Benítez, Guillermo Ignacio

Desarrollo de métodos bioinformáticos para la caracterización estructural y funcional de proteínas ancestrales

Relación sociedad-naturaleza: los efectos del turismo como actividad estructurante del espacio urbano en San Carlos de Bariloche entre el 2001 y el 2018

TESIS DOCTORAL

Gonzalo E. Barrios Garcia

gonzaloebarriosgarcia@gmail.com

Resumen

Existe cierto consenso general sobre el turismo como una industria limpia que resulta relativamente fácil lograr su despliegue territorial y que una vez alcanzado implica grandes beneficios económicos para las localidades donde se realiza. Sin embargo, estas visiones no suelen considerar la cantidad de recursos que insume su viabilidad ni los impactos en el corto y largo plazo tanto para las poblaciones locales como para el ambiente natural. Fueron necesarias diversas políticas materiales y prácticas simbólicas que pusieron el énfasis en los majestuosos paisajes naturales para que ciudades de la Patagonia andina argentina, en general, y la ciudad de San Carlos de Bariloche, en particular, se ligaran casi exclusivamente con el turismo como principal motor de desarrollo. El escenario general de cambios climático y de readecuación del modo de producción capitalista, hacen que las ciudades turísticas dependientes de los atractivos naturales se encuentren transitando transformaciones profundas en una posición de gran vulnerabilidad. A nivel nacional, este proceso es acompañado por un fuerte crecimiento económico posterior a la crisis del 2001, que a nivel local se materializó en un periodo marcado por el crecimiento tanto poblacional como urbano. Partiendo de un marco teórico desde la ecología política y la geografía crítica y por medio de una combinación de herramientas cuantitativas y análisis cualitativo que implicaron entre otras cosas: análisis de imágenes satelitales, entrecruzamiento de bases de datos estadísticos georreferenciados, relevamiento de una gran cantidad de fuentes documentales que fue desde la normativa existente y diversos planes de desarrollo provinciales, municipales y federales, hasta informes técnicos e entrevistas a informantes claves; esta tesis busca presentar y problematizar las formas en que el turismo como actividad económica impacta de manera concreta en la configuración del espacio urbano, lo que a su vez trae consecuencias diferenciadas para el ambiente que le rodea, así como para las

poblaciones locales que allí viven para el periodo de tiempo comprendido entre el 2001 y el 2018. Entre los principales resultados concluimos que, la ciudad de Bariloche, año a año atrae, por sus paisajes naturales, grandes contingentes de visitantes turísticos. El modo particular con que se desarrolló la actividad implicó una serie de impactos negativos tanto en el ambiente como en la infraestructura urbana, pero con una configuración espacial diferenciada. La actividad turística se despliega territorialmente sobre el producto que ofrece: el entorno natural, es decir, el bosque y el lago, afectando su calidad. La configuración urbana resultante de baja densidad, a su vez impacta en diversos problemas tales como la calidad del servicio de transporte, los costes en la provisión de servicios o la falta de ellos, contaminación de suelo y agua e incluso dificultad en el acceso a la vivienda y empleo. El Estado resulta un actor fundamental en permitir y garantizar la reproducción ampliada del capital en el territorio. A través de sus diversas manifestaciones se pone en evidencia su rol central en el despliegue de la actividad. Finalmente, frente a los emergentes de estas problemáticas, las alternativas propuestas a nivel local profundizan el modo de desarrollo lo que redunda en nuevos impactos negativos. De este modo, se puede concluir que se ha construido una imagen de destino natural que se ve amenazado por el devenir de la misma actividad que lo promociona y que impacta negativamente en la calidad de vida de las personas que lo habitan.

Desarrollo de métodos bioinformáticos para la caracterización estructural y funcional de proteínas ancestrales

Lic. Guillermo Ignacio Benítez

Director: Dr. Gustavo Parisi

Índice general

Publicaciones

Como resultado del presente trabajo se han publicado los siguientes artículos:

Matias Sebastian Carletti, Alexander Miguel Monzon, Emilio Garcia-Rios, Guillermo Benitez, Layla Hirsh, Maria Silvina Fornasari, Gustavo Parisi, Revenant: a database of resurrected proteins, Database, 2020

Guillermo Ignacio Benítez, Ana Julia Velez Rueda, Leandro Matías Sommese, Sebastián M. Ardanaz, Estefanía L. Borucki, Nicolas Palopoli, Luis E. Iglesias, Gustavo Parisi. Structural and evolutionary analysis unveil functional adaptations in the promiscuous behavior of serum albumins, Biochimie, Volume 197, 2022.

Diego Javier Zea, Juan Mac Donagh, Guillermo Benitez, Cristian Guisande Donadio, Julia Marchetti, Nicolas Palopoli, María Silvina Fornasari, Gustavo Parisi. Evolutionary rates in human amyloid proteins reveal their intrinsic metastability, bioRxiv 2022.09.07.506994; doi: https://doi.org/10.1101/2022.09.07.506994

Conformational epistasis impairs AlphaFold structural predictions. Luciana Rodriguez Sawicki, Guillermo Benitez, Matias Carletti, Nicolas Palopoli, Maria Silvina Fornasari, Gustavo Parisi, bioRxiv 2022.11.15.516638; doi: https://doi.org/10.1101/2022.11.15.516638

Artículos en colaboración:

Velez Rueda, A, Benitez, G, Marchetti, J, Hasenahuer, M, Fornasari, MS, Palopoli, N and Parisi, G. Bioinformatics calls the school: Use of smartphones to introduce Python for Bioinformatics in High Schools. 15(2): e1006473. Plos Comp Biology (ISSN 1553-734X, DOI doi.org/10.1371/journal.pcbi.1006473). 2019

Necci, M, Piovesan, D, CAID Predictors, András Hatos, Borbála Hajdu-Soltész, Alexander Miguel Monzon, Nicolas Palopoli, Lucía Álvarez, Burcu Aykac-Fas, Claudio Bassot, Guillermo Ignacio Benítez, Martina Bevilacqua, Anastasia Chasapi, Lucia Chemes, Norman Davey, Radoslav Davidović, Keith Dunker, Arne Elofsson, Julien Gobeill, Nicolás S. González Foutel, Mainak Guharoy, Tamas Horvath, Valentin Iglesias, Andrey V. Kajava, Orsolya Panna Kovacs, John Lamb, Matteo Lambrughi, Tamas Lazar, Jeremy Y. Leclercq, Emanuela Leonardi, Sandra Macedo-Ribeiro, Mauricio Macossay-Castillo, Emiliano Maiani, Jose A. Manso, Cristina Marino-Buslje, Elizabeth Martínez-Pérez, Bálint Mészáros, Ivan Mičetić, Giovanni Minervini, Nikoletta Murvai, Marco Necci, Christos Ouzounis, Mátyás Pajkos, Lisanna Paladin, Rita Pancsa, Elena Papaleo, Gustavo Parisi, Emilie Pasche, Pedro José Barbosa Pereira, Vasilis J. Promponas, Jordi Pujols, Federica Quaglia, Patrick Ruch, Eva Schad, Beata Szabo, Tamás Szaniszló, Stella Tamana, Agnes Tantos, Nevena Veljkovic, Salvador Ventura, Wim Vranken, Zsuzsanna Dosztányi, Peter Tompa, Tosatto, SCE. Critical Assessment of Protein Intrinsic Disorder Prediction. Nature Methods 2021.

András Hatos, Borbála Hajdu-Soltész, Alexander Miguel Monzon, Nicolas Palopoli, Lucía Álvarez, Burcu Aykac-Fas, Claudio Bassot, Guillermo Ignacio Benítez, Martina Bevilacqua, Anastasia Chasapi, Lucia Chemes, Norman Davey, Radoslav Davidović, Keith Dunker, Arne Elofsson, Julien Gobeill, Nicolás S. González Foutel, Mainak Guharoy, Tamas Horvath, Valentin Iglesias, Andrey V. Kajava, Orsolya Panna Kovacs, John Lamb, Matteo Lambrughi, Tamas Lazar, Jeremy Y. Leclercq, Emanuela Leonardi, Sandra Macedo-Ribeiro, Mauricio Macossay-Castillo, Emiliano Maiani, Jose A. Manso, Cristina Marino-Buslje, Elizabeth Martínez-Pérez, Bálint Mészáros, Ivan Mičetić, Giovanni Minervini, Nikoletta Murvai, Marco Necci, Christos Ouzounis, Mátyás Pajkos, Lisanna Paladin, Rita Pancsa, Elena Papaleo, Gustavo Parisi, Emilie Pasche, Pedro José Barbosa Pereira, Vasilis J. Promponas, Jordi Pujols, Federica Quaglia, Patrick Ruch, Eva Schad, Beata Szabo, Tamás Szaniszló, Stella Tamana, Agnes Tantos, Nevena Veljkovic, Salvador Ventura, Wim Vranken, Zsuzsanna Dosztányi, Peter Tompa, Silvio C. E. Tosatto, Damiano Piovesan. DisProt: intrinsic protein disorder annotation in 2020. Nucleic Acids Research 48 (D1), D269-D276 2020 (ISSN 0305-1048).

1 Introducción

1.1 Evolución biológica

La evolución se puede definir, como lo postuló Charles Darwin (1859) en su libro *On the Origin of Species*, como un proceso de descendencia con modificaciones. En base a esto podemos asumir que todos los organismos que existen en la Tierra actualmente son producto de este proceso y son descendientes de organismos diferentes que existieron en el pasado (organismos ancestrales). Se cree que el origen de la vida ocurrió hace aproximadamente 4000 millones de años (Lane, 2016), constituídos por sistemas con distintos niveles de organización molecular, los cuales comenzaron oportunamente a estar sujetos a los condicionamientos dados por la evolución biológica. Estos sistemas llegaron al punto de organizarse en los primeros organismos con organización celular. Uno de estos organismos es el considerado último ancestro común universal (LUCA del inglés *last universal common ancestor*, Theobald, 2010), del cual desciende toda la diversidad biológica observada a día de hoy. La teoría de la evolución no sólo ha afectado nuestra comprensión de la historia biológica de los organismos al ofrecer una explicación mecanística del proceso, sino que al ubicar al hombre en un contexto natural, ha impactado enormemente en ciencias humanísticas como la sociología, filosofía y psicología (Mayr, 1997).

Básicamente, podemos describir la evolución biológica como el resultado de dos procesos: por un lado, la generación de variantes genéticas y por otro lado, la selección de estas variantes. Este último proceso, sin embargo, se ha debatido ampliamente debido a la importancia relativa que tiene en la evolución (ver por ejemplo (Nei, 2005)). En cambio, la generación de variantes genéticas es considerada uno de los pilares de la evolución; este proceso ocurre mayormente en forma aleatoria (Li, 2006). Considerando todas las posibles variantes de los polímeros de ácidos nucleicos o aminoácidos (espacio secuencial) y viendo

la diversidad secuencial de la actualidad, podemos asumir que en el origen de la vida hubo una fracción reducida de estas posibles combinaciones, lo que terminó siendo un condicionamiento inicial en la variabilidad secuencial. Si pensamos, por ejemplo, en una secuencia de ADN de 100 nucleótidos de longitud, sabemos que su espacio secuencial es $4x10^{100}$, y considerando un peso molecular promedio para los nucleótidos que lo componen es posible encontrar que esta cantidad inmensa de moléculas tendrían un peso mayor que varias Tierras juntas. Esto es un indicio de que, aunque hayan pasado millones de años de historia evolutiva y de la gran cantidad de organismos que la habitan y habitaron, es imposible recorrer completamente el espacio secuencial.

En este sentido, la evolución es un proceso que depende fuertemente de su historia y de las condiciones iniciales del proceso. Una enorme variedad de acontecimientos a lo largo del tiempo moldearon y determinaron el proceso evolutivo condicionando dicha diversidad biológica. Estas contingencias históricas provocaron que la diversidad biológica que observamos actualmente en la Tierra sea sólo una representante de las tantas que podrían haberse originado en otras circunstancias históricas (Gould, 1990).

El proceso evolutivo, desde un punto de vista bioquímico, está dado por mutaciones al azar en el material genético, las cuales, si son fijadas con el paso de las generaciones, pueden conducir a que haya una diversificación genética dentro de una población. Es decir, el proceso evolutivo es sinónimo de la existencia de un cambio genético dentro de una población. Estas mutaciones están sometidas a dos procesos: la selección natural y la deriva génica. Según la teoría clásica del Neodarwinismo, la selección juega un rol fundamental, siendo el punto de vista más extremo de esto el seleccionismo, el cual propone que la selección es la única fuerza que direcciona el proceso evolutivo. Entonces, desde este punto de vista, que una mutación sea fijada o no, depende de la aptitud relativa (*fitness*) que el alelo mutante le confiere al organismo, respecto al alelo original (*wild-type*). Para los seleccionistas, esta aptitud relativa siempre es el resultado del fenotipo particular del organismo, que es el producto de la interacción entre el genotipo y el ambiente. Por otro lado, la deriva génica considera que las mutaciones que ocurren son básicamente al azar,

pudiendo incluso perder alelos beneficiosos. Kimura (Kimura, 1968; Kimura and Ohta, 1974) propone que la mayoría de estas mutaciones son neutrales y están determinadas por la deriva génica. Para esta teoría neutral de la evolución molecular, si bien acepta la existencia de la selección, considera que lo que más aporta en el proceso de evolución son estos cambios neutrales aleatorios. A pesar de estas discrepancias centradas en el papel del factor dominante en el proceso evolutivo, seleccionistas y neutralistas coinciden en que la mayoría de las mutaciones nuevas serán deletéreas, es decir selectivamente desventajosas y que serán removidas de la población por selección purificadora. Donde sí se diferencian es en la suposición de la proporción relativa entre las mutaciones ventajosas y neutras. Por un lado, tenemos a los seleccionistas, que asumen que la mayoría de las mutaciones van a ser selectivamente positivas, mientras que, por otro lado, los neutralistas consideran que la mayoría de las mutaciones no deletéreas serán neutrales y muy pocas ventajosas.

Podemos distinguir dos tipos de mutaciones que ocurren dentro de las regiones codificantes de un gen: las mutaciones sinónimas (o silenciosas), que como resultado no vemos ningún cambio en los aminoácidos; y por otro lado existen las mutaciones no sinónimas (no silenciosas), por las cuales podemos observar un reemplazo de un aminoácido por otro. El primer grupo de mutaciones se puede considerar como selectivamente neutra, ya que no producen un cambio en el fenotipo del organismo. Sin embargo hay ciertas ocasiones en las que este tipo de mutación pueden estar sujetas a selección, como en el caso de existir el denominado *codon bias* (Akashi, 2003), proceso por el cual se pueden introducir ciertas alteraciones en el momento de la traducción debido a la deficiencia relativa de anticodones, tales como un cambio en la velocidad de la síntesis de la proteína, lo que puede generar cambios en su plegamiento. Por otro lado, el segundo grupo de mutaciones, al generar un cambio de aminoácidos, puede conllevar a un cambio en el fenotipo del organismo, dando lugar a posibles cambios en la funcionalidad o estabilidad de la proteína. Las distintas presiones de selección (positivas o negativas) dictaminan si estos cambios terminan siendo aceptados o no, dependiendo si el cambio de aminoácido resulta adaptativamente ventajoso o deletéreo, respectivamente (Li, 2006). Es

además sabido que el impacto en el fenotipo suele ser multifactorial y aislar el efecto de una mutación no suele ser sencillo (Chandler *et al.*, 2014).

A la hora de estudiar la evolución de las proteínas, podemos decir que su historia está dada, tanto por las sustituciones o mutaciones de los nucleótidos, como también por inserciones y deleciones de los mismos. Este conjunto es la fuente de variabilidad sobre la cual la evolución puede actuar, generar y condicionar cambios secuenciales, estructurales, conformacionales y funcionales de acuerdo al impacto en el fenotipo. Debido a que el proceso evolutivo tiene un carácter mayormente divergente, se puede considerar que esta serie de cambios empezaron teniendo lugar en las proteínas del ya mencionado LUCA, a partir del cual, mediante eventos de especiación, duplicación, neofuncionalización, entre otros, dieron como resultado los genes y proteínas de todos los organismos actuales(Pál *et al.*, 2006; Webster *et al.*, 2003).

La especiación es un fenómeno que ocurre cuando dos poblaciones diferentes de una misma especie acumulan mutaciones de manera diferencial, llegando a distinguirse al punto de no poder generar descendencia fértil entre ellas, lo que da como resultado la generación de dos especies diferentes. Los genes y proteínas que surgen de este proceso en cada especie se denominan ortólogos (Fig. 1.1). Por ejemplo, el citocromo C del humano es ortólogo al del gorila, ya que ambas proteínas provienen de un mismo ancestro en común, manteniendo su función biológica, pero presentando diferencias a nivel secuencial.

Otra forma en la que dos genes pueden divergir a partir de un mismo gen es en el proceso de duplicación génica, el cual ocurre, como su nombre lo indica, cuando un gen se duplica, estando presente más de una vez en el material genético del mismo organismo. Ahora bien, si bien estos dos genes son idénticos, también son independientes entre sí, es decir, ambos van a sufrir y acumular mutaciones de manera diferencial. Estos genes se denominan parálogos (Fig. 1.1), que, si bien provienen de un ancestro común, a diferencia de los ortólogos, ambas copias están presentes en el mismo organismo.

Ahora bien, el hecho de tener dos copias de un mismo gen puede dar paso a que alguno de los dos tenga más flexibilidad a la hora de mutar, ya que la presión selectiva solo

se ejerce sobre el fenotipo, por lo que solo es necesario que uno de los genes mantenga su función original. Esta flexibilidad puede dar lugar a que una de las copias acumule las suficientes mutaciones como para distinguirse de la otra, a tal punto incluso de poder cumplir con otra función (Tokuriki and Tawfik, 2009), proceso conocido como neofuncionalización. Este proceso puede ocurrir por dos mecanismos: 1) cambios en la secuencia de aminoácidos en la proteína que expresa el gen (por ej. produciendo una proteína con una nueva actividad bioquímica) o 2) cambios en el patrón de expresión del gen (por ejemplo, el gen comienza a expresarse en un tejido donde el gen ancestral no se expresaba). Un ejemplo clásico de paralogía y neofuncionalización ocurre con la hemoglobina y la mioglobina, donde si bien ambas proteínas tienen como actividad unir oxígeno, sus funciones biológicas son completamente distintas: mientras que la hemoglobina es una proteínas cooperativa y alostérica cuya función está relacionada con el transporte de oxígeno en la sangre, la mioglobina funciona como *buffer* de oxígeno en el músculo esquelético (Fig. 1.2).

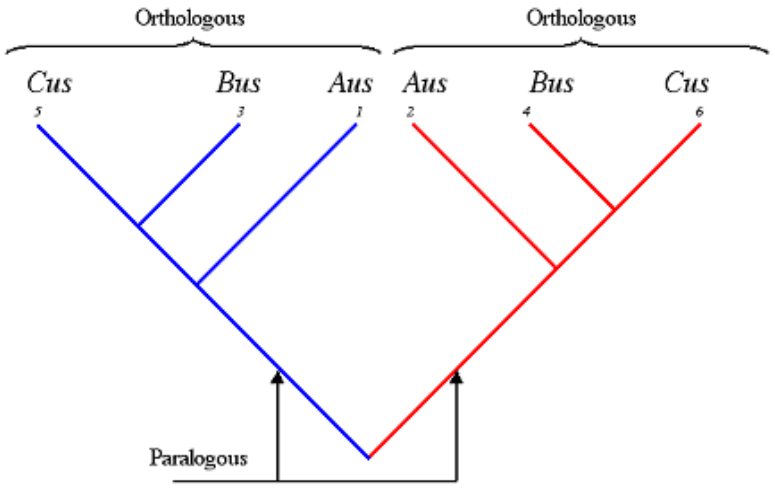


Figura 1.1. Evolución de genes ortólogos y parálogos.
Por un proceso de duplicación génica surgen dos genes parálogos (color azul y color rojo) que divergen de manera independiente por procesos de especialización generando ortólogos en los distintos organismos.

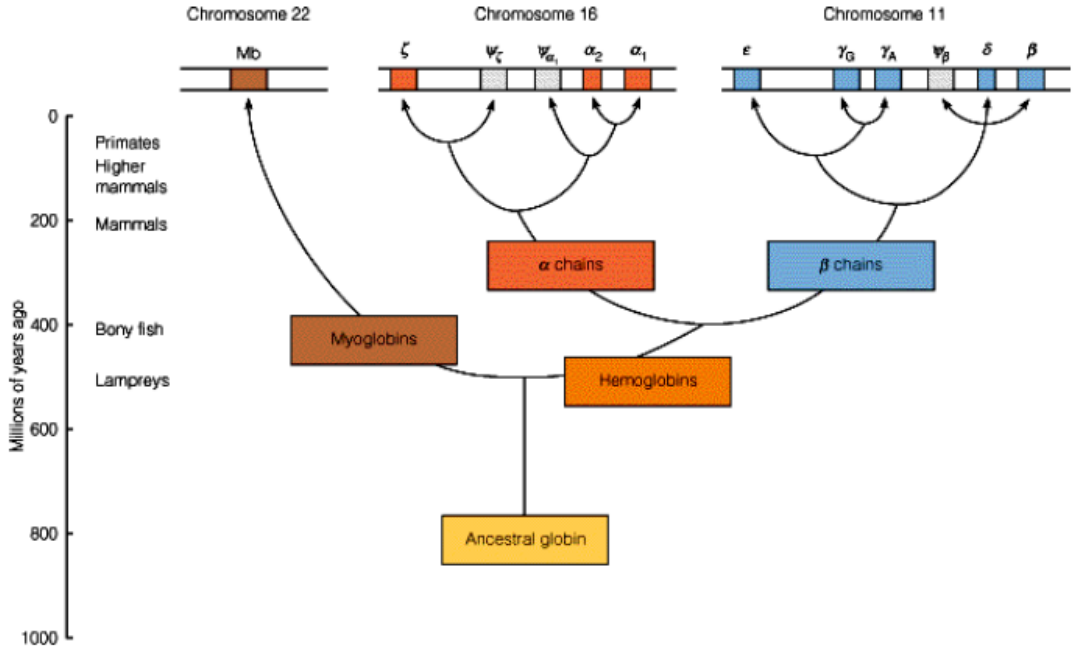


Figura 1.2. Ejemplo de evolución de genes parálogos referido a la familia de las globinas.
La hemoglobina y la mioglobina son dos proteínas homólogas que surgen por duplicación génica de un único gen ancestral (*ancestral globin*). Todos los vertebrados tienen mioglobina en su músculo y hemoglobina en su sangre. La Mioglobina difiere de la hemoglobina por las subunidades α y β de la hemoglobina, lo que indica que la mioglobina divergió (TD ≈ 600-800 millones de años) antes de que surgieran los genes α y β (TD ≈ 500 millones de años). Mamíferos, reptiles, aves, anfibios y peces óseos tienen subunidades α y β distintas, mientras que los vertebrados más primitivos, los Agnatha (peces sin mandíbula), contienen solo un tipo de subunidad de hemoglobina. Este patrón de distribución es consistente con las fechas estimadas de las duplicaciones genéticas: la mioglobina y la hemoglobina divergieron antes de la separación de agnatos y vertebrados mandibulares, mientras que la duplicación que da lugar a los genes α y β ocurrió en el antepasado de todos los vertebrados mandibulares luego de su divergencia con respecto a los agnatos.

Tal como se dijo previamente, todos los organismos, genes y proteínas actuales y extintos están relacionados entre sí a través del proceso de divergencia desde LUCA. La filogenia se encarga de estudiar estas relaciones, tanto a nivel macroscópico, como a nivel microscópico. A partir de esto se han desarrollado análisis filogenéticos, que utilizan un conjunto de conceptos y técnicas para representar las relaciones evolutivas estimando un árbol filogenético (Woese, 2000).

En sus inicios, estos análisis utilizaban comparaciones de caracteres morfológicos para estimar las relaciones evolutivas entre las especies (Linnaeus, 1758). Con el avance de la biología molecular en conjunto con el crecimiento de bases de datos moleculares

(Pfam (Punta *et al.*, 2012), Uniprot (Wu *et al.*, 2006), etc) se produjo un aumento considerable de la información molecular disponible, tales como secuencias de nucleótidos o de aminoácidos. El uso de esta información molecular permitió estimar relaciones filogenéticas más confiables desde el punto de vista estadístico debido a la mayor cantidad de información utilizada (Woese, 2004). Por otro lado, permiten estudiar las relaciones evolutivas entre genes o proteínas y formular y probar hipótesis sobre el proceso evolutivo que dio origen a las diversas estructuras y funciones de estas biomoléculas (Kumar and Filipski, 2001).

El procedimiento mayormente utilizado para estimar relaciones filogenéticas entre genes y proteínas involucra la comparación de patrones de mutaciones entre las secuencias involucradas en el análisis. Con este objetivo, los métodos filogenéticos consideran la similitud entre las secuencias y las asumen homólogas. El término homología se utiliza para indicar la similitud originada por la presencia de un ancestro común. Así, secuencias o estructuras similares pueden evidenciar a nivel molecular la presencia de un ancestro común. Si el ancestro común es muy antiguo, la divergencia secuencial puede ser tal que la cantidad de variación acumulada en el tiempo entre las secuencias no permite extraer información suficiente acerca de su relación evolutiva. Sin embargo, este problema ocurre en mucha menor medida cuando se comparan estructuras proteicas. Es un hecho ampliamente demostrado que las estructuras proteicas evolucionan mucho más lentamente que las secuencias que las codifican (Lesk and Chothia, 1980; Chothia and Lesk, 1986; Wood and Pearson, 1999; Overington *et al.*, 1990, 1992; Worth *et al.*, 2009).

La forma más básica de representar las relaciones evolutivas entre organismos, genes y proteínas es a través de un árbol bifurcante. Estas representaciones reciben el nombre de árboles filogenéticos debido a que su diagrama representa la estructura de un árbol, e incluso, varios términos para denominar sus partes son análogas a las partes de un árbol (Fig. 1.3). Por un lado, los nodos terminales, también llamados hojas, representan los taxones actuales, que también son llamados Unidades Taxonómicas Operacionales (OTUs por sus siglas en inglés *operational taxonomics units*), término genérico para representar la

variedad de taxones actuales que pueden ser comparados (una familia de organismos, organismos de diferentes familias, etc.). A su vez, los nodos terminales pueden también representar biomoléculas (proteínas o ácidos nucleicos) y no representar organismos propiamente dichos. De ahí, que los árboles filogenéticos pueden ser utilizados para estudiar la evolución de organismos o de moléculas.

Por otro lado, los nodos internos representan taxones ancestrales, organismos ya extintos y también son llamados como Unidades Hipotéticas Taxonómicas (HTUs por sus siglas en inglés) cuya denominación hace referencia a que estos taxones ancestrales son los antecesores hipotéticos. Los nodos se relacionan entre sí por medio de ramas que se bifurcan donde un ancestro genera dos descendientes. El patrón de ramificación del árbol filogenético, es decir el orden y ubicación de los nodos, se denomina topología.

La raíz de un árbol filogenético permite representar el nodo que corresponde al ancestro común (AC) de todos los OTUs estudiados, a la vez que posibilita visualizar la dirección del proceso evolutivo, desde la raíz hacia las hojas del árbol. En cambio, en un árbol sin raíz no es posible indicar que nodo corresponde al AC de todos los OTUs y tampoco la direccionalidad del proceso evolutivo. Básicamente, para que un árbol tenga raíz, es necesario que uno o más OTUs formen un grupo externo u *outgroup* del cual sabemos (o creemos saber) que es el más distante evolutivamente con respecto a todos los restantes OTUs. El resto de los OTUs que están más cerca evolutivamente entre sí forman un grupo interno o *ingroup*. Así el nodo raíz divide a los taxones *outgroup* e *ingroup* representando, como se dijo, el AC.

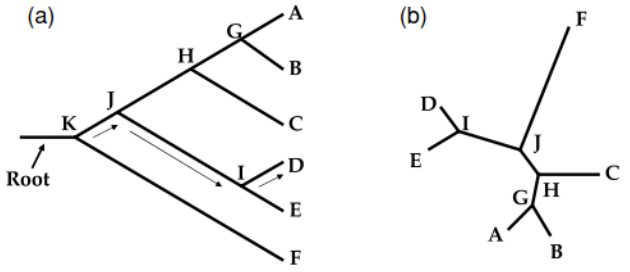


Figura 1.3. Estructura de un árbol filogenético con raíz (a) y un árbol filogenético sin raíz (b).
Los dos árboles tienen la misma topología. Un árbol con raíz generalmente se dibuja con la raíz a la izquierda.
Las letras A, B, C, D, E y F representan los nodos externos o OTUs. Las letras G, H, I, J y K representan los nodos internos o HTUs, donde la K es el nodo raíz. Él árbol raíz carece de este nodo raíz (*root*). Las líneas entre cada nodo son las ramas. En el árbol con raíz, la flecha indica la dirección del proceso evolutivo (por ejemplo, desde el nodo raíz K hasta el nodo externo D). En el árbol sin raíz, por el contrario, la dirección del proceso evolutivo es desconocida.

1.2 Resucitación de proteínas ancestrales

La reproducibilidad de la evolución a lo largo de la historia en la Tierra hoy en día es imposible de lograr y se puede considerar como un proceso único. Sin embargo, al tener acceso a numerosas evidencias de este proceso (tal como lo son los fósiles), podemos estimar ciertas características de los distintos tipos de organismos que habitaron la Tierra (Hunt, 2010; Koonin, 2012). Con relación a esto, se han encontrado fósiles que datan desde la época de las primeras formas de vida (~3500 millones de años (Schopf *et al.*, 2007)) en adelante.

Por el contrario, las evidencias moleculares más antiguas son relativamente recientes, debido a los cambios y degradación, tanto del ADN como de las proteínas. Se estima que los primeros se degradan a una velocidad de ~$5.5x10^{-6}$ bases por año (Allentoft *et al.*, 2012). Si bien hay publicaciones que aseguran que se pueden recuperar restos de ADN de aproximadamente 40 millones de años, estos fueron criticados por su falta de reproducibilidad. Incluso considerando que esto fuera posible, en lo que a la historia evolutiva se refiere, 40 millones de años es una cantidad relativamente pequeña. Por otro lado, las proteínas ancestrales en muestras fosilizadas presentan una supervivencia mayor (Hendy *et al.*, 2018), pero sigue presentando el mismo problema que con el ADN, ya que se las puede considerar relativamente actuales. Además, la disponibilidad de ambos tipos de moléculas es baja.

En los últimos años y debido a esta carencia de información molecular de organismos ancestrales de edades comparables a las provenientes de muestras fósiles, las estrategias para estimar las secuencias ubicadas en los nodos internos de un árbol filogenético (proteínas ancestrales) prometen ser métodos poderosos para inferir la sucesión ordenada de sustituciones de aminoácidos, o lo que se denomina camino evolutivo (Thornton, 2004; Gumulya and Gillam, 2017), para así llegar a estimar la secuencia completa de la proteína ancestral.

Estas estrategias nacieron prácticamente con el estudio de la evolución molecular, ya que en la década de los 60s, Pauling y Zuckerkandl propusieron que las proteínas modernas contenían suficiente información para inferir las secuencias de las proteínas ancestrales (Pauling and Zuckerkandl, 1963). Posteriormente, gracias al desarrollo de los análisis filogenéticos basados en métodos estadísticos surgieron los métodos modernos de reconstrucción de secuencias ancestrales (ASR por sus siglas en inglés de *Ancestral Sequence Reconstruction*) (Fitch, 1971; Yang *et al.*, 1995; Zhang and Nei, 1997; Pagel, 1999; Clemente *et al.*, 2009; Liberles, 2007). Estos son métodos computacionales que permiten estimar las secuencias de las proteínas ancestrales a partir de secuencias de proteínas homólogas actuales (Liberles, 2007). De manera simplificada, para realizar dicha estimación, estos métodos requieren un alineamiento múltiple de secuencias (MSA por su siglas en inglés de *multiple sequence alignment*) de proteínas pertenecientes a organismos actuales y su estimación filogenética correspondiente, obteniendo las secuencias de los nodos internos. Como paso posterior, estas secuencias ancestrales reconstruidas pueden ser sintetizadas en el laboratorio usando técnicas de biología molecular, siendo clonadas y expresadas en un sistema apropiado para finalmente ser purificadas. Estos últimos pasos son en sí mismos otro proceso, denominado resucitación de la proteína ancestral. Una vez purificadas, estas proteínas se pueden caracterizar experimentalmente de distintas formas, tanto biofísica como bioquímicamente (obtención de estructura, análisis de actividad, estabilidad, etc)(Fig. 1.4).

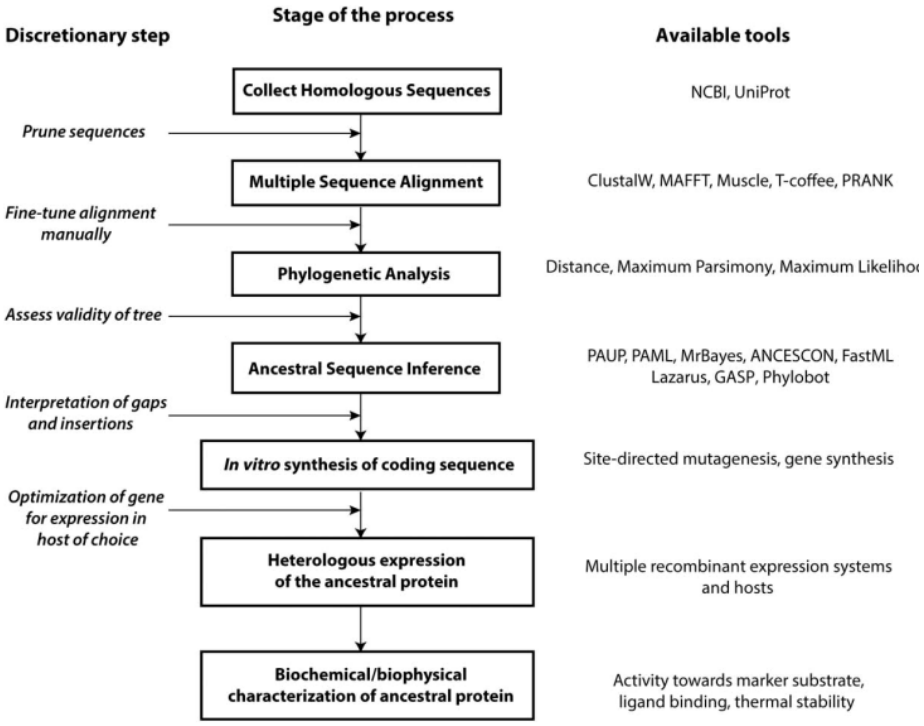


Figura 1.4. El proceso de la técnica de ASR y de la resucitación
La inferencia de las secuencias ancestrales desde homólogos actuales involucra varias etapas, muchas de las cuales involucran pasos de optimización manual (Anisimova *et al.*, 2010, 2013). Las secuencias recuperadas de bases de datos públicas primero deben ser revisadas y alguna que otra secuencia podrá ser removida. Las secuencias seleccionadas por lo tanto son alineadas y utilizadas para construir un árbol filogenético. Puede ser necesario un refinamiento manual del alineamiento y las relaciones evolutivas indicadas por el árbol deben corresponderse con la filogenia aceptada para los organismos en cuestión. Las secuencias ancestrales son entonces estimadas para cada nodo interno del árbol basándose en MSA optimizado y el árbol filogenético. Dependiendo de las relaciones evolutivas de las secuencias, de la precisión del curado manual de los *gap* y del algoritmo usado para la inferencia, es posible que se necesite tomar decisiones sobre si conservar o eliminar inserciones secuenciales del antepasado. Luego se sintetiza la secuencia de ADN que codifica la proteína ancestral y se clona en un plásmido de expresión, lo que permite la expresión heteróloga y la caracterización de la proteína ancestral. Para cada etapa, como se indica a la derecha, muchas herramientas de bioinformática están disponibles públicamente. Figura extraída de (Gumulya and Gillam, 2017).

Para comprender un poco más en profundidad cómo es que las técnicas de ASR permiten predecir las secuencias de proteínas ancestrales, a continuación detallaremos brevemente los pasos generales involucrados estos métodos.

Alineamiento múltiple de secuencias (MSA)

El primer paso en la reconstrucción de proteínas ancestrales requiere de un conjunto de secuencias de proteínas homólogas, para poder estimar un árbol filogenético a fin de

mostrar sus relaciones evolutivas (Fig. 1.4). Cabe destacar que se pueden utilizar secuencias actuales de ADN o proteínas dependiendo de qué tipo de modelos evolutivos serán aplicados para las inferencias filogenéticas: modelos basados en nucleótidos (*nucleotide-based*), en aminoácidos (*aminoacid-based*) o en codones (*codon-based*). Generalmente, los estudios de ASR usan secuencias de proteínas debido a que el uso de secuencias de ADN acarrea diferentes sesgos metodológicos del alineamiento en sí: 1) serán menos robustos cuando las regiones alineadas contengan muchas inserciones y deleciones; 2) no toman en cuenta la redundancia del código genético y las implicancias de ésta redundancia para la conservación de la función (Kumar and Filipski, 2007).

Las secuencias de ADN o proteínas homólogas que se utilizan para el MSA generalmente son obtenidas a través de búsquedas de similitud secuencial (BlastN, BlastP, PsiBlast (Altschul *et al.*, 1997)) usando distintos tipos de bases de datos, según lo requerido (GenBank (Benson *et al.*, 2013), UniProt (UniProt Consortium, 2008), DDBJ (Mashima *et al.*, 2016), SwisProt (Bairoch and Apweiler, 2000), entre otras). El conjunto de secuencias homólogas obtenidas de esta forma, debería representar, en la medida de lo posible, una diversidad que abarque a diferentes linajes evolutivos o dominios del mundo vivo. Finalmente el MSA se realiza utilizando un programa bioinformático que es seleccionado según el criterio de los investigadores contemplando el algoritmo utilizado para el alineamiento y las características mismas del programa; entre los programas más comúnmente usados están ClustalW (Sievers *et al.*, 2011), MUSCLE (Edgar, 2004), T-Coffee (Notredame *et al.*, 2000) or PRANK (usado para secuencias con inserciones (Maiolo *et al.*, 2018)).

El alineamiento en sí mismo, implica la aceptación de una hipótesis evolutiva. Los aminoácidos alineados en un alineamiento proteico asumen posiciones homólogas. De esta forma, la calidad del MSA es crucial para realizar una estimación filogenética fiable que determinará a su vez la calidad de la reconstrucción ancestral (Wang *et al.*, 2011). Así, frecuentemente se usan sólo secuencias que están completas (no truncadas) o también secuencias altamente conservadas debido a que tienen residuos críticos tales como

posiciones de sitios activos (Cole *et al.*, 2013). Siguiendo con está lógica, las secuencias que puedan incorporar cierto grado de incertidumbre en el alineamiento serán generalmente descartadas: secuencias que son dudosamente homólogas (por baja similitud secuencial, por ejemplo) así como secuencias que contienen muchas inserciones o deleciones. Por otro lado, en general el alineamiento debe ser curado manualmente con el objetivo de mejorar la confiabilidad de las posiciones con *gaps* (Anisimova *et al.*, 2010)). Para este procedimiento se puede usar como guía la ausencia o presencia de *gaps* dentro de uno de los grupos taxonómicos particulares usados en el alineamiento. Por último, es importante incluir un grupo reducido de secuencias que estén relacionadas con el resto, pero de manera más lejana (*outgroup*), para poder enraizar el árbol, ya que, como se dijo previamente, este tipo de árboles nos dan una direccionalidad en el tiempo.

Estimación filogenética

Una vez aceptado el MSA, se procede a la construcción del árbol filogenético que representará las relaciones evolutivas entre las secuencias. Para realizar esta estimación no solo se necesita el alineamiento, sino que requiere de algunas suposiciones evolutivas más. En general estas suposiciones están dadas por el modelo evolutivo que se utiliza para evaluar la filogenia. Dichos modelos están basados en matrices de sustitución, que indican la probabilidad de que un nucleótido o un aminoácido sea reemplazado por otro. Las matrices de sustitución más utilizadas con respecto a las proteínas son matrices empíricas, las cuales fueron confeccionadas analizando una gran cantidad de secuencias. Entre estas las más comunes son las de Dayhoff (Elleman, 1978), JTT (Jones *et al.*, 1992), WAG (Wang *et al.*, 2008) y LG (Jones *et al.*, 1992; Whelan and Goldman, 2001; Le and Gascuel, 2008). Debido a la variedad de modelos disponibles, la selección de un modelo de manera objetiva se realiza utilizando un criterio estadístico que estima cuán bien el modelo explica los cambios observados en el MSA. Existen muchos ejemplos en los cuales el uso de modelos

de evolución poco adecuados a la realidad llevan a conclusiones erróneas (Sullivan and Swofford, 1997; Pupko, Huchon, *et al.*, 2002).

La selección del método computacional para realizar el análisis filogenético es otro punto crucial que va a determinar la confiabilidad de la estimación filogenética. Aunque existen varios métodos disponibles ninguno garantiza que el árbol filogenético estimado sea el "real". Estos métodos pueden ser clasificados de dos maneras (Lemey *et al.*, 2009) :

- Por el enfoque algorítmico que utilizan: están por un lado los que se basan en un algoritmo de agrupamiento (*clustering*) que en general dan como resultado una única topología del árbol estimado, y por otro lado los que utilizan un algoritmo basado en un criterio de optimización los cuales evalúan un conjunto de posibles topologías.
- Por el tipo de datos que usan: pueden ser estados de caracteres discretos o matrices de distancia de a pares.

Según el algoritmo que utilizan los métodos pueden tener ventajas que radican en la selección de un mejor árbol o en los recursos computacionales necesarios para los cálculos. Los métodos basados en algoritmos de optimización utilizan un criterio de mejor adecuación (*goodness-of-fit criterion*) y evalúan diferentes topologías para un número dado de taxones y de esta manera intentan encontrar el árbol óptimo para el criterio empleado. Los métodos basados en estados de caracteres discretos consideran que cada posición en las secuencias alineadas es un "carácter" y los diferentes nucleótidos o aminoácidos de cada posición son los "estados". En contraste, los métodos basados en distancias calculan alguna medida de la disimilitud entre cada par de secuencias para producir una matriz de distancia de a pares. La Tabla 1.1 extraída del texto filogenia de Lemey y colaboradores (Lemey *et al.*, 2009)) resume la clasificación de los métodos de reconstrucción del árbol filogenético según los datos y los algoritmos utilizados.

Una diferencia importante entre los métodos basados en caracteres y los basados en distancias es que los primeros consideran explícitamente el estado original del carácter del taxón, es decir que para datos moleculares consideran qué nucleótido, codón o aminoácido hay en cada posición de cada secuencia perteneciente al MSA. De este modo,

pueden ser utilizados para reconstruir el estado del carácter en el nodo ancestral. En cambio, los métodos basados en distancia descartan el estado del carácter del taxón aunque lo consideran para calcular las distancias. De allí que se pierde la información necesaria para reconstruir el estado del carácter en el nodo ancestral. Finalmente, si observamos la (Tabla 1.1) podemos deducir que, en el contexto de las técnicas de reconstrucción ancestral detalladas en la próxima sección, se utilizan métodos basados en criterios de búsqueda óptima (Máxima Parsimonia (MP), Máxima Verosimilitud (MV) o métodos Bayesianos (By)).

Tabla 1.1 Clasificación de métodos análisis filogenéticos y sus enfoques estratégicos.

	Algoritmo de Criterio de Optimización	**Algoritmo de Agrupamiento**
Estado del Carácter	Máxima Verosimilitud (MV)	
	Inferencia Bayesiana (By)	
	Máxima Parsimonia (MP)	
Matrices de Distancia	Fitch–Margoliash	UPGMA
		Agrupamiento de Vecinos (NJ del inglés *Neighbor-joining*)

Por último, un método que se usa generalmente para evaluar la fiabilidad de esta estimación, es el método de bootstrapping (Efron, 1994). Este método consiste básicamente en un remuestreo de la población para realizar la inferencia con estos datos. En el caso de los árboles filogenéticos, cada columna del alineamiento es considerado un carácter de la población, por lo que lo cada remuestreo lo que hace es reacomodar las columnas, generando un nuevo alineamiento, a partir del cual se vuelve inferir el árbol. Esto se hace una cantidad N de veces, siendo 1000 el mínimo recomendable. La evaluación consiste en comparar las topologías de cada árbol generado, donde se hace foco en cada nodo interno. Si un nodo en particular aparece en todas las topologías, se dice que es robusto estadísticamente. Aunque no hay un valor de corte definido, en general se aceptan como buenos los nodos que aparecen en un 50% o más de los árboles generados.

Reconstrucción de las secuencias ancestrales (ASR)

Una vez obtenidas las secuencias, el MSA y la estimación filogenética, el próximo paso es inferir las secuencias correspondientes a uno o varios nodos internos del árbol filogenético. Evidentemente estos nodos ancestrales corresponden a momentos específicos y particulares de la historia evolutiva de la familia en cuestión. Como se dijo anteriormente la reconstrucción ancestral de proteínas trata de estimar las propiedades secuenciales y a partir de ellas, mediante la resucitación, se estudian propiedades estructurales, bioquímicas, biofísicas de proteínas pertenecientes a organismos ancestrales para responder preguntas biológicas específicas. Por ejemplo, sabiendo que la temperatura de la Tierra era en promedio ~80 grados hace 3000 millones de años, ¿las proteínas ancestrales correspondientes a esa época eran más estables que las actuales? Muchas proteínas que degradan antibióticos son altamente específicas, ¿lo fueron siempre? o ¿los estados ancestrales de dichas proteínas eran promiscuos?

Para realizar la inferencia ancestral se utilizan métodos que primeramente fueron desarrollados para análisis filogenéticos basados en caracteres discretos que, como se dijo anteriormente, consideran explícitamente los estados de los caracteres. Diferentes herramientas bioinformáticas (*softwares*), algunas de las cuales también utilizan para realizar las estimaciones filogenéticas, se pueden utilizar para llevar a cabo la ASR: PAUP (*Phylogenetic Analysis Using Parsimony*; (Swofford and Sullivan, 2009)), PAML (*phylogenetic analysis by maximum likelihood*) (Yang, 2007), MrBayes (Ronquist *et al.*, 2012), ANCESCON (Cai *et al.*, 2004), FastML (Pupko, 2002). La Tabla 1.2 provee información de cada uno de estos *softwares*, sus características principales y se cita un ejemplo de su uso. Además, en la misma tabla se nombra qué tipo de modelos evolutivos están disponibles para llevar a cabo la reconstrucción ancestral, que como dijimos antes pueden estar basados en DNA, codones o proteínas. Para seleccionar el modelo evolutivo más adecuado a los datos (el MSA y la filogenia) en general se utilizan criterios estadísticos para comparar diferentes modelos, tales como Akaike Information Criterion (AIC) y el

likelihood ratio test (LRT; (Posada and Buckley, 2004)). Así, para seleccionar el modelo más adecuado mediante alguno de dichos criterios se utilizan *softwares* tales como ProTest (Abascal *et al.*, 2005). Por otro lado, los *softwares* de la Tabla 1.2 que permiten indicar el supuesto de la variación de las velocidades de sustitución entre los sitios (ASRV del inglés *Among Site Rate Variation*) en la forma de distribución discreta gamma (*γ-distribution*), obteniendo una mejor adecuación del modelo evolutivo a la realidad y por lo tanto esto redunda en una mejor estimación filogenética (Yang, 1996). Con respecto a la reconstrucción de secuencias ancestrales, ASRV es especialmente crítica cuando se reconstruyen secuencias altamente divergentes (Pupko, Pe'er, *et al.*, 2002) lo que generalmente ocurre en los trabajos de resucitación (e.g. Thornton, 2001; Chang *et al.*, 2002).

Tabla 1.2 Comparación de herramientas para la estimación filogenética y la ASR. Tabla extraída y adaptada de (Gumulya and Gillam, 2017)

Software	**Modelos evolutivos**	**Método de inferencia**	**Características**	**Referencia (ejemplo)**
PAUP	ADN, proteína	MP, MV y métodos de distancia. Todos incorporados en la versión 4.0	Varias opciones para métodos filogenéticos.	(Chandrasekharan *et al.*, 1996)
PAML	ADN, proteína	MV	*Likelihood ratio tests,* estimación de las tasas de mutaciones sinónimas y no sinónimas,detección de selección darwiniana positiva, estimación de los tiempos de divergencia de las especies bajo modelos del reloj molecular.	(Malcolm *et al.*, 1990; Akanuma, 2017; Risso *et al.*, 2015)
MrBayes	ADN, codón, proteína	Métodos Bys empíricos y jerárquicos	Incorpora análisis de incertidumbres (Por ej. en topologías y modelos), muchos modelos evolutivos disponibles, estimación de sitios bajo selección positiva	(Steindel *et al.*, 2016; Thomson *et al.*, 2005)
ANCESCON	Proteína	Métodos basados en distancia	Considera la tasas de variaciones entre los sitios (ASRV)	(Butzin *et al.*, 2013)
Lazarus	ADN, proteína	Métodos Bys	Utiliza PAML, corre	(Finnigan *et al.*,

		empíricos y MV	paralelamente a PAML	2012)
FastML	ADN, codón, proteína	MV	Servidor web de fácil utilización, reconstruye estados ancestrales que son *gaps* en el alineamiento (indel: inserciones o deleciones)	(Baker *et al.*, 2013)
Phylobot	ADN, proteína	Métodos Bys empíricos y MV	Servidor web de fácil utilización, automatizado	(Bar-Rogovsky *et al.*, 2013)
GASP	Proteína	MV	Puede inferir secuencias ancestrales desde alineamientos con *gaps*.	

A continuación describiremos brevemente cómo funcionan los algoritmos desarrollados bajo estos diferentes enfoques para realizar la reconstrucción ancestral.

ASR basado en el criterio de Máxima Parsimonia (MP)

El criterio de la máxima parsimonia adopta el principio de selección de la hipótesis más simple (“parsimonia”). En el contexto de ASR el objetivo de la parsimonia es encontrar los estados ancestrales en cada nodo interno de un árbol filogenético (es decir, la distribución de los estados ancestrales en todo el árbol) que minimiza el número de cambios en los caracteres que serían necesarios para explicar las diferencias observadas entre las secuencias de los nodos externos. El método bajo este principio llamado de Máxima Parsimonia corresponde a los primeros algoritmos desarrollados formalmente para reconstruir secuencias ancestrales. Estos algoritmos fueron desarrollados por Fitch, Sankoff y Rousseau (Fitch, 1971; Sankoff, 1975; Sankoff and Rousseau, 1975). Tomando como ejemplo el algoritmo más simple desarrollado por Fitch y utilizando como datos secuencias de ADN, el método propone penalizar por igual a cualquiera de los cambios entre los estados del carácter (A, C, G y T). En lo que respecta a proteínas, existen varios ejemplos de uso del método de MP para reconstruir diversas familias de proteínas ancestrales: lisozimas ancestrales (Malcolm *et al.*, 1990), la proteína L1 de ratón (Adey *et al.*, 1994), ribonucleasas de bóvidos y de artiodáctilos (Stackhouse *et al.*, 1990) (Jermann *et al.*, 1995).

Para reconstruir una posición específica de la cadena de las secuencias ancestrales, el procedimiento algorítmico asigna a cada nodo interno de un árbol bifurcante con raíz un conjunto de estados del carácter que son compatibles con el principio de la parsimonia (es decir el mínimo número de cambios). El algoritmo consta de dos etapas, la primera procesa el árbol desde las "hojas" hacia la raíz, esto es visitando cada nodo del árbol solamente después de que su descendientes hayan sido visitados. El objetivo es determinar el conjunto posible de estados de los caracteres de un dado ancestro basado en los estados del carácter observado en sus descendientes. La segunda etapa consiste en visitar el árbol desde la raíz hacia las hojas con el objetivo de finalmente determinar los estados ancestrales para los nodos internos.

En la primera etapa, llamada recorrido de pos orden (*postorder traversal*), el algoritmo comienza por asignar un conjunto de caracteres a las hojas del árbol. Por ejemplo, si una hoja posee una posición dada de la secuencia una Adenina (A), el conjunto {A} es asignado a esa hoja. Paso siguiente, el algoritmo procesa a un nodo interno del cual sus dos descendientes ya han sido visitados. Así, a dicho nodo interno se le asigna como conjunto de estados del carácter a la intersección de los dos conjuntos asignados a sus nodos descendientes (si la intersección no es un conjunto vacío), y si la intersección es un conjunto vacío se le asigna la unión de los dos conjuntos.

En la segunda etapa, llamada recorrido de pre orden (*preorder traversal*), los estados del carácter son asignados a cada descendiente basados en identificar qué caracteres comparten con sus predecesores. En el comienzo de este recorrido, los estados ancestrales de la raíz corresponden a los ya asignados como conjunto de estados de caracteres en el primer recorrido. Si hay más de un posible estado para el conjunto de caracteres en la raíz, debido a que la raíz no tienen ningún nodo predecesor, se requiere que se asignen los estados ancestrales del carácter de manera arbitraria para proseguir la reconstrucción y al final existirán varias reconstrucción igualmente parsimoniosas. Para asignar los estados ancestrales de los nodos descendientes de la raíz se evalúa la coincidencia de los estados ancestrales asignados a la raíz con los estados de los

caracteres del nodo descendiente. Los estados que coinciden son asignados como estados ancestrales al nodo descendiente, y de los estados que no coinciden entonces se le asignará el estado del nodo descendiente. De esta misma manera se evalúan todo el resto de los nodos internos del árbol. Existen descripciones detalladas de este algoritmo en utilizando árboles filogenéticos y caracteres hipotéticos (ver por ejemplo (Pupko *et al.*, 2007; Joy *et al.*, 2016))

ASR basado en modelos probabilísticos (MV y By)

Máxima verosimilitud (MV) o *Maximum likelihood* (ML). Estos métodos estiman el valor de un parámetro como aquél que haga máxima la probabilidad de obtener los datos observados (Cam, 1990). En el contexto de la reconstrucción ancestral la MV toma como parámetro los estados del carácter en los nodos internos, ancestrales, y se propone a encontrar los valores del parámetro que maximizan la probabilidad de obtener los estados actuales observados de los caracteres (Joy *et al.*, 2016). A su vez la MV supone la veracidad de las hipótesis: el modelo de evolución y la topología. En otras palabras, si hablamos de reconstrucción de secuencias ancestrales, en los métodos basados en la MV los parámetros a optimizar serían los nucleótidos o aminoácidos en cada nodo ancestral que hagan máxima la probabilidad de obtener las secuencias homólogas actuales dados la estimación filogenética y el modelo de evolución molecular. Los primeros métodos de MV para reconstruir secuencias de nucleótidos y proteínas fueron desarrollados a mediados de la década del noventa (Yang *et al.*, 1995; Koshi and Goldstein, 1996).

Antes de exponer una descripción general de los algoritmos basados en la verosimilitud de los datos es importante aclarar las dos suposiciones principales en las que se basan. En primer lugar, se asume que los modelos evolutivos son simétricos, es decir que la probabilidad de observar la mutación de un estado *i* a uno *j* es la misma que la mutación de un estado *j* a uno *i*. Esta suposición implica que la ubicación de la raíz del árbol no sea importante y por lo tanto se puede implementar la ASR bajo el método de MV sobre un árbol sin raíz o seleccionando arbitrariamente cualquier nodo como raíz dado que la

posición de la raíz no afecta la verosimilitud (Felsenstein, 1981). En segundo lugar, se asume que cada sitio evoluciona de manera independiente, lo que implica que la probabilidad de observar todo el alineamiento de las secuencias actuales se puede calcular como el producto de las probabilidades de cada sitio.

Para describir el algoritmo de la ASR basado en la MV, en principio se considerará, por simplicidad, a un solo sitio de la secuencia. Así para ilustrar dicho algoritmo primero consideremos la siguiente ecuación:

$$P(y|d,M,T,\theta) = \frac{L(d|y,M,T,\theta)}{\sum_i L(d|y_i,M,T,\theta)}$$

Ecuación 1

- La letra *d* se refiere a los datos observados: los estados para un solo sitio de las secuencias alineadas.
- La letra y_i se refiere al vector de los estados ancestrales asignados a todos los nodos de la filogenia.
- La letra *T* se refiere a la filogenia obtenida por MV y *M* se refiere al modelo evolutivo utilizado.
- La letra *θ* se refiere al vector de parámetros del modelo de MV y las longitud de las ramas del árbol.
- La expresión *P(y|d,M,T,θ)* es la probabilidad condicional (probabilidad posterior o PP) de asignar el vector de estados ancestrales *y* a todos los nodos internos, dados los datos, el árbol obtenido por MV, el modelo y los parámetros del modelo de MV.
- La expresión *L(d|y,M, T, θ)* es la verosimilitud de los datos observados, dados el *set* de estados ancestrales (*y*), el modelo, la topología del árbol y los parámetros del modelo. El denominador precedido por la sumatoria, corresponde a la verosimilitud de los datos sumados sobre todos las asignaciones posibles de estados ancestrales.

Si consideramos un alineamiento con n sitios, la PP de asignar determinados estados ancestrales a todos los nodos internos en todos los sitios se calcula multiplicando la PP de cada uno de los estados ancestrales para cada sitio, como muestra la siguiente ecuación:

$$\prod_{s=1}^{n} P\left(y_s | d_s, M, T, \theta \right) = \frac{L\left(d_s | y_s, M, T, \theta \right)}{\sum_i L_k\left(d_s | y_{s_i}, M, T, \theta \right)}$$

Ecuación 2

Dentro del modelo de MV existen dos variantes o enfoques diferentes para determinar las secuencias ancestrales en los nodos internos de un árbol filogenético: la reconstrucción marginal (*marginal ML reconstruction*) y la reconstrucción conjunta (*joint ML reconstruction*). En el enfoque de la reconstrucción marginal se trabaja asignando el estado ancestral (en un nodo interno) más probable de un sitio de la secuencia tomando en cuenta sólo a sus descendientes inmediatos. Dicho de otra forma, se estiman los estados ancestrales de cada nodo, tomando los nodos desde las hojas hacia la raíz del árbol, que maximizan la probabilidad de obtener los estados del carácter de sus descendientes inmediatos (Ecuación 1). En cambio, en el enfoque de la reconstrucción conjunta se intenta encontrar la combinación de estados ancestrales a través del árbol entero (es decir la combinación de estados ancestrales de todos los nodos internos del árbol) que maximiza la probabilidad de los datos (las secuencias homólogas actuales alineadas) (Ecuación 2). Los algoritmos para la ASR marginal fueron introducidos por Koshi y Goldstein (Koshi and Goldstein, 1996) y los algoritmos para la ASR conjunta fueron introducidos por Pupko *et al*. (Pupko *et al.*, 2000). Estas versiones de ambos enfoques se implementan con una complejidad computacional equivalente que aumenta cuando aumenta el número de taxones del alineamiento múltiple. Ambas variantes están implementadas en el paquete PAML (Yang, 1997, 2007).

Las reconstrucciones marginales y conjuntas pueden llevar a diferentes resultados en la reconstrucción ancestral (Fig. 1.6), y la selección del método está sujeta a la pregunta filogenética del investigador. Por ejemplo, la reconstrucción conjunta es más apropiada si la pregunta se relaciona con conocer la variación de los estados ancestrales a lo largo del árbol, en cambio, la reconstrucción marginal debería ser utilizada si el objetivo es resucitar un ancestro específico.

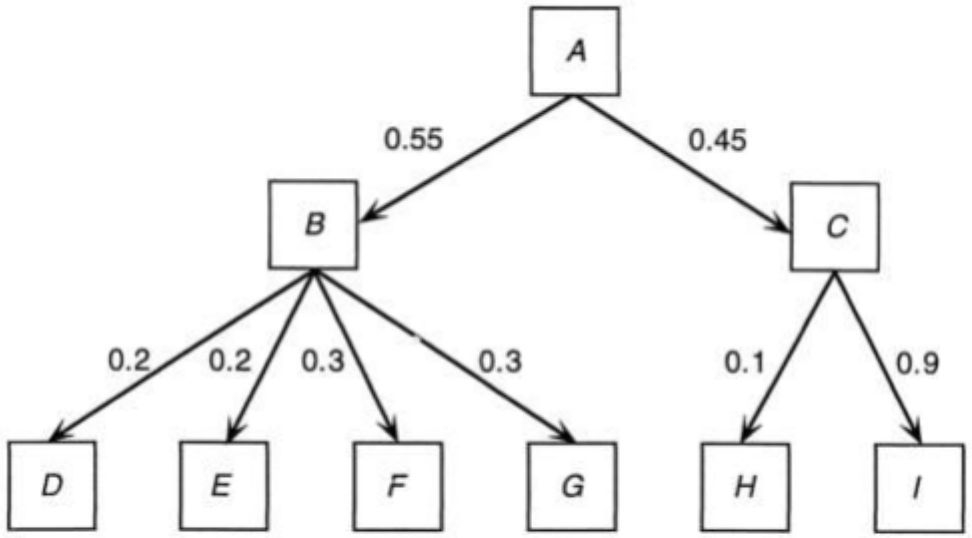


Figura 1.6 Ilustración Numérica de la diferencia entre ML *marginal* y ML *joint*.
La probabilidad de cambio de cualquier estado a otro está marcada en las flechas. Vemos que el primer paso es el más probable (ML *marginal*) es A → B (0.55). Sin embargo, el camino completo con mayor probabilidad (ML *joint*) es A → C → I (0.45 x 0.9 = 0.405).

Inferencia Bayesiana (By) o *Bayesian Inference* (BI). En la inferencia Bayesiana, la estimación de la topología o árbol filogenético tanto como el modelo evolutivo constituyen variables que aportan a la confiabilidad de los datos obtenidos. Esto es, en los métodos Bys la estimación de la filogenia y el mejor modelo evolutivo que describe los datos son estimados conjuntamente con los caracteres ancestrales. Utilizando el teorema de Bayes uno podría obtener:

$$P(S|D,\theta) = \frac{P(D|S,\theta)\,P(S|\theta)}{P(D|\theta)}$$

Ecuación 3

- *S* estados ancestrales,
- *D* datos observados (actuales)
- *θ* representa filogenia y modelo evolutivo

P(D|S,θ) es la verosimilitud de los datos observados mientras que P (S | θ) es la probabilidad a priori de los estados ancestrales dados un determinado modelo y árbol filogenético. Finalmente, *P(D|θ)* es la probabilidad de los datos para un determinado modelo y árbol considerando todos los posibles estados ancestrales.

Una de las primeras formas de aplicar Bayes en la estimación ancestral fue derivado por Yang y colaboradores e implementado en el programa PAML (Yang, 1997) extendiéndose luego a otros programas como MrBayes (Ronquist and Huelsenbeck, 2003). En principio parecería que tener en cuenta la incertidumbre de la topología y del modelo incrementa la capacidad predictiva de los métodos bayesianos a los que utilizan ML. Sin embargo, los métodos Bys no muestran una mejor capacidad predictiva sobre los métodos de máxima verosimilitud (Gumulya and Gillam, 2017; Eick *et al.*, 2017; Hanson-Smith *et al.*, 2010).

2 Tasas evolutivas en proteínas amiloides humanas revelan su metaestabilidad intrínseca

2.1 Resumen

Este capítulo está dedicado al estudio de un conjunto de proteínas humanas que en determinadas condiciones metabólicas y/o fisiológicas producen amiloides. Los amiloides son estados condensados con propiedades de “sólidos” donde predominan las interacciones inter-catenarias. Los amiloides forman fibras que en la mayoría de los casos producen diversas patologías en humanos. Sin embargo, se han caracterizado amiloides funcionales, haciendo más complejo el efecto de la presencia de amiloides en las células. Utilizando técnicas de análisis evolutivo, encontramos que las proteínas que forman amiloides, tanto funcionales como patológicos, están entre las proteínas que más rápidamente evolucionan en el hombre, comparadas con un amplio conjunto de aproximadamente 16.000 proteínas de referencia del proteoma humano. Encontramos que las tasas evolutivas de las proteínas amiloidogénicas podrían ser moduladas por factores asociados con transiciones metaestables como sobresaturación y diversidad conformacional.

Por otro lado, también se estudió particularmente la interacción con las chaperonas, como una posible causa de este aumento en la velocidad de evolución. Para esto estimamos los caminos evolutivos, dado un árbol de 7 especies, y para cada uno de los cambios de aminoácidos desde los nodos ancestrales hasta las proteínas actuales, calculamos los cambios energéticos de estas mutaciones. Si bien encontramos algunos indicios de que el patrón de sustitución es progresivamente desestabilizante, esto es, que las proteínas progresivamente se tornan más inestables incrementando aún más su necesidad de interaccionar con chaperonas, los resultados no son concluyentes en este sentido.

2.2 Introducción

Las fibrillas de amiloides son arreglos insolubles altamente ordenados y estrechamente empaquetados que pueden adoptar la mayoría de las proteínas (Dobson, 1999), aunque con diferentes tendencias (López de la Paz and Serrano, 2004). Surgen cuando una proteína soluble o un fragmento de una proteína se agrega en una estructura cuaternaria con la forma de un filamento, donde cada monómero se adhiere a otro a través de un conjunto de hojas beta dispuestas de manera perpendicular al eje de la fibra. Este arreglo es llamado arquitectura beta cruzada (Rambaran and Serpell, 2008). Una colección de moléculas proteicas que se estructuran de esta manera forman subunidades individuales llamadas protofilamentos, cuyas asociaciones variables en número resultan en un ensamblaje supramolecular largo, recto y sin ramificaciones (Serpell et al., 2000; Harrison et al., 2007). Debido a la existencia de un red bastante poblada de enlaces de hidrógeno entre las hojas beta, las fibrillas son altamente estables y casi independientes de las características de los aminoácidos que componen su secuencia. Contrario con lo que ocurre con el plegamiento de estados nativos que muestran una gran diversidad de formas, tamaño, composición y arreglos (Parisi *et al*., 2021), las fibrillas de amiloides provenientes de diferentes proteínas son marcadamente similares.

Los amiloides han sido asociados desde hace tiempo con la ocurrencia de enfermedades en humanos. Un ejemplo conocido dentro de los alrededor de 50 desordenes similares es la acumulación de fibrillas de amiloides derivadas de ciertas proteínas, o fragmentos de las mismas, como la transtiretina en la amiloidosis transtiretina (ATTR) (da Costa *et al*., 2015), la proteína precursora de amiloides y la proteína tau en la enfermedad de Alzheimer (AD) (Murrell *et al*., 2000), la alfa-sinucleína (SNCA) en la enfermedad de Parkinson (Stefanis, 2012) y la proteína priónica (PrPc) en las encefalopatías espongiformes (Weissmann *et al*., 2002). En estos casos los amiloides pueden ocurrir intracelularmente en el citoplasma o núcleo o estar asociados con agregados particulares como los agresomas, o los cuerpos de Lewy o Russel (Stefani and Dobson, 2003). También son frecuentemente

encontrados en compartimentos extracelulares donde forman agregados más grandes denominados placas. Los amiloides también han sido asociados con el estado funcional de las proteínas (Fowler *et al*., 2007). Han sido encontrados en varias especies, incluido el humano, involucrando tanto eucariotas como procariotas. Los roles biológicos de estos amiloides funcionales son diversos, e incluyen el almacenamiento de proteínas como en varias hormonas peptídicas humanas (Maji *et al*. 2009; Jacob *et al*., 2016), la formación de biofilms y la participación en la invasión al hospedador (Louros *et al*., 2016) y la producción de melatonina (McGlinchey *et al*., 2009), entre otras. La existencia de proteínas donde las fibrillas de amiloides se comportan como otra conformación del estado nativo de algunas proteínas (Fowler *et al*., 2007; Hervás and Oroz, 2020; Jackson and Hewitt, 2017; Maury, 2009) generó preguntas sobre los factores realmente asociados al origen de las enfermedades cuando las proteínas amiloides se acumulan en la célula.

La propensión a formar agregados amiloides ha sido correlacionada con diferentes determinantes de secuencia como la hidrofobicidad, la propensión a hebras beta y la carga neta baja (López de la Paz and Serrano, 2004). La acumulacion de residuos con estas propiedades en una o varias regiones cortas (regiones propensas a agregarse, APR por sus siglas en inglés *amyloid propensity regions*) fue extensamente utilizado para predecir la propensión a agregación amiloide a nivel proteómico con TANGO (Fernandez-Escamilla *et al*., 2004; Maurer-Stroh *et al*., 2010; Zambrano *et al*., 2015) (Ventura *et al*., 2004; Walsh *et al*., 2014). La mayoría de los APRs que ocurren en proteínas globulares están escondidos en el núcleo hidrofóbico de la estructura nativa. Sin embargo diferentes procesos, como la ocurrencia de mutaciones, estrés fisiológico, pérdida de estabilidad, o disfunción de la capacidad de la proteína para detectar proteínas mal plegadas generan que estos APRs queden expuestos al solvente, aumentando las chances de que se estabilicen auto ensamblándose para formar fibrillas de amiloides (Beerten *et al*., 2012; Buck *et al*., 2013). Investigaciones extensas también se han enfocado en los APRs y el rol del resto de la proteína para influenciar la capacidad para transformar la proteína en amiloides, como la tendencia de los residuos de formar estructuras globulares ordenadas o la ocurrencia de

regiones desordenadas (Tartaglia *et al*., 2008; Zhang *et al*., 2013) o estudios relacionando la estabilidad termodinámica de los APRs en la forma globular y amiloide de la proteína (Langenberg *et al*., 2020).

En el presente trabajo de tesis, estudiamos las tasas evolutivas de proteínas humanas con capacidades reportadas para formar amiloides funcionales y patológicos. Usamos diferentes conjuntos de ortólogos, cubriendo desde ~8000 a ~16000 proteínas del proteoma humano. Luego de remover varios factores que introducen ruido en la interpretación de la señal evolutiva (típicamente denominados en inglés "*confounding factors"*) encontramos que las proteínas formadoras de amiloides están dentro de las proteínas más rápidas del proteoma humano. Inesperadamente, las proteínas formadoras de amiloides son ampliamente expresadas en los tejidos humanos y también se encuentran de manera abundante tanto a nivel de proteínas como de sus RNA mensajeros. Esta información no concuerda con la observación ya bien establecida de que las proteínas de alta expresión evolucionan lento (Drummond *et al*., 2006; Park and Choi, 2009; Zhang and Yang, 2015; Drummond *et al*., 2005). Las tasas de evolución de los APRs son más lentas comparadas con el resto de las proteínas en los amiloides patológicos pero no en los amiloides funcionales. La influencia de las chaperonas en los amiloides podría explicar las tasas evolutivas más rápidas observadas al permitir la aceptación de sustituciones con menos efectos estabilizantes (Tokuriki and Tawfik, 2009; Agozzino and Dill, 2018). Nuestros resultados muestran que un parámetro global como el de la tasa evolutiva puede diferenciar proteínas que adoptan conformaciones amiloides en humanos y puede ofrecer una explicación mecanicista sobre la propensión a amiloides de estas proteínas.

2.3 Resultados

Usando el conjunto de datos de ortólogos humano-ratón encontramos un mayor dN (tasa de sustitución no sinónima por sitio. Ver Métodos específicos del capítulo) para las proteínas amiloides (Wilcoxon rank-sum test p-value < 0.01) comparado al del conjunto de datos de referencia (Fig. 2.1). (Todos los resultados mostrados aquí son para el conjunto de datos de humano-ratón, que son representativos de los resultados obtenidos para los otros conjuntos (ver Métodos específicos del capítulo). Inesperadamente para estas proteínas de rápida evolución, los amiloides también mostraron altos niveles de expresión comparados con las proteínas de referencia (Wilcoxon rank-sum test p-value < 0.05), así como también alta abundancia (Wilcoxon rank-sum test p-value < 0.001, Fig 2.1).

Como la localización extracelular (Feyertag *et al*., 2017) y la presencia de puentes disulfuro (Feyertag and Alvarez-Ponce, 2017) han sido también encontrados como factores que caracterizan a las proteínas de rápida evolución, dividimos el conjunto de datos basados en estas variables. Anotamos a las proteínas como contenedoras de puentes disulfuro siguiendo las anotaciones UniProt y usamos la base de datos MetazSecKB (Meinken *et al*., 2015) para definir subgrupos de proteínas relacionadas a membrana, y proteínas intra y extracelulares. Encontramos que las proteínas amiloidogénicas evolucionan igual de rápido que las proteínas secretadas (con o sin puentes disulfuro) (Wilcoxon rank-sum test P > 0.01). Además no observamos ninguna influencia de los puentes disulfuro en las tasas evolutivas de los conjuntos de datos amiloides (Fig. 2.2)

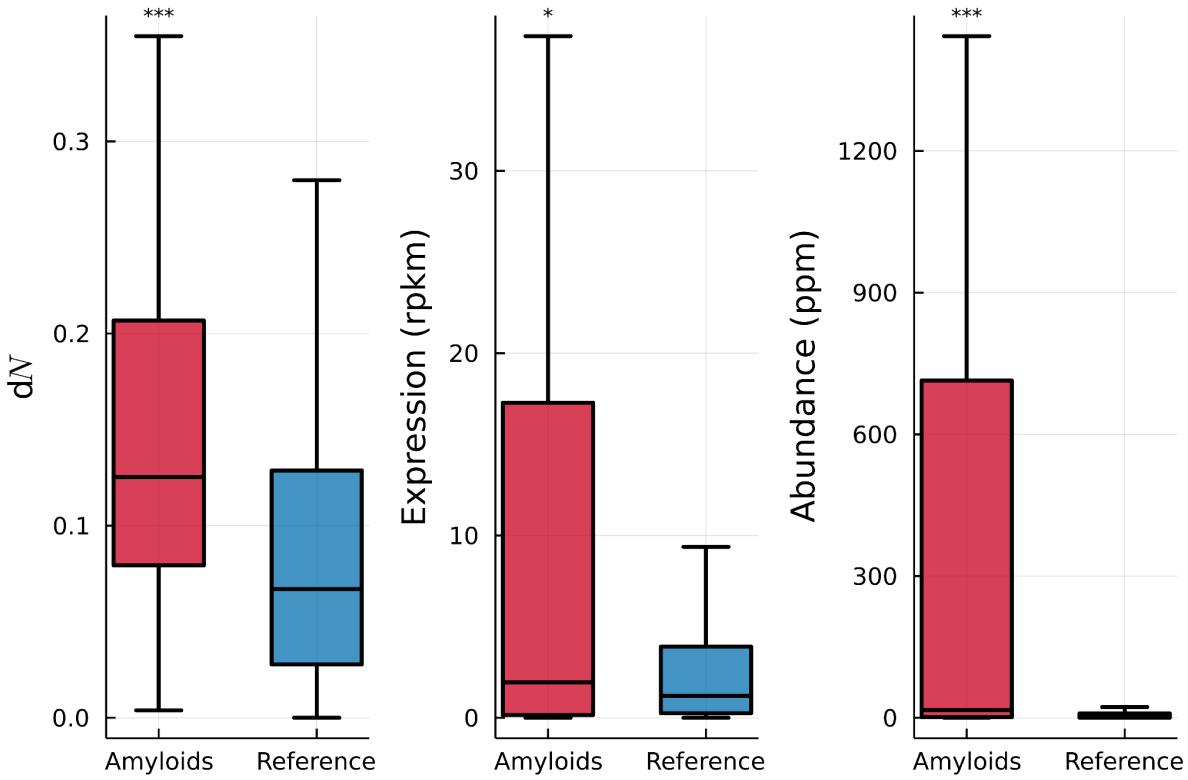


Figura 2.1. Las proteínas amiloidogénicas evolucionan más rápido que el resto de las proteínas (panel izquierdo, Amiloides n=62, Referencia n=14.502) y se expresan más (panel central, Amiloides n=72, Referencia n=17.087) y son más abundantes (panel derecho, Amiloides n=77, Referencia n=17.345). Las estrellas indican los resultados de los test Wilcoxon rank-sum: ***: p-value<= 0.001; ** p-value <= 0.01; * p-value <= 0.05; ns: p > 0.05.

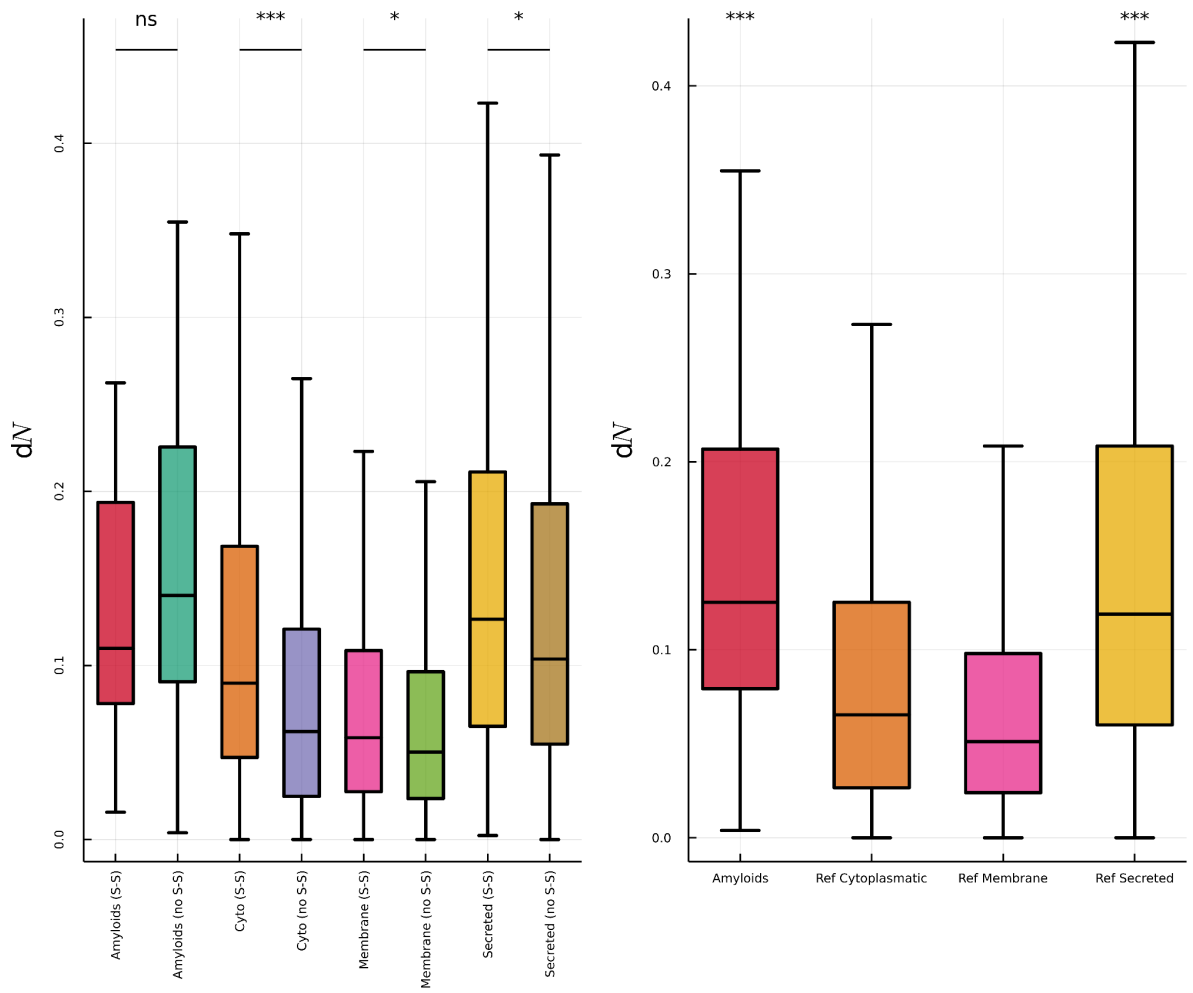


Figura 2.2 Los puentes disulfuro (S-S) no afectan relativamente las tasas evolutivas de las proteínas amiloides, por lo tanto estos conjuntos de datos son indistinguibles entre sí, pero si afectan las tasas del resto de los datos del conjunto de referencia (panel izquierdo: Amiloides con S-S = 30, Amiloides sinS-S = 27, Citoplasmáticas con S-S = 1303, Citoplasmáticas sin S-S = 9751, Membrana con S-S = 274, Membrana sin S-S =1468, Secretadas con S-S = 860, Secretadas sin S-S = 380). Sin embargo, los amiloides y las proteínas secretadas en general evolucionan más rápido que las proteínas de membrana y citoplasmáticas, pero no se evidencia diferencia estadística entre ellas (panel derecho: ambas poblaciones son rápidas cuando se comparan con proteínas de membrana y citoplasmáticas). Las estrellas indican los resultados de los test Wilcoxon rank-sum: ***: p-value<= 0.001; ** p-value <= 0.01; * p-value <= 0.05; ns: p > 0.05. "Ref" se refiere al conjunto de datos de referencia. "S-S" y "no S-S" indican las proteínas con y sin puentes disulfuro respectivamente.

Las proteínas secretadas son sintetizadas en el retículo endoplasmático. En el proceso de secreción interactúan con chaperonas y enzimas asistentes del plegado para promover el correcto plegamiento y reconocer proteínas mal plegadas (Braakman and Hebert, 2013). Estos controles de calidad han sido destacados como responsables de la falta de correlación entre los niveles de expresión (o abundancia de proteína) y las tasas evolutivas de las proteínas secretadas, probablemente debido a la presión relajada en los errores en el

proceso de traducción (Feyertag *et al*., 2017). Los mismos mecanismos de control podrían ser responsables por las rápidas tasas evolutivas de las proteínas secretadas (Fig. 2.2) debido al efecto buffer de las sustituciones levemente desestabilizantes impuestas por las chaperonas en sus clientes (Tokuriki and Tawfik, 2009; Williams and Fares, 2010; Alvarez-Ponce *et al*., 2019). Como el 70% de las proteínas amiloidogénicas de nuestro conjunto de datos son proteínas secretadas, asumimos que el mismo mecanismo podría también explicar sus tasas evolutivas rápidas. Está bien documentado que mecanismos de control de calidad similares se aplican a amiloides citoplasmáticos; por ejemplo los sistemas de chaperonas se activan para evitar toxicidad fibrilar amiloide, evitando y en algunos casos revirtiendo la formación de fibrillas (Yerbury *et al*., 2007; Wilson *et al*., 2008; Månsson *et al*., 2014; Landreh *et al*., 2015; Wentink *et al*., 2019; Hervás and Oroz, 2020). En concordancia, no encontramos una diferencia estadística significativa entre las tasas evolutivas de las proteínas amiloidogénicas secretadas y citoplasmáticas (Wilcoxon rank-sum test, p-value = 0.8061).

¿Por qué las proteínas amiloidogénicas se agregan y las secretadas no, considerando que ambas están bajo un fuerte control de calidad? Como la estabilidad de proteínas ha sido usada para caracterizar la agregación y el mal plegado (Knowles *et al*., 2014), mapeamos ambos conjuntos de proteínas en la base de datos ProTherm (Kumar *et al*., 2006). Usando energías libres de Gibbs de desplegamiento (ΔG^0) y temperaturas de fusión (Tm) como medida de estabilidad proteica (Becher *et al*., 2018), no encontramos ninguna diferencia significativa entre las proteínas amiloidogénicas y las secretadas. Además, como derivado de la ocurrencia APR (estimada con predicciones TANGO) encontramos resultados contra intuitivos porque las proteínas secretadas muestran el mismo contenido APR que el conjunto de datos amiloide. Aunque las energías libres de Gibbs de desplegamiento y las temperaturas de fusión han sido ampliamente utilizadas para estimar estabilidad proteica y la tendencia a desplegarse y agregarse de las proteínas, medir dichas tendencias es más complejo debido a la naturaleza metaestable de las proteínas (Baldwin *et al*., 2011). La

metaestabilidad es una condición que caracteriza a un estado de la proteína que es cinéticamente estable pero tiene una tendencia termodinámica a cambiar. El estado nativo de las proteínas se describe como un conjunto de estructuras que están cinéticamente atrapadas en el mínimo local de energía libre bajo condiciones fisiológicas. Sin embargo, la metaestabilidad significa que las barreras cinéticas pueden ser sobrepasadas bajo ciertas condiciones celulares que permiten al estado nativo adoptar arreglos más estables, como las fibrillas amiloides (Baldwin *et al*., 2011) o conformaciones cinéticamente controladas asociadas con la función proteica (Lee *et al*., 2000; Ghosh and Ranjan, 2020). Como los diagramas de energía de Gibbs dependen fuertemente de la concentración de proteínas, podemos usar la abundancia de proteínas y sus niveles de expresión como indicadores de la agregación proteica (Ciryam *et al*., 2015; Buell, 2022). Además, los grandes cambios conformacionales y en flexibilidad, las cavidades con interacciones inusuales entre residuos, las modificaciones postraduccionales, y el clivado de proteínas han sido asociados con transiciones a estados metaestables (Ghosh and Ranjan, 2020). La alta concentración de proteínas puede también aumentar la concentración absoluta de conformaciones poco pobladas, quizás regiones APR expuestas y parcialmente desplegadas, que podrían disparar la formación de fibrillas (Faravelli *et al*., 2022). Tomando los niveles de expresión, el desorden (o flexibilidad proteica), la diversidad conformacional de las proteínas y la abundancia de proteínas como indicadores de la propensión a agregación, encontramos que estos parámetros son significativamente mayores en el conjunto de datos amiloide que en las proteínas secretadas (Fig. 2.3). Como en trabajos anteriores (Zea *et al*., 2013), medimos la diversidad conformacional de la proteína usando el máximo normalizado RMSD (max RMSD100, Carugo and Pongor, 2001) calculado entre todos los pares de sus confórmeros estructurales conocidos obtenidos de la base de datos CoDNaS (Monzon *et al*., 2016). Cuando las correlaciones parciales fueron estimadas para controlar el efecto de dN sobre diferentes parámetros del tipo *confounding factors* en el conjunto de datos amiloides (n = 25), usando valores transformados a logaritmo encontramos correlaciones significativas de dN con RMSD100 (Pearson's $\rho = 0.49$, p-value < 0.05, abundancia

(Pearson's ρ = 0.47, p-value < 0.05) e interacciones proteína-proteína (Pearson's ρ = -0.495, p-value < 0.05).

Cuando la abundancia de proteínas y el max RMSD100 fueron combinados en un modelo de regresión lineal usando las transformaciones logarítmicas de las variables encontramos que explican ~24% de la varianza de dN (p-value <0.02). La varianza observada en dN aumentó a ~26% al incluir las interacciones proteína-proteína en el modelo (p-value = 0.073).

Como se explicó en este trabajo, la diversidad conformacional y la abundancia de proteínas pueden ser tomadas como indicadores de transiciones de metaestabilidad. También analizamos la influencia de la sobresaturación, un parámetro más específico asociado a las transiciones en proteínas metaestables. El valor de sobresaturación (σf y σu para los estados plegados y desplegados de las proteínas respectivamente) es una combinación lineal entre la tendencia a agregación y la concentración de proteínas. Identifica proteínas con alta tendencia a agregarse y fue usado exitosamente para marcar agregación patológica (Ciryam *et al*., 2013). Las proteínas se denominan sobresaturadas cuando sus concentraciones exceden sus niveles de solubilidad, aumentando la propensión a agregarse. Encontramos que σf, pero no σu, es mayor en el conjunto de datos amiloides comparado con el de las proteínas secretadas. También observamos una alta correlación entre dN y σf (Spearman's ρ=0.639, p-value < 0.01, n= 17, power = 0.81). Aunque el número de proteínas amiloidogénicas es reducido debido a la disponibilidad de σf, cuando agregamos σf al modelo previo encontramos que la varianza explicada en dN fue ~38% (Adjusted R^2 = 0.388, p-value = 0.012).

Aplicando el mismo modelo a proteínas secretadas se obtuvo un modelo lineal no significativo (Adjusted R^2 = 0.008, p-value = 0.198).

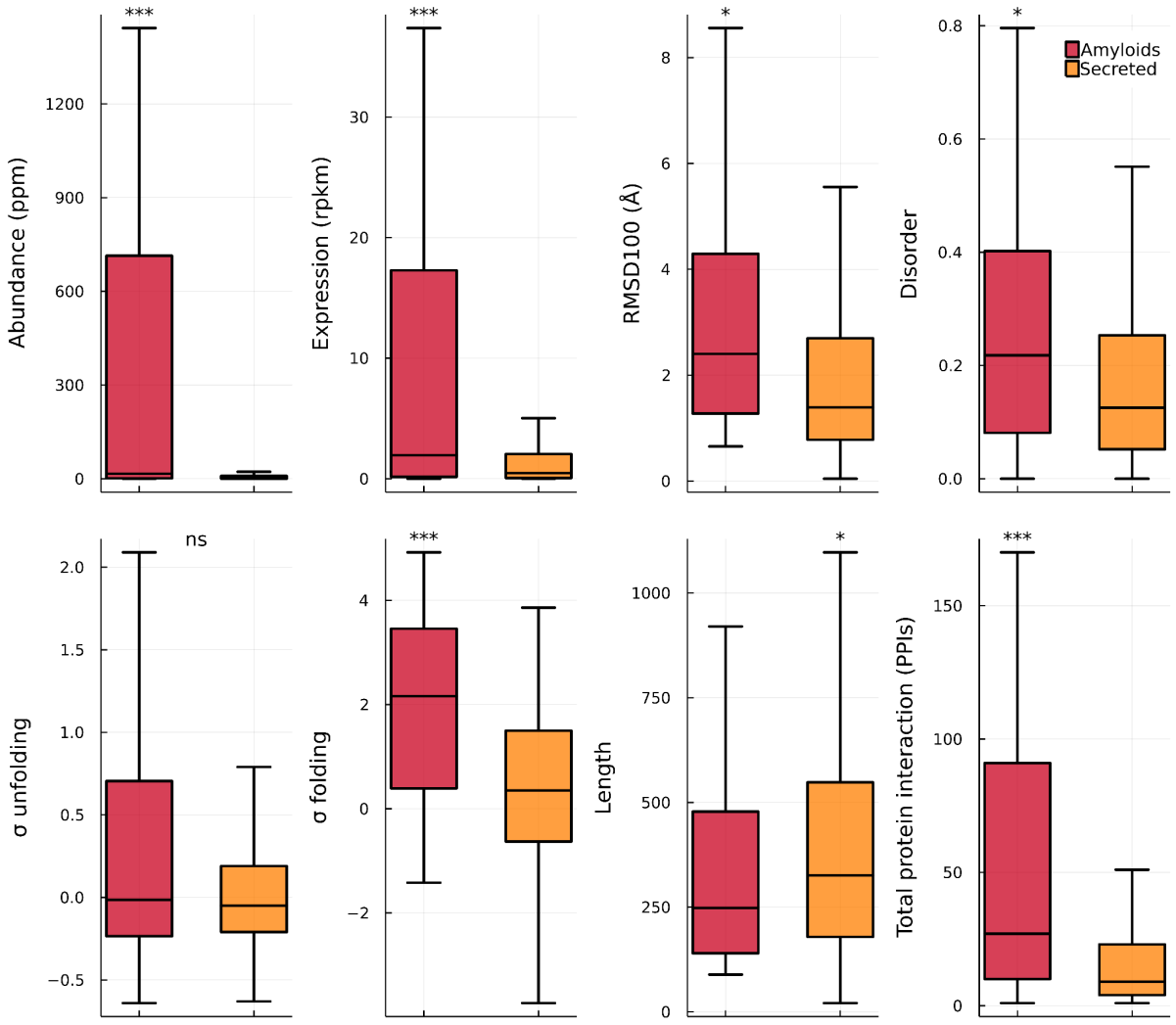


Figura 2.3. Diferencias observadas entre proteínas amiloides y secretadas (Amiloides = 77, Secretadas = 1829, excepto por RMSD100, Amiloides = 34, Secretadas = 389). Aunque tenían tasas evolutivas similares, los amiloides se comportan diferente que las proteínas secretadas con respecto a las variables que están correlacionadas a las tasas evolutivas. Las estrellas indican los resultados de los test Wilcoxon rank-sum: ***: p-value<= 0.001; ** p-value <= 0.01; * p-value <= 0.05; ns: p > 0.05

2.3.1 Uso de ASR y evolución de amiloides

Inicialmente, quisimos estudiar particularmente el efecto de las chaperonas en la estabilidad de las proteínas amiloides con un enfoque vertical, es decir, estudiando los ancestros, tanto de los amiloides como de las proteínas de referencia. Para esto utilizamos el conjunto de datos correspondiente al árbol de 7 especies, donde se realizó para cada proteína del conjunto de datos (ya sea amiloide o de referencia) un alineamiento múltiple con las secuencias correspondientes de cada especie. De esta forma, en conjunto con el árbol filogenético de las 7 especies, se realizó un ASR para cada proteína, infiriendo de esta forma las secuencias de los ancestros. Paso seguido se obtuvo el camino evolutivo, es decir, las mutaciones que ocurrieron en orden desde la raíz hasta la proteína humana. Una vez obtenidas estas mutaciones se procedió a calcular el cambio en la diferencia de energía libre de Gibbs de plegamiento de la proteína mutada con respecto a la diferencia en la energía libre de Gibbs de plegamiento de la proteína *wild-type* (es decir, el ΔΔG de plegamiento) usando el *software* FoldX (Delgado, 2019). Este cálculo se realizó para cada mutación y para cada proteína, siempre considerando el camino desde la raíz a la proteína actual humana. Una vez obtenidos todos los valores, se separó al conjunto en 3 grupos: amiloides, proteínas de referencia que interaccionan con chaperonas y proteínas de referencia que no interaccionan con chaperonas y se graficó la distribución de cada grupo, obteniendo un comportamiento diferencial de los amiloides y las proteínas que interaccionan con chaperonas, con respecto a las proteínas de referencia que no interaccionan con chaperonas (Fig 2.4). Es decir, los amiloides y las proteínas que interaccionan con chaperonas tenían un comportamiento similar, dándonos un indicio de que los amiloides, además de evolucionar más rápido, también permitían mutaciones desestabilizantes debido al *buffer* otorgado por las chaperonas.

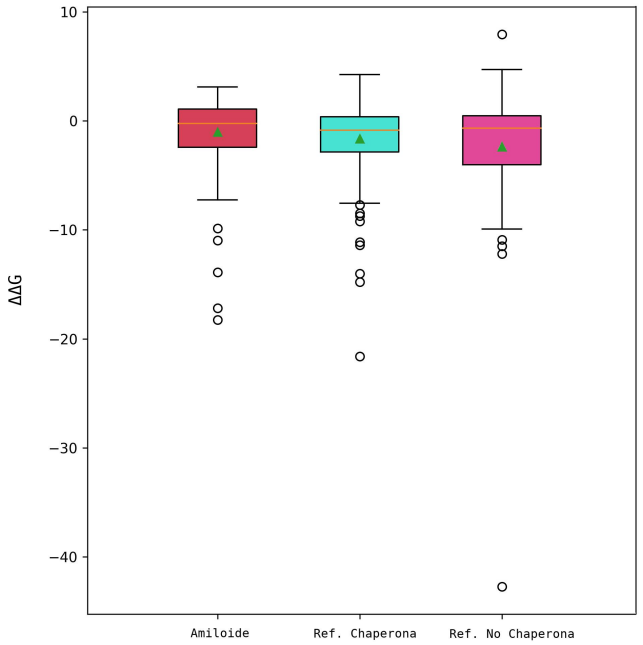


Figura 2.4. Distribución de los valores de ddG de distintas poblaciones. Cada columna contiene la información de los valores de ddG para todas las mutaciones ocurridas desde el ancestro raíz al ser humano. (Amilode: proteínas amiloidogénicas; Ref. Chaperona: proteínas de referencia que interaccionan con chaperonas; Ref. No Chaperona: proteínas de referencia que no interaccionan con chaperonas). No se observan diferencias entre las distribuciones de amiloides y proteínas que interaccionan con chaperonas (p-value: 0.321), pero sí entre cada una de estas dos y las proteínas que no interaccionan con chaperonas (p-value: 0.0058 y 0.012, respectivamente).

Por último, a continuación se estudiaron en particular los nodos ancestrales por separado, para poder determinar con más precisión cómo se modificó este comportamiento en los amiloides (Fig 2.5). Sin embargo, al momento de analizar estos resultados, descubrimos que no existía diferencia significativa entre los amiloides y ninguno de los otros grupos en cada nodo, resultado probable que las diferencias vistas de manera global sean pequeñas acumulaciones.

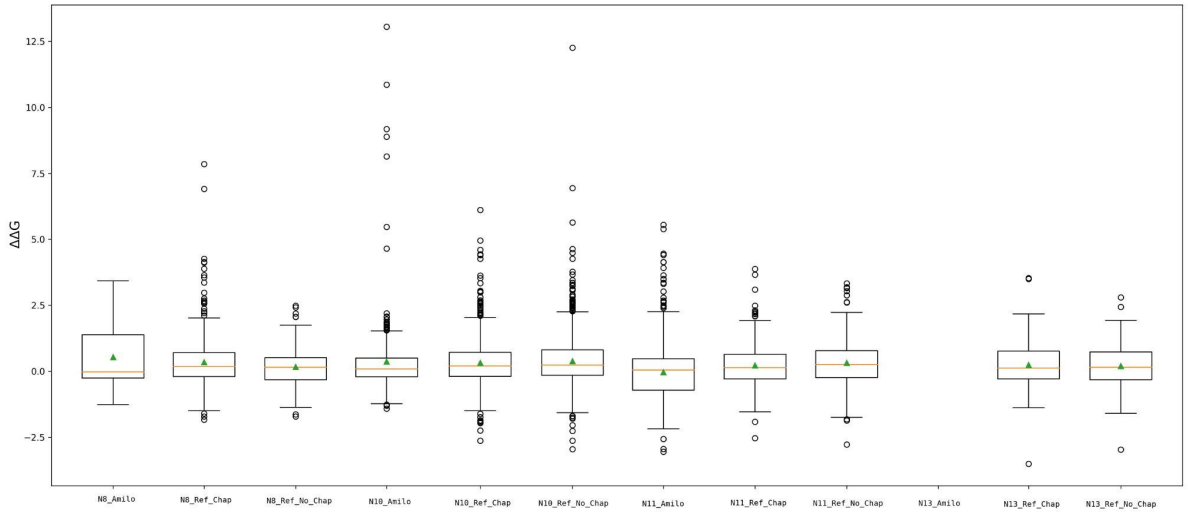


Figura 2.5. Distribución de los valores de ddG de distintas poblaciones. Cada columna contiene la información de los valores de ddG para todas las mutaciones ocurridas en cada nodo desde el ancestro raíz al ser humano. (Referencias similares a la figura 2.4, indicando al inicio de los nombres el número de nodo). Para este caso no se observan diferencias significativas.

2.4 Discusión

Hemos explorado las tasas evolutivas de una población de proteínas humanas (n=81) con evidencia experimental de agregación como fibrillas amiloides. Encontramos que las proteínas amiloidogénicas evolucionan más rápido que el conjunto de datos de referencia, a pesar de ser abundantes y con alta expresión. Aunque no observamos diferencias significativas en las tasas evolutivas de las proteínas secretadas y amiloides (Fig. 2.2), encontramos diferencias sustanciales en otras características. En primer lugar, la diversidad conformacional es mayor en proteínas amiloidogénicas, también evidenciado por la alta presencia de regiones desordenadas o altamente flexibles (Fig. 2.3). Esta diversidad conformacional aumentada podría aumentar las chances de exponer APRs en confórmeros levemente desplegados, dirigiendo a la proteína hacia la formación de fibrillas. En segundo lugar, las proteínas amiloidogénicas son más abundantes y más expresadas que las proteínas de secreción (Fig. 2.3). La concentración y la solubilidad de proteínas y sus efectos en la formación de amiloides fueron estudiados ampliamente con anterioridad (Ciryam *et al*., 2013). Para permanecer soluble, las proteínas abundantes requieren un apoyo constante de mecanismos de control de calidad como lo son las chaperonas moleculares. Esta asistencia y la participación bien documentada de las chaperonas en evitar y revertir la formación de amiloides podría explicar la tasa de evolución acelerada en las proteínas amiloidogénicas, seguido de los efectos buffer de las sustituciones levemente desestabilizantes (Tokuriki and Tawfik, 2009). Más aún, la correlación positiva entre el máximo RMSD100 y las tasas evolutivas indica que un aumento en la diversidad conformacional acelera las tasas evolutivas, compatible con una alta propensión a agregación. En tercer lugar, un modelo lineal combinando un parámetro proteico intrínseco como una medida de la diversidad conformacional (por ejemplo RMSD100) con una condición celular (como el valor de sobresaturación) explica mejor la variación en las tasas evolutivas observada en las proteínas amiloidogénicas pero no en las secretadas.

Evidencia reciente sugiere que las proteínas amiloidogénicas representan un proteoma metaestable, fuertemente dirigido hacia la formación de fibrillas amiloides. En este conjunto de proteínas particular, las propiedades intrínsecas y las condiciones celulares pueden liberar la traba cinética de sus estados nativos hacia formas más estables como fibrillas amiloides. Nuestros resultados muestran que las tasas evolutivas reflejan este comportamiento particular, mostrando la importancia de la metaestabilidad sobre otros factores moduladores. En el futuro sería interesante evaluar la metaestabilidad proteica como un factor de modulación general de las tasas evolutivas a nivel proteómico (Ciryam, 2015; Kundra, 2017; Ciryam, 2019).

2.5 Métodos específicos del capítulo

Hemos obtenido 81 proteínas amiloidogénicas humanas de la base de datos Amypro (Varadi *et al*., 2018). Amypro ofrece un conjunto de proteínas curado a mano con evidencia experimental de la capacidad de una proteína de formar amiloides. En Amypro también se encuentra disponible la anotación de los APRs para cada proteína. Para cada proteína derivada de Amypro, buscamos ortólogos en la base de datos OMA (Altenhoff *et al*., 2018) y descargamos los ortólogos uno a uno optimizando la ocurrencia de las 81 proteínas amiloidogénicas junto con el máximo número de proteínas humanas no pertenecientes a ese conjunto. Para estimaciones de las tasas evolutivas en proteínas obtuvimos un conjunto (*set* 1) con 14297 proteínas con sus correspondientes proteínas ortólogas en 5 especies (*Homo sapiens, Cercocebus atys, Macaca fascicularis, Pan troglodytes and Rhinopithecus bieti*). Este conjunto de datos contenía 12 y 33 proteínas amiloides funcionales y patológicas, respectivamente, y el resto de las proteínas (14250) formaron el conjunto de referencia. De la misma manera, obtuvimos otros dos conjuntos (*set* 2 y *set* 3) con 7 y 12 especies y con 8011 y 11789 proteínas, respectivamente. En todos los casos, las tasas evolutivas fueron obtenidas usando el programa Rate4site (Pupko *et al*., 2002) y el promedio por proteína usando las tasas por sitio no normalizadas fueron derivadas para cada proteína. Dos programas de alineamientos diferentes fueron usados para cada conjunto (T-coffee (Notredame *et al*., 2000) y Muscle (Edgar, 2004)) originando respectivamente los *set* 1a y *set* 1b. La misma nomenclatura fue usada para el resto de los conjuntos. Para el análisis de las regiones codificantes trabajamos con dos conjuntos, uno fue obtenido usando las secuencias codificantes del *set* 1 (al que llamamos *set* 4) y el programa Codeml fue utilizado para estimar dN. El otro conjunto de cADN (*set* 5) se conformó por 16478 alineamientos de a pares entre ortólogos de humanos y de ratón, también obtenidos de OMA. El *set* 5 fue usado para estimar dN con el programa Yn00. Codeml y Yn00 son parte del paquete PAML (Yang, 1997). Como la tasa evolutiva puede ser influenciada por varios factores (Zhang and Yang, 2015; Rocha, 2006), utilizamos el

análisis de regresión de componente principal (PCR por sus siglas en inglés, utilizando tasas de evolución como una variable dependiente) para estimar la contribución relativa de cada factor (Drummond *et al*., 2006). Una extensa lista de de variables putativas confusas han sido estudiadas (largo, presencia de desorden, nivel de expresión, abundancia de proteína, interacción proteína-proteína, presencia de genes de mantenimientos, presencia de genes relacionados a enfermedades).

Para el proceso de reconstrucción de las secuencias ancestrales se utilizó el software ClustalO para realizar los alineamientos múltiples. Luego se utilizó PAML para obtener los estados ancestrales, usando como modelo de evolución JTT con una distribución Gamma de 4 categorías de velocidades. Finalmente, para el cálculo del ΔΔG de plegamiento se usó el comando BuildModel del software FoldX.

3 Caracterización de una tiorredoxina atípica por reconstrucción ancestral

3.1 Resumen

Este capítulo describe los resultados sobre el estudio de una tiorredoxina atípica del parásito cestodo *Echinococcus granulosus*, la proteína EgIsTRP (*E. granulosus Iron Sulfur Thioredoxin Protein*), la cual posee propiedades biológicas únicas (Bisio *et al*., 2016). Se ha demostrado que EgIsTRP no tiene función oxidorreductasa y por el contrario funciona como una proteína que forma *clusters* de hierro y azufre. El estudio de esta proteína surgió de una colaboración con el grupo del Dr. Massimo Bellanda, Universidad de Padova, Italia. Con el objetivo de caracterizar el origen de su cambio de función utilizamos métodos de reconstrucción ancestral para evaluar los cambios de aminoácidos y lograr así una explicación mecanística de su posible adaptación biológica. Para esto construímos árboles filogenéticos a partir de proteínas homólogas cercanas para estimar secuencias ancestrales. Además, analizamos el *ensemble* proveniente de la caracterización por técnicas de Resonancia Magnética Nuclear (RMN) en el contexto de la colaboración con el Dr. Bellanda. Nuestra hipótesis de trabajo propone que EgIsTRP posee una dinámica y flexibilidad atípicas en esta familia de proteínas, con la principal consecuencia de producir un cambio en el pka de la Cys amino terminal, esencial para la actividad oxidorreductasa. Para probar esta hipótesis nos propusimos utilizar reconstrucción ancestral para resucitar estados ancestrales de EgIsTRP, sintetizar estos estados ancestrales, expresarlos, purificarlos y caracterizarlos.

3.2 Introducción

La familia de las tiorredoxinas es una familia numerosa de proteínas mayormente asociada a reacciones redox en una forma tiol-dependiente (Holmgren, 1989). Se encuentran en la gran mayoría de los organismos y resulta esencial para los mamíferos. En los mamíferos la tiorredoxina actúa en vías relacionadas con la proliferación celular y la expresión génica (Atkinson and Babbitt, 2009). El plegamiento de la tiorredoxina se encuentra altamente conservado y consiste en una hoja plegada beta de 4 láminas rodeada de 3 o 4 hélices alfa (Fig. 3.1). Según la base de datos CATH (Orengo *et al.*, 1999), la superfamilia de la tiorredoxina contiene ~20200 especies diferentes, ~335 clusters secuenciales (>35%identidad) y 1250 términos GO, lo que habla de la gran extensión filogenética y variabilidad funcional (Fig. 3.2). Por su parte, la base de datos Interpro en la familia de las tiorredoxina contiene ~55000 secuencias y ~10000 especies distintas que contienen dicho dominio.

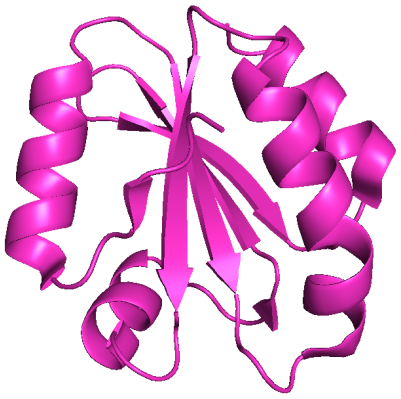


Figura 3.1 Tiorredoxina humana (PDB ID = 4POK) mostrando su plegamiento típico.

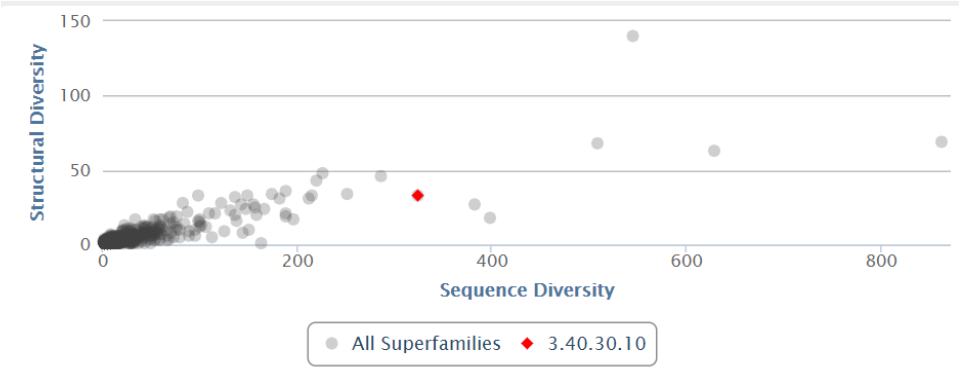


Figura 3.2 El gráfico muestra la diversidad secuencial y estructural del dominio correspondiente a la tiorredoxina. El mismo fue extraído de la base de datos CATH (htt://cathdb.info) y muestra la variabilidad de la superfamilia estructural de la tiorredoxina (punto rojo) con respecto al resto de las superfamilias en CATH. Como vemos, la superfamilia de la tiorredoxina se caracteriza por poseer proteínas con gran diversidad secuencial y cierta variabilidad estructural.

Las proteínas con el plegamiento tiorredoxina poseen dos características conservadas: un motivo CXXC (dos cisteínas separadas por dos residuos) ubicado en el loop o al comienzo de la hélice α1 y una cis-prolina adyacente a la hebra β3. El motivo CXXC o sus variantes son habitualmente los residuos del sitio activo en las tiorredoxinas y estas dos cisteínas dan las bases de la actividad redox, dependiente de la presencia de un puente disulfuro formado entre las mencionadas cisteínas. La prolina adyacente a la hebra β3 otorga estabilidad estructural al plegamiento y participa en el posicionamiento del sustrato en el sitio activo para promover la catálisis. Las tiorredoxinas contienen dos residuos cargados conservados (un ácido aspártico y una lisina ubicados en la cadena β2 y la cadena β3, respectivamente). Son parte de una región cargada presente entre la hoja β y la hélice α2 retorcida. Esta región está protegida del medio por el disulfuro presente en el estado oxidado de la proteína. Este ácido aspártico en algunos casos actúa como el residuo clave que activa como nucleófilo a la cisteína C-terminal del motivo CXXC. La enzima tiorredoxina reductasa mantiene reducida a la tiorredoxina a expensas de la oxidación del NADPH. Por otra parte, una enzima íntimamente relacionada con la tiorredoxina y que pertenece a la misma superfamilia estructural (Lillig *et al.*, 2008), la glutaredoxina, utiliza glutation como cofactor para mantener el estado redox adecuado, en vez de NADPH (Fig.

3.3). Estas enzimas también deben su comportamiento redox a la presencia de un puente disulfuro; sin embargo, la clase II de glutaredoxinas no posee actividad oxidoreductasa (dependiente de puente de disulfuro) y su rol principal está ligado a la homeostasis del hierro (Lill *et al.*, 2012). Por ejemplo, la glutaredoxina de mitocondria de*Saccharomyces cerevisiae* denominada Grx5 perteneciente a la clase II, es indispensable para ensamblar los componentes hierro-azufre, participa del transporte de Fe/S y es esencial para el ensamblado de apoproteÍnas (Lill, 2009).

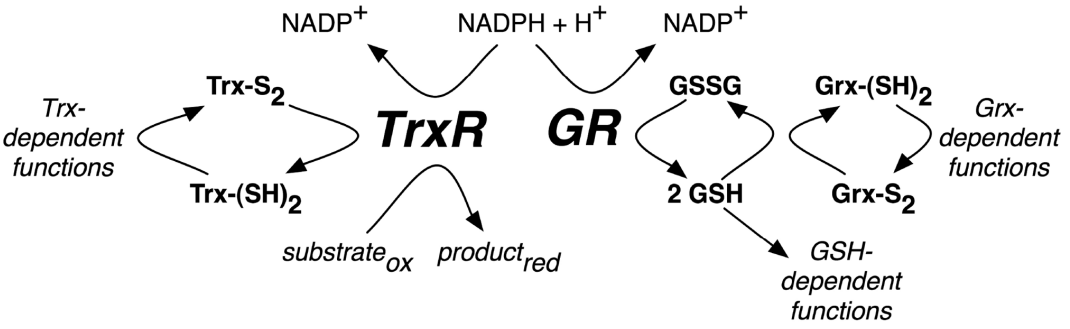


Figura 3.3 En los mamíferos (vía superior), los electrones de NADPH se transfieren a una oxidoreductasa flavoenzima, ya sea tiorredoxina reductasa (TrxR) o glutatión reductasa (GR). Luego, los electrones se transfieren de la oxidorreductasa flavoenzima al portador de electrones apropiado, ya sea tiorredoxina oxidada (Trx-S2) o disulfuro de glutatión (GSSG) convirtiéndolos en tiorredoxina reducida (Trx- [SH] 2) o glutatión (GSH), respectivamente. Trx-(SH)2 y GSH luego suministran equivalentes reductores para varias reacciones diferentes, incluidas aquellas que dependen de la glutaredoxina (Grx). Figura extraída de (Kuntz *et al.*, 2007).

Recientemente se ha descrito una tiorredoxina proveniente del helminto *Echinococcus granulosus* con características biológicas atípicas, EgIsTRP, que sería un ortólogo lejano de Grx5, ya que la función principal de la EgIsTRP no es funcionar como una oxidoreductasa, sino que forma *clusters* Fe/S y estaría relacionada con el almacenamiento y regulación del Fe/S en los helmintos (Bisio *et al.*, 2016). Además, en consonancia con esta hipótesis también se encontró que la expresión de EgIsTRP en mutantes nulos de Grx5 de *S. cerevisiae* revierte el fenotipo. En términos generales, las proteínas con plegamiento tiorredoxina sirven principalmente como oxidorreductasas de tiol/disulfuro y sólo un pequeño subconjunto es capaz de unir iones metálicos. TrxA de *Escherichia coli* y Trx1 humano son capaces de coordinar Fe/S cuando se introducen mutaciones no naturales específicas en

sus secuencias. EgIsTRP coordina Fe/S a través de los residuos Cys 34 y Cys 37, correspondientes a las posiciones homólogas de las cisteínas nucleófila y de resolución, respectivamente, de las tiorredoxinas. En el estado reducido en las tiorredoxinas, la cisteína de resolución está parcialmente ocluida, evitando la unión de Fe/S. Por lo tanto debe haber adaptaciones permanentes o inducidas en la estructura de EgIsTRP para posicionar la cisteína de resolución en la ubicación espacial correcta para coordinar la Fe/S.

El objetivo principal de este capítulo es estudiar y comprender estas adaptaciones y su relación con la flexibilidad de la proteína, que aportará información sobre los determinantes secuenciales que derivan en un cambio de función relevante para la proteína.

3.3 Resultados

3.3.1 Características secuenciales de la EgIsTRP

En búsquedas de similitud secuencial utilizando la secuencia de la EgIsTRP de *E. granulosu*s aparecen pocas proteínas homólogas cercanas con un porcentaje de identidad promedio de alrededor de 30% (*E. granulosus, Echinococcus multilocularis, Taenia asiatica, Taenia solium, Hydatigera taeniaeformis, Hymenolepis diminuta, Hymenolepis microstoma, Hymenolepis nana*) para luego aparecer secuencias de tiorredoxinas canónicas mucho más alejadas evolutivamente. Las secuencias recuperadas de estas búsquedas utilizando los algoritmos Blast y PsiBlast por un lado (para recuperar proteínas a partir de base de datos de proteínas) y TBlastN por otro (para recuperar proteínas que estén anotadas en genomas o transcriptomas), fueron alineadas utilizando el programa ClustalO. El alineamiento resultante se muestra en la (Fig. 3.4).

Lo más importante para destacar de la información secuencial, es que la proteína EgIsTRP y el conjunto de proteínas homólogas cercanas difieren de las tiorredoxinas canónicas (aquellas que funcionan como oxidoreductasas en una forma tiol/puente disulfuro dependiente como se describió anteriormente). La diferencia radica en que el motivo secuencial típico y ampliamente conservado del sitio activo de las tiorredoxinas canónicas (WCGPC, o en forma más general WCXXC) se encuentra reemplazado por el patrón también conservado C[YF]ACC, donde los corchetes indican distintos posibles aminoácidos.

El resto del alineamiento muestra diferentes grados de conservación, mostrando una indudable relación de homología entre la EgIsTRP y sus homólogas cercanas con las tiorredoxinas canónicas. Para profundizar este estudio realizamos, con el alineamiento de la Figura (Fig. 3.4), una estimación filogenética utilizando el programa Phyml (Guindon *et al.*, 2010) (con el modelo LG+F+Gamma *distribution*). El árbol obtenido se muestra en la (Fig.3.5). Es interesante mencionar, que los organismos contenidos en el *cluster* rojo poseen el motivo homólogo a las Cys del sitio activo canónico correspondiente a CIACYF,

mientras que el azul se puede dividir en dos motivos caracterizando a dos *subclusters*: CYPCCY y CFACCF. La EgIsTRP se encuentra en este último cluster. El gran cluster verde contiene tiorredoxinas canónicas de diversos organismos.

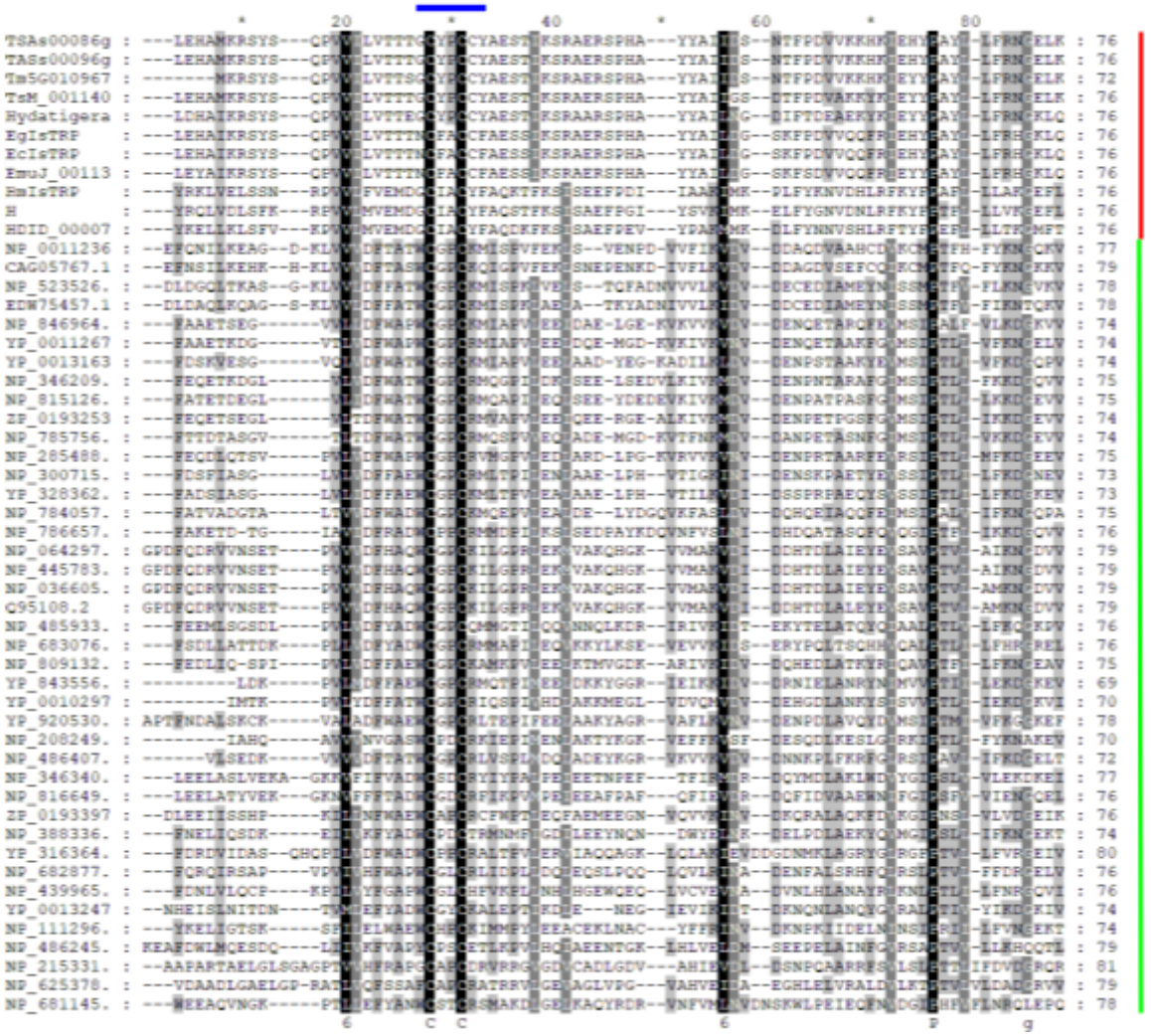


Figura 3.4 Alineamiento de las principales secuencias homólogas a EgIsTRP. En un primer bloque de secuencias (indicadas por una línea roja a la derecha) se encuentran los homólogos más cercanos a EgIsTRP mayormente de las familias *Taeniidae* y *Hymenolepididae*. El resto de las secuencias pertenecen a tiorredoxinas canónicas. El patrón secuencial característico de esta superfamilia se muestra con una línea azul sobre las secuencias. Se destacan por su conservación relativa las Cys que forman parte del sitio activo. Entre y a ambos lados de estas dos Cys se encuentran distintos aminoácidos que evidenciaría la divergencia funcional de estas proteínas.

Como mencionamos anteriormente, la EgIsTRP muestra actividad formadora de *clusters* Fe/S (Fig. 3.6) donde dos cadenas de EgIsTRP, exponiendo sus Cys al solvente,

interaccionarían con el Fe/S como lo hacen otras proteínas como la ferredoxina (Fukuyama, 2004).

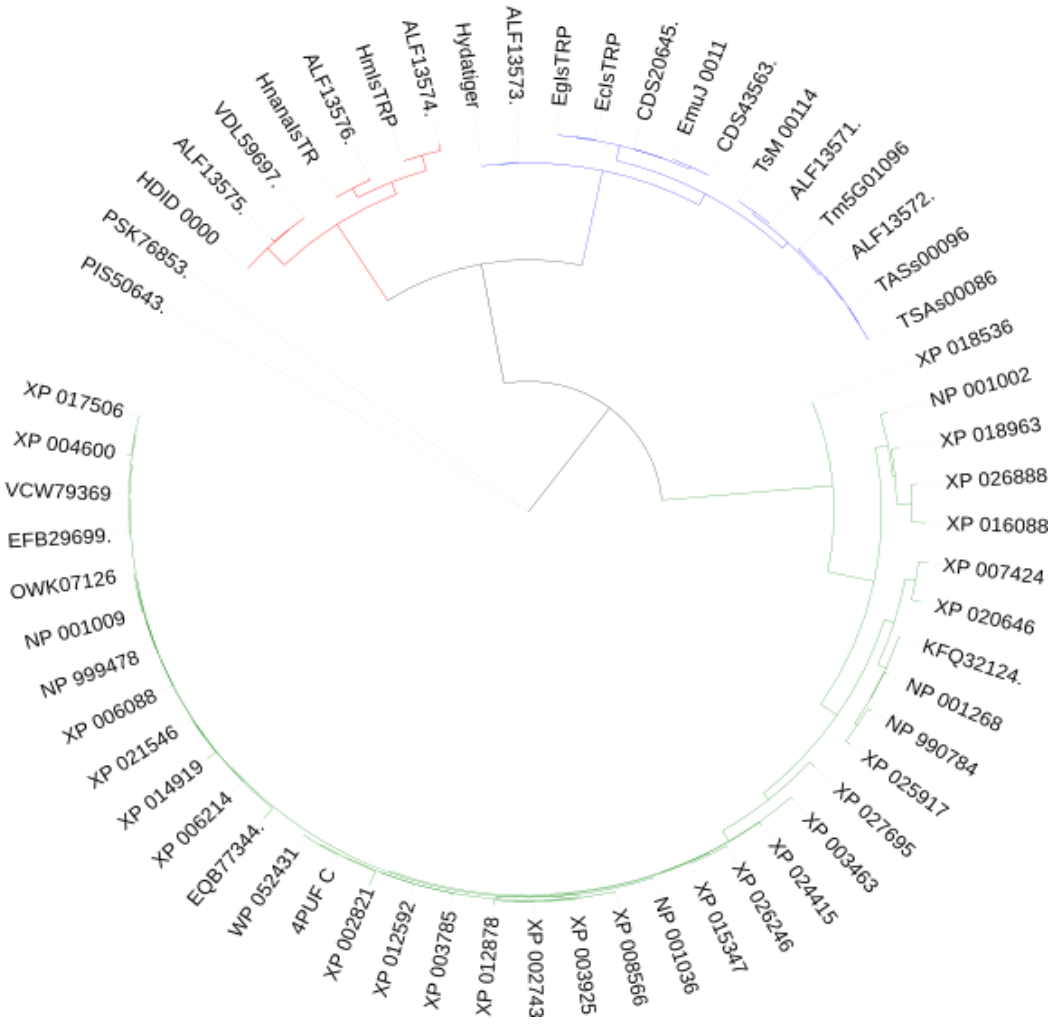


Figura 3.5 Estimación filogenética para las secuencias conteniendo las proteínas cercanas evolutivamente a EgIsTRP así como las muy alejadas evolutivamente. El cluster rojo y azul contienen las secuencias cercanas a EgIsTRP. EgIsTRP se encuentra en el cluster azul que corresponde a organismos de la Familia *Taeniidae* mientras que las proteínas homólogas cercanas se encuentran en el cluster rojo conteniendo proteínas de la Familia *Hymenolepididae*. El cluster verde, claramente diferenciado de los dos anteriores contiene tiorredoxinas canónicas de diversos organismos mayoremente vertebrados.

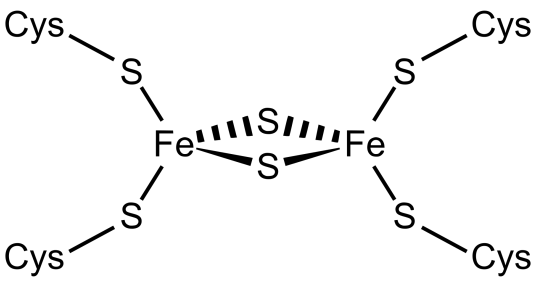


Figura 3.6 : Ejemplo de formación de clusters Fe/S como los forma la ferredoxina.

Sin embargo, un miembro cercano (*Hymenolepis diminuta* presente en el *cluster* rojo del árbol filogenético (Fig. 3.5) muestra la típica reacción redox de las tiorredoxinas y tiene actividad negativa a la formación de *clusters* Fe/S (Pórfido y Bellanda, resultados no publicados).

3.3.2 Relación estructura-función de EgIsTRP y tiorredoxinas

El grupo del Dr. Bellanda en la Universidad de Padova (Italia) obtuvo la estructura de la EgIsTRP por NMR (datos sin publicar). Esta misma muestra una gran similitud estructural al resto de las tiorredoxinas canónicas (Fig. 3.7) evidenciada por su bajo RMSD ~2 Å.

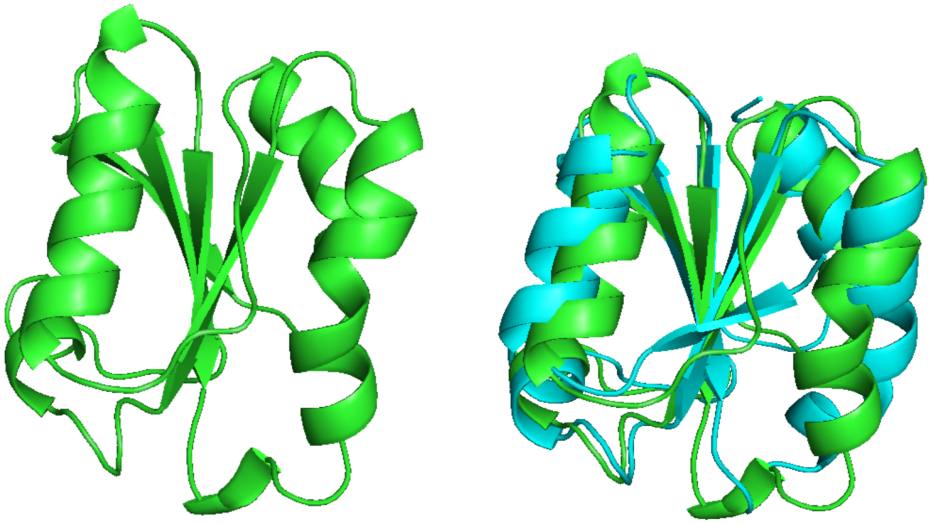


Figura 3.7 Estructura de la EgIsTRP mostrando el típico plegamiento de las familias de las tiorredoxinas (panel izquierdo). En el panel derecho mostramos la estructura de la EgIsTRP (en cian) alineada estructuralmente con la tiorredoxina humana (PDB: 1ERU, en verde). El RMSD resultante es de 2.2 Å.

Como mencionamos anteriormente, la superfamilia estructural de las tiorredoxinas es sumamente numerosa y con gran divergencia funcional (Atkinson and Babbitt, 2009). A pesar de la alta conservación del plegamiento en toda la superfamilia, variaciones secuenciales definen la gran divergencia funcional de la superfamilia. Así, distintas variaciones secuenciales tanto en el motivo canónico para las tiorredoxinas, CXXC, como en distintas partes de su plegamiento, generan una enorme variabilidad funcional. Básicamente la variabilidad funcional depende del potencial redox definido por las Cys del motivo secuencial típico de esta superfamilia. Se ha encontrado que diversos condicionamientos secuenciales y estructurales modulan dicho potencial (Mössner *et al.*, 2000). De las dos Cys correspondientes al motivo secuencial típico de esta superfamilia, la Cys del extremo amino terminal es la más afectada por diversas interacciones fisicoquímicas produciendo un *shift* en su pka de hasta ~2 unidades de pH (Jeng *et al.*, 1995), condicionando de esta forma sus propiedades biológicas (Moutevelis and Warwicker, 2004). Las particularidades de esta Cys en estos entornos fisicoquímicos aumentaría su capacidad nucleofílica para favorecer su capacidad redox como se ve en la (Fig. 3.8).

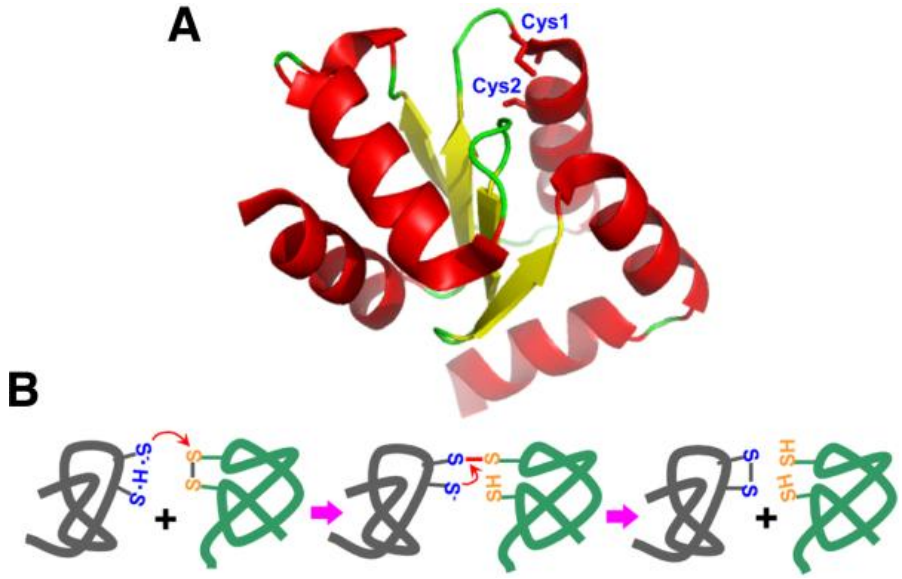


Figura 3.8 La figura muestra el funcionamiento de una tiorredoxina canónica alternando entre su estado reducido y oxidado por acción sobre otros tioles de otras proteínas. Figura extraída de (Fomenko *et al.*, 2008)

Sin embargo, como mencionamos anteriormente, la EgIsTRP no posee actividad oxidoreductasa, sino por el contrario forma *clusters* Fe/S. Para comprender las bases de dicho comportamiento, realizamos estudios secuenciales y estructurales utilizando la estructura de la EgIsTRP provista por el Dr. Bellanda.

La estructura de la EgIsTRP está representada por 20 confórmeros NMR, o sea los mejores modelos que representan el mejor ajuste a los datos de NMR. La máxima diferencia estructural entre dichos confórmeros es de 1.65 Å, medido utilizando el programa Mammoth (Ortiz *et al.*, 2002) a partir de una comparación de a pares entre todos los confórmeros de la EgIsTRP (Fig. 3.9)

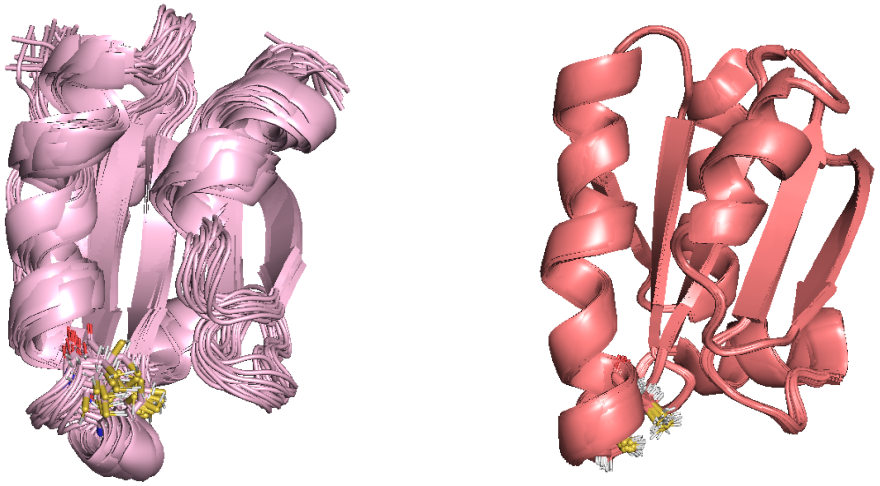


Figura 3.9 A la izquierda se muestra los 20 confórmeros correspondientes a la EgIsTRP destacando la presencia de las Cys del sitio activo. Para su comparación, a la derecha se muestran 40 confórmeros correspondientes a la tiorredoxina humana en su estado reducido (PDB 1TRV).

Por el contrario, la estructura de la tiorredoxina humana parece más rígida, ya que la comparación entre sus confórmeros en la estructura NMR da un RMSD máximo de 0.35 Å. Otra particularidad de la EgIsTRP que resalta a la vista en la figura anterior, es que las Cys parecen estar más expuestas al solvente comparadas con la tiorredoxina humana. Esto posiblemente se deba a la presencia de una Pro y una Gly, aminoácidos fuertemente conservados en toda las tiorredoxinas con actividad oxidoreductasa. La Pro induce un cambio de dirección de la hélice alfa amino terminal ocultando del solvente las Cys del sitios activo. La Pro y la Gly están ausentes en las secuencias de la EgIsTRP y homólogas cercanas.

Al analizar más en detalle la dinámica de la EgIsTRP representada por la colección de los 20 confórmeros del NMR, encontramos que la proteína alterna entre dos conformaciones, cerrada y abierta (Fig. 3.10)

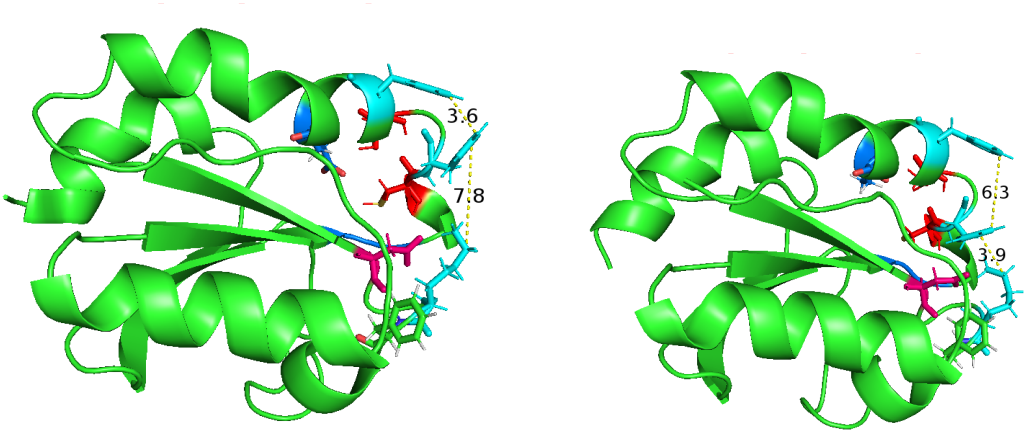


Figura 3.10 conformaciones alternativas de la EgIsTRP. Estas conformaciones se pueden describir en función de 3 residuos: Phe 35, Phe 39 y Lys 63 (en cian). En la figura de la izquierda se aprecia que la Phe 35 establece una interacción pi-pi con la Phe 39 (distancia mínima registrada 3.6 Å), mientras que en la otra conformación (imagen de la derecha) la Phe 35 establece una interacción pi-catión con la Lys 63. En rojo se destacan las Cys del sitio activo.

Estas conformaciones a su vez condicionan las interacciones de distintos aminoácidos sobre las Cys del sitio activo, mayormente sobre la Cys aminoterminal (Cys 34 en la EgIsTRP). En la figura 3.11 vemos que distintos aminoácidos (Thr 30, Thr 31, Glu 41, Asp 60 y Lys 63) se encuentran entre los que interaccionan alternativamente entre las conformaciones descritas arriba

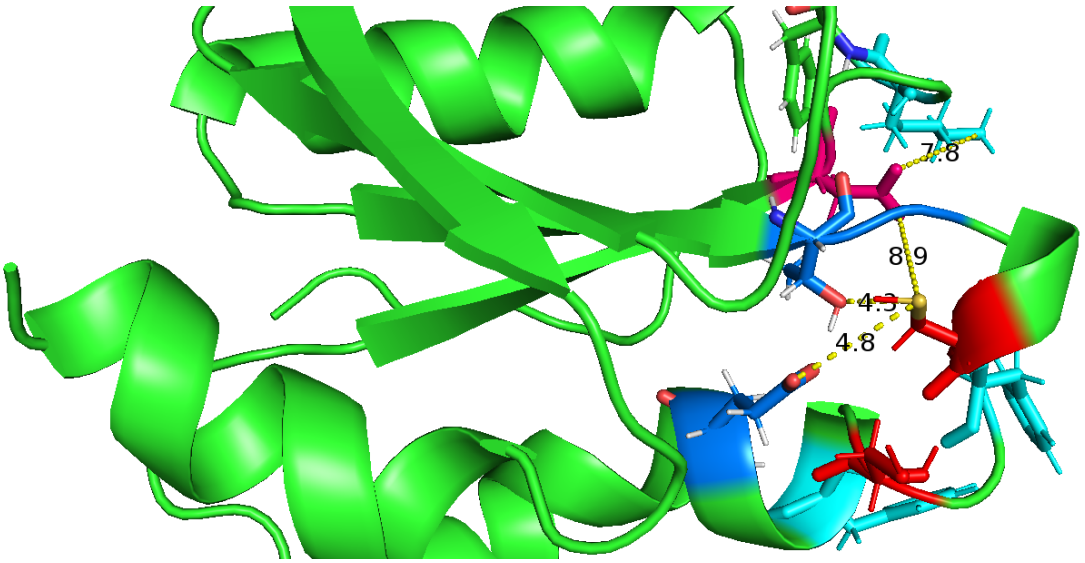


Figura 3.11 En rojo se observa la Cys 34 interaccionando con el Glu 41 (distancia 4.8 Å) con la Thr 30 (distancia 4.3 Å) y con el Asp 60 (distancia 8.9 Å). Estas distancias fluctúan haciéndose más chicas o grandes en las distintas conformaciones.

Para evaluar en mayor profundidad el efecto de estos cambios conformacionales sobre las propiedades fisicoquímicas de las Cys del sitios activo (Cys 34 y 37) estimamos el pka en cada una de las conformaciones utilizando el programa Propka (Unni *et al.*, 2011; Dolinsky *et al.*, 2004). En la tabla 3.1 vemos que el pka de la Cys 34 varía apreciablemente en función de los cambios conformacionales (pka min = 7.83, pka max= 10.96) mientras que la Cys 37 casi no varía de manera apreciable su valor.

Estos cambios en el pka están correlacionados en mayor medida con la variación de la distancias relativas de los residuos mencionados más arriba, encontrándose que la mayor correlación (~0.45) es con la Thr 31 y con el Glu 41. Es interesante notar que la Lys 63, muy alejada de la Cys 34, ejerce su efecto a través del apantallamiento relativo del Asp 60 que si se encuentra cercano a la Cys 34.

Creemos que estos cambios de pka, que muestran un cambio (*shift*) hacia pka básicos (promedio en las conformaciones ~9.2 y máximo 10.96) podrían explicar la ausencia de capacidad redox en la EgIsTRP, como fue anteriormente propuesto para proteínas similares (Su *et al.*, 2007).

Tabla 3.1 pka de la Cys 34 y 37 de la EgIsTRP en función de los distintos confórmeros del ensemble del NMR.

Modelo	pk Cys 34	pk Cys 37
1	10.64	9.34
2	9.48	9.87
3	10.38	9.82
4	10.96	9.75
5	8.26	9.02
6	8.36	9.19
7	7.83	9.52
8	8.27	9.35
9	10.03	9.68
10	8.95	9.71
11	9.87	9.34
12	8.89	9.87
13	9.09	9.68
14	10.75	9.4
15	8.09	9.38
16	8.23	9.02
17	8.18	9.78
18	10.46	9.35
19	8.32	9.75
20	8.83	9.4

Figura 3.12 Alineamiento secuencial de EgIsTRP (EgIsTRP y EcISTRP) junto a homólogos cercanos y tiorredoxinas canónicas.

En la figura 3.12 hemos alineado algunos representantes cercanos de EgIsTRP (secuencias con nombres en rojo, distintas especies de *Echinococcus*). A su vez el alineamiento contiene géneros cercanos como *Hymenolepis* (secuencias con nombres verdes) y *Taenia* (secuencias con nombres naranjas). Además, hemos incluido para su comparación secuencias de tiorredoxinas canónicas (secuencias con nombres en cian). En las secuencias podemos observar en gris en sitio catalítico con la perfecta conservación de sus Cys mientras que en rosado se muestran los residuos mencionados arriba y que jugarían un rol esencial en la biología de la EgIsTRP.

3.3.3 Estudios de reconstrucción ancestral

En la base de datos Revenant (capítulo 4) se encuentran 7 tiorredoxinas ancestrales resucitadas y cristalizadas (RV9 a RV15). Estas tiorredoxinas ancestrales representan determinados ancestros comunes a determinados organismos actuales (por ejemplo, el ancestro común a todos los Eucariotas, el ancestro común a todos los Procariotas, etc) y han sido reconstruidas y resucitadas por métodos de máxima verosimilitud con el objeto de estudiar su actividad redox y su estabilidad térmica (Perez-Jimenez *et al.*, 2011). En la figura 3.13 se muestra una imagen original de dicho *paper* en la que se observa la ubicación

relativa de las distintas entradas Revenant además de la información sobre a qué ancestro representa.

Para establecer el origen evolutivo de la EgIsTRP y de las proteínas homólogas cercanas, incluimos estas secuencias en el set de proteínas del árbol de la figura 3.14. Utilizando el programa Phyml con el modelo evolutivo Jones y una distribución Gamma con 4 categorías de velocidades, estimamos el árbol de Máxima Verosimilitud. La filogenia resultante muestra que este conjunto de proteínas similares que contiene a EgIsTRP posee un ancestro común con procariotas. Como las distancias evolutivas del grupo de secuencias cercanas a EgIsTRP son muy lejanas con secuencias de procariotas y de distancias comparables a otros grupos de organismos (por ejemplo metazoos y hongos), decidimos comenzar nuestro análisis de reconstrucción utilizando el último ancestro común de los organismos cercanos a EgIsTRP (ver árbol de la figura 3.14).

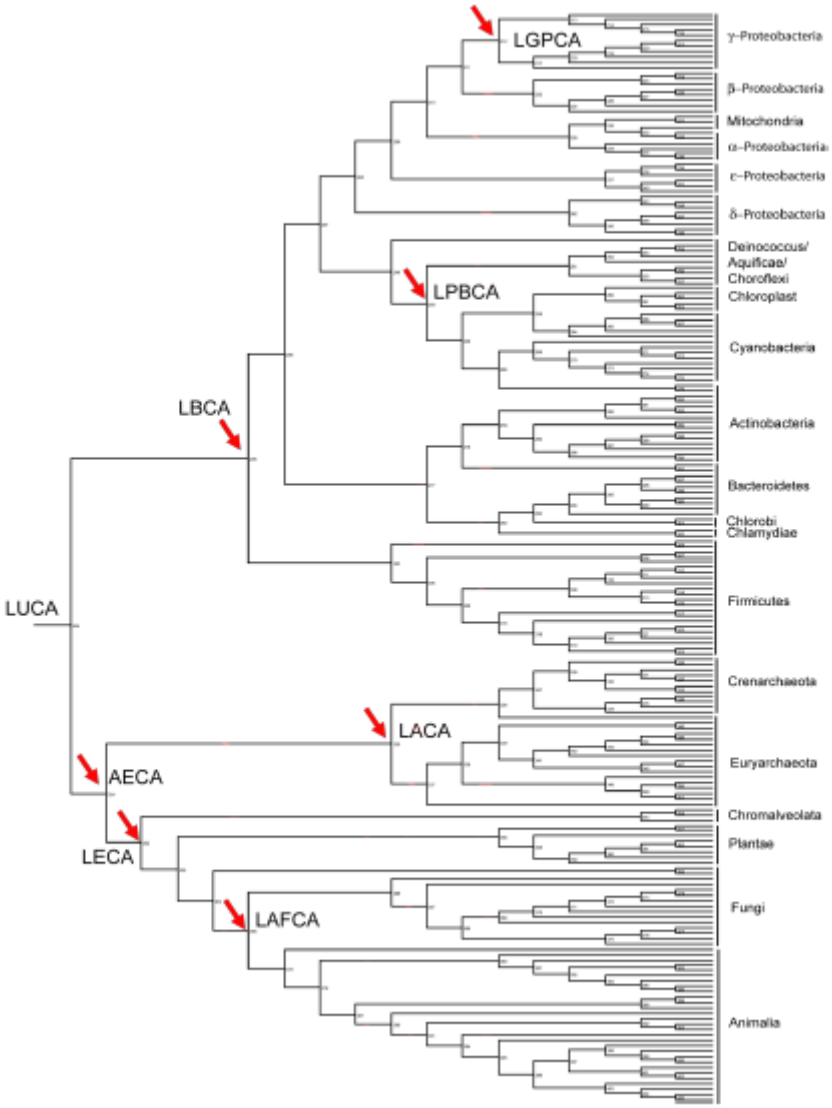


Figura 3.13 Figura tomada del artículo "Single-molecule paleoenzymology probes the chemistry of resurrected enzymes" (Perez-Jimenez *et al.*, 2011) para mostrar la ubicación relativa de los ancestros reconstruidos (indicados por flechas rojas). LGPCA= *last common ancestor of gamma proteobacterias,* LPBCA= *last common ancestor of cyanobacteria and deinococcus thermus (origin of photosynthetic bacteria),* LBCA= *last common ancestor of bacteria,* LUCA= *last universal common ancestor,* LACA= *last archaeal common ancestor,* AECA= *archaea/eukarya common ancestor,* LECA= *last eukaryotic common ancestor,* LAFCA=*last common ancestor of animals and fungi.*

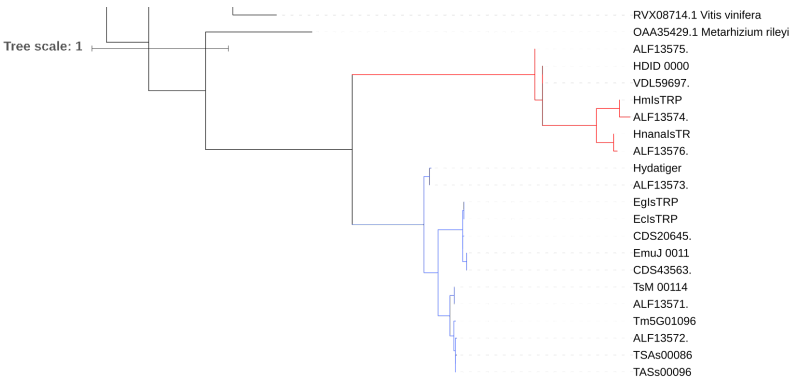


Figura 3.14 Árbol filogenético de las secuencias homólogas cercanas a EgIsTRP. En el cluster azul se encuentra la EgIsTRP (códigos: EgIsTRP y EcIsTRP) junto con secuencias del género *Taenia*, mientras que el cluster rojo contiene secuencias del género *Hymenolepis*.

Antes de pasar a la etapa de reconstrucción decidimos formular nuestra hipótesis para explicar el comportamiento diferencial entre las proteínas del *cluster* rojo y azul de la última figura. Recordemos que la EgIsTRP no presenta actividad oxidoreductasa y sí participa de la formación de *clusters* Fe/S (Bisio *et al.*, 2016). Por el contrario, las proteínas del *cluster* rojo, mayormente del orden Hymenolepis muestran actividad oxidoreductasa y no participan de la formación de *clusters* Fe/S (Bellanda y colaboradores, resultados sin publicar). Sabemos que las propiedades fisicoquimicas del entorno de las Cys (especialmente la del amino terminal) condicionan el pka de la Cys en principio llevándolo a valores muy básicos con lo cual no se podría llevar a cabo la reacción redox. Vimos en la sección anterior que la EgIsTRP es mucho más flexible que la tiorredoxina canónica humana, y dicha diversidad conformacional en EgIsTRP conducía a la extrema variación del pka de la Cys (Cys 34). De esta forma creemos que el aumento de la flexibilidad podría ser una adaptación funcional de la EgIsTRP para que funcione como formadora de *clusters* de Fe/S y no como oxidoreductasa.

Para profundizar en esta hipótesis, y debido a que no hay en la actualidad ninguna proteína cristalizada del *cluster* rojo (*Hymenolepis*) decidimos utilizar un método de predicción de flexibilidad con base secuencial denominado Dynamine (Kosciolek *et al.*,

2017; Cilia *et al.*, 2014, 2013). Dynamine es un método que predice la flexibilidad a nivel de cada aminoácido en la secuencia de una proteína. Consiste en un modelo de regresión lineal entrenado con un set de proteínas con estructura estimada por NMR el cual dispone de parámetros de la dinámica a nivel del esqueleto carbonado en posición específicas. El modelo así entrenado predice la flexibilidad de cada posición en una proteína asignándole una probabilidad, cuanto mayor, más rígido será la posición.

Según Dynamine, el coeficiente de correlación entre la flexibilidad predicha y la movilidad estimada por S^2_{RCI} a partir de corrimientos químicos estimados por NMR (Cilia *et al.*, 2014) en promedio es ~0.6 con una dispersión bastante amplia (Fig. 3.15). En nuestro caso utilizando la base de datos CoDNaS (Monzon *et al.*, 2016, 2013) para estimar la diversidad conformacional de diversas tiorredoxinas a partir de datos experimentales (estructuras cristalográficas y NMR), encontramos que el coeficiente de correlación entre el RMSD promedio que caracteriza la diversidad conformacional de cada proteína y un *score* que mide la rigidez de la proteínas (simplemente es el promedio de sitios rígidos que tienen una probabilidad > 0.8) da ~0.33. Por ejemplo, el *score* para la EgIsTRP es 0.54 mientras que la tiorredoxina humana es 0.67, reflejando el hallazgo de la reducida diversidad conformacional de la tiorredoxina humana (RMSDmax = 0.35Å) con respecto a la EgIsTRP (RMSDmax = 1.65Å).

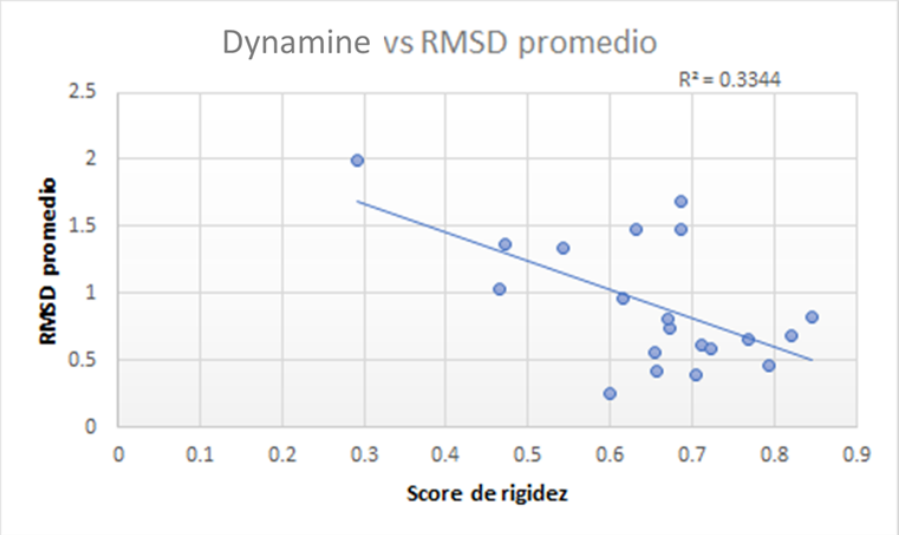


Figura 3.15 Correlación entre tiorredoxinas con diversidad conformacional conocida (RMSD promedio entre los confórmeros) y *score* de rigidez estimado usando el programa Dynamine

A pesar de esta relativamente baja correlación, el Dynamine brinda *scores* de flexibilidad perfectamente diferenciables cuando se acumula la información sobre la extensión de la diversidad conformacional expresada como RMSD. En nuestra experiencia, RMSD < 0.8 Å, representan a proteínas rígidas a nivel de su movilidad en sus cadenas carbonadas (Marino-Buslje et al., 2019) sobre todo teniendo en cuenta que el error experimental en cristalografía es aproximadamente ~0.5Å (Burra *et al.*, 2009). Considerando este *cutoff* encontramos que los *scores* para proteínas con un RMSD promedio menor a 0.8Å es de 0.72 ± 0.074 mientras que las proteínas con RMSD promedio mayor a 0.8Å es igual a 0.55 ± 0.12.

A la luz de la capacidad de Dynamine de diferenciar entre proteínas flexibles y rígidas, decidimos utilizar Dynamine para explorar la flexibilidad de las tiorredoxinas canónicas y las homólogas cercanas a EgIsTRP. Cuando calculamos nuestro *score* de rigidez para las proteínas de los clusters mostrados en la figura 3.14, vemos que el *cluster* azul que contiene a EgIsTRP y homólogas cercanas tiene *scores* de rigidez bajos como muestra la figura 3.16. La secuencia de EgIsTRP (códigos EglsTRP y EclsTRP correspondiente a distintas secuenciaciones del mismo genoma) muestran los menores *scores* del grupo (0.54).

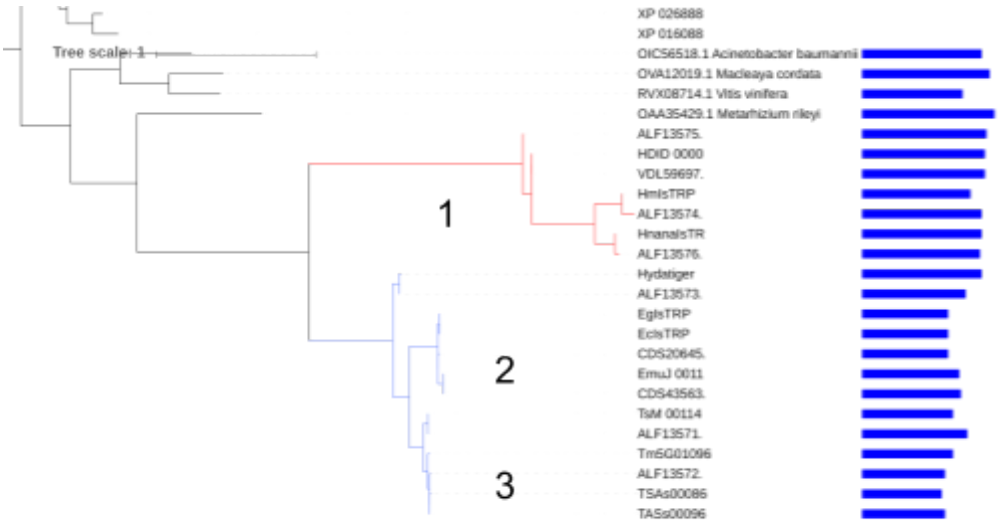


Figura 3.16 Árbol filogenético conteniendo los homólogos cercanos a EgIsTRP mostrando en forma de barras los scores de rigidez (barras azules). Este *score* es el promedio de posiciones rígidas por proteína considerando como corte una probabilidad > 0.8. Cuanto menor es el *score* mayor es la flexibilidad de la proteína. Los números (1, 2 y 3) indican los nodos ancestrales comunes a 3 grupos principales de organismos: 1 corresponde al *cluster* de *Hymenolepis*, 2 al de *Echinococcus* y el 3 al de *Taenia*.

La fuerte correlación entre flexibilidad y clustering filogenético nos motivó a explorar con Dynamine los estados ancestrales de la EgIsTRP. Para esto reconstruimos las proteínas ancestrales de todos los nodos internos de la filogenia mostrada en la figura 3.13. Para realizar esto procedimos a curar a mano el alineamiento eliminando inserciones o deleciones especie-específicos generados por las secuencias *outgroup* y asegurando que los dominios estén correctamente alineados. Paso seguido utilizamos el paquete de software PAML 4.9 (Yang, 1997) para obtener las secuencias de los nodos, utilizando el árbol que obtuvimos anteriormente, el alineamiento curado y configurando los parámetros con el modelo evolutivo Jones (Jones *et al.*, 1992) y una distribución Gamma con 8 categorías de velocidades. Para cada posición de las secuencias reconstruidas se tomó el aminoácido con mayor probabilidad de ocurrencia en ese nodo. A cada una de las secuencias ancestrales reconstruidas por este método le determinamos la rigidez utilizando nuevamente el programa Dynamine. En la figura 3.17 se puede ver la evolución de la rigidez desde el nodo ancestral común a los clusters rojo y azul de la figura 3.14 hasta las

secuencias de los clusters que contienen 3 grupos principales de organismos: *Echinococcus* (EgIsTRP), *Hymenolepis* y *Taenia*.

El árbol filogenético (Fig. 3.13) realizado para establecer el origen evolutivo de EgIsTRP indica que posee un ancestro común con procariotas. Al realizar el mismo estudio de rigidez utilizando el Dynamine en las entradas Revenant para tiorredoxinas, encontramos que el último ancestro común a procariotas da un *score* de rigidez igual a 0.69 (Fig. 3.17), lo que sugeriría que de ser cierto el origen procariótico de EgIsTRP, existió una adaptación funcional hacia un aumento de la flexibilidad en los homólogos cercanos a EgIsTRP (*Echinococcus* y *Taenia*). El promedio del *score* de rigidez en Bacteria, para las tiorredoxinas estudiadas en la estimación filogenética, da 0.67 mostrando que la mayoría de las tiorredoxinas bacterianas posee una rigidez predicha mayor que el cluster de EgIsTRP (0.54).

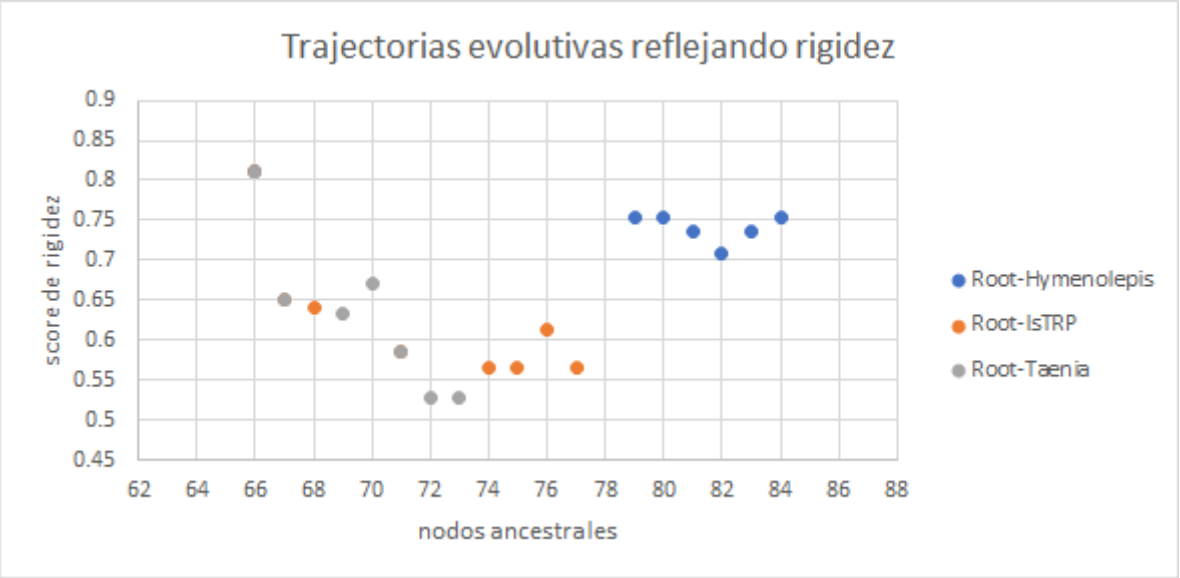


Figura 3.17 Evolución de la rigidez predicha por Dynamine sobre las secuencias reconstruidas según la trayectoria seguida por los nodos ancestrales hasta los organismos actuales. El nodo 66 es el nodo ancestral común a los 3 grupos de organismos destacados en la figura 3.16

La mayoría de los organismos muestran cierta rigidez en sus tiorredoxinas (Bacteria=0.67, Archaea=0.72, Metazoa=0.67, Fungi=0.70) sacando a las plantas que poseen un *score* = 0.51. Al analizar estos datos junto a las entradas Revenant (estados

ancestrales) encontramos que AECA=0.76, LACA=0.80, LAFCA=0.71, LBCA=0.69, LECA=0.73, LGPCA=0.62, LPBCA=0.68 muestran que la rigidez de las tiorredoxinas es una característica ancestral y que una reducida rigidez ha sido probablemente una adaptación funcional en determinados organismos como las Plantas y en el caso de los homólogos cercanos a EgIsTRP.

3.3.4 Estudios de resucitación

Para investigar esto más en profundidad, redujimos el árbol filogenético de la figura 3.5, haciendo énfasis en la EgIsTRP y sus homólogas más cercana, obteniendo el árbol de la figura 3.18.

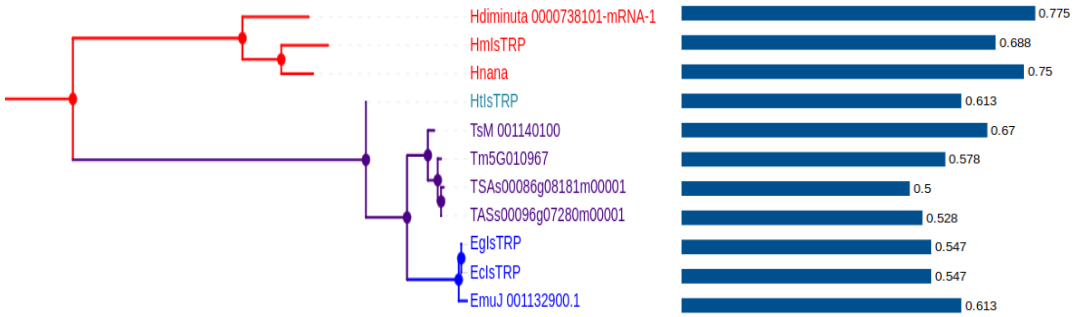


Figura 3.18. Árbol filogenético de las tiorredoxinas, reducido a EgIsTRP y sus homólogas cercanas. En rojo, la familia *Hymenolepidiae*, en violeta la familia *Taeniidae* y en azul el género *Echinococcus*. En cian se encuentra marcada la proteína de *H. taeniaeformis*, que es la más cercana a la familia *Hymenolepidiae*.

Al analizar el árbol para dilucidar qué nodo debíamos resucitar, nos dimos cuenta de que nos faltaba algo de información, ya que conocíamos la actividad de la EgIsTRP y del *cluster* rojo teníamos la información de HmIsTRP, la cual muestra la típica actividad redox de las tiorredoxinas y no es capaz de coordinar clusters Fe/S (Pórfido et al, resultados no publicados), sin embargo, no teníamos información de la familia *Taeniidae*. Basándonos en el *score* de la proteína HtIsTRP (el cual es bajo) podíamos suponer un comportamiento similar a EgIsTRP. Conocer el comportamiento de HtIsTRP nos permitiría decidir qué nodo ancestral reconstruir y resucitar. Obtener información de la proteína correspondiente al nodo

ancestral nos permitiría comprender si EgIsTRP y el resto de los integrantes de la familia *Taeniidae* evolucionaron aumentando su flexibilidad y permitiendo una adaptación funcional. O por el contrario, si el ancestro en común de EgIsTRP y los *Taeniidae* era ya una proteína flexible y con diversidad conformacional, y HmIsTRP y sus homólogas cercanas evolucionaron a conformaciones más rígidas y funciones específicas. Con este fin, expresamos y purificamos HtIsTRP, para luego realizar su caracterización estructural y funcional. En acuerdo con lo observado para EgIsTRP, en un ensayo de determinación de actividad redox, HtIsTRP no presentó esta actividad (Fig. 3.19) (datos sin publicar de nuestro grupo).

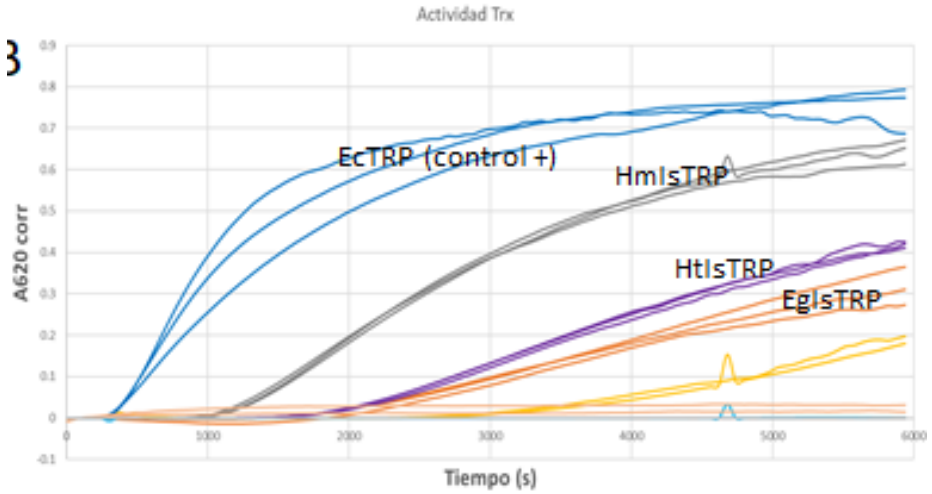


Figura 3.19. Medida de actividad redox de diferentes proteínas. La proteína HtIsTRP no presenta actividad.

Considerando la información in silico y experimental recopilada al momento, decidimos resucitar el nodo que une a familia *Hymenolepididae* y la familia *Taeniidae* (el que aparece como raíz en la figura 3.18), con la finalidad de obtener información sobre la trayectoria evolutiva y las características estructurales que determinaron el cambio de flexibilidad y la adaptación funcional en estas tiorredoxinas atípicas.

Como se dijo anteriormente, para reconstruir las secuencias de los nodos internos se utilizó el programa PAML. Como resultado obtenemos las probabilidades posteriores (PP) de cada sitio y ocurrió que en dos sitios importantes había cierta incertidumbre, por lo

que se decidió tratar de sintetizar todas las combinaciones. Estas posiciones fueron la 27 y 57 del alineamiento, donde las posibilidades eran M y L para la primera posición (con PPs de 0.58 y 0.25 respectivamente) y K e I para la segunda posición (con PPs de 0.368 y 0.365, respectivamente).

Con esta información se sintetizó el gen de la tiorredoxina ancestral (ATRX) en el vector de expresión pET28a. Asimismo, se diseñaron primers que mediante técnicas de clonado permitieron obtener variantes de ATRX, tales como ATRX K57I y ATRX M27L.

La proteína ATRX en el vector de expresión pET28a para *E. coli* se sobreexpresa pero se obtiene en cuerpos de inclusión. Se evaluaron diversas condiciones de crecimiento y expresión de modo de mejorar la cantidad de proteína en la fracción soluble de *E. coli* sin obtener mejoras. Se procedió a purificar ATRX desde los cuerpos de inclusión disueltos usando agentes desnaturalizantes en altas concentraciones. La proteína se intentó renaturalizar mediante diálisis utilizando distintas soluciones así como disminuyendo progresivamente la cantidad de agente desnaturalizante. Lamentablemente, el resultado siempre fue la proteína precipitada. Por otro lado, se intentó aislar la pequeña proporción de ATRX que se encuentra soluble. Para llevar a cabo este objetivo se utilizaron soluciones de lisis que mejoran la solubilidad proteica, mediante el agregado de Arg/Glu, glicerol, alta fuerza iónica. A pesar de obtener una pequeña cantidad de proteína soluble luego de la lisis celular, la misma se pierde en los pasos sucesivos de purificación (cromatografía de afinidad con Ni y cromatografía de exclusión molecular). Asimismo se practicaron distintas técnicas cromatográficas con el objetivo de mejorar el rendimiento, sin obtener resultados positivos.

Se subclonó la secuencia en otros dos vectores de expresión en E.coli que utilizan proteínas de fusión, pGEX (proteína de fusión a GST) y pETSUMO (proteína de fusión a SUMO). En el caso de la proteína fusionada a GST, la proteína mayoritariamente se encuentra en cuerpos de inclusión. En el caso de la proteína fusionada a SUMO, se evidenció un aumento significativo de la proteína de fusión en la fracción soluble. Se procedió a la purificación de esta proteína de fusión mediante una cromatografía de afinidad

con Ni y una posterior cromatografía de exclusión molecular (CEC). A pesar de obtener la proteína de fusión en la fracción soluble, en el gel de poliacrilamida realizado con las fracciones de la CEC, se observa un claro proceso de degradación de la proteína. Esto mismo fue confirmado por espectrometría de masa, donde se pudo observar la especie esperada y otras especies de menor peso molecular. Por último, las variantes ATRX K57I y ATRX M27L fueron clonadas en pETSUMO, expresadas y purificadas. Nuevamente, se observa un patrón de degradación de la proteína de fusión obtenida. En el caso de ATRX K57I, se continuó la purificación hasta el clivaje de la proteína de fusión SUMO, y la obtención de ATRX K57I. El rendimiento de esta proteína fue muy bajo, probablemente debido a la degradación sufrida. A pesar de esto, se realizó un ensayo de actividad oxidorreductasa con la proteína ATRX K57I obtenida así como con la proteína de fusión SUMOATRX y sus variantes. En todos los casos no se observó actividad oxidorreductasa.

Teniendo en cuenta el trabajo realizado y los múltiples intentos por obtener la proteína ancestral utilizando las diversas estrategias de expresión y purificación, concluímos que dada la baja probabilidad posterior de la secuencia reconstruida, la secuencia obtenida es poco soluble e incapaz de adoptar un plegamiento, por ende vemos degradación proteica. Los resultados negativos respecto a la actividad oxidorreductasa de la proteína ancestral no son concluyentes ya que la proteína parece no estar plegada correctamente. Por otro lado, las dificultades abordadas en la obtención de la proteína y el bajo rendimiento de la misma hacen difícil proyectar otro tipo de estudios que implican mayor concentración proteica.

3.4 Discusión

En este capítulo hemos derivado importante información sobre la relación estructura-función de la EgIsTRP. Hemos encontrado que la flexibilidad diferencial de esta proteína podría ayudar a comprender su biología. En este sentido es importante destacar que sin la ayuda de los estados ancestrales la información recabada estaría incompleta. Hemos encontrado que las proteínas más cercanas a EgIsTRP (*Echinococcus* y *Taenia*) presentan un aumento de flexibilidad sobre otras proteínas también de helmintos pero de otros géneros (*Hymenolepis*) y que este aumento podrían explicar el comportamiento biológico diferencial entre esos géneros. El haber predicho la flexibilidad en los estados ancestrales desde el ancestro común a estos géneros nos permitió elaborar la hipótesis que sobre la EgIsTRP actuó una adaptación funcional sobre sus ancestros promoviendo el aumento de flexibilidad. Además, al utilizar los estados ancestrales resucitados de diversas tiorredoxinas nos permitió elaborar la hipótesis más general, de que originalmente las tiorredoxinas tendieron a ser proteínas rígidas y que sólo en algunos casos (Plantas y proteínas relacionadas a EgIsTRP) se ve un aumento de la flexibilidad.

A pesar de estos resultados negativos, hemos obtenido experiencia válida y criterio para el futuro abordaje de reconstrucciones ancestrales. Claramente, el caudal de información sobre las proteínas actuales es fundamental para la obtención de una predicción confiable y el futuro éxito de la resucitación de la proteína ancestral.

3.5 Métodos específicos del capítulo

En lo respectivo a los métodos experimentales, como se comenta más arriba, estos estuvieron a cargo del Dr. Massimo Bellanda, en la Universidad de Padova.

El alineamiento que se utilizó para realizar esta investigación se infirió a partir de secuencias cercanas a la tiorredoxina atípica provistas por el Dr. Bellanda y se le agregaron secuencias de tiorredoxinas canónicas obtenidas mediante el uso de BLAST. En total 60 secuencias fueron usadas para inferir la filogenia usando el software IQ-TREE, utilizando el modelo de evolución LG+F con una distribución Gamma de 4 velocidades.

Este mismo alineamiento, junto con la filogenia inferida fueron utilizados para llevar a cabo la reconstrucción de secuencias ancestrales usando los mismos parámetros (modelo de evolución LG+F con una distribución Gamma de 4 velocidades) mediante el software PAML, obteniendo de esta forma la secuencia más probable para cada nodo.

4 Revenant: una base de datos de proteínas resucitadas

4.1 Resumen

Como ya se ha comentado, la evolución biológica es un proceso complejo fuertemente dependiente de su historia. La diversidad biológica observada en la actualidad fue producto de muchos cambios y condicionamientos ocurridos a lo largo de la historia. El espacio secuencial solo está explorado en parte y fue condicionado por las secuencias exploradas en el inicio de la vida. Esto fue un proceso probablemente único e irrepetible, por lo que se han desarrollado técnicas bioinformáticas para poder estimar cómo fue este proceso. Gracias a las técnicas de reconstrucción ancestral y a las técnicas de biología molecular que nos permiten resucitar proteínas de organismos extintos, tenemos la posibilidad de estudiar esta historia, conocer el pasado y entender cómo se llegó a la diversidad biológica de la actualidad.

En el presente capítulo presentamos el desarrollo de una base de datos de proteínas resucitadas. Revenant (Carletti *et al*.,, 2020, contiene en la actualidad 211 proteínas resucitadas de las cuales 55 tienen estructura cristalográfica. Además, siempre que estén disponibles, Revenant contiene los árboles filogenéticos, secuencias, estructuras y datos biofísicos (estabilidad relativa, constantes cinéticas y de equilibrio) de las proteínas resucitadas. En Revenant la proteína más antigua resucitada se corresponde con una antigüedad de unos ~4200 millones de años y la más moderna a unos ~1.7 millones de años.

4.2 Introducción

Revenant DB es una base de datos de proteínas resucitadas. La misma nos permite acceder a las secuencias obtenidas por la técnica de ASR y mediante técnicas de resucitación a diversos parámetros y propiedades de dichas secuencias (estructuras cristalográficas, parámetros que caracterizan la cinética enzimática, datos de estabilidad térmica, etc). Es una base de datos curada manualmente, derivada de la información contenida en numerosas publicaciones científicas. Como mencionamos anteriormente, hoy en día existen muchos estudios de reconstrucción y resucitación que intentan contestar diversas preguntas de índole biológico, evolutivo, bioquímico, biofísico e incluso biotecnológico (por ejemplo (Bridgham *et al.*, 2009; Hobbs *et al.*, 2012; Perez-Jimenez *et al.*, 2011; Akanuma *et al.*, 2013; Hart *et al.*, 2014; Steindel *et al.*, 2016; Blanchet *et al.*, 2017; Risso *et al.*, 2017)).

Las secuencias depositadas en Revenant DB representan el estado más probable de un estado ancestral, siguiendo el camino evolutivo (o trayectoria) desde diferentes ancestros comunes ubicados en la raíz y el resto de los nodos internos de una variedad de árboles filogenéticos que representan la historia evolutiva de diferentes familia de proteínas. Cada proteína ancestral con su secuencia única tiene asociado un conjunto de datos que no sólo reflejan las propiedades de la proteínas resucitada, sino que además contiene información sobre la metodología ASR utilizada, la edad estimada de su nodo interno al cual corresponden, métodos de estimación filogenética, secuencias y organismos utilizados en la reconstrucción, etc. Además, Revenant se encuentra vinculada con otras bases de datos (GO (Ashburner *et al.*, 2000), PDB (Rose *et al.*, 2011), Uniprot (UniProt Consortium, 2019), CodNas (Monzon *et al.*, 2016)) que permiten relacionar a cada proteína ancestral resucitada con información biológica relevante.

4.3 Resultados

4.3.1 Implementación del servidor web

Revenant se encuentra disponible en línea desde la URL: http://revenant.inf.pucp.edu.pe/, y es de acceso libre y gratuito para su uso académico. El servidor web de Revenant fué realizado en colaboración con el grupo de Inteligencia Artificial del Departamento de Ingeniería, Pontificia Universidad Católica del Perú dirigido por la Dra. Layla Hirsh Martínez. El mismo fue diseñado con el objetivo de proveer una interfaz gráfica para que el usuario pueda explorar las proteínas resucitadas.

El servidor web posee una arquitectura de microservicios basadas en el *framework* Spring Boot. Este *framework* facilita la creación de aplicaciones web JAVA mediante la generación de código automático para tareas estándares, como por ejemplo el acceso a bases de datos. Los datos en el servidor se intercambian usando el formato JavaScript Object Notation (JSON) mediante un servicio RESTful que utiliza directamente HTTP para obtener datos o indicar la ejecución de operaciones sobre los datos.

4.3.2 Campos de la base de datos y contenidos de cada campo

Cada una de las proteínas resucitadas de un nodo interno (ancestral) particular con una única secuencia corresponde a una entrada de Revenant. Usando la información bibliográfica y el procedimiento del curado a mano de cada referencia, todas las entradas han sido anotadas con diferente datos relevantes: El nodo ancestral usado en la reconstrucción, su edad cronológica estimada, las metodologías de ASR usadas para inferir la secuencia ancestrales (modelo evolutivo y método computacional), las secuencias usadas para el MSA (cantidad de secuencias, bases de datos de las cuales fueron obtenidas y taxones a los que pertenecen esas secuencias) y la estimación del árbol filogenético (modelo evolutivo y método computacional), el programa usado para la estimación filogenética, el MSA y la ASR, disponibilidad de la estructura cristalizada (código

PDB asociado) y su caracterización de ligando. Además, varias entradas tienen asociados parámetros bioquímicos (i. e. Km, kcat) y biofísicos (i.e temperatura de desnaturalización, *ΔGunfolding*). Todas las proteínas de *Revenant* están asociadas con otras base de datos tales como PDB (Rose *et al.*, 2011), Pfam (Punta *et al.*, 2012), UniProt (UniProt Consortium, 2019), Gene Ontology (Ashburner *et al.*, 2000) y NCBI taxonomy (https://www.ncbi.nlm.nih.gov/taxonomy) que aportan información biológica relevante y, por otro lado, tienen asociado el código Pubmed (https://www.ncbi.nlm.nih.gov/pubmed/) correspondiente a la cita bibliográfica primaria donde fue resucitada la proteína.

4.3.3 Búsqueda de una proteína en Revenant

Para facilitar la búsqueda en la base de datos, el servidor web de Revenant ofrece diferentes métodos de búsqueda. El método más rápido consiste en buscar una proteína por medio de un código Revenant, de un código PDB o del nombre de la familia de proteína a la cual pertenece la proteína resucitada. Esta forma de búsqueda está disponible desde la página principal (*home*) (Fig. 4.1) y su utilización direcciona a una página de resultados en la que se que muestra una lista de entradas Revenant junto con una breve descripción de las mismas (descripción del nodo interno, edad estimada si está disponible y disponibilidad de estructura cristalográfica, por ejemplo ver (Fig. 4.2).

Figura 4.1 Página principal (*home*) de Revenant.

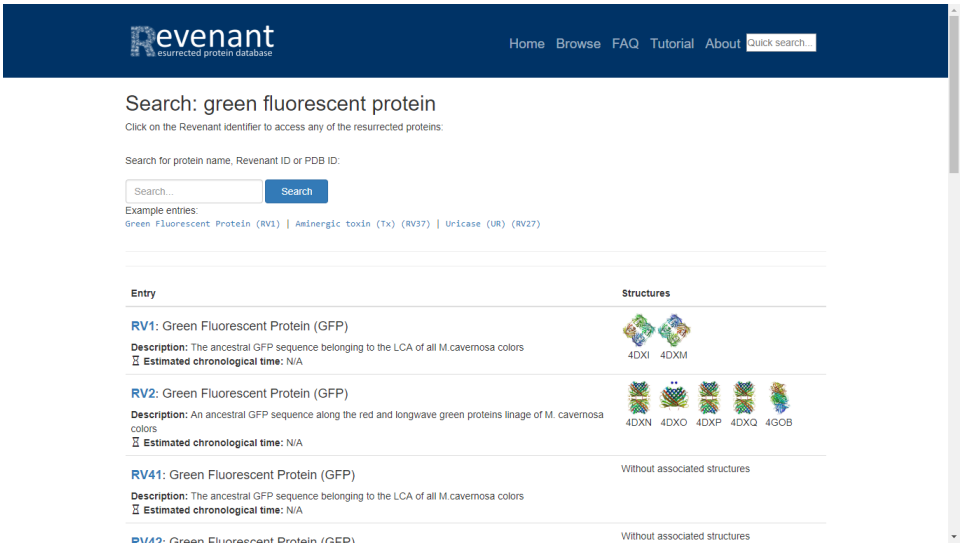


Figura 4.2 Página de resultados de Revenant. Esta página se abre solo cuando la búsqueda se realiza desde el campo de búsqueda del *home*.

El segundo método de búsqueda permite al usuario explorar todas las proteínas depositadas en Revenant por medio de la utilización de la página de "Búsqueda" (*Browse*, que está accesible desde las pestañas ubicadas en margen superior derecho del servidor). Para este último método, el usuario puede optar por dos formas de visualización: a) Una lista de todas las proteínas que muestra todas las entradas de Revenant de acuerdo con su ID (RVID) y también muestra una información mínima que describe a cada entrada (Fig. 4.3); b) Una línea del tiempo que muestra a todas las entradas de Revenant ordenadas por su edad cronológica estimada y además muestra eventos biológicos relevantes en la historia de vida del planeta Tierra (Fig. 4.3). Viendo esta línea del tiempo, por ejemplo, vemos que la entrada más antigua de Revenant data de 4200-3500 millones de años atrás (Mya por sus siglas en inglés) y la entrada más joven data de 8.8-6.3 Mya.

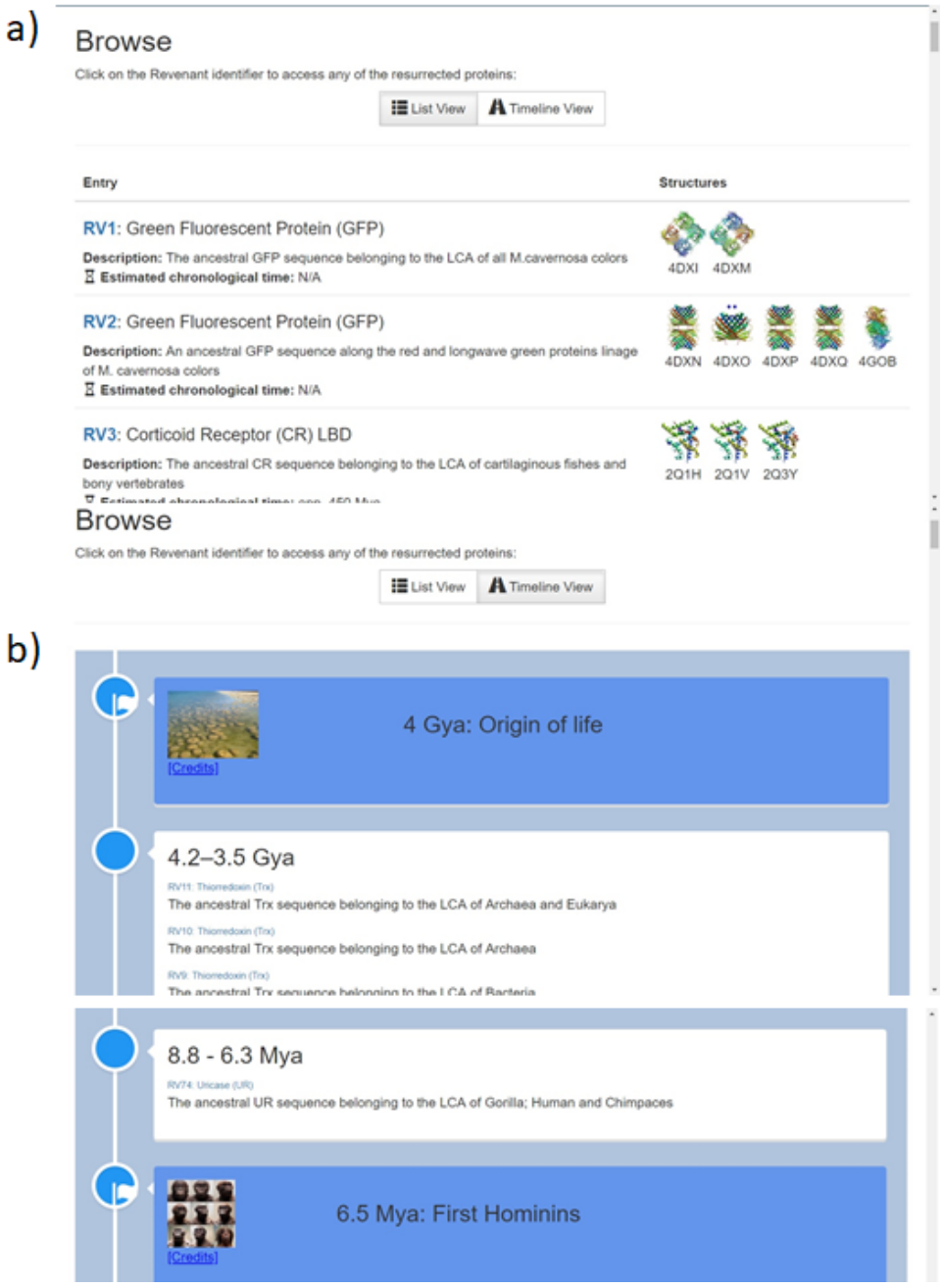


Figura 4.3 Página de búsqueda (browse). La base de datos se puede navegar por a) una lista de todas las entradas de Revenant que muestran su ID correspondiente, el nombre de la proteína, la descripción del nodo y la edad estimada. b) una línea de tiempo que muestra todas las entradas de Revenant ordenadas por su tiempo cronológico estimado (Gya, Mya y Kya) y los eventos biológicos relevantes en la historia de la vida en la Tierra caracterizados con imágenes (por ejemplo, 4 Gya: Origen de la vida, 6.5 Mya: Primeros Homínidos).

4.3.4 Página principal para una proteína en Revenant

Una vez realizada la búsqueda por alguno de los métodos mencionados anteriormente y habiendo seleccionado la entrada de interés (a criterio del investigador), el

servidor muestra toda la información que posee de la proteína resucitada dentro de la base de datos. En la (Fig. 4.4) se muestra el ejemplo de la entrada de la proteína ancestral Malato Deshidrogenasa. La página de resultados está compuesta por diferentes secciones (de arriba a abajo) que consisten en la información más general a la más particular sobre la proteína resucitada. La parte superior (Fig. 4.4.a-c) es una recopilación de información relevante sobre la secuencia ancestral obtenida por el enfoque ASR y se puede dividir en tres partes: la primera parte muestra una breve descripción de la resurrección ancestral donde se menciona el tiempo geológico estimado de la proteína; la siguiente parte “Reconstrucción de secuencia ancestral (ASR)” ofrece información sobre las herramientas y los modelos utilizados para la reconstrucción de secuencia, incluida la secuencia obtenida con su nombre correspondiente; finalmente, la sección "secuencia" muestra la secuencia obtenida por ASR e incluye una utilidad de búsqueda para patrones de secuencia que luego se resaltan en la secuencia correspondiente. En la parte inferior de cada entrada podemos encontrar información estructural (Fig. 4.4.d-e). Al hacer clic en cada una de las pestañas con las diferentes estructuras para una entrada específica de Revenant, la herramienta de visualización 3D mostrará la estructura. Luego, se muestra información adicional de la estructura, como el método de cristalización, la resolución, los términos GO, la secuencia correspondiente al campo SEQRES de la base de datos de PDB, la presencia y el tipo de los ligandos co-cristalizados. Debajo de la sección de la estructura, algunas de las proteínas Revenant tienen anotaciones sobre sus parámetros bioquímicos y / o biofísicos (Fig. 4.4.f-g). Estos parámetros aparecen en las secciones "Parámetros fisicoquímicos de proteínas" y "Parámetros de parámetros termodinámicos de proteínas". Algunos de estos, como kcat y Km, son específicos del sustrato y otros, como Tm o Topt, son específicos del pH, por lo que también están listados. En la parte inferior de la página, en cada entrada de Revenant, se muestra la cita primaria correspondiente de la proteína resucitada (Fig. 4.4.h).

a) RV24: Malate Dehidrogenase (MDH)

The ancestral MDH sequence belonging to the LCA of all modern apicomplexan
Estimated Chronological Time: 700-900 Mya

b) **ANCESTRAL SEQUENCE RECONSTRUCTION (ASR):**

Protein family sequences: MDH and LDH protein sequences
MSA number of sequences: 1844
Sequence database: UniPROTKB
MSA Software: MUSCLE

Phylogeny Sofware: PhyML 3.0
Phylogeny Method: Maximum Likelihood (ML)
Phylogeny Evolutionary model: LG substitution matrix and estimating the gamma parameter (12 categories) and empirical amino acid frequencies

ASR Software: Codeml module of PAML (version no informed)
ASR Method: Maximum Likelihood (ML)
ASR Substitution model: LG substitution matrix DNA-based and amino acid-based models

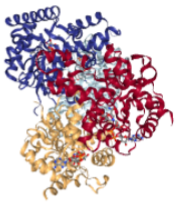
c) 318 Sequence name: AncMDH2

Search in sequence.. (Regex supported)

```
  1 MTQRKKISLI GAGNIGGTLA HLIAQKELGD VVLFDIVEGM PQGKALDISH SSPIMGSNVK ITGTNNYEDI KGSDVVIITA GIPRKPGMSR DDLLSVNAKI
101 MKDVAENIKK YCPNAFVIVV TNPLDVMVYV LHKYSGLPHN KVCGMAGVLD SSRFRYFLAE KLNVSPNDVQ AMVIGGHGDT MVPLTRYCTV GGIPLTEFIK
201 QGWITQEEID EIVERTRNAG GEIVNLLKTG SAYFAPAASA IEMAESYLKD KKRILPCSAY LEGQYGVKDL FVGVPVIIGK NGVEKIIELE LTEEEQEMFD
301 KSVESVRELV ETVKKLNA
```

d) STRUCTURES OF THE RESURRECTED PROTEINS: 5

4PLH | 4PLT | 4PLV | 4PLW | 4PLY



4PLV PDB

Crystal structure of ancestral apicomplexan malate dehydrogenase with lactate.

Exp. Method: X-RAY DIFFRACTION
Classification: OXIDOREDUCTASE
Resolution: 1.85
Taxonomy ID: 5794
Source: Apicomplexa

e) **Chains for 4PLV: 4**

A | B | C | D

Chain Length: 331
Molecular weight: 36459
Biological process: GO:0005975 GO:0019752 GO:0055114
Cellular component: N/A
Molecular Function: GO:0000166 GO:0003824 GO:0016491 GO:0016616

```
  1 MTQRKKISLI GAGNIGGTLA HLIAQKELGD VVLFDIVEGM PQGKALDISH SSPIMGSNVK ITGTNNYEDI KGSDVVIITA GIPRKPGKSD
 91 KEWSRDDLLS VNAKIMKDVA ENIKKYCPNA FVIVVTNPLD VMVYVLHKYS GLPHNKVCGM AGVLDSSRFR YFLAEKLNVS PNDVQAMVIG
181 GHGDTMVPLT RYCTVGGIPL TEFIKQGWIT QEEIDEIVER TRNAGGEIVN LLKTGSAYFA PAASAIEMAE SYLKDKKRIL PCSAYLEGQY
271 GVKDLFVGVP VIIGKNGVEK IIELELTEEE QEMFDKSVES VRELVETVKK LNALEHHHHH H
```

Ligands for chain A: 4

Chain	ID	Formula	Ligand name
A	2OP	C3 H6 O3	(2S)-2-HYDROXYPROPANOIC ACID
A	NAI	C21 H29 N7 O14 P2	1;4-DIHYDRONICOTINAMIDE ADENINE DINUCLEOTIDE

f) **PROTEIN PHYSIOCHEMICAL PARAMETERS:**

Substrate	Kcat (s-1)	Kcat/KM (s-1 mM-1)	KM (mM)
Oxaloacetate	7.5±1	3.3x10^4±6.6x10^4	0.4x10^-3±0.5x10^-3
Pyruvate	0.07±0.006	3.2x10^-3±0.4x10^-3	22.0±4.4

Kcat (s-1): Catalytic Constant - **Kcat/KM (s-1 mM-1):** Catalytic Efficiency - **KM (mM):** Michaelis-Menten Constant

g) **PROTEIN THERMODYNAMIC PARAMETERS:**

pH	Tm (°C)	Topt (°C)	Teq (°C)	ΔGunf (KJ/mol)

There are no entries for this protein yet!

Tm (°C): Melting Temperature - **Topt (°C):** Optimum Temperature - **Teq (°C):** Equilibrium Temperature - **ΔGunf (KJ/mol):** Gibbs' free energy of unfoldign

h) Primary citation

An atomic-resolution view of neofunctionalization in the evolution of apicomplexan lactate dehydrogenases
Author(s): Boucher; J.I.; Jacobowitz; J.R.; Beckett; B.C.; Classen; S.; Theobald; D.L.
Journal: Elife 2014. **Page(s):** e02304-e02304, **DOI**: 10.7554/eLife.02304
PubMed ID: 24966208 | **PubMed Central ID:** PMC4109310

Figura 4.4 Página principal de una entrada de Revenant. Se ejemplifica un RV ID de la familia malato deshidrogenasa. a) Descripción del nodo y su edad estimada b) Metodología utilizada para ASR: MSA, inferencia filogenética y enfoque ASR; c) Secuencia inferida por ASR: longitud de la secuencia, nombre de la secuencia y la secuencia misma en formato Fasta; d) Estructuras de cristal vinculadas con sus respectivos códigos PDB, visualización 3D; e) Información biológica obtenida de la base de datos de PDB: enlaces a términos GO, secuencia de estructura cristalina y ligandos; f) Caracterización bioquímica: parámetros cinéticos; g) Caracterización biofísica: los parámetros termodinámicos aún no existen datos para este ejemplo); h) Cita primaria donde la proteína resucitada se describió por primera vez relacionada con la base de datos PubMed y los códigos DOI.

4.4 Discusión

La base de datos Revenant es una base curada a mano de todas las proteínas ancestrales reconstruidas y resucitadas conocidas al momento. Su sitio web ofrece una interfaz amigable con varias alternativas de búsquedas y de exploración para recorrer su contenido. Creemos que la información contenida en Revenant puede ser utilizada para explorar distintos aspectos de las proteínas ancestrales con base experimental y ofrecer ejemplos biológicos para desarrollar o mejorar diversas herramientas bioinformáticas.

4.5 Métodos específicos del capítulo

4.5.1 Construcción de Revenant a partir fuentes bibliográficas primarias

Como mencionamos, Revenant es una base de datos curada a mano. Esto implica que cada entrada de Revenant tiene su validación experimental asociada a su correspondiente cita bibliográfica. De esta forma, para construir Revenant se identificaron publicaciones científicas relacionadas con la resurrección de proteínas ancestrales a través de la implementación de técnicas de "raspado web" (i.e *web-scraping)* y "minería de texto" (i.e *text-mining*) usando librerías estándares de Python (Fig. 4.5.a.). La primer técnica se utilizó básicamente para acceder de manera sucesiva a diferentes páginas web y recuperar información, para lo cual se armó librerías propias donde se utilizaron *urls* relacionados con páginas de revistas científica especializadas, páginas con artículos de divulgación científica y páginas de diferentes laboratorios especializados en evolución y biología molecular. De esta manera se identificaron un conjunto grande de posibles artículos relacionados con proteínas ancestrales resucitadas. Una vez obtenidos estos artículos, se procedió a utilizar la técnica de *text mining* para acceder a los textos y realizar una inspección automatizada

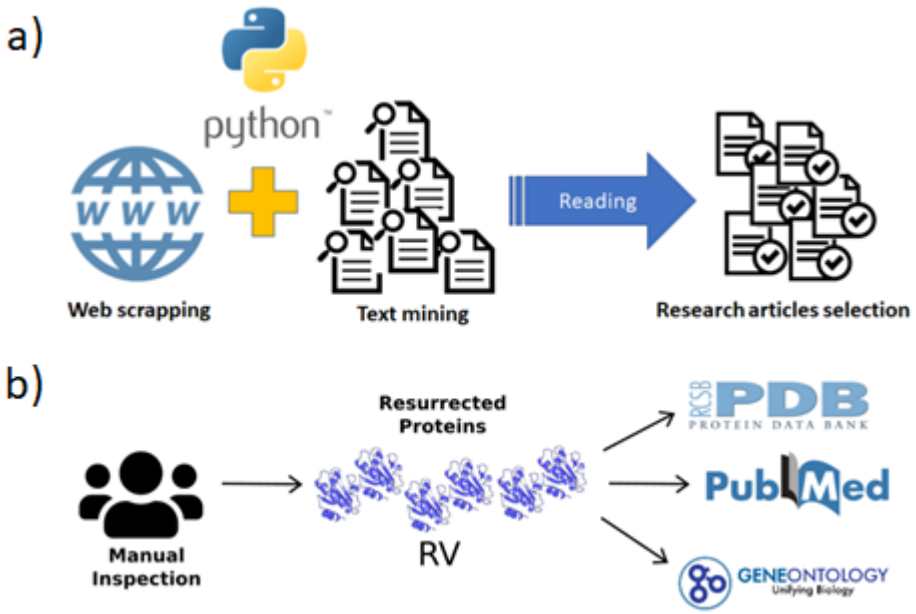


Figura 2.5. Esquema que representa el desarrollo de Revenant DB. a) Implementación de técnicas de "raspado web" (i.e *web-scraping)* y "minería de texto" (i.e *text-mining*) usando librerías estándares de Python. b) Inspección manual para identificar cada una de las citas primarias donde la proteína resucitada fue obtenida primero.

por medio de librerías de palabras claves (en idioma inglés) relacionadas con técnicas de ASR, resucitación y cristalización: *Ancestral (o Anc), Ancestral Proteins, Resuscitation, x-ray crystallography, Nuclear magnetic resonance (o NMR), Ancestral Sequence Reconstruction (o ASR), Maximum Likelihood (o ML), Bayesian,* entre otras. Aquí, cabe aclarar que las librerías disponibles en la web utilizaban palabras claves relacionadas con ciencias sociales (o redes sociales) por lo que se tuvo que desarrollar librerías con diccionarios propios. Así, mediante esta última técnica se logró filtrar los artículos inicialmente obtenidos, ya que los artículos que no tenían coincidencia con ninguna de dichas palabras claves podrían ser descartados. Para los artículos en los que sí había coincidencias con las palabras claves, se realizó una contabilización básica del número de veces en que aparecían dichas palabras y esto permitió establecer un orden de prioridad para proceder en el próximo paso. A diferencia de esta primera etapa, que como se dijo consistió mayormente en una inspección automatizada mediantes las técnicas mencionadas de programación, el paso subsiguiente consistió en la inspección manual de cada artículo.

Cada artículo fue inspeccionado de manera manual (Fig. 4.5.b) con el objetivo de corroborar que cada proteína ancestral identificada haya sido obtenida utilizando las técnicas de ASR y resurrección (descritas en la introducción de éste trabajo). Primero fueron inspeccionados los artículos con menor frecuencia de palabras claves con el objetivo de descartarlos rápidamente si no estaban relacionados con proteínas resucitadas. También en este paso de verificación se encontraron ciertas ambigüedades en la coincidencia de algunas palabras claves en relación con la metodología utilizada en algunas publicaciones científicas que no correspondían a metodologías de ASR. Por otro lado, de algunas publicaciones seleccionadas se identificó la publicación primaria donde se obtuvo la proteína resucitada, porque en algunos casos la proteína resucitada fue caracterizada en

una publicación pero fue obtenida por la técnica de ASR en una publicación previa. Una vez se identificaron inequívocamente cada una de las publicaciones primarias donde se obtuvieron una o varias proteínas resucitadas se comenzó el curado manual para anotar información relevante para la construcción de la base de datos.

4.5.2 Curado manual

Usando la información bibliográfica y el procedimiento del curado a mano de cada referencia, todas las entradas han sido anotadas con diferente datos relevantes: el nodo ancestral usado en la reconstrucción, su edad cronológica estimada, las metodologías de RSA usadas para inferir la secuencia ancestral, las secuencias usadas para el ASR y la estimación del árbol filogenético (cantidad de secuencias, bases de datos de las cuales fueron obtenidas y taxones a los que pertenecen esas secuencias), el programa usado para la estimación filogenética, el AMS y la ASR, disponibilidad de la estructura cristalizada y su caracterización de ligando. Además se anotó la cita primaria correspondiente donde la proteína fué obtenida por ASR. Para obtener una mejor caracterización de la proteína resucitada, también anotamos aquellas proteínas que tuvieran asociados parámetros bioquímicos que caracterizan la actividad enzimática (es decir, Km, kcat) y biofísicos (es decir, temperatura de desnaturalización, ΔGunfolding).

4.5.3 Asociación con otras fuentes de información biológica

Con el objetivo de aumentar la anotación biológica de Revenant, las proteínas presentes en Revenant DB han sido asociadas con otras bases de datos biológicas con el objetivo de hacer más rica la información para el usuario y por otra parte dar la posibilidad de realizar correlaciones de la información secuencial y estructurales con información biológica (función y participación en procesos biológicos) y fisicoquímica de las proteínas ancestrales. La vinculación se realizó mayormente de manera manual:

1. Los códigos de Protein Data Bank (PDB), la mejor base de datos de información estructural (httP://rcsb.org), de las estructuras cristalizadas fueron obtenidos por el procedimiento de curado manual de las publicaciones científicas (primarias o secundarias).
2. Los códigos de Gene Ontology (consorcio dedicado a la anotación funcional de proteínas y genes, http://geneontology.org/) fueron recuperados de la vinculación de PDB con esta base de datos.
3. Los códigos de pFam, bases de datos que acumula información de familias de proteínas homólogas, estructuras, relaciones funcionales, etc https://pfam.xfam.org/) fueron obtenidos utilizando un algoritmo de búsqueda secuencial dentro de esta base de datos y recuperando el primer hit.
4. Las entradas de CodNas (base de datos de diversidad conformacional de proteínas, http://ufq.unq.edu.ar/codnas/) fueron obtenidas identificando las secuencias de Revenant con más de un PDB asociado (i.e que tienen diversidad conformacional) y buscando el PDB dentro de dicha base de datos.
5. Los códigos de NCBI y Uniprot (principales bases de datos secuenciales fuertemente conectadas a bases de datos adicionales) fueron obtenidos de las listas de números de acceso de las proteínas homólogas utilizadas para los MSA para estimar el árbol filogenético (primera etapa de la ASR).

Toda la información anotada descrita anteriormente, fue estructurada en tablas basadas en un esquema conceptual de la base de datos (Fig. 4.5). Posteriormente, las tablas fueron transformadas a formato MySQL como fuente de datos para implementar la subida al servidor web. El servidor web se diseñó para visualizar la información contenida en Revenant DB de la forma más amigable posible.

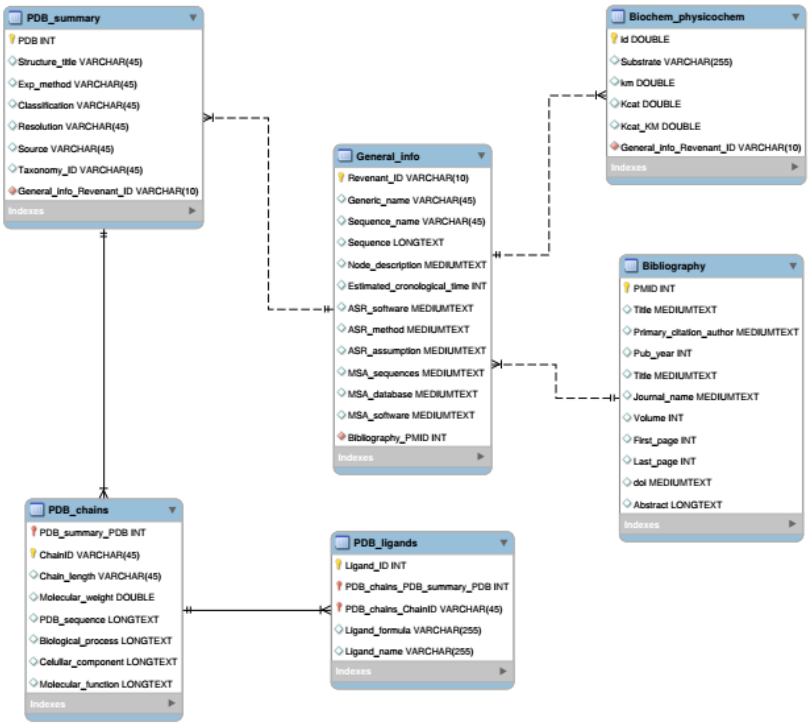


Figura 4.5 Esquema conceptual de la base de datos

Discusión general y perspectivas a futuro

En este trabajo pudimos estudiar varios casos en los que usamos información provista por proteínas ancestrales para poder estudiar y entender distintas propiedades y características de proteínas actuales. Esto se pudo lograr debido a que las proteínas ancestrales aportan información inestimable e inalcanzable por otros métodos.

Por un lado, pudimos realizar una extensa investigación sobre una tiorredoxina atípica, la EgIsTRP de helmintos, en la cual encontramos que la flexibilidad puede ser un parámetro muy importante para poder entender la función biológica diferencial que realizan, donde esta proteína y sus homólogas más cercanas (*Echinococcus* y *Taenia*) muestran una mayor flexibilidad en relación a las otras tiorredoxinas de helmintos, pero de otros géneros (*Hymenolepis*). Además, esto también se ve reflejado de alguna manera en relación a las tiorredoxinas canónicas (como la del humano, por ejemplo), ya que la diversidad conformacional de la EgIsTRP es mayor que la de la tiorredoxina humana. El uso de las técnicas de ASR fueron cruciales para esta investigación debido a que logramos obtener e inferir información esencial para plantear y/o probar ciertas hipótesis, imposible de obtener con otros métodos. Gracias a esta investigación también pudimos pulir y poner a punto varios detalles de esta técnica, lo que nos sirvió no solo en este caso, sino también para las siguientes investigaciones. Además, utilizar los estados ancestrales resucitados permitiría una mejor validación para la hipótesis, donde si bien, experimentalmente no se obtuvieron buenos resultados, si nos sirvió como aprendizaje.

Luego, hemos estudiado las tasas evolutivas de una población de proteínas humanas con evidencia experimental de agregación como fibrillas amiloides. Encontramos que este tipo de proteínas tiene un comportamiento inesperado, ya que a pesar de ser proteínas que son abundantes y tienen una alta tasa de expresión, también ocurre que tienen una tasa de evolución mayor que la de las proteínas de referencia utilizadas en el análisis. Al investigar más en detalle, encontramos que existe un grupo dentro de las proteínas de referencia

(proteínas secretadas) que no presentaban una diferencia en la tasa de evolución con respecto a los amiloides, pero que sí mostraban diferencias en otras características, haciendo, de esta forma, que los amiloides tengan un comportamiento propio. Una de estas características fue la diversidad conformacional, donde se vio que era mayor en las proteínas amiloidogénicas, lo cual puede deberse, en parte, por la gran cantidad de regiones desordenadas o altamente flexibles que poseen. Esta característica es la que podría ser la responsable de facilitar la exposición de las APRs, dirigiendo a la proteína hacia la formación de fibrillas. Otra diferenciación entre las proteínas amiloidogénicas y las proteínas secretadas se da, como decía anteriormente, en la abundancia y nivel de expresión, siendo mayor en ambas en las proteínas amiloidogénicas. Se sabe, además, que para que las proteínas con gran abundancia requieren de un apoyo constante de mecanismos de control de calidad para permanecer solubles. Las chaperonas en general tienden a evitar y revertir la formación de amiloides. En este punto quisimos estudiar si además esta interacción con las chaperonas le otorgaba a las proteínas amiloidogénicas una capacidad mayor para adquirir mutaciones desestabilizantes, lo cual en parte también iría de la mano con su alta tasa de evolución y gran diversidad conformacional. Para esto recurrimos a las técnicas de ASR, de modo de poder estudiar la estabilidad a lo largo de la evolución, teniendo en cuenta un árbol de 7 especies y recorriendo el camino desde la raíz hasta el ser humano. Nuevamente, hoy en día no hubiera sido posible realizar estos análisis sin las técnicas de ASR, por lo que la información que nos brinda es inestimable. Con estos análisis, demostramos además que podemos inferir información estructural de estos estados ancestrales sin la necesidad de tener que resucitar la proteína. Si bien pudimos obtener un indicio de que las proteínas amiloidogénicas tienden a tener mutación menos estables que las proteínas de referencia que no interaccionan con chaperonas, no fue posible obtener resultados concluyentes.

Por último, se creó una base de datos con información curada a mano que engloba a la gran cantidad de proteínas ancestrales reconstruidas y resucitadas conocidas al día de hoy,

Revenant (Carletti *et al*., 2020), así como también muchas de las características experimentales, tanto biofísicas como bioquímicas de las mismas. Toda esta información puede resultar muy útil a la hora de estudiar proteínas relacionadas, tal como se demostró en el caso de la tiorredoxina atípica. Además, esta base puede ofrecer ejemplos biológicos para desarrollar o mejorar diversas herramientas bioinformáticas. Además, esta base puede ofrecer ejemplos biológicos para desarrollar o mejorar diversas herramientas bioinformáticas. Este trabajo en particular, al ser una de las primeras incursiones en el mundo de las proteínas ancestrales fue muy provechoso, debido a que pudimos comprender y observar detalles más específicos de las técnicas y las herramientas, así como sus distintos usos y alcances.

El trabajo realizado durante esta tesis ha sido una experiencia enriquecedora, ya que, a pesar de varias dificultades y de proyectos truncados, hemos podido incursionar en un ámbito nuevo para el grupo, como lo es el estudio de proteínas ancestrales, mediante técnicas computacionales y posteriormente experimentales. Así como también, hemos podido demostrar que este tipo de estudios, no solo es útil, sino que imprescindible en muchos aspectos, debido a que ofrece información única. Como una breve perspectiva al futuro, se puede decir que se podrían encarar dos frentes, donde por un lado se puedan establecer las condiciones experimentales para poder estudiar en el futuro diversos sistemas en los que se requiera resucitar a las proteínas; y por otro lado, en cuanto a lo computacional, algo que se puede intentar hacer es mejorar las técnicas de reconstrucción usando mayor información en los modelos de evolución, como puede ser información estructural. Sea como fuere, todavía queda recorrer un largo camino en el uso y aprovechamiento de las técnicas de reconstrucción ancestral y resucitación.

Bibliografía

Abascal,F. *et al.* (2005) ProtTest: selection of best-fit models of protein evolution. *Bioinformatics*, **21**, 2104–2105.

Adey,N.B. *et al.* (1994) Molecular resurrection of an extinct ancestral promoter for mouse L1. *Proc Natl Acad Sci USA*, **91**, 1569–1573.

Agozzino,L. and Dill,K.A. (2018) Protein evolution speed depends on its stability and abundance and on chaperone concentrations. *Proc Natl Acad Sci USA*, **115**, 9092–9097.

Akanuma,S. (2017) Characterization of reconstructed ancestral proteins suggests a change in temperature of the ancient biosphere. *Life (Basel)*, **7**.

Akanuma,S. *et al.* (2013) Experimental evidence for the thermophilicity of ancestral life. *Proc Natl Acad Sci USA*, **110**, 11067–11072.

Akashi,H. (2003) Translational selection and yeast proteome evolution. *Genetics*, **164**, 1291–1303.

Allentoft,M.E. *et al.* (2012) The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proc. Biol. Sci.*, **279**, 4724–4733.

Altenhoff,A.M. *et al.* (2018) The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. *Nucleic Acids Res.*, **46**, D477–D485.

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Anisimova,M. *et al.* (2010) Finding the balance between the mathematical and biological optima in multiple sequence alignment. *Trends Evol. Biol.*, **2**, 7.

Anisimova,M. *et al.* (2013) State-of the art methodologies dictate new standards for phylogenetic analysis. *BMC Evol. Biol.*, **13**, 161.

Anisimova,M. and Gascuel,O. (2006) Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst. Biol.*, **55**, 539–552.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Atkinson,H.J. and Babbitt,P.C. (2009) An atlas of the thioredoxin fold class reveals the complexity of function-enabling adaptations. *PLoS Comput. Biol.*, **5**, e1000541.

Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.

Baker,C.R. *et al.* (2013) Following gene duplication, paralog interference constrains transcriptional circuit evolution. *Science*, **342**, 104–108.

Bar-Rogovsky,H. *et al.* (2013) The evolutionary origins of detoxifying enzymes: the mammalian serum paraoxonases (PONs) relate to bacterial homoserine lactonases. *J. Biol. Chem.*, **288**, 23914–23927.

Beerten,J. *et al.* (2012) Aggregation prone regions and gatekeeping residues in protein sequences. *Curr. Top. Med. Chem.*, **12**, 2470–2478.

Benner,S.A. *et al.* (2007) Molecular paleoscience: systems biology from the past. *Adv. Enzymol. Relat. Areas Mol. Biol.*, **75**, 1–132, xi.

Benson,D.A. *et al.* (2013) GenBank. *Nucleic Acids Res.*, **41**, D36-42.

Berman,H.M. *et al.* (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Berrera,M. *et al.* (2003) Amino acid empirical contact energy definitions for fold recognition in

the space of contact maps. *BMC Bioinformatics*, **4**, 8.
Best,R.B. *et al.* (2006) Relation between native ensembles and experimental structures of proteins. *Proc Natl Acad Sci USA*, **103**, 10901–10906.
Bisio,H. *et al.* (2016) A New Class of Thioredoxin-Related Protein Able to Bind Iron-Sulfur Clusters. *Antioxid. Redox Signal.*
Blanchet,G. *et al.* (2017) Ancestral protein resurrection and engineering opportunities of the mamba aminergic toxins. *Sci. Rep.*, **7**, 2701.
Boehr,D.D. *et al.* (2009) The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.*, **5**, 789–796.
Boucher,J.I. *et al.* (2014) An atomic-resolution view of neofunctionalization in the evolution of apicomplexan lactate dehydrogenases. *elife*, **3**.
Breen,M.S. *et al.* (2012) Epistasis as the primary factor in molecular evolution. *Nature*, **490**, 535–538.
Bridgham,J.T. *et al.* (2009) An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. *Nature*, **461**, 515–519.
Bridgham,J.T. *et al.* (2006) Evolution of hormone-receptor complexity by molecular exploitation. *Science*, **312**, 97–101.
Brookes,J.C. *et al.* (2012) System among the corticosteroids: specificity and molecular dynamics. *J. R. Soc. Interface*, **9**, 43–53.
Buck,P.M. *et al.* (2013) On the role of aggregation prone regions in protein evolution, stability, and enzymatic catalysis: insights from diverse analyses. *PLoS Comput. Biol.*, **9**, e1003291.
Burra,P.V. *et al.* (2009) Global distribution of conformational states derived from redundant models in the PDB points to non-uniqueness of the protein structure. *Proc Natl Acad Sci USA*, **106**, 10505–10510.
Butzin,N.C. *et al.* (2013) Reconstructed ancestral Myo-inositol-3-phosphate synthases indicate that ancestors of the Thermococcales and Thermotoga species were more thermophilic than their descendants. *PLoS ONE*, **8**, e84300.
Cai,W. *et al.* (2004) Reconstruction of ancestral protein sequences and its applications. *BMC Evol. Biol.*, **4**, 33.
Cam,L.L. (1990) Maximum likelihood: an introduction. *International Statistical Review / Revue Internationale de Statistique*, **58**, 153.
Cano,R.J. and Borucki,M.K. (1995) Revival and identification of bacterial spores in 25- to 40-million-year-old Dominican amber. *Science*, **268**, 1060–1064.
Cappellini,E. *et al.* (2014) Biochemistry. Unlocking ancient protein palimpsests. *Science*, **343**, 1320–1322.
Carletti, M. S., Monzon, A. M., Garcia-Rios, E., Benitez, G., Hirsh, L., Fornasari, M. S., & Parisi, G. (2020). Revenant: a database of resurrected proteins. *Database : the journal of biological databases and curation*, 2020, baaa031.
Carroll,S.M. *et al.* (2011) Mechanisms for the evolution of a derived function in the ancestral glucocorticoid receptor. *PLoS Genet.*, **7**, e1002117.
Chandler,C.H. *et al.* (2014) Causes and consequences of genetic background effects illuminated by integrative genomic analysis. *Genetics*, **196**, 1321–1336.
Chandrasekharan,U.M. *et al.* (1996) Angiotensin II-forming activity in a reconstructed ancestral chymase. *Science*, **271**, 502–505.
Chang,B.S.W. *et al.* (2002) Recreating a functional ancestral archosaur visual pigment. *Mol. Biol. Evol.*, **19**, 1483–1489.
Chinen,A. *et al.* (2005) Reconstitution of ancestral green visual pigments of zebrafish and

molecular mechanism of their spectral differentiation. *Mol. Biol. Evol.*, **22**, 1001–1010.
Chothia,C. and Lesk,A.M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J.*, **5**, 823–826.
Cilia,E. *et al.* (2013) From protein sequence to dynamics and disorder with DynaMine. *Nat. Commun.*, **4**, 2741.
Cilia,E. *et al.* (2014) The DynaMine webserver: predicting protein dynamics from sequence. *Nucleic Acids Res.*, **42**, W264-70.
Ciryam, P., Kundra, R., Morimoto, R. I., Dobson, C. M., & Vendruscolo, M. (2015). Supersaturation is a major driving force for protein aggregation in neurodegenerative diseases. *Trends in pharmacological sciences*, **36**, 72–77.
Ciryam, P., Antalek, M., Cid, F. *et al*. (2019) A metastable subproteome underlies inclusion formation in muscle proteinopathies. *acta neuropathol commun* **7**, 197.
Clemente,J.C. *et al.* (2009) Optimized ancestral state reconstruction using Sankoff parsimony. *BMC Bioinformatics*, **10**, 51.
Clifton,B.E. and Jackson,C.J. (2016) Ancestral Protein Reconstruction Yields Insights into Adaptive Evolution of Binding Specificity in Solute-Binding Proteins. *Cell Chem. Biol.*, **23**, 236–245.
Cole,M.F. *et al.* (2013) Reconstructing evolutionary adaptive paths for protein engineering. *Methods Mol. Biol.*, **978**, 115–125.
da Costa,G. *et al.* (2015) Transthyretin amyloidosis: chaperone concentration changes and increased proteolysis in the pathway to disease. *PLoS ONE*, **10**, e0125392.
DeGiorgio,M. and Rosenberg,N.A. (2016) Consistency and inconsistency of consensus methods for inferring species trees from gene trees in the presence of ancestral population structure. *Theor. Popul. Biol.*, **110**, 12–24.
Delgado, J., Radusky, L. G., Cianferoni, D., & Serrano, L. (2019). FoldX 5.0: working with RNA, small molecules and a new graphical interface. *Bioinformatics*, **35**, 4168–4169.
Devamani,T. *et al.* (2016) Catalytic promiscuity of ancestral esterases and hydroxynitrile lyases. *J. Am. Chem. Soc.*, **138**, 1046–1056.
Dobson,C.M. (1999) Protein misfolding, evolution and disease. *Trends Biochem. Sci.*, **24**, 329–332.
Dolinsky,T.J. *et al.* (2004) PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res.*, **32**, W665-7.
Domingues,F.S. *et al.* (2000) Structure-based evaluation of sequence comparison and fold recognition alignment accuracy. *J. Mol. Biol.*, **297**, 1003–1013.
Drummond,D.A. *et al.* (2006) A single determinant dominates the rate of yeast protein evolution. *Mol. Biol. Evol.*, **23**, 327–337.
Drummond,D.A. *et al.* (2005) Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci USA*, **102**, 14338–14343.
Durbin,R. *et al.* (1998) Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids Cambridge Univeristy Press (ed).
Dutheil,J. (2008) Ancestral Sequence Reconstruction: methods and applications.
Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.
Efron, B. and Tibshirani, R.J. (1994) An Introduction to the Bootstrap. Chapman and Hall/CRC, UK.
Eick,G.N. *et al.* (2012) Evolution of minimal specificity and promiscuity in steroid hormone receptors. *PLoS Genet.*, **8**, e1003072.
Eick,G.N. *et al.* (2017) Robustness of reconstructed ancestral protein functions to statistical

uncertainty. *Mol. Biol. Evol.*, **34**, 247–261.
Elleman,T.C. (1978) A method for detecting distant evolutionary relationships between protein or nucleic acid sequences in the presence of deletions or insertions. *J. Mol. Evol.*, **11**, 143–161.
Felsenstein,J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, **17**, 368–376.
Felsenstein,J. (2004) Inferring Phylogenies 2nd ed. Sinauer Associates Is An Imprint Of Oxford University Press, Sunderland, Mass.
Fernandez-Escamilla,A.-M. *et al.* (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol.*, **22**, 1302–1306.
Field,S.F. *et al.* (2006) Adaptive evolution of multicolored fluorescent proteins in reef-building corals. *J. Mol. Evol.*, **62**, 332–339.
Field,S.F. and Matz,M.V. (2010) Retracing evolution of red fluorescence in GFP-like proteins from Faviina corals. *Mol. Biol. Evol.*, **27**, 225–233.
Finnigan,G.C. *et al.* (2012) Evolution of increased complexity in a molecular machine. *Nature*, **481**, 360–364.
Fitch,W.M. (1971) Toward defining the course of evolution: minimum change for a specific tree topology. *Syst. Zool.*, **20**, 406.
Fogolari,F. *et al.* (2007) Scoring predictive models using a reduced representation of proteins: model and energy definition. *BMC Struct. Biol.*, **7**, 15.
Fomenko,D.E. *et al.* (2008) Functional diversity of cysteine residues in proteins and unique features of catalytic redox-active cysteines in thiol oxidoreductases. *Mol. Cells*, **26**, 228–235.
Fowler,D.M. *et al.* (2007) Functional amyloid--from bacteria to humans. *Trends Biochem. Sci.*, **32**, 217–224.
Fukuyama,K. (2004) Structure and function of plant-type ferredoxins. *Photosyn. Res.*, **81**, 289–301.
Fuller,P.J. *et al.* (2000) Specificity in mineralocorticoid versus glucocorticoid action. *Kidney Int.*, **57**, 1256–1264.
Gaucher,E.A. *et al.* (2003) Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. *Nature*, **425**, 285–288.
Gerlt,J.A. and Babbitt,P.C. (2009) Enzyme (re)design: lessons from natural evolution and computation. *Curr. Opin. Chem. Biol.*, **13**, 10–18.
Goldman,N. and Yang,Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.*, **11**, 725–736.
Gould,S.J. (1990) Wonderful Life: The Burgess Shale and the Nature of History Norton, W. W. & Company, Inc.
Guindon,S. *et al.* (2009) Estimating maximum likelihood phylogenies with PhyML. *Methods Mol. Biol.*, **537**, 113–137.
Guindon,S. *et al.* (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.*, **59**, 307–321.
Gumulya,Y. and Gillam,E.M.J. (2017) Exploring the past and the future of protein evolution with ancestral sequence reconstruction: the “retro” approach to protein engineering. *Biochem. J.*, **474**, 1–19.
Hanson-Smith,V. *et al.* (2010) Robustness of ancestral sequence reconstruction to phylogenetic uncertainty. *Mol. Biol. Evol.*, **27**, 1988–1999.
Harms,M.J. *et al.* (2013) Biophysical mechanisms for large-effect mutations in the evolution of steroid hormone receptors. *Proc Natl Acad Sci USA*, **110**, 11475–11480.

Harms,M.J. and Thornton,J.W. (2010) Analyzing protein structure and function using ancestral gene reconstruction. *Curr. Opin. Struct. Biol.*, **20**, 360–366.
Harms,M.J. and Thornton,J.W. (2014) Historical contingency and its biophysical basis in glucocorticoid receptor evolution. *Nature*, **512**, 203–207.
Harrison,R.S. *et al.* (2007) Amyloid peptides and proteins in review. *Rev. Physiol. Biochem. Pharmacol.*, **159**, 1–77.
Hart,K.M. *et al.* (2014) Thermodynamic system drift in protein evolution. *PLoS Biol.*, **12**, e1001994.
Hedges,S.B. *et al.* (2006) TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics*, **22**, 2971–2972.
Hedges,S.B. and Kumar,S. (2009) The Timetree of Life Oxford University Press, USA, Oxford.
Hendy,J. *et al.* (2018) A guide to ancient protein studies. *Nat. Ecol. Evol.*, **2**, 791–799.
Hervás,R. and Oroz,J. (2020) Mechanistic Insights into the Role of Molecular Chaperones in Protein Misfolding Diseases: From Molecular Recognition to Amyloid Disassembly. *Int. J. Mol. Sci.*, **21**.
Hobbs,J.K. *et al.* (2012) On the origin and evolution of thermophily: reconstruction of functional precambrian enzymes from ancestors of Bacillus. *Mol. Biol. Evol.*, **29**, 825–835.
Hochberg,G.K.A. and Thornton,J.W. (2017) Reconstructing ancient proteins to understand the causes of structure and function. *Annu. Rev. Biophys.*, **46**, 247–269.
Holmgren,A. (1989) Thioredoxin and glutaredoxin systems. *J. Biol. Chem.*, **264**, 13963–13966.
Hudson,W.H. *et al.* (2016) Distal substitutions drive divergent DNA specificity among paralogous transcription factors through subdivision of conformational space. *Proc Natl Acad Sci USA*, **113**, 326–331.
Huelsenbeck,J.P. and Rannala,B. (1997) Phylogenetic methods come of age: testing hypotheses in an evolutionary context. *Science*, **276**, 227–232.
Hunt,G. (2010) Evolution in fossil lineages: paleontology and The Origin of Species. *Am. Nat.*, **176 Suppl 1**, S61-76.
Ingles-Prieto,A. *et al.* (2013) Conservation of protein structure over four billion years. *Structure*, **21**, 1690–1697.
Ivics,Z. *et al.* (1997) Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. *Cell*, **91**, 501–510.
Iwabata,H. *et al.* (2005) Thermostability of ancestral mutants of Caldococcus noboribetus isocitrate dehydrogenase. *FEMS Microbiol. Lett.*, **243**, 393–398.
Jackson,M.P. and Hewitt,E.W. (2017) Why are Functional Amyloids Non-Toxic in Humans? *Biomolecules*, **7**.
Jacob,R.S. *et al.* (2016) Amyloid formation of growth hormone in presence of zinc: Relevance to its storage in secretory granules. *Sci. Rep.*, **6**, 23370.
Jeng,M.F. *et al.* (1995) Proton sharing between cysteine thiols in Escherichia coli thioredoxin: implications for the mechanism of protein disulfide reduction. *Biochemistry*, **34**, 10101–10105.
Jermann,T.M. *et al.* (1995) Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. *Nature*, **374**, 57–59.
Jobb,G. *et al.* (2004) TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. *BMC Evol. Biol.*, **4**, 18.
Jones,D.T. *et al.* (1992) The rapid generation of mutation data matrices from protein

sequences. *Comput. Appl. Biosci.*, **8**, 275–282.
Joy,J.B. *et al.* (2016) Ancestral Reconstruction. *PLoS Comput. Biol.*, **12**, e1004763.
Kim,H. *et al.* (2015) A hinge migration mechanism unlocks the evolution of green-to-red photoconversion in GFP-like proteins. *Structure*, **23**, 34–43.
Kim,H. *et al.* (2013) Acid-base catalysis and crystal structures of a least evolved ancestral GFP-like protein undergoing green-to-red photoconversion. *Biochemistry*, **52**, 8048–8059.
Kimura,M. (1968) Evolutionary rate at the molecular level. *Nature*, **217**, 624–626.
Kimura,M. and Ohta,T. (1974) On some principles governing molecular evolution. *Proc Natl Acad Sci USA*, **71**, 2848–2852.
Kohn,J.A. *et al.* (2012) Deciphering modern glucocorticoid cross-pharmacology using ancestral corticosteroid receptors. *J. Biol. Chem.*, **287**, 16267–16275.
Konno,A. *et al.* (2011) Tracing protein evolution through ancestral structures of fish galectin. *Structure*, **19**, 711–721.
Koonin,E.V. (2012) The Logic Of Chance: The Nature And Origin Of Biological Evolution (ft Press Science) 1st ed. Ft Press, Upper Saddle River, N.J.
Kosciolek,T. *et al.* (2017) Predictions of Backbone Dynamics in Intrinsically Disordered Proteins Using De Novo Fragment-Based Protein Structure Predictions. *Sci. Rep.*, **7**, 6999.
Koshi,J.M. and Goldstein,R.A. (1996) Probabilistic reconstruction of ancestral protein sequences. *J. Mol. Evol.*, **42**, 313–320.
Kratzer,J.T. *et al.* (2014) Evolutionary history and metabolic insights of ancient mammalian uricases. *Proc Natl Acad Sci USA*, **111**, 3763–3768.
Kumar,S. (2005) Molecular clocks: four decades of evolution. *Nat. Rev. Genet.*, **6**, 654–662.
Kumar,S. and Filipski,A. (2001) Molecular Phylogeny Reconstruction. In, John Wiley & Sons, Ltd (ed), *Encyclopedia of life sciences*. John Wiley & Sons, Ltd, Chichester, UK.
Kumar,S. and Filipski,A. (2007) Multiple sequence alignment: in pursuit of homologous DNA positions. *Genome Res.*, **17**, 127–135.
Kundra, R., Ciryam, P., Morimoto, R. I., Dobson, C. M., & Vendruscolo, M. (2017). Protein homeostasis of a metastable subproteome associated with Alzheimer's disease. *Proceedings of the National Academy of Sciences of the United States of America*, **114**, E5703–E5711.
Kuntz,A.N. *et al.* (2007) Thioredoxin glutathione reductase from Schistosoma mansoni: an essential parasite enzyme and a key drug target. *PLoS Med.*, **4**, e206.
Lane,N. (2016) The vital question: energy, evolution, and the origins of complex life. *Choice Reviews Online*, **53**, 53-2198-53–2198.
Langenberg,T. *et al.* (2020) Thermodynamic and Evolutionary Coupling between the Native and Amyloid State of Globular Proteins. *Cell Rep.*, **31**, 107512.
Lartillot,N. *et al.* (2009) PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics*, **25**, 2286–2288.
Le,S.Q. and Gascuel,O. (2008) An improved general amino acid replacement matrix. *Mol. Biol. Evol.*, **25**, 1307–1320.
Lemey,P. *et al.* eds. (2009) The phylogenetic handbook: A practical approach to phylogenetic analysis and hypothesis testing Cambridge University Press, Cambridge.
Lesk,A.M. and Chothia,C. (1980) How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J. Mol. Biol.*, **136**, 225–270.
Li,W.-H. (2006) Molecular Evolution PB Sinauer Associates, Inc.

Li,Y. *et al.* (2005) Structural and biochemical mechanisms for the specificity of hormone binding and coactivator assembly by mineralocorticoid receptor. *Mol. Cell*, **19**, 367–380.
Liberles,D.A. (2007a) Ancestral Sequence Reconstruction (oxford Biosciences) 1st ed. Oxford University Press, Oxford.
Liberles,D.A. (2007b) Ancestral Sequence Reconstruction (oxford Biosciences) 1st ed. Oxford University Press, Oxford.
Lill,R. (2009) Function and biogenesis of iron-sulphur proteins. *Nature*, **460**, 831–838.
Lill,R. *et al.* (2012) The role of mitochondria in cellular iron-sulfur protein biogenesis and iron metabolism. *Biochim. Biophys. Acta*, **1823**, 1491–1508.
Lillig,C.H. *et al.* (2008) Glutaredoxin systems. *Biochim. Biophys. Acta*, **1780**, 1304–1317.
López de la Paz,M. and Serrano,L. (2004) Sequence determinants of amyloid fibril formation. *Proc Natl Acad Sci USA*, **101**, 87–92.
Louros,N.N. *et al.* (2016) Intrinsic aggregation propensity of the CsgB nucleator protein is crucial for curli fiber formation. *J. Struct. Biol.*, **195**, 179–189.
Ma,B. and Nussinov,R. (2016) Protein dynamics: Conformational footprints. *Nat. Chem. Biol.*, **12**, 890–891.
Maddison,W.P. (1997) Gene Trees in Species Trees. *Syst. Biol.*, **46**, 523–536.
Maiolo,M. *et al.* (2018) Progressive multiple sequence alignment with indel evolution. *BMC Bioinformatics*, **19**, 331.
Maji,S.K. *et al.* (2009) Functional amyloids as natural storage of peptide hormones in pituitary secretory granules. *Science*, **325**, 328–332.
Malcolm,B.A. *et al.* (1990) Ancestral lysozymes reconstructed, neutrality tested, and thermostability linked to hydrocarbon packing. *Nature*, **345**, 86–89.
Marchler-Bauer,A. *et al.* (2013) CDD: conserved domains and protein three-dimensional structure. *Nucleic Acids Res.*, **41**, D348-52.
Marino-Buslje,C. *et al.* (2017) On the dynamical incompleteness of the Protein Data Bank. *Brief. Bioinformatics*.
Mashima,J. *et al.* (2016) DNA data bank of Japan (DDBJ) progress report. *Nucleic Acids Res.*, **44**, D51-7.
Maurer-Stroh,S. *et al.* (2010) Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nat. Methods*, **7**, 237–242.
Maury,C.P.J. (2009) The emerging concept of functional amyloid. *J. Intern. Med.*, **265**, 329–334.
Mayr,E. (1997) This is biology: The science of the living world Belknap Press of Harvard University Press, Cambridge, Mass.
McKeown,A.N. *et al.* (2014) Evolution of DNA specificity in a transcription factor family produced a new gene regulatory module. *Cell*, **159**, 58–68.
McGlinchey,R.P. *et al.* (2009) The repeat domain of the melanosome fibril protein Pmel17 forms the amyloid core promoting melanin synthesis. *Proc Natl Acad Sci USA*, **106**, 13731–13736.
Mehta,R.S. *et al.* (2016) The probability of monophyly of a sample of gene lineages on a species tree. *Proc Natl Acad Sci USA*, **113**, 8002–8009.
Miyazaki,J. *et al.* (2001) Ancestral residues stabilizing 3-isopropylmalate dehydrogenase of an extreme thermophile: experimental evidence supporting the thermophilic common ancestor hypothesis. *J. Biochem.*, **129**, 777–782.
Monzon,A.M. *et al.* (2013) CoDNaS: a database of conformational diversity in the native state of proteins. *Bioinformatics*, **29**, 2512–2514.
Monzon,A.M. *et al.* (2016) CoDNaS 2.0: a comprehensive database of protein

conformational diversity in the native state. *Database (Oxford)*, **2016**.
Monzon,A.M., Zea,D.J., Fornasari,M.S., *et al.* (2017) Conformational diversity analysis reveals three functional mechanisms in proteins. *PLoS Comput. Biol.*, **13**, e1005398.
Monzon,A.M., Zea,D.J., Marino-Buslje,C., *et al.* (2017) Homology modeling in a dynamical world. *Protein Sci.*, **26**, 2195–2206.
Mössner,E. *et al.* (2000) Influence of the pK_a value of the buried, active-site cysteine on the redox properties of thioredoxin-like oxidoreductases. *FEBS Lett.*, **477**, 21–26.
Mount,D. (2004) Bioinformatics: Sequence and Genome Analysis CSHL Press (ed).
Moutevelis,E. and Warwicker,J. (2004) Prediction of pKa and redox properties in the thioredoxin superfamily. *Protein Sci.*, **13**, 2744–2752.
Murrell,J.R. *et al.* (2000) Early-onset Alzheimer disease caused by a new mutation (V717L) in the amyloid precursor protein gene. *Arch. Neurol.*, **57**, 885–887.
Nei,M. (2005) Selectionism and neutralism in molecular evolution. *Mol. Biol. Evol.*, **22**, 2318–2342.
Notredame,C. *et al.* (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.
Orengo,C.A. *et al.* (1999) The CATH Database provides insights into protein structure/function relationships. *Nucleic Acids Res.*, **27**, 275–279.
Ortiz,A.R. *et al.* (2002) MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci.*, **11**, 2606–2621.
Ortlund,E.A. *et al.* (2007) Crystal structure of an ancient protein: evolution by conformational epistasis. *Science*, **317**, 1544–1548.
Overington,J. *et al.* (1992) Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. *Protein Sci.*, **1**, 216–226.
Overington,J. *et al.* (1990) Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. *Proc. Biol. Sci.*, **241**, 132–145.
Pagel,M. (1999) Inferring the historical patterns of biological evolution. *Nature*, **401**, 877–884.
Pál,C. *et al.* (2006) An integrated view of protein evolution. *Nat. Rev. Genet.*, **7**, 337–348.
Pamilo,P. and Nei,M. (1988) Relationships between gene trees and species trees. *Mol. Biol. Evol.*, **5**, 568–583.
Parisi,G. *et al.* (2021) “Protein” no longer means what it used to. *Current Research in Structural Biology*, **3**, 146–152.
Park,D. and Choi,S.S. (2009) Why proteins evolve at different rates: the functional hypothesis versus the mistranslation-induced protein misfolding hypothesis. *FEBS Lett.*, **583**, 1053–1059.
Perez-Jimenez,R. *et al.* (2011) Single-molecule paleoenzymology probes the chemistry of resurrected enzymes. *Nat. Struct. Mol. Biol.*, **18**, 592–596.
Perica,T. *et al.* (2014) Evolution of oligomeric state through allosteric pathways that mimic ligand binding. *Science*, **346**, 1254346.
Phillips,P.C. (2008) Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems. *Nat. Rev. Genet.*, **9**, 855–867.
Piovesan,D. *et al.* (2017) DisProt 7.0: a major update of the database of disordered proteins. *Nucleic Acids Res.*, **45**, D1123–D1124.
Piovesan,D. *et al.* (2018) MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins. *Nucleic Acids Res.*, **46**, D471–D476.
Posada,D. (2006) ModelTest Server: a web-based tool for the statistical selection of models of nucleotide substitution online. *Nucleic Acids Res.*, **34**, W700-3.

Posada,D. and Buckley,T.R. (2004) Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Syst. Biol.*, **53**, 793–808.
Punta,M. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290-301.
Pupko,T., Pe'er,I., *et al.* (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. *et al.* (2000) A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Mol. Biol. Evol.*, **17**, 890–896.
Pupko,T., Huchon,D., *et al.* (2002) Combining multiple data sets in a likelihood analysis: which models are the best? *Mol. Biol. Evol.*, **19**, 2294–2307.
Pupko,T. *et al.* (2007) Probabilistic models and their impact on the accuracy of reconstructed ancestral protein sequences. In, Liberles,D.A. (ed), *Ancestral Sequence Reconstruction*. Oxford University Press, pp. 43–57.
Pupko,T. *et al.* (2002) Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, **18 Suppl 1**, S71-7.
Rambaran,R.N. and Serpell,L.C. (2008) Amyloid fibrils: abnormal protein assembly. *Prion*, **2**, 112–117.
Risso,V.A. *et al.* (2017) De novo active sites for resurrected Precambrian enzymes. *Nat. Commun.*, **8**, 16113.
Risso,V.A. *et al.* (2013) Hyperstability and substrate promiscuity in laboratory resurrections of Precambrian β-lactamases. *J. Am. Chem. Soc.*, **135**, 2899–2902.
Risso,V.A. *et al.* (2015) Mutational studies on resurrected ancestral proteins reveal conservation of site-specific amino acid preferences throughout evolutionary history. *Mol. Biol. Evol.*, **32**, 440–455.
Rocha,E.P.C. (2006) The quest for the universals of protein evolution. *Trends Genet.*, **22**, 412–416.
Rodríguez,F. *et al.* (1990) The general stochastic model of nucleotide substitution. *J. Theor. Biol.*, **142**, 485–501.
Ronquist,F. *et al.* (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.*, **61**, 539–542.
Ronquist,F. and Huelsenbeck,J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–1574.
Rose,P.W. *et al.* (2011) The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res.*, **39**, D392-401.
Sankoff,D. (1975) Minimal mutation trees of sequences. *SIAM J. Appl. Math.*, **28**, 35–42.
Sankoff,D. and Rousseau,P. (1975) Locating the vertices of a steiner tree in an arbitrary metric space. *Math. Program.*, **9**, 240–246.
Schopf,J.W. *et al.* (2007) Evidence of Archean life: Stromatolites and microfossils. *Precambrian Res.*, **158**, 141–155.
Serpell,L.C. *et al.* (2000) The protofilament substructure of amyloid fibrils. *J. Mol. Biol.*, **300**, 1033–1039.
Shi,Y. and Yokoyama,S. (2003) Molecular analysis of the evolutionary significance of ultraviolet vision in vertebrates. *Proc Natl Acad Sci USA*, **100**, 8308–8313.
Sievers,F. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.

Sippl,M.J. (1993) Recognition of errors in three-dimensional structures of proteins. *Proteins*, **17**, 355–362.
Söding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–960.
Stackhouse,J. *et al.* (1990) The ribonuclease from an extinct bovid ruminant. *FEBS Lett.*, **262**, 104–106.
Stadler,T. *et al.* (2016) Does Gene Tree Discordance Explain the Mismatch between Macroevolutionary Models and Empirical Patterns of Tree Shape and Branching Times? *Syst. Biol.*, **65**, 628–639.
Starr,T.N. *et al.* (2017) Alternative evolutionary histories in the sequence space of an ancient protein. *Nature*, **549**, 409–413.
Starr,T.N. and Thornton,J.W. (2016) Epistasis in protein evolution. *Protein Sci.*, **25**, 1204–1218.
Stefanis,L. (2012) α-Synuclein in Parkinson's disease. *Cold Spring Harb. Perspect. Med.*, **2**, a009399.
Stefani,M. and Dobson,C.M. (2003) Protein aggregation and aggregate toxicity: new insights into protein folding, misfolding diseases and biological evolution. *J. Mol. Med.*, **81**, 678–699.
Steindel,P.A. *et al.* (2016) Gradual neofunctionalization in the convergent evolution of trichomonad lactate and malate dehydrogenases. *Protein Sci.*, **25**, 1319–1331.
Su,D. *et al.* (2007) A conserved cis-proline precludes metal binding by the active site thiolates in members of the thioredoxin family of proteins. *Biochemistry*, **46**, 6903–6910.
Suchard,M.A. and Redelings,B.D. (2006) BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. *Bioinformatics*, **22**, 2047–2048.
Sullivan,J. and Joyce,P. (2005) MODEL SELECTION IN PHYLOGENETICS. *Annu. Rev. Ecol. Evol. Syst.*, **36**, 445–466.
Sullivan,J. and Swofford,D.L. (1997) Are Guinea Pigs Rodents? The Importance of Adequate Models in Molecular Phylogenetics. *Journal of Mammalian Evolution*.
Sun,G. *et al.* (2002) Archaefructaceae, a new basal angiosperm family. *Science*, **296**, 899–904.
Swofford,D.L. and Sullivan,J. (2009) Phylogeny inference based on parsimony and other methods using PAUP. In, Lemey,P. *et al.* (eds), *The phylogenetic handbook: A practical approach to phylogenetic analysis and hypothesis testing*. Cambridge University Press, Cambridge, pp. 267–312.
Tartaglia,G.G. *et al.* (2008) Prediction of aggregation-prone regions in structured proteins. *J. Mol. Biol.*, **380**, 425–436.
Theobald,D.L. (2010) A formal test of the theory of universal common ancestry. *Nature*, **465**, 219–222.
Thomson,J.M. *et al.* (2005) Resurrecting ancestral alcohol dehydrogenases from yeast. *Nat. Genet.*, **37**, 630–635.
Thornton,J.W. (2001) Evolution of vertebrate steroid receptors from an ancestral estrogen receptor by ligand exploitation and serial genome expansions. *Proc Natl Acad Sci USA*, **98**, 5671–5676.
Thornton,J.W. (2004) Resurrecting ancient genes: experimental analysis of extinct molecules. *Nat. Rev. Genet.*, **5**, 366–375.
Thornton,J.W. *et al.* (2003) Resurrecting the ancestral steroid receptor: ancient origin of estrogen signaling. *Science*, **301**, 1714–1717.
Tokuriki,N. and Tawfik,D.S. (2009) Protein dynamism and evolvability. *Science*, **324**,

203–207.
Tokuriki,N. and Tawfik,D.S. (2009) Chaperonin overexpression promotes genetic variation and enzyme evolution. *Nature*, **459**, 668–673.
Tsai,C.J. *et al.* (1999) Folding funnels, binding funnels, and protein function. *Protein Sci.*, **8**, 1181–1190.
Ugalde,J.A. *et al.* (2004) Evolution of coral pigments recreated. *Science*, **305**, 1433.
UniProt Consortium (2008) The universal protein resource (UniProt). *Nucleic Acids Res.*, **36**, D190-5.
UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.
Unni,S. *et al.* (2011) Web servers and services for electrostatics calculations with APBS and PDB2PQR. *J. Comput. Chem.*, **32**, 1488–1491.
Varadi,M. *et al.* (2018) AmyPro: a database of proteins with validated amyloidogenic regions. *Nucleic Acids Res.*, **46**, D387–D392.
Ventura,S. *et al.* (2004) Short amino acid stretches can mediate amyloid formation in globular proteins: the Src homology 3 (SH3) case. *Proc Natl Acad Sci USA*, **101**, 7258–7263.
Walsh,I. *et al.* (2014) PASTA 2.0: an improved server for protein aggregation prediction. *Nucleic Acids Res.*, **42**, W301-7.
Wang,H.-C. *et al.* (2008) A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. *BMC Evol. Biol.*, **8**, 331.
Wang,L.-S. *et al.* (2011) The impact of multiple protein sequence alignment on phylogenetic estimation. *IEEE/ACM Trans Comput Biol Bioinform*, **8**, 1108–1119.
Webster,A.J. *et al.* (2003) Molecular phylogenies link rates of evolution and speciation. *Science*, **301**, 478.
Wei,G. *et al.* (2016) Protein ensembles: how does nature harness thermodynamic fluctuations for life? the diverse functional roles of conformational ensembles in the cell. *Chem. Rev.*, **116**, 6516–6551.
Weissmann,C. *et al.* (2002) Transmission of prions. *Proc Natl Acad Sci USA*, **99 Suppl 4**, 16378–16383.
Whelan,S. and Goldman,N. (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.*, **18**, 691–699.
Whittington,A.C. and Moerland,T.S. (2009) Ancestral sequence reconstruction and homology modeling link temperature adaptation and conservation of function with sequence evolution in parvalbumin. *Biophys. J.*, **96**, 651a.
Wiederstein,M. and Sippl,M.J. (2007) ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res.*, **35**, W407-10.
Wilke,C.O. (2004) Molecular clock in neutral protein evolution. *BMC Genet.*, **5**, 25.
Willerslev,E. *et al.* (2004) Long-term persistence of bacterial DNA. *Curr. Biol.*, **14**, R9-10.
Wilson,C. *et al.* (2015) Kinase dynamics. Using ancient protein kinases to unravel a modern cancer drug's mechanism. *Science*, **347**, 882–886.
Woese,C.R. (2004) A new biology for a new century. *Microbiol. Mol. Biol. Rev.*, **68**, 173–186.
Woese,C.R. (2000) Interpreting the universal phylogenetic tree. *Proc Natl Acad Sci USA*, **97**, 8392–8396.
Wood,T.C. and Pearson,W.R. (1999) Evolution of protein sequences and structures. *J. Mol. Biol.*, **291**, 977–995.
Worth,C.L. *et al.* (2009) Structural and functional constraints in the evolution of protein

families. *Nat. Rev. Mol. Cell Biol.*, **10**, 709–720.
Wu,C.H. *et al.* (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.*, **34**, D187-91.
Yang,Z. *et al.* (1995) A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics*, **141**, 1641–1650.
Yang,Z. (1996) Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol Evol (Amst)*, **11**, 367–372.
Yang,Z. *et al.* (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, **155**, 431–449.
Yang,Z. (2000) Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. *J. Mol. Evol.*, **51**, 423–432.
Yang,Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.*, **13**, 555–556.
Yang,Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.*, **24**, 1586–1591.
Zambrano,R. *et al.* (2015) AGGRESCAN3D (A3D): server for prediction of aggregation properties of protein structures. *Nucleic Acids Res.*, **43**, W306-13.
Zhang,J. and Nei,M. (1997) Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. *J. Mol. Evol.*, **44 Suppl 1**, S139-46.
Zhang,J. and Rosenberg,H.F. (2002) Complementary advantageous substitutions in the evolution of an antiviral RNase of higher primates. *Proc Natl Acad Sci USA*, **99**, 5486–5491.
Zhang,J. and Yang,J.-R. (2015) Determinants of the rate of protein sequence evolution. *Nat. Rev. Genet.*, **16**, 409–420.
Zhang,T. *et al.* (2013) Intrinsically semi-disordered state and its role in induced folding and protein aggregation. *Cell Biochem. Biophys.*, **67**, 1193–1205.
Zhu,G. *et al.* (2005) The selective cause of an ancient adaptation. *Science*, **307**, 1279–1282.