive Generative Models. *Molecular Biology and Evolution* 42.
- Del Amparo R, Arenas M. 2022. Consequences of Substitution Model Selection on Protein Ancestral Sequence Reconstruction. *Molecular Biology and Evolution* 39:msac144.
- Del Amparo R, Arenas M. 2023. Influence of substitution model selection on protein phylogenetic tree reconstruction. *Gene* 865:147336.
- Doria-Rose NA, Learn GH, Rodrigo AG, Nickle DC, Li F, Mahalanabis M, Hensel MT, McLaughlin S, Edmonson PF, Montefiori D, y col. 2005. Human immunodeficiency virus type 1 subtype B ancestral envelope protein is functional and elicits neutralizing antibodies in rabbits similar to those elicited by a circulating subtype B envelope. *Journal of Virology* 79:11214-11224.
- Echave J. 2019. Beyond Stability Constraints: A Biophysical Model of Enzyme Evolution with Selection on Stability and Activity. *Molecular Biology and Evolution* 36:613-620.
- Echave J, Wilke CO. 2017. Biophysical Models of Protein Evolution: Understanding the Patterns of Evolutionary Sequence Divergence. *Annu Rev Biophys* 46:85-103.
- Eyre-Walker A. 1998. Problems with Parsimony in Sequences of Biased Base Composition. *Journal of Molecular Evolution* 47:686-690.
- Felsenstein J. 2004. *Inferring phylogenies*. MA: Sinauer Associates: Sunderland.
- Ferreiro D, Khalil R, Sousa SF, Arenas M. 2024. Substitution Models of Protein Evolution with Selection on Enzymatic Activity. *Molecular Biology and Evolution* 41:msae026.
- Ferreiro D, Pazos E, Arenas M. 2026. Trends in substitution models of protein evolution for phylogenetic inference. *Molecular Phylogenetics and Evolution* 214:108473.

- Fitch W. 1971. Toward defining the course of evolution: minimal change for a specific tree topology. *Systematic Zoology* 20:406-416.
- Galtier N. 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. *Molecular Biology and Evolution* 18:866-873.
- Gao F, Bhattacharya T, Gaschen B, Taylor J, Moore JP, Novitsky V, Yusim K, Lang D, Foley B, Beddows S, y col. 2003. Consensus and ancestral state HIV vaccines. *Science* 299:1515-1518.
- Gaschen B, Taylor J, Yusim K, Foley B, Gao F, Lang D, Novitsky V, Haynes B, Hahn BH, Bhattacharya T, y col. 2002. Diversity considerations in HIV-1 vaccine selection. *Science* 296:2354-2360.
- Gaucher EA. 2007. Experimental resurrection of ancient biomolecules: gene synthesis, heterologous protein expression, and functional assays. In: Liberles DA, editor. *Ancestral Sequence Reconstruction*. Oxford: Oxford University Press.
- Gaucher EA, Thomson JM, Burgan MF, Benner SA. 2003. Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. *Nature* 425:285-288.
- Griffiths RC, Marjoram P. 1997. An ancestral recombination graph. In: Donnelly P, Tavaré S, editors. *Progress in population genetics and human evolution*. Berlin: Springer-Verlag. p. 257-270.
- Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology* 59:307-321.
- Hanson-Smith V, Kolaczkowski B, Thornton JW. 2010. Robustness of Ancestral Sequence Reconstruction to Phylogenetic Uncertainty. *Molecular Biology and Evolution* 27:1988-1999.
- Holder M, Lewis PO. 2003. Phylogeny estimation: traditional and Bayesian approaches. *Nature Reviews Genetics* 4:275-284.
- Huelsenbeck JP, Bollback JP. 2001. Empirical and hierarchical Bayesian estimation of ancestral states. *Systematic Biology* 50:351-366.
- Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogeny. *Bioinformatics* 17:754-755.
- Iwabata H, Watanabe K, Ohkuri T, Yokobori S-i, Yamagishi A. 2005. Thermostability of ancestral mutants of *Caldococcus noboribetus* isocitrate dehydrogenase. *FEMS Microbiology Letters* 243:393-398.
- Jermann TM, Opitz JG, Stackhouse J, Benner SA. 1995. Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. *Nature* 374:57-59.
- Joy JB, Liang RH, McCloskey RM, Nguyen T, Poon AFY. 2016. Ancestral Reconstruction. *PLOS Computational Biology* 12:e1004763.
- Kosakovsky Pond SL, Frost SD, Muse SV. 2005. HYPHY: Hypothesis testing using phylogenies. *Bioinformatics* 21:676-679.
- Kothe DL, Decker JM, Li Y, Weng Z, Bibollet-Ruche F, Zammit KP, Salazar MG, Chen Y, Salazar-Gonzalez JF, Moldoveanu Z, y col. 2007. Antigenicity and immunogenicity of HIV-1 consensus subtype B envelope glycoproteins. *Virology* 360:218-234.
- Kothe DL, Li Y, Decker JM, Bibollet-Ruche F, Zammit KP, Salazar MG, Chen Y, Weng Z, Weaver EA, Gao F, y col. 2006. Ancestral and consensus envelope immunogens for HIV-1 subtype C. *Virology* 352:438-449.
- Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* 35:4453-4455.
- Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Molecular Biology and Evolution* 35:1547-1549.
- Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* 25:2286-2288.
- Lemmon AR, Moriarty EC. 2004. The importance of proper model assumption in bayesian phylogenetics. *Systematic Biology* 53:265-277.
- Liberles DA. 2007. *Ancestral Sequence Reconstruction*: Oxford University Press.
- Liberles DA, Teichmann SA, Bahar I, Bastolla U, Bloom J, Bornberg-Bauer E, Colwell LJ, de Koning AP, Dokholyan NV, Echave J, y col. 2012. The interface of protein structure, protein biophysics, and molecular evolution. *Protein Science* 21:769-785.

- Maddison WP, Maddison DR. 2019. Mesquite: a modular system for evolutionary analysis. Version 3.61.
- Malcom BA, Wilson KP, Matthews BW, Kirsch JF, Wilson AC. 1990. Ancestral lysozymes reconstructed, neutrality tested, and thermostability linked to hydrocarbon packing. *Nature* 345:86-89.
- Martin DP, Lemey P, Posada D. 2011. Analysing recombination in nucleotide sequences. *Mol Ecol Resour* 11:943-955.
- Martin DP, Varsani A, Roumagnac P, Botha G, Maslamoney S, Schwab T, Kelz Z, Kumar V, Murrell B. 2021. RDP5: a computer program for analyzing recombination in, and removing signals of recombination from, nucleotide sequence datasets. *Virus Evol* 7:veaa087.
- McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. 2004. The fine-scale structure of recombination rate variation in the human genome. *Science* 304:581-584.
- Merkl R, Sterner R. 2016a. Ancestral protein reconstruction: techniques and applications. *Biological Chemistry* 397:1-21.
- Merkl R, Sterner R. 2016b. Reconstruction of ancestral enzymes. *Perspectives in Science* 9:17-23.
- Miyazaki J, Nakaya S, Suzuki T, Tamakoshi M, Oshima T, Yamagishi A. 2001. Ancestral Residues Stabilizing 3-Isopropylmalate Dehydrogenase of an Extreme Thermophile: Experimental Evidence Supporting the Thermophilic Common Ancestor Hypothesis. *The Journal of Biochemistry* 129:777-782.
- Modesti M, Kanaar R. 2001. Homologous recombination: from model organisms to human disease. *Genome Biology* 2:REVIEWS1014.
- Moshe A, Pupko T. 2019. Ancestral sequence reconstruction: accounting for structural information by averaging over replacement matrices. *Bioinformatics* 35:2562-2568.
- Neuhauser C, Krone SM. 1997. The genealogy of samples in models with selection. *Genetics* 145:519-534.
- Pauling L, Zuckerkandl E. 1963. Chemical paleogenetics: molecular "restoration studies" of extinct forms of life. *Act Chem. Scand.* 17:S9-S16.
- Perez-Jimenez R, Ingles-Prieto A, Zhao ZM, Sanchez-Romero I, Alegre-Cebollada J, Kosuri P, Garcia-Manyes S, Kappock TJ, Tanokura M, Holmgren A, y col. 2011. Single-molecule paleoenzymology probes the chemistry of resurrected enzymes. *Nat Struct Mol Biol* 18:592-596.
- Perez-Losada M, Arenas M, Galan JC, Palero F, Gonzalez-Candelas F. 2015. Recombination in viruses: Mechanisms, methods of study, and evolutionary consequences. *Infect Genet Evol* 30C:296-307.
- Pupko T, Pe'er I, Hasegawa M, Graur D, Friedman N. 2002. A branch-and-bound algorithm for the inference of ancestral amino-acid sequences when the replacement rate varies among sites: Application to the evolution of five gene families. *Bioinformatics* 18:1116-1123.
- Pupko T, Pe'er I, Shamir R, Graur D. 2000. A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Molecular Biology and Evolution* 17:890-896.
- Randall RN, Radford CE, Roof KA, Natarajan DK, Gaucher EA. 2016. An experimental phylogeny to benchmark ancestral sequence reconstruction. *Nat Commun* 7:12847.
- Risso VA, Gavira JA, Mejia-Carmona DF, Gaucher EA, Sanchez-Ruiz JM. 2013. Hyperstability and substrate promiscuity in laboratory resurrections of Precambrian beta-lactamases. *Journal of the American Chemical Society* 135:2899-2902.
- Sankoff D. 1975. Minimal mutation trees of sequences. *SIAM Journal on Applied Mathematics* 28:35-42.
- Schierup MH, Hein J. 2000. Consequences of recombination on traditional phylogenetic analysis. *Genetics* 156:879-891.
- Schmitt AO, Schuchhardt J, Ludwig A, Brockmann GA. 2007. Protein evolution within and between species. *Journal of Theoretical Biology* 249:376-383.
- Severino R, Cuellar J, Gutierrez-Seijo J, Maestro-Lopez M, Sanchez-Pulido L, Santiago C, Moreno-Paz M, Valpuesta JM, Parro V. 2025. Ancestral Chaperonins Provide the First Structural Glimpse into Early Multimeric Protein Evolution. *Molecular Biology and Evolution* 42.
- Shih PM, Occhialini A, Cameron JC, Andralojc PJ, Parry MA, Kerfeld CA. 2016. Biochemical characterization of predicted Precambrian RuBisCO. *Nat Commun* 7:10382.

- Sianga-Mete R, Hartnady P, Mandikumba WC, Rutherford K, Currin CB, Phelanyane F, Stefan S, Weaver S, Pond SLK, Martin DP. 2025. Viral genome sequence datasets display pervasive evidence of strand-specific substitution biases that are best described using non-reversible nucleotide substitution models. *eLife* 12:RP87361.
- Stackhouse J, Presnell SR, McGeehan GM, Nambiar KP, Benner SA. 1990. The ribonuclease from an extinct bovid ruminant. *FEBS Letters* 262:104-106.
- Swofford DL, Maddison WP. 1987. Reconstructing ancestral character states under Wagner parsimony. *Mathematical Biosciences* 87:199-229.
- Thomson JM, Gaucher EA, Burgan MF, De Kee DW, Li T, Aris JP, Benner SA. 2005. Resurrecting ancestral alcohol dehydrogenases from yeast. *Nature Genetics* 37:630-635.
- Thornton JW, Need E, Crews D. 2003. Resurrecting the ancestral steroid receptor: ancient origin of estrogen signaling. *Science* 301:1714-1717.
- Tinh NH, Dang CC, Vinh LS. 2025. nT4X and nT4M: Novel Time Non-reversible Mixture Amino Acid Substitution Models. *Journal of Molecular Evolution* 93:136-148.
- Trudeau DL, Kaltenbach M, Tawfik DS. 2016. On the Potential Origins of the High Stability of Reconstructed Ancestral Proteins. *Molecular Biology and Evolution* 33:2633-2641.
- Williams PD, Pollock DD, Blackburne BP, Goldstein RA. 2006. Assessing the accuracy of ancestral protein reconstruction methods. *PLoS Comput Biol* 2:e69.
- Wilson C, Agafonov RV, Hoemberger M, Kutter S, Zorba A, Halpin J, Buosi V, Otten R, Waterman D, Theobald DL, et al. 2015. Kinase dynamics. Using ancient protein kinases to unravel a modern cancer drug's mechanism. *Science* 347:882-886.
- Wong KM, Suchard MA, Huelsenbeck JP. 2008. Alignment uncertainty and genomic analysis. *Science* 319:473-476.
- Yamashiro K, Yokobori S, Koikeda S, Yamagishi A. 2010. Improvement of *Bacillus circulans* beta-amylase activity attained using the ancestral mutation method. *Protein Eng Des Sel* 23:519-528.
- Yang Z. 2006. *Computational Molecular Evolution*. Oxford, England.: Oxford University Press.
- Yang Z. 2014. *Molecular Evolution: A Statistical Approach*. Oxford, England.: Oxford University Press.
- Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* 24:1586-1591.
- Yang Z, Kumar S, Nei M. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* 141:1641-1650.
- Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular Biology and Evolution* 17:32-43.
- Zakas PM, Brown HC, Knight K, Meeks SL, Spencer HT, Gaucher EA, Doering CB. 2017. Enhancing the pharmaceutical properties of protein drugs by ancestral sequence reconstruction. *Nature Biotechnology* 35:35-37.
- Zhang J, Nei M. 1997. Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. *Journal of Molecular Evolution* 44 Suppl 1:S139-146.