Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Departamento de Química Biológica

Desarrollo, diseño, implementación y aplicación de herramientas y métodos Bioinformáticos para el estudio de proteínas.

Tesis a presentar para optar al título de Doctor de la Universidad de Buenos Aires en el área de Química Biológica

Lic. María Inés Freiberger

Director de tesis: Dr. Diego U. Ferreiro

Consejero de estudios: Dr. Marcelo A. Martí

Lugar de trabajo: Laboratorio de Fisiología de Proteínas, IQUIBICEN-CONICET, FCEyN, UBA

Ciudad Autónoma de Buenos Aires, 31 de octubre de 2023

Resumen

Las proteínas son las macromoléculas clave que modulan todos los procesos que ocurren en los seres vivos. La diversidad de estos polímeros está dada por variaciones en las secuencias de aminoácidos, las que condicionan la formación de estructuras tridimensionales, sus dinámicas y su función.

En este trabajo de tesis diseñamos, implementamos y aplicamos, herramientas bioinformáticas basadas en el "Principio de mínima frustración", que establece que las proteínas se pliegan minimizando sus conflictos energéticos. Sin embargo, este principio no impide que haya conflictos remanentes en la estructura proteica. Es más, se ha demostrado que estos conflictos están relacionados a ciertos aspectos funcionales de las proteínas.

Nos enfocamos en el estudio del plegado y evolución de proteínas que contienen repeticiones de la familia ankirina. También estudiamos proteínas que contienen regiones desordenadas denominadas *fuzzy* y familias de proteínas evolutivamente relacionadas. Logramos caracterizar los patrones de frustración de las regiones *fuzzy* y de los sitios de unión proteína-ligando, los cuales se encuentran enriquecidos de interacciones altamente frustradas. De los estudios de la conservación de la frustración en familias de proteínas, fuimos capaces de especificar residuos implicados en el cambio conformacional, del estado no activado al estado activado, del factor de elongación bacteriano.

Caracterizamos la estructura exón-intrón de la familia de ankirinas, encontrando que el largo más frecuente de exones es de 99 nt, lo cual se corresponde con el largo de una repetición de ankirina (33 aminoácidos). Además vimos que los eventos de barajado de exones y de empalme alternativo actúan en estas proteínas, pero no pudimos detectar patrones claros y recurrentes de ocurrencia específica, indicando que estos eventos son variados. Caracterizamos el mecanismo de plegado de proteínas de la familia de ankirinas de diferentes largos usando el campo de fuerza *AWSEM*.

En síntesis, caracterizamos y logramos predecir sitios funcionales de las proteínas usando el concepto de frustración local. Caracterizamos la arquitectura de los genes que codifican ankirinas y los eventos que están implicados en su evolución. Por último analizamos y caracterizamos la dinámica y mecanismos de plegado de proteínas ankirinas de diferentes largos.

Estos resultados ofrecen nuevas herramientas y metodologías para el estudio de la biología estructural y la evolución de proteínas desde diferentes enfoques.

Palabras Claves: Evolución; Plegado de proteínas; Función proteica; Frustración; Proteínas repetitivas; Ankirinas; Genómica; Dinámica molecular.

Abstract

Development, design, implementation and application of bioinformatics tools and methods for the study of proteins.

Proteins are key macromolecules that modulate all processes occurring in living organisms. The diversity of these polymers is determined by variations in the amino acid sequences, which determine the formation of three-dimensional structures, their dynamics, and their function.

In this thesis, we designed, implemented, and applied bioinformatics tools based on the "Principle of Minimal Frustration", which says that proteins fold by minimizing their energy conflicts. However, this principle does not imply residual conflicts in the protein structure. In fact, it has been shown that these conflicts are related to certain functional aspects of proteins.

We focused on the study of folding and evolution of proteins containing ankyrin repeat motifs. We also studied proteins containing fuzzy regions, which are intrinsically disordered, and families of evolutionarily related proteins.

We were able to characterize the frustration patterns in the fuzzy regions and protein-ligand binding sites, which are enriched with highly frustrated interactions. By studying the conservation of frustration in protein families, we were able to identify residues involved in the conformational change from the non-activated state to the activated state of the bacterial elongation factor.

We characterized the exon-intron structure of the ankyrin family, finding that the most frequent exon length is 99 nt, corresponding to the length of an ankyrin repeat (33 amino acids). We also observed exon shuffling and alternative splicing events in these proteins, but we could not detect clear and recurrent patterns of specific occurrence, indicating that these events are diverse. We characterized the folding mechanism of ankyrin proteins of different lengths using the AWSEM force field.

In summary, we characterized and predicted functional sites of proteins using the concept of local frustration. We characterized the gene architecture encoding ankyrins and the events

involved in their evolution. Lastly, we analyzed and characterized the dynamics and folding mechanisms of ankyrin proteins of different lengths. These results provide new tools and methodologies for the study of structural biology and protein evolution from different perspectives.

Keywords: Evolution; Protein folding; Protein function; Frustration; Repetitive proteins; Ankyrins; Genomics; Molecular dynamics.

Agradecimientos

En la vida no estamos solos, siempre existe una primer persona que confía en nosotros y esa persona es la que nos da la valentía para seguir adelante y cumplir nuestros sueños. En mi caso, esa primer persona fue Pablo. Luego, durante el camino que transitamos, se van sumando personas que nos contagian con su amor a la ciencia, en mi caso tuve el honor de cruzarme con Gonzalo y Diego, dos personas que han contribuido a mi vida científica y a las cuales les voy a agradecer por siempre.

También les quiero agradecer a mi familia, en especial a mi mamá (Claudia), a mi hermana (Evangelina) y a mi abuela (Inés), tres mujeres son muy importantes en mi vida y siempre me han dado el amor, el cariño y las fuerzas para seguir por este camino tan hermoso que es la ciencia. También a mi cuñado Maxi que siempre me apoyó y a mis tres amadas sobrinas (Julia, Emilia y Victoria) que todos los días me demuestran su cariño y apoyo. A mi abuelo (Aurelio), a mi tío (Anselmo) y a mi papá (Rubén), que son y fueron personas que han cultivado en mi, desde muy chica, una de las partes esenciales del perfil científico, la curiosidad.

A todos mi amigos y amigas que les debo muchísimo, que me han soportado durante tantos años, a Emanuel y Guada, a las gurisas de Crespo (Flor, Meli y Pauli), a los gurises de la facultad (Nahuel, Diego y Analía) a Darío. A todos mis amigos académicos que me hice en la UBA y también en otros laboratorios y a todas las personas que contribuyeron en este proyecto. Especialmente a César, que es un genio, muy buen amigo y siempre esta con sus tips, a Ezequiel y Lucio mis compañeros de doctorado y a Nacho.

También le quiero agradecer a Camila que me bancó durante casi todo el doctorado, que me apoyó, que me aconsejó, básicamente, estuvo siempre ahí.

Por último, quiero agradecer a todxs lxs argentinxs, al sindicato y la federación de Luz y Fuerza, al CCAD de la Universidad Nacional de Córdoba ya que en este trabajo utilizó recursos computacionales del CCAD de la Universidad Nacional de Córdoba (https://ccad.unc.edu.ar/), que forman parte del SNCAD del MinCyT de la República Argentina.

Publicaciones

1. **Freiberger, M.I.***, Ruiz-Serra, V.*, Pontes, C. et al. Local energetic frustration conservation in protein families and superfamilies. Nat Commun 14, 8379 (2023).

2. Zheng, Y., Li, Q., **Freiberger, M. I.**, Song, H., Hu, G., Zhang, M., Li, J. (2023). Predicting the dynamic interaction between intrinsically disordered proteins. bioRxiv, 2023-12.

3. Parra, R. G.*, **Freiberger, M. I***., Poley-Gil, M., Fernandez-Martin, M., Radusky, L. G., Wolynes, P. G., Valencia, A. (2023). Frustraevo: A Web Server To Localize And Quantify The Conservation Of Local Energetic Frustration In Protein Families. bioRxiv, 2023-11.

4. **Freiberger, M. I.***, Clemente, C. M.*, Valero, E., Pombo, J. G., Leonetti, C. O., Ravetti, S., ..., Ferreiro, D. U. (2022). FrustraPocket: A protein–ligand binding site predictor using energetic local frustration. bioRxiv, 2022-12

5. Rausch AO*, **Freiberger MI***, Leonetti CO, Luna DM, Radusky LG, Wolynes PG, Ferreiro DU, Gonzalo Parra R. FrustratometeR: an R-package to compute local frustration in protein structures, point mutants and MD simulations. Bioinformatics. 2021 Mar 15

6. **Freiberger, M. I.**, Wolynes, P. G., Ferreiro, D. U., & Fuxreiter, M. (2021). Frustration in fuzzy protein complexes leads to interaction versatility. The Journal of Physical Chemistry B, 125(10), 2513-2520

7. Gianni, S., **Freiberger, M. I.**, Jemth, P., Ferreiro, D. U., Wolynes, P. G., & Fuxreiter, M. (2021). Fuzziness and frustration in the energy landscape of protein folding, function, and assembly. Accounts of chemical research, 54(5), 1251-1259

8. Clemente, C. M.*, **Freiberger, M. I*.**, Ravetti, S., Beltramo, D. M., & Garro, A. G. (2021). An in silico analysis of Ibuprofen enantiomers in high concentrations of sodium chloride with SARS-CoV-2 main protease. Journal of Biomolecular Structure and Dynamics, 1-12

9. Galpern, E. A., **Freiberger, M. I.**, & Ferreiro, D. U. (2020). Large Ankyrin repeat proteins are formed with similar and energetically favorable units. Plos one, 15(6), e0233865

10. **Freiberger, M. I.***, Guzovsky, A. B.*, Wolynes, P. G., Parra, R. G., & Ferreiro, D. U. (2019). Local frustration around enzyme active sites. Proceedings of the National Academy of Sciences, 116(10), 4037-4043

Índice general

Capítulo 1

Introducción

1.1. Plegado de proteínas, paisajes energéticos y frustración local

Las proteínas son unas de las biomoléculas orgánicas más fascinantes que se encuentran en la biosfera, además están implicadas en todos los procesos biológicos que ocurren en todos los seres vivos. Son polímeros formados por la unión mediante enlaces peptídicos de unidades estructurales llamadas aminoácidos. En la naturaleza existen 20 tipos diferentes de aminoácidos genéticamente codificados. La composición química y el largo de las proteínas es variable, es decir, pueden ser péptidos pequeños de tan solo unos 20 aminoácidos hasta proteínas de miles. La estructura tridimensional, la dinámica, la función y la localización celular de una proteína están determinados por las propiedades físico-químicas de los aminoácidos que la componen y el entorno, lo cual hace que estudiarlas sea un desafío. La estructura primaria de una proteína es la secuencia continua de los aminoácidos de la cadena polipeptídica. La estructura secundaria son plegados locales, que adopta la estructura primaria, en forma de hélices alfa u hojas beta que se forman mediante la unión de puentes de hidrógeno entre los átomos del enlace peptídico. Estas estructuras secundarias pueden interactuar entre ellas mediante diferentes tipos de interacciones, de manera tal que aminoácidos que estaban lejos en secuencia queden próximos en localización espacial, logrando así un plegado global (estructura terciaria). El proceso mediante el cual una proteína adquiere una estructura espacial

definida a partir de su estructura primaria se conoce como "plegado proteico".Este proceso es robusto, es decir, que cada vez que una proteína es sintetizada en el interior de una célula en las mismas condiciones fisiológicas, va a adoptar el mismo conjunto de estructuras que se corresponden con su estado nativo. También los paisajes energéticos son robusto a mutaciones en la estructura primaria, permitiendo la exploración de secuencias que pueden dar lugar a nuevas funciones (Morcos *et al.*, 2014). Sin embargo, no cualquier secuencia de aminoácidos puede plegarse, la capacidad de una cadena polipeptídica de plegarse y tener una actividad biológica relevante está asociada a su éxito evolutivo, es decir, secuencias que han sido positivamente seleccionadas debido a la acción de la selección natural.

El estudio del proceso de plegado proteico comenzó hace muchos años y aún sigue siendo un tema abierto. Actualmente, sabemos que la estructura tridimensional que adopta una proteína, en su mayoría, esta determinada por la secuencia de aminoácidos y el entorno celular, también conocemos que es un proceso colaborativo debido a las interacciones que se forman entre los aminoácidos y que existen proteínas llamadas chaperonas y cofactores que ayudan a las proteínas a plegarse. Lo que no conocemos con precisión son los mecanismo mediante por el cual una secuencia de proteínas se pliega o se despliega o cual es la relación entre la estructura y la función. En 1968, Cyrus Levinthal plantea que si una proteína tuviera que explorar al azar todas las conformaciones posibles para alcanzar la conformación correcta necesitaría de un tiempo mayor a la edad del universo debido al gran número de grados de libertad que tiene una cadena polipeptídica desplegada (Levinthal, 1968). Unos años más tarde, Christian B. Anfinsen, sugirió que el proceso de plegado es una búsqueda sesgada de un mínimo de energía libre, ya que pudo demostrar que la estructura de algunas proteínas podía ser adquirida espontáneamente a partir de sus cadenas desplegadas (Anfinsen, 1972).

La teoría utilizada para describir cómo las proteínas se pliegan es la teoría de "Paisajes Energéticos", en la cual se explica que una proteína es un sistema evolucionado de forma que las interacciones del estado nativo son energéticamente más favorables que las interacciones que se forman al azar durante el proceso de plegado. Por lo tanto cada vez que se forma una interacción presente en el estado nativo, la energía baja más que en el caso de que se forme una interacción al azar. Debido a la cooperatividad entre las interacciones nativas el paisaje energéticos de las proteínas tiene forma de embudo corrugado (Fig. 1.1) en donde existe un

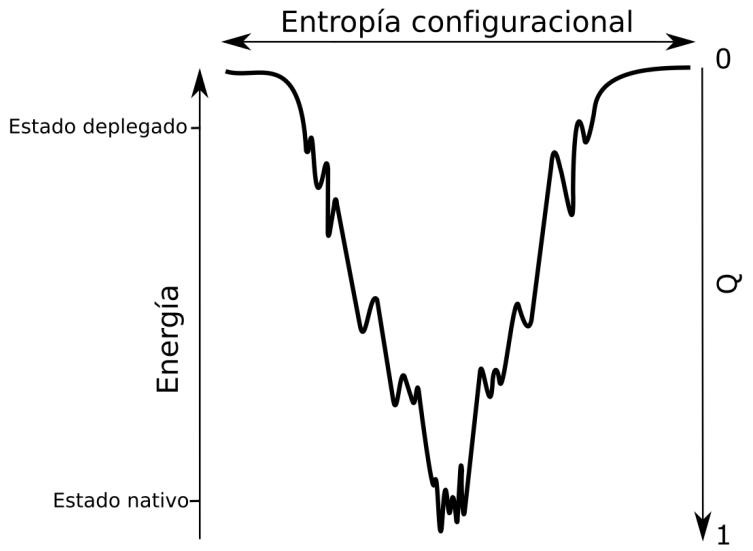

Figura 1.1: Esquema de un paisaje energético con forma de embudo corrugado que describe el proceso de plegado de una proteína natural. El ancho representa la entropía conformacional; la profundidad el cambio total en la energía entre el estado desplegado y el nativo y Q la fracción plegada.

fuerte sesgo energético hacia el estado nativo. Además, en la figura 1.1, podemos observar que el paisaje energético de una proteína natural posee algunos valles o mínimos locales que la proteína explora durante el proceso de plegado, los cuales representan estados estables, que pueden corresponderse a conformaciones intermedias transitorias o estados metaestables. Por lo tanto, durante el proceso de plegado una proteína no se pliega probando interacciones al azar sino que lo hace minimizando sus conflictos energéticos internos a medida que va adoptando conformaciones cada vez más parecidas a su estado nativo, el cual es un conjunto de estructuras de mínima energía, es decir, que cumplen con el "Principio de mínima frustración" (Bryngelson y Wolynes, 1987). Sin embargo, este principio no implica que todas las interacciones de una proteína tienen que estar energéticamente minimizadas, sino que pueden existir ciertos conflictos remanentes en la estructura, es decir, contactos que están energéticamente frustrados. Más aún, se ha demostrado que esta frustración presente en los estados nativos de las proteínas han sido evolutivamente seleccionados ya que estos conflictos tienen implicancias funcionales como por ejemplo, sitios catalíticos (Freiberger *et al.*, 2019), sitios alostéricos (Das y Plotkin, 2013), sitios de unión a proteínas (Gianni *et al.*, 2014), a cofactores, a ligandos y a la dinámica de la proteína (Ferreiro *et al.*, 2014). Además la frus-

tración esta relacionada directamente a cómo las proteínas se pliegan, ya que, las trampas o "desniveles" en el paisaje energético pueden demostrar la existencias de intermediarios de plegado, dado a que estas trampas son manifestaciones de la frustración topológica o energética (Ferreiro *et al.*, 2018).

Para poder calcular la frustración en estructuras de proteínas en 2007 Ferreiro y colaboradores desarrollaron un método llamado "*Protein Frustratometer*" (Ferreiro *et al.*, 2007). Este algoritmo esta basado en la teoría de paisajes energéticos y calcula la frustración energética local a nivel de contacto y de residuo único. Para calcular el índice de frustración (FI), a nivel de contacto o de residuo, el algoritmo primero calcula la energía de la proteína nativa, luego define los residuos en contacto como aquellos residuos cuya la distancia entre C_β (excepto glicina que carece de este átomo y se define la distancia desde el C_α), es menor a 9.5 $\mathring{A}$ y que estén a más de 2 residuos de distancia en secuencia. Luego, para cada contacto definido va a generar como máximo 2000 señuelos, los señuelos pueden ser calculados de dos maneras diferentes, para el caso del índice *mutational*, solamente se modifica la identidad de los aminoácidos en contacto por los 19 restantes. Por otro lado, para el índice *configurational*, no solo se modifica la identidad de los residuos en contacto sino también la distancia del contacto, lo que modifica la exposición al solvente de los residuos y su configuración espacial. Por último a nivel de residuo, para este modo no se definen contactos en la estructura, solamente, para cada residuo, se modifica su identidad. Una vez generados todos los señuelos, se calcula la distribución de la diferencia entre la energía nativa y la energía de cada señuelo, el FI, por contacto o por residuo, se define como un *Z-score* de la distribución de la diferencia de las energías (ec. 1.1). Este algoritmo fue luego fue implementado como un servicio web que se encuentra disponible para su uso de forma libre (Parra *et al.*, 2016). En el marco de esta tesis implementamos el "*Protein Frustratometer*" como un paquete de R y además le agregamos nuevas funcionalidades.

$$FI_{ij} = (\mathcal{H}^N_{ij} - < \mathcal{H}^U_{i'j'} >) / \sqrt{1/N \sum_{k=1}^{n} (\mathcal{H}^U_{i'j'} - < \mathcal{H}^U_{i'j'} >)^2} \tag{1.1}$$

Donde FI_{ij} es el índice de frustración par los residuos i y j en contacto, $\mathcal{H}^N$ es la energía nativa, n es la cantidad de señuelos y $\mathcal{H}^U_{i'j'}$ es la energía de los señuelos.

1.2. Evolución de proteínas

El estudio de la evolución de proteínas es una disciplina que busca comprender cómo se producen los cambios en la secuencia y estructura de los genes que codifican proteínas, así como también cambios en la secuencia de las proteínas y cómo estos cambios afectan su estructura y, a su vez, su función. Se han identificado varios mecanismos que contribuyen a la diversidad genética y la aparición de nuevas proteínas (Force *et al.*, 1999; Hughes, 1994; Ohno, 2013; Xu *et al.*, 2012). Uno de ellos es la acumulación de mutaciones puntuales, donde cambios aleatorios en la secuencia de ADN generan variaciones en la secuencia de aminoácidos de las proteínas codificadas (Saks *et al.*, 1998). Otro mecanismo importante en la evolución de proteínas, es la duplicación o deleción de material genético, es decir, cuando un fragmento de material genético es duplica o elimina, las copias adicionales pueden acumular mutaciones y evolucionar independientemente, adquiriendo nuevas funciones o especializaciones a lo largo del tiempo (Conant y Wolfe, 2008; Force *et al.*, 1999; Lynch y Conery, 2000). Por otro lado, la transferencia horizontal de genes, que implica la transferencia de material genético entre organismos no relacionados, puede introducir nuevos genes en una especie y contribuir a la evolución de proteínas (Reeves, 1993; Zhaxybayeva y Gogarten, 2004). El barajado de exones es otro mecanismo interesante que puede conducir a la formación de proteínas diferentes. Este mecanismo también implica la duplicación de exones, pero a diferencia del mecanismo de duplicación, el barajado de exones no solo tiene en cuenta la simetría de los exones sino que también a los intrones (regiones no codificantes). Estableciendo que solamente los exones simétricos pueden dulpicarse y debe insertarse en intrones de la misma clase (Patthy, 1996). Por último el empalme alternativo, en este proceso los exones, pueden ser reordenados o combinados de diferentes maneras durante la maduración del ARN mensajero, generando variantes de proteínas con funciones distintas (Gilbert, 1978; Rosenfeld *et al.*, 1982; Sharp, 2005). Es importante destacar que el empalme alternativo, aunque no es un proceso evolutivo en sí mismo, es un mecanismo regulador que permite la generación de múltiples transcriptos de ARN mensajero a partir de un gen, a través de la combinación diferencial de exones. Esto amplía aún más la diversidad de proteínas que pueden ser producidas a partir de un solo gen

.

En resumen, la evolución de los genes eucariotas que codifican proteínas involucra una interacción compleja entre mutaciones, duplicaciones, transferencias genéticas y otros mecanismos que conducen a la diversificación de las secuencias, estructuras y funciones de las proteínas. El estudio de estos mecanismos nos brinda una visión más completa de cómo han surgido y diversificado las proteínas a lo largo de la historia evolutiva. En este trabajo de tesis nos centraremos en aquellas mutaciones que dan lugar a la formación de nuevos genes que codifican proteínas. Para poder entender como son los mecanismos evolutivos es necesario conocer como es la estructura de un gen que codifica para una proteína.

1.2.1. Estructura de los genes que codifican proteínas

En la figura 1.2A, se muestra un esquema de la estructura de un gen eucariota que codifica para una proteína. Un gen está compuesto de una parte que se transcribe y otra que no se transcribe. La región "río arriba" de la parte que se transcribe (región 5') contiene en su mayoría señales que regulan el proceso de iniciación de la transcripción, llamada "Región Promotor". Por otro lado, la región 3', contiene señales para la terminación del proceso de transcripción y una adición de cadena poli-A (Patthy, 2009).

Típicamente, en eucariotas, la parte que se transribe consiste en los exones y los intrones. Los intrones, son secuencias que son removidas del transcripto primario (pre-ARNm) durante el proceso de pre-ARNm a ARNm (ARN mensajero), lo cual se conoce como maduración del ARNm (Adams *et al.*, 1996; Moore *et al.*, 1993). Las secuencias que se conservan en el ARNm maduro, los exones, se traducen en proteínas. Dentro de los exones, existe una región específica conocida como marco abierto de lectura (ORF por sus siglas en inglés, *Open Reading Frame*). El ORF es la secuencia de nucleótidos en el ARNm que se traduce en una secuencia de aminoácidos durante el proceso de síntesis de proteínas. Esta secuencia de aminoácidos en el ORF determina la composición y estructura de la proteína resultante.

Los intrones se clasifican en 3 fases, según la posición en el que interrumpen el marco de lectura (1.2B). Si un intrón interrumpe el marco de lectura entre dos codones consecutivos se denomina de fase 0, si lo interrumpe entre el nucleótido 1 y 2 de un mismo codón es de fase 1 y por último si el intrón interrumpe el marco de lectura entre el nucleótido 2 y 3 de un mismo codón es de fase 2 (Patthy, 1987).

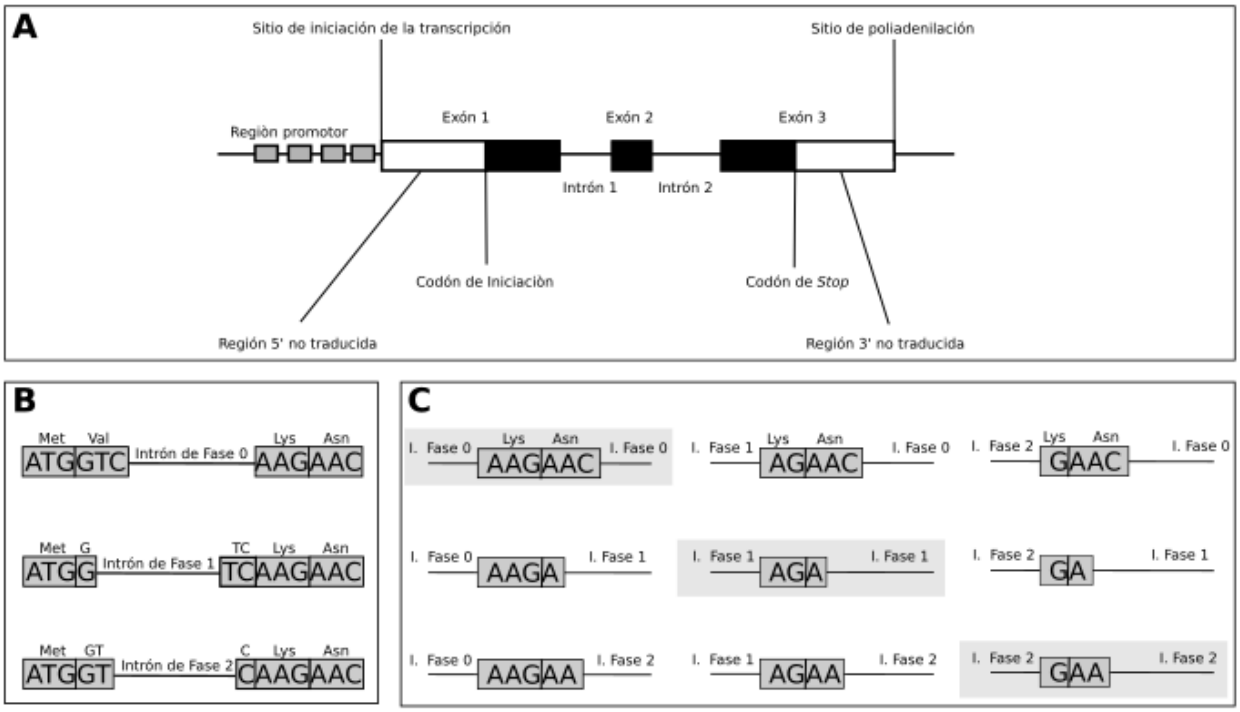

Figura 1.2: A) Esquema de un gen eucariota que codifica para proteína, los rectángulos representan exones, los cuadrados chicos representan elementos regulatorios en la región promotora y las líneas representan los intrones. Esta figura se la extrajo de [Patthy, L. (2009)]. B) Clasificación de los intrones según su fase, en rectángulos se representan los codones de un mismo exón, en líneas los intrones. C) Clase de exones, en rectángulos se representan los codones de un mismo exón, en líneas los intrones y en fondo gris la clase de exones que se clasifican como simétricos.

Por otro lado, los exones se clasifican según la fase de los intrones flanqueantes, en total hay 9 clases (1.2C). Hay 3 clases que se denominan clases simétricas las cuales son, (0-0), (1-1) y (2-2), estos son los únicos que pueden insertarse en intrones, de la misma fase, bajo eventos de duplicación en tándem en intrones adyacentes y pueden eliminarse o duplicarse sin alterar el marco de lectura. Las demás clases, (0-1), (0-2), (1-0), (1-2), (2-0) y (2-1) son asimétricas (Patthy, 1987).

1.2.2. Duplicación de genes

El primer evento de duplicación de genes se observó en el *bar locus* de *Drosophila* en 1936 por Bridges (Bridges, 1936). No obstante, el significado evolutivo de la duplicación de ge-

nes fue reconocido unos pocos años antes (Haldane, 1932; Muller, 1935), sugiriendo que la duplicación redundante de un gen puede adquirir mutaciones divergentes y eventualmente emerger como un nuevo gen. Años más tarde en los años 1970s y gracias al desarrollo de los métodos de secuenciación y de herramientas para el estudio de procesos evolutivos, se pudieron detectar más genes que evolucionaron mediante este mecanismo como por ejemplo la hemoglobina (Braunitzer *et al.*, 1961; Itano, 1957; Schroeder *et al.*, 1963).

Los tipos reconocidos de duplicación de genes son, (1) duplicación parcial o interna de un gen, (2) duplicación completa de un gen, (3) duplicación parcial de un cromosoma, (4) duplicación entera de un cromosoma y (5) poliploidía o duplicación de un genoma (Innan y Kondrashov, 2010). Cabe aclarar que los genes que se duplican no están exentos a que otros mecanismos evolutivos actúen sobre ellos para dar origen a nuevas funciones y que también puede suceder que la copia pierda su funcionalidad mediante el truncamiento de la traducción o transcripción. En esta tesis nos centraremos en los eventos de duplicación que ocurren dentro de un mismo gen.

La duplicación de un exón se refiere a la duplicación de uno o más exones en un gen y es uno de los tipos más importantes de duplicaciones internas de un gen. Muchas de las proteínas que conocemos actualmente muestran repeticiones internas de secuencias de aminoácidos, y las repeticiones a menudo corresponden a dominios funcionales o estructurales dentro de las proteínas (Barker *et al.*, 1978).

1.2.3. Barajado de exones

El barajado de exones es un mecanismo importante en la evolución de los genes que codifican proteínas. Este proceso involucra la duplicación de exones dentro de un gen, así como la unión de exones de diferentes genes a través de la recombinación en intrones. Este proceso puede contribuir a la generación de proteínas con características novedosas o a la adición de dominios funcionales provenientes de otros genes. Además ha ocurrido a lo largo de la evolución y han desempeñado un papel fundamental en la creación de nuevos genes y en la generación de diversidad en las proteínas. La comprensión de este mecanismo es importante para el estudio de la evolución de las proteínas y su implicación en la diversidad biológica.

Como ya se comentó anteriormente, solamente los exones que son simétricos son los únicos

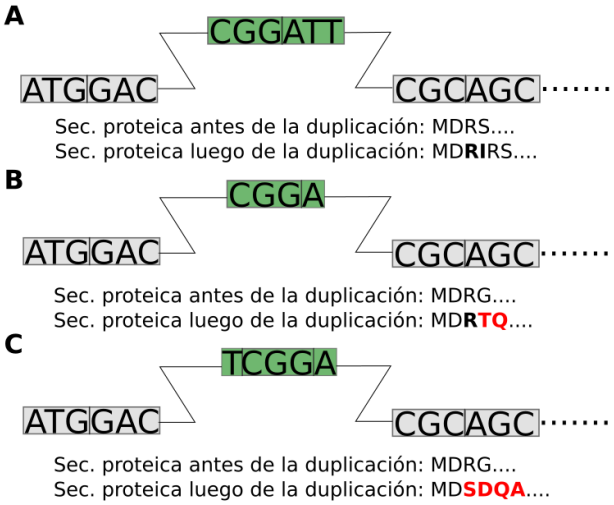

Figura 1.3: Duplicación de un exón e inserción en un intrón de fase 0, A) Duplicación de un exón de clase 0-0 (verde). B) Duplicación de un exón de clase 0-1 (verde). C) Duplicación de un exón de clase 2-1 (verde). En cajas grises los codones de los exones flanqueantes a la inserción del exón duplicado y los intrones se representan con líneas entre los exones. En rojo se muestra como se corre el marco de lectura cuando se inserta un exón que no es de la misma clase que la fase del intrón en el cual se inserta.

que pueden ser duplicados en tandém o eliminados sin ocasionar un corrimiento en el marco de lectura. Por lo tanto, la restricción del barajado de exones está dada por la fase de los intrones y la clase de los exones ya que la duplicación de un exón asimétrico interrumpiría el marco de lectura consecuente al punto de inserción (Fig. 1.3).

La inserción de exones simétricos también esta restringida, ya que, los exones de clase (0-0) solo pueden insertarse en intrones de fase 0, los exones de clase (1-1) solo pueden insertarse en intrones de fase 1 y lo mismo para los exones de clase (2-2) que solo se pueden insertar en intrones de fase 2 (Patthy, 1987). Esto sugiere que luego de que ocurra un evento de duplicación por barajado de exones la fase de los intrones y la clase de los exones del nuevo gen es no aleatoria, lo que indica que esta propiedad puede ser una característica importante para detectar eventos de duplicación por barajado de exones. Un aspecto importante de la combinación aleatoria de exones es que el impacto de la inserción de un exón puede mitigarse inicialmente mediante eventos de empalme alternativo.

1.2.4. Empalme Alternativo

Como ya se mencionó, los genes eucariotas están compuestos de exones e intrones. El empalme es el proceso mediante el cual se eliminan los intrones del pre-ARNm para dar lugar a un ARN maduro que solo contiene a los exones (Leff *et al.*, 1986). El empalme alternativo es es un tipo de empalme mediante el cual los exones se combinan para generar distintos transcriptos de ARNm. Su significancia evolutiva es que un solo gen puede codificar una gran cantidad de transcriptos y de proteínas que pueden tener funciones diferentes (Gilbert, 1978).

Los eventos de empalme alternativo son procesos reguladores que ocurren durante la trans-

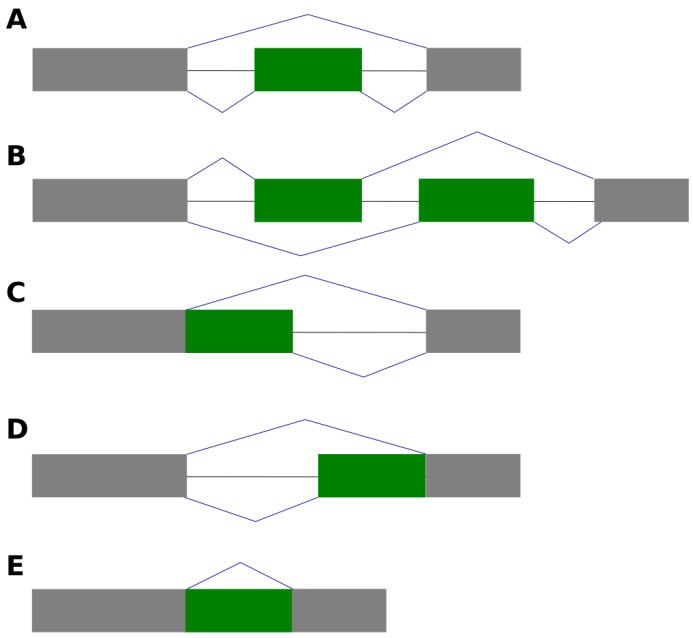

Figura 1.4: Mecanismo de empalme alternativo. A) Salto de exón. B) Exones mutuamente excluyentes. C) Sitio donante alternativo. D) Sitio aceptor alternativo. E) Retención de intrón. En verde el exón implicado en el evento de empalme alternativo, en gris los exones que no se ven afectados por este evento y los intrones se representan con líneas entre los exones.

cripción del ARN precursor mensajero (pre-ARNm) y dan lugar a la generación de múltiples variantes de ARNm a partir de un mismo gen. Estos eventos se pueden clasificar en cinco modos básicos. Salto de exón, en este modo, un exón completo se omite durante el proceso de empalme, lo que resulta en la ausencia de ese exón en la forma madura del ARNm. Exones mutuamente excluyentes, dos exones diferentes se encuentran en la misma posición en el gen,

pero solo uno de ellos se retiene en el ARNm final después del empalme. Esto significa que solo una de las variantes de ARNm tendrá ese exón en particular, mientras que la otra lo omitirá. Sitio donante alternativo, se produce una variación en el sitio donante del empalme 5'. Esto implica un cambio en el límite 3' del exón río arriba, lo que resulta en diferentes variantes de ARNm con distintas longitudes de exones río arriba. Sitio aceptor alternativo, en este caso, se produce un cambio en el sitio aceptor del empalme 3'. Esto implica un cambio en el límite 5' del exón río abajo, lo que da lugar a variantes de ARNm con diferentes longitudes de exones río abajo. Retención de intrón, un intrón que normalmente se eliminaría durante el proceso de empalme se retiene en la forma madura del ARNm. Esto puede resultar en la presencia de una secuencia de intrón en la proteína final o en la generación de una proteína truncada.

Las isoformas productos del empalme alternativo pueden tener características diferentes a la de referencia como la función, la afinidad por un ligando u otra proteína, la composición de dominios, la localización celular y la vida media (Light y Elofsson, 2013). También se producen isoformas que no tienen función o dan como producto proteínas anormales, esto puede estar asociado a que hay un sesgo hacia la producción de isoformas que eliminan dominios completos (Kriventseva *et al.*, 2003).

1.3. Proteínas Repetitivas

Las proteínas repetitivas están compuestas por repeticiones en tándem de uno o más motivos estructurales y funcionales, que suelen tener una longitud de entre 20 y 40 aminoácidos. Estas repeticiones se encuentran consecutivas en la secuencia de la proteína y pueden estar presentes en distintas combinaciones y números. Las proteínas repetitivas desempeñan un papel importante en numerosos procesos biológicos. Debido a su estructura modular, estas proteínas pueden tener diversas funciones, como mediar interacciones con otras moléculas y regulación de procesos celulares (Blatch y Lässle, 1999; Kobe y Kajava, 2001; Sedgwick y Smerdon, 1999). Además, las proteínas repetitivas son muy abundantes en los organismos y se estima que aproximadamente el 50 % de las proteínas contiene al menos una repetición en tándem. La presencia de repeticiones en tándem en las proteínas proporciona flexibilidad y

plasticidad estructural, lo que les permite adaptarse a diferentes funciones y contribuir a la diversidad y complejidad de los sistemas biológicos (Delucchi *et al.*, 2020; Marcotte *et al.*, 1999).

Estas proteínas pueden clasificarse en 5 clases (Tabla 1.1), que depende del largo de la unidad que se repite (Kajava, 2012), las más estudiadas son las de las clases III y IV. El la figura 1.5B se muestran ejemplos de proteínas pertenecientes a cada clase. En el marco de esta tesis

Clase	Largo unidad que se repite (AA)	Descripción
I	< 5	Pueden formar agregados insolubles con estructuras cristalinas
II	< 5	Pueden formar largas hélices enrolladas de estructuras fibrosas
III	Entre 5-40	Pueden formar estructuras abiertas del tipo solenoide
IV	Entre 5-40	Pueden formar estructuras cerradas del tipo toroidal
V	> 50	Pueden formar dominios que se pliegan en forma independiente

Tabla 1.1: Clasificación de las proteínas repetitivas

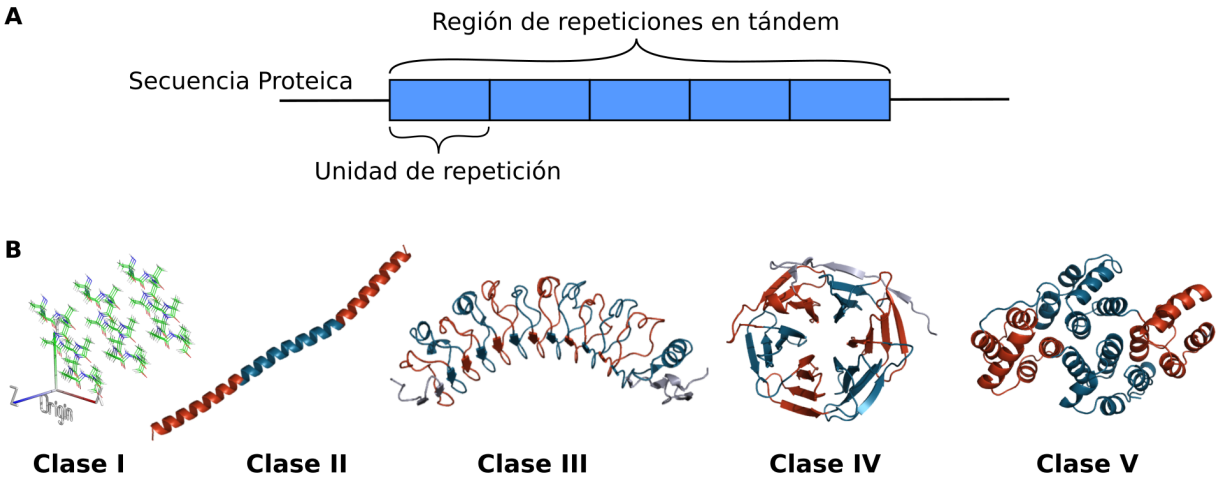

Figura 1.5: A) Esquema de la organización de una proteína que contiene repeticiones en tándem. B) Ejemplo de estructuras para las 5 clases diferentes de proteínas repetitivas.

nos centraremos en las repeticiones que forman estructuras del tipo solenoide (clase III). Las proteínas de esta clase se pueden diferenciar por la composición de estructuras secundarias,

que pueden ser del tipo β (como por ejemplo las WD-40), de composición mixta α/β (como por ejemplo las *Leucine rich*) o de composición todo del tipo α (como por ejemplo, las ankirinas, *HEAT*, armadillo y las TPR). Además se pueden diferenciar por la disposición y orden de las repeticiones, pueden estar dispuestas en una secuencia lineal continua o pueden formar estructuras superenrolladas en forma de hélice. También se pueden diferenciar por su estructura primaria, es decir, por el largo de las repeticiones y por la composición de aminoácidos de cada repetición. Como por ejemplo, las repeticiones de la familia ankirina se caracterizan por tener un largo de 33 aminoácidos y un motivo lineal TPLH (Parra *et al.*, 2015), mientras que las repeticiones de la familia *leucin rich* son de largo de 24 aminoácidos y están compuestas de motivos lineales LRR.

Uno de los desafíos que presentan estas proteínas es el estudio del plegado ya que, a diferencia de las proteínas globulares, no forman interacciones entre residuos distales en secuencias, es decir, que las interacciones entre los aminoácidos se dan de forma local o entre repeticiones vecinas (Main *et al.*, 2003a). Esto tiene un efecto en la estabilidad estructural, en el empaquetamiento de las repeticiones y en su función. Por lo tanto, mutaciones de cambio de aminoácidos en las repeticiones pueden tener implicaciones en la estabilidad y la función de la proteína, ya que puede influir en el empaquetamiento de las repeticiones y en su capacidad de interactuar con otras proteínas o moléculas. Todas estas características de las proteínas repetitivas, las hacen un buen modelo para estudiar el plegado, ya que es más fácil cuantificar y localizar el efecto de modificaciones en la secuencia, debido a pueden alterar su estabilidad, estructura y función.

El hecho de que las proteínas repetitivas formen contactos locales no significa que el plegado de cada repetición sea independiente. Recientemente se ha demostrado que esto no ocurre así, sino que el plegado es cooperativo en donde las interacciones vecinas cooperan en la estabilización de la estructura de las repeticiones (Espada *et al.*, 2015). Para estudiar el plegado de estas proteínas se han generado diferentes tipos de variantes, en la cuales se eliminaron o insertaron repeticiones enteras con el objetivo de ver si estas mutantes eran capaces de plegarse. Se llegó a la conclusión de que los arreglos resultantes, de eliminar o insertar repeticiones, mantienen su plegado y que para el caso de las Ankirinas y TPR incrementa la estabilidad de la proteína a medida que se aumenta la cantidad de repeticiones (Main *et al.*,

2003a). Además para las ankirinas se observó que, a mayor cantidad de repeticiones, más parecidas en secuencia son entre repeticiones vecinas y que las repeticiones que constituyen arreglos largos son energéticamente más favorables que las que forman arreglos más cortos (Galpern *et al.*, 2020). También se observó que los arreglos de largo entre 22 y 24 son los que presentan menor cantidad de inserciones y deleciones de residuos y son las de menor energía, esto puede estar asociado a que la cantidad de repeticiones necesarias para completar una superhélice es de 23 (Galpern *et al.*, 2020).

A lo largo de los años se han realizado diversos tipos de experimentos en diferentes proteínas con el objetivo de caracterizar cómo es el proceso de plegado de las proteínas repetitivas. Los mecanismos de plegado de estas proteínas generalmente comienzan con una nucleación de algunas de sus repeticiones que luego se propaga a las demás. Recientemente se ha demostrado que hay proteínas que presentan un plegado totalmente cooperativo, mientras que otras presentan un plegado no cooperativo. Esta diferencia en el plegado esta dada por la simetría en secuencia y por la energía de las interacciones entre las repeticiones, cuanto mayor es la similitud y más fuerte son las interacciones más cooperativos es el plegado (Galpern *et al.*, 2022).

1.3.1. Plegado de proteínas repetitivas

Como ya se mencionó, las repeticiones sólo se pliegan en elementos estructurales similares que se empaquetan y forman estructuras superhelicoidales extendidas, estabilizadas únicamente por interacciones dentro de las repeticiones y los vecinos adyacentes (Main *et al.*, 2003a). Por lo tanto, las proteínas repetitivas carecen de interacciones distales, lo cual puede afectar a que el plegado de estas proteínas sea cooperativo, ya que la alta cooperatividad en el plegado de las proteínas globulares se debe a los contactos que se forman entre aminoácidos distales en secuencia (Onuchic y Wolynes, 2004). Sin embargo se ha demostrado experimentalmente que el plegado es cooperativo en las proteínas repetitivas y que depende de las interacciones que forman las repeticiones con sus repeticiones vecinas más cercanas (Lowe, 2007; Mello y Barrick, 2004; Mosavi *et al.*, 2002b; Tang *et al.*, 1999). También se ha demostrado, mediante experimentos de simulaciones de recocido simulado de proteínas repetitivas naturales

de diferente largo, que a medida que aumenta el número de repeticiones, la cooperatividad tiende a romperse (Ferreiro *et al.*, 2005). Por otro lado, la existencia de interacciones de corto alcance no significa que una repetición por sí sola sea un dominio de plegado independiente, ya que necesita de la formación de contactos con las repeticiones vecinas para estabilizar su estructura (Ferreiro *et al.*, 2008).

La teoría de paisaje energético de las proteínas, nos dice que la energía de una proteína baja más cuando adquiere conformaciones más parecidas al estado nativo. En el caso de las proteínas repetitivas, debido a la similitud entre las repeticiones, da lugar a diferentes conformaciones con energías libres similares lo que genera rutas paralelas de plegado (Ferreiro *et al.*, 2008). Los experimentos de simulaciones de dinámica molecular basadas en paisajes perfectamente embudados sugieren que los "dominios" se pliegan por un mecanismo por nucleación-propagación. Una vez que tiene lugar la nucleación inicial, los módulos estructurales individuales de la proteína repetitiva se pliegan en serie de una forma altamente cooperativa que da lugar al plegado completo de los "dominios". Usando modelos de Gõ, se ha estudiado el plegado de la proteína P-16 demostrando que esta proteína muestra un plegado de dos estados y que las repeticiones localizadas en el C-Terminal son las primeras en plegarse (Tang *et al.*, 2003). Otra características de las proteínas repetitivas es que pliegan mucho más lento de lo esperado, suelen plegarse en al menos tres ordenes de magnitud más lento (Ivankov *et al.*, 2003), esto está relacionado a su orden de contactos (Mello *et al.*, 2005). La mayoría de los dominios repetitivos adquieren un estado plegado consolidado cuando interactúan con sus ligandos.

Las proteínas repetitivas también se pueden diseñar y esta capacidad de diseño se relaciona a su regularidad estructural y a la modularidad de las repeticiones, lo que permite crear proteínas con una estructura específica, ya que si se eliminan o agregan repeticiones el arreglo resultante puede plegarse en una conformación estable (Main *et al.*, 2003a). Una de las estrategias usadas para el diseño de proteínas repetitivas es el diseño por consenso, que se refiere a proteínas diseñadas utilizando información y patrones obtenidos directamente de la conservación de secuencia de las repeticiones presentes en la naturaleza. Usando esta metodología se han diseñado proteínas de diferentes familias de proteínas repetitivas, ANK (Binz *et al.*, 2003; Mosavi *et al.*, 2002a; Plückthun, 2015), TPR (Main *et al.*, 2003b), LRR (Stumpp *et al.*,

2003), HEAT(Urvoas *et al.*, 2010), ARM (Varadamsetty *et al.*, 2012). Las proteínas repetitivas se diseñan con el propósito de estudiar la relación entre la secuencia y la estructura de proteínas repetitivas naturales y para explorar cómo estas repeticiones pueden afectar la función de la proteína.

1.3.2. Familia de proteínas con repeticiones de ankirina

Las repeticiones de ankirinas (ANKs) fueron identificadas por primera vez en las proteínas Notch, Swi6 y cdc10 en *Drosophila melanogaster*. Breeden y Nasmyth detectaron que estas proteínas tenían en común la presencia de la misma repetición de 33 aminoácidos de largo y que sus estructuras tridimensionales eran similares (Breeden y Nasmyth, 1987). Las ankirinas se encuentran en los tres súper reinos, bacterias, arqueas y en eucariotas, y también se han encontrado en varios tipos de virus, siendo más abundantes en organismos eucariotas, presentes en aproximadamente el 6 % de la secuencias de proteínas (Björklund *et al.*, 2006; Mosavi *et al.*, 2004). Aún no se ha demostrado que las ankirinas tengan algún tipo de actividad enzimática, sino que su función típica es mediar la interacción proteína-proteína y se encuentran involucradas en muchos procesos metabólicos. Se han encontrado mutaciones en las proteínas repetitivas de ankirina en varias enfermedades humanas (Chagula *et al.*, 2020). La secuencia consenso de la repetición de ankirinas está defina por una serie de firmas. El motivo más prevalente es el TPLH que está localizado en la posición 4 al 7 (según la fase definida por Peng, se denomina fase al aminoácido que es considerado el que comienzan y terminan las repeticiones dado un arreglo de varias repeticiones). La prolina ubicada en la posición 5 es la responsable de la formación de la estructura en forma de L la cual es estabilizada mediante enlaces de hidrógeno de la cadena lateral de la histidina de la posición 7 (ubicada en la hélice) y la treonina en la posición 4 (ubicada en el *β-harping*). El núcleo hidrofóbico está formado por los residuos en las posiciones 6, 8, 9, 10, 17, 18, 20, 21 y 22, los cuales tiene baja exposición al solvente. A diferencia de las repeticiones ubicadas en los terminales, en los cuales estos residuos tiene una mayor exposición al solvente, son reemplazados por residuos polares (Mosavi *et al.*, 2002a).

Generalmente se las describe como un arreglo lineal de copias en tándem de un motivo de aproximadamente 33 aminoácidos de largo. Un estudio reciente ha revelado que el 80 % de

estos arreglos están compuestos de al menos 7 repeticiones, y existe una longitud de arreglo particularmente abundante que consta de 22 a 24 unidades de repetición (Galpern *et al.*, 2020). En la figura 1.6 se muestran ejemplos de diferentes estructuras de proteínas que contienen repeticiones de ankirinas. Se puede observar que el motivo ANK adopta una conformación hélice–turn–hélice–loop (Main *et al.*, 2005; Mosavi *et al.*, 2004; Sedgwick y Smerdon, 1999). Las repeticiones de ANKs tienden a apilarse juntas en forma lineal (Fig. 1.6), en dónde

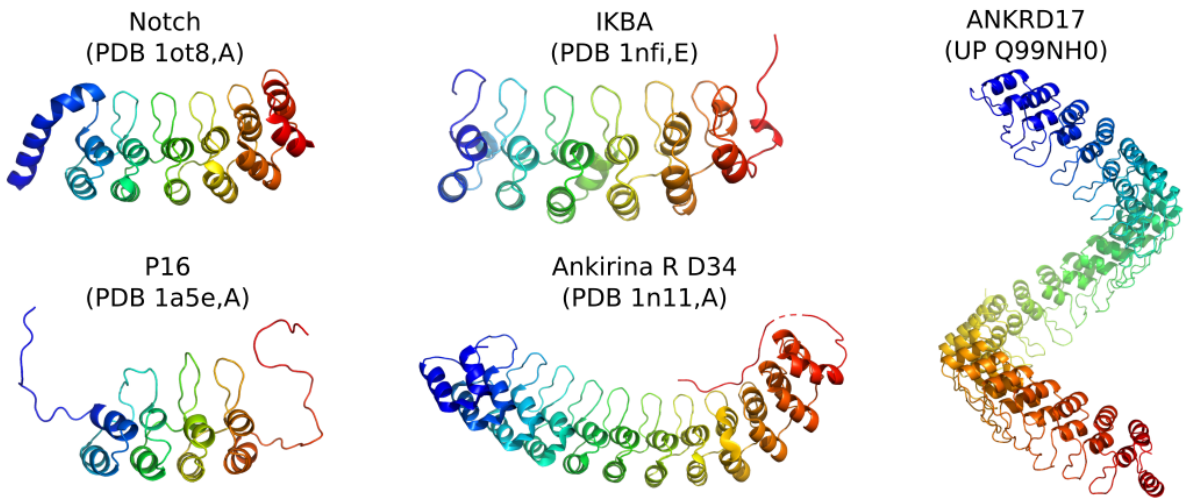

Figura 1.6: Ejemplo de estructuras de proteínas que contienen diferentes cantidades de repeticiones de Ankirina. Se uso el método de color arco iris para colorear los residuos.

predominan las interacciones hidrofóbicas entre las hélices y los enlaces de hidrógeno entre las regiones *hairpin–loop*, los aminoácidos no conservados se localizan preferentemente en la superficie (Forrer *et al.*, 2003; Mosavi *et al.*, 2002a), creando así superficies no polares que son altamente complementarias en forma y terminan ensamblándose para generar esa forma helicoidal extendida. El empaquetamiento de múltiples repeticiones tiende a orientar los motivos que tienen forma de *β-harping* uno al lado del otro formando enlaces de hidrógeno con la repeticiones vecinas. Este empaquetamiento se estabiliza aún más por las interacciones de enlaces de hidrógeno entre residuos polares conservados y átomos de la cadena principal de módulos de repetición ANK adyacentes. Todas estas interacciones locales entre repeticiones se combinan para dar lugar a una estructura no globular estable (Jacob y Harrison, 1998; Sedgwick y Smerdon, 1999). La interfase entre repeticiones consiste mayormente de

interacciones hidrofóbicas que estabilizan la hélice y de enlaces de hidrógeno que conectan al *β-harping*.

Mediante el estudio del proceso de plegado de varias proteínas naturales se ha demostrado que las ankirinas, al igual que las proteínas globulares, también presentan un plegado cooperativo. Como por ejemplo, la proteína P16 (Fig. 1.6), un supresor de tumores compuesto de 4 repeticiones de ankirinas, mediante varios ensayos experimentales (dicroísmo circular (Boice y Fairman, 1996; Lowe, 2007; Tevelev *et al.*, 1996), espectroscopia de fluorescencia y cromatografía en filtración de gel (Tang *et al.*, 1999)) se observó que tiene una transición de plegado cooperativo (Tang *et al.*, 1999).

Otra proteína muy estudiada, que también presenta una transición de plegado cooperativo entre sus repeticiones y un plegado de dos estados, es el receptor Notch (Fig. 1.6), una proteína conformada por 6 repeticiones de ankirinas bien definidas a diferencia de las dos últimas repeticiones del C-Terminal que muestra una región desordenada (Zweifel y Barrick, 2001). Los estudios *in silico* de modelos de Gõ (Ferreiro *et al.*, 2005) y estudios experimentales (Bradley y Barrick, 2002), han demostrado que esta proteína tiene dos estados de transición, en el cual el primero es más estable que el segundo, y puede estar formado por el plegado de las repeticiones 5 y 6 y la interfaz entre ellas o por el plegado de la repetición 2 y la primer hélice de la repetición 3 y la interfaz entre ellas. Para el caso de la proteína I$\kappa\beta\alpha$ (Fig. 1.6), proteína conformada por 6 repeticiones de ankirinas, de las cuales las dos repeticiones localizadas en el C-Terminal se encuentran desplegadas debido a la falta de interacciones estabilizadoras. Los estudios *in silico* del plegado de I$\kappa\beta\alpha$, usando modelos de Gõ, muestran que tiene dos temperatura de plegado, a diferencia de las proteínas anteriormente descriptas que solo tienen una temperatura de plegado (Ferreiro *et al.*, 2005).

Debido a la conservación en la estructura de las repeticiones de ankirinas y al plegado cooperativo, el efecto de mutaciones puntuales puede ser muy significativo, característica que hace atractivo el estudio de estas proteínas. El efecto de esas mutaciones puede impactar directamente en el proceso de plegado así como también puede cambiar la afinidad de unión de las proteínas con las que forma interacciones. Por ejemplo, se han identificado mediante mutaciones puntuales, dos residuos de I$\kappa\beta\alpha$ (Y254 y T257) que al mutarlos por sus contrapartes del consenso provocan un aumento significativo en su capacidad de plegado (Truhlar *et al.*,

2008). Sin embargo, el aumento de la capacidad de plegado producto de esas mutaciones no es favorable porque se degrada más lentamente. Además no solamente se ve afectada la afinidad de unión con NF$\kappa\beta\alpha$ ya que se ve disminuida, sino que también afecta al proceso de despojamiento molecular, mediante el cual se libera al factor NF-$\kappa\beta$ (Potoyan *et al.*, 2016).

1.3.3. Evolución de las proteínas con repeticiones de ankirinas

Las ankirinas no solo son atractivas porque se encuentran en los tres súper reinos o por su conservación estructural o su capacidad de mediar interacciones proteína-proteína, sino también porque aún no esta totalmente claro cómo es que estas proteínas evolucionan. La secuencia de aminoácidos del dominio ankirinas y la cantidad de repeticiones es variable entre ortólogos, aunque los residuos específicos que son críticos para la estructura se conservan (Javadi y Itzhaki, 2013). Es decir, que a lo largo de la evolución las ankirinas han conservado su estructura terciaria característica, pero han modificado algunos residuos impactando en la especificidad y afinidad con las proteínas con las que forman interacciones (Islam *et al.*, 2018). Los residuos implicados en las interacciones proteína-proteína, son 7 residuos no conservados ubicados en las posiciones 3, 5, 6, 9, 17, 18 y 30 de las repeticiones (Parra *et al.*, 2015).

Con respecto al mecanismo evolutivo involucrado en la generación de nuevos genes con diferentes cantidad de repeticiones en las diferentes proteínas de la familia de ankirinas, hay muchas hipótesis y estudios que tratan de entender esta problemática pero aún no esta totalmente resuelta (Björklund *et al.*, 2006). Se consideran dos posibles escenarios, que las ankirinas tienen un origen muy antiguo, antes de la evolución de las eucariotas o que evolucionaron independientemente en los diferentes linajes por evolución convergente que seleccionó positivamente el diseño modular de algunas de estas proteínas (Al-Khodor *et al.*, 2010). La hipótesis más estudiada en cuanto a la expansión en el número de repeticiones de ankirinas es que ocurre a través de la duplicación génica de los dominios repetitivos (Andrade *et al.*, 2001; Björklund *et al.*, 2006; Pâques *et al.*, 1998; Schüler y Bornberg-Bauer, 2016). Además, las tasas de inserción o deleción de dominios repetitivos es muy alta, es decir, que estas proteínas pueden modificar su estructura y estabilidad, así como también su función muy rápidamente (Schüler y Bornberg-Bauer, 2016). Por ejemplo, en las ankirinas identificadas

en el arroz (OsANK), en la que se han encontrado nueve eventos de duplicación segmentaria (Huang *et al.*, 2009).

1.4. Objetivos

Este trabajo de tesis tiene como objetivos, desarrollar métodos computacionales y técnicas avanzadas para el estudio del plegado, la función y evolución de familias de proteínas. Se desarrollaran e implementarán algoritmos basados en la "Teoría del paisaje energético de las proteínas", que serán aplicados para el estudios de los patrones energético de las regiones *fuzzy* y para analizar la conservación de la frustración en proteínas evolutivamente relacionadas. Se recopilará y analizará de manera integral los datos genómicos existentes relacionados con proteínas de la familia de ankirinas, con el objetivo de identificar patrones y clasificar estas proteínas según sus características estructurales y funcionales. Se implementarán sistemas de simulación computacional de reacciones de plegado de proteínas, permitiendo el estudio detallado de sus propiedades termodinámicas y cinéticas. Al alcanzar estos objetivos, se espera contribuir al avance del conocimiento en el campo de la proteómica y la genómica, generando nuevas herramientas y enfoques para el análisis del plegado, la función y la evolución de proteínas.

Capítulo 2

Métodos

2.1. Frustración energética local

La teoría de paisajes energéticos nos dice que la forma del paisaje energético de una proteína natural es de un embudo corrugado en dónde existe un fuerte sesgo energético hacia el estado nativo. Además, nos dice que las proteínas se pliegan minimizando sus conflictos energéticos, lo cual se conoce como "Principio de mínima frustración" (Bryngelson y Wolynes, 1987). Sin embargo este principio no implica que no puedan existir ciertos conflictos remanentes en la estructura proteica. En los últimos años se ha demostrado que la presencia de estos conflictos en la estructura están asociados a varios aspectos funcionales (Ferreiro *et al.*, 2007, 2014; Freiberger *et al.*, 2019; Gianni *et al.*, 2014).

En el 2007, Ferreiro y colaboradores (Ferreiro *et al.*, 2007), desarrollaron un algoritmo, basado en el paisaje energético de las proteínas, para localizar y cuantificar frustración en estructuras proteicas. Posteriormente, dicho algoritmo fue implementado de forma general como un servicio web, disponible para el público y fue llamado *Protein Frustratometer* (Parra *et al.*, 2016). Actualmente se pueden calcular los patrones de frustración para cualquier estructura proteica en formato PDB, ya sea obtenida por cristalización o por algún método de modelado de proteínas, accediendo a la dirección web http://www.frustratometer.tk/.

Para localizar frustración el *Frustratometer* lo hace de dos maneras diferentes, una es a nivel de residuo y la otra a nivel de contacto. Para calcular la frustración a nivel de contacto, primero define los residuos que están en contacto en la estructura, para ello mide la distancia

euclídea (DE) entre C_β, excepto para la glicina que utiliza el Cα debido que carece de C_β. Según la distancia que hay entre los carbonos, se definen tres tipos de contactos, los de corto alcance son los que la $DE_{C_\beta} < 6{,}5$Å, los de largo alcance son los que 6,5Å$< DE_{C_\beta} < 9{,}5$Åy por último los mediados por agua que están definidos igual que los de largo alcance pero donde la exposición al solvente (SASA) tiene que ser mayor a 0,05 (SAS $> 0{,}05$) y para terminar de definir el contacto se tienen que cumplir la condición de que la distancia en secuencia tiene que ser mayor a 2 aminoácidos. Una vez definido los pares de aminoácidos en contacto en la estructura proteica, se generan 2000 señuelos (*decoys*) en los que se modifican algunos parámetros de la interacción nativa. Luego se mide la energía de la interacción nativa y de los señuelos usando el potencial de AMW (*Associative Memory Hamiltonian optimized with Water-mediated interactions* (Papoian *et al.*, 2004)) y el índice de frustración (FI) es calculado mediante un Z-*score* definido como $Fi = \frac{E_N - <E_D>}{\sigma D}$ donde E_N es la energía nativa, $< E_D >$ es la media de la distribución de la energía de los señuelos y σD es su desvío estándar.

Existen 2 variantes para el índice de frustración a nivel de contacto y difieren en la forma en que se crean los señuelos.

Mutational: Dados dos residuos en contacto, los señuelos se crean cambiando de forma aleatoria la identidad de los mismos y conservando los demás parámetros en sus valores nativos.

Configurational: dados dos residuos en contactos, los señuelos se crean cambiando aleatoriamente no solo la identidad de los aminoácidos, sino también se modifica la accesibilidad al solvente y la distancia de la interacción.

Por último la frustración también se puede calcular a **nivel de residuo único** (*single residue level*), en este caso no se definen residuos en contacto en la estructura sino que dado un residuo i, solamente se cambia la identidad del mismo conservando los demás parámetros en sus valores nativos.

El índice de frustración (FI) se puede clasificar en tres tipos lo cual depende del valor del mismo. Para la frustración a nivel de contacto los grupos se definen como, si el FI $\geq 0{,}78$ se clasifica como mínimamente frustrado, si -1 $<$ FI $< 0{,}78$ se clasifica como neutro y si FI $<$ -1 el contacto es altamente frustrado. Para la frustración a nivel de residuos estos valores cambian, para los mínimamente frustrados el FI $\geq 0{,}55$, para los neutros -1 $<$ FI $< 0{,}55$ y para los altamente frustrados FI $<$ -1.

2.1.1. Función de distribución radial

Esta función es ampliamente utilizada en estadística para calcular la probabilidad de encontrar una partícula r desde una partícula de referencia en un sistema de partículas, es decir, describe la variación de la densidad como función de la distancia medida desde una partícula de referencia (Fig. 2.1A). Para calcular la $g(r)$ se utiliza la siguiente fórmula 2.1:

$$g(r) = 4\pi r^2 \rho dr \tag{2.1}$$

donde ρ es la densidad numérica y está definida como ρ = N/V, donde N es el número total de objetos en un volumen V. Para poder utilizar la función de distribución radial

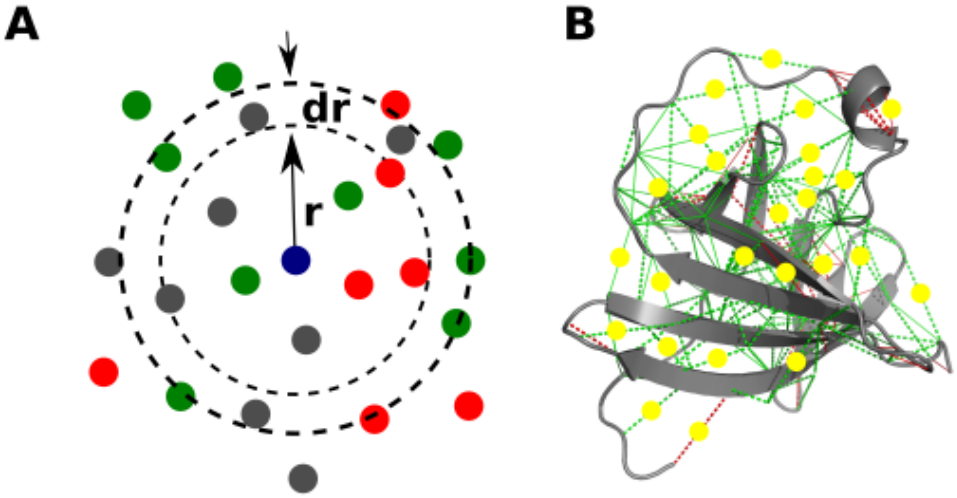

Figura 2.1: A) Ejemplo de un sistema de partículas, el círculo de color azul representa la partícula de referencia, en círculos de línea punteada negra la distancia $r+dr$, la rojo se representan los contactos altamente frustrados, en verde los mínimamente frustrados y en gris los neutros. B) Ejemplo del frustratograma de una proteína (PdbID: 1a1x), El esqueleto de la proteína se representa en *cartoon* en color gris, los contactos mínimamente frustrados se representan con líneas verdes, las interacciones altamente frustradas se representan con líneas rojas. Las interacciones neutras se omitieron para mejor visualización y en círculos amarillos se muestran algunas partículas virtuales generadas entre los residuos en contacto.

se generaron partículas virtuales (VP) entre los residuos en contacto (Fig. 2.1B), cada VP contiene información sobre su disposición tridimensional (coordenadas en los ejes x, y y z) y

el valor del índice de frustración del contacto. Para generar los sistemas de partículas para cada proteína se utilizaron *scripts* en lenguaje Python2,7.

2.2. Simulaciones de dinámica molecular de grano grueso

Las proteínas y otras biomoléculas pueden ser simuladas *in silico*, es decir utilizando métodos computacionales, que pueden ser por ejemplo campos de fuerza a nivel atomístico o a nivel de grano grueso. En el marco de esta tesis nos centraremos en simulaciones del tipo de grano grueso.

Para las simulaciones de dinámica molecular se utilizó el campo de fuerza de *AWSEM-MD*. El cual es un modelo de memoria asociativa, mediada por agua, estructura y energía (*AWSEM*). Es un campo de fuerza basado en estructuras locales, es decir, que contiene un término de sesgo de estructura local basado en bioinformática, que tiene en cuenta muchos fragmentos de secuencias cortas de diferentes estructuras que están modulados por su secuencia local. Para predecir la estructura de una proteína *de novo*, primero se definen una colección de fragmentos cortos (denominados memorias) que van a ser parte del sesgo de estructura local. A modo de definir los fragmentos se generan alineamientos de secuencia locales y globales de la secuencia de la proteína en estudio con la secuencia de todas las proteínas de la base de datos de PDB (*Protein Data Bank*). Las proteínas resultantes de estos alineamientos, se denominan homólogos, en caso de que no se detecte ningún homólogo, la simulación se correr en modo *single-memory* (memoria única) y solamente se utiliza como memoria los fragmentos de la proteína de entrada de la simulación.

2.2.1. Modelo

Como se comentó *AWSEM* es un campo de grano grueso, a cada uno de los aminoácidos que componen la proteína se los representa usando el modelo estructural de tres cuentas por residuo (C_α,C_β,O) excepto para la glicina que carece de C_β.

Además utiliza un Hamiltoniano, definido por los siguientes potenciales de la ec.2.2:

$$V_{total} = V_{backbone} + V_{contact} + V_{burial} + V_{helical} + V_{FM} \tag{2.2}$$

donde, el potencial de contactos ($V_{contact}$), depende del tipo de aminoácido y del término de interacción terciaria, actúa en pares de residuos que están a más de 10 residuos de distancia en la secuencia. El termino potencial de entierro (V_{burial}) al igual que el $V_{contact}$ son obtenidos por optimización que maximiza la relación entre la temperatura de plegado y la transición vítrea (T_f/T_g). El potencial de hélice $V_{helical}$ es el término explícito de los puentes de hidrógeno, la fuerza de la interacción depende de la propensión helicoidal de los 2 residuos que participan en la interacción. El potencial V_{FM} es el término bioinformático que hace uso de toda la información experimental de PDB, está compuesto de la suma de todos los fragmentos alineados y la suma de todos los posibles pares de átomos C_α y C_β contenidos en los fragmentos y separados por más de 2 residuos. Por último, el $V_{backbone}$ es el potencial responsable de la geometría del esqueleto de la proteína (ec. **??**), este potencial esta definido por el potencial V_{C_α}, el cual garantiza la conectividad de la cadena mediante una serie de enlaces armónicos.

2.2.2. Coordenada Q_w

A lo largo de las simulaciones la conformación de la proteína que se está analizando va a sufrir cambios, esto se debe a la propia dinámica de la proteína. Para poder captar esos cambios en las trayectorias de dinámica molecular se utilizó la coordenada Q_w (ec. 2.3), el valor Q_w se utiliza para evaluar el grado de plegado y la similitud estructural de las conformaciones de proteínas simuladas con sus estados nativos.

$$Q_w = \frac{1}{N_p} \sum_i \sum_j exp \frac{-(r_{i,j} - r^u_{i,j})^2}{2\sigma^2_{i,j}} \tag{2.3}$$

Las posiciones de los aminoácidos i y j en la cadena polipeptídica se representan por los subíndices de i y j. N_p es el número total de pares (i,j). La distancia entre los C_α de los residuos i y j se representa con $r^u_{i,j}$ es la misma distancia pero medida en la estructura nativa

de referencia. La constante σ_{ij} esta definida como $(1 + |i - j|)^{0.15}$ que es la anchura de la separación en secuencia de los residuos. Los valores que puede tomar Q_w son entre 0 y 1. Un valor de Q_w cercano a 1 indica una conformación mas cercana al estado nativo y un valor de Q_w cercano a cero es una conformación desplegada.

2.2.3. Predicción de la estructura de las proteínas

Para todas las simulaciones de dinámica molecular que se corrieron en esta tesis se utilizó el paquete de *AWSEM-MD*.

Las estructuras de las proteínas que se analizaron se descargaron de la base de datos de PDB. Los correspondientes PdbIDs son, para la proteína de ankirinas diseñada con cuatro repeticiones idénticas del consenso se usó el PdbID: 1N0R. Para I$\kappa\beta\alpha$ PdbID: 1NFI (para generar el archivo que solo contiene 4 repeticiones se eliminaron manualmente las dos últimas repeticiones), para P16 se usó PdbID: 1A5E y por último para la proteína Notch se utilizó el PdbID: 1OT8.

También se usaron 3 modelos por homología que se obtuvieron utilizando el software *Modeller* (Webb y Sali, 2017) que genera modelos basados en una proteína que se denomina molde (*template*). Fueron modelados utilizando como molde la proteína 1N0R, el primer modelo contiene solamente 3 repeticiones idénticas del consenso, el segundo contiene 5 y el tercer modelo contiene 6.

Se modificaron, del código fuente del *AWSEM*, algunos de los valores por defecto de los parámetros definidos por el campo de fuerza para adaptarlos a los sistemas de proteínas que fueron evaluados. Los parámetros modificados fueron: **Energía del fragmento (defecto:20):** se utilizó 16, **Tamaño del fragmento (defecto: 9 aminoácidos):** se usó 12, El valor de temperatura a la cual se comenzó la simulación fue 1200 y el valor final 200, las variaciones de temperatura se realizaron a lo largo de 16 millones de pasos. A modo de estimar la T_f, que es la temperatura a la cual se observa una transición entre el estado plegado y desplegado, es decir es el punto medio de la pendiente de transición entre ambos estados, se analizaron los valores de Q_w en función de la temperatura.

Luego de estimar la T_f, se realizaron simulaciones a esta temperatura y a otras temperaturas cercanas a la T_f, usando el método de muestreo de *Umbrella Sampling*. *Umbrella Sampling*

es un método computacional, que tiene como objetivo la exploración de las estructuras cercanas a un valor de Q_w determinado. Los valores de Q_w que se usaron fueron entre 0 y 1, con intervalos de 0,05, es decir se que se muestrearon 21 valores de Q_w. Para cada intervalo de Q_w se simuló el sistema durante 1.000.000 de pasos. Las 20 trayectorias obtenidas para los diferentes sesgos de *umbrella* son integradas mediante el método *WHAM* (Weighted Histogram Analysis Method), que es un método computacional que se utiliza para analizar y calcular propiedades termodinámicas de sistemas físicos, particularmente en sistemas en equilibrio, como líquidos, sólidos o sistemas biológicos (Kumar *et al.*, 1992), obteniéndose de esta forma los valores de Energía Libre proyectados en diferentes coordenadas como Q_w y el Radio de Giro (Rg). Se basa en la utilización de histogramas ponderados, en simulaciones de dinámica molecular, se generan histogramas que representan la probabilidad de encontrar el sistema en ciertos estados. El *WHAM* utiliza estos histogramas, pero los pondera para dar más importancia a ciertas regiones del espacio de fase, corrigiendo así las fluctuaciones estadísticas y los sesgos. Además es capaz de manejar múltiples ventanas de simulación con diferentes condiciones termodinámicas (por ejemplo, diferentes temperaturas o presiones) y combina los datos de todas las ventanas para calcular el perfil de energía libre final.

2.3. Construcción de la base de datos de Ankirinas

Al trabajar con grandes volúmenes de datos, cualquiera sea su tipo, siempre es necesario la construcción de una base de datos que almacene toda la información contenida en los datos a analizar. El modelo de base de datos que mejor se ajusta a nuestros datos son la de entidad-relación (ER).

Modelo Entidad-Relación o ER: este tipo de modelos de bases de dato describe a los datos como entidades, relaciones y atributos. Para poder entender como es la estructura de este modelo hay que comprender los aspectos básicos del mismo. **Entidades:** es el objeto básico del modelo ER, que puede ser un objeto con existencia física (por ejemplo, un producto, una persona, etc.) o puede ser un objeto con una existencia conceptual (por ejemplo, una materia de la facultad, un cargo laboral, etc.). **Atributos:** son parte de la entidad y cada entidad tiene su atributo, es decir, son propiedades particulares que la describen. Por ejemplo, la

entidad ALUMNO se puede describir por, su Nº de legajo, su nombres y apellidos, la carrera que cursa y sus notas. Los valores que contienen los atributos que describen a cada entidad son los datos almacenados en la base de datos.

Los atributos de las base de datos ER pueden ser de varios tipos simple o compuesto, mono valor o multivalor, y almacenado o derivado. Los atributos se pueden utilizar para identificar a cada entidad de forma unívoca la cual se denomina **clave o llave primaria**. Una llave primaria puede estar compuesta por uno o más atributos de la entidad y tiene que cumplir con el requisito de que identifica de forma única a cada fila de una tabla. También pueden ser **clave o llave foránea** la cual identifica una columna o grupo de columnas en una tabla que se refiere a una columna o grupo de columnas en otra tabla.

Y por último para que sea un modelo ER, tienen que existir **relaciones entre las entidades**, es decir, que dos entidades pueden estar asociadas definiendo una relación mediante algunos de sus atributos. Existen 3 tipos cardinalidades en una relación lo cual indica el número de entidades con las que puede estar relacionada una entidad dada. **Una a una:** una entidad de A está asociada únicamente con una entidad de B y una entidad de B está asociada solo con una entidad de A. **Una a muchas:** una entidad en A está asociada con varias entidades de B, pero una entidad de B puede asociarse únicamente con una entidad de A. **Muchas a Muchas:** una entidad en A está asociada con varias entidades de B y una entidad en B está vinculada con varias entidades de A.

2.3.1. Base de datos de ankirinas

Siguiendo los pasos descriptos anteriormente, diseñamos y construimos una base de datos ER que contiene datos de secuencias de proteínas y genómicos, así como también todas las anotaciones disponibles en las bases de datos de UniProt KB, GenBank y EMBL, para todas las proteínas con repeticiones de ankirinas detectadas usando el método descripto en (Galpern *et al.*, 2020). En la figura 2.2 se puede ver el diagrama de Entidad-Relación de nuestra base de datos. La base de datos contiene 8 tablas en las cuales están almacenadas, las secuencias de proteínas y genómicas, anotaciones sobre los exones e introness, organismos, IDs correspondientes a las bases de datos biológicas de donde fueron descargados los datos, entre otras.

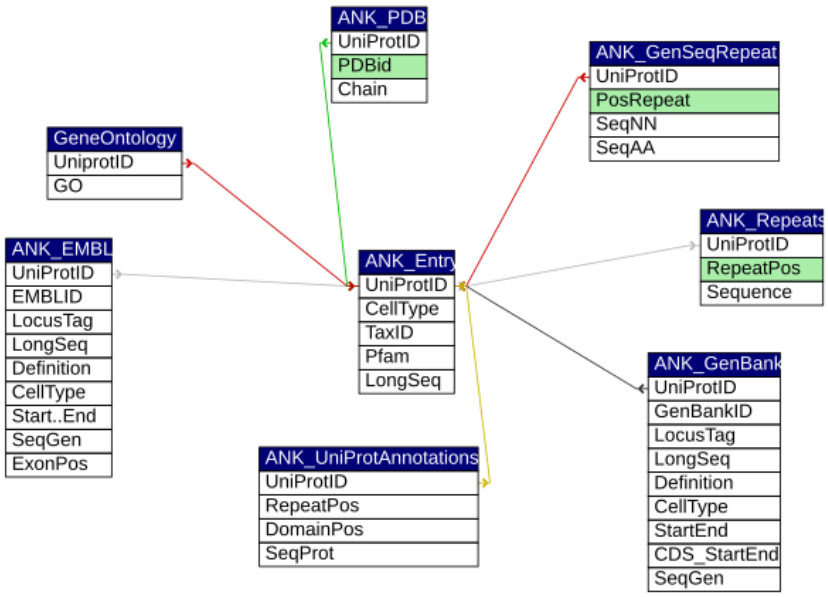

Figura 2.2: Diagrama de Entidad-Relación de nuestra base de datos de ankirinas.

Descripción general de la base de datos

Se descargaron de la base de datos de UniProt KB (Consortium, 2019) las secuencias de proteínas, anotaciones como, el tipo celular (Procariota, Eucariota, Virus y archea), el TaxID (identificación del organismo según su taxa), Pfam (ID de la familia de proteína) y el largo de la secuencia, estos atributos pertenecen a la tabla **ANK Entry**. Se descargaron de las bases de datos de EMBL (Kanz *et al.*, 2005) y GenBank (Sayers *et al.*, 2019) las secuencias de genes, anotaciones de la localización de los exones e intrones, localización del gen en el cromosomas (*Locus Tag*), una breve definición, el ID del gen, el largo de la secuencia codificante y se almacenaron en las tablas **ANK EMBL** y **ANK GenBank**. Se generó una tabla, llamada **ANK Repeats**, con las anotaciones de las posiciones de comienzo y de fin de cada repetición. Y por último con las anotaciones sobre las posiciones de comienzo y fin de cada repetición, generamos otra tabla, llamada **ANK GenSeqRepeat**, que contiene la secuencia de proteínas de cada repetición y su correspondiente secuencia génica. Todas las tablas se relacionan a través del atributo llamado UniProtID que es la identificación de cada proteína almacenada en la base de datos de UniProt. Una vez diseñado el diagrama de ER de la base de dato se la construyó usando un motor de MySQL. Mediante el uso de stripts en Python 2,7 se descargaron de las bases de datos biológicas previamente descriptas, todos los datos que

luego fueron almacenados en las tablas correspondientes. La base de datos está almacenada en https://drive.google.com/file/d/1IKmHeKlcv_bJO297rhg2TWUgCwI4FGUe/view?usp=sharing.

Capítulo 3

Diseño, desarrollo, implementación y testeo de herramientas bioinformáticas para el análisis de patrones energéticos de proteínas

3.1. FrustratometeR: implementación del *Protein Frustratometer* como un paquete de R

En el marco de esta tesis, hemos implementado la herramienta del *Protein Frustratometer* como un paquete R, al cual llamamos FrustratometeR. Uno de los motivos principales del porqué se implementó como un paquete de R es debido a que la última versión *stand alone* del algoritmo está programada en Perl, un lenguaje de programación que está casi obsoleto y además la implementación como un paquete de R es más fácil de instalar, de usar y de agregar nuevas funcionalidades. FrustratometeR no solo calcula la frustración energética local en estructuras proteicas, sino que además ofrece la posibilidad de analizar la frustración en trayectorias de de simulaciones de dinámica molecular y evaluar el efecto de las variantes de aminoácidos. En las figuras 3.1 y 3.2, se muestra un resumen las funcionalidades del FrustratometeR y el código mínimo necesario para generar los gráficos de la figura

3.2C. Se puede instalar y correr en una computadora personal, en supercomputadoras o clusters, la implementación del paquete de R está disponible en un repositorio de GitHub: https://github.com/proteinphysiologylab/frustratometeR. En las siguientes secciones se detallaran las funciones principales del FrustratometeR.

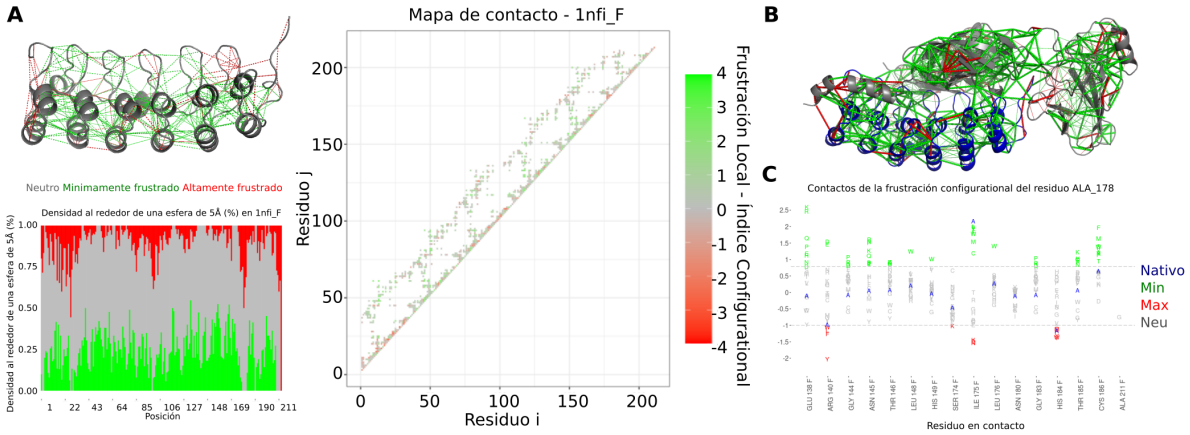

Figura 3.1: Frustración en IκBα (PDB ID 1NFI, F). (A) Mapa de contactos, gráfico 5Adens y representación pymol. (B) Frustración de IκBα en complejo con NfκB (cadenas A, B, F). (C) Cambios de frustración al mutar un residuo específico a todas las alternativas canónicas de aminoácidos. Eje X: se muestran todos los residuos que interaccionan con el residuo de interés para todos los mutantes. Eje Y: se muestran los valores de frustración y se colorean en función de su clasificación de frustración. La variante nativa aparece en azul. Cada mutante está representado por su código de aminoácido de 1 letra para identificar a qué variante corresponde.

En las siguientes secciones y capítulos de esta tesis, se utilizarán todas las nuevas funcionalidades del FrustratometeR, en diferentes proteínas y familias de proteínas, para realizar diferentes tipos de análisis de la frustración local.

3.1.1. Funciones del FrustratometeR

Calculo de la frustración

En la figura 3.3 se muestra como es el flujo de trabajo de la función *calculate frustration()* la cual calcula la frustración de una estructura proteica y también se muestran las funciones para generar los gráficos necesarios para visualizar los resultados. La función *calculate_frustration()*

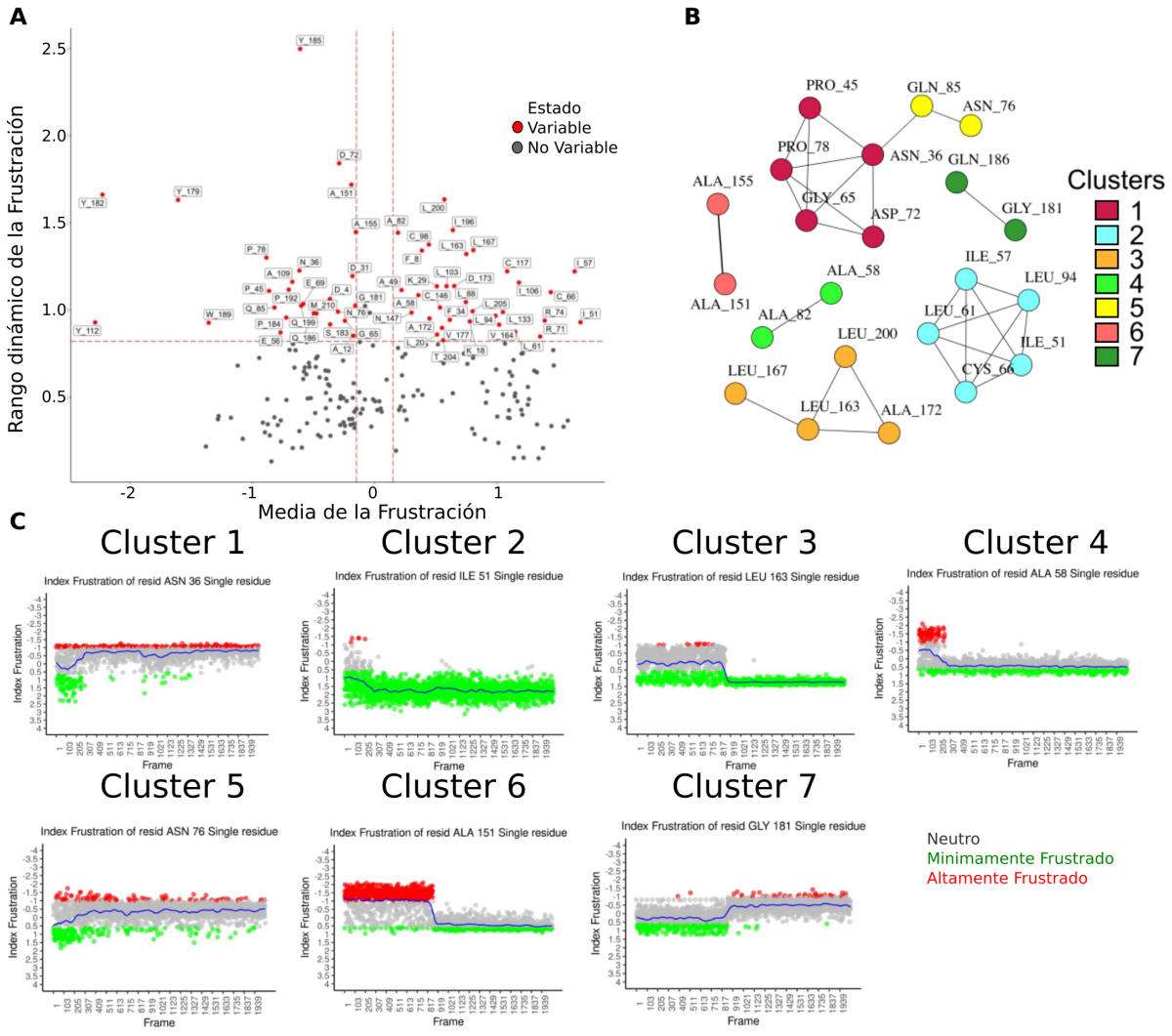

Figura 3.2: Frustración en IκBα (PdbID 1NFI, F). (A) Se identifican los residuos que varían su frustración a lo largo de los fotogramas en simulaciones de dinámica molecular basándose en sus valores de frustración media y rango dinámico. B) Los residuos variables se conectan entre sí en una red de correlación que luego se agrupa para encontrar módulos con un comportamiento dinámico similar. C) Valores de frustración en función del tiempo (fotogramas de simulación) para residuos representativos en los *Clusters* 1-7. Cada uno de los *clusters* contiene residuos, que a lo largo de una trayectoria de dinámica molecular, en cada fotograma, tienen una frustración similar (Ver Anexo para visualizar en detalle cada uno de los *clusters*). Por ejemplo, todos los residuos que conforman el *cluster* 1, comienzan con un índice de frustración neutro y luego, a lo largo de la simulación, su índice de frustración cambia a mínimamente frustrado. Dado que la entrada procede de una simulación, los residuos se renumeran consecutivamente desde 1 y no como en el PDB original.

(Fig. 3.3) es la que calcula la frustración local a partir de un PdbFile o PdbID de una estructura proteica, creando un "Objeto de Frustración del Pdb" que se utiliza para generar

visualizaciones, obtener datos de frustración y otros procesos cómo analizar mutaciones puntuales. El PdbFile puede ser descargado de PDB (Protein Data Bank) o también pude ser un modelo de proteína (obtenido por *AlphaFold2* (Jumper *et al.*, 2020), Meta AI (Lin *et al.*, 2023) o cualquier otra herramienta de predicción de estructura de proteínas). Usando esta función se puede calcular la frustración en los tres modos (*configurational, mutational o singleresidue*, ver métodos), la generación de los gráficos, en este caso es opcional, a diferencia de la versión *Protein Frustratometer* del servicio web que son generados automáticamente. Los cálculos y pasos para calcular la frustración a partir de un Pdb se encuentran detallados en la sección de Métodos.

El código mínimo necesario para calcular la frustración es:

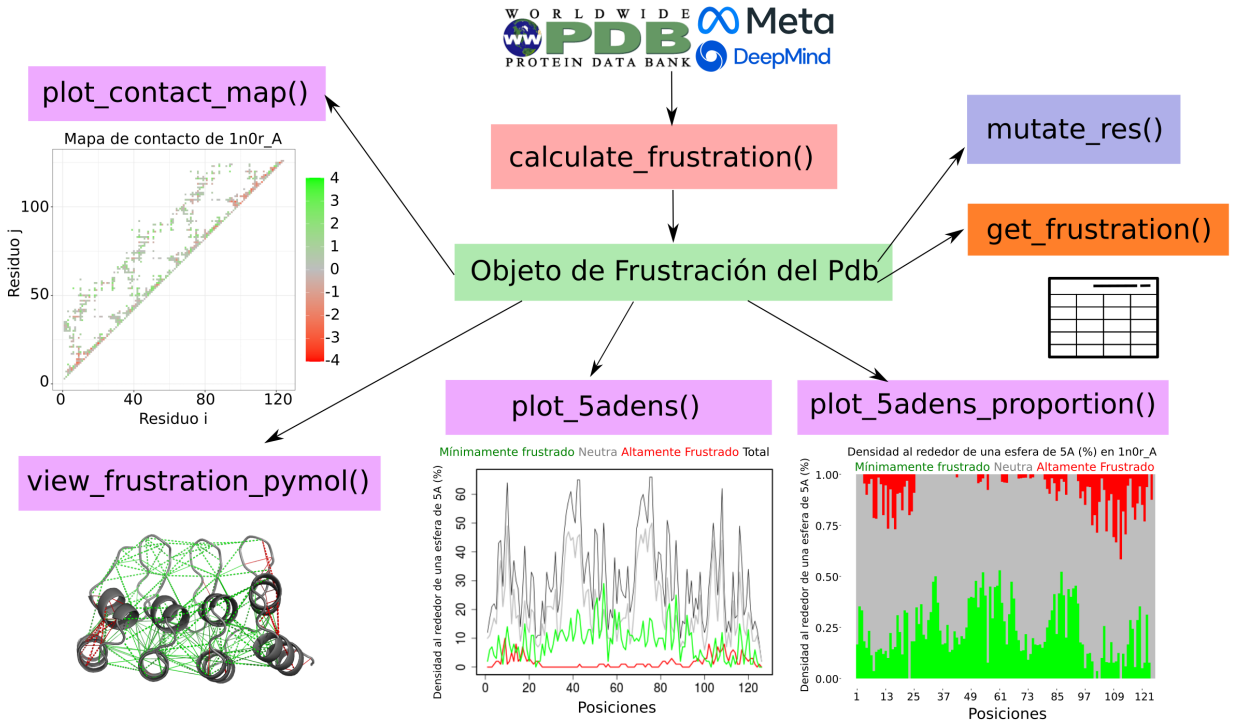

Figura 3.3: Flujo de trabajo de la función *calculate_frustration()* que se utiliza para calcular la frustración local. En verde los objetos de frustración, en rosa procesos principales, en celeste otros procesos, en naranja obtención de datos y en magenta visualizaciones.

```
library(frustratometeR)
calculate_frustration(PdbFile = PdbFile, Mode = Mode, Chain = Chain,
```

```
Electrostatics_K = Electrostatics_K, SeqDist = SeqDist,
ResultsDir = ResultsDir, Graphics = Graphics, Visualization = Visualization)
```

Los parámetros usados son, *PdbFile* es el PdbID, *Mode* es el modo de frustración por defecto se calcula el configurational, *Chain* es la cadena si este parámetro se omite la frustración se calcula para toda la estructura, *Electrostatics_K* su valor por defecto es *NULL* pero si se utiliza un valor numérico se creará un nuevo término en el hamiltoniano, que tendrá en cuenta las interacciones electrostáticas para calcular la frustración. *SeqDist* separación en secuencia utilizada para calcular las densidades locales de los aminoácidos por defecto es 12, *ResultsDir* es el directorio en donde se almacenarán los cálculos y gráficos, *Graphics* los valores posibles son *True* o *False*, en caso de usar *True* se generarán todos los gráficos, en caso contrario se usa *False* y por último *Visualization* al igual que *Graphics* los valores son *True* o *False*, en caso de usar *True* se generarán todos los *scripts* de visualizaciones.

Medición de la frustración de mutaciones puntuales

El diagrama de la figura 3.4 muestra que a partir del objeto Frustración del Pdb obtenido al ejecutar *calculate_frustration()*, explicado en la sección anterior, se pueden analizar las 20 posibles variantes de aminoácidos en una determinada posición. Para calcular la medición de la frustración de mutaciones puntuales, primero se calcula la frustración de la proteína nativa en uno de los tres modos de frustración. Luego, usando la función *mutate_res()* (Fig. 3.4), utilizando el índice de frustración especificado previamente en *calculate_frustration()* se genera un modelo de la proteína para cada una de las 19 mutantes posibles y vuelve a calcular la frustración con *calculate_frustration()*, pero ahora de la proteína mutada. Se implementan dos modos para generar los mutantes: threading (no modifica las coordenadas del *backbone*) y *Modeller* (realiza una optimización energética generando un modelo de homología con *Modeller* (Webb y Sali, 2017)). Los resultados obtenidos pueden ser visualizados en gráficos como que se muestran en la figura 3.4.

El código mínimo necesario para calcular la predicción de la frustración para el índice *configurational* para una mutación puntual es:

```
library(frustratometeR)
Pdb_conf <- calculate_frustration(PdbID = PdbID, Chain = Chain,
```

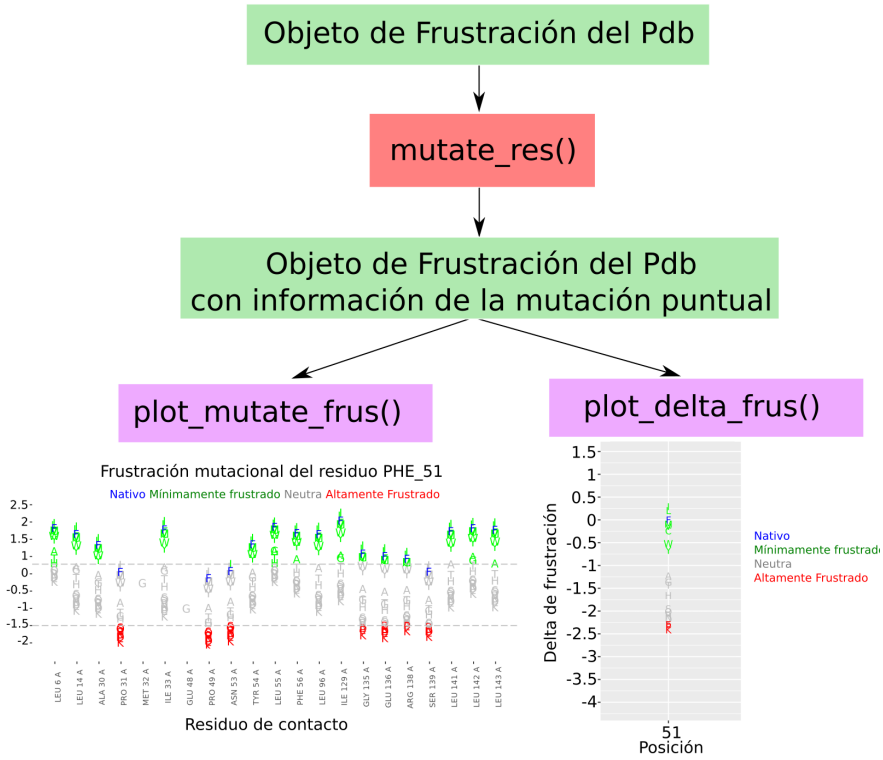

Figura 3.4: Flujo de trabajo de la función *mutate_res()* que se utiliza para calcular la predicción de la frustración de mutaciones puntuales. En verde los objetos de frustración, en rosa procesos principales y en magenta visualizaciones.

```
Mode = "configurational")
Pdb_conf <- mutate_res(Pdb = Pdb_conf, Resno = NoRes, Chain = Chain)
```

Los parámetros para la función *calculate_frustration()* ya fueron detallados en la sección anterior. Para la función *mutate_res()* *Pdb* es el calculo de la frustración obtenida con *calculate_frustration()* y almacenada en *Pdb_conf*, *Resno* es el número del aminoácido para el cuál se quiere calcular la predicción de la frustración y *Chain* es la cadena en donde está el aminoácido de interés.

Análisis de la frustración local en simulaciones de dinámica molecular

El diagrama de la figura 3.5A muestra cómo, a partir de un directorio que contiene las estructuras de cada fotograma de una simulación de dinámica molecular (MD), la frustración local puede ser analizada con la función *dynamic_frustration()*. Para ello, primero se genera un objeto de frustración de la simulación que puede utilizarse para obtener visualizaciones y

otros procesamientos. Para crear este objeto de frustración, la función *dynamic_frustration()*, calcula la frustración para cada uno de los fotogramas en el directorio, usando la función *calculate_frustration()*. Si el modo de frustración seleccionado es *configurational* o *mutational*, además genera un archivo en formato GIF juntando todos de los mapas de contacto de cada fotograma. En este GIF se puede ver no solo como varían los contactos sino también como cambia la frustración de los contactos a lo largo de la simulación.

El código mínimo necesario para calcular la frustración de una simulación de dinámica molecular para el índice *configurational* es:

```
library(frustratometeR)
OrderList <- c("pdb1,pdb", "pdb2,pdb", .... ,"pdbn.pdb")
Dynamic_conf <- dynamic_frustration(PdbsDir = PdbsDir, OrderList = OrderList,
ResultsDir = ResultsDir)
```

Los parámetros PdbsDir y ResultsDir ya fueron explicados, *OrderList* es la lista secuencial de los fotogramas de una MD.

El diagrama de la figura 3.5B muestra uno de los posibles procesos de un objeto de frustración de la simulación obtenido con *dynamic_frustration()*, la función *dynamic_res()* se utiliza para analizar la frustración de un determinado residuo a lo largo de una MD, esto devuelve un objeto de frustración de la dinámica modificado, añadiendo los datos resultantes del proceso. Esta función se puede utilizar para visualizar específicamente como cambia la frustración de un solo residuo a lo largo de una MD. Si el modo de frustración calculado es *configurational* o *mutational* utilizando la función *plot_dynamic_res_5adens_proportions()*, mediante un gráfico se puede observar como cambia la frustración alrededor de de una esfera de 5Åde un residuo en particular a lo largo de una MD. Por otro lado si el modo de frustración calculado es *singleresidue* utilizando la función *plot_res_dynamics()* se puede observar como cambia la frustración a nivel de residuo único de un residuo en particular a lo largo de una MD.

Otro de los procesos que pueden aplicarse al objeto de frustración de la simulación obtenido con *dynamic_frustration()*, es *detect_dynamic_clusters()* que encuentra grupos de residuos con dinámicas de frustración similares. La función *detect_dynamic_clusters()* modifica el objeto de frustración de la simulación para que pueda ser utilizado para obtener la información de los *clusters* y múltiples visualizaciones. Los residuos que se van a agrupar en *clusters*

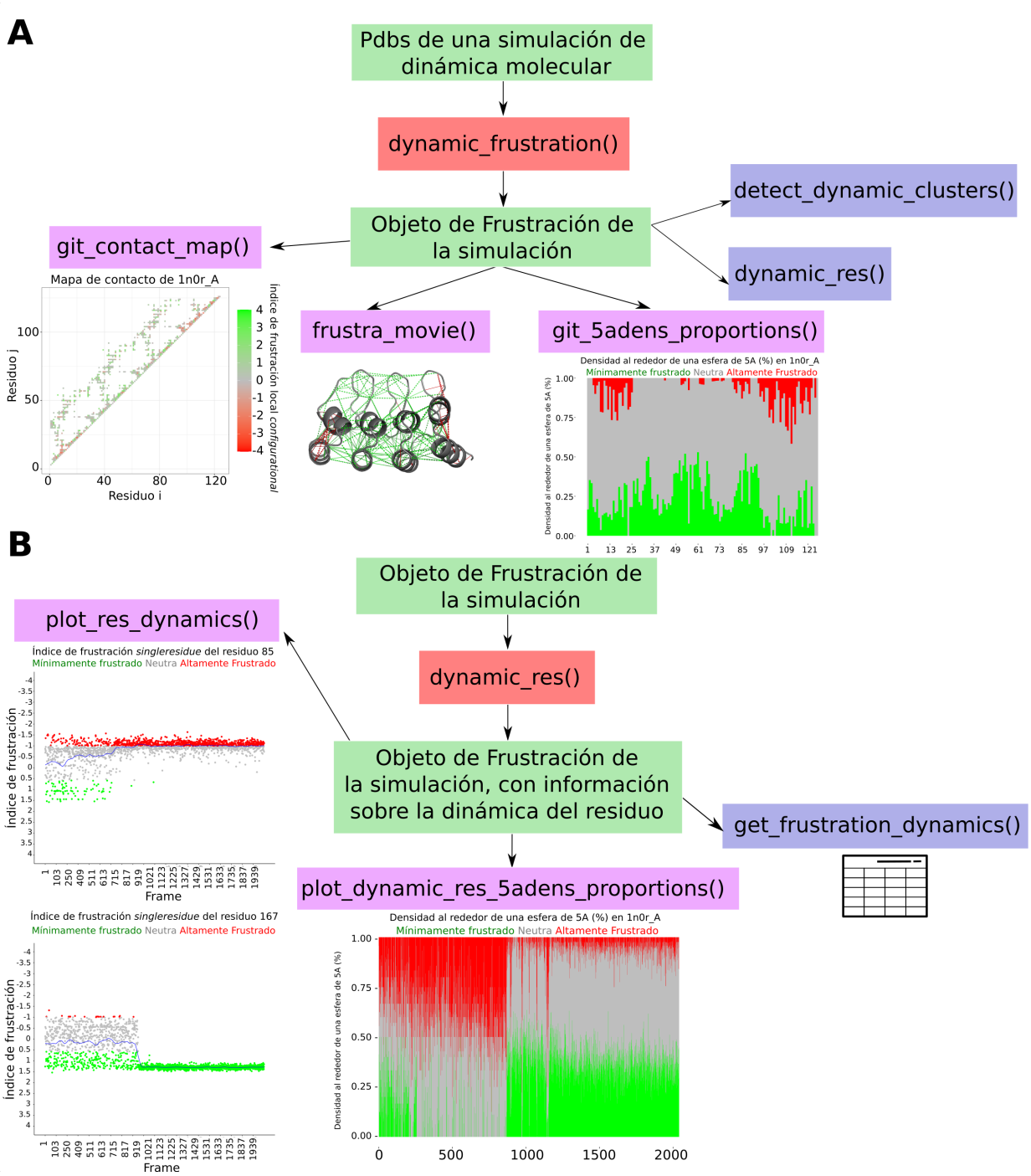

Figura 3.5: Flujo de trabajo de las funciones, A) *dynamic_frustration()* y *dynamic_res()* que se utilizan para analizar la frustración de simulaciones de dinámica molecular. En verde los objetos de frustración, en rosa procesos principales, en celeste otros procesos y en magenta visualizaciones.

se definen por residuos que cumplen con dos condiciones, primero que el valor de la media de frustración del residuo supere el umbral definido en *FiltMean* (por defecto es 0.15) y

que el rango dinámico sea mayor al umbral definido en *MinFrstRange* (por defecto 0.7), para mayor detalle ver el código necesario para calcular la frustración de una simulación de dinámica molecular. Para todos los residuos que superen estos dos filtros se calcula un análisis de componentes principales (PCA), el parámetro *Ncp* se utiliza para definir cuántos componentes se conservan (valor por defecto *Ncp*=10). Los valores de correlación se calculan (utilizando el test de correlación de *spearman* por defecto o también se puede seleccionar *pearson*) en el espacio PCA entre todos los pares de residuos. El parámetro *MinCorr* define qué coeficientes de correlación se mantendrán para definir que dos residuos están conectados en la red. La matriz de adyacencia resultante se utiliza para generar un grafo no dirigido y el método de agrupación de *Leiden* (Traag *et al.*, 2019) para detectar los *clusters*. El parámetro *LeidenResol* define la resolución a la que se detectarán los *clusters* (valor por defecto 1.00). Dicho en otras palabras los *clusters* son residuos que presentan cambios correlacionados en su frustración energética local y variantes a lo largo de una trayectoria de dinámica molecular. En la figura 3.6 se muestra el flujo de trabajo de la función.

El código necesario para calcular la frustración de una simulación de dinámica molecular para el índice *singleresidue* es:

```
library(frustratometeR)
OrderList <- c("pdb1,pdb", "pdb2,pdb", .... ,"pdbn.pdb")
Dynamic_sing <- dynamic_frustration(PdbsDir = PdbsDir, OrderList = OrderList,
ResultsDir = ResultsDir, Mode = "singleresidue")
Dynamic_sing <- detect_dynamic_clusters(Dynamic = Dynamic_sing, CorrType = "spearman",
FiltMean = 0.15, MinFrstRange = 0.7, MinCorr = 0.95)
```

Los parámetros utilizados en esta función ya fueron explicados.

3.2. FrustraEvo: Una herramienta para estudiar los patrones energéticos en familias de proteínas

FrustraEvo es una herramienta que calcula la conservación de la frustración local, es decir, calcula la conservación de los patrones energéticos de la frustración local para proteínas re-

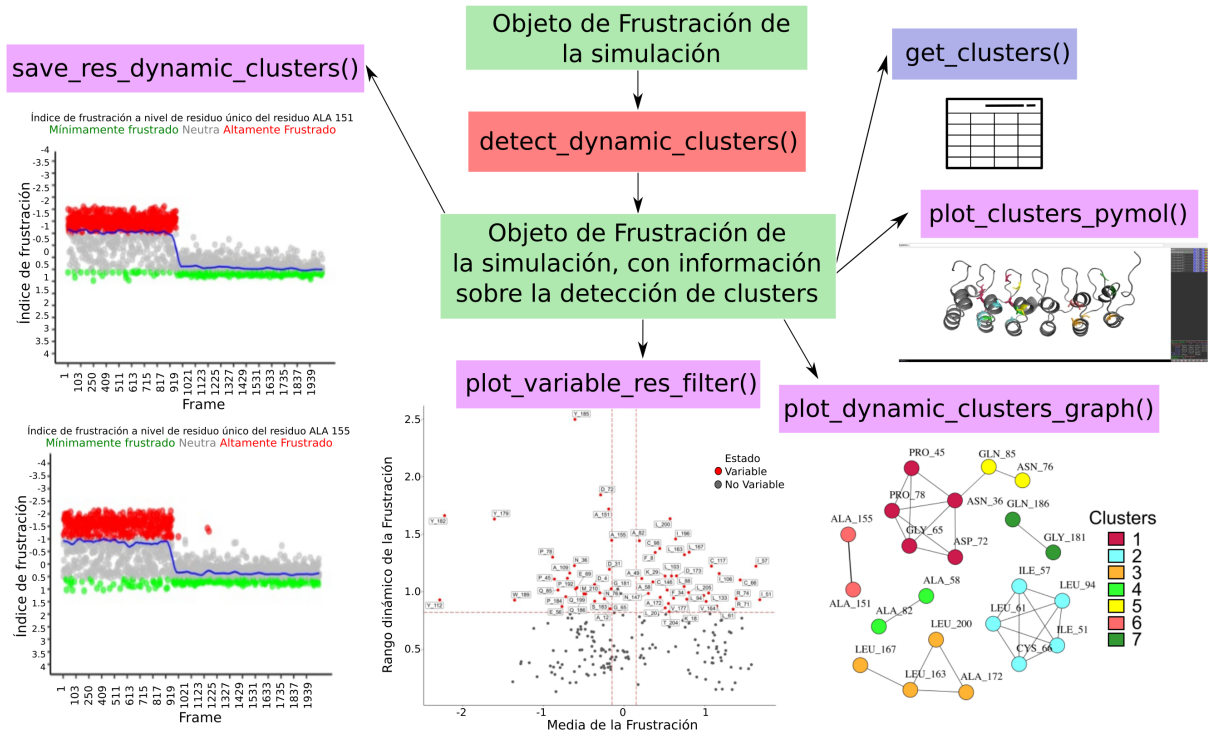

Figura 3.6: Flujo de trabajo de la función *detect_dynamic_clusters()* que se utiliza para analizar la frustración de simulaciones de dinámica molecular. En verde los objetos de frustración, en rosa procesos principales, en celeste otros procesos y en magenta visualizaciones.

lacionadas evolutivamente. FrustraEvo puede utilizarse para calcular la frustración a nivel de residuos único (*singleresidue*) o a nivel de contacto (*mutational* o *configurational*). Como archivos de entrada FrustraEvo recibe un conjunto de secuencias alineadas (MSA) y sus correspondientes estructuras. Los patrones de frustración local para todas las estructuras proteicas individuales son calculados por FrustratometeR. En la figura 3.7C se muestra el flujo de trabajo del FrustraEvo. El *pipeline* recibe los archivos de entrada (mencionados anteriormente), el identificador de cada secuencia de proteína dentro del MSA se debe corresponder con el nombre del archivo de la estructura proteica (sin la extensión .pdb). Las secuencias dentro del MSA deben coincidir exactamente con las contenidas en los archivos PDB.

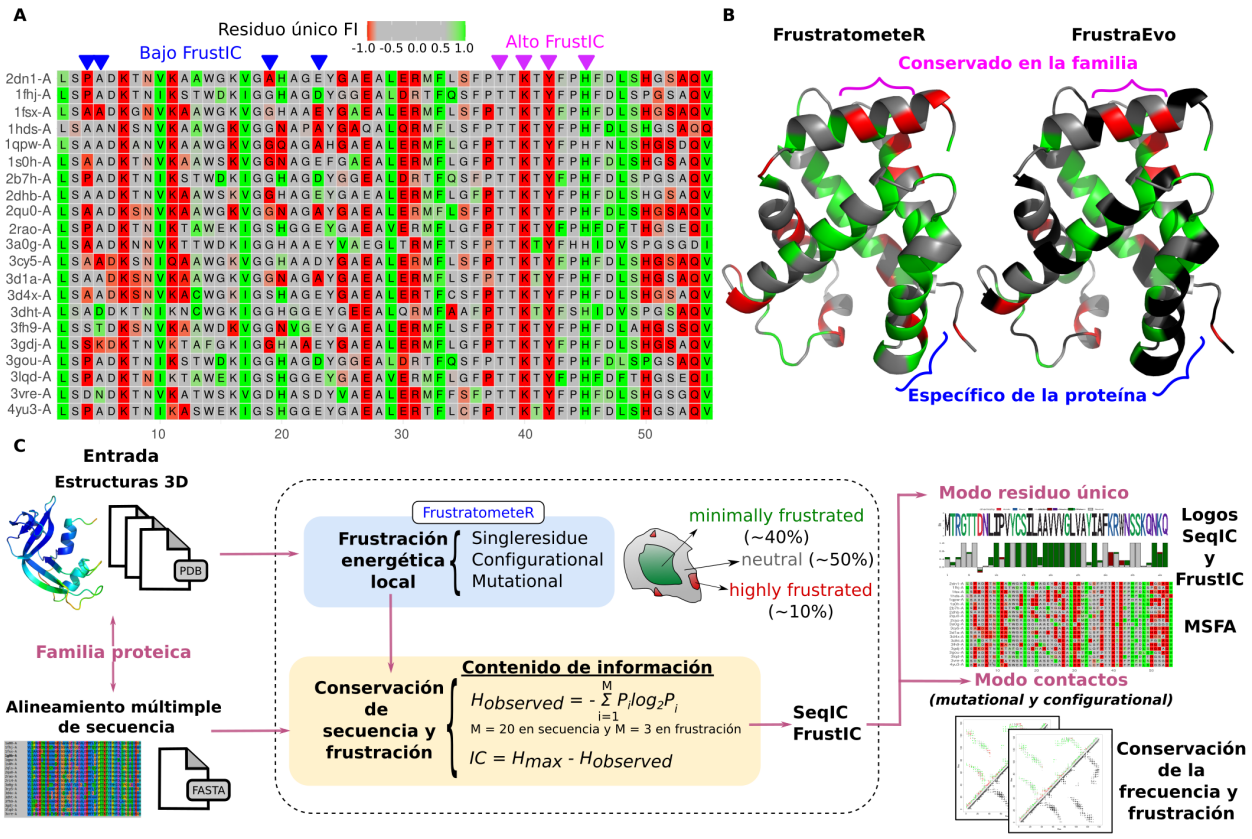

Figura 3.7: Análisis de la frustración local en familias de proteínas. A) Alineamiento múltiple de secuencia y coloreado según la clasificación del índice de frustración a nivel de residuo (MSFA), es decir, consiste en el SRFI (*singleresidue frustation index*) calculado a partir de estructuras proteicas individuales mapeadas en el MSA. Los asteriscos en magenta marcan los residuos conservados en frustración (alto FrustIC) y los azules los residuos no conservados en frustración (bajo FrustIC). Los residuos mínimamente frustrados están coloreados en verde, los neutros en gris y los altamente frustrados en rojo. B) Comparación entre los valores SRFI calculados por FrustratometeR (izquierda) y la conservación de los estados de frustración basados en sus valores FrustIC calculados por FrustraEvo (derecha) visualizados en la misma estructura (globina α humana, PDB 2DN1, cadena A). En la representación del FrustratometeR los residuos están coloreados según sus estados de frustración. En la representación del FrustraEvo, los residuos con FrstIC>0,5 están coloreados según su estado de frustración más informativo, mientras que los residuos con FrstIC$\leq$0,5 están coloreados en negro. C) Resumen del flujo de trabajo de FrustraEvo para analizar una única familia de proteínas. Las figuras son a modo ilustrativo, en secciones más adelante se visualizarán con mayor detalle y definición.

3.2.1. Cálculos de la conservación de secuencia y frustración

Los valores de la conservación de secuencia y frustración, SeqIC y FrustIC respectivamente, se calculan utilizando conceptos de la teoría de la información. SeqIC se calcula a partir de los

residuos alineados en el MSA, basándose en la distribución de las identidades de aminoácidos. FrustIC se calcula para los residuos alineados en un MSA o contactos equivalentes entre proteínas en el MSA, basándose en los estados de frustración asignados a los residuos a partir de las estructuras. Debido a las inserciones o deleciones que pueden estar presentes en diferentes proteínas de una misma familia, se selecciona una proteína de referencia para definir sobre qué residuos o contactos se calculan los cálculos de la conservación. También es importante definir una proteína de referencia porque todos los archivos de salida van a estar numerados usando la numeración de la proteína de referencia. Los cálculos se realizan sólo sobre los residuos presentes en la proteína de referencia o para los contactos entre residuos que están presentes en la estructura de referencia. La estructura de referencia puede ser definida por el usuario o el *pipeline* va a seleccionar la proteína con mayor cobertura en el MSA.

Conservación de secuencia: Se calcula a partir del MSA utilizando el paquete R ggseqlogo (Wagih, 2017). Corresponde a los valores SeqIC y se calcula usando la entropía de secuencia correspondiente al contenido de información basado en la entropía de Shannon y se calcula mediante las ecuaciones 3.2 y 3.1. La fórmula (ec. 3.2) calcula la suma de la multiplicación de cada probabilidad P_i por el logaritmo en base 2 de esa misma probabilidad $log_2 P_i$ para todos los eventos i en la distribución. El resultado es la entropía, que representa la cantidad promedio de información requerida para describir la distribución de probabilidad. Cuanto mayor sea la entropía, mayor será la incertidumbre o la diferencia en la distribución.

$$\mathrm{H}(p_i \ ... \ p_n \ ) = -\sum_{i=1}^{n} P_i log_2 P_i \tag{3.1}$$

$$\mathrm{R} = H_{max} - \mathrm{H}(p_i \ ... \ p_n) \tag{3.2}$$

Donde $\mathrm{H}(p_i \ ... \ p_n \ )$ es la entropía de la distribución de probabilidad con n eventos, P_i es la probabilidad de que el sistema este en el estado i y n es el número de valores posibles, los valores de n son los 20 posibles aminoácidos, es decir, que el valor de H_{max} es $log_2(20)$.

FrustraEvo modo *singleresidue*: El MSA se procesa de forma que sólo se conservan las columnas que tienen un aminoácido (sin posiciones vacías) en la estructura de referencia, de lo contrario se eliminan. El contenido de información de la frustración (FrustIC) basado en la distribución de estados de frustración para cada columna del MSA se calcula utilizando

las fórmulas de contenido de información de Shannon (ecuaciones 3.3 y 3.4).

$$H_{observed} = -\sum_{i=1}^{n} P_i log_2 P_i \tag{3.3}$$

$$\mathrm{R} = H_{max} - H_{observed} \tag{3.4}$$

Donde P_i es la probabilidad de que el sistema este en el estado i. Para la frustración los posibles estados son 3, mínimamente frustrado, altamente frustrado y neutro. Por lo tanto, el contenido de información puede calcularse como en la ecuación 3.4. Debido a que en frustración la probabilidad de que se produzcan estados no es la misma para cada clase, debe utilizarse una distribución de probabilidad de fondo de los estados para estimar H_{max}. Hemos utilizado la distribución de estados descrita por Ferreiro et al. (Ferreiro *et al.*, 2007) para calcular H_{max} en los cálculos de FrustIC.

Además genera un alineamiento múltiple de secuencia en el que cada aminoácido está coloreado por su índice de frustración *singleresidue* (MSFA), se utiliza para comparar los valores generados por el FrustratometeR con los del FrustraEvo.

FrustraEvo modo contacto: De forma similar al modo de *singleresidue*, se toma una estructura de referencia para definir los contactos a evaluar. Teniendo en cuenta que se eliminaron del MSA (MSARef) todas las columnas que no tienen aminoácidos en la estructura de referencia, FrustraEvo calcula la frecuencia de un contacto entre las columnas i, j en el MSARef, para cada estructura en el conjunto de datos. Donde i,j $\in$[1, N], siendo N el número de columnas en el MSARef. Posteriormente, usando las ecuaciones 3.3 y 3.4, FrustraEvo calculará, para cada posible contacto, de acuerdo con los pares de columnas dentro del MSARef, las contribuciones de contenido de información de cada estado de frustración. El FrustIC de un contacto dado se calculará como la suma de las contribuciones individuales de cada estado de frustración.

Las frecuencias de fondo para los estados mínimamente, neutro y altamente frustrado se definen como, mínimamente frustrado 0,4, altamente frustrado 0,1 y neutro 0,5, en correspondencia con las frecuencias observadas por Ferreiro et al. (Ferreiro *et al.*, 2007) para los índices *configurational* y *mutational*.

Implementación del algoritmo

FrustraEvo está implementado en Python3 y en R v.4, El código fuente del FrustraEvo está depositado en un repositorio de GitHub (proteinphysiologylab/FrustraEvo), también está disponible como una imagen en Docker (proteinphysiologylab/frustraevo) y como un servicio web en el siguiente enlace https://frustraevo.qb.fcen.uba.ar/.
Los archivos de entrada son un MSA y una estructura en formato PDB por cada secuencia en el MSA. Los archivos de salida están distribuidos en 3 directorios, *AuxFiles*, que contiene archivos intermediarios que se van generando a medida que se va ejecutando el *pipeline* (MSAs, archivos .log, tabla de posiciones). Otro directorio es *Data* que contiene, para cada proteína, todos los cálculos de frustración, sesiones de pymol para visualizar mapeado en las estructuras la conservación de la frustración por residuo. Y por último *OutPutFiles*, que contiene los logos de frustración y de secuencia, los mapas de contactos, el alineamiento múltiple de secuencias basado en frustración (MSFA) y los datos crudos sobre el calculo del contenido de información para cada índice de frustración (*mutational*, *configurational* y *singleresidue*). Los gráficos de los logos de frustración, de secuencia y los mapas de contactos se elaboran con el paquete R ggplot2.

3.2.2. Caso de estudio: Hemoglobina

Aunque descienden de un ancestro común, las secuencias pertenecientes a un grupo de proteínas de una misma familia presentan una cantidad variable de diferencias acumuladas a lo largo de su historia evolutiva. Muchos de estos cambios a nivel de secuencia de los miembros de una familia de proteínas, han sido moldeados por diferentes requisitos funcionales que, en algunos casos también pueden tener un impacto a nivel estructural. En esta sección se analizará cómo los cambios a nivel de secuencia de proteínas de una misma súperfamilia repercuten a nivel estructural analizando sus patrones de frustración local. Para ello estudiamos una subparte del árbol evolutivo de las globinas (Fig. 3.8A), una de las superfamilias proteicas mejor caracterizadas (Hardison, 2012). La hemoglobina es una proteína responsable del transporte de oxígeno y dióxido de carbono a través de la sangre de mamíferos y otros vertebrados con mandíbulas, es una molécula tetrámerica compuesta por dos polipéptidos

α-globina y dos β-globina, cada uno con un grupo hemo asociado (Fig. 3.8B). Dado que las subunidades α-globina y β-globina están relacionadas evolutivamente, el estudio de un conjunto de hemoglobinas de diferentes especies implica el análisis de la ortología (comparación de diferentes α-globinas entre especies) y de la paralogía (comparación de α-globina y β-globina dentro de la misma especie) al mismo tiempo.

Para realizar este análisis hemos descargado un conjunto de datos no redundante de 21

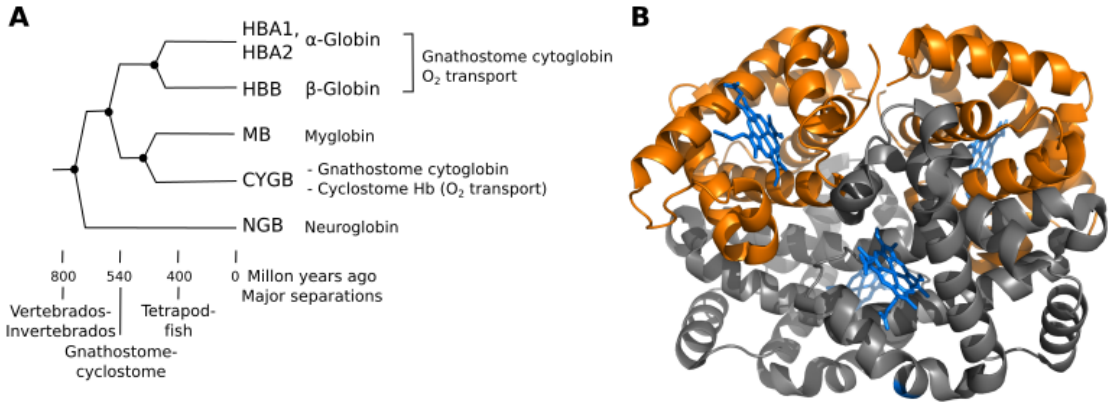

Figura 3.8: A) Modelo de evolución de los genes globina de los vertebrados. Los tiempos deducidos de duplicación y divergencia se muestran a lo largo del eje horizontal, y los genes globina humanos contemporáneos se muestran a la derecha. Los principales eventos de duplicación en la evolución de los genes de la globina se señalan a lo largo del árbol, y el momento de origen de algunos de los principales grupos de animales se indica a lo largo del eje horizontal. Para simplificar, sólo se muestra el árbol filogénetico (figura obtenida de (Hardison, 2012)). B) Estructura tetramérica de la hemoglobina de conejo (PdbID: 2rao), en color gris se representan las subunidades β y en naranja las subunidades α.

hemoglobinas de mamíferos (tabla 3.1). Seleccionamos mamíferos para asegurarnos que las estructuras sean tetraméricas y que la función sea conservada. Hemos dividido al conjunto de datos en dos, uno que contiene solamente estructuras α-globina y el otro solo estructuras β-globina y hemos calculado sus patrones de frustración utilizando FrustratometeR (Rausch *et al.*, 2021). Todas las estructuras fueron obtenida de forma experimental y descargadas del *Protein Data Bank* (PDB).

PdbID	Organismo	Especie	Holo/Apo state
1fsx	*Bos taurus* (Bovino)	Mamífero	Deoxy
3d4x	*Felis catus (Felis silvestris catus)* Gato	Mamífero	Deoxy
2dn1	*Homo sapiens* (Humano)	Mamífero	Deoxy
3vre	*Mammuthus primigenius (Siberian woolly mammoth)*	Mamífero	Deoxy
3d1a	*Capra hircus* (Cabra)	Mamífero	Methemoglobin
2dhb	*Equus caballus* (Caballo)	Mamífero	Sin información
3fh9	*Pteropus giganteus* (Zorro volador indio)	Mamífero	Sin información
3gou	*Canis lupus familiaris* (Perro)	Mamífero	Oxy
3cy5	*Bubalus bubalis* (Búfalo de agua doméstico)	Mamífero	Oxy
3gdj	*Camelus dromedarius* (Camello)	Mamífero	Oxy
3a0g	*Cavia porcellus* (Cobayo)	Mamífero	Oxy
2b7h	*Cerdocyon thous*	Mamífero	Oxy
1fhj	*Chrysocyon brachyurus* (Lobo de crin)	Mamífero	Oxy
1s0h	*Equus asinus (Equus africanus asinus)* (Burro)	Mamífero	Oxy
4yu3	*Helogale parvula*	Mamífero	Oxy
3lqd	*Lepus europaeus* (Liebre europea)	Mamífero	Oxy
1hds	*Odocoileus virginianus*	Mamífero	Oxy
2rao	*Oryctolagus cuniculus* (Conejo)	Mamífero	Oxy
2qu0	*Ovis aries* (Oveja)	Mamífero	Oxy
3dht	*Rattus norvegicus* (Rata)	Mamífero	Oxy
1qpw	*Sus scrofa* (Cerdo)	Mamífero	Oxy

Tabla 3.1: Conjunto de datos de diferentes especies de 21 hemoglobinas de mamíferos.

Resultados

Con la finalidad de investigar el vínculo entre la divergencia en secuencias y la divergencia en la frustración energética local decidimos usar el FrustraEvo en el conjunto de datos de las 21 estructuras de hemoglobina no redundantes de mamíferos. Para ello obtuvimos las secuencias de las subunidades α y β, resultando un total de 42 secuencias (21 secuencias de α y 21 de β). Las 42 secuencias fueron alineadas usando *t-coffee*, a este alineamiento múltiple de secuencias (MSA) lo llamamos $\alpha\beta$MSA. Luego dividimos el set de datos en dos y volvimos a alinear usando *t-coffee*. Uno solo contiene subunidades α al que llamamos αMSA y el otro solo contienen subunidades β y lo llamamos βMSA. Por lo tanto como resultado tenemos 3 MSAs, $\alpha\beta$MSA, αMSA y βMSA.

En la figura 3.9, se muestran los logos de secuencia y de frustración para cada uno de los conjuntos de datos. Los colores del logo de secuencia discriminan la identidad diferentes aminoácidos y los colores del logo de frustración representan la clasificación del índice de frustración, en rojo se representan los altamente frustrados, en verde los mínimamente frustrados y en gris los neutros. Arbitrariamente se definió que un residuo conserva su estado de frustración cuando el valor de contenido de información de frustración (FrustIC) es mayor a 0,5. En el $\alpha\beta$MSA (3.9A), se observa que el nivel de frustración se conserva mayoritariamente en posiciones mínimamente frustradas (n=35, $FrustIC_{medio}$=1,02) y en posiciones neutras (n=34,$FrustIC_{medio}$=0,85)y solamente 3 posiciones están muy frustradas ($FrustIC_{medio}$=0,72). Algunas posiciones en el $\alpha\beta$MSA muestran cambios en la identidad del aminoácido que dan lugar a señales de conservación de frustración diferenciales cuando los dos linajes se analizan por separado. Un ejemplo de ello es la posición 39 en el $\alpha\beta$MSA, que se corresponde con una Lys altamente frustrada en el linaje α (Lys40α) (Fig. 3.9B), pero para el linaje β se corresponde con una Glu cuyo estado es neutro (Glu39β) (Fig. 3.9C). Esto sugiere que esta posición corresponde a una adaptación funcional que se produjo después de la divergencia de las dos familias y que existen más restricciones funcionales asociadas a esta posición en la familia α en comparación con la misma posición en la familia β. Este residuo representa un ejemplo de una *Specificity Determining Positions* (SPD), que son posiciones que se conservan de forma diferencial en distintas subfamilias (Rausell *et al.*, 2010). En este caso, vemos que el análisis de frustración evolutiva proporciona una explicación funcional

basada en los valores de FrustIC en cada familia.

En total, hay 12 posiciones altamente frustradas en las α globinas ($FrustIC_{medio}$=0,87, fi-

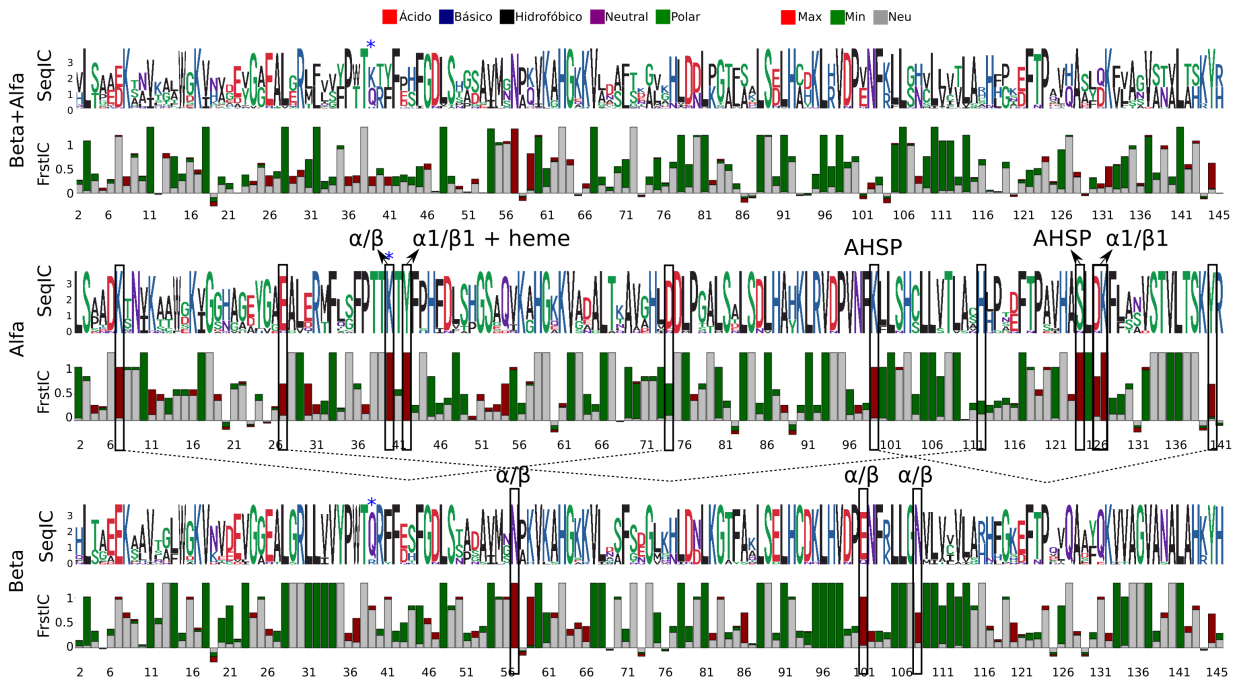

Figura 3.9: Resultados de FrustraEvo basados en el *Singleresidue Frustration Index* (SRFI) para, A) α/β-globinas B) solo α y C) solo β. Los rectángulos denotan posiciones funcionalmente relevantes explicadas con más detalle en la Tabla 8.1. En asteriscos azules marcamos la posición 39 en el MSA α/β, que corresponde a una Lys40α altamente frustrada pero a una Glu39β neutra. Para una mejor visualización del MSFA se recortaron ambos logos.

gura 3.9B) y 8 en las β globinas ($FrustIC_{medio}$=0,88, 3.9C). La mayoría de estas posiciones están frustradas sólo en una de las dos familias. Se ha descrito que estos loci están implicados en interacciones proteína-proteína dentro de la estructura tetramérica de la hemoglobina o se dan entre las α-globinas y la proteína estabilizadora de αHb (AHSP), una chaperona que previene la toxicidad de la α-globina cuando está aislada (Bisconte *et al.*, 2015; Mollan *et al.*, 2010). Varias otras posiciones altamente frustradas y conservadas desde el punto de vista de la frustración (FrustIC>0,5) en las α-globinas corresponden a residuos que forman parte de los puentes salinos intracadena (Shaanan, 1983), que son críticos para el alosterismo y el efecto Bohr, como explica Perutz (Perutz, 1970) (Fig. 3.9B), Tabla del Anexo 8.1).

En la figura 3.10, se muestran los alineamiento múltiple de secuencia y frustración (MSFA) para las subunidades α (Fig. 3.10A) y β (Fig. 3.10B). El MSFA es un archivo de salida

del FrustraEvo que se utiliza para comparar los resultados de FrustratrometeR a través de múltiples secuencias de proteínas. Cada celda está coloreada según su SRFI en las estructuras correspondientes, es decir, los residuos mínimamente frustrados están coloreados en tonos verdes, los neutros en gris y los altamente frustrados en rojo. Podemos ver, marcado en círculos azules (fig 3.10), que hay posiciones que conservan su estado de frustración en todas las especies (alta conservación) y en algunos casos también la identidad del aminoácido es la misma para todas las especies. Al analizar aquellas posiciones en las cuales se conserva la frustración pero el aminoácido en algunas especies cambia, como por ejemplo, para las α-globinas las posiciones 10, 17, 29, y 48 y para las β-globinas las posiciones 33, 54, 101 y 112. Esto podría estar indicando que estas posiciones tengan que cumplir con un requerimiento de frustración, es decir, que las mutaciones de cambio de aminoácido que se produzcan en esas posiciones deben mantener su estado de frustración.

3.2.3. Caso de estudio: Factor de elongación bacteriano, RfaH

El factor de elongación bacteriano RfaH, presente en *E. coli*, promueve la expresión de genes distantes en operones largos que codifican factores de virulencia, mediante la unión específica a ARN polimerasas (RNAP) en pausa y un elemento de ADN denominado *ops* a través de su dominio N-terminal (NTD) (Artsimovitch y Landick, 2002). La unión espuria de RfaH a RNAP en ausencia de *ops* se ve impedida por su dominio C-terminal (CTD), que se une al NTD como una α-*harping* (αCTD) y constituye un estado autoinhibido (Burmann *et al.*, 2012). La activación se produce tras el reclutamiento de RfaH a RNAP por el ADN *ops* expuesto en su superficie (Zuber *et al.*, 2018). Tras el reclutamiento, el αCTD se disocia del NTD y sufre una reorganización estructural en un β-barril (βCTD) que se une al ribosoma en la transcripción y traducción acopladas, mientras que la estructura del NTD permanece inalterada(Burmann *et al.*, 2012).

Tanto los experimentos como las simulaciones de dinámica molecular han establecido que el cambio de conformacional de RfaH es un proceso reversible (Zuber *et al.*, 2019), en el que las interacciones interdominio con la NTD favorecen la conformación αCTD (Ramírez-Sarmiento *et al.*, 2015; Tomar *et al.*, 2013) y promueven la ruptura del β-barril tras la disociación de la RNAP para volver al estado autoinhibido (Galaz-Davison *et al.*, 2021). Además,

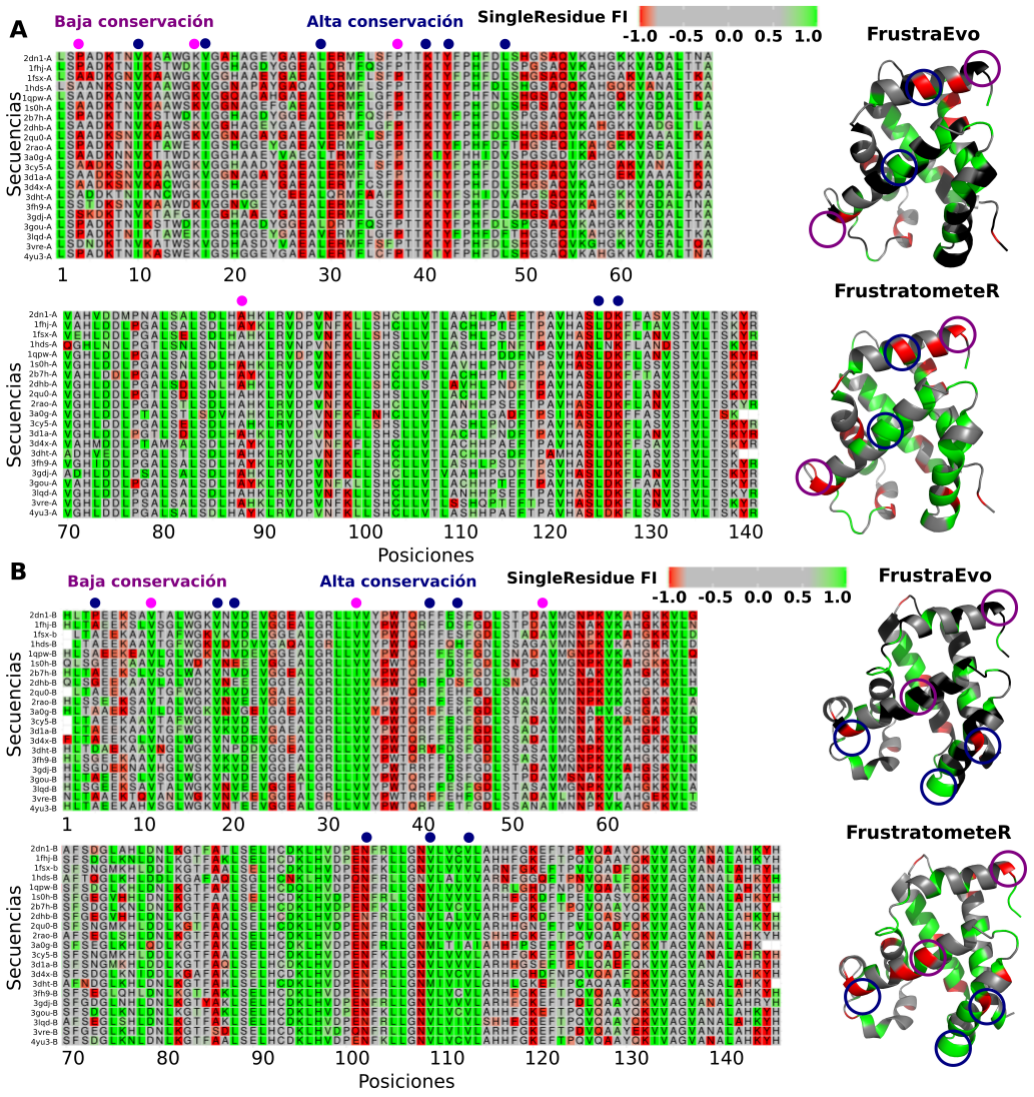

Figura 3.10: Alineamiento múltiple de secuencia y frustración (MSFA). A) α-globina. B) βglobina. A la derecha de cada MSFA se muestra la conservación de los estados de frustración basados en sus valores FrustIC calculados por FrustraEvo (izquierda) y la comparación entre los valores SRFI calculados por FrustratometeR (derecha) visualizados en la misma estructura (globina α y β humana, PDB 2DN1, cadena A y B). Los residuos están coloreados según sus estados de frustración en la representación FrustratometeR. Los residuos con FrstIC $> 0{,}5$ están coloreados según su estado de frustración más informativo en la representación FrustraEvo, mientras que los residuos con FrstIC $\leq 0{,}5$ están coloreados en negro. En círculos azules se muestran ejemplos de posiciones en las cuales la frustración es la misma entre especies (alto FrustIC) y en magenta se muestran algunos ejemplos de posiciones en las cuales la frustración es diferente entre especies (alto FrustIC).

una mutante de interacción interdominio (E48S) permite que RfaH libre coexista en ambas conformaciones en equilibrio 1:12. Estos resultados establecen que la interacción de RfaH con

RNAP *ops*-pausada cambia el equilibrio entre las dos conformaciones, de modo que debería existir un conjunto de interacciones altamente frustradas que favorezcan su transición de la conformación αCTD a la βCTD. Por lo tanto, utilizamos el análisis de la conservación de la frustración para comprender mejor los determinantes tridimensionales de este cambio en el plegado en la familia RfaH.

Resultados

Para detectar residuos involucrados en el cambio conformacional de RfaH, descargamos un conjunto de secuencias de proteínas RfaH no redundantes y relacionadas evolutivamente. En la actualidad la única proteína con estructuras disponibles para ambas conformaciones metamórficas de RfaH de la de *Escherichia coli* (UniProtID: P0AFW0). En la figura 3.11 se muestran las conformaciones αCTD (Fig. 3.11A) y βCTD (Fig. 3.11B) de *E. Coli*. También se ha demostrado que cuatro homólogos de *Salmonella typhimurium* (identidad de secuencia: 88 %), *Klebsiella pneumoniae* (80 %), *Vibrio cholerae* (64 %) y *Yersinia enterocolitica* (43 %) son capaces de sustituir la función de la RfaH de *E. coli* in vivo (Carter *et al.*, 2004). Además, la mutación deletérea de RfaH en *Y. pestis* e *Y. pseudotuberculosis* presenta defectos lipopolisacáridos similares a los de *E. coli* ΔRfaH y, por tanto, un comportamiento metamórfico (Carter *et al.*, 2004). Para realizar este estudio se seleccionaron proteínas homólogas de RfaH metamórfica.

La selección de ortólogos metamórficos de RfaH no es una tarea sencilla, ya que la mayoría de las secuencias relacionadas con la RfaH carecen de información experimental. Primero recuperamos todas las secuencias de la entrada IPR010215 en la base de datos InterPro (Blum *et al.*, 2021) y las agrupamos al 90 % de identidad utilizando CD-HIT (Huang *et al.*, 2010), lo que nos dio un total de 1004 secuencias. Como estrategia para determinar qué secuencias son probablemente homólogas metamórficas para RfaH de *E. Coli* seguimos los siguientes criterios, 1) para cada secuencia de RfaH, debe haber al menos una secuencia de NusG del mismo organismo en la base de datos Uniprot (Consortium, 2019), 2) debe predecirse que la secuencia de proteína RfaH completa se pliega en la estructura del dominio C-terminal (CTD) autoinhibido y plegado en la conformación α de RfaH (Fig. 3.11A); y

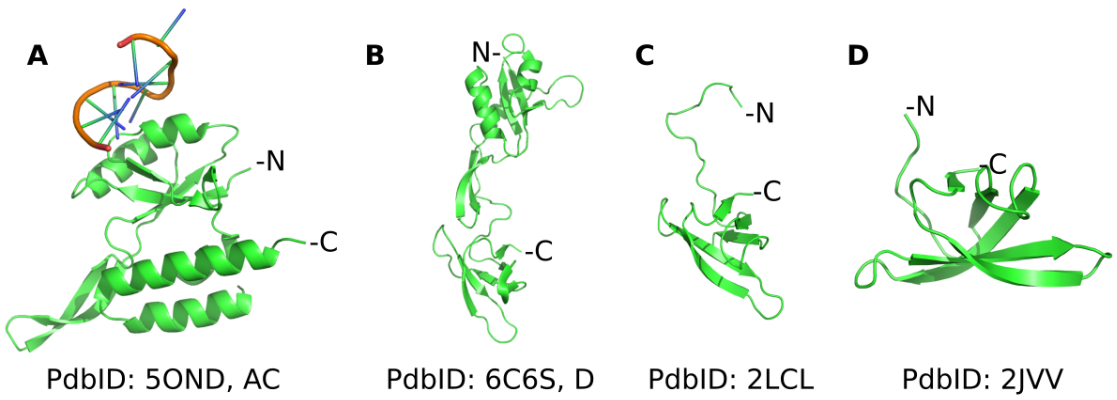

Figura 3.11: Estructura 3D de, A) RfaH de *E. Coli* unida al ADN. B) RfaH de *E. Coli*. C) Estructura del dominio C-terminal (CTD) autoinhibido. D) CTD aislado de RfaH en la estructura canónica βCTD. En verde esta coloreado el esqueleto de la proteína.

3) debe predecirse que el CTD aislado de RfaH se pliegue en la estructura canónica βCTD (Fig. 3.11C-D). Se consideró que el CTD comenzaba a partir del primer residuo que formaba estructura secundaria en el PDB 2LCL (residuo 115, patrón KVII). Seleccionamos aleatoriamente secuencias del conjunto de datos de 1004 entradas hasta completar 30 proteínas que cumplían los criterios mencionados. Para asegurar un conjunto de datos de alta calidad de homólogos potencialmente metamórficos, se comprobó manualmente que cada secuencia que se añadía al conjunto de datos a utilizar en el análisis cumplía los criterios mencionados. Por último, se generaron modelos *AlphaFold2* (Jumper *et al.*, 2020) (pLDDT medio= 71,97 +- 3,8) para cada secuencia del conjunto de homólogos; mínimo =65,1 y máximo = 79,4). En *AlphaFold2*, pLDDT (*predicted Local Distance Difference Test*) es una métrica que se utiliza para evaluar la calidad y confiabilidad de las predicciones de estructura de proteínas realizadas por el modelo, los valores fluctúan entre 0 y 100. Cabe destacar que las proteínas metamórficas suelen contener regiones con puntuaciones pLDDT más bajas (enlazadores y dominio metamórfico) debido a su diversidad conformacional. Por este motivo, consideramos todas las estructuras sin aplicar ningún otro filtro de calidad basado en su puntuación pLDDT media. Para cada modelo de *AlphaFold2* calculamos sus patrones de conservación de la frustración utilizando tanto el SRFI como los FIs de contactos *mutational* y *configurational*.

En la figura 3.12, se muestran los resultados del FrustraEvo para las 30 secuencias de RfaH, el dominio metamórfico esta definido desde la posición 110 a la 162 (marcado en recuadro

naranja en la figura 3.12). En el MSFA (fig. 3.12A), se observan posiciones en las que la clasificación del índice de frustración (neutro, altamente frustrado y mínimamente frustrado) y la secuencia es la misma para todos los organismos, como por ejemplo las posiciones 4, 21, 41, 51 y 72. También se observan posiciones en las que la secuencia y la clasificación del índice de frustración son diferentes entre organismos, como por ejemplo, las posiciones 2, 14, 15, 45 y 83. Por último se observan posiciones en las cuales la frustración es la misma para todas las especies pero la secuencia en algunas especies es diferente, como por ejemplo, la posiciones 7, 20, 69, 78, 79, 92, 93, 118, 141, 142 y 143. Las posiciones en las cuales la clasificación del índice de frustración es la misma para todas las especies pero la secuencia es diferente, nos puede estar diciendo que en esas posiciones es importante que la frustración local sea mantenga, ya que hay un cambio de aminoácido pero este cambio no afecta significativamente al índice de frustración del residuo.

En el logo de secuencia y en el de frustración (fig. 3.12B) en general observamos que hay muchos residuos que están conservados y mínimamente frustrados y solamente dos altamente frustrados (FrustIC $>$ 0,5), los cuales están localizados en el dominio αCTD. Contrario al resultado que esperábamos ver, ya que lo esperado era que los residuos altamente frustrados estén localizados en el dominio CTD. Estos resultados nos lleva a la pregunta de que si la falta de residuos altamente frustrados en el dominio CTD de la conformación autoinhibida se debe a que esta conformación es más estable que la conformación canónica βCTD. Para responder a esta pregunta, hemos decidido llevar a cabo un análisis de la frustración a nivel de contacto. En la figura 3.13A se muestran los mapas de contactos para la familia de RfaH para los índices *mutational* y *configurational*, podemos ver en ambos índices, en el dominio CTD, la presencia de contactos mínimamente frustrados entre residuos del dominio NTD y entre residuos del dominio αCTD. Esto nos lleva a pensar que posiblemente el dominio CTD de la conformación autoinhibida está siendo estabilizado por contactos entre ambos dominios (NTD y αCTD). Esto es coherente con la estabilización a través de interacciones interdominio (Ramírez-Sarmiento *et al.*, 2015; Tomar *et al.*, 2013) que desencadenan el cambio de plegado de RfaH hacia la conformación βCTD (Burmann *et al.*, 2012; Shi *et al.*, 2017). A partir de los resultados de la conservación de frustración local para los índices *mutational* y *configurational* encontramos 9 residuos interdominio (L6, F51, L96, F126, I129, L141, L142, L145,

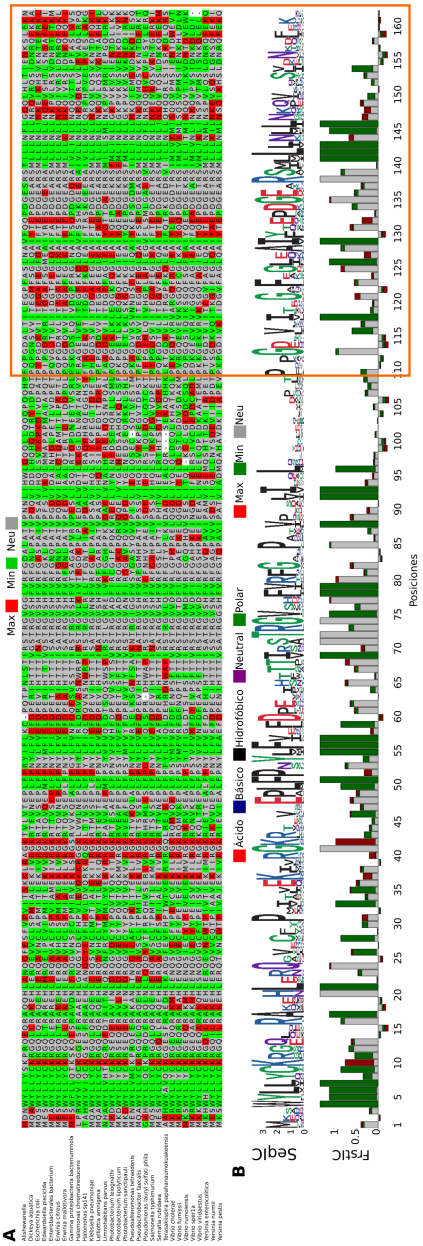

Figura 3.12: Resultados del FrustraEvo para la familia de RfaH a nivel de residuo único usando como referencia la secuencia de *E. Coli*. A) MSFA. B) Logo de secuencia (arriba) y logo de frustración (abajo). En el logo de frustración y en el MSFA en color verde se representan los residuos mínimamente frustrados, en gris los neutros y en rojo los altamente frustrados. En el logo de secuencia en rojo se representan los aminoácidos ácidos, en azul los básicos, en negro los hidrofóbicos, en violeta los neutros y en verde los polares.

I146) que contribuyen a la estabilización de la interfaz interdominio de RfaH entre NTD y αCTD.. La metodología usada para detectar los residuos interdominio fue la siguiente, seleccionamos aquellos residuos que: 1) establecen contactos interdominio según los mapas de

contactos obtenidos por FrustraEvo; 2) por cada residuo se seleccionaron todos los contactos cuyo FrustIC $> 0{,}5$; 3) los contactos filtrados según el paso 2 deben estar presentes en más del 50 % de los modelos analizados; 4) del total de contactos filtrados según los filtros en los puntos 2 y 3, el 50 % de ellos tienen que estar mínimamente frustrados; y 5) el residuo debe de tener al menos 3 interacciones mínimamente frustradas con otros residuos CTD. En el recuadro de la figura 3.14A, muestra una ampliación de la interfaz entre los dominios NTD y αCTD, se muestra en azul y con etiqueta en blanco los residuos de interfaz.

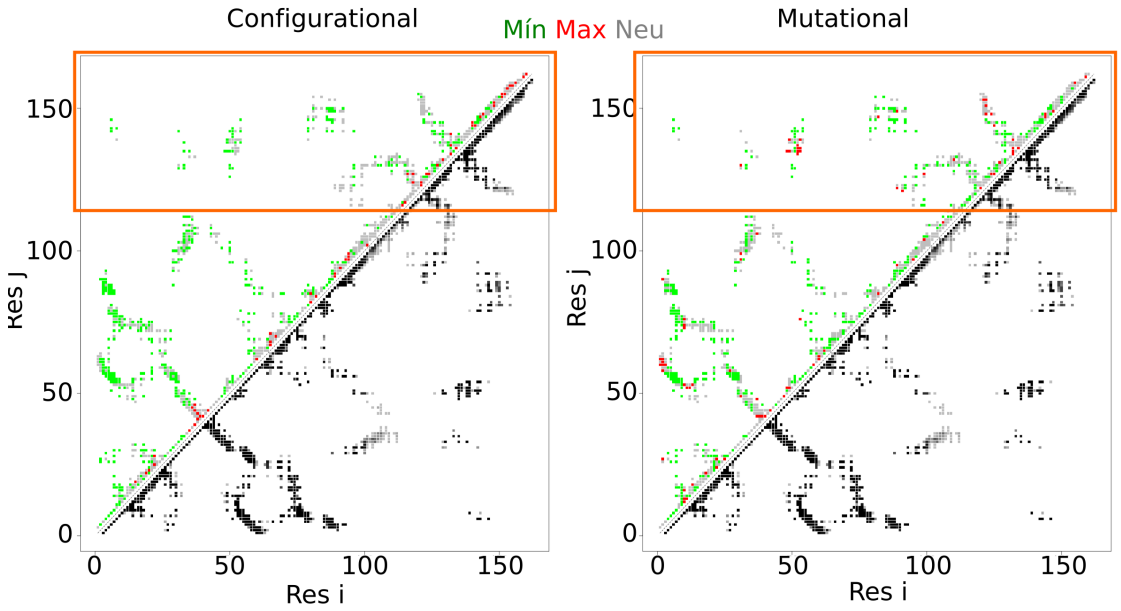

Figura 3.13: Mapas de contactos de la frustración local para el índice *mutational* y *configurational* de la familia de RfaH. Por encima de la diagonal principal se representa el estado de frustración que más se conserva para cada contacto, en color verde se representan los contactos mínimamente frustrados, en rojo los altamente frustrados y en gris lo neutros. Por debajo de la diagonal principal se representa en escala de grises el valor de la conservación del estado de frustración local. En el recuadro naranja se muestra el dominio CTD.

Luego, utilizamos FrustratometeR para predecir los cambios en la frustración de mutar individualmente cada uno de los residuos interdominio por los otros 19 aminoácidos (Fig. 3.14B y figura 1 del Anexo (8.9)) La figura 3.14B ilustra cómo cambia la frustración local del índice *mutational* al mutar F51. La mayoría de los 21 contactos formados por F51 (Fig. 3.14B, letras azules) están mínimamente frustrados, con la excepción de 5 que son neutros. En general, algunas mutaciones producen valores de frustración similares en todos los contactos (por ejemplo, F51M), mientras que otras pasan de interacciones mínimamente frustradas a

neutras o altamente frustradas (por ejemplo, F51K). El mismo efecto se observa al repetir el análisis para los 8 residuos interdominio restantes (figura 1 del Anexo (8.9)), lo que conduce a la identificación de dos tipos de mutaciones: 1) "Mutaciones de frustración similar" (SFMs), que mantendrían la naturaleza estabilizadora de las identidades aminoacídicas nativas (L6I, F51M, L96W, F126W, I129V, L141V, L142V, L145M, I146V) y 2) "Mutaciones Altamente Frustradas" (HFMs), que maximizarían el índice de frustración local con sus residuos vecinos (L6D, F51K, L96K, F126N, I129E, L141D, L142K, L145E, I146D). Generamos dos secuencias mutantes RfaH de E. coli que contenían todos los SFM o HFM y predecimos sus estructuras con *AlphaFold2*. Las estructuras con SFMs muestran una estructura similar a la del salvaje con una conformación αCTD (Fig. 3.15A) mientras que las que contienen el conjunto de HFMs muestran un cambio conformacional similar a βCTD (Fig. 3.15B). Esto último sugiere que la desestabilización de la interfaz entre NTD y αCTD podría inducir el comportamiento metamórfico del NTD de RfaH.

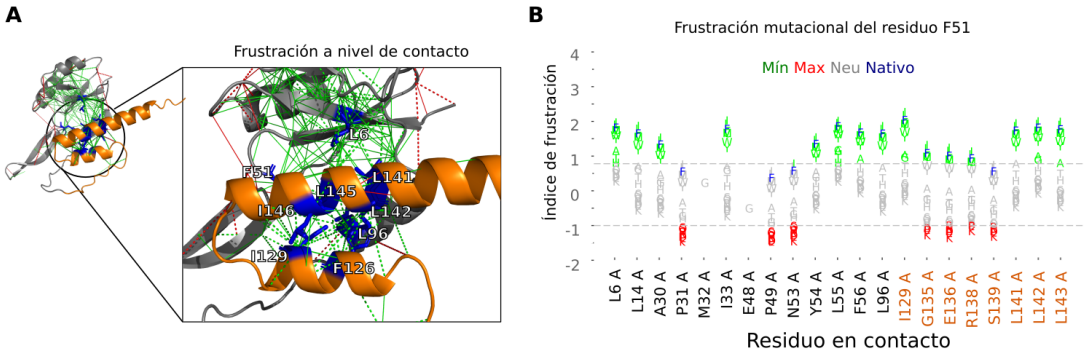

Figura 3.14: Análisis de frustración de cambio conformacional metamórfico de una proteína. A) Frustratograma de la conservación del índice *mutational*, la estructura es de RfaH de *E. Coli*. Las líneas rojas corresponden a interacciones altamente frustradas y las verdes a interacciones mínimamente frustradas. El esqueleto de la proteína coloreado en naranja corresponde a la región interdominio (CTD) y los residuos en azul y representados en *sticks* son los 9 residuos de interfaz. B) Cambios en la frustración local para las 19 mutaciones del residuo de fenilalanina 51 (F51) utilizando FrustratometeR. El eje x muestra los residuos con los que el residuo, ya sea silvestre (*Phe*) o mutado, establece contactos en la estructura. En el eje y se muestra el índice de frustración *mutational* de los contactos. La identidad del aminoácido de tipo silvestre se muestra en azul y las variantes se colorean según su estado de frustración.

El análisis anterior podría no reflejar todos los posibles cambios que podría introducir la

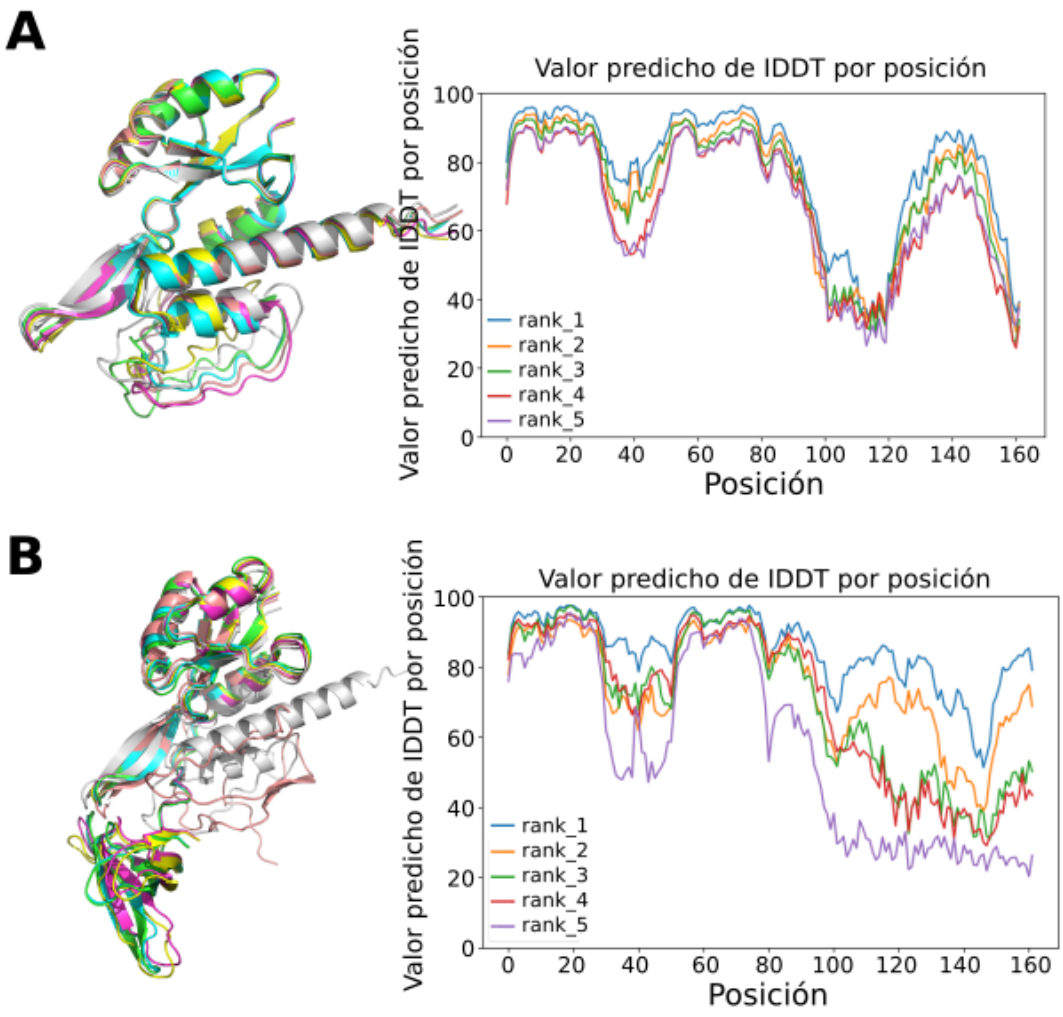

Figura 3.15: Los 5 Modelos de *AlphaFold2* de las secuencias generadas y la salvaje de RfaH de *E. Coli* A) SFMs y B) HFMs. En blanco la salvaje de RfaH de *E. Coli* en azul el top 1 de los modelos de *AlphaFold*, en naranja el top 2, en verde el top 3, en rojo el top 4 y en violeta el top 5.

naturaleza. Por lo tanto, investigamos qué cambios de secuencia fueron introducidos por la evolución al comparar la RfaH con NusG, su homólogo no metamórfico. Podemos ver que 6 de los 9 residuos interdominio cambian su identidad en el alineamiento de secuencias RfaH/NusG (es decir, mutaciones similares a NusG: L6V, I129V, L141V, L145I, I146F y L142S, Fig. 2 del Anexo 8.10A). Introdujimos estas mutaciones en la secuencia RfaH, generamos los modelos de *AlphaFold2* y descubrimos que 3 de las 5 mejores predicciones de estructura de *AlphaFold2* muestran un plegado similar al αCTD y dos muestran, al final de la región metamórfica un plegado similar al βCTD (Fig. 3.16). Los perfiles de frustración local de estas mutaciones

(Fig. 2 del Anexo 8.10B) muestran que 4 de ellas son SFM (L6V, I129V, L141V, L145I), una cambia los valores de frustración de mínimamente frustrado a neutros (I146F) y sólo una (L142S) es un HFM. Cuando sólo se introduce L142S en RfaH, *AlphaFold2* devuelve 3 modelos con un plegado αCTD y 2 tienen un plegado βCTD (Fig. 3.16A). De los 115 residuos que cambian su identidad en el alineamiento de la secuencia RfaH/NusG, L142S es el único caso en el que el αCTD cambia para adoptar una conformación similar al βCTD (Fig. 3.16B) tras la predicción de la estructura de *AlphaFold2*. La conformación CTD adoptada por el mutante L142S es menos similar a la de NusG en comparación con la obtenida cuando se introducen las 6 mutaciones similares a NusG. Por lo tanto, parece que no sólo es necesaria la introducción de frustración para desencadenar un cambio conformacional en RfaH, sino que también es necesario afinar los contactos mínimamente frustrados en la región interdominio.

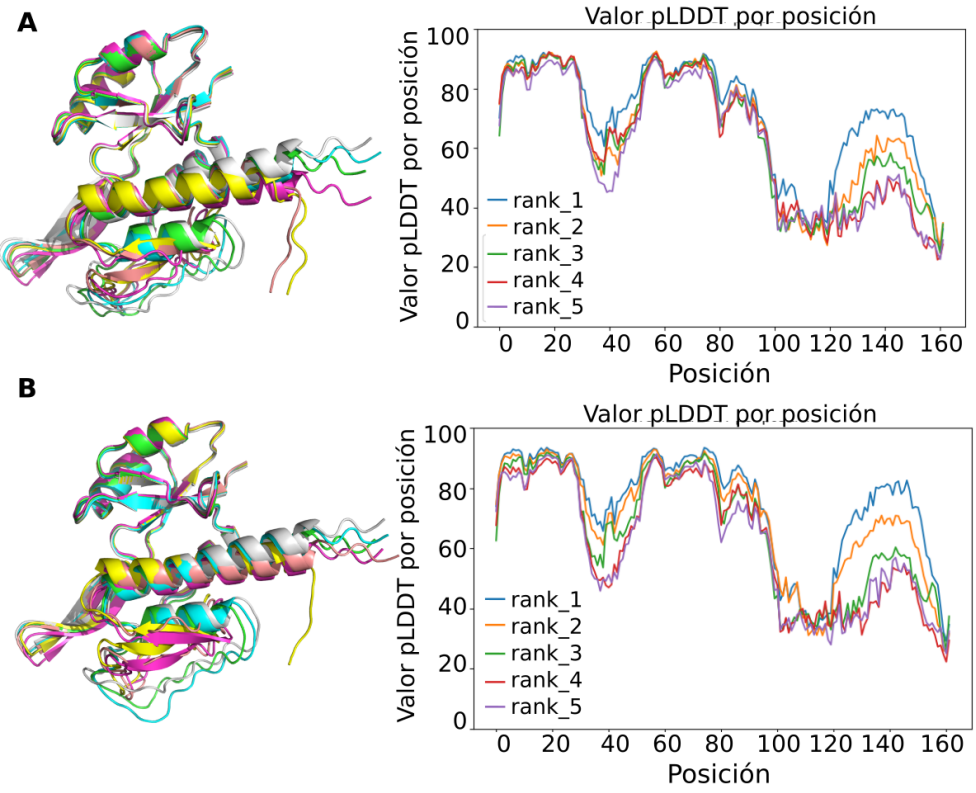

Figura 3.16: Los 5 Modelos de *AlphaFold2* de la secuencia con mutaciones Nusg. A) Los 6 residuos que cambian la identidad del aminoácido entre RfaH y NusG y B) Solamente la mutante L142S. En blanco la salvaje de RfaH de *E. Coli* en azul el top 1 de los modelos de *AlphaFold*, en naranja el top 2, en verde el top 3, en rojo el top 4 y en violeta el top 5.

3.3. FrustraPocket: Una herramienta para la predicción de sitios de unión proteína-ligando

La identificación de los sitios de unión proteína-ligando (LBS, por sus siglas en inglés) constituye un paso importante hacia el esclarecimiento de las funciones de las proteínas y la obtención de información para el diseño racional de fármacos dirigidos a proteínas específicas. Los bolsillos de interacción proteína-ligando son regiones específicas de una proteína que tienen la capacidad de unirse a una molécula pequeña o ligando. Estos bolsillos están formados por un conjunto de residuos de aminoácidos que presentan interacciones con el ligando permitiendo una unión específica y estable. El problema de la unión proteína-molécula pequeña es una cuestión central en el campo del diseño de fármacos y la bioinformática estructural. Se trata de predecir cómo una proteína y una molécula pequeña pueden interactuar entre sí, lo que es fundamental para entender los mecanismos de reconocimiento molecular y el desarrollo de fármacos. También predecir cómo son las conformaciones y orientaciones relativas de la proteína y el ligando en su complejo, la identificación de los residuos de la proteína involucrados en la unión y la estimación de la afinidad y estabilidad de la interacción. Debido a los desafíos que presenta la identificación de los residuos de unión proteína-ligando, en los últimos años se han desarrollado varios métodos para predecir estos residuos. La mayoría se basan en una definición geométrica de los bolsillos de la proteína a los que se unen los pequeños ligandos. Recientemente, producto del crecimiento de la inteligencia artificial (IA), la utilización de algoritmos de IA en la predicción de residuos de interacción proteína-ligando es una área activa de investigación en bioinformática y biología computacional. En particular, el aprendizaje automático supervisado se ha utilizado para predecir residuos de interacción proteína-ligando utilizando características estructurales y secuenciales de la proteína, así como información sobre el ligando (Harren *et al.*, 2022; Schneider y Clark, 2019; Schneider *et al.*, 2020; Zhavoronkov *et al.*, 2019). Se pueden utilizar diferentes tipos de características, como propiedades físico-químicas de los aminoácidos, información de estructura secundaria y accesibilidad al solvente, entre otros. El modelo de aprendizaje automático se entrena con ejemplos etiquetados de residuos de interacción proteína-ligando y residuos de no interacción proteína-ligando para aprender a distinguir entre ellos. Por lo tanto, esta área de investigación continúa evo-

lucionando con el desarrollo de nuevos enfoques y algoritmos para mejorar la precisión y la capacidad de generalización de los modelos predictivos. Aquí presentamos una herramienta llamada FrustraPocket que está basada en el concepto de frustración en proteínas y utiliza un sencillo algoritmo de aprendizaje automático para predecir sitios de unión proteína-ligando. Además demostramos que los sitios de unión de pequeños ligandos están enriquecidos en interacciones altamente frustradas en el estado no unido.
El código y la implementación del algoritmo se encuentra disponible en, GitHub: https://github.com/CamilaClemente/FrustraPocket/
Docker container: https://hub.docker.com/r/proteinphysiologylab/frustrapocket

3.3.1. Conjunto de datos, diseño e implementación del algoritmo

Conjunto de datos y frustración local. Para caracterizar las LBSs, seleccionamos de la base de datos BioLiP (Yang *et al.*, 2012) todas las proteínas con anotaciones de *ECNumber* con el fin de seleccionar sólo proteínas enzimáticas. Las enzimas se clasificaron según su estado oligomérico y finalmente se seleccionó un conjunto de datos no redundante de 1007 proteínas monoméricas y enzimáticas. Para eliminar redundancia se utilizó la base de datos de UniProt, es decir, que nos quedamos con un solo PDB por UniProt, el PDB seleccionado fue el de mejor resolución. Las estructuras de las proteínas se descargaron del PDB (https://www.rcsb.org/), los patrones de frustración y la densidad local (LD) se calcularon utilizando el FrustratometeR de proteínas (Rausch *et al.*, 2021).
Métodos de extracción de características. Con el fin de construir el conjunto de datos para el entrenamiento y las pruebas de nuestro para el algoritmo de aprendizaje automático, utilizamos el índice de frustración para todos los residuos del conjunto de datos. Además, los residuos se clasificaron como los que están anotados de que formar interacciones proteína-ligando (clase 1) y los que no forman interacciones proteína-ligando (clase 0). Dado que la cantidad de residuos que son de interacción proteína-ligando (clase 1) es un porcentaje muy pequeño con respecto a la longitud total de la proteína, para evitar un conjunto de datos desequilibrado, hemos utilizado la estrategia de *undersampling* implementado el algoritmo de *NearMiss* y de *oversampling* implementado el algoritmo de *SMOTE*. *NearMiss* tiene como

objetivo reducir el número de muestras en la clase mayoritaria (la clase con más ejemplos) para igualarla con la clase minoritaria (la clase con menos ejemplos). NearMiss selecciona cuidadosamente las muestras de la clase mayoritaria que están çerca"de las muestras de la clase minoritaria en el espacio de características. *SMOTE* es una técnica que crea muestras sintéticas de la clase minoritaria para equilibrar el conjunto de datos. El conjunto de datos final contiene 97,246 aminoácidos (48,623 de la clase 1 y 48,623 de la clase 0) y las características utilizadas figuran en la tabla 3.2.

Característica	**Descripción**
Número de contactos	Número de contactos que forma cada residuo
Local Density	Valor del *Local density* del residuo
% de contactos altamente frustrados	Porcentaje de contactos altamente frustrados que forma el residuo con otros residuos.
% de contactos altamente frustrados alrededor de una esfera de 5Å	Porcentaje de contactos altamente frustrados dentro de una esfera de 5Åalrededor del Cα
% de contactos mínimamente frustrados alrededor de una esfera de 5Å	Porcentaje de contactos mínimamente frustrados dentro de una esfera de 5Åalrededor del Cα
% de contactos neutros alrededor de una esfera de 5Å	Porcentaje de contactos neutros dentro de una esfera de 5Åalrededor del Cα
Clases	Indica si el aminoácido es un residuo de interacción P-L (1) o no (0)

Tabla 3.2: Características utilizadas para el entrenamiento y el testeo de *XGBoost*.

Construcción y evaluación del modelo. El método de aprendizaje automático usado fue *XGBoost* (*Extreme Gradient Boosting*) el cual es una biblioteca de código abierto desarrollada para implementar el algoritmo de *Gradient Boosting* en *Python*. El algoritmo de *Gradient Boosting*, es una técnica de aprendizaje automático en la que se combinan múltiples modelos de árboles de decisión débiles para formar un modelo más robusto y preciso. En cada iteración, se ajusta un nuevo árbol de decisión al residuo de los modelos anteriores, lo que permite al modelo mejorar gradualmente su capacidad de predicción. Este algoritmo fue desarrollado para maximizar su precisión y escalabilidad, así como para ampliar los límites de la potencia de cálculo mejorando su rendimiento y velocidad computacional. Además, la implementación de estos modelos ha dado muy buenos resultados en trabajos anteriores relacionados con este

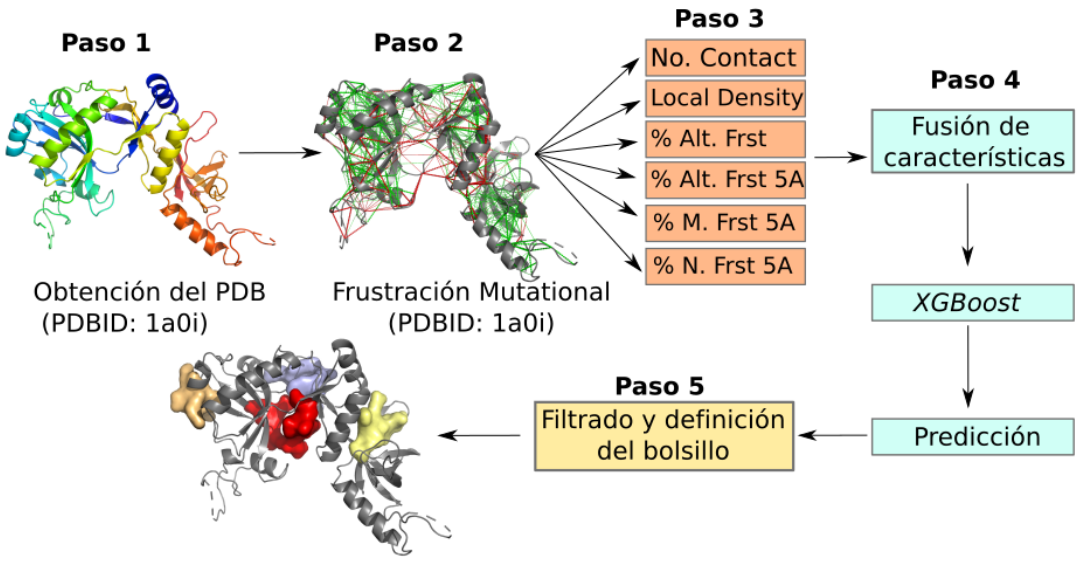

Figura 3.17: **El flujo de trabajo del algoritmo FrustraPocket**. **Paso 1:** Obtención del la estructura en formato PDB, si no es proporcionada por el usuario, se descarga del *Protein Data Bank*. **Paso 2:** Calculo de la frustración *mutational* con el *frustratometeR* (Rausch *et al.*, 2021). **Paso 3:** todas las características necesarias se obtienen a partir de los resultados de frustraR. **Paso 4:** se realizan predicciones utilizando el modelo definido. Se filtran los resultados de la predicción y se definen los bolsillos.

tema (Chen *et al.*, 2020; XU *et al.*, 2020) y también porque nuestro conjunto de datos es de tipo estructurado.

Implementación.

El flujo de trabajo utilizado para construir FrustraPocket se esquematiza en la figura 3.17. La entrada de FrustraPocket es una estructura de proteína en formato PDB.

El archivo de entrada es únicamente un archivo en formato PDB no necesariamente tiene que ser una estructura obtenida experimentalmente, también puede ser un modelo generado por cualquier método de modelado, o solamente un PdbID que luego el algoritmo se encargará de descargarlo de la base de datos del *Protein Data Bank*. **Paso 1:** En caso de que el usuario no provea la estructura en formato pdb, el pipeline lo va a descargar del *Protein Data Bank* **Paso 2:** Cálculo del MFI (Índice de Frustración *mutational*) (Rausch *et al.*, 2021) y la correspondiente proporción de interacciones altamente frustradas por residuo (MFI_hprop) y el LD (*Local Density*) (Davtyan *et al.*, 2012) de la proteína. *FrustratometeR* calcula el porcentaje de los diferentes tipos de contacto de frustración (es decir, altamente frustrado,

neutro o mínimamente frustrado) alrededor de una esfera de 5 Å, centrada en el átomo Cα del residuo (5Adens). **Paso 3:** Ejecuta la predicción utilizando el modelo *XGBoost* ML. **Paso 4:** Una de las ventajas que tiene FrustraPocket es que define pocos bolsillos, luego de que se predicen los residuos, se filtran y se definen los bolsillos, un bolsillo está definido por al menos 5 residuos cercanos predichos como residuos de la clase 1, en caso de que no haya residuos de clase 1 cerca uno del otro no se define bolsillo. Los archivos de salida incluyen el cálculo de frustración, un *script* de *pymol* para visualizar los bolsillos en la estructura de la proteína y el centro de masa para cada bolsillo.

3.3.2. Resultados

Los sitios de unión proteína-ligando están rodeados espacialmente por interacciones altamente frustradas

Para analizar la distribución de la frustración local en los sitios de unión proteína-ligando, recopilamos todas las entradas de la base de datos BioLiP (Yang *et al.*, 2012), dividimos el conjunto de datos según el estado oligomérico de las proteínas y se seleccionaron las proteínas monoméricas (1007 entradas no redundantes). Se seleccionaron enzimas monoméricas porque sus patrones de frustración local ya fueron caracterizados y analizados (Freiberger *et al.*, 2019) también seleccionamos monómeros debido a su simplicidad topológica. A continuación, calculamos los patrones de frustración local utilizando el paquete FrustratometeR (Rauer *et al.*, 2021). Para cuantificar los patrones de frustración local, calculamos las función de distribución radial g(r) para las distintas clases de contactos, clasificadas por el índice de frustración. La g(r) calcula la densidad de VPs (ver métodos) correspondientes a los distintos tipos de contactos en función de la distancia relativa al Cα de los residuos de unión.

En la figura3.18A mostramos la función de distribución de radial $g(r)$ para aquellos residuos que están anotados como residuos de unión proteína - ligando. Podemos ver un enriquecimiento de interacciones neutras y altamente frustradas, en relación con la topología de contactos de la proteína (línea negra). La distribución de las interacciones alrededor de los residuos de unión muestra dos picos característicos, uno situado alrededor de 1 Å, correspondiente a las interacciones de los propios residuos de unión (primera cáscara), y un segundo pico

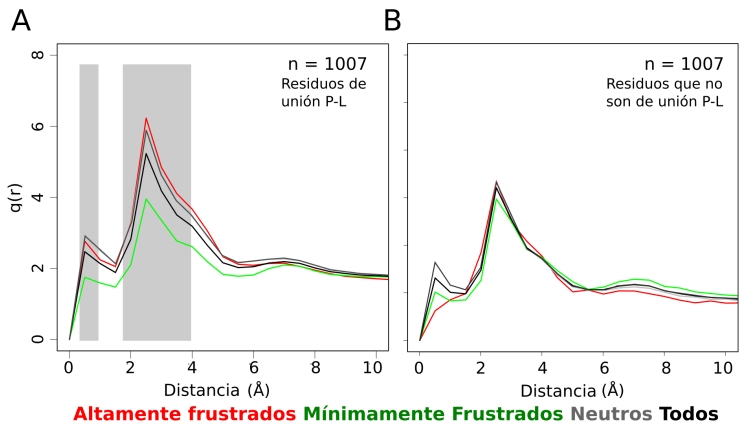

Figura 3.18: Función de distribución radial g(r), entre el Cα de **A)** los residuos de unión anotados y el centro de masa de los contactos y **B)** los residuos de control. Verde, contactos mínimamente frustrados; rojo, contactos altamente frustrados; gris, contactos neutros; negro, todos los contactos. Los gráficos g(r) se ajustaron en los rangos de sus ejes para mejorar la visualización; en todos los casos los valores g(r) se normalizaron de forma que g(20) = 1, **A** Residuos anotados como residuos de unión proteína - ligando. **B** Residuos de control, definidos como residuos seleccionados aleatoriamente que no están anotados como residuos de unión proteína - ligando. En gris se muestran la primera y la segunda cáscara, respectivamente.

entre 2 y 4 Å, que comprende las interacciones entre los residuos que coordinan la unión (segunda cáscara). Sin embargo, el enriquecimiento de interacciones altamente frustradas en la segunda cáscara es superior al esperado por la topología de la proteína (línea negra). Se observa que tanto en la primera como en la segunda cáscara hay una disminución de las interacciones mínimamente frustradas. Estos resultados muestran que los sitios específicos para el reconocimiento proteína-ligando están típicamente frustrados en el estado no unido. En la figura3.18B, observamos que la $g(r)$ para el conjunto de control, que se generó utilizando residuos aleatorios que no están involucrados en los sitios de unión, no muestra ningún enriquecimiento, en contraste con lo que se observa para los residuos de unión al ligando figura 3.18A. Basándonos en esto y en trabajos anteriores en los que observamos un enriquecimiento de interacciones altamente frustradas alrededor de los residuos de interacción proteína-ligando y también en los sitios catalíticos (Freiberger *et al.*, 2019) decidimos usar esta característica para predecir la interacción proteína-ligando y los sitios catalíticos combinando la información de los patrones de frustración local y la densidad local de residuos en una estructura proteica.

Rendimiento del modelo

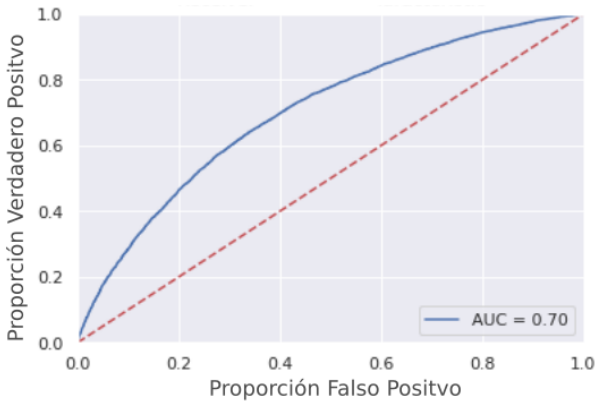

Figura 3.19: La curva ROC de *XGBoost* aplicada a los patrones de frustración local de las enzimas monoméricas. La curva ROC está definida como, la tasa de verdaderos positivos (TP) que se define como TP/(TP+FN) y la tasa de falsos positivos (FP) que se define como FN/(FN+TP), donde FN es el falso negativo. La tasa de verdaderos negativos es la probabilidad de que un verdadero positivo dé positivo y la tasa de falsos negativos es la probabilidad de que un verdadero positivo no sea detectado.

Para evaluar la eficacia del modelo *XGBoost*, implementado en este trabajo, utilizamos la curva ROC (Característica operativa del receptor) que se muestra en la figura 3.19. Para obtener la curva ROC es necesario calcular la tasa de verdaderos positivos (TP) que se define como TP/(TP+FN) y la tasa de falsos positivos (FP) que se define como FN/(FN+TP), donde FN es el falso negativo. El área debajo de la curva (AUC) obtenido por *XGBoost* fue de 0,70, lo que indica que nuestro método no selecciona los residuos al azar. En la tabla 3.3 se muestran los valores de las métricas para la exactitud, precisión, recuperación, *kappa* y *f1-score*. Estas métricas son utilizadas en la evaluación de modelos de clasificación y permiten medir el desempeño de los modelos en la predicción de clases o categorías. La exactitud representa la proporción de predicciones correctas realizadas por el modelo sobre el total de instancias. Se calcula dividiendo el número de predicciones correctas entre el número total de instancias en el conjunto de datos. La precisión mide la proporción de predicciones positivas que fueron realmente correctas. La recuperación mide la proporción de instancias positivas que fueron correctamente identificadas por el modelo. El coeficiente *kappa* es una métrica que ajusta la exactitud observada por la cantidad de coincidencias esperadas al azar y varía entre -1 y 1, donde 1 representa una concordancia perfecta, 0 representa concordancia

aleatoria y -1 representa discordancia total. El *F1-score* es una métrica que combina la precisión y la recuperación en una sola medida y se calcula como la media armónica de la precisión y la recuperación. Los valores de las métricas (tabla 3.3) indican que nuestro modelo detecta correctamente aproximadamente el 65-70 % de los residuos de unión ligando-proteína, basándose únicamente en sus patrones de frustración local en los estados no unidos.

Métrica	**Valor**
Área bajo la curva (AUC)	0,70
Precisión	0,64
Recuperación	0,66
Kappa	0,29
f1-score	0,65

Tabla 3.3: Métricas usadas para evaluar el rendimiento del modelo.

Ejemplo de uso del FrustraPocket. Como ejemplo, hemos aplicado FrustraPocket a la ADN ligasa ATP-dependiente (PdbID: 1A0I). En la figura 3.20A-B representa el primer paso de la herramienta donde figura3.20A muestra el frustratograma de frustración local *mutational* y figura3.20B muestra el LD de cada residuo de la proteína. Fig.3.20C representa la salida de los nueve bolsillos predichos de la proteína y figura3.20D el centro de masa para cada bolsillo predicho.

3.4. Conclusiones del capítulo

Presentamos un paquete R fácil de usar para calcular la frustración local energética en estructuras proteicas. El paquete incluye nuevas funcionalidades para evaluar el efecto de las mutaciones en la frustración local, así como para analizar la frustración local a lo largo de trayectorias de dinámica molecular. Su sencilla interfaz, junto con las nuevas funcionalidades implementadas, facilitará el análisis de la frustración a mayores escalas y puede utilizarse para incluir *FrustratometeR* como parte de diferentes *pipelines* para el análisis estructural de proteínas.

Hemos desarrollado el FrustraEvo, una herramienta para el análisis de patrones energéticos entre familias de proteínas basado en la conservación de los niveles de frustración. Estos

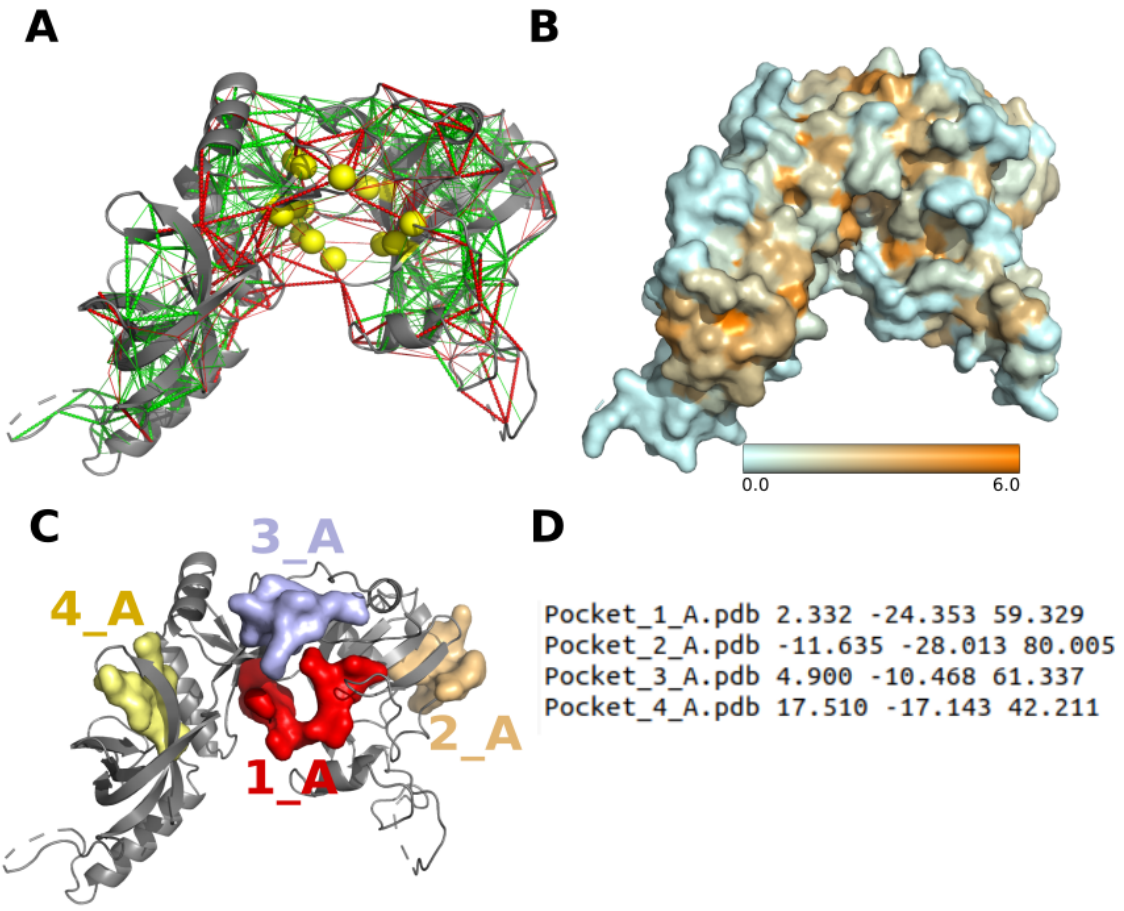

Figura 3.20: **Ejemplo PdbID: 1A0I. A.** Los esqueletos de las proteínas se muestran como *catoon* en color gris, los contactos mínimamente frustrados se representan con líneas verdes, las interacciones altamente frustradas con líneas rojas. Se han omitido las interacciones neutras para facilitar la interpretación. **B.** Los residuos con menor densidad local se muestran en azul y los residuos con mayor densidad local se muestran en naranja. **C.** Salida de la herramienta, en diferentes colores los bolsillos predichos. **D.** Salida de la herramienta, un archivo en formato txt que contiene los centro de masa de cada bolsillo.

patrones revelan restricciones fisicoquímicas relacionadas con la estabilidad y la función, proporcionando una interpretación biofísica del impacto de la divergencia de secuencias a lo largo de escalas de tiempo evolutivas. La comparación de los patrones de frustración entre familias de proteínas apunta a regiones de diversidad funcional (hemoglobina). El mejor ejemplo es Leu142, que se ha propuesto que desempeña un papel importante en impedir que el dominio metamórfico pase de la conformación totalmente alfa a la totalmente beta. En el caso de las familias de globinas α y β estrechamente relacionadas que constituyen la molécula de hemoglobina, las funciones diferenciales en el transporte de oxígeno y la contribución al

ensamblaje del complejo cuaternario se traducen en una diferencia significativa en el número y la ubicación de posiciones energéticamente conservadas correspondientes a sitios de interacción proteína-proteína y puentes salinos. Para la proteína metamórfica RfaH, el análisis de contactos mínimamente frustrados conservados condujo a la identificación de residuos claves implicados en transiciones conformacionales.

La evolución de las familias de proteínas se ve constreñida por un estrecho margen de estabilidad y energética de plegabilidad en el contexto de exigentes requisitos funcionales. En estos márgenes, demasiadas regiones mínimamente frustradas pueden dificultar la evolución funcional, mientras que la presencia de demasiadas regiones altamente frustradas impedirá que se produzca el plegado (Ferreiro *et al.*, 2014). El análisis de la conservación de la frustración dentro de las familias de proteínas puede utilizarse para definir los límites teóricos para preservar la función a lo largo de la evolución, revelando la interacción entre secuencia, estructura, dinámica y función (Sánchez *et al.*, 2022). Ahora que pueden obtenerse modelos de estructuras proteicas de alta calidad para los miembros de cualquier familia de proteínas, nuestra estrategia de análisis de conservación de la frustración se erige como una valiosa herramienta para aumentar el nivel de anotación funcional en las bases de datos biológicas (Rauer *et al.*, 2021). Además, prevemos que la frustración local energética, tanto a nivel de proteínas individuales como de familias, podría utilizarse como una medida novedosa para complementar el entrenamiento de los métodos de aprendizaje automático más avanzados, como los modelos de lenguaje de proteínas, como forma de mejorar sus capacidades predictivas/generativas. Por último, los ejemplos aquí mostrados ilustran cómo el enfoque que presentamos puede guiar aplicaciones tales como la construcción de proteínas artificiales o la predicción del riesgo de nuevas variantes de proteínas patógenas naturales.

La identificación de sitios de unión de pequeños ligandos en la estructura de una proteína es un tema central de la fisiología proteica y ha sido objeto de un número creciente de estudios en la última década. Actualmente existen muchos predictores, la mayoría de ellos basados en la geometría y hasta ahora no existía ningún método basado en el análisis de la proteína solo energético. La frustración energética local es un concepto basado en la biofísica relacionado con diversos aspectos funcionales de las proteínas (Ferreiro *et al.*, 2007, 2011; Freiberger *et al.*, 2019; Kumar *et al.*, 2013; Lindstrom y Dogan, 2018; Parra *et al.*, 2015; Stelzl *et al.*, 2020),

especialmente a las interacciones entre proteínas o proteínas y sus ligandos. De ahí que consideremos que la frustración es un concepto intuitivo que puede mejorar la predicción de las LBSs, como hemos demostrado en esta sección.

Capítulo 4

Proteínas *fuzzy*

Como ya se mencionó anteriormente las proteínas luego de ser sintetizadas se pliegan adoptando un conjunto de estructuras de mínima energía libre que se denomina estado nativo. Sin embargo, esto no implica que las estructuras que forman parte del estado nativo estén estructuralmente bien definidas, más aún se ha demostrado que en naturaleza existen muchas proteínas o regiones de proteínas que son desordenadas en su estado nativo (Tompa, 2002). A este tipo de proteínas se las conoce como proteínas intrínsecamente desordenadas (IDPs, por sus siglas en inglés). La función más reconocida de estas proteínas es la de reconocimiento molecular (Tompa, 2005) y están involucradas en muchos procesos biológicos. Como por ejemplo, la proteína $p27^{kip1}$ involucrada en la regulación de la progresión de la fase G_1 a la fase S en el ciclo celular. El domino desordenado de $p27^{kip1}$, provoca la detención del ciclo celular mediante la unión y el cierre del complejo ciclina A/Cdk2, La disociación de $p27^{kip1}$ de este complejo causa la progresión a la fase S (Galea *et al.*, 2008).

Las proteínas intrínsecamente desordenadas, por definición son no estructuradas y dinámicas en forma aislada (Adamski *et al.*, 2019). Sin embargo, cuando se encuentran y se unen a uno de sus blancos fisiológicos, se suele desencadenar un cambio conformacional.

A diferencia de las proteínas globulares, pueden someterse a un plegado con plantilla (*templated folding*, en inglés) al unirse con sus blancos, lo que lleva a una conformación mejor definida en el estado unido, lo cual indica que en un contexto funcional las IDPs pueden adquirir un estructura bien definida (Wright y Dyson, 2009). El plegado con plantilla se puede describir mediante un paisaje de energía libre similar a un embudo, que se forma a

partir de interacciones intramoleculares e intermoleculares, en contraste con las proteínas de plegado autónomo, cuyo embudo puede generarse únicamente mediante interacciones intramoleculares. Se cree que las interacciones intermoleculares de las proteínas desordenadas con sus blancos complementan con el paisaje rugoso que surgiría solo de sus interacciones intramoleculares.

Los modos de unión observados de las proteínas desordenadas van desde ordenarse casi por completo, lo cual se denomina una transición de "desordenada a ordenada" (*disorder-to-order*, en inglés) hasta formar estados más bien desordenados en el complejo unido, lo cual se conoce como una transición de "desordenada a desordenada" (*disorder-to-disorder*, en inglés). Las estructuras también pueden cambiar a través de una modificación postraduccional, por la variación de las condiciones celulares o dependiendo del blanco al que se unen (Miskei *et al.*, 2020), en este caso de denomina como "plegado condicional" (*conditional folding*, en inglés). Muchas de las proteínas desordenadas a menudo muestran diferentes estructuras cuando están unidas a diferentes blancos este fenómeno se lo denomina *fuzzy binding* (Fuxreiter, 2018). El *fuzzy binding* permite que las proteínas desordenadas interactúen no con todas las biomoléculas, sino específicamente solo con un conjunto definido de blancos. La base física de esta promiscuidad controlada aún no ha sido revelada. Todas estas observaciones sugieren la idea de que las interacciones de las proteínas desordenadas pueden ser *fuzzy* y que su versatilidad funcional explota la diversidad de muchos subestados diferentes (Tompa y Fuxreiter, 2008).

Todas las proteínas están gobernadas por dos aspectos, por un lado su habilidad de plegarse y por el otro su capacidad de llevar a cabo una función específica. Aunque en algunos casos estos aspectos se contraponen en la mayoría de las veces llegan a un equilibrio. Como ya se mencionó, lo largo de los años se han estudiado los patrones de frustración de varios aspectos funcionales de las proteínas y en todos los casos se demostró que los sitios analizados están enriquecidos de interacciones altamente frustradas (Ferreiro *et al.*, 2007, 2014; Freiberger *et al.*, 2019; Gianni *et al.*, 2014).

Todos estos análisis arriba mencionados se realizaron en proteínas globulares. Pero, ¿Qué sucede para el caso de las proteínas desordendas? ¿Tienen el mismo comportamiento, en términos energéticos, que las proteínas globulares? Es decir, el uso de la teoría del paisaje

energético de las proteínas ¿Se puede aplicar también para estudiar los complejos de proteínas desordenadas?. Las proteínas desordenadas con frecuencia se las describe como interactores que pueden unir una variedad limitada de moléculas denominadas blancos. Los complejos con diferentes blancos suelen presentar modos de unión distintos, en los que intervienen regiones que permanecen desordenadas en el estado unido. Aunque se ha establecido la importancia biológica de la *fuzzyness* (Fuxreiter, 2018; Sharma *et al.*, 2015), comprender cómo se reconcilian la diversidad y la especificidad requiere la aplicación cuantitativa de la teoría del paisaje energético. La intuición de que las interacciones mediadas por regiones desordenadas siempre deben ser débiles se contradice con la existencia de complejos de proteínas desordenadas con altas afinidades (Borgia *et al.*, 2018).
En este capítulo aplicaremos el concepto de frustración con el objetivo de entender la dinámica de las interacciones *fuzzy* y demostrar que es posible aplicar de la teoría del paisaje energético de las proteínas a los complejos de proteínas desordenadas.

Base de datos de estructuras

Para este análisis seleccionamos un conjunto de datos aplicando el protocolo descripto en (Horvath *et al.*, 2020; Miskei *et al.*, 2020).
Las regiones *fuzzy* se clasifican en tres grupos según el tipo de unión a otra molécula.
***DORs*, modo de unión de desordenada a ordenada**: son estructuras que tienen al menos una región que sufren una transición de desordenada a ordenada cuando se unen a un blanco específico. Esto significa que hay un cambio en la conformación de la proteína en respuesta a la unión de otra molécula, en este caso pasan de tener una conformación desordenada a una ordenada. Para encontrar proteínas que tengan este tipo de transición recolectamos estructuras cristalinas del PDB con una resolución superior a 3 Å, pero que tengan una densidad de electrones faltante para al menos cinco residuos consecutivos. Excluimos las secuencias de proteínas con modificaciones postraduccionales o que contenían aminoácidos no estándar. Luego recolectamos todas las estructuras en estado unido disponibles en PDB que involucren la región desordenada en todos los complejos. Analizamos los residuos de la interfaz y seleccionamos aquellas regiones en las que al menos 1 residuo medie una interacción interproteica (dentro de 4,5 Ådesde la interfaz). Para el análisis de g(r) se

seleccionó un dataset no redundante de 83 complejos, para mayor detalle sobre el calculo de la g(r) ver los métodos.

***DDRs*, modo de unión de desordenada a desordenada**: este tipo de estructura contienen una o más regiones que sufren una transición de desordenada a desordenada cuando se unen a un blanco específico. Es decir que la transición ocurre cuando las regiones desordenadas aún exhiben heterogeneidad conformacional en los estados ligados, ya sea plegándose en conformaciones alternativas (Gógl *et al.*, 2016) o fluctuando mientras interactúan con sus blancos (Lukhele *et al.*, 2013). Debido a que para hacer análisis de frustración se requiere contar con la estructura completa de la proteína así como de la región a analizar, este grupo fue excluido.

***CDRs*, modo de unión que depende del contexto**: son estructuras que tienen una multiplicidad de modos de unión de regiones de desordenada a ordenada (DOR) y regiones de desordenada a desordenada (DDR), que se observaron solo en un estado (ya sea ordenado o desordenado) en sus complejos, los cuales dependen del blanco o del contexto. Recolectamos de PDB estructuras con regiones dependiente del contexto, con una longitud mínima de cinco residuos consecutivos, es decir estructuras cuyas regiones *fuzzy* presentan ambos modos de transición DOR y DDR, el modo de transición depende del blanco al que se unen. Este conjunto de datos contiene 93 regiones desordenadas no redundantes, representadas en formas ordenadas y desordenadas en 750 estructuras complejas (1505 cadenas), seleccionamos solamente un conjunto de datos no redundantes de 77 complejos.

Las tablas con la información completa sobre las estructuras utilizadas y las anotaciones correspondientes a las regiones *fuzzy* para ambas clases (DORs y CDRs) se encuentran en el Anexo en la sección Proteínas *Fuzzy*.

Caracterización de los patrones de frustración de las regiones *fuzzy*

Para poder caracterizar los patrones energéticos de las proteínas seleccionadas para este estudio, primero calculamos la frustración de todos los monómeros y complejos proteicos de las proteínas usando el paquete de R del *Frustratometer* (Rausch *et al.*, 2021) (ver métodos). Para evaluar la densidad de frustración en las regiones *fuzzy* utilizamos la función de distribución radial (*g(r)*) (ver métodos).

Analizamos sistemáticamente la frustración en los estados libre y unido de 160 proteínas que forman complejos *fuzzy*. Dividimos el conjunto de datos según el modo de unión, es decir, proteínas cuyo estado de transición, luego de la unión, es de desordenada a ordenada (DORs) y las que tienen una multiplicidad de modos de unión (CDRs) aquellas que la transición de la región desordenada depende del contexto y de la molécula a la que se unan.

Proteínas con transición desordenada a ordenada (DORs)

Para las proteínas DORs (Fig. 4.1), se utilizó un total de 83 estructuras de complejos proteicos que contienen 97 regiones *fuzzy* no redundantes. En la figura 4.1A se muestra un ejemplo de los patrones de frustración local (frustratograma) en un complejo proteico con sus regiones *fuzzy* coloreadas en amarillo. En la figura 4.1B se muestra la *g(r)*, para el índice de frustración configurational. La *g(r)* fue calculada usando como partículas de referencia los C_α de los residuos de la región *fuzzy*. La línea verde representa los contactos mínimamente frustrados, la roja los contactos altamente frustrado, los contactos neutros se representan en gris y la línea negra representa a todos los contactos. Para todos los tipos de contacto los valores de *g(r)* se normalizaron de manera que g(20) = 1.

Para el índice *configurational* (Fig. 4.1B (izquierda)), pudimos observar que las regiones de proteínas que originalmente se encuentran desordenadas en la estructura, pero que ahora adoptan una estructura bien definida al unirse a una molécula, aún exhiben interacciones altamente frustradas entre 2 y 4 Å. Por otro lado, la densidad de contactos mínimamente frustrados en las regiones *fuzzy* es mucho menor. Estos resultados indican que el plegado de las regiones desordenadas, tras la unión, a menudo está lejos de ser óptimo. También hemos encontrado que las proteínas DORs, para el índice *mutational* (Fig. 4.1C (izquierda)), muestran un enriquecimiento de interacciones altamente frustradas en las regiones *fuzzy*, pero en este caso el enriquecimiento es menor que en el configurational.

Para demostrar que el enriquecimiento de contactos altamente frustrados es característica de las regiones *fuzzy*, se seleccionaron de forma aleatoria residuos que no sean residuos definidos como *fuzzy*. Las interacciones encontradas en estas regiones estructuradas de las mismas proteínas (elegidas como controles aleatorios), también muestran un enriquecimiento de con-

tactos altamente frustrados, pero este enriquecimiento es mucho menor que el visto en las regiones desordenadas (Fig. 4.1B (derecha)). La frustración de las interacciones de las regiones ordenadas sigue siendo significativamente mayor que la que se suele encontrar en los complejos formados a partir de proteínas totalmente estructuradas (Ferreiro *et al.*, 2007). Estos resultados indican que el plegado de las proteínas DOR en las regiones *fuzzy* también imponen restricciones en la parte plegada de la proteína. Luego comparamos el valor de la

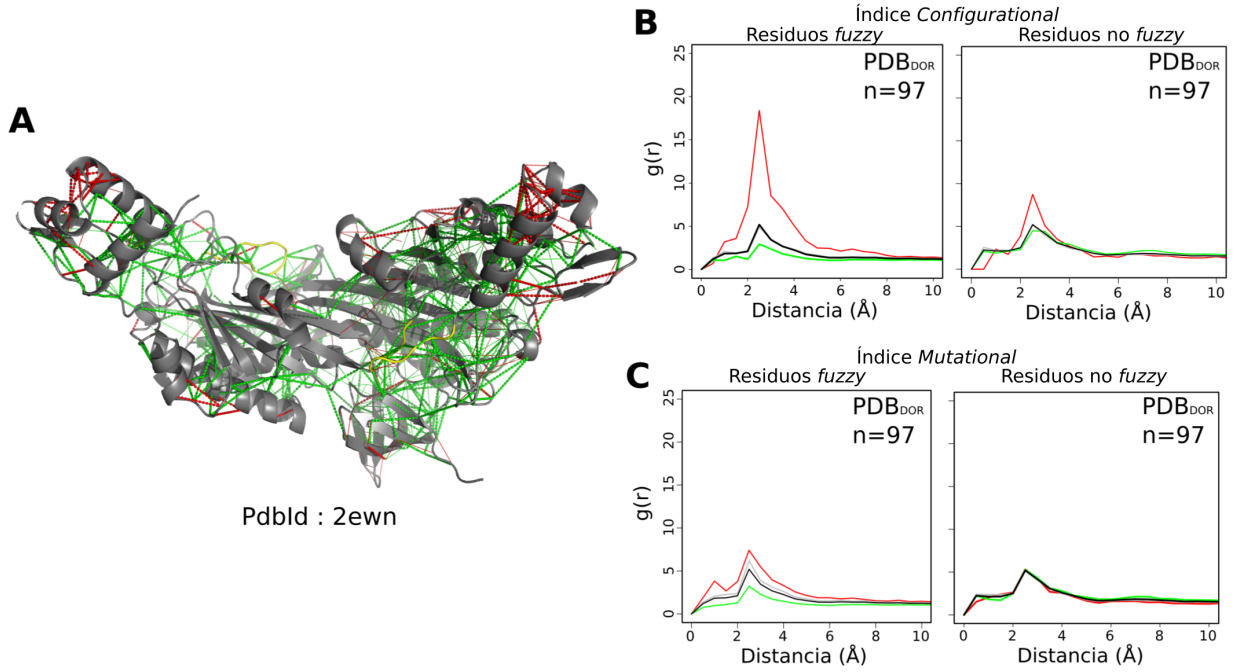

Figura 4.1: Frustración local en complejos de proteínas desordenadas donde las regiones experimentan una transición de un estado desordenado a uno ordenado. A) Ejemplos de patrones de frustración en una proteína que experimenta una transición de desordenada a ordenada luego de la unión. El esqueleto de la proteína se muestra en *cartoon* en color gris, los contactos mínimamente frustrados se representan con líneas verdes, las interacciones altamente frustradas se representan con líneas rojas. Las interacciones neutras se omitieron para mejor visualización. La región de transición de desordenada a ordenada está coloreada de amarillo. B y C) A la izquierda la función de distribución radial de los contactos entre la proteína y los residuos en la región transición de desordenada a ordenada. A la derecha mostramos la función de distribución radial de los contactos entre residuos de regiones estructuradas que no son residuos *fuzzy*. En verde se representan los contactos mínimamente frustrados; en rojo los altamente frustrado; en gris los contactos neutros y en negro todos los contactos. En todos los casos, los valores de g(r) se normalizaron de manera que g(20) = 1, B) índice configurational. C) índice mutational.

frustración de aquellos residuos que están involucrados en la interfaz de unión (unión) con

aquellos que no median interacciones intermoleculares (no unión). Se consideran interacciones altamente frustradas si el índice de frustración local es menor a -1 (ver métodos). La figura 4.2 compara la densidad del índice de frustración *configurational* para residuos *fuzzy* involucrados en contactos de unión (azul) y de no unión intermolecular (rosa), en los complejos DOR. Observamos que aquellos residuos que no forman contactos con el blanco exhiben un índice de frustración más alto que aquellos que forman directamente contactos intermoleculares (Fig. 4.2), lo que indica que la unión en sí disminuye la frustración de las proteínas desordenadas. Estos resultados indican que el plegado de las regiones desordenadas es menos óptimo que sus interacciones de interfaz frustradas.

Debido a que las regiones *fuzzy* de las proteínas analizadas están desordenadas en forma

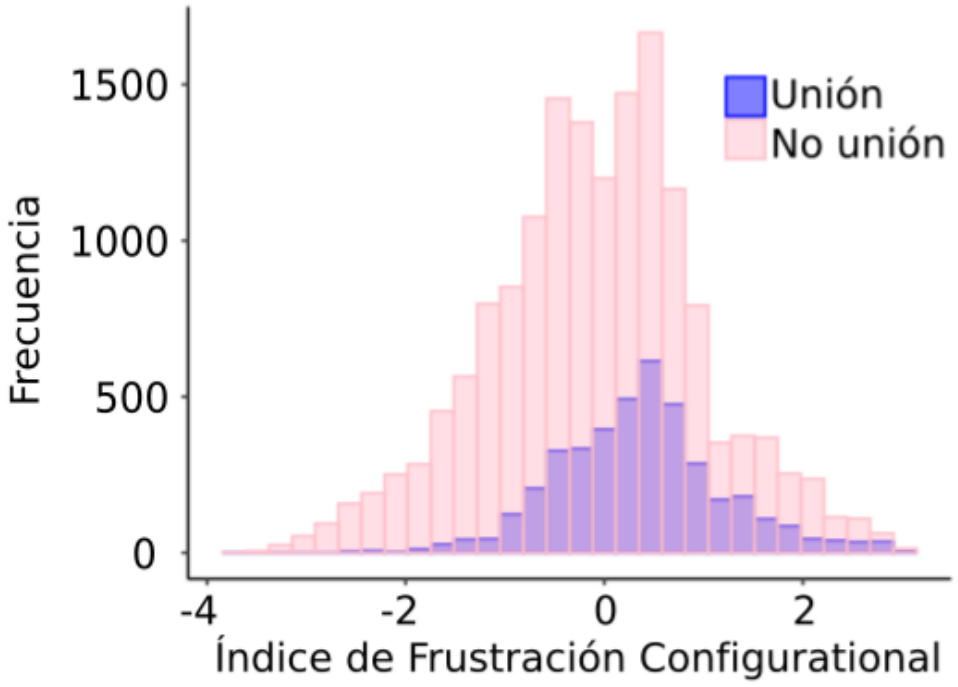

Figura 4.2: Distribución de índices de frustración local para contactos de unión (rosado) y para contactos de no unión (azul) de los residuos *fuzzy* de proteínas del tipo DOR. Las interacciones altamente frustradas se definen por un índice de frustración inferior a -1, las neutras entre los valores de -1 y 0,78 y las mínimamente frustradas valores mayores a 0,78.

libre (no unida), no podemos calcular su frustración en el monómero porque esta región está ausente en la estructura. A modo de poder comparar la frustración del complejo con la del estado libre, simplemente removimos los blancos de la estructura y calculamos su frustra-

ción. En la figura 4.3 se muestra la correlación del índice *configurational* del complejo y del monómero. Para los contactos entre residuos que forman de las regiones *fuzzy* (Fig. 4.3A), sobre la diagonal principal podemos ver que hay muchos contactos que mantienen la misma frustración en el monómero y en el complejo. Por debajo de la diagonal principal, vemos una línea de puntos paralela a esta, lo que indica que hay una disminución de la frustración en el complejo y esa disminución es constante para todos esos contactos, el valor de la disminución de la frustración es cercano a 2, lo que indica que probablemente la clasificación del índice de frustración para esos contactos este cambiando. Además podemos ver que hay una disminución de la frustración local en el complejo con respecto al monómero. Lo mismo vemos en los residuos de regiones estructuradas (Fig.4.3B). Lo que indica que las interacciones con el blanco reducen la frustración de los contactos, es decir que la unión de un blanco a una región *fuzzy* produce cambios que benefician a la estabilidad de la proteína.

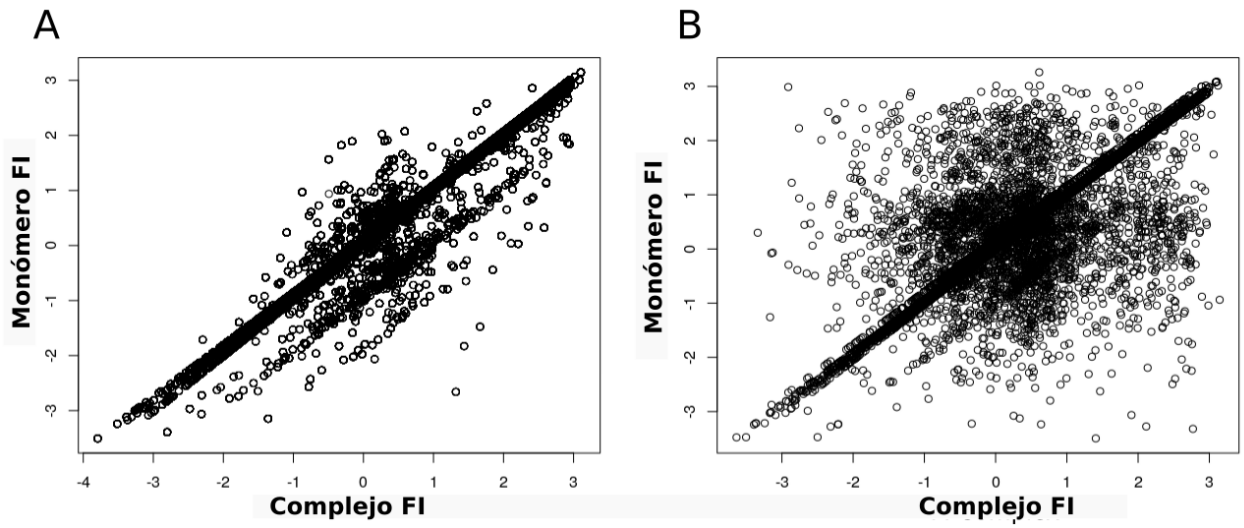

Figura 4.3: Correlación entre los índices de frustración *configurational* en las forma monomérica y unida para los contactos A) residuos involucrados en las regiones desordenadas (DOR), B) residuos de regiones estructuradas.

Proteínas con transición desordenada a ordenada (CDRs)

Para las proteínas CDRs, se utilizó un total de 77 estructuras de complejos proteicos que contienen 93 regiones *fuzzy* no redundantes. La creciente evidencia experimental indica que

las interacciones frustradas en las proteínas desordenadas a menudo se manifiestan formando complejos ordenados con algunos blancos pero formando complejos desordenados con otros blancos (Miskei *et al.*, 2020; Stamos *et al.*, 2014), a lo que llamamos "plegado condicional". El plegado de las proteínas desordenadas que depende del contexto de unión, también pueden depender del tipo de blanco al que se está uniendo, a modificaciones postraduccionales o las condiciones celulares (Korennykh *et al.*, 2009). De las proteínas analizadas, encontramos que estos complejos exhiben interacciones altamente frustradas (Fig. 4.5A) de manera similar a las DOR (Fig. 4.1A). Las distribuciones de los índices de frustración *mutational* se muestran en la Figura 4.5C. Estos indican un pequeño enriquecimiento de interacciones altamente frustradas alrededor de las regiones *fuzzy*, relativo a la topología de la proteína (líneas negras). Los contactos altamente frustrados se pueden encontrar tanto en las regiones estructuradas de las proteínas (Fig. 4.5A) como en las regiones *fuzzy* fuera de la interfaz de unión (Fig. 4.4). Por lo tanto, variar los grados de plegados con diferentes blancos también da como resultado interacciones subóptimas en el estado unido.

Al igual que para las proteínas clasificadas como DOR, usando la misma estrategia de remover los blancos de las proteínas, calculamos la correlación de la frustración del monómero y del complejo (fig 4.6). También observamos que en forma de complejo la frustración local es menor que en el monómero, indicando que la unión con de la proteína con el blanco reduce la frustración, como vimos en la sección anterior en la figura 4.3.

De lo observado anteriormente sobre las proteínas CDRs, surge la siguiente pregunta, ¿Cómo es el patrón de frustración de la misma proteína unida a diferentes blancos? Por lo tanto, planteamos que la frustración específica del blanco facilita la selección del mismo. Para responder a la pregunta, de que si la frustración cambia según el blanco al que se unen la regiones *fuzzy*, examinamos algunos complejos en los que la misma región desordenada interactúa con diferentes blancos de unión.

En la figura 4.7A podemos ver la unión diferencial de las regiones *fuzzy* con los residuos 39–47 (en amarillo) de la subunidad gamma del factor 2 de iniciación de la traducción (UniprotID: Q980A5). En las estructuras proteicas, PdbID: 3cw2 y PdbID: 3i1f, podemos ver que eif2g se pliega en dos conformaciones diferentes (Fig. 4.7A). Mientras que la estructura de la región *fuzzy* en 3cw2 está estabilizada por las interacciones intramoleculares entre Glu39-Thr46 y

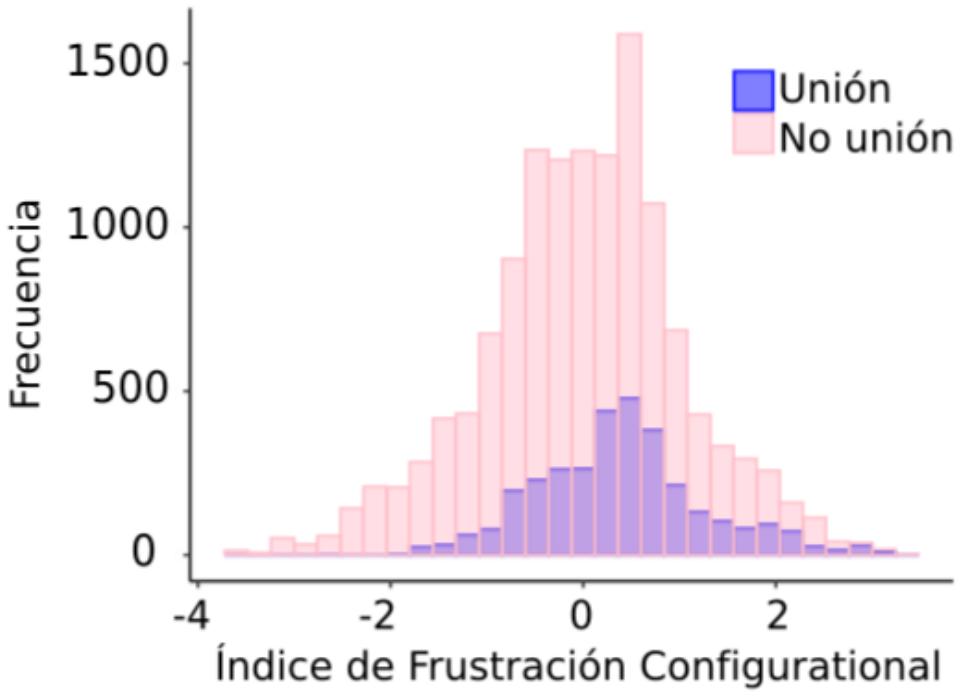

Figura 4.4: Distribución de índices de frustración local para contactos de unión (rosado) y para contactos de no unión (azul) de los residuos *fuzzy* de proteínas del tipo DOR. Las interacciones altamente frustradas se definen por un índice de frustración inferior a -1, las neutras entre los valores de -1 y 0,78 y las mínimamente frustradas valores mayores a 0,78.

Glu40-Gly44, la estructura 3i1f está estabilizada por una interacción entre Glu39 y Arg43, En la estructura 3i1f, la cadena principal Gly44 forma un enlace de hidrógeno con una cadena lateral Lys42 mientras que en la estructura 3cw2 Lys42 interactúa con Asp-283 del dominio estructurado.

La figura 4.7A muestra las estructuras y los patrones de frustración locales para estas diferentes estructuras proteicas. En general, ambas estructuras poseen una red extendida de interacciones mínimamente frustradas, con parches de interacciones altamente frustradas en la superficie. Cada estructura muestra diferentes patrones de frustración para la región *fuzzy* de los complejos. Vemos que en las estructuras alternativas se han elegido diferentes formas de resolver los conflictos energéticos. Este intercambio de interacciones frustradas se visualiza en los mapas de contactos (Fig 4.7A).

Otro ejemplo (Fig. 4.7B), que ilustra la naturaleza de la unión *fuzzy* es la región de residuos 369–382 de la proteína quinasa 10 activada por mitógeno (UniprotID: P53779). En la estruc-

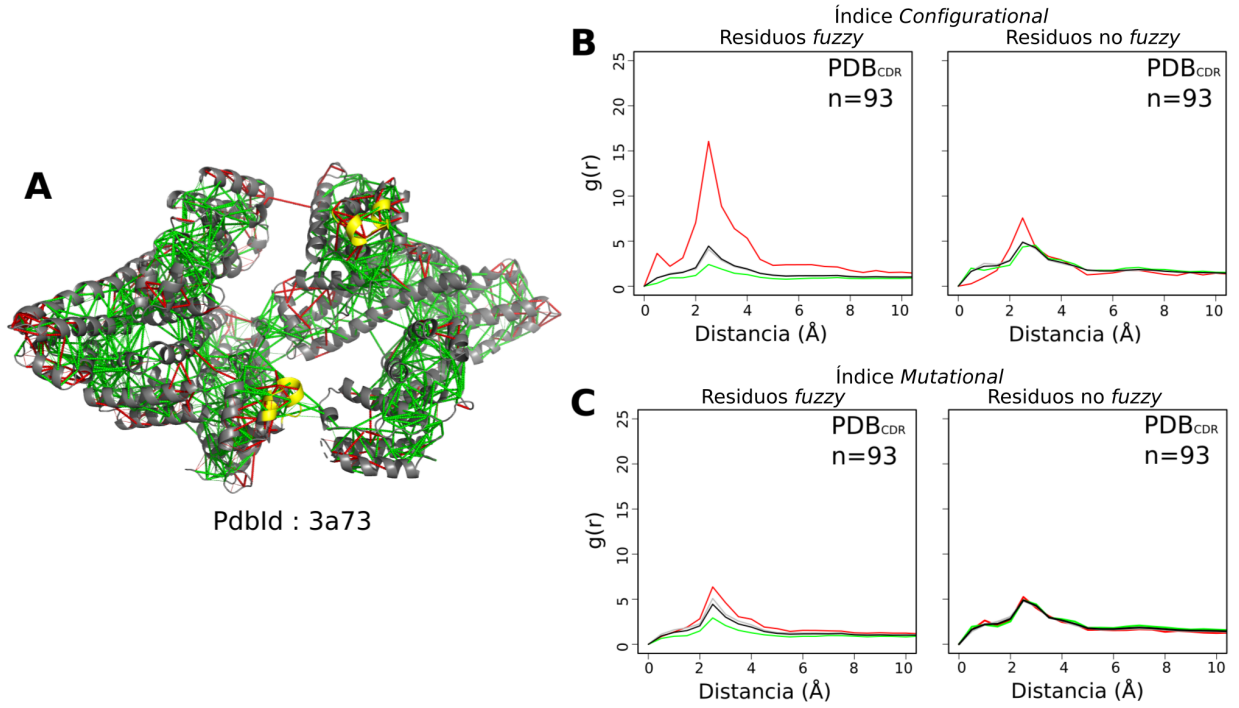

Figura 4.5: Frustración local en complejos de proteínas desordenadas donde las regiones experimentan una multiplicidad de modos de unión. A) Ejemplos de patrones de frustración en una proteína que experimenta una transición de desordenada a ordenada luego de la unión. El esqueleto de la proteína se muestra en *cartoon* en color gris, los contactos mínimamente frustrados se representan con líneas verdes, las interacciones altamente frustradas se representan con líneas rojas. Las interacciones neutras se omitieron para mejor visualización. La región de transición de desordenada a ordenada está coloreada de amarillo. B y C) A la izquierda la función de distribución radial de los contactos entre la proteína y los residuos en la región transición de desordenada a ordenada. A la derecha mostramos la función de distribución radial de los contactos entre residuos de regiones estructuradas que no son residuos *fuzzy*. En verde se representan los contactos mínimamente frustrados; en rojo los altamente frustrado; en gris los contactos neutros y en negro todos los contactos. En todos los casos, los valores de g(r) se normalizaron de manera que g(20) = 1, B) índice configurational. C) índice mutational.

tura PdbID: 4h3b, el plegado de la región *fuzzy* está estabilizado por muchas interacciones intramoleculares. Como podemos observar en el mapa de contactos, algunas de estas interacciones se forman entre cadenas laterales con átomos de la cadena principal, por ejemplo, Gln374 y Pro372, Gln379 y Leu380, o Glu382 NE2 y Glu382 C–O. Por el contrario, en la estructura PdbID: 3v6r, hay muchos menos contactos intramoleculares (Glu369, Leu380 y Asp381).

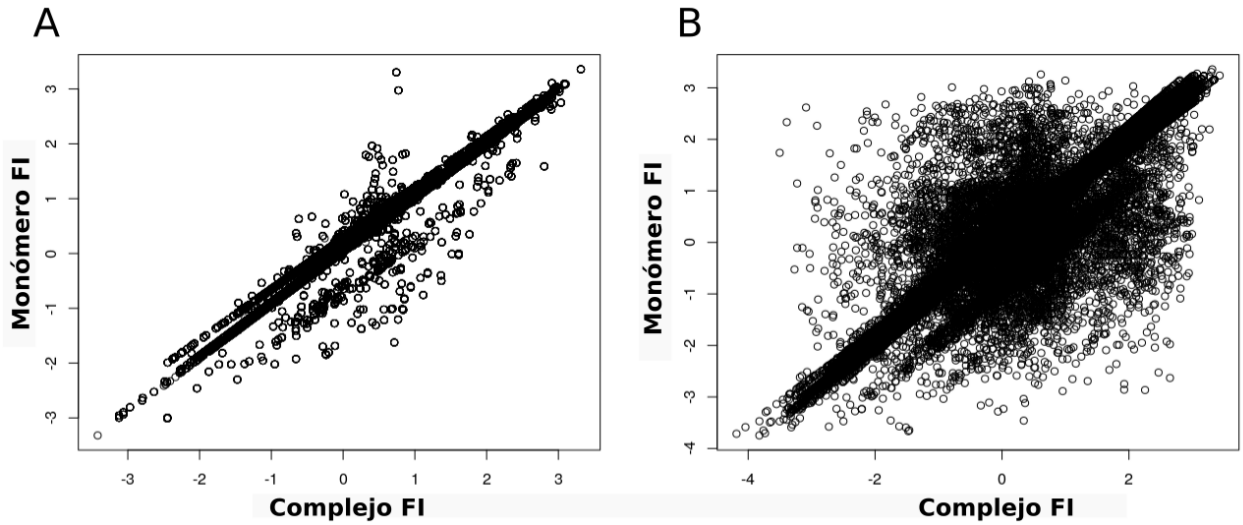

Figura 4.6: Correlación entre los índices de frustración *configurational* en las formas monomérica y unida para los contactos A) residuos involucrados en las regiones desordenadas (CDR), B) residuos de regiones estructuradas.

4.1. Conclusiones del capítulo

El carácter estructurado de los complejos de proteínas desordenadas con sus blancos específicos ha llevado a la intuición errónea de que, al final, el funcionamiento requiere siempre una única conformación bien definida. La presencia de una estructura bien definida, sin embargo, no se correlaciona con la afinidad de las interacciones. En este trabajo hemos realizado un análisis sistemático de los complejos de muchas regiones proteicas desordenadas. Demostrando que, incluso después de la unión, la energética de la interacción dista mucho de ser óptima en las regiones desordenadas, de acuerdo con los datos experimentales (Hadži *et al.*, 2017). En consonancia con resultados anteriores para casos individuales (Toto *et al.*, 2016), hemos encontrado que tanto las regiones desordenadas como las estructuradas de los complejos están enriquecidas en interacciones altamente frustradas en los complejos unidos de proteínas desordenadas. Los contactos de interfaz disminuyen el nivel de frustración en la proteína desordenada una vez unida en comparación con la frustración del estado libre, pero las interacciones a menudo siguen siendo subóptimas y quedan conflictos energéticos por resolver (Figuras 4.3 y 4.6). Estos resultados corroboran la rugosidad del paisaje energético que describen los complejos de regiones desordenadas. Ilustramos a través de dos ejemplos, que las regiones desordenadas muestran distintos patrones de frustración con diferentes blancos, racionalizan-

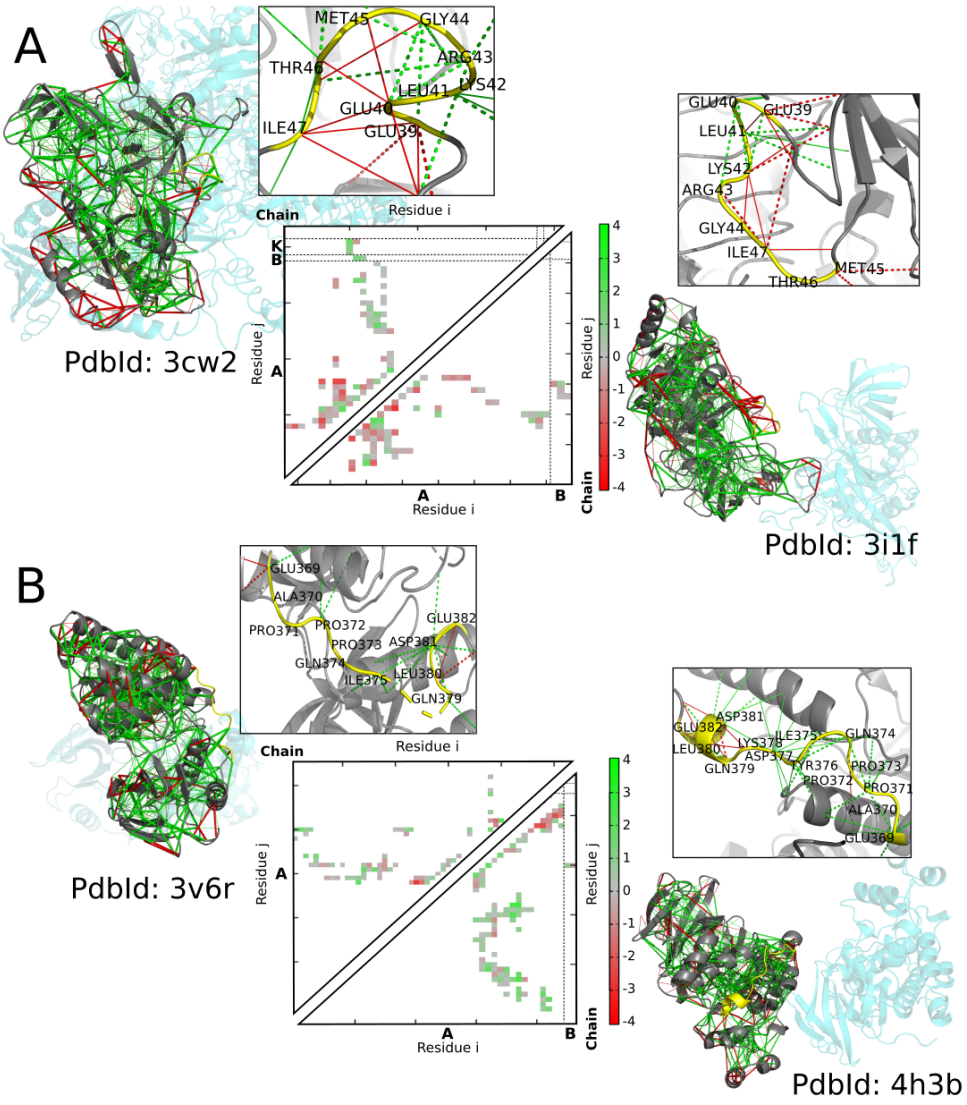

Figura 4.7: A) Estructuras para la subunidad gamma del factor 2 de iniciación de la traducción (eif2g), PdbID: 3cw2 (arriba) y PdbID: 3i1f (abajo). Mapa de contacto de 3cw2 (arriba de la diagonal) y 3i1f (debajo de la diagonal). B) Estructura de la proteína para la proteína quinasa 10 activada por mitógeno, PdbID: 3v6r (arriba) y PdbID: 4h3b (abajo). Mapa de contacto de 3v6r (arriba de la diagonal) y 4h3b (debajo de la diagonal). Los patrones de frustración locales de la proteína, con las interacciones mínimamente frustradas mostradas en verde, las neutras mostradas en gris y las interacciones altamente frustradas mostradas en rojo. La región *fuzzy* se muestra con el esqueleto en amarillo. Para el mapa de contactos: verde, contactos mínimamente frustrados; rojo, altamente frustrado; contactos neutros en gris.

do cómo la frustración local permite definir tanto la especificidad como la versatilidad.

El plegado y la unión acoplados de las regiones desordenadas conducen a contactos subópti-

mos, lo que permite la unión a diferentes blancos. La alta frustración local explica por qué las regiones desordenadas son capaces de manifestar varios modos de unión diferente (Horvath *et al.*, 2020). Varias observaciones sugieren que los requisitos contrastados del plegado físico y la función biológica pueden dar lugar a una frustración energética local en las proteínas naturales. La heterogeneidad conformacional fomenta la adaptabilidad, que a menudo puede ser objeto de selección evolutiva. La robustez del paisaje puede ser modulada por las interacciones con blancos fisiológicos. Los contactos subóptimos con el blanco suelen dar lugar a uniones *fuzzy* y dan lugar a complejos proteicos con modos de unión alternativos y específicos del blanco.

Capítulo 5

Análisis genómico de las proteínas con repeticiones de Ankirina en eucariotas

En este capítulo estudiaremos proteínas que contienen repeticiones de ankirina (AR), ya que son un buen modelo para estudiar cómo evolucionan los dominios proteicos. Se cree que estas proteínas han evolucionado por transferencia horizontal, porque algunas proteínas AR encontradas en organismos procariotas tienen una alta identidad de secuencia con AR encontradas en animales (Al-Khodor *et al.*, 2010; Bork, 1993). También se cree que estas proteínas evolucionaron por duplicación y deleción de repeticiones internas dando lugar a proteínas de longitud variable (Andrade *et al.*, 2001; Björklund *et al.*, 2006; Pâques *et al.*, 1998; Schüler y Bornberg-Bauer, 2016). En los organismos eucariotas, la estructura de los genes viene determinada por la organización exón-intrón, la clase de los exones y la fase de los intrones. Se ha demostrado que la arquitectura exón-intrón, está más conservada que la secuencia. Asimismo, el análisis de la arquitectura exón-intrón es una clave importante para estudiar la evolución de las proteínas así como para detectar homólogos (Betts *et al.*, 2001). Una forma de caracterizar la estructura de los genes de una familia de proteínas es conocer cómo es el largo de los exones e intrones y su clase y fase respectivamente. Es importante caracterizar la estructura exón-intrón de los genes, para poder responder preguntas acerca de cuáles son los procesos y mecanismos evolutivos, que actúan sobre las proteínas, que dan lugar a nuevos genes. En este capítulo caracterizaremos la estructura exón-intrón de los genes de proteínas que contienen repeticiones de ankirinas y cómo están distribuidos las repeticiones

en los exones. También analizaremos los posibles mecanismos por los cuales estas proteínas evolucionan, como lo son el barajado de exones y el empalme alternativo.

5.1. Estructura exón-intrón

En esta sección se caracterizará la estructura exón-intrón de todos los genes eucariotas anotados en nuestra base de datos construida en el marco de esta tesis (ver métodos), así como también se analizará cómo están codificadas y distribuidas las repeticiones de ankirinas en los genes. Para llevar a cabo el análisis, se seleccionó un total de 529343 exones y 460693 intrones, correspondientes a 68650 proteínas provenientes de 11662 TaxIDs diferentes. La cantidad de repeticiones de ankirinas analizadas fue de 429024.

5.1.1. Distribución del largo de exones e intrones

¿Qué longitud tienen los exones de los genes que contienen repeticiones de ankirina? ¿Cómo es la fase del intrón y la clase de exón? Para responder a estas preguntas se van a utilizar las anotaciones contenidas en la base de datos de ankirinas, que desarrollamos en marco de esta tesis (ver métodos). Dado que los genes que codifican repeticiones de ankirina (AR) pueden contener también regiones que no codifican para AR, los exones se clasificaron en exones que no codifican AR y exones que codifican AR. Aproximadamente el 48 % de los exones analizados codifican repeticiones ANK. Para caracterizar la estructura exón-intrón calculamos la longitud del exón utilizando las anotaciones EMBL de nuestra base de datos. La figura 5.1A muestra el histograma de la distribución de la longitud de los exones para las dos clases (en violeta AR y en amarillo no AR). La mayoría de los exones tienen una longitud inferior a 300 nt y vemos que hay un pico (marcado en asterisco) que indica que la frecuencia más alta es de 99 nt. La mayoría de los exones de longitud 99 nt (Fig. 5.1B) codifican para repeticiones, pero también vemos una alta frecuencia de exones que no codifican para repeticiones de ankirinas. Esto puede ser un artefacto del método utilizado para detectar repeticiones, ya que en la mayoría de los casos estos métodos no son muy exactos y puede estar fallando en la detección de algunas repeticiones. Es muy común que los métodos no detecten todas las repeticiones de ankirina, esto se debe a la diversidad estructural y secuencial de estas

repeticiones, así como a las limitaciones de los métodos de detección tradicionales, el cual sigue siendo un problema que aún no se ha resuelto por completo (Kajava, 2012).

Estos resultados indican que la longitud de exones más frecuente que codifican para repeticiones de ankirina es de 99 nt. Esta longitud de exones coincide con la longitud de una repetición de ankirina que es de aproximadamente 33 aa de longitud.

Por otra parte, realizamos un análisis de la longitud de los intrones, los cuales se clasificaron

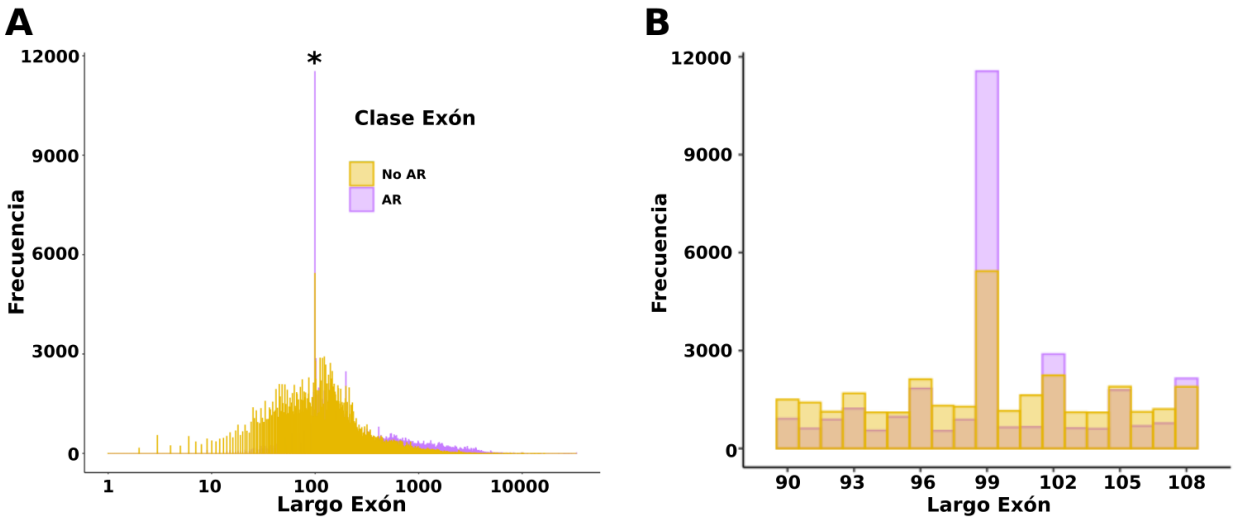

Figura 5.1: Distribución de la longitud de los exones de las proteínas de repetición de ankirina. En amarillo exón que no codifica para la repetición ANK y en violeta exón que codifica para la repetición ANK. A) Todos los exones. B) Longitud de los exones entre 90 y 108.

en función de si los exones flanqueantes codifican o no repeticiones de ankirina. Los intrones flanqueados por dos exones que codifican para repetición de ankirina son de clase A, los intrones de clase B están flanqueados por un exón que codifica para repeticiones de ankirina y los intrones de clase C por dos exones que no codifican para repeticiones de ankirina. La figura 5.2 muestra la distribución de la longitud de los intrones para cada clase de intrón, hemos encontrado que la longitud de los intrones son en su mayoría menor a 100 nt de largo. También se observa que hay una mayor abundancia de intrones de clase B y C. Estos resultados indican que la mayoría de los intrones son cortos y que las proteínas analizadas contienen regiones que no codifican repeticiones de ankirina, según nuestras anotaciones, ya que la clase más abundante es la de intrones que no están flanqueados por exones que codifican ankirinas.

Una vez caracterizada la longitud de los exones e intrones, analizaremos la fase de los intrones

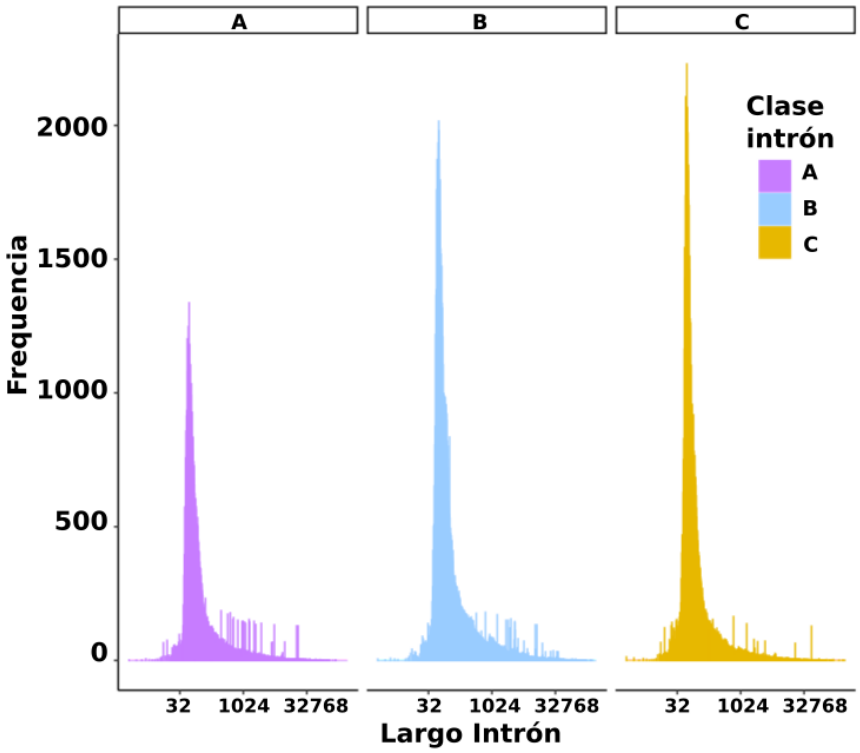

Figura 5.2: Distribución de la longitud de los intrones de los genes que codifican proteínas que contienen repeticiones de ankirina. En violeta se representan intrones flanqueados por dos exones que codifican para ANK. En celeste, intrones flanqueados por un exón que no codifica para repetición ANK y uno que si codifica para repetición ANK. En amarillo intrones flanqueados por exones que no codifican para la repetición ANK.

y la clase de los exones. En la figura 5.3A se muestra la distribución de las fases de los intrones, la mayor frecuencia es la fase 0 y el número de intrones de fase 1 y fase 2 es similar. Para la distribución de las clases de exones (Fig. 5.3B), observamos que la mayor frecuencia es la clase 0-0 estos exones se clasifican como exones simétricos. La abundancia de intrones de fase 0 podría ser una evidencia del mecanismo de barajado de exones, ya que este funciona únicamente si los intrones están todos en la misma fase (Patthy, 1996).

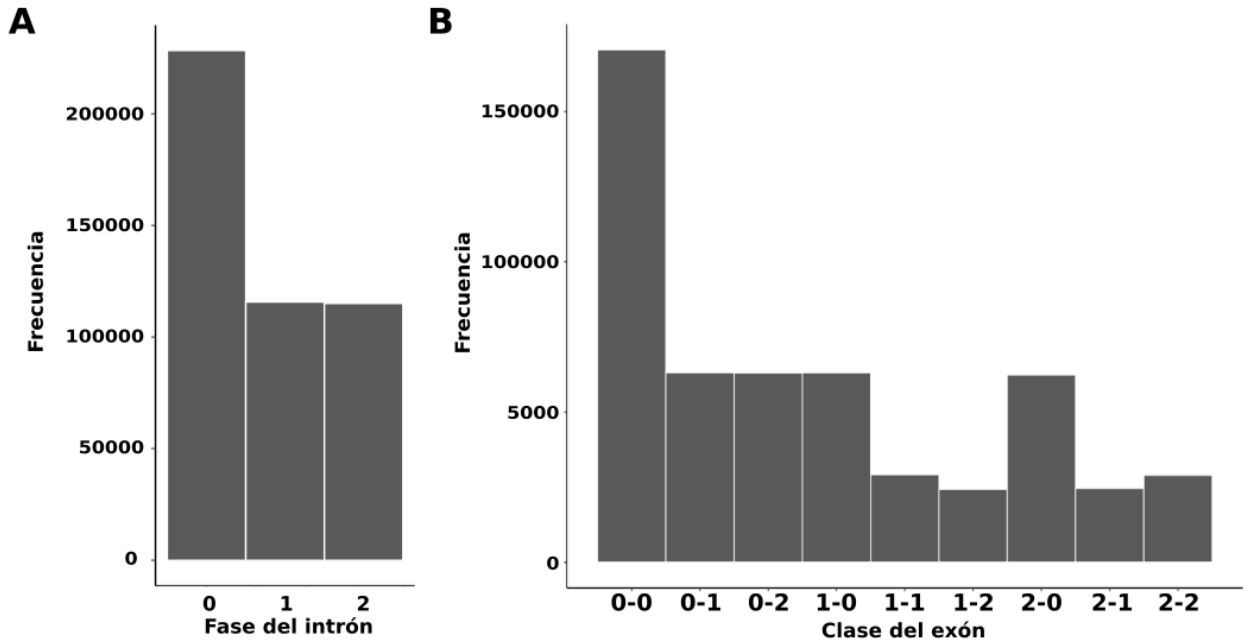

Figura 5.3: Análisis de la distribución de A) las fases de los intrones y B) de las clases de los exones de los genes que codifican proteínas que contienen repeticiones de ankirina.

5.1.2. Organización de las repeticiones de ankirina en los exones

Con el fin de caracterizar cómo se codifican las repeticiones en los exones, definimos dos clases de repeticiones, la primera denominada "Repetición Parcial" que son repeticiones codificadas por dos o más exones consecutivos y "Repetición Completa" que son repeticiones completamente codificadas por un solo exón (Fig. 5.4A). Por lo tanto, los exones pueden codificar más de una repetición completa, pero sólo un máximo de 2 repeticiones parciales, una en cada extremo del exón (Fig. 5.4A).

Se analizó un total de 429024 repeticiones, la mayoría de las repeticiones (n=302784) están codificadas por un solo exón y 126240 de las repeticiones están codificadas por dos o más

exones consecutivos. Para caracterizar la distribución de las repeticiones en los exones comparamos el número de repeticiones codificadas por exón y la longitud del mismo. Hemos encontrado que hay una alta frecuencia de exones de 99 nt de longitud que codifican para dos repeticiones parciales (Fig. 5.4B). Además se observa que a medida que aumenta la longitud del exón, aumenta la cantidad de repeticiones completas codificadas por el exón. Hemos encontrado que el número máximo de repeticiones codificadas por un exón es de 73 y la longitud del exón es de 8359 nt. En la figura 5.4C se muestra la distribución del largo de exones que codifican repeticiones parciales y completas. Para el caso de las repeticiones parciales podemos ver un pico en 99 nt y para las repeticiones completas vemos un pico en exones de largo 198 nt. Estos resultados indican que la distribución de las repeticiones de ankirinas en los exones es variada y que los exones cortos codifican para repeticiones parciales mientras que los exones más largos codifican repeticiones completas.

Encontramos que 47460 proteínas de nuestro conjunto de datos contienen al menos una repetición de ankirina codificada por dos exones consecutivos. La detección de repeticiones codificadas por dos exones consecutivos puede deberse a que las anotaciones utilizadas en este estudio de dónde empieza y termina una repetición se anotaron utilizando una fase estructural definida en (Parra *et al.*, 2015). Realizamos un análisis de la fase de aquellos intrones que se encuentran en medio de los exones que codifican para la misma repetición y también analizamos en qué posición se interrumpe la repetición. Se analizaron un total de 66438 intrones. En la figura **??** se muestra un histograma de las posiciones en las cuales el intrón interrumpe el marco de lectura de la repetición y en color se representa la fase del intrón y un logo de la secuencia de todas las repeticiones analizadas. Observamos que la posición con mayor frecuencia donde el intrón divide la repetición es en el aminoácido 7, este aminoácido no está conservado en la secuencia consenso. Además, la mayoría de los intrones, que interrumpen el marco de lectura en el residuo 7, están en fase 0, esto significa que no se interrumpe la secuencia de codificación y asegura que se mantenga el marco de lectura correcto para la traducción.

Estos resultados indican que en algunos casos la fase estructural definida en (Parra *et al.*, 2015) no se corresponde con la fase genómica. La fase genómica se refiere a la posición de inicio de una repetición de ankirina en relación con la estructura exón-intrón. Esto refiere

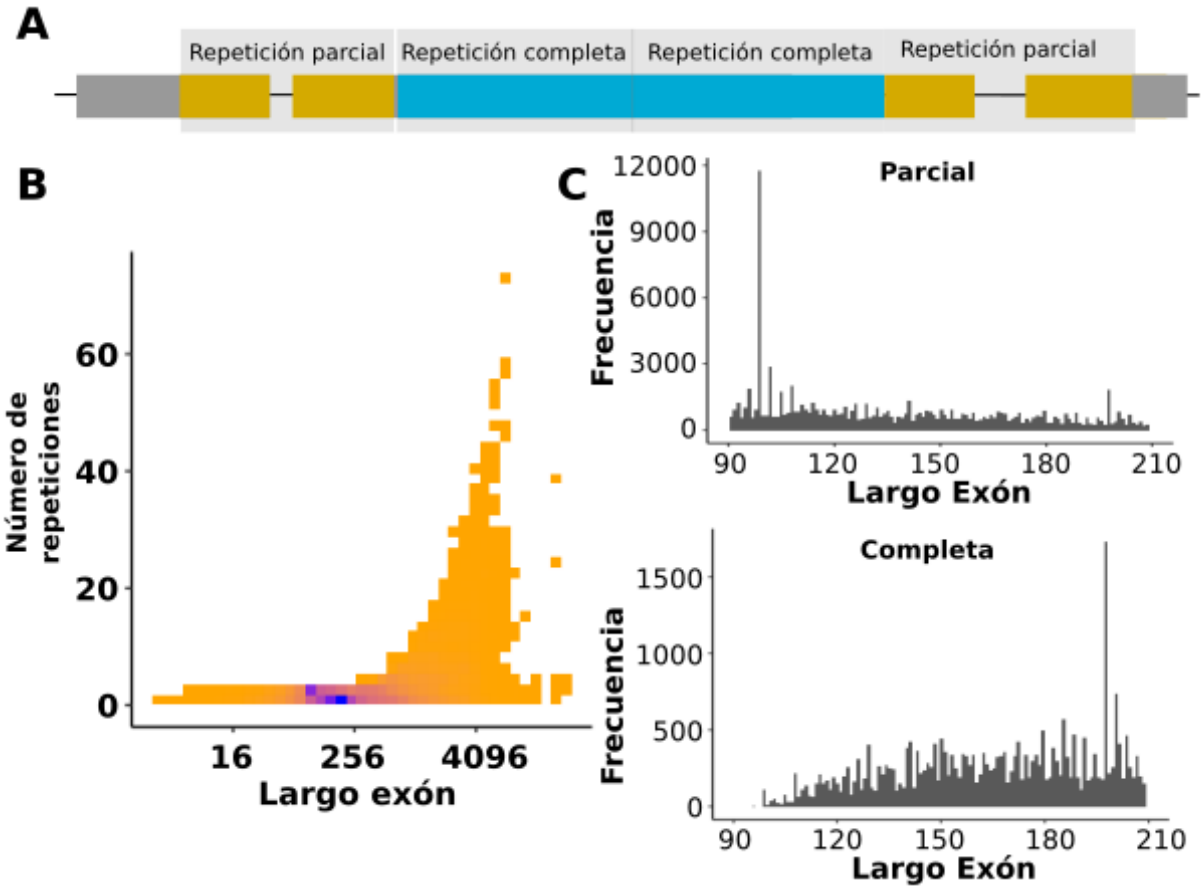

Figura 5.4: A) Esquema de la repetición parcial y completa. B) Mapa de calor para las repeticiones parciales y para las repeticiones completas. C) Histograma de la distribución del largo de exones que codifican repeticiones parciales (arriba) y repeticiones completas (abajo).

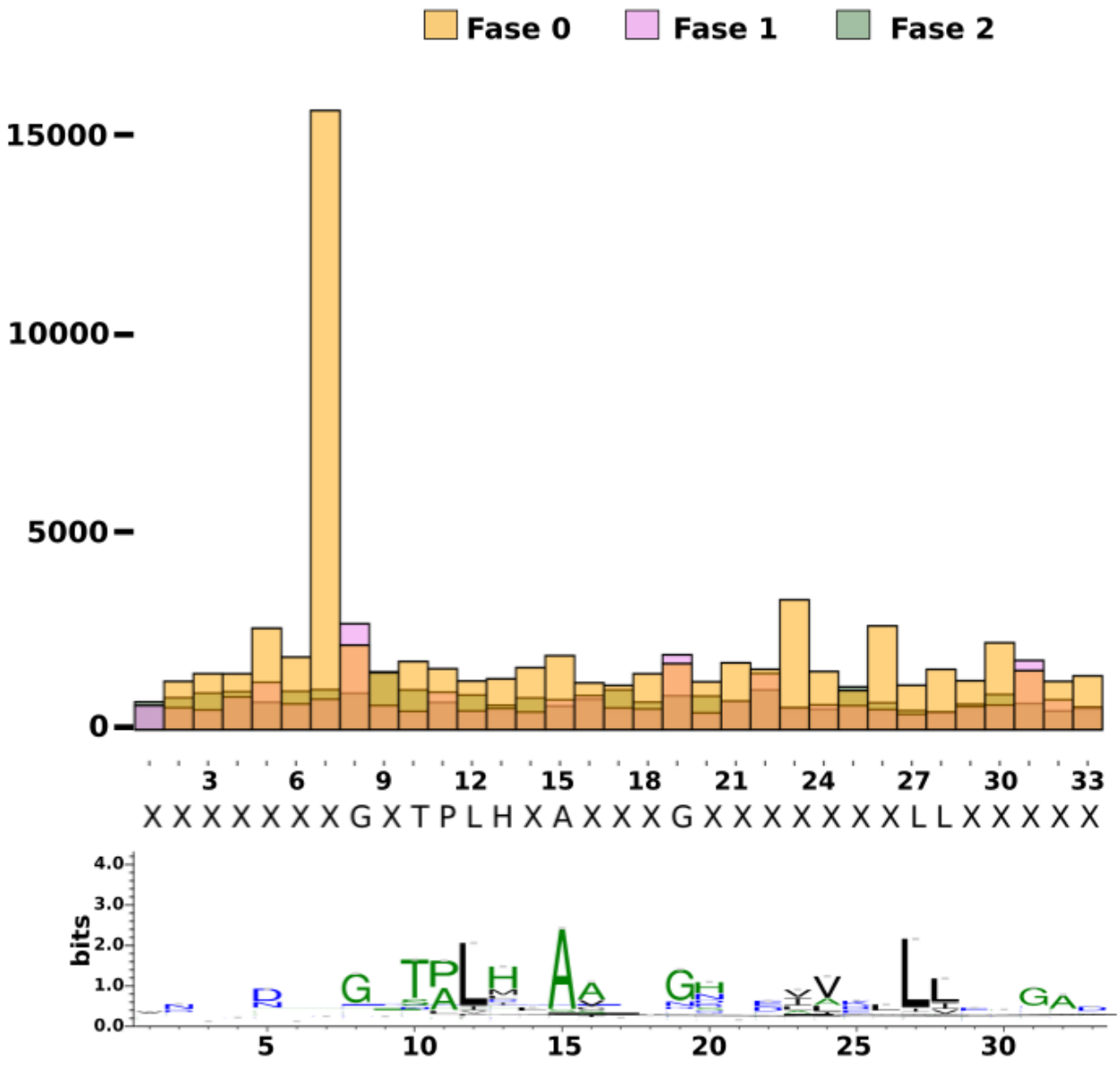

Figura 5.5: Fase del intrón y lugar en la secuencia de una repetición de ankirina, donde el intrón interrumpe el marco de lectura. En naranja intrón Fase 0, en violeta intrón Fase 1 y en verde intrón Fase 2, Abajo, el logo de secuencia de las repeticiones analizadas.

a que la fase genómica de las repeticiones de ankirina codificada por dos exones diferentes, comienza en un exón y termina en el otro. Además, la ubicación precisa de la repetición de ankirina en la fase genómica puede influir en cómo se realizan eventos de empalme alternativo. Por lo tanto, cuando decimos que en algunas repeticiones la fase estructural no se corresponde con la fase genómica, es porque la numeración de las repeticiones usadas en este análisis está definida por la fase estructural, es decir, que si ambas coincidieran no deberíamos de tener repeticiones codificadas por más de un exón. También se observa que la posición 7 de la repetición donde el intrón interrumpe el marco de lectura, en su mayoría son de fase 0, lo que significa que, en caso de que ocurra un evento de empalme alterativo no se va a ver afectado el marco de lectura. Además está posición se corresponde a un punto de inserción de aminoácidos detectado en (Galpern *et al.*, 2020).

5.1.3. Las ankirinas, ¿Evolucionan por medio del mecanismo de barajado de exones?

Como ya mencionó, el barajado de exones es un mecanismo que genera nuevos genes mediante la duplicación de un exón y su posterior inserción en un intrón de la misma fase que la clase del exón dejando una nueva estructura exón-intrón. Este mecanismo tiene ciertas limitaciones, una de ellas es que el exón que se esta duplicando se puede insertar solamente en un intrón de la misma fase que la clase del exón, por lo cual los exones simétricos son los que pueden duplicarse. En la figura 5.6 se muestra un ejemplo de un gen que evolucionó por barajado de exones. Antes de que la duplicación del exón ocurra (fig. 5.6A), el gen estaba compuesto de 4 exones, dos de clase 0-0, uno de clase 0-1 y uno de clase 1-0, es decir, dos introns de fase 0 y un intrón de fase 1, por lo tanto su arquitectura exón-intrón era 0 0 1. Luego del evento de barajado de exones (fig. 5.6B), en donde el exón que se está duplicando es el primero (en gris) y vemos que se insertó en el intrón 1 (en verde), dejando una nueva estructura exón-intrón de 0 0 0 1. Para el caso de la figura 5.6C, el exón que se duplicó es un exón simétrico pero se insertó en un intrón de fase 1, lo que significa que no cumple con las reglas del barajado de exones. Por lo tanto, este gen no evolucionó por este mecanismo, ya que los exones simétricos deben de insertarse en un intrón de la misma fase que la clase

que el exón que se duplicó. Como se ha visto en los apartados anteriores, los exones más

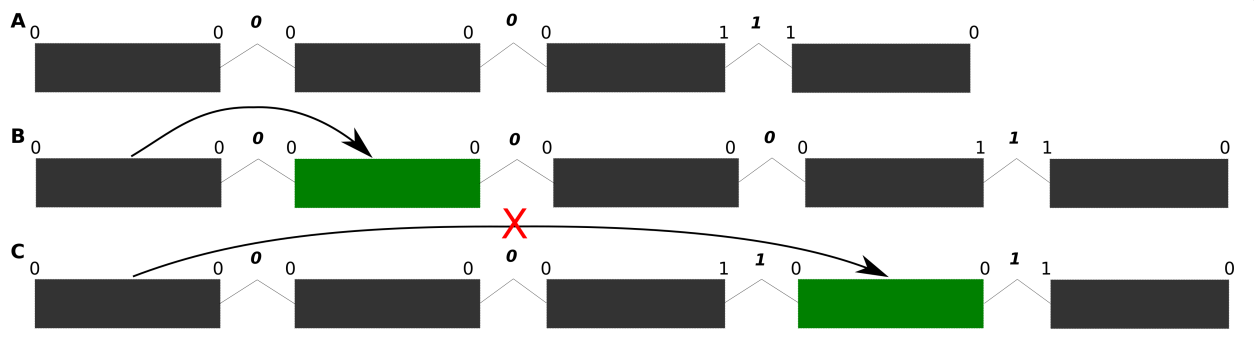

Figura 5.6: A) Esquema del gen ancestral. B) Esquema de un gen que evolucionó por barajado de exones. C) Esquema de un gen que no evolucionó por barajado de exones. En rectángulos grises se muestran los exones del gen ancestral y en verde el exón que se duplicó. Los números en *italic* representan la clase del exón y los números en *italic* y negrita representan la fase del intrón.

En lineas se representan los intrones.

frecuentes son los de longitud 99 nt, pero también hay exones de diferentes longitudes, así como también la longitud de los intrones es variada. En la sección anterior, en el análisis de la estructura exón-intrón, vimos que hay una alta frecuencia de exones simétricos de clase 0-0, lo cual nos lleva a pensar que quizás las ankirinas evolucionan por el mecanismo de barajado de exones. Para poder evaluar si estas proteínas evolucionan por este mecanismo debemos encontrar genes que tengan una arquitectura exón-intrón no aleatoria, es decir, que estén compuestos de exones de las misma clase y de clase simétrica como se vio en la figura 5.6. Primero comenzaremos caracterizando la variabilidad de la longitud de los exones e intrones y la identidad en la secuencia entre exones de un mismo gen, analizamos para cada gen la variabilidad de la longitud de sus exones e intrones (eq. 5.1) y su identidad de secuencia. Donde, Δ es la longitud del exón, α la longitud del intrón, γ la fase del intrón, f_i es la frecuencia absoluta de la longitud del exón o intrón y n es el total de exones o intrones por proteína. En este análisis, utilizando la ec. 5.1, calculamos como varía la longitud de los exones e intrones por proteína, esta variabilidad va de 0 a 1, es decir, 0 cuando todos los exones o intrones de la misma proteína tienen la misma longitud y 1 cuando todas las longitudes son diferentes. También aplicamos la ec. 5.1 para calcular la variabilidad de la fase de los intrones de un gen. Para el caso en que la frecuencia máxima sea igual a 1, la variabilidad es igual 1.

$$\text{var} = 1 - (\frac{\text{Frecuencia de la moda}}{n}) \tag{5.1}$$

En la figura 5.7A-C, se muestran la variabilidad de la longitud del exón (Δ), intrón (α) y la fase del intrón (γ) respectivamente, vemos que la mayoría de las proteínas tienen valores de variabilidad de exón e intrón cercanos a 1, esto indica que los exones e introns de la misma proteína tienen longitudes diferentes. Sin embargo, pudimos detectar algunos genes que presentan una baja variabilidad en la longitud de sus exones e introns. Consideramos que la alta variabilidad de la longitud de los exones e introns puede deberse a que el método utilizado es muy restrictivo, ya que calculamos la variabilidad sobre la longitud exacta del exón o intrón. Es necesario que la metodología sea restrictiva porque queremos evaluar si hay exones son duplicados dentro del mismo gen, por lo tanto, si un exón se duplicó e insertado en un intrón lo que se espera es que la longitud de ambos exones sea igual. Con respecto a la variabilidad de la fase de los introns, se observa que la mayoría es inferior a 0,5 y que hay algunos genes en los que la variabilidad es 0, lo que indica que todos los introns, de estos genes, son de la misma fase.

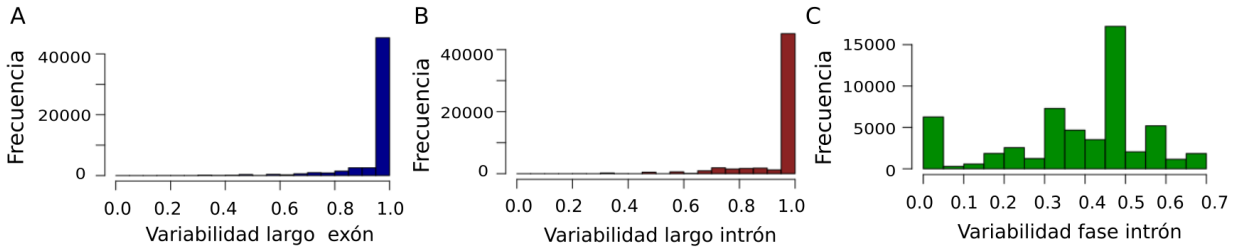

Figura 5.7: A) Distribución de la variabilidad del largo de exones (Δ) por gen. B) Distribución de la variabilidad del largo de introns (α) por gen. C) Distribución de la variabilidad de la fase de los introns (γ) por gen.

Para detectar y analizar eventos de barajado de exones de genes que codifican repeticiones de ankirinas, seleccionamos de nuestro conjunto de datos un total de 2662 genes que tienen más de 3 exones que codifican para la repeticiones de ankirina y la variabilidad de la fase del intrón (γ) es igual a 0, es decir, la fase de los introns es la misma para todos. Para medir el porcentaje de identidad en secuencia entre los exones de un mismo gen, realizamos

alineamientos de secuencias entre exones del mismo gen utilizando clustal omega (Sievers y Higgins, 2014) esta herramienta genera una matriz del porcentaje de identidad de secuencia entre pares de exones por gen. A continuación, agrupamos los genes por número de exones y calculamos una única matriz por grupo como la media del porcentaje de identidad de secuencia entre pares de exones. Sólo se seleccionaron aquellos grupos en los que el número de genes era superior a 40 (Fig. 5.8).

Para el grupo de 4 exones (Fig. 5.8A), los exones 1, 3 y 4 podemos ver que la media del porcentaje de identidad en secuencia entre ellos es alta (el porcentaje de identidad ente el exón 1 y 3 es : 85,84, entre el 1 y el 4 es: 97,43). También vemos un alto porcentaje de identidad de secuencia entre los exones 2 , 3 y 4 (el porcentaje de identidad ente el exón 2 y 3 es : 87,61, entre el 2 y el 4 es: 100). En el grupo de 8 exones, se puede observar una valor alto de media del porcentaje de la identidad en secuencia del primer exón con casi todos los exones y que los exones interiores son más similares entre sí. En el grupo de 9 exones, se ve algo similar que en el grupo de 8 exones, en donde los exones internos son los más parecidos entre sí. En el resto de los grupos, la media del porcentaje de identidad de secuencia es bajo. Estos resultados podrían estar indicando que posiblemente algunos genes de ankirinas evolucionan por eventos de barajado de exones porque vemos alto porcentaje de identidad de secuencia entre exones en algunos de los grupos analizados. Sin embargo, no podemos detectar un patrón claro de cómo se están duplicando los exones, es decir, no podemos ver el orden en el cual se duplica y se inserta un exón. Solamente podemos ver, para el grupo 4, los exones que se están duplicando es el segundo y se inserta en la posición 4 porque en estas posiciones vemos que la media del porcentaje de identidad en secuencia es alta. Que no podamos observar un patrón de repetición está asociado a que los valores del porcentaje de identidad se están promediando, es por ello que se decidió analizar genes individuales.

En la figura 5.9A-F se muestran ejemplos individuales para cada grupo analizado, para el gen de 4 exones (fig. 5.9A) se observa que los exones más parecidos son los del extremo 5', para el gen de 5 exones, también se ve que los más parecidos son los exones del extremo 5' pero también el exón 1 tiene un porcentaje de identidad alto entre los exones 2 y 3. Para los genes de 6, 7 y 8 exones (fig. 5.9C-F), se observa que los exones internos son los más similares entre sí.

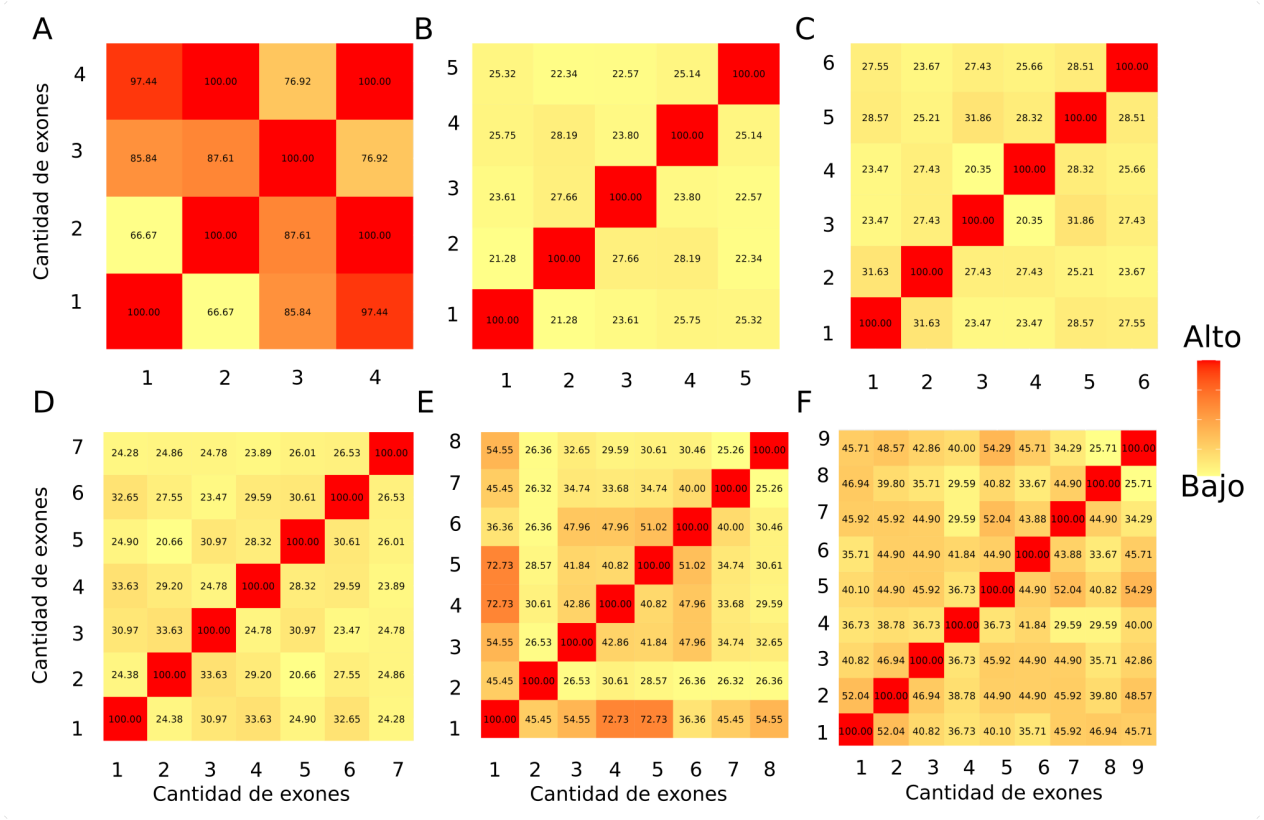

Figura 5.8: Media de la puntuación de clustal omega para los diferentes grupos definidos por la cantidad de exones por gen que codifican para repeticiones de ankirina. A) Grupo 4, n=515; B) Grupo 5, n=228; C) Grupo 6, n=146; D) Grupo 7, n=101; E) Grupo 8, n=100; F) Grupo 9, n=74, Los colores indican el valor medio del porcentaje de identidad, en rojo alto y en amarillo bajo porcentaje de identidad. En cada cuadrante en números se muestra la media del porcentaje de identidad en secuencia.

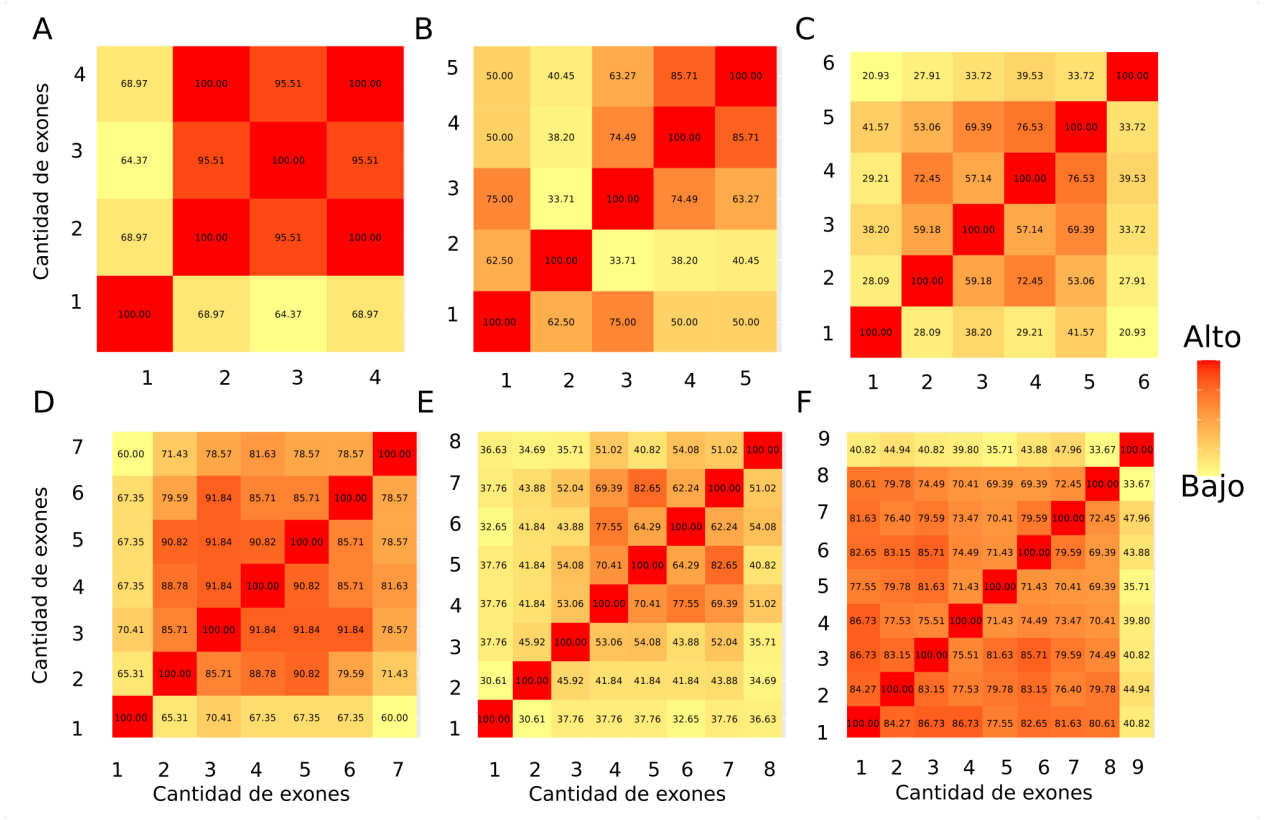

Figura 5.9: Puntuación de clustal omega para genes individuales. A) Grupo 4, GenBankID: KOB66863.1; B) Grupo 5, GenBankID: EQC30471.1; C) Grupo 6, GenBankID: EQC24718.1; D) Grupo 7, GenBankID: EEB94296.1 ; E) Grupo 8, GenBankID: EQC36829.1 ; F) Grupo 9, GenBankID: KXZ42561.1. Los colores indican el valor medio del porcentaje de identidad, en rojo alto y en amarillo bajo porcentaje de identidad. En cada cuadrante en números se muestra el porcentaje de identidad en secuencia.

A modo de control, para comprobar que lo que vimos en el análisis anterior, no se debe a solamente el largo de los exones a su clase y a la fase de los intrones, seleccionamos de nuestro conjunto de datos aquellos genes compuestos de más de 3 exones que codifican para repeticiones de ankirinas y su variabilidad de la longitud de los exones (Δ) es inferior a 0,5. Hemos encontrado un total de 579 genes.(Fig. 5.10).

En la figura 5.10A-F, se muestra la matriz de la media del porcentaje de identidad de secuen-

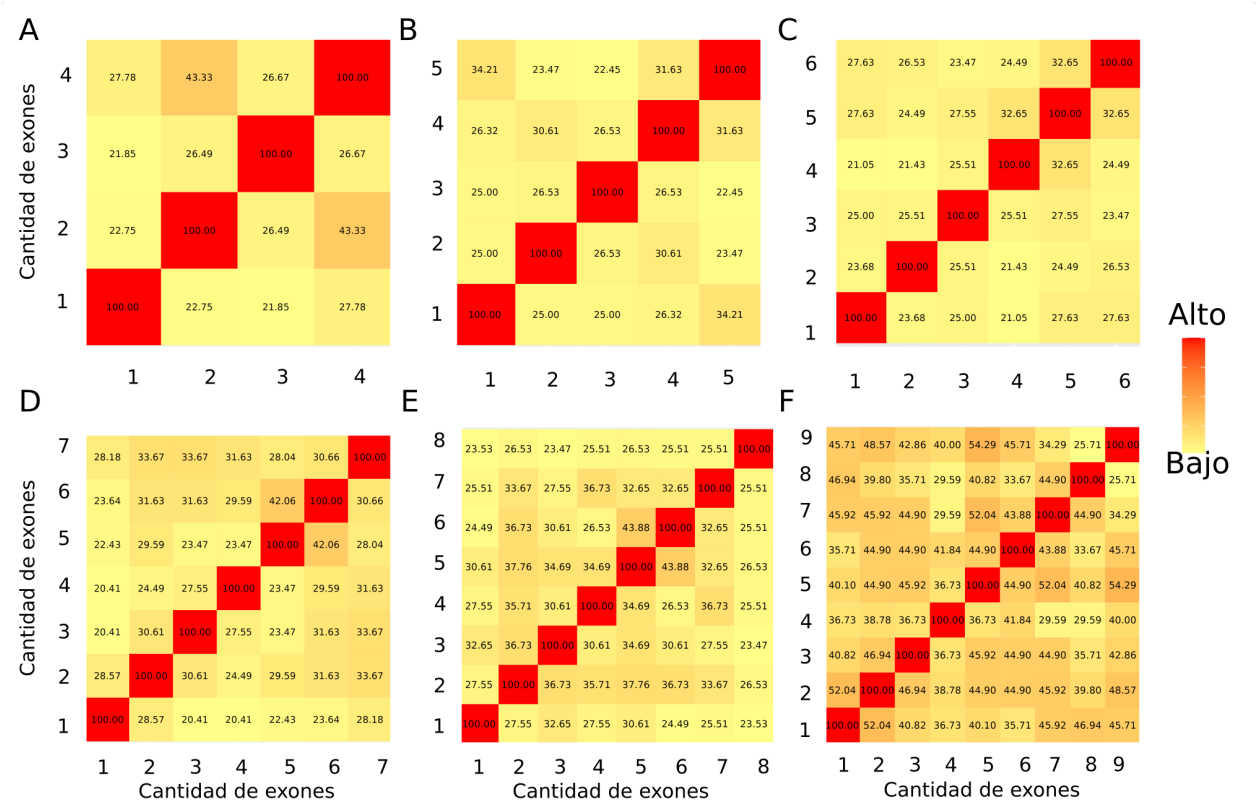

Figura 5.10: Media de la puntuación de clustal omega para los diferentes grupos definidos por la cantidad de exones por gen que codifican para repeticiones de ankirina. A) Grupo 4, n=79; B) Grupo 5, n=40; C) Grupo 6, n=72; D) Grupo 7, n=49; E) Grupo 8, n=75; F) Grupo 9, n=63, Los colores indican el valor medio del porcentaje de identidad, en rojo alto y en amarillo bajo porcentaje de identidad. En cada cuadrante en números se muestra la media del porcentaje de identidad en secuencia.

cia entre pares de exones (en color) por grupos. Observamos que la media de la identidad de secuencia entre exones es baja en la mayoría de los casos. Sin embargo, por ejemplo para los exones del grupo 4 (Fig.5.10A), se observa que entre los exones 2 y 4 la media del porcentaje de identidad de secuencia es un poco mayor que el resto. A medida que aumenta el número

de exones, por ejemplo para los grupos 8 y 9, más exones se parecen entre sí, pero no se puede observar un patrón claro de duplicación. Ahora bien, si comparamos estos resultado con los resultados de agrupar genes según la fase de los intrones, en este ultimo se vio mayor señal de la media de la identidad en secuencia.

Al analizar la identidad en secuencia de genes individuales (fig. 5.11), vemos que son muy diferentes a lo visto en la figura 5.9, en la cual los intrones de los genes son todos de la misma fase. En este caso vemos que el gen que contiene 4 exones que codifican para repeticiones de ankirinas, los que tienen mayor porcentaje de identidad en secuencia son solamente el 3 y el 4. Para el resto de los genes, el porcentaje de identidad entre exones, en su mayoría, es baja.

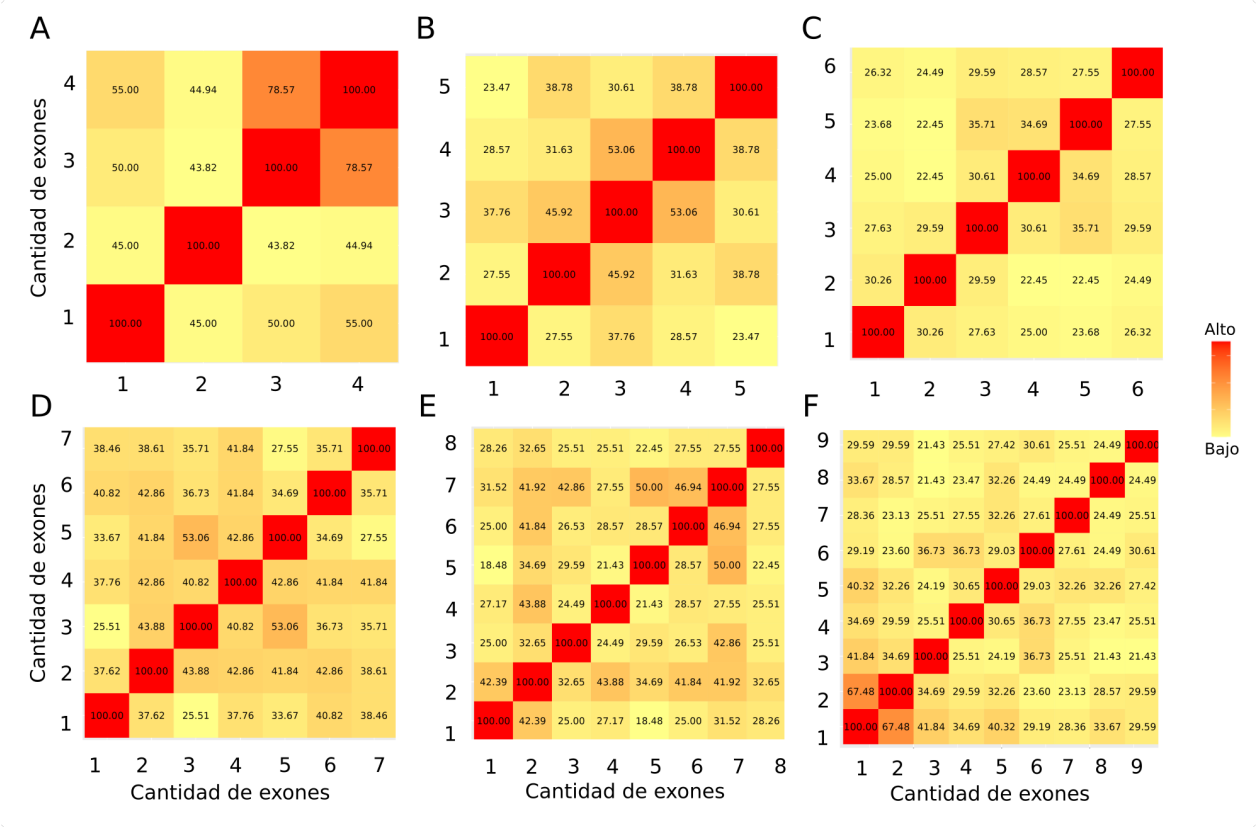

Figura 5.11: Puntuación de clustal omega para genes individuales. A) Grupo 4, GenBankID: EQC26698.1; B) Grupo 5, GenBankID: EFB28192.1; C) Grupo 6, GenBankID: AAN05492.1; D) Grupo 7, GenBankID: ETV89257.1 ; E) Grupo 8, GenBankID: OQR88973.1 ; F) Grupo 9, GenBankID: OQR85846.1. Los colores indican el valor medio del porcentaje de identidad, en rojo alto y en amarillo bajo porcentaje de identidad. En cada cuadrante en números se muestra el porcentaje de identidad en secuencia.

Organismo	Número de proteínas
Homo sapiens	134
Mus musculus	63
Rattus norvegicus	16
Arabidopsis thaliana	7
Drosophila melanogaster	7
Caenorhabditis elegans	7
Oryza sativa subsp. japonica	5
Gallus gallus	2
Bos taurus	2
Danio rerio	2
Xenopus tropicalis	1
Chlamydomonas reinhardtii	1
Canis lupus familiaris	1

Tabla 5.1: Número de proteínas de diferentes organismos que presentan eventos de empalme alternativo anotados con evidencia experimental en UniProtKB.

5.2. Empalme Alternativo

Como ya se mencionó anteriormente, el empalme alternativo es un mecanismo por el cual un mismo gen puede dar lugar a diferentes transcriptos. Los exones susceptibles a eventos de empalme alternativo se pueden generar a partir de tres mecanismo, el barajado de exones, la conversión de intrones y transición de exones constitutivos (Keren *et al.*, 2010). En esta sección se analizarán si los diferentes transcriptos, producto del empalme alternativo, afectan a las repeticiones de ankirinas, es decir, si son o no eliminadas luego del evento de empalme alternativo. En los casos en los que se eliminan repeticiones, luego del evento de empalme alternativo, se caracterizará cuáles son las que se eliminan, es decir, ¿Se eliminan las repeticiones localizadas en los extremos o las internas del arreglo de repeticiones?. De nuestra base de datos se encontraron un total de 248 proteínas con evidencia experimental de empalme alternativo según las anotaciones de UniProtKB. En la Tabla 5.1 se muestra el número de proteínas por organismo, en su mayoría las proteínas analizadas son de *Homo sapiens* y *Mus musculus*.

En la figura 5.12, en el diagrama de dispersión se muestra la cantidad de repeticiones de ankirinas antes y después de que suceda un evento de empalme alternativo. Lo que observa-

mos fue que los eventos de empalme alternativo de las repeticiones es variado, observamos casos en los que se eliminan todas las repeticiones, algunas o no se elimina ninguna (Fig. 5.12). En la mayoría de los casos de empalme alternativo, el exón que codifica la repetición se conserva. Aproximadamente el 60 % de las variantes de los eventos empalme alternativo conservan todas sus repeticiones, es decir, que estos eventos no afectan a ninguno de los exones que codifican para esas repeticiones. De las 248 proteínas analizadas, hay un total de 471 variantes de empalme alternativo.

Para evaluar las posiciones en las que las repeticiones son eliminadas luego de que ocurra

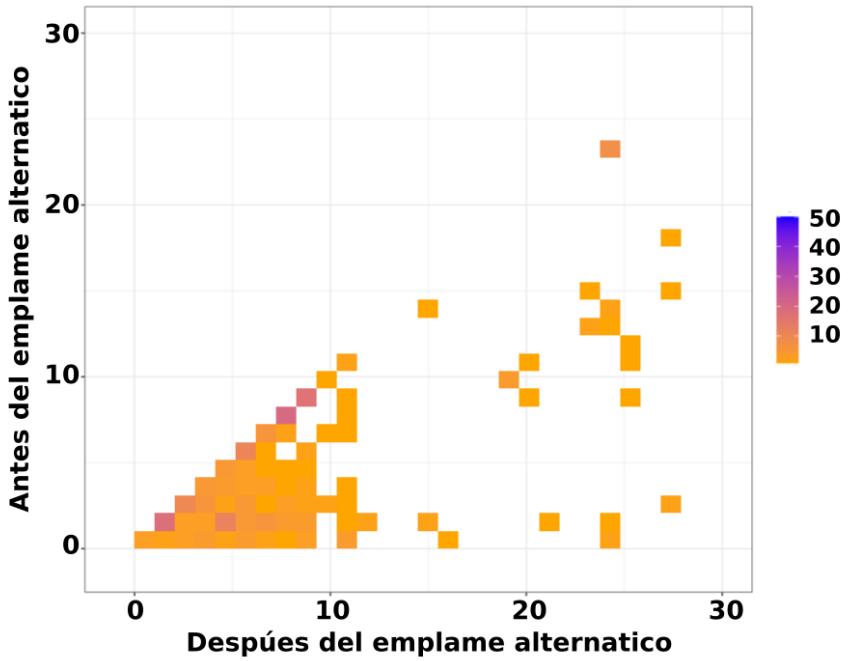

Figura 5.12: Diagrama de dispersión, en mapa de calor, entre el número de repeticiones antes y después del evento de splicing alternativo.

un evento de empalme alternativo, agrupamos a todas las proteínas según la cantidad de repeticiones que contienen en su secuencia. En la figura 5.13, se muestran todos los eventos de splicing alternativo para cada uno de los grupos. En algunos casos el número de proteínas es muy bajo, pero también hay grupos en los que el número de proteínas es superior a 20. Por ejemplo, si analizamos los grupo que contiene 5, 6, 7, 8 y 9 repeticiones (fig. 5.13D-H), vemos que en su mayoría, las frecuencias más altas son en repeticiones localizadas en los extremos de los arreglos, esto indica que las repeticiones de los extremos, mayormente, son

las que están involucradas en los eventos de empalme alternativo.

En la figura 5.14, se muestra un ejemplo de una proteína con sus diferentes variantes de

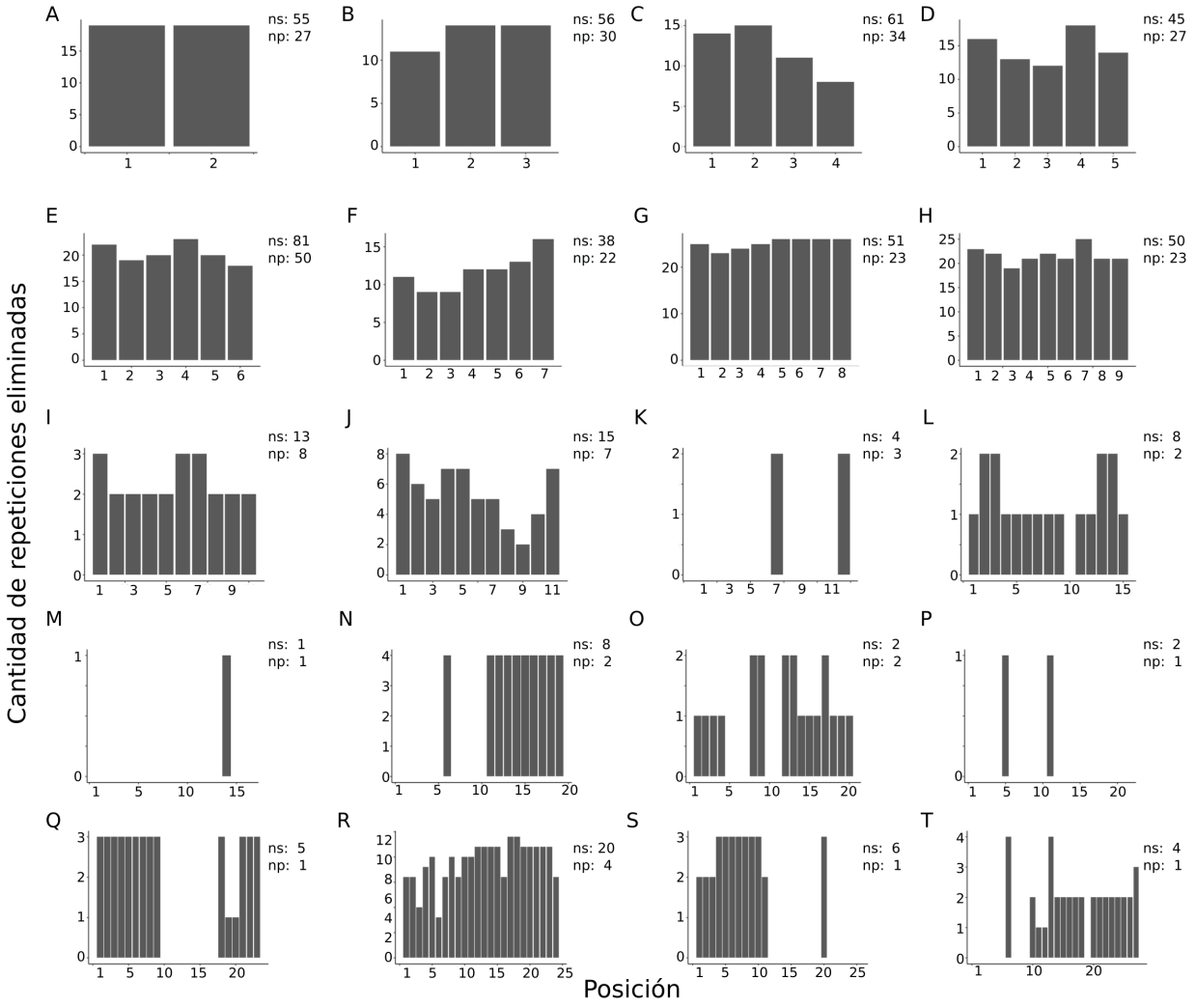

Figura 5.13: Posición de las repeticiones que se eliminan después de que se produzca un evento de empalme alternativo. Donde ns es el número de eventos de empalme alternativo analizados y np es el número de proteínas analizadas.

empalme alternativo, su UniProtID: Q8C8R3. Esta proteína tiene una longitud de 3898 AA, está compuesta por una región ANK (57-819) y una región que no contiene repeticiones de ankirina (820-3898). Esta proteína tiene 8 isoformas, en este ejemplo se observan diferentes variantes de empalme alternativo, que implican la eliminación de diferentes repeticiones. En la isoforma Q8C8R3-2 todas las repeticiones se conservan. En el caso de las isoformas Q8C8R3-7 y Q8C8R3-8 sólo se eliminan las repeticiones C-terminales. En la isoforma Q8C8R3-6, la

repetición número 6, está truncada en los primeros aminoácidos y se eliminan las repeticiones C-terminales. Finalmente en la isoforma 5, se eliminan los últimos 7 aminoácidos de la repetición 17 y los primeros 7 aminoácidos de la repetición 18. En el MSA vemos lo comentado anteriormente, la fase estructural con la que se definieron las repeticiones a veces no es la misma que la de los exones, por lo tanto hay algunas repeticiones definidas por dos o más exones y que el intrón que se inserta entre los exones que codifican para una repetición mayoritariamente se localiza entre los residuos 7 y 8 como se ve en este caso. También se observa que al final de las isoformas 7 y 8 las secuencias no alinean bien, esto es debido a que en esas isoformas cuando ocurre el evento de empalme alternativo hay un corrimiento del marco de lectura.

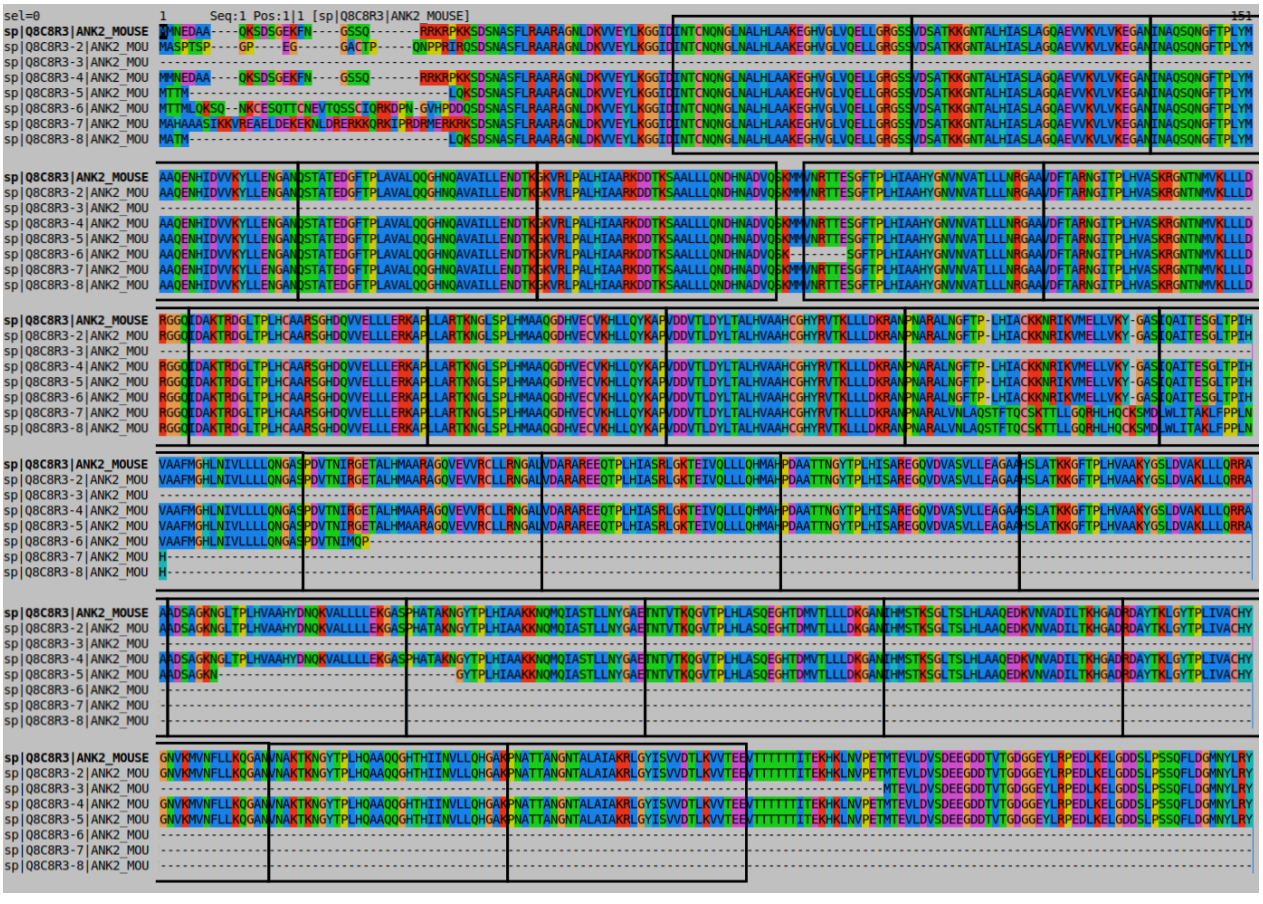

Figura 5.14: Alineamiento Múltiple de Secuencias (MSA) de las variantes de empalme alternativo de la proteína UniprotID: Q8C8R3, en cajas negras se muestran las repeticiones. Solo se muestra la región que contiene las repeticiones de ankirinas, la región No-Ankirina es omitida para una mejor visualización.

5.3. Conclusiones del capítulo

En este capítulo analizamos y caracterizamos la arquitectura exón-intrón de 68,650 proteínas de repeticiones de ankirina en 11,662 organismos eucariotas, utilizando un conjunto de datos anotado en Galpern (2020) (Galpern *et al.*, 2020). Nuestros hallazgos revelaron varios patrones interesantes en relación con la longitud de los exones y la distribución de fases. En primer lugar, observamos que la longitud de exón más frecuente es de 99 nucleótidos (nt), y una proporción significativa de exones con esta longitud codifica para repeticiones de ankirinas. Además, notamos que los exones con longitudes que eran múltiplos de 3 nucleótidos eran más comunes. Esta observación es significativa, ya que influye directamente en la fase del intrón adyacente. Específicamente, encontramos que la mayoría de los intrones pertenecían a la fase 0. Además, en términos de clase de exón, la categoría predominante fue la clase 0-0. Esta clase de exón se considera asociada con la duplicación de exones y/o reclutamiento de exones, según informes de estudios previos realizados por (De la Llosa *et al.*, 1985; Jones *et al.*, 1985; Yamada *et al.*, 1984). La presencia de exones simétricos fue especialmente notable, lo que sugiere una preferencia evolutiva por intercambiar exones simétricos. Los exones simétricos son favorables porque no interrumpen el marco de lectura corriente abajo, y son el único tipo de exones que se pueden insertar en intrones de la misma fase, según sugieren (Fedorov *et al.*, 1998; Kolkman y Stemmer, 2001; Patthy, 1987).

De hecho, la distribución no aleatoria de las clases de exones después de eventos de duplicación de exones o barajado de exones se ha observado en estudios anteriores, como se menciona en (Fedorov *et al.*, 1998). Esto indica que ciertas clases de exones son más propensas a estar involucradas en estos eventos, lo que sugiere un orden preferencial de disposición de los exones. Además, la identificación de más de 60 elementos móviles diferentes codificados por exones de clase 0-0, 1-1 y 2-2, según informes de (Kolkman y Stemmer, 2001), respalda la idea que el barajado de exones es un mecanismo para generar diversidad genética. Estos elementos móviles pueden contribuir a la diversidad funcional y estructural de las proteínas mediante la reorganización de exones. Considerando la alta frecuencia de intrones de fase 0 y la observación de que los genes compuestos por cuatro exones de la misma clase, tienen un alto porcentaje de identidad en secuencia, es razonable considerar la posibilidad de eventos

de barajado de exones. La abundancia de exones simétricos en la arquitectura exón-intrón también respalda la noción de una tendencia a mantener repeticiones. Los exones simétricos, al poder ser insertados en intrones sin interrumpir el marco de lectura corriente abajo, tienen más probabilidades de preservarse durante los procesos evolutivos, incluidos los eventos de empalme alternativo, como sugirieron (Fedorov *et al.*, 1998; Kolkman y Stemmer, 2001; Patthy, 1987).

La asignación de las ubicaciones de las repeticiones en los exones brinda información valiosa sobre cómo se distribuyen las repeticiones dentro de la arquitectura exón-intrón. La observación de que la mayoría de las repeticiones están codificadas por un solo exón (repetición completa) es consistente con hallazgos previos en el estudio realizado por (Street *et al.*, 2006). Además, la identificación de repeticiones codificadas por dos exones consecutivos (repetición parcial) agrega mayor complejidad a la distribución de las repeticiones. La conservación de la posición donde los intrones interrumpen el marco de lectura dentro de estas repeticiones parciales indica un patrón consistente. Esta observación sugiere que, en algunos casos, la fase estructural definida en (Parra *et al.*, 2015) no necesariamente se corresponde con la fase genómica. Estas discrepancias pueden surgir debido a variaciones en la estructura exón-intrón y los mecanismos específicos involucrados en la generación y mantenimiento de las repeticiones. El uso de un conjunto de datos más amplio que comprende genes de diferentes organismos en su estudio en comparación con el utilizado en (Street *et al.*, 2006) brinda una perspectiva más amplia sobre la distribución de las repeticiones dentro de los exones. A pesar de las diferencias en el conjunto de datos, los resultados obtenidos en los análisis son consistentes con los hallazgos de (Street *et al.*, 2006), lo que respalda aún más la solidez y generalización de estas observaciones. En general, la caracterización de la distribución de repeticiones en los exones contribuye a nuestra comprensión de la arquitectura exón-intrón y arroja luz sobre los mecanismos subyacentes a la generación, mantenimiento e implicaciones funcionales de las repeticiones.

El enfoque para detectar eventos de barajado de exones examinando la distribución de las clases de exones y seleccionando proteínas con una distribución no aleatoria es un enfoque razonable. La presencia de una distribución constante de clases de exones, particularmente con todos los intrones siendo de la misma fase, sugiere una posible participación de los

mecanismos de barajado de exones. Si bien las similitudes observadas entre exones consecutivos en ambos casos proporcionan evidencia intrigante, se necesitan estudios adicionales para confirmar la ocurrencia de eventos de duplicación de exones o barajado de exones. Análisis adicionales, como genómica comparativa, estudios filogenéticos o validación experimental, pueden ayudar a dilucidar los mecanismos evolutivos involucrados en estos procesos. Es importante continuar investigando y explorando estas proteínas para reunir más evidencia y establecer una comprensión más concluyente de su dinámica evolutiva, incluido el posible papel de la duplicación de exones y la recombinación de exones en su evolución.

La observación de que aproximadamente la mitad de las proteínas conservan todas sus repeticiones después del empalme alternativo indica una preferencia por mantener las regiones repetitivas en el producto final de la proteína. Esto sugiere que la presencia de repeticiones podría desempeñar un papel funcional y que su eliminación podría tener consecuencias potenciales para la función y estabilidad de la proteína. Por otro lado, los eventos de empalme alternativo que conducen a la eliminación de algunas repeticiones, especialmente aquellas ubicadas cerca de los extremos, resaltan la flexibilidad potencial en la regulación del contenido repetitivo en estas proteínas. La eliminación de repeticiones específicas mediante el empalme alternativo podría modular las interacciones proteicas, la localización celular u otros aspectos funcionales asociados con las regiones repetitivas.

De hecho, el estudio de la arquitectura exón-intrón proporciona información valiosa sobre los mecanismos evolutivos que dan forma a las familias de proteínas, incluidas las proteínas con repeticiones de ankirina. Si bien la duplicación, eliminación y transferencia horizontal de genes han sido reconocidos tradicionalmente como mecanismos importantes en la evolución de las proteínas, nuestros hallazgos sugieren que los eventos de empalme alternativo y el barajado de exones también pueden desempeñar roles significativos. Los eventos de empalme alternativo pueden generar múltiples isoformas de transcriptos a partir de un solo gen, lo que expande la diversidad proteica dentro de una familia génica. La observación de que el empalme alternativo puede afectar el número y la disposición de las repeticiones de ankirina resalta la contribución potencial de este mecanismo a la evolución de estas proteínas. El empalme alternativo proporciona un mecanismo para ajustar finamente la función y expresión de las proteínas en respuesta a condiciones celulares o ambientales específicas.

Por otro lado, el barajado de exones implica el reordenamiento de exones entre diferentes genes, lo que lleva a la creación de nuevas estructuras y funciones génicas. La distribución no aleatoria de las clases de exones y la similitud observada entre exones consecutivos en tu estudio sugieren la posibilidad de eventos de recombinación de exones en la evolución de las proteínas con repeticiones de ankirina. Se necesitan investigaciones y análisis adicionales para confirmar y caracterizar la extensión de la recombinación de exones en esta familia de proteínas.

Capítulo 6

Análisis el paisaje energético de proteínas de la familia de Ankirina y de sus mecanismos de plegado

Como ya se mencionó anteriormente, no conocemos con exactitud los mecanismos por el cual las proteínas se pliegan, por lo tanto para describir como las proteínas se pliegan utilizamos la teoría de paisaje energético. Esta teoría explica que, durante el proceso de plegado, las interacciones presentes en el estado nativo son energéticamente más favorables que las interacciones aleatorias. Este sesgo energético es producto a la cooperatividad de las interacciones presentes en el estado nativo y que además cada vez que se forma una interacción nativa, la energía baja más que en el caso de formarse una interacción no nativa o al azar ("Principio de mínima frustración"). La morfología de embudo corrugado del paisaje energético de las proteínas, producto de la frustración energética local, es un factor que limita la plegabilidad de las proteínas. Para que una proteína pueda plegarse de forma robusta y en tiempos acotados, la diferencia de energía entre el estado nativo y el desplegado tiene que ser significativamente mayor que la rugosidad del embudo.

En la sección anterior hemos analizado y caracterizado cómo son las proteínas con repeticiones de ankirinas a un nivel genómico. En este capítulo nos centraremos en tratar de caracterizar el plegado y el paisaje energético de estas proteínas. Vamos a aplicar a varios miembros de la familia ANK de diferentes largos y con algunas variaciones estructurales, un

método de simulación de dinámica molecular del tipo de grano grueso. El método aplicado, es un campo de fuerza de memoria asociativa, mediada por agua, estructura y energía, que se llama *AWSEM* (Ver métodos). Este método, está determinado por las interacciones físicas que ocurren en la molécula y además se basa en el término de sesgo de estructura local, es decir, que sesga las secuencias locales (largo >= 9 aa) hacia conformaciones encontradas en proteínas que contienen secuencias de fragmentos análogos. Aplicamos este método a una proteína diseñada que contiene 4 repeticiones idénticas al consenso de ankirina (PdbID: 1N0R) y a modelos que generamos que contienen 3, 5 y 6 peticiones idénticas al consenso. Además caracterizamos las proteínas naturales de I$\kappa\beta\alpha$ (PdbID: 1NFI), la proteína Notch (PdbID: 1OT8) y P16 (PdbID: 1A5E).

6.1. Recocido simulado y análisis de *Umbrella Sampling* de proteínas de la familia ANK

El recocido simulado (en inglés, *simulated annealing*), es un método probabilístico para encontrar el mínimo global de una función de coste que puede poseer varios mínimos locales. Funciona emulando el proceso físico por el cual un sólido se enfría lentamente para que, cuando finalmente su estructura se "congele", esto ocurra en una configuración de energía mínima. Para evaluar el proceso de plegado de proteínas se implementa este algoritmo. Se parte de un sistema que está constituido por la proteína totalmente desplegada que se conoce como *random coil* y consiste en hacer evolucionar al sistema desde una temperatura alta hasta temperaturas lo suficientemente bajas como para que la proteína se pliegue.

En el diagrama 6.1 se muestra el típico modelo de plegado de dos estados, en el cual se muestra que una proteína pasa de un estado totalmente desplegado (U) a un estado completamente plegado (N). Mientras que el diagrama 6.2, muestra que durante el proceso de plegado pueden existir intermediarios de plegado. Por lo tanto, para describir completamente el mecanismo de plegado y caracterizar las rutas de plegado no solo se estudian los estados iniciales y finales, sino también hay que estudiar los estados intermedios y los estados de transición que los conectan. Un estado intermediario es una conformación estable pero que no está totalmente plegada ni totalmente desplegada.

$$U \longleftrightarrow N \tag{6.1}$$

$$U \longleftrightarrow I \longleftrightarrow N \tag{6.2}$$

6.2. Recocido simulado de miembros de la familia ANK

Como ya mencionó, para realizar las simulaciones de recocido simulado se usó el campo de fuerza *AWSEM* implementado en *OpenMM* (Lu *et al.*, 2021). *OpenMM*, es una caja de herramientas que se diseñó en 2010 y se usa para simulaciones de dinámica molecular, su ventaja es que se puede correr en diferentes arquitecturas de hardware (Eastman *et al.*, 2017). El campo de fuerza *AWSEM* a diferencia de un modelo de Gõ, representa a cada uno de los aminoácidos que componen la proteína usando el modelo estructural de tres cuentas (ver métodos) y tiene en cuenta efectos como la energía de van der Waals, la solvatación y las interacciones electrostáticas. Además *AWSEM* incorpora la presencia de agua en el proceso de plegado, considerando explícitamente las interacciones mediadas por agua y contiene un término de sesgo de estructura local. Debido a su descripción detallada y sus términos de energía más complejos, *AWSEM* puede ser computacionalmente más costoso que un modelo de Gõ. También *AWSEM* considera la frustración como parte de su modelo, ya que tiene en cuenta cómo las interacciones entre los aminoácidos pueden contribuir a la estabilidad o inestabilidad de la estructura proteica. Debido a estas diferencias, utilizar el campo de fuerza de *AWSEM* puede aportar una descripción más detallada que un modelo de Gõ y la inclusión de la frustración puede ayudarnos a comprender mejor el proceso de plegado de proteínas. La coordenada de plegado elegida es Q_w, el cual representa el grado de plegado y la similitud estructural de las conformaciones simuladas de la proteína con su estado nativo (ver mértodos). Se utilizó a Q_w como coordenada de plegado debido a que la dinámica de plegado de una proteína se puede monitorear a través de cómo cambia el valor de Q_w a lo largo del tiempo durante una simulación *AWSEM*.

El campo de fuerza *AWSEM* tiene ciertos parámetros definidos por defecto, como por ejemplo, el tamaño del fragmento de las memorias y la energía del fragmento, para las ankirinas, estos parámetros fueron modificados y para todas las simulaciones se usaron los valores que

se muestran en la tabla 6.1). El tamaño del fragmento se modificó debido a que las proteínas repetitivas forman contactos locales, por lo tanto las interacciones vecinas que cooperan en la estabilización de la estructura de las repeticiones son importantes, por lo cual consideramos que agrandando el tamaño del fragmento se lograban incluir estas interacciones. Debido a que se modificó el tamaño del fragmento es necesario modificar la energía del fragmento para que sea equivalente al definido por defecto, porque sino estaríamos aumentando el sesgo hacia la formación de estructuras secundarias. En la figura 6.1, se muestran los resultados

Parámetro	Valor por defecto	Valor para ankirinas
Largo del fragmento	9aa	12aa
Energía del fragmento	20	16

Tabla 6.1: Parámetros modificados del campo de fuerza *AWSEM*

de los cálculos realizados para encontrar el valor de la energía del fragmento para el tamaño del fragmento de 12aa. Las pruebas que se hicieron consistieron en, usando una trayectoria de una simulación de recocido simulado de la proteína 1N0R para la cual se usaron los parámetros por defecto (puntos en negro), se calculó el valor de la energía modificando el tamaño del fragmento a 12aa, los valores de energía del fragmento medidos fueron de 15 (en amarillo), 16 (en rosa) y 17 (en azul). Se puede observar en la figura 6.1A que el valor de energía que más se aproxima al de defecto es 15, Sin embargo, cuando se hicieron pruebas de recocido simulado de 1N0R, usando 15 para el valor de la energía del fragmento, no se obtuvó una buena predicción de la estructura de 1N0r (fig. 6.1B). Por otro lado, cuando se realizaron pruebas de recocido simulado de 1N0R, usando 16 para la energía del fragmento (fig. 6.1C), la predicción de la estructura fue la correcta, por este motivo, se seleccionó 12aa para el tamaño del fragmento y 16 para el valor de la energía del fragmento. Además para todos los experimentos de esta sección se usó para la memoria de los fragmentos únicamente la estructura de la proteína correspondiente, es decir, se corrieron en modo de memoria única (en inglés, *single memory*). Este concepto de memorias es un término bioinformático del potencial (ver métodos) el cual afecta localmente a la estructura a adoptar la conformación de la memoria utilizada. Las simulaciones del tipo memoria única son aquellas que se corren utilizando como guía para el proceso de plegado a la estructura de la proteína que se va a

analizar (que puede ser una estructura resuelta experimentalmente o un modelo de la proteína). Además es importante aclarar que las temperaturas expresadas no se corresponden con temperaturas reales. Cada recocido simulado se comenzó a una temperatura de 1200 y se enfrió hasta la temperatura de 400 la cantidad de pasos fue de 16.000.000.

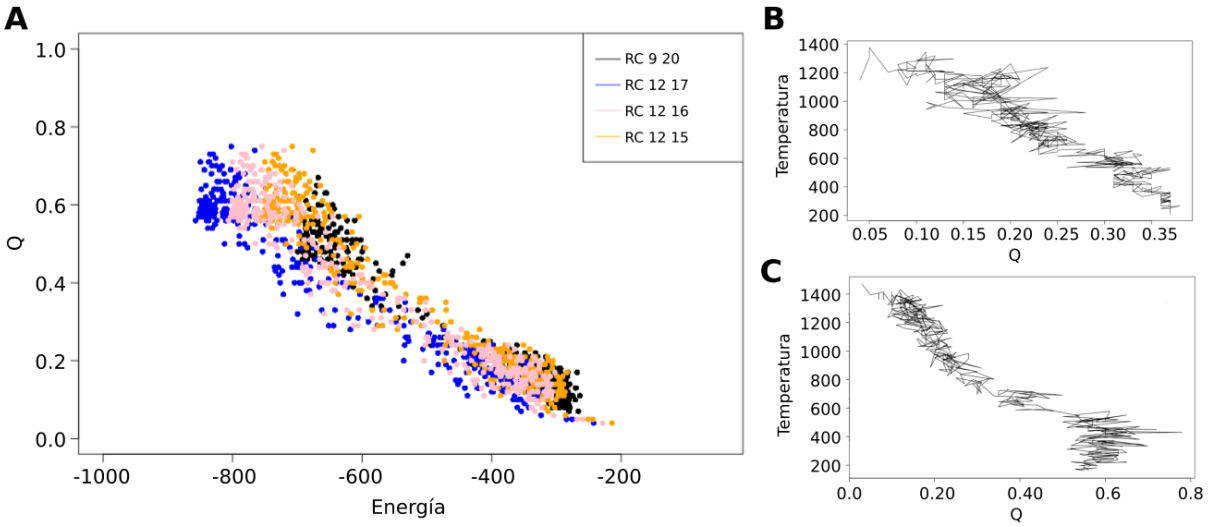

Figura 6.1: Resultados de las pruebas que se realizaron en 1N0R para definir la energía del fragmento. A) Comparación entre los diferentes valores energías del fragmento para un fragmento de largo 12aa, 15 se representa en amarillo, 16 en rosa, 17 en azul en negro se representan los valores por defecto (largo del fragmento de 9aa y energía del fragmento de 20). B) Recocido simulado de 1N0R usando largo del fragmento de 12aa y energía del fragmento de 15, C) Recocido simulado de 1N0R usando largo del fragmento de 12aa y energía del fragmento de 16.

6.2.1. Ankirinas diseñadas por consenso

La primer proteína diseñada por consenso fue la proteína 1N0R, como ya se mencionó esta proteína contiene 4 repeticiones idénticas al consenso de ANK. En esta sección se analizarán y mostrarán los resultados de experimentos de recocido simulado para la proteína 1N0R y modelos de proteínas de 3, 5 y 6 de la secuencia consenso.

En la figura 6.2A-D se muestra el mejor valor de Q_w obtenido en cada corrida de recocido simulado, su valor de RMSD y un alineamiento estructural entre el cristal y la estructura que se corresponde con el mejor valor de Q_w, para las proteínas 3ANK, 1N0R, 5ANK y 6ANK. Como ya se mencionó las simulaciones de dinámica molecular se corrieron en modo

de memoria única, por lo tanto las memorias de los fragmentos para recocido simulado se obtuvieron de la proteína de referencia para cada caso. Los mejores valores de Q_w de los recocidos simulados para la proteína 3ANK (fig. 6.2A) fluctúan entre 0,62 y 0,88 mientras que los valores de RMSD entre 1,023 y 1,529 Å. En esta proteína, se observa que todos los valores de Q_w son mayores a 0,62, es decir que en todas las corridas se obtuvieron modelos similares al del estado nativo. Esto se justifica al ver que las estructuras de las repeticiones están correctamente formadas así como también el *β-harping* y la orientación de las repeticiones es muy parecida a la del modelo para todos los valores de Q_w, excepto para la corrida 10, donde se ve que la orientación de la primera repetición, con respecto al modelo, no es la correcta.

Para la proteína 1N0R (fig. 6.2B), para todos los mejores valores de Q_w, las estructuras secundarias de hélice-α de las repeticiones están bien formadas. Sin embargo, solamente para los valores de $Q_w > 0{,}6$, la formación del *β-harping* es la correcta y la orientación de las repeticiones es más parecida a la del cristal que el resto de los valores de Q_w. Con respecto a los valores de RMSD, podemos ver que para los valores de $Q_w > 0{,}6$ el valor de RMSD es chico y a medida que el valor de Q_w baja el valor de RMSD aumenta. Lo que podría llamar nuestra atención es que, a valores de Q_w mayores a 0,65 el RMSD es chico, indicando que a estos valores de Q_w se comienza a observar un correcto plegado de 1N0R.

Los mejores valores de Q_w para la proteína 5ANK (fig. 6.2C) fluctúan entre 0,34 y 0,76 mientras que los valores de RMSD entre 1,808 y 16,459 Å. En esta proteína, se observa que, solamente para valores de Q>0,4 las estructuras de las repeticiones están correctamente formadas. Mientras que para valores de Q_w más bajos, solamente se ve que algunas estructuras secundarias de las repeticiones internas están correctamente formadas y solamente para valores de Q_w mayores a 0,6 la orientación de las repeticiones y el *β-harping* es la correcta. En este caso la proteína analizada tiene 5 repeticiones de ankirina y se comienzan a ver valores de Q_w por debajo de 0,4, esto puede estar asociado a que la proteína es más grande y las posibles conformaciones y orientaciones de las repeticiones son mayores. Podemos decir que en algunas simulaciones no se obtuvo una correcta predicción de la estructura de la proteína. Sin embargo, vemos que en otras trayectorias los valores de Q_w fueron altos y las estructuras son muy parecidas al modelo.

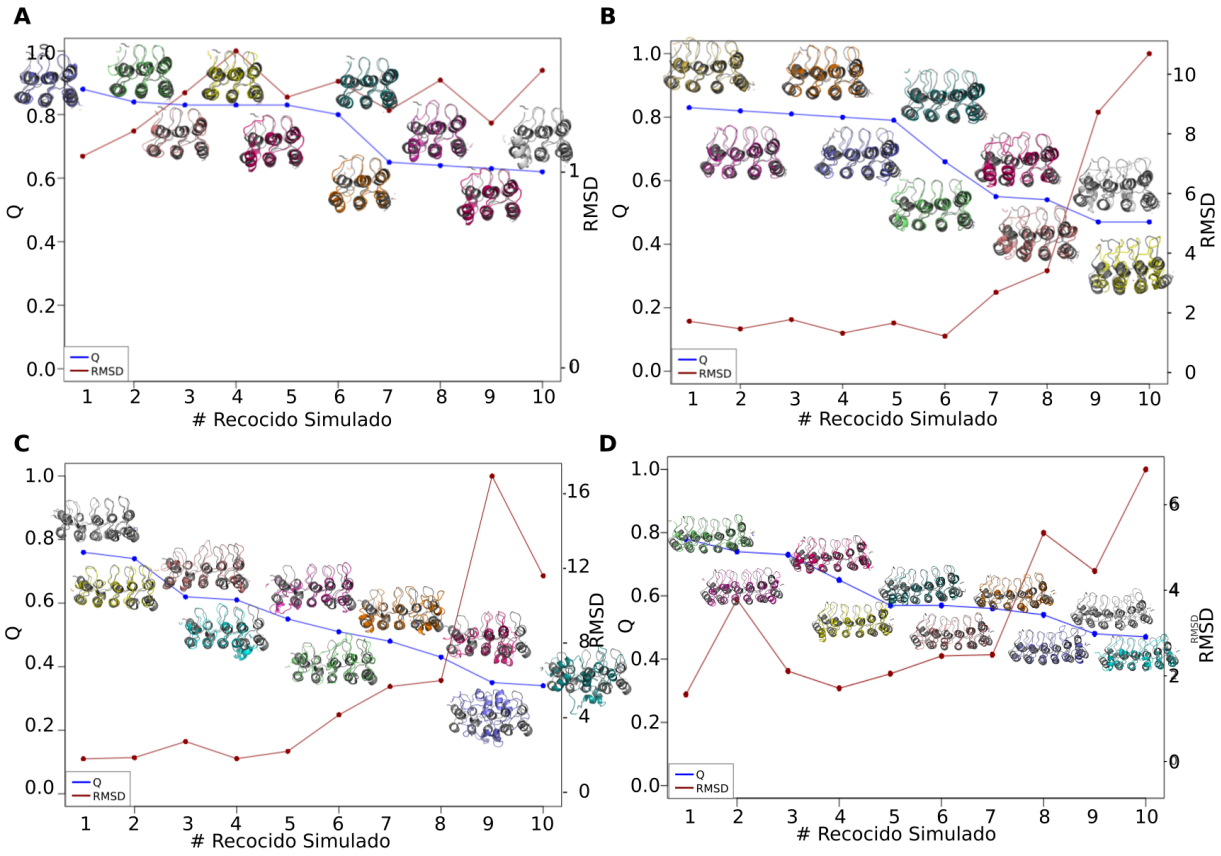

Figura 6.2: Resultados de las 10 corridas de recocido simulado están ordenados de forma descendente. Cada punto azul representa el valor de Q_w más alto de cada corrida y en rojo el valor de RMSD. Para cada valor de Q_w se muestra un alineamiento estructural entre la proteína cristalizada (en gris) y la conformación de mayor Q_w (en los diferentes colores). A) Modelo de 3 repeticiones del consenso (3ANK). B) 1N0R. C) Modelo de 5 repeticiones del consenso (5ANK). D) Modelo de 6 repeticiones del consenso (6ANK).

Por último, para el modelo de proteína de 6ANK (fig. 6.2D) los valores de Q_w fluctúan entre 0,47 y 0,78 mientras que los valores de RMSD entre 1,801 y 6,251 Å. Podemos ver que las estructuras de hélice-α de las repeticiones están correctamente formadas para todos los valores de Q_w y solamente para los valores de $Q_w > 0.55$ se ve que la estructura del *β-harping* también es la correcta. Con respecto a la orientación de las repeticiones, para todos los valores de Q_w, las estructuras de hélice-α están correctamente orientadas, las diferencias de orientación se observa mayormente en la estructura del *β-harping*.

Del análisis de los resultados de recocido simulado vimos que a medida que aumentábamos la cantidad de repeticiones el mejor valor de Q_w de cada simulación disminuía, sin embargo, para la proteína de 6ANK el mejor valor de Q_w aumentó y a valores de $Q_w > 0{,}55$, la

orientación de las repeticiones es la correcta, cuando en las demás proteínas analizadas este valor era de 0,6. Debido a estos resultados nos surge la pregunta de que quizás a medida de vayamos agregando repeticiones del consenso la predicción de la estructura, utilizando este modelo, mejore o quizás las proteínas pares con más de 4 repeticiones tienen paisajes energéticos menos frustrado.

De los resultados de los recocidos simulados (fig. 6.2A-D) se observó a valores mayores a 0,6 de Q_w la orientación y formación de las estructuras secundarias es muy similar a la de la proteína de referencia, definiendo un estado nativo a valores de $Q_w > 0.6$. En la figura 6.3A-D se muestran los mapas de calor de los mapas de contactos para los valores de Q_w de entre 0,3 y 0,6. Para el modelo de 3ANK (fig. 6.3A) se observan en estos intervalos de Q_w que las repeticiones están formadas pero aún hay formación de contactos no nativos entre repeticiones vecinas. En 1N0R (fig. 6.3B) vemos una correcta formación de todas las repeticiones, con la diferencia de que las repeticiones 2, 3 y 4 forman interacciones no nativas. Este resultado sugiere que la primer repetición en formarse completamente es la repetición del N-Terminal. En los modelos de 5 y 6 repeticiones (fig. 6.3C-D), se observa que las repeticiones están correctamente formadas. En el modelo de 5ANK, las repeticiones que forman contactos no nativos entre ellas son la repetición 3 y la 4. Por otro lado, el modelo de 6ANK, las dos repeticiones localizadas en el C-Terminal son las que están formando algunos contactos no nativos.

6.2.2. Proteínas naturales con repeticiones de ankirina.

Cuando hablamos de proteínas naturales, hacemos referencia a proteínas que se han encontrado en la naturaleza, a diferencia de las proteínas analizadas en la sección anterior. En esta sección se analizarán 3 proteínas miembros de la familia de ankirinas, las cuales son, I$\kappa\beta\alpha$, en este caso se analizarán dos estructuras, una de 4 repeticiones y una de 6 repeticiones, la proteína P16 y Notch.

I$\kappa\beta\alpha$ es una proteína con repeticiones de Ankirina que inhibe la actividad del factor de transcripción NF$-\kappa\beta$ (Ferreiro y Komives, 2010). La proteína NF$-\kappa\beta$ es un factor de transcripción que regula la expresión de diversos genes involucrados en la respuesta inflamatoria y la inmunidad, así como en la regulación del crecimiento celular y otros procesos biológicos.

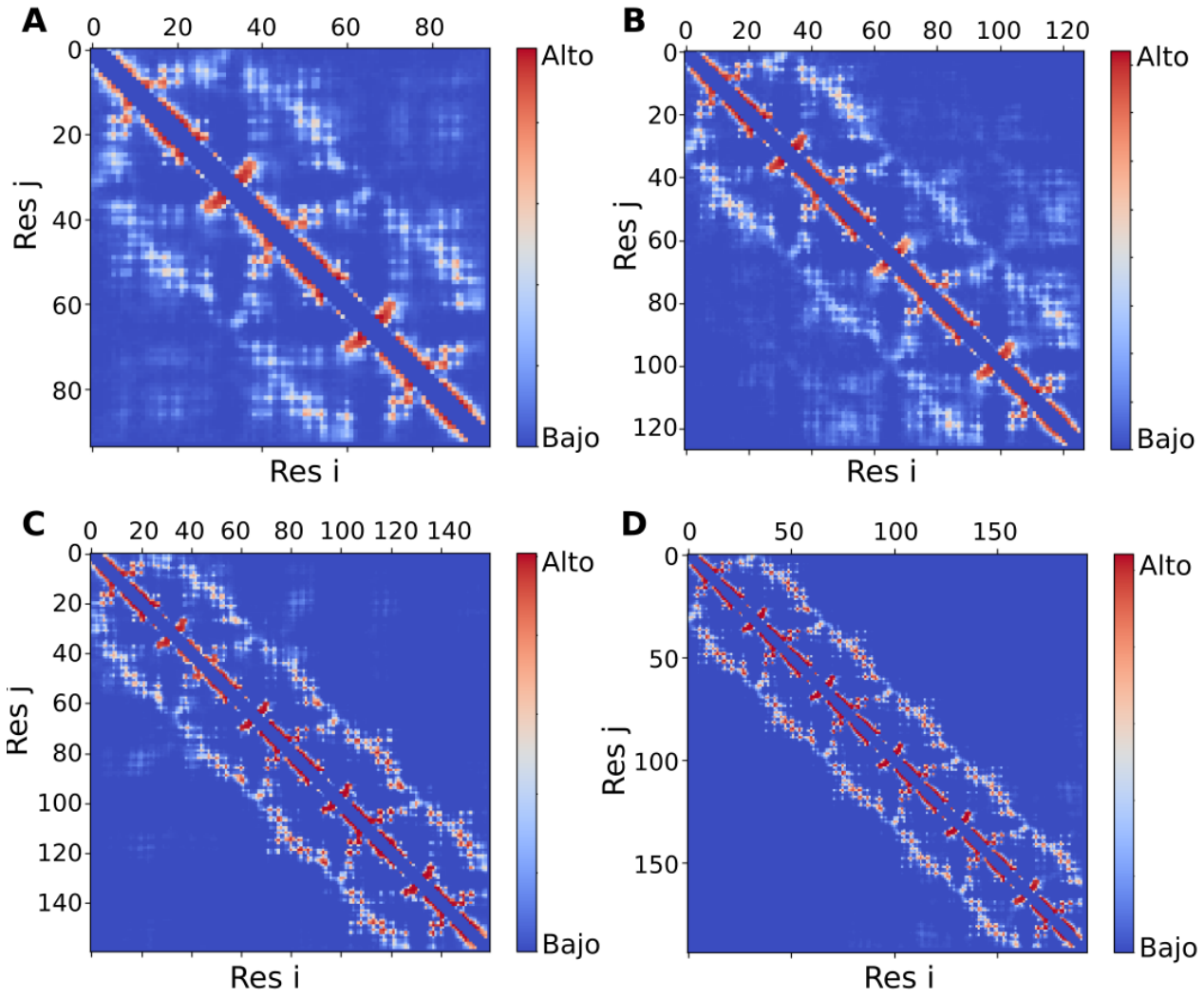

Figura 6.3: Mapa de calor de los mapas de contacto para los valores de Q_w de entre 0,3 y 0,6 para las proteínas A) Modelo de 3 repeticiones del consenso (3ANK). B) 1N0R. C) Modelo de 5 repeticiones del consenso (5ANK). D) Modelo de 6 repeticiones del consenso (6ANK).

La proteína I$\kappa\beta\alpha$ contiene 6 repeticiones de ankirina, de las cuales la ultimas dos repeticiones son más flexibles que las demás, y esta flexibilidad es crucial para la función de la proteína ya que regula la degradación de I$\kappa\beta\alpha$ y también regula la actividad de NF$-\kappa\beta$, ya que NF$-\kappa\beta$ está normalmente inactivo en el citoplasma de las células porque está unido a I$\kappa\beta\alpha$. Para las memorias de los fragmentos se utilizó del PdbID 1NFI la cadena E, este cristal contiene 6 repeticiones de ankirina, debido a que analizamos dos estructuras (una que contiene 4 y otra que contiene las 6 repeticiones), para la estructura de 4 repeticiones se seleccionaron las 4 primeras, por lo que fueron eliminadas manualmente las últimas dos.

Los valores del mejor Q_w de cada corrida varían entre 0,34 y 0,77 y el RMSD varía entre 1,520 y 15,106 Å. En la figura 6.4A podemos ver que en todas las estructuras la formación de la estructura secundaria de hélice α es la correcta para los valores de $Q_w > 0{,}4$, pero solamente, se observa una correcta orientación de las repeticiones para el valor de Q_w de 0,77. Si comparamos a esta proteína con 1N0R, ya que ambas tienen 4 repeticiones, pero con la diferencia de que una es natural y la otra diseñada, vemos que los valores de Q_w obtenidos de las simulaciones para I$\kappa\beta\alpha$ son más chicos que para 1N0R. Si comparamos las estructuras, en I$\kappa\beta\alpha$ solamente vemos una correcta orientación de las repeticiones para el valor de Q_w de 0,77, mientras que para 1N0R, se comienza a ver a valores más bajos.
Esto que se observa, ¿Puede deberse a que la simetría en secuencia de 1N0R es mayor que la de I$\kappa\beta\alpha$ y está afectando al plegado?. En la figura 6.5A se muestra en mapa de calor el mapa de contacto para los valores de Q_w de entre 0,3 y 0,6 para I$\kappa\beta\alpha$ de 4 repeticiones, si comparamos los mapas de calor de I$\kappa\beta\alpha$ de 4 repeticiones con 1N0R (fig. 6.3B), vemos diferencias en la formación de contactos no nativos, en 1N0R vemos que los contactos no nativos se dan en mayor frecuencia entre las repeticiones localizadas en el C-Terminal, mientras que en I$\kappa\beta\alpha$ de 4 repeticiones se dan entre las repeticiones internas. Estos resultados sugieren, que 1N0R comienza a plegarse desde el N-Terminal y luego se van plegando las demás repeticiones, mientras que en I$\kappa\beta\alpha$ de 4 repeticiones primero se pliegan las repeticiones del N y C-Terminal y por último las internas. Claramente se observa una diferencia en el mecanismo de plegado entre ambas proteínas lo que está afectando al plegado, además estos resultados estarían indicando que posiblemente el paisaje energético de I$\kappa\beta\alpha$ de 4 repeticiones sea más frustrado que el de 1N0R, por eso I$\kappa\beta\alpha$ de 4 repeticiones que queda atrapada con mayor frecuencia en valores de $Q_w < 0{,}6$.

Los experimentos de recocido simulado para I$\kappa\beta\alpha$ con 6 repeticiones se muestran en la figura 6.4B. Para las memorias de los fragmentos se utilizó el PdbID 1NFI la cadena E completa del cristal, es decir, con las 6 repeticiones. De los resultados se ve que para los valores de $Q_w > 0{,}4$ la formación de la estructura de hélice α de las repeticiones es la correcta, pero la localización de las repeticiones es correcta solamente para valores de valores de Q_w de 0,68 y 0,7. Si comparamos ambas I$\kappa\beta\alpha$, de 4 y 6 repeticiones, vemos que los resultados de los valores obtenidos de Q_w, son similares. Con respecto a las estructuras lo que se observó en ambas

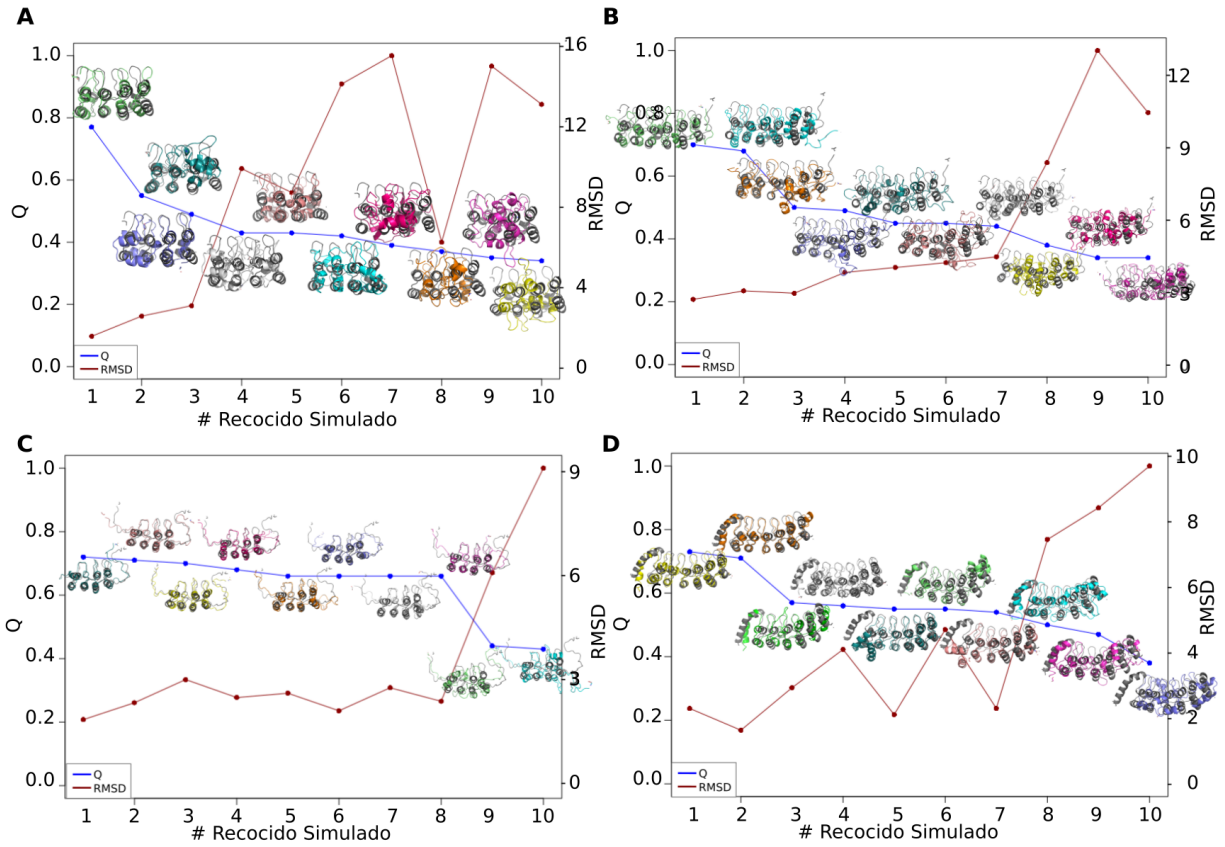

Figura 6.4: Resultados de las 10 corridas de recocido simulado están ordenados de forma descendente. Cada punto azul representa el valor de Q_w más alto de cada corrida y en rojo el valor de RMSD. Para cada valor de Q_w se muestra un alineamiento estructural entre la proteína cristalizada (en gris) y la conformación de mayor Q_w (en los diferentes colores). A) I$\kappa\beta\alpha$ de 4 repeticiones. B) I$\kappa\beta\alpha$ de 6 repeticiones. C) P16. D) Notch.

proteínas también es similar, una correcta formación de la estructura de las hélices-α de las repeticiones para valores de $Q_w > 0{,}4$ y una correcta orientación de las repeticiones para valores de $Q_w > 0{,}6$. Esto indica que, no se observaron diferencias entre ambas proteínas, en la predicción de la estructura, con la metodología aplicada.

Si comparamos I$\kappa\beta\alpha$ con el modelo de 6ANK, en los mapas de calor de la frecuencia de contactos para valores de Q_w de entre 0,3 y 0,6, figuras 6.5B y 6.3D respectivamente, vemos que I$\kappa\beta\alpha$ forma contactos no nativos entre repeticiones internas (repeticiones 3 y 4), mientras que 6ANK forma solamente algunos contactos no nativos entre las repeticiones del C-Terminal. Además se observa que la mayoría de los mejores valores de Q_w de cada corrida están por debajo de 0,5 (fig. 6.4B) a diferencia de 6ANK que en este caso la mayoría tiene

valores mayores a 0,5 (fig. 6.2D). Estos resultados indican que el paisaje energético de la proteína I$\kappa\beta\alpha$ es más frustrado que 6ANK ya que mayormente queda atrapada en valores bajos de Q_w.

Otra proteína natural que estudiamos es P16, que es un supresor de tumores compuesto de 4 repeticiones de ankirinas. Para las memorias de los fragmentos se utilizó el PdbID: 1A5E. En la figura 6.4C, se ve que en todos los valores de Q_w la formación de la estructura de hélice α de las repeticiones es la correcta, excepto para los valores de Q_w de 0,43 y 0,44. Con respecto a la orientación de las repeticiones en la estructura solamente es correcta solamente para valores de Q>0,6.

En la figura 6.5C se muestra en mapa de calor el mapa de contactos para la proteína P16, se observa que todas las repeticiones están formadas y que las repeticiones 2, 3 y 4 forman contactos no nativos. La frecuencia de contactos no nativos es baja si la comparamos con 1N0R (fig. 6.3B) y con I$\kappa\beta\alpha$ de 4 repeticiones (fig. 6.5A), además los valor de Q_w de los resultados de recocido simulado en su mayoría fueron mayores a 0,6. Estos resultados indican que P16 comienza a plegarse por el N-Terminal y que el paisaje energético de P16 tiene poca frustración, porque se vio que la proteína no queda atrapada en una conformación que se corresponde con un mínimo local. Por último analizamos la proteína Notch, que es una proteína de señalización compuesta de 6 repeticiones de ankirinas, una de sus características más importantes es que la repetición ubicada en el N-Terminal posee rasgos intrínsecos (Ehebauer *et al.*, 2005). En esta sección se analizarán los resultados de los experimentos de recocido simulado para Notch. Para las memorias de los fragmentos se utilizó el PdbID: 1OT8. En la figura 6.4D se muestran los mejores valores de Q_w obtenidos en las deferentes corridas de recocido simulado, los valores de Q_w fluctúan entre 0,38 y 0,73 mientras que los valores de RMSD entre 1,668 y 9,888 Å. Además se ve que en todos los valores de $Q_w > 0{,}4$, la formación de la estructura de hélice α de las repeticiones es la correcta, pero la localización de las repeticiones es correcta solamente para valores de Q>0,6.

Si comparamos Notch con I$\kappa\beta\alpha$, ya que ambas tienen 6 repeticiones, vemos que los valores de Q_w de las trayectorias son similares, en ambas proteínas se observó que el valor de Q_w varía entre 0,35 y 0.78 aproximadamente. Sin embargo, para Notch comenzamos a ver una estructura parecida a la nativa para valores de $Q_w > 0{,}55$ y para I$\kappa\beta\alpha$ lo vemos a valores de

$Q_w \geq 0{,}68$.

En la figura 6.5D se muestra en mapa de calor el mapa de contactos para valores de Q_w de entre 0,3 y 0,6, para Notch vemos que solamente la repetición ubicada en el C-Terminal forma contactos no nativos. Esto indica que a valor bajos de Q_w la proteína adquiere conformaciones muy similares a su estado nativo.

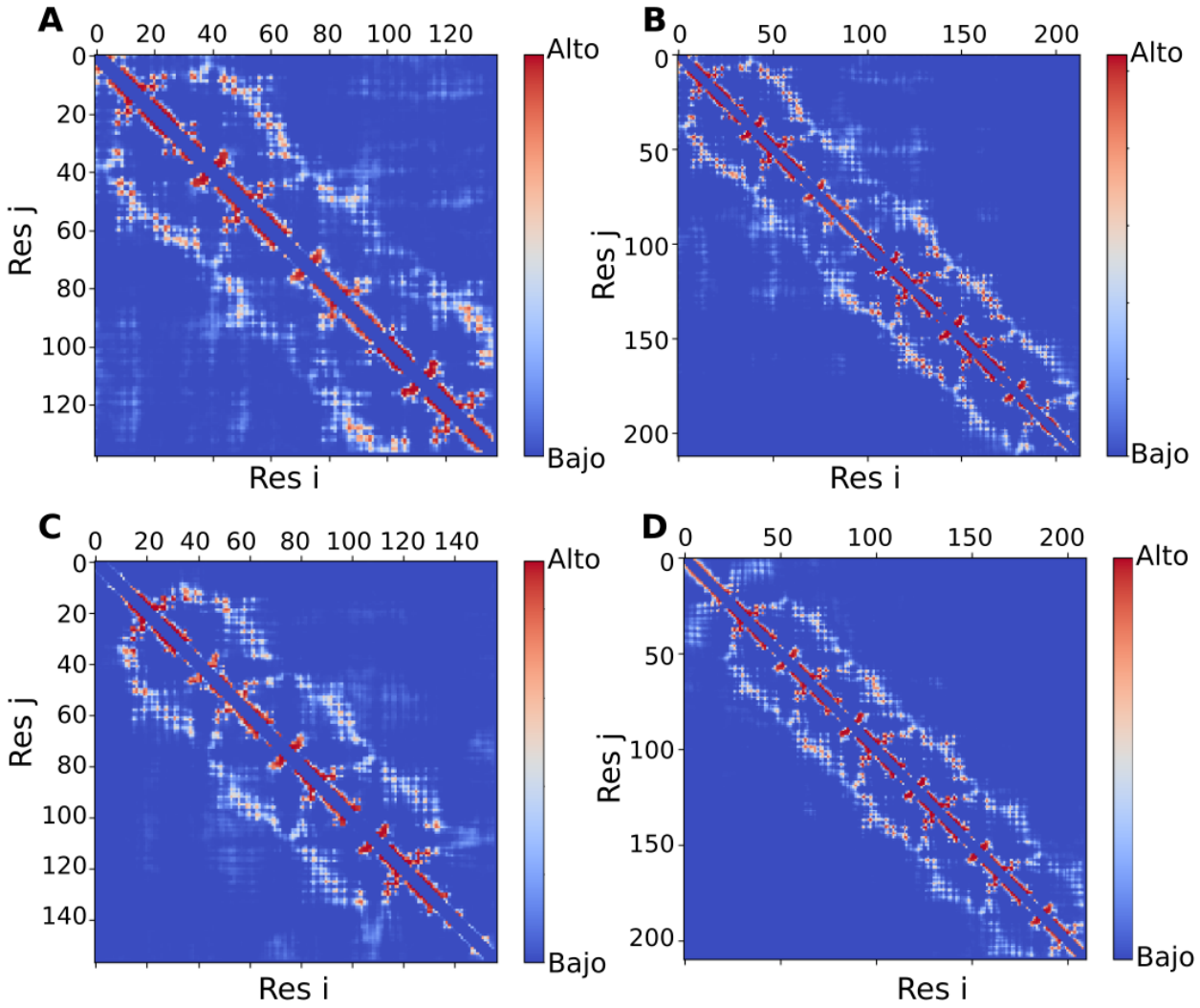

Figura 6.5: Mapa de calor de los mapas de contacto para los valores de Q_w de entre 0,3 y 0,6 para las proteínas A) I$\kappa\beta\alpha$ de 4 repeticiones. B) I$\kappa\beta\alpha$ de 6 repeticiones. C) P16. D) Notch.

6.3. Análisis de *Umbrella Sampling* en proteínas de la familia de ankirinas

El método de *Umbrella Sampling* es un método de muestreo aumentado que se utiliza para calcular una aproximación del perfil de energía libre de un proceso. Para ello, al hamiltoniano del sistema (ver métodos) se le agregan términos armónicos para restringir al sistema y de este modo obtener un valor sesgado de la energía del sistema. Se imponen potenciales de restricción a una molécula de estudio para limitarla a un lugar definido a lo largo de una variable colectiva, permitiendo así un muestreo del espacio de las fases en lugares que no son accesibles por la molécula debido a la presencia de barreras energéticas.

En la práctica, se corren simulaciones de dinámica molecular a la temperatura de plegado (T_f) usando como coordenada colectiva el valor de Q_W que toma valores entre 0 (totalmente desplegada) y 1 (totalmente plegada). Para aplicar el método de *Umbrella Sampling* tenemos que anclar al sistema en nuestra variable colectiva, es decir, en un determinado valor de Q_W, para poder explorar todas aquellas estructuras que posean ese valor cercano de Q_W.

Recientemente se ha usado el campo de fuerza de *AWSEM* para realizar estudios de *Umbrella Sampling* en proteínas metamórficas (Galaz-Davison *et al.*, 2021). En esta sección pondremos a prueba el campo de fuerza de *AWSEM* para hacer análisis de *Umbrella Sampling* en proteínas de la familia de Ankirinas. Se corrieron 21 trayectorias para los diferentes sesgos de *umbrellas*. Para cada trayectoria se simuló el sistema durante 1 millón de pasos. Luego las 21 trayectorias fueron integradas usando el método WHAM (*Weighted Histogram Analysis Method*) que es un método de análisis de histogramas ponderados. Para encontrar los parámetros que mejor se ajustaron a nuestras proteínas, se realizaron pruebas a distintos valores del potencial armónico, se corrieron entre 5 y 10 temperaturas cercanas a la temperatura de plegado y se probaron diferentes valores para las energía de las memorias.

6.3.1. Ankirinas diseñadas por consenso

Como ya se mencionó, se corrieron simulaciones de dinámica molecular del tipo grano grueso para hacer una exploración del paisaje energético de las proteínas de estudio, usando como coordenada global de reacción el parámetro Q_w. En este apartado se analizarán los resultados

de *Umbrella Sampling* para las proteínas, 1N0R, 3ANK,5ANK y 6ANK.

Para la proteína 1N0R, se realizaron simulaciones a una temperatura constante igual a la temperatura de plegado (T_f) que se estimó en 709 ºK, para 3ANK la T_f se estimó en 703 ºK, para 5ANK en 712 ºK y para 6ANK en 721 ºK y el potencial armónico usado fue de 1200 kcal/(mol$\mathring{A}^2$). Es importante destacar las temperaturas expresadas no son temperaturas reales, esto se debe a que el solvente no es modelado de forma explícita y que los valores del peso del sesgo del *Umbrella Sampling* también modifican los valores de las temperaturas a las que se realizan las simulaciones. Para calcular los perfiles de energía libre se usaron los resultados de todas las trayectorias. Los valores de energía libre están expresados en Kcal/mol. Se puede observar que los valores de Q_w (fig. 6.6A-D) para todas las proteínas están correctamente muestreados, porque vemos un solapamiento de los histogramas a los diferentes valores de Q_w. Además podemos ver que los histogramas para los valores de Q_w más bajos (entre 0 y 0,3) los histogramas son más angostos y tienen mayor frecuencia. Esto puede significar que en aproximadamente para el valor de Q_w de 0,3 puede haber una barrera energética. En

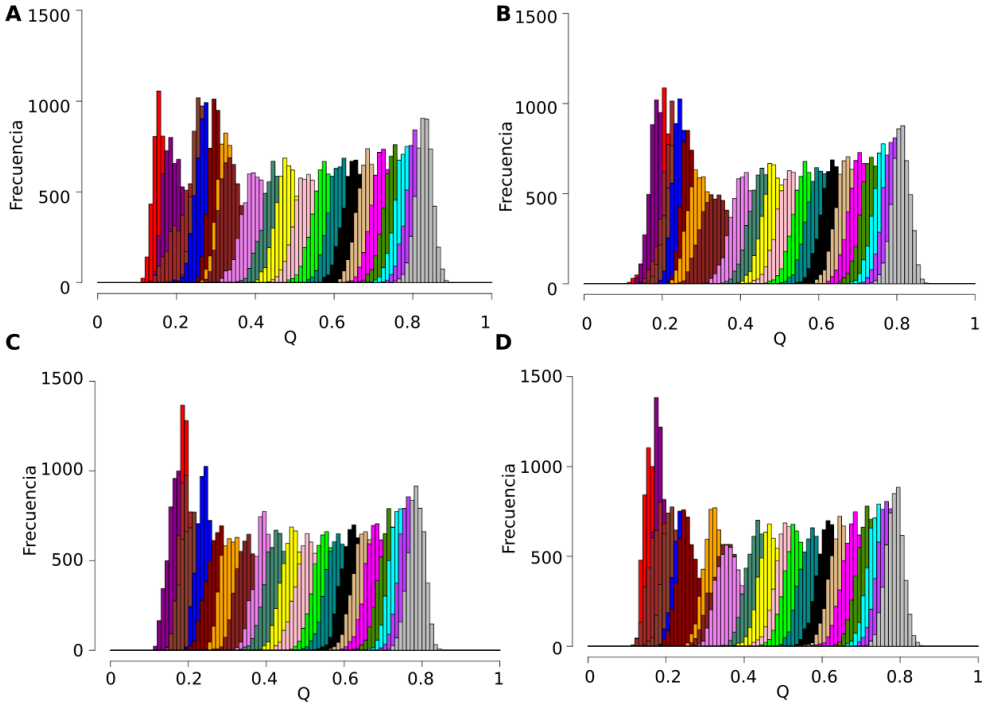

Figura 6.6: Histogramas del muestreo de los valores de Q_w de las proteínas. A) 3ANK. B) 1N0R. C) 5ANK. D) 6ANK. Cada color representa los intervalos de Q_w muestreados.

la figura 6.7A-D, se muestran los perfiles de energía libre en función del valor de Q_w para

las proteínas 3ANK, 1N0R, 5ANK y 6ANK. Se observa que el mecanismo de plegado es de dos estados, en el cual los estados desplegado y plegados están separados por una barrera energética. Para la proteína 3ANK (fig. 6.7A) se ve una barrera energética muy baja al valor de Q_w de 0,35 y el valor de la barrera es de 0,83 Kcal/mol y además se puede observar una pequeña barrera de 0,24 Kcal/mol a Q_w de 0,5. En la proteína 1N0R (fig. 6.7B) se observa una barrera energética muy baja de 0,905 Kcal/mol al valor de Q_w de 0,35. En las proteínas 5ANK (fig. 6.7C) y 6ANK (fig. 6.7D), también vemos una barrera energética muy baja, de 0,35 y de 0,17 respectivamente, ambas al valor de 0,35 de Q_w. Lo que vemos en común, para las 4 proteínas es que la barrera energética está a valor de Q_w de 0,35. Por otro lado, para las proteínas 5ANK y 6ANK, vemos que los valores de Q_w del estado plegado son bajos, 0,45 y 0,4 respectivamente, esto puede deberse a que, como vimos en los análisis de recocido simulado, para valores de Q_w mayores a 0,4, la estructura secundaria de las repeticiones ya estaban correctamente formadas, solamente la localización era incorrecta. Lo que nos puede estar indicando es que, el estado plegado está poblado de conformaciones con la correcta formación de las repeticiones pero no con la correcta localización de las repeticiones en la estructura. Debido a que las barreras energética observadas en estas proteínas son muy bajas, lo cual además puede estar relacionada a un error en la medición porque los valores no superan el 1 Kcal/mol, decidimos usar una segunda coordenada de reacción, el radio de giro (Rg) de las estructuras.

El la figura 6.8A-D, se muestran los valores de energía libre en función de Q_w y del Rg. Lo que se observa para las proteínas 3ANK, 1N0R y 5ANK, figuras figura 6.8A-C respectivamente, es bastante similar. Se ve una región de estructuras de baja energía, que se corresponden al estado desplegado (U), a valores de radio de giro de entre aproximadamente 12 y 14Å. En las estructuras que se corresponden con el estado desplegado, no se ven proteínas totalmente desplegada y elongada, sino que se ven proteínas que está en estado colapsado. Entre el estado desplegado y plegado, se puede ver una barrera energética de aproximadamente 8 Kcal/mol, además se ve una región más ancha que se corresponde al estado plegado y con respecto a las estructuras, se observa una correcta localización y formación de las estructuras secundarias. Para la proteína 6ANK, en el estado desplegado a valores cercanos a 18Åde radio de giro, se ve una región más compacta y separada del estado nativo por una barre-

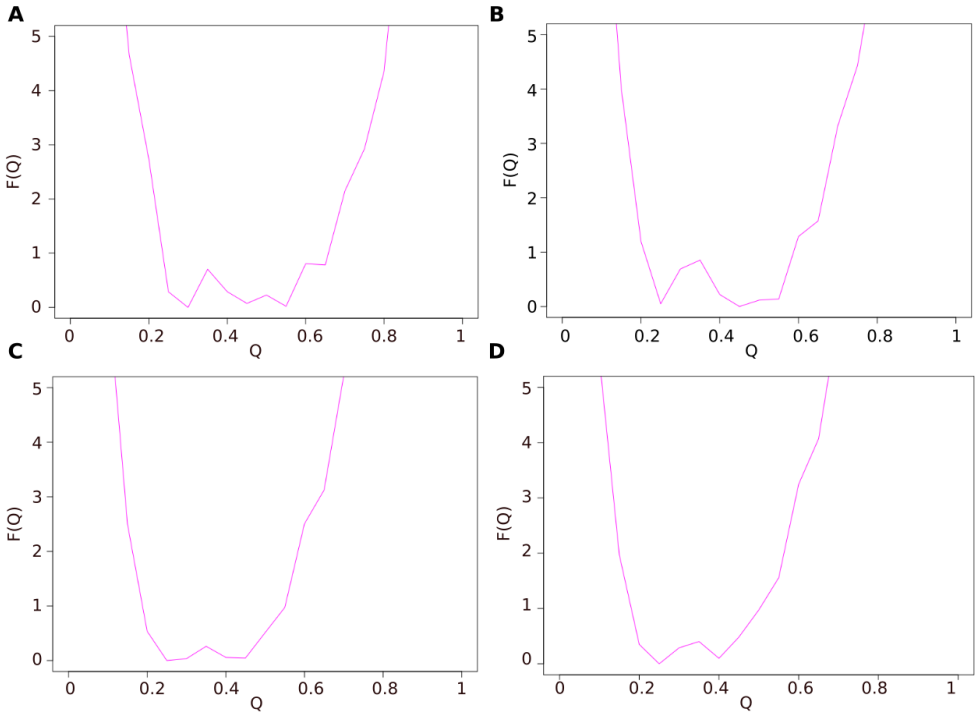

Figura 6.7: Perfiles de energía libre en función de Q_w para las proteínas. A) 3ANK. B) 1N0R. C) 5ANK. D) 6ANK.

ra de casi 10 Kcal/mol. El estado plegado de 6ANK es más amplio, en donde se observan conformaciones en las que las repeticiones localizadas en el C-Terminal están desplegadas, esto significa que cuando la proteína tiene un mayor radio de giro, cercano a 25Å, dónde se observa una conformación más bien elongada, comienza a plegarse hacia el estado nativo sin barrera energética. Estos resultados demuestran que el estado desplegado de las ankirinas es una conformación compactada, en la cual se observa la presencia de algunas estructuras locales similares a la estructura de una repetición. Por otro lado, en el estado plegado se observan conformaciones en las cuales la estructura secundaria de las repeticiones están bien plegadas pero la localización de las misma no es la correcta y también se ven conformaciones en las cuales la estructura de la proteína es similar a la del estado nativo.

Dada una estructura con un valor específico de Q_w se puede evaluar, para cada uno de los residuo, de todas las interacciones que forma, cuantas de ellas están presentes en el estado nativo, lo cual se define como el Qw local. A fin de evaluar los mecanismos de plegado se agruparon todas las estructuras de las trayectorias según sus valores de Q_w y se calculo el valor promedio del Qw local para cada residuo para cada valor de Q_w. En la figura 6.9A-D,

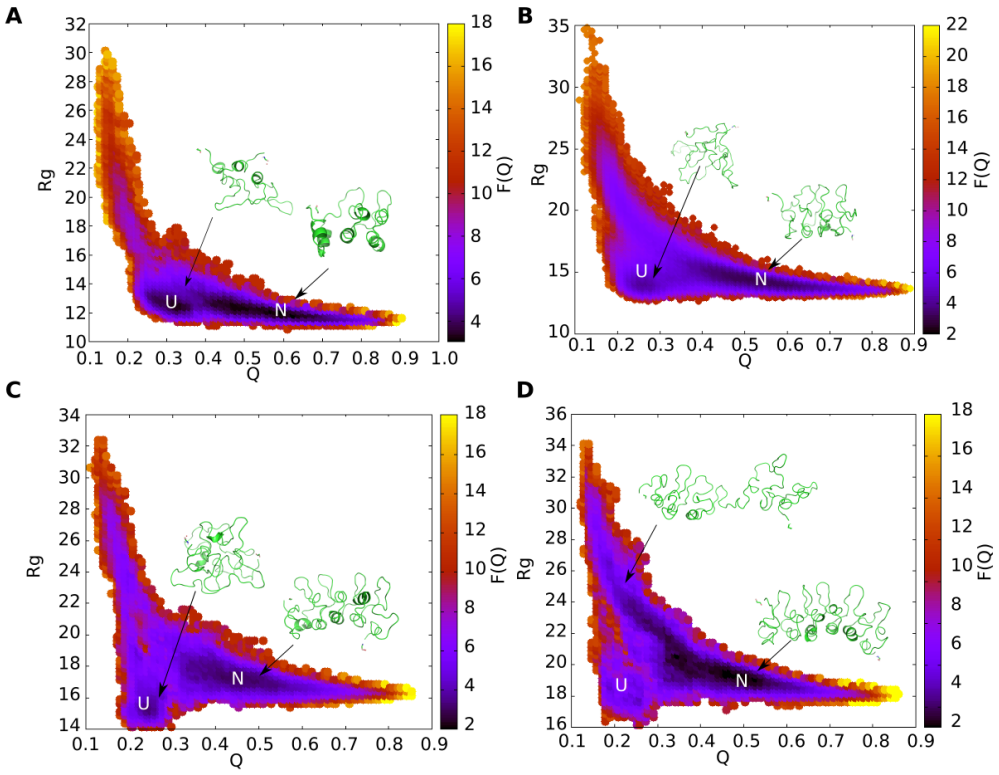

Figura 6.8: Perfiles de energía libre en función de Q_w y el radio de giro (Rg) para las proteínas. A) 3ANK. B) 1N0R. C) 5ANK. D) 6ANK. Las regiones correspondientes con el estado desplegado y plegado se marcan en los gráficos con las letras U y N respectivamente.

se observa el mecanismo de plegado de las proteínas 3ANK, 1N0R, 5ANK y 6ANK respectivamente. Cada repetición está marcada en un recuadro de líneas punteadas negras. Vemos que las repeticiones ubicadas en el C-Terminal son las primeras en plegarse, esto esta relacionado a que estas repeticiones forman menor cantidad de contactos nativos que las demás. Esto indica que el plegado no es cooperativo, porque no vemos que todas las repeticiones se forman al mismo tiempo, sino que comienzan a plegarse las del C-Terminal y luego las demás. También se observa que a valores mayores de 0,4 de Q_w la estructura de las repeticiones ya está casi completamente formada.

6.3.2. Ankirinas naturales

En este apartado se analizarán los resultados de las simulaciones de *Umbrella Sampling* para las proteínas I$\kappa\beta\alpha$ (de 4 y 6 repeticiones), P16 y Notch. La metodología aplicada fue la misma que se aplicó para las proteínas diseñadas por consenso.

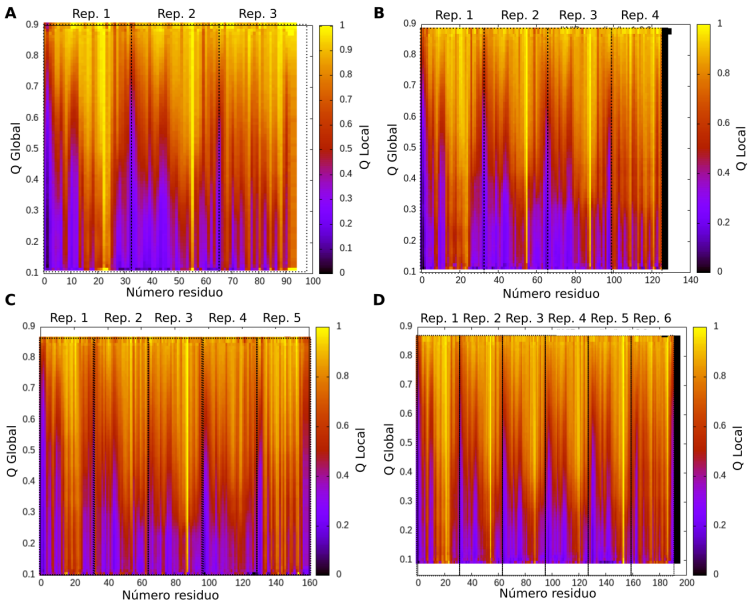

Figura 6.9: Mecanismo de plegado por residuo de las proteínas A) 3ANK. B) 1N0R. C) 5ANK. D) 6ANK. Valor del Q_w global (eje y) y en colores se representa el promedio del valor del Q_w local. Los recuadros de líneas punteadas negras marcan los límites entre repeticiones adyacentes.

Para la proteína I$\kappa\beta\alpha$ de 4 repeticiones, se realizaron simulaciones a una temperatura constante igual a la temperatura de plegado (T_f) que se estimó en 671 ^{o}K, para la de 6 repeticiones la T_f se estimó en 675 ^{o}K. Para la proteína P16 la T_f se estimó en 670 ^{o}K y para Notch en 675 ^{o}K y el potencial armónico usado fue de 1500 kcal/(mol$\mathring{A}^2$). Para calcular los perfiles de energía libre se usaron los resultados de todas las trayectorias. Los valores de energía libre están expresados en Kcal/mol.

Se puede observar que los valores de Q_w (fig. 6.10A-D) para todas las proteínas están correctamente muestreados, por el solapamiento de los histogramas a los diferentes valores de Q_w. Para la proteína I$\kappa\beta\alpha$ de 4 repeticiones, podemos ver que los histogramas entre los valores de Q_w entre 0,1 y 0,25 y entre 0,35 y 0,4 son más altos y mas angostos que el resto, esto puede significar la existencia de una barrera energética porque la proteína esta quedando atrapada en ese valor Q_w. Para el resto de las proteínas podemos ver que los histogramas entre 0,1 y 0,2 son más angostos y más altos.

En la figura 6.11A-D, se muestran los perfiles de energía libre en función del valor de Q_w para las proteínas, I$\kappa\beta\alpha$ de 4 repeticiones, I$\kappa\beta\alpha$ de 6 repeticiones, P16 y Notch, respectivamente.

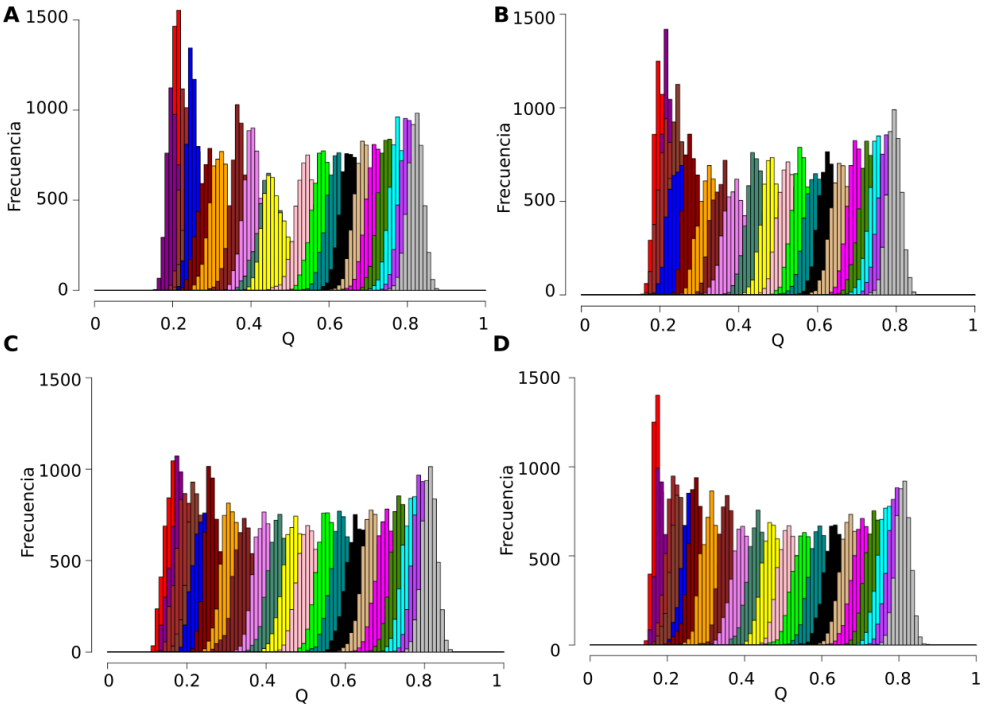

Figura 6.10: Histogramas del muestreo de los valores de Q_w de las proteínas. A) I$\kappa\beta\alpha$ de 4 repeticiones. B) I$\kappa\beta\alpha$ de 6 repeticiones. C) P16. D) Notch.

Para la proteína I$\kappa\beta\alpha$ de 4 repeticiones (fig. 6.11A) se ven dos barreras energéticas muy bajas, una al valor de Q_w de 0,35 y el valor de la barrera es de 0,29 Kcal/mol y la otra barrera de 0,35 Kcal/mol a Q_w de 0,45. En la proteína I$\kappa\beta\alpha$ de 6 repeticiones (fig. 6.11B) también se observan dos barreras energéticas muy bajas, la primera de 0,24 Kcal/mol al valor de Q_w de 0,35 y la segunda a Q_w de 0,45 con un valor de 0,21 Kcal/mol. En las proteínas P16 (fig. 6.11C) y Notch (fig. 6.11D), vemos, en ambas proteínas, una sola barrera energética muy baja, de 0,50 y 0,28 respectivamente, ambas al valor de 0,35 de Q_w. Al igual de como se observó para las proteínas de ankirinas diseñadas, al ser barreras energéticas tan bajas puede deberse a errores de medición. Por este motivo, decidimos usar una segunda coordenada de reacción, el radio de giro (Rg) de las estructuras.

En la figura 6.12A-D, se muestran los valores de energía libre en función de Q_w y del Rg. Para las proteínas I$\kappa\beta\alpha$ de 4 y 6 repeticiones y para Notch (figs. 6.12ABD) se ve una región de estructuras de baja energía entre los valores de 0,2 y 0,3 de Q_w y de radio de giro de entre 12 y 14. Estas regiones se encuentra separadas, por una barrera energética alta, con el estado plegado, a un valor de aproximadamente 0,35 de Q_w. Si observamos la estructuras del estado

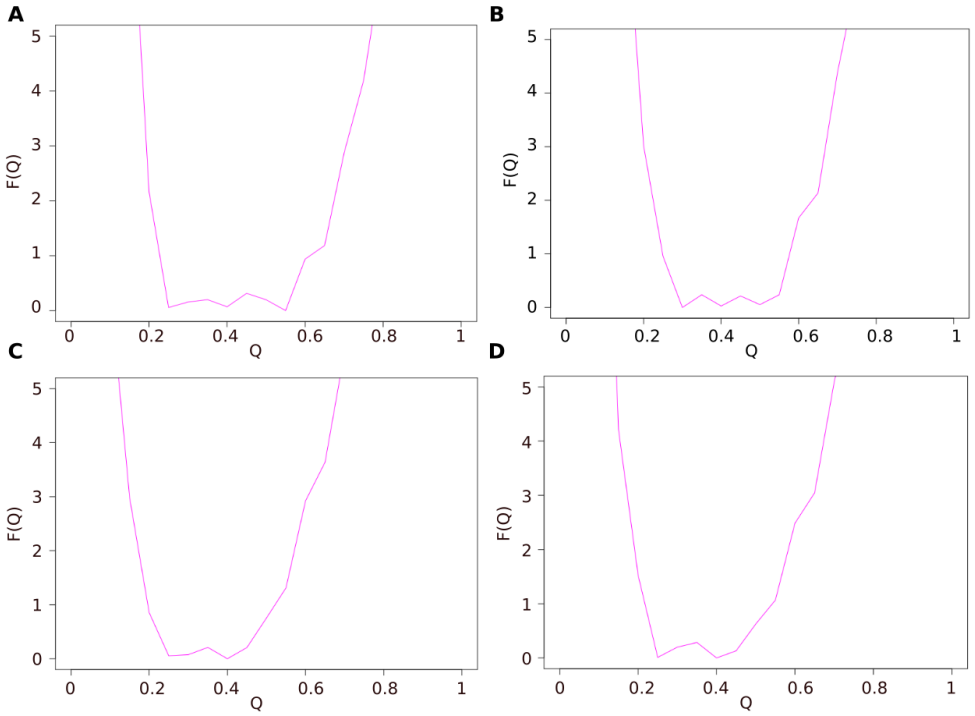

Figura 6.11: Perfiles de energía libre en función de Q_w para las proteínas. A) I$\kappa\beta\alpha$ de 4 repeticiones. B) I$\kappa\beta\alpha$ de 6 repeticiones. C) P16. D) Notch.

desplegado, vemos algo similar a lo que se vio para las proteínas diseñadas, una proteína en estado colapsado.

Para la proteína I$\kappa\beta\alpha$ de 4 repeticiones (fig. 6.12A), vemos un intermediario de plegado a un valor de Q_w de entre 0,3 y 0,35 y de radio de giro de entre 17 y 19 Å, en donde vemos una estructura elongada en la cual las repeticiones internas no están formadas y las terminales están parcialmente formadas. También vemos otra región de baja energía entre 0,35 y 0,45 que se corresponde al radio de giro entre 13 y 15Å, separada del estado nativo por una barrera energética más baja cercana a 3 Kcal/mol, en donde vemos una correcta formación de la estructura secundaria de las repeticiones pero están mal orientadas con respecto al estado nativo.

En I$\kappa\beta\alpha$ de 6 repeticiones (fig. 6.12B), vemos algo similar que para la de 4 repeticiones, un estado desplegado colapsado, un intermediario de plegado a valor de Q_w de 0,35 y radio de giro de entre 17 y 19Å, en donde vemos que las repeticiones internas (3, 4 y 5) están parcialmente formadas, mientras que las demás están completamente desplegadas. En la proteína Notch (fig. 6.12C), es difícil definir si hay un estado intermediario porque no se ve una barrera

energética clara con el estado nativo. Se observa un región muy ancha de baja energía que representa el estado nativo al valor de Q_w de aproximadamente 0,35 y radio de giro de 23Å. Al igual que se observó para 6ANK, una conformación elongada con repeticiones parcialmente formadas que comienza a plegarse hacia el estado nativo sin una barrera energética entre una conformación elongada y el estado nativo.

Por último analizamos la proteína P16, en donde vemos un estado desplegado a Q_w de entre 0,3 y 0,25 y radio de giro de entre 23 y 25, en este caso, a diferencia de las demás proteínas analizadas, si vemos un estado desplegado elongado. Sin embargo, también se ve una barrera energética cercana a 5 Kcal/mol entre el estado desplegado y plegado, con respecto a la región nativa, la igual que en 6ANK y Notch se ve que es muy ancha, lo cual indica que cuando la proteína adquiere una estructura eloganda con la formación parcial de repeticiones, se pliega sin barrera energética hacia el estado nativo.

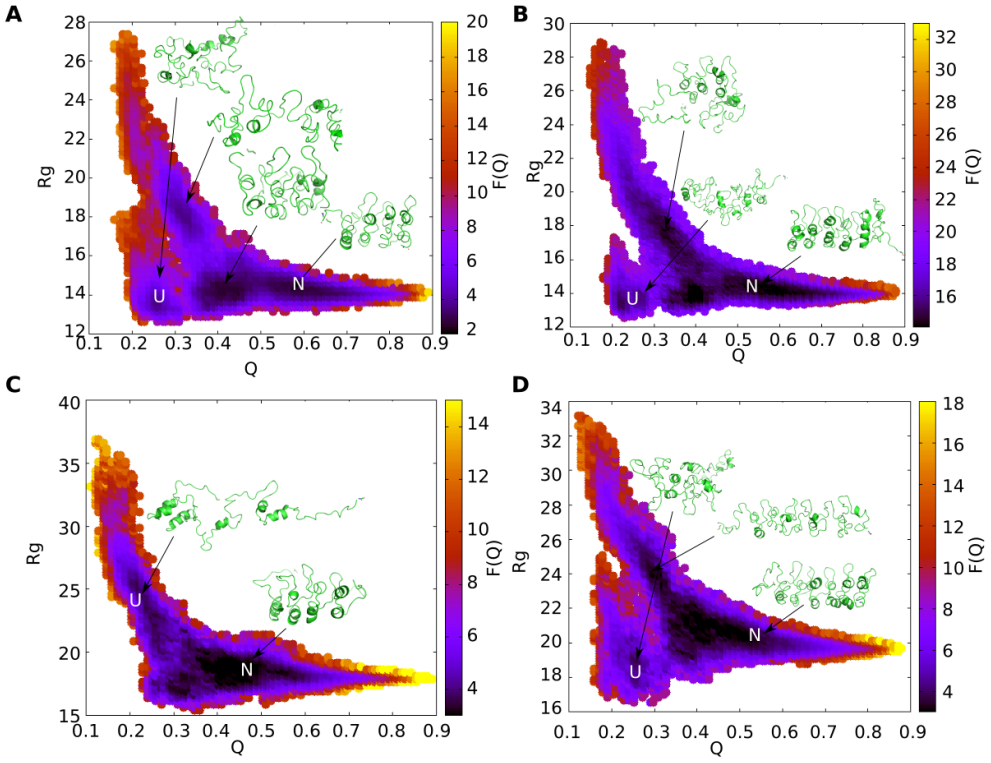

Figura 6.12: Perfiles de energía libre en función de Q_w y el radio de giro (Rg) para las proteínas. A) I$\kappa\beta\alpha$ de 4 repeticiones. B) I$\kappa\beta\alpha$ de 6 repeticiones. C) P16. D) Notch. Las regiones correspondientes con el estado desplegado y plegado se marcan en los gráficos con las letras U y N respectivamente.

En la figura 6.13A-D, se observa el mecanismo de plegado de las proteínas I$\kappa\beta\alpha$ de 4 re-

peticiones y de 6 repeticiones, P16 y Notch respectivamente. Cada repetición está marcada en un recuadro de líneas punteadas negras. Para I$\kappa\beta\alpha$ de 4 repeticiones vemos que las repeticiones internas son las últimas en plegarse, lo mismo veíamos en la figura 6.12A, que en el estado intermediario, las repeticiones internas eran las que no estaban plegadas. Para I$\kappa\beta\alpha$ de 6 repeticiones, se observa que las repeticiones parecerían estar formándose todas al mismo tiempo, lo mismo pasa para P16 y Notch. Si solo tenemos en cuenta los contactos de las repeticiones, ya que P16 tienen una región desordenada en sus terminales y Notch tienen una región que no es ankirina en el N-Terminal, se observa que a valores de Q_w mayores a aproximadamente 0,35 ya se comienzan a formar una gran cantidad de interacciones nativas en todas las repeticiones.

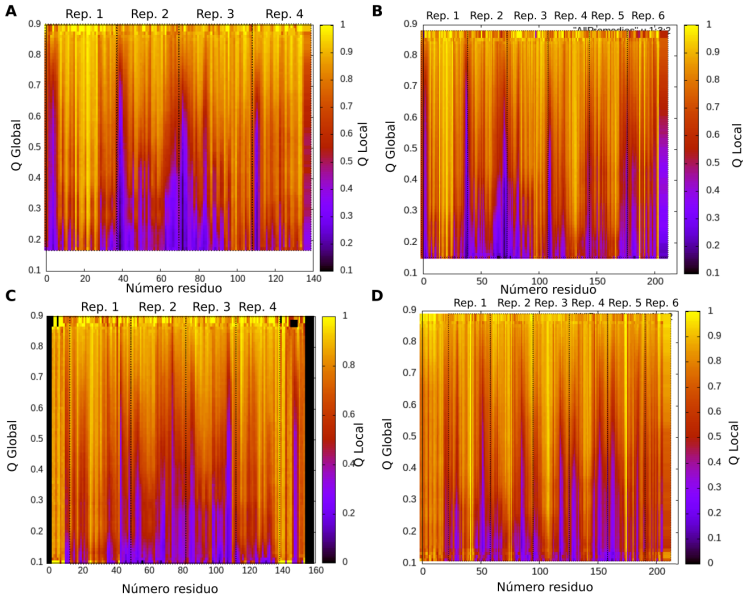

Figura 6.13: Mecanismo de plegado por residuo de las proteínas A) I$\kappa\beta\alpha$ de 4 repeticiones. B) I$\kappa\beta\alpha$ de 6 repeticiones. C) P16. D) Notch. Valor del Q_w global (eje y) y en colores se representa el promedio del valor del Q_w local. Los recuadros de líneas punteadas negras marcan los límites entre repeticiones adyacentes.

6.4. Conclusiones del capítulo

En este capítulo se aplicó el campo de fuerza de *AWSEM* para estudiar el plegado y los mecanismo de plegado de un conjunto de proteínas de la familia de ankirinas. Los resultados del recocido simulado de proteínas diseñadas por consenso lo que observamos fue que para

todas las proteínas se pudo obtener una correcta predicción de la estructura. La proteína de 3 repeticiones fue en la que se obtuvieron los mejores resultados, ya que el mejor valor de Qw obtenido fue de 0,88 y el más bajo de 0,62, probablemente sea debido a que las posibilidades conformacionales están más restringidas debido al largo de la estructura. Al analizar los contactos no nativos que se forman entre los valores de Qw entre 0,3 y 0,6 para todas las proteínas diseñadas, vimos que son diferentes, que cuanto mayor es la cantidad de repeticiones menor es la cantidad de contactos no nativos entre los valores de Qw analizados. Para las proteínas naturales, de los resultados del recocido simulado, también se obtuvieron una correcta predicción de la estructura de las proteínas analizadas. En I$\kappa\beta\alpha$ de 4 repeticiones se observaron los mejores valores de Q_w que fue de 0,77. Tanto para las diseñadas por consenso como para las naturales a valores de Q_w mayores a 0,50, cantidad de contactos no nativos disminuye casi en su totalidad, indicando que por encima de este valor ya se puede observar un correcto plegado de la proteína.

Al comparar los resultados de recocido simulado entre las proteína 1N0R y I$\kappa\beta\alpha$ de 4 repeticiones, proteínas con la misma cantidad de repeticiones, podemos observar resultados de valor de Q_w muy diferentes. En 1N0R los valores son más altos que en I$\kappa\beta\alpha$, esto puede estar dado a la simetría en secuencia de 1N0R.

El campo de fuerza de *AWSEM* esta basado en la teoría de paisajes energéticos de las proteínas, través de términos en su Hamiltoniano, *AWSEM* captura las interacciones esenciales que contribuyen a la estabilidad y la energía libre de las conformaciones de las proteínas. Este campo de fuerza tiene en cuenta las interacciones no nativas, las cuales contribuyen a la dinámica y flexibilidad de la proteína durante el proceso de plegado y pueden atrapar temporalmente a la proteína en conformaciones no nativas o mínimos locales del paisaje energético. Para llegar a la estructura nativa, la proteína debe superar estas barreras energéticas, lo que implica la formación de contactos nativos y la disminución de contactos no nativos. Es decir, que el análisis de contactos no nativos reflejan la complejidad del proceso de plegado de proteínas en un paisaje energético. Además se ha demostrado que la simetría desempeña un papel importante en el plegado de proteínas y que es más fácil encontrar secuencias con paisajes energéticos en embudo capaces de plegarse rápidamente si la estructura es simétrica (Wolynes, 1996). Lo mismo sucede si comparamos el modelo de 6ANK con I$\kappa\beta\alpha$ de 6

repeticiones, en donde se observaron mejores valores de Q_w para la proteína que contiene 6 repeticiones del consenso. Estos resultados pueden estar reflejando que la simetría en secuencia en las proteínas influye en el plegado ya que pudimos demostrar que las proteínas simétricas en secuencias obtuvieron mejores valores de Q_w.

De los análisis de *Umbrella Sampling*, de todos los perfiles de energía libre analizados, pudimos obtener barreras energéticas muy bajas cuando usamos solamente la coordenada de Q_w, esto se debe a que *AWSEM* está diseñado para plegar proteínas sin barreras energéticas altas. Si observamos los perfiles de energía libre agregando la coordenada de radio de giro, obtuvimos barreras energéticas más altas y en la mayoría de las proteínas no se observó en el estado desplegado una estructura totalmente elongada, sino que se observó un estado colapsado con un valor de radio de giro similar al del estado plegado. Esto puede estar indicando que *AWSEM* esta detectando dos estados desplegados uno elongado y otro colapsado.

Capítulo 7

Conclusiones Generales

En esta tesis se han diseñado, implementado y testeado una variedad de herramientas, experimentos y técnicas para el estudio de diferentes tipos y familias de proteínas a un nivel de secuencia y estructura de proteínas, así como también a nivel genómico. Los resultados obtenidos aportan nuevos conocimientos para el campo del estudio de la biología de las proteínas y no solo por toda la información nueva que aporta al campo, sino que también por todas las nuevas estrategias, utilizadas para el estudio de las mismas en el marco de esta tesis, se pueden aplicar para analizar cualquier proteína o familia de interés.

Una de las herramientas desarrolladas es un paquete R llamado FrustratometeR que facilita el cálculo de la frustración local energética en estructuras proteicas. Este paquete incluye nuevas funcionalidades que permiten evaluar el efecto de las mutaciones en la frustración local y analizar la frustración a lo largo de trayectorias de dinámica molecular. La sencilla interfaz del paquete y las nuevas funcionalidades implementadas hacen que el análisis de la frustración sea más accesible y útil para el estudio de proteínas a diferentes escalas.

Además, hemos desarrollado una herramienta llamada FrustraEvo basada en el FrustratometeR, que analiza patrones energéticos en familias de proteínas y revela restricciones físico-químicas relacionadas con la estabilidad y la función. Este enfoque proporciona una interpretación biofísica del impacto de la divergencia de secuencias a lo largo de escalas evolutivas. La conservación de la frustración dentro de las familias de proteínas puede utilizarse para definir regiones o residuos que están involucrados en la función y estabilidad de las proteínas.

Utilizando el FrustratometeR, hemos aplicado este enfoque al estudio de las regiones desordenadas de las proteínas, llegando a la conclusión de que las regiones *fuzzy* presentan un enriquecimiento de interacciones altamente frustradas. Las interacciones altamente frustradas generan un paisaje energético rugoso que abarca múltiples mínimos locales, lo que permite diversas actividades biológicas. A lo largo de la evolución, las proteínas han minimizado los conflictos energéticos sin comprometer su función, especialmente en el caso de las proteínas desordenadas, que se describen como un conjunto de conformaciones en lugar de una conformación bien definida.

Nuestros resultados demuestran que, incluso después de la unión, las interacciones en las regiones desordenadas de las proteínas siguen siendo subóptimas y presentan conflictos energéticos por resolver. Esto sugiere que las proteínas desordenadas no necesariamente requieren una conformación bien definida para su función, ya que la especificidad puede surgir de diferentes patrones de frustración en conformaciones altamente heterogéneas. La frustración y la rugosidad del paisaje energético permiten la versatilidad funcional junto con la especificidad.

En resumen, nuestro trabajo muestra que el análisis de la frustración local energética en proteínas y su conservación a lo largo de la evolución proporciona información valiosa sobre la adaptabilidad funcional, la especificidad de la interacción y la diversidad funcional de las proteínas con regiones *fuzzy*. Además, nuestro paquete FrustratometeR y la herramienta FrustraEvo tienen el potencial de mejorar el análisis estructural de proteínas.

Se analizó y caracterizó la arquitectura exón-intrón de proteínas con repeticiones de ankirinas en organismos eucariotas. Se encontró que la longitud de exón más frecuente era de 99 nt y que la mayoría de los exones de esta longitud codificaban para repeticiones de ankirinas. También se observó que los exones con una longitud múltiplo de 3 eran más frecuentes, lo que afectaba directamente a la fase del intrón, siendo la mayoría de los intrones de fase 0. Se identificaron patrones en la distribución de las clases de exones, encontrando que la mayoría de los exones pertenecían a la clase 0-0, relacionada con la duplicación y reclutamiento de exones. La presencia de exones simétricos y la alta frecuencia de intrones de fase 0 sugieren la posibilidad de eventos de barajado de exones. Se analizó la distribución de repeticiones en los exones y se observó que la mayoría de las repeticiones están codificadas por un solo exón. Sin embargo, también se encontraron repeticiones parciales codificadas por dos exones

consecutivos. Se observó que la posición de interrupción del marco de lectura por los intrones está conservada en algunas repeticiones, lo que indica que la fase estructural y genómica no siempre coinciden. Se estudiaron los eventos de empalme alternativo y se encontró que aproximadamente la mitad de las proteínas conservaron todas sus repeticiones después de este evento, mientras que en la otra mitad se eliminaron algunas repeticiones, principalmente las cercanas a los terminales.

Se utilizó el campo de fuerza de *AWSEM* para estudiar el plegado de proteínas de la familia de ankirinas. Se observó que *AWSEM* fue capaz de predecir correctamente la estructura de las proteínas diseñadas por consenso y las proteínas naturales analizadas. Los análisis de *Umbrella Sampling* revelaron barreras energéticas muy bajas en los perfiles de energía libre. Sin embargo, cuando incluimos una segunda coordenada de reacción, el radio de giro, pudimos encontrar barreras energéticas más altas y una mejor descripción de los mecanismos de plegado de las proteínas estudiadas. También se observó que las proteínas presentaban estados desplegados tanto elongados como colapsados, lo que podría indicar la presencia de dos estados desplegados distintos. Estos resultados demuestran que, el estado no nativo de baja energía es una conformación colapsada que luego se estira y se comienzan a formar las repeticiones de los terminales para finalmente plegarse hacia el estado nativo sin barrera energética.

Capítulo 8

Anexo

8.1. FrustratometeR

Para ver en detalle el conjunto de datos y los *scripts* usado para realizar el análisis de este capitulo puede visitar el siguiente enlace https://github.com/mariafreiberger/PhD-Thesis/tree/main/Chapter1

8.1.1. Caso de estudio:Hemoglobina

8.1.2. Caso de estudio: Factor de elongación bacteriano, RfaH

8.2. Proteínas *fuzzy*

Para ver en detalle el conjunto de datos y los *scripts* usado para realizar el análisis de las proteínas analizadas visitar el siguiente enlace https://github.com/mariafreiberger/PhD-Thesis/tree/main/Chapter2

8.3. Análisis genómico de las proteínas con repeticiones de Ankirina en eucariotas

Para ver en detalle el conjunto de datos y los *scripts* usado para realizar el análisis de este capitulo puede visitar el siguiente enlace

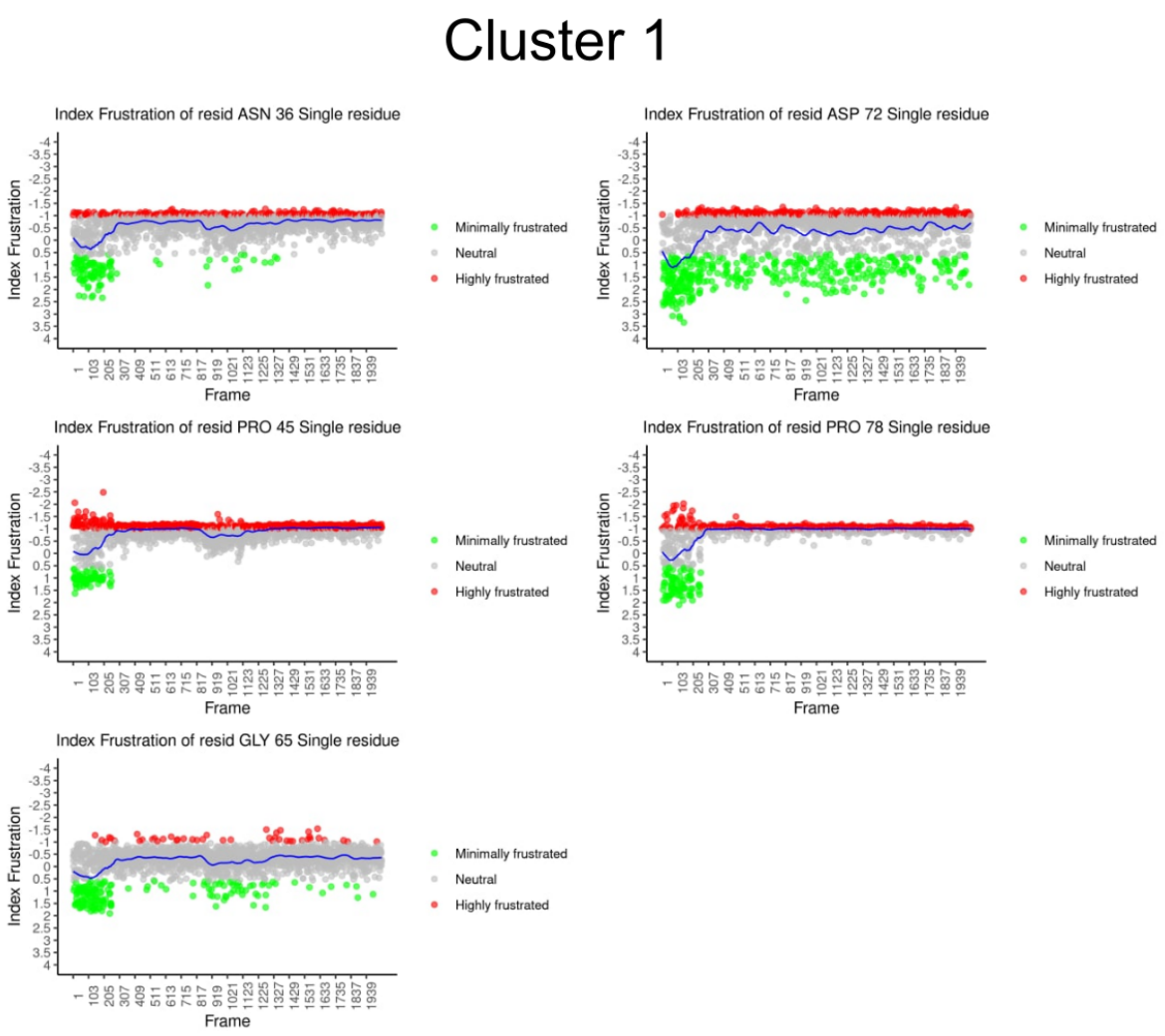

Figura 8.1: Valores de frustración en función del tiempo (fotogramas de simulación) para residuos en el *Cluster* 1.

https://github.com/mariafreiberger/PhD-Thesis/tree/main/Chapter3

8.4. Análisis el paisaje energético de proteínas de la familia de Ankirina y de sus mecanismos de plegado

Para ver en detalle el conjunto de datos y los *scripts* usado para realizar el análisis de este capitulo puede visitar el siguiente enlace

https://github.com/mariafreiberger/PhD-Thesis/tree/main/Chapter4

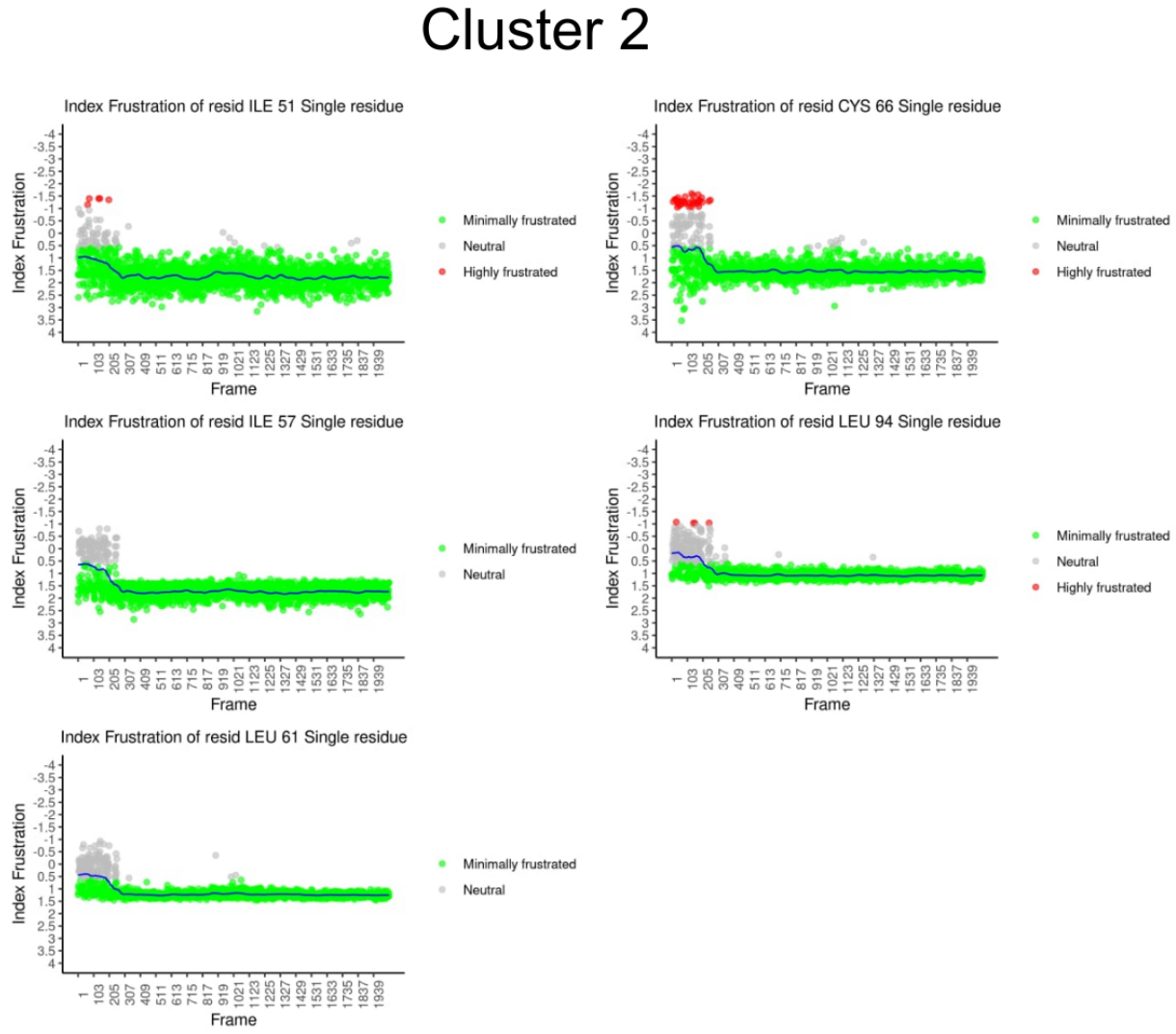

Figura 8.2: Valores de frustración en función del tiempo (fotogramas de simulación) para residuos en el *Cluster* 2.

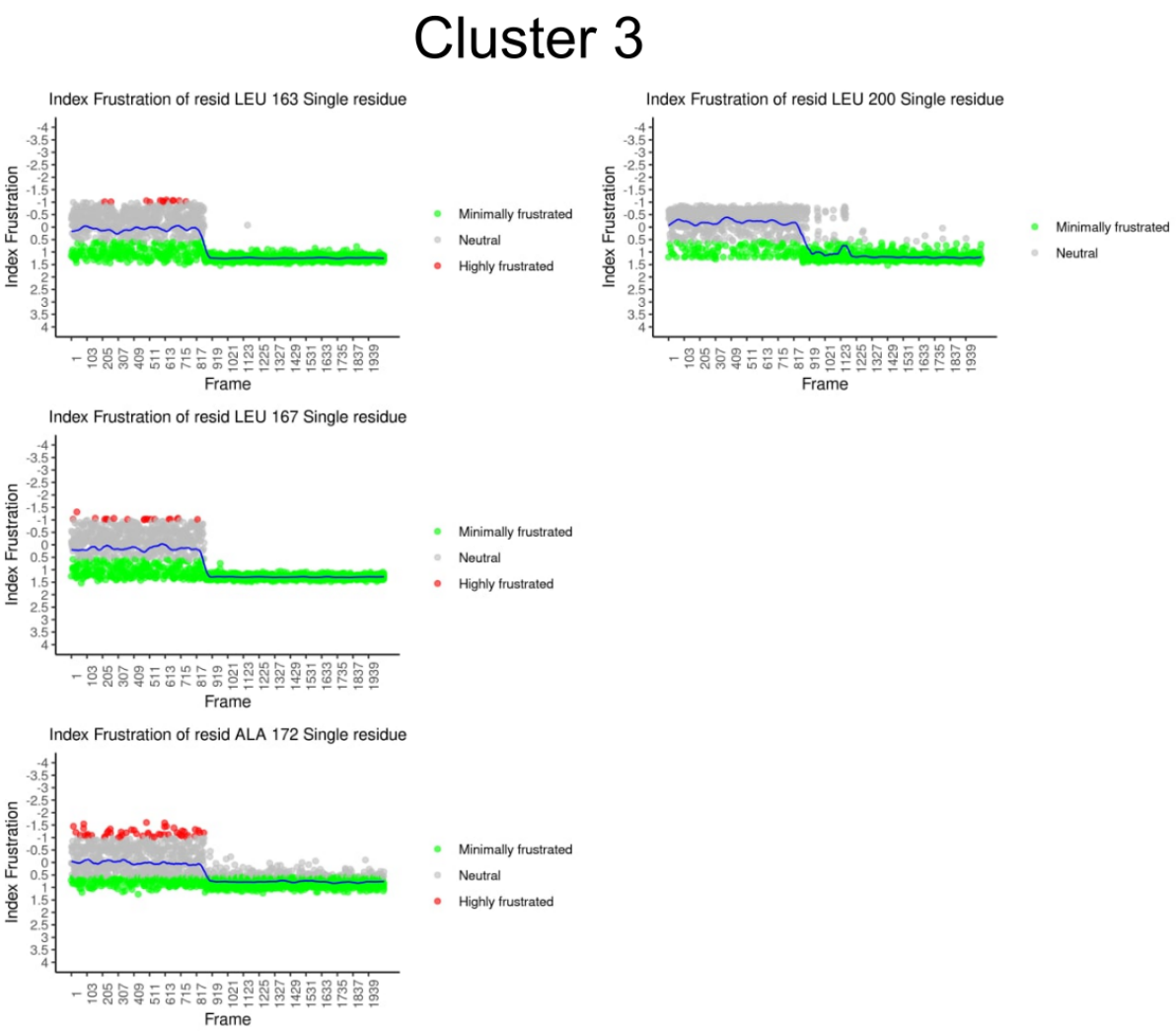

Figura 8.3: Valores de frustración en función del tiempo (fotogramas de simulación) para residuos en el *Cluster* 3.

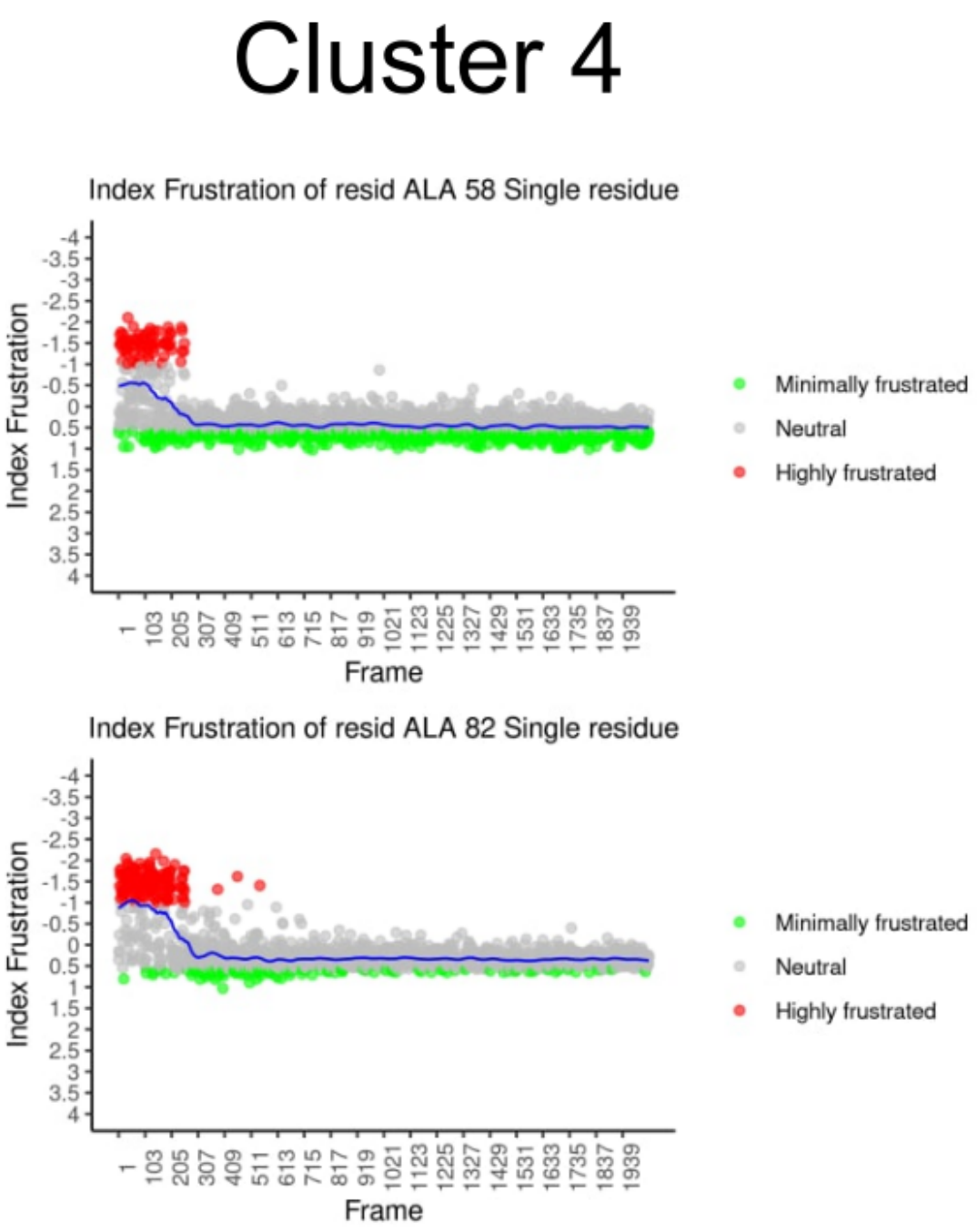

Figura 8.4: Valores de frustración en función del tiempo (fotogramas de simulación) para residuos en el *Cluster* 4.

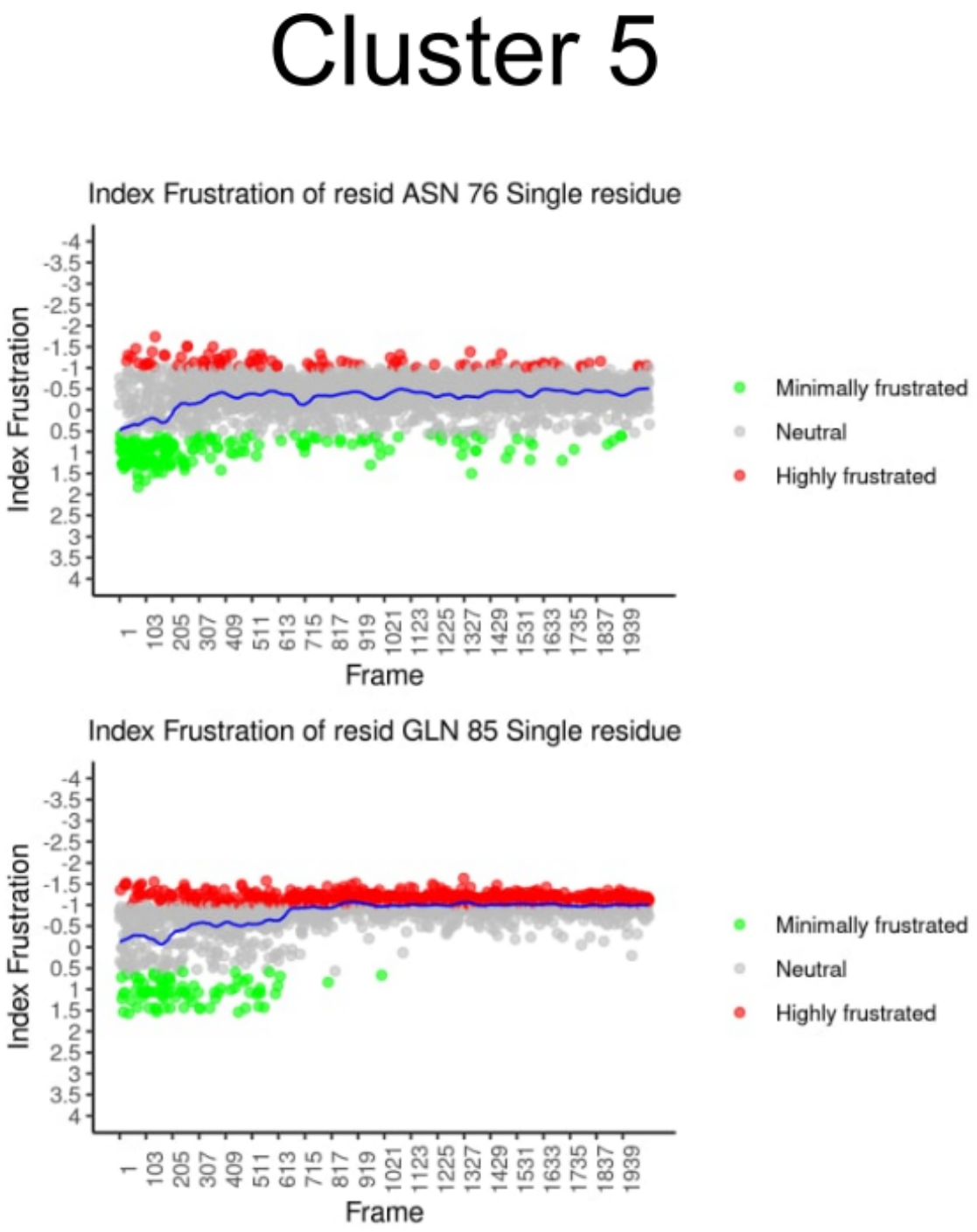

Figura 8.5: Valores de frustración en función del tiempo (fotogramas de simulación) para residuos en el *Cluster* 5.

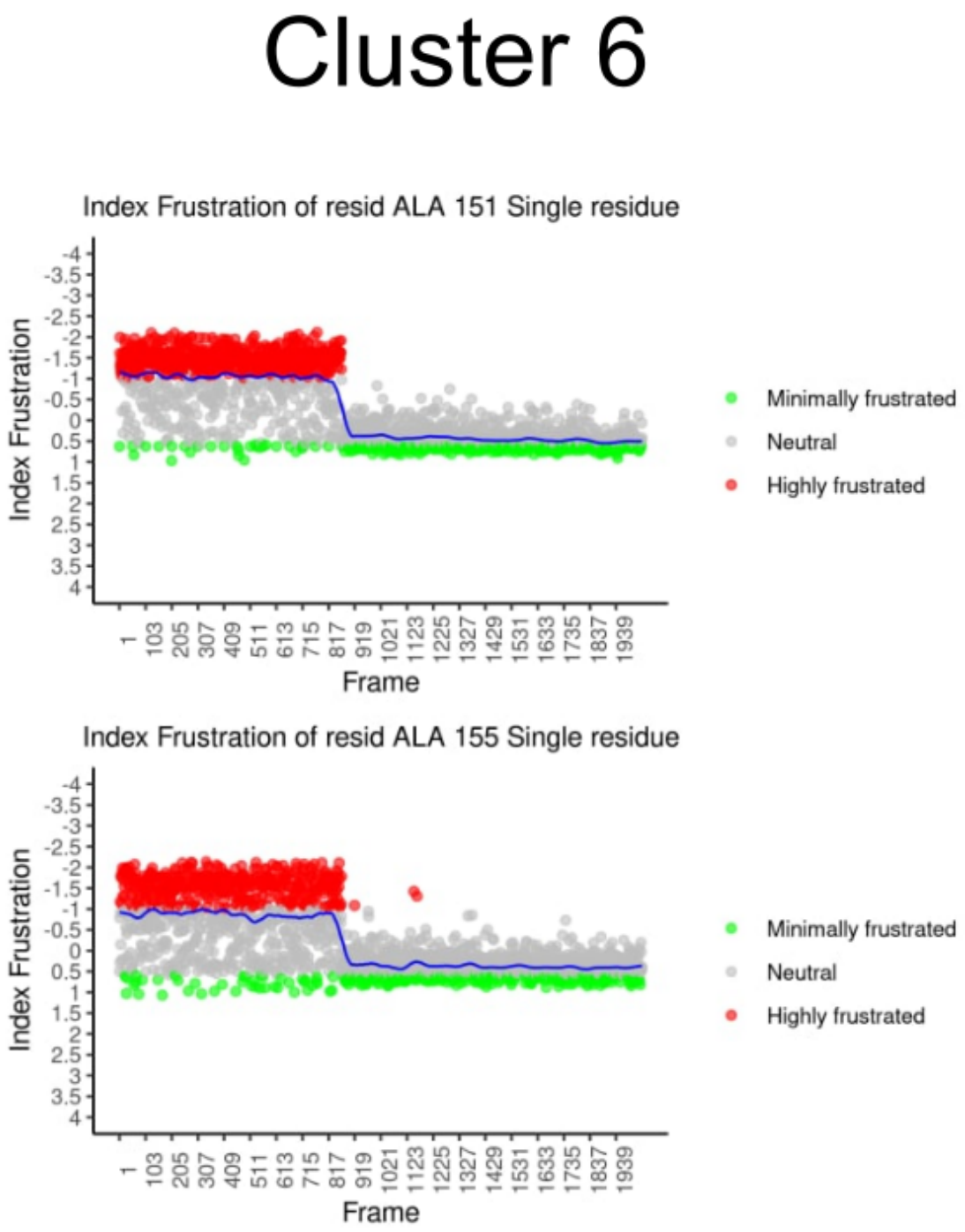

Figura 8.6: Valores de frustración en función del tiempo (fotogramas de simulación) para residuos en el *Cluster* 6.

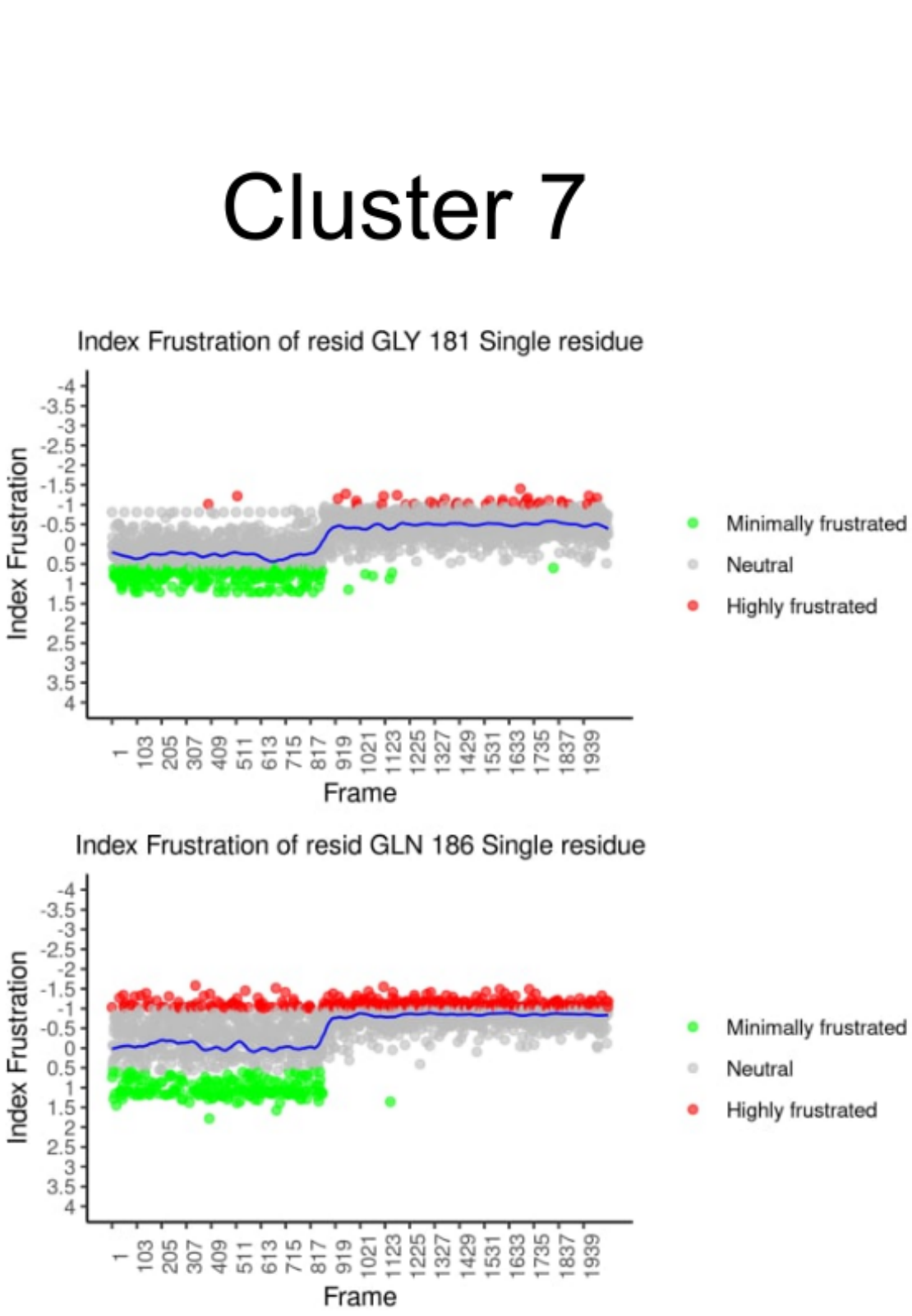

Figura 8.7: Valores de frustración en función del tiempo (fotogramas de simulación) para residuos en el *Cluster* 7.

Residuo	Conservación de la frustración	Bibliografía
Lys7α	1,02	Forma un puente salino con Asp74α (conservado y mínimamente frustrado) (Shaanan, 1983).
Glu27α	0,7	Forma un puente salino con His112α (Shaanan, 1983) (conservado y altamente frustrado).
Glu30α	0,7	Forma un puente salino con His50α (Shaanan, 1983) (conservado y altamente frustrado).
Tyr42α	1,29	Participa en una de las interacciones fuertes en la interfaz α1-β2, estableciendo un enlace de hidrógeno con β2, La ausencia de este enlace se observa realmente en una hemoglobina anormal, y desplaza el equilibrio alostérico hacia la forma oxi, pero no inhibe la formación de la estructura desoxi. En su ausencia no se forma la estructura deoxi cuaternaria y se inhiben todos los efectos cooperativos(Kavanaugh *et al.*, 2005). Tyr42α es también uno de los aminoácidos que interactúa con el grupo hemo en la hemoglobina humana (Kavanaugh *et al.*, 2005).
Lys99α	0,84	Forma puente salino con Arg141α. Lys99α da lugar a anemias cuando muta por aminoácidos que interrumpen las interacciones proteína-proteína con AHSP (Mollan *et al.*, 2010).
Ser124α	1,02	Cuando se muta por una Pro puede dar lugar a una interacción interrumpida con AHSP debido a cambios en la conformación de la hélice que interactúa en la α-globina, dando lugar a la α-talasemia (Bisconte *et al.*, 2015).
Asp126α	0,57	Desempeña un papel en la estabilización de la interfaz α1/α2 y se ha encontrado en la Hb Sassari (mutación D126H) dando lugar a una mayor afinidad por el oxígeno, posiblemente debido a perturbaciones en el equilibrio entre los dos estados cuaternarios (Sanna *et al.*, 1994).
Lys127α	1,29	La mutación de la Lys127α vecina también puede conducir a un aumento de las moléculas de unión al oxígeno, como ocurre con la variante natural Hb Waikato (mutación L127Q) (Moore *et al.*, 2021).
Asn57β	1,29	Presenta un aumento de la afinidad por el oxígeno observado en valores de pH ácido, que puede estar relacionado con una mayor disociación de la molécula en dímeros (Giardina *et al.*, 1978).
Glu101β	1,02	Interfaz con α-globina en humanos (Fermi, 1975; Shaanan, 1983).
Asn108β	0,7	Interfaz con α-globina en humanos (Fermi, 1975; Shaanan, 1983).

Tabla 8.1: Estado de conservación de la frustración de residuos funcionales relevantes en hemoglobinas.

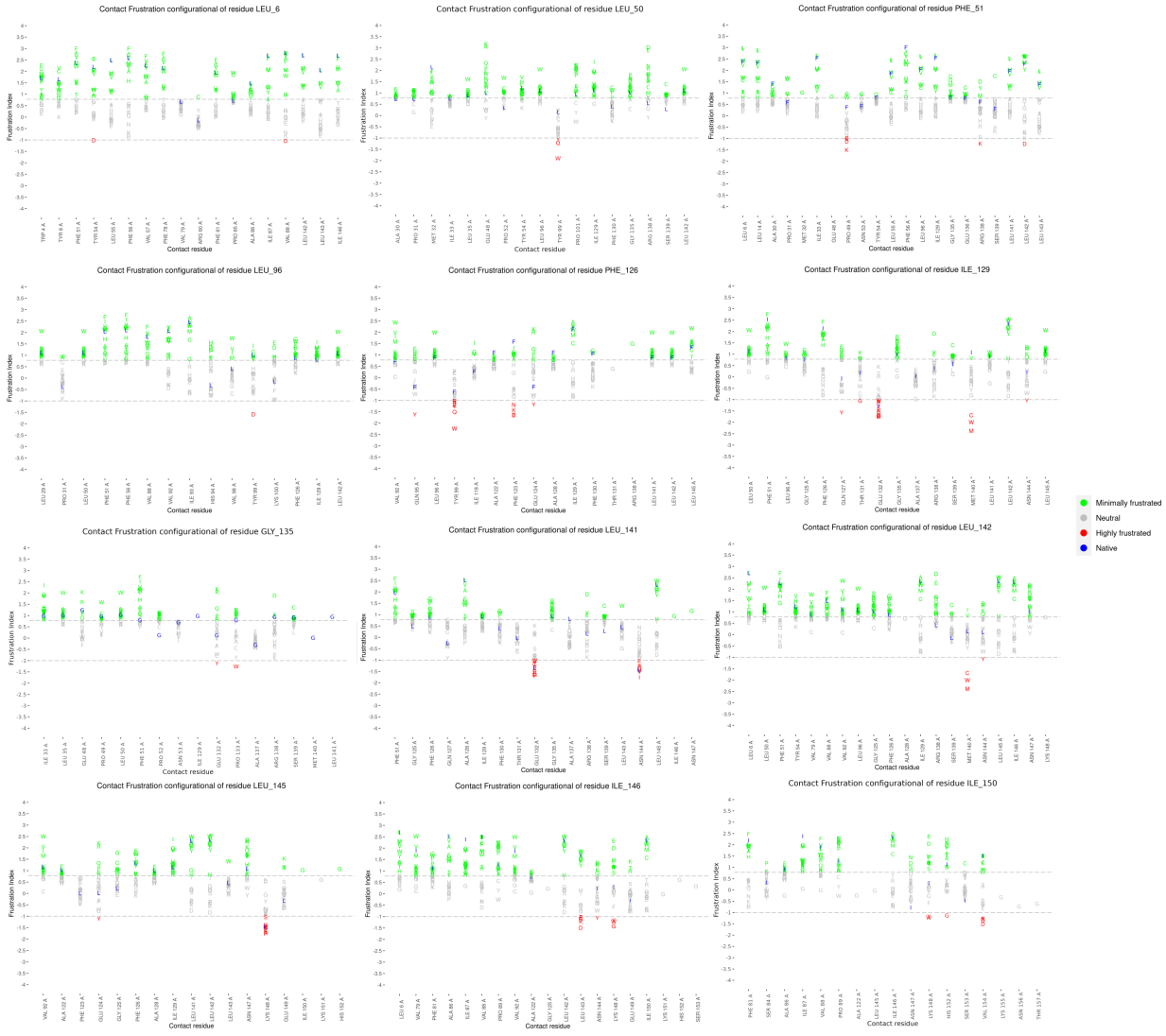

Figura 8.8: Cambios en la frustración local para las 19 mutaciones de los residuos interdominio utilizando FrustratometeR. El eje x muestra los residuos con los que el residuo, ya sea salvaje o mutado, establece contactos en la estructura. En el eje y se muestra el índice de frustración *configurational* de los contactos. La identidad del aminoácido de tipo silvestre se muestra en azul y las variantes se colorean según su estado de frustración.

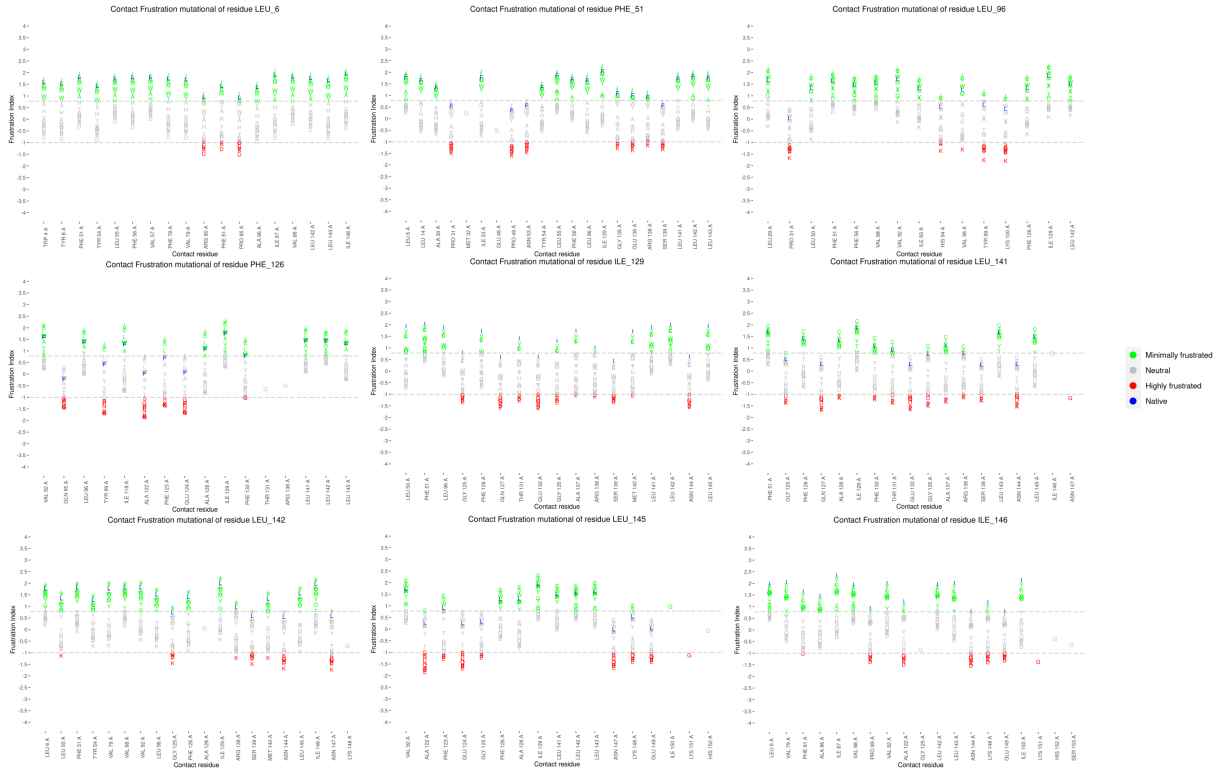

Figura 8.9: Cambios en la frustración local para las 19 mutaciones de los residuos interdominio utilizando FrustratometeR. El eje x muestra los residuos con los que el residuo, ya sea salvaje o mutado, establece contactos en la estructura. En el eje y se muestra el índice de frustración *mutational* de los contactos. La identidad del aminoácido de tipo salvaje se muestra en azul y las variantes se colorean según su estado de frustración.

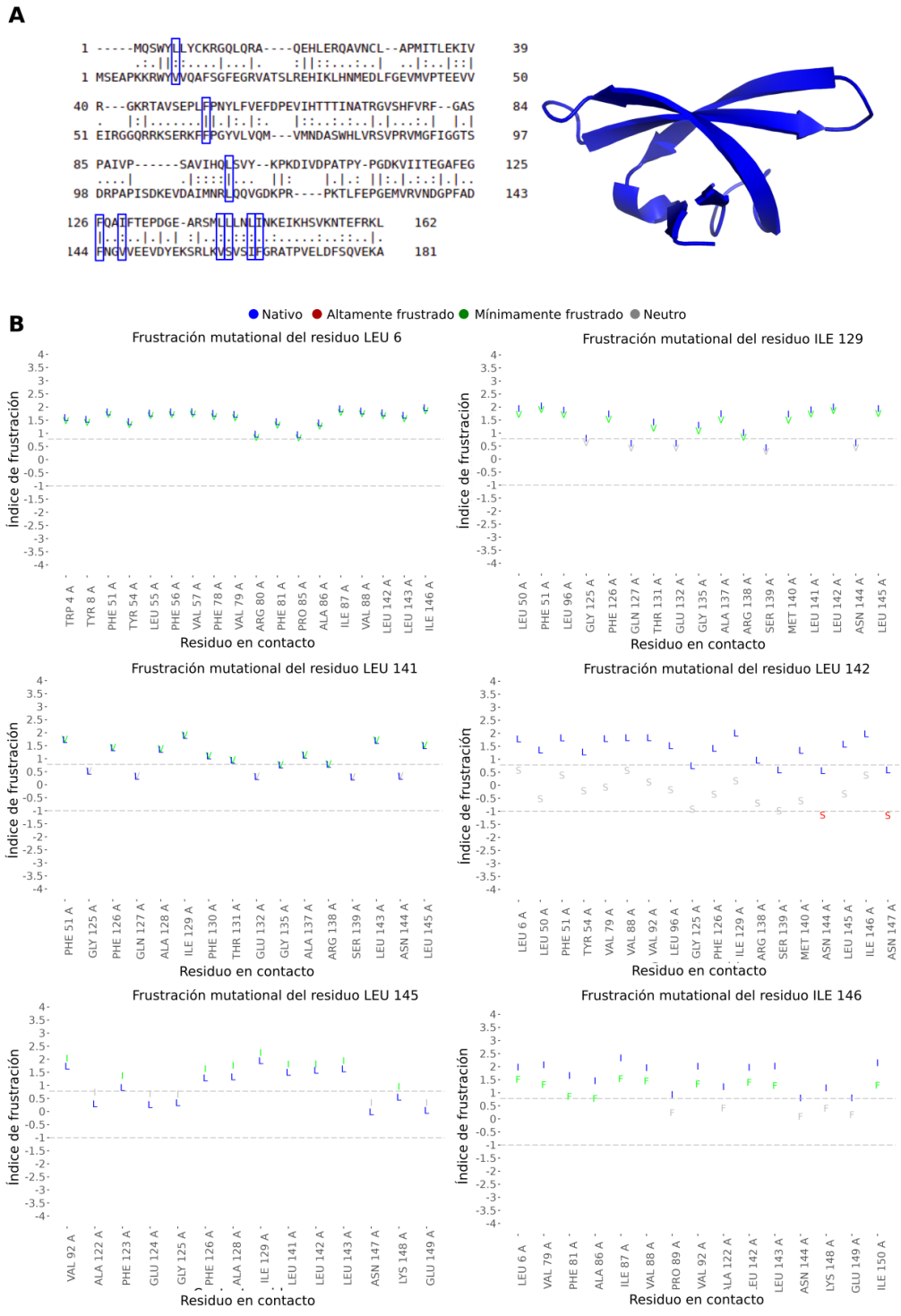

Figura 8.10: A) Alineamiento de secuencias entre RfaH y NusG. B) Cambios en la frustración de contactos para el índice *mutational* entre los aminoácidos nativos de la RfaH y las posiciones equivalentes de la NusG. Eje X: se muestran todos los contactos posibles que forman la proteína nativa y las mutantes. Eje Y: valores de frustración, los aminoácidos alternativos se representan en letras y se colorean en función de su valor de frustración. La variante nativa aparece en azul.

Bibliografía

Adams, M. D., Rudner, D. Z., y Rio, D. C. (1996). *Biochemistry and regulation of pre-mRNA splicing*. Current opinion in cell biology, 8(3):331–339.

Adamski, W., Salvi, N., Maurin, D., Magnat, J., Milles, S., Jensen, M. R., Abyzov, A., Moreau, C. J., y Blackledge, M. (2019). *A unified description of intrinsically disordered protein dynamics under physiological conditions using NMR spectroscopy*. Journal of the American Chemical Society, 141(44):17817–17829.

Al-Khodor, S., Price, C. T., Kalia, A., y Kwaik, Y. A. (2010). *Functional diversity of ankyrin repeats in microbial proteins*. Trends in microbiology, 18(3):132–139.

Andrade, M. A., Perez-Iratxeta, C., y Ponting, C. P. (2001). *Protein repeats: structures, functions, and evolution*. Journal of structural biology, 134(2-3):117–131.

Anfinsen, C. B. (1972). *The formation and stabilization of protein structure*. Biochemical Journal, 128(4):737.

Artsimovitch, I. y Landick, R. (2002). *The transcriptional regulator RfaH stimulates RNA chain synthesis after recruitment to elongation complexes by the exposed nontemplate DNA strand*. Cell, 109(2):193–203.

Barker, W. C., Ketcham, L. K., y Dayhoff, M. O. (1978). *A comprehensive examination of protein sequences for evidence of internal gene duplication*. Journal of Molecular Evolution, 10(4):265–281.

Betts, M. J., Guigó, R., Agarwal, P., y Russell, R. B. (2001). *Exon structure conservation*

despite low sequence similarity: a relic of dramatic events in evolution? The EMBO journal, 20(19):5354–5360.

Binz, H. K., Stumpp, M. T., Forrer, P., Amstutz, P., y Plückthun, A. (2003). *Designing repeat proteins: well-expressed, soluble and stable proteins from combinatorial libraries of consensus ankyrin repeat proteins.* Journal of molecular biology, 332(2):489–503.

Bisconte, M. G., Caldora, M., Musollino, G., Cardiero, G., Flagiello, A., La Porta, G., Lagona, L., Prezioso, R., Qualtieri, G., Gaudiano, C., et al. (2015). *α-Thalassemia associated with hb instability: a tale of two features. the case of Hb Rogliano or α1 Cod 108 (G15) Thr→ Asn and Hb Policoro or α2 Cod 124 (H7) Ser→ Pro.* Plos one, 10(3):e0115738.

Björklund, Å. K., Ekman, D., y Elofsson, A. (2006). *Expansion of protein domain repeats.* PLoS computational biology, 2(8):e114.

Blatch, G. L. y Lässle, M. (1999). *The tetratricopeptide repeat: a structural motif mediating protein-protein interactions.* Bioessays, 21(11):932–939.

Blum, M., Chang, H.-Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., Nuka, G., Paysan-Lafosse, T., Qureshi, M., Raj, S., et al. (2021). *The InterPro protein families and domains database: 20 years on.* Nucleic acids research, 49(D1):D344–D354.

Boice, J. A. y Fairman, R. (1996). *Structural characterization of the tumor suppressor p16, an ankyrin-like repeat protein.* Protein science, 5(9):1776–1784.

Borgia, A., Borgia, M. B., Bugge, K., Kissling, V. M., Heidarsson, P. O., Fernandes, C. B., Sottini, A., Soranno, A., Buholzer, K. J., Nettels, D., et al. (2018). *Extreme disorder in an ultrahigh-affinity protein complex.* Nature, 555(7694):61–66.

Bork, P. (1993). *Hundreds of ankyrin-like repeats in functionally diverse proteins: mobile modules that cross phyla horizontally?* Proteins: Structure, Function, and Bioinformatics, 17(4):363–374.

Bradley, C. M. y Barrick, D. (2002). *Limits of cooperativity in a structurally modular protein: response of the Notch ankyrin domain to analogous alanine substitutions in each repeat.* Journal of molecular biology, 324(2):373–386.

Braunitzer, G., Gehring-Müller, R., Hilschmann, N., Hilse, K., Hobom, G., Rudloff, V., y Wittmann-Liebold, u. B. (1961). *Die Konstitution des normalen adulten Humanhämoglobins.*

Breeden, L. y Nasmyth, K. (1987). *Similarity between cell-cycle genes of budding yeast and fission yeast and the Notch gene of Drosophila.* Nature, 329(6140):651–654.

Bridges, C. B. (1936). *The bar"gene.ª duplication.* Science, 83(2148):210–211.

Bryngelson, J. D. y Wolynes, P. G. (1987). *Spin glasses and the statistical mechanics of protein folding.* Proceedings of the National Academy of sciences, 84(21):7524–7528.

Burmann, B. M., Knauer, S. H., Sevostyanova, A., Schweimer, K., Mooney, R. A., Landick, R., Artsimovitch, I., y Rösch, P. (2012). *An α helix to β barrel domain switch transforms the transcription factor RfaH into a translation factor.* Cell, 150(2):291–303.

Carter, H. D., Svetlov, V., y Artsimovitch, I. (2004). *Highly divergent RfaH orthologs from pathogenic proteobacteria can substitute for Escherichia coli RfaH both in vivo and in vitro.* Journal of bacteriology, 186(9):2829–2840.

Chagula, D. B., Rechciński, T., Rudnicka, K., y Chmiela, M. (2020). *Ankyrins in human health and disease–an update of recent experimental findings.* Archives of Medical Science: AMS, 16(4):715.

Chen, C., Zhang, Q., Yu, B., Yu, Z., Lawrence, P. J., Ma, Q., y Zhang, Y. (2020). *Improving protein-protein interactions prediction accuracy using XGBoost feature selection and stacked ensemble classifier.* Computers in Biology and Medicine, 123:103899.

Conant, G. C. y Wolfe, K. H. (2008). *Turning a hobby into a job: how duplicated genes find new functions.* Nature Reviews Genetics, 9(12):938–950.

Consortium, U. (2019). *UniProt: a worldwide hub of protein knowledge.* Nucleic acids research, 47(D1):D506–D515.

Das, A. y Plotkin, S. S. (2013). *SOD1 exhibits allosteric frustration to facilitate metal binding affinity.* Proceedings of the National Academy of Sciences, 110(10):3871–3876.

Davtyan, A., Schafer, N. P., Zheng, W., Clementi, C., Wolynes, P. G., y Papoian, G. A. (2012). *AWSEM-MD: protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing.* The Journal of Physical Chemistry B, 116(29):8494–8503.

De la Llosa, P., Chene, N., y Martal, J. (1985). *Involvement of lysine residues in the binding of ovine prolactin and human growth hormone to lactogenic receptors.* FEBS letters, 191(2):211–215.

Delucchi, M., Schaper, E., Sachenkova, O., Elofsson, A., y Anisimova, M. (2020). *A new census of protein tandem repeats and their relationship with intrinsic disorder.* Genes, 11(4):407.

Eastman, P., Swails, J., Chodera, J. D., McGibbon, R. T., Zhao, Y., Beauchamp, K. A., Wang, L.-P., Simmonett, A. C., Harrigan, M. P., Stern, C. D., et al. (2017). *OpenMM 7: Rapid development of high performance algorithms for molecular dynamics.* PLoS computational biology, 13(7):e1005659.

Ehebauer, M. T., Chirgadze, D. Y., Hayward, P., Martinez Arias, A., y Blundell, T. L. (2005). *High-resolution crystal structure of the human Notch 1 ankyrin domain.* Biochemical Journal, 392(1):13–20.

Espada, R., Parra, R. G., Sippl, M. J., Mora, T., Walczak, A. M., y Ferreiro, D. U. (2015). *Repeat proteins challenge the concept of structural domains.* Biochemical Society Transactions, 43(5):844–849.

Fedorov, A., Fedorova, L., Starshenko, V., Filatov, V., y Grigor'ev, E. (1998). *Influence of exon duplication on intron and exon phase distribution.* Journal of molecular evolution, 46:263–271.

Fermi, G. (1975). *Three-dimensional Fourier synthesis of human deoxyhaemoglobin at 2· 5 Å resolution: refinement of the atomic model.* Journal of Molecular Biology, 97(2):237–256.

Ferreiro, D. U., Cho, S. S., Komives, E. A., y Wolynes, P. G. (2005). *The energy landscape*

of modular repeat proteins: topology determines folding mechanism in the ankyrin family. Journal of molecular biology, 354(3):679–692.

Ferreiro, D. U., Hegler, J. A., Komives, E. A., y Wolynes, P. G. (2007). *Localizing frustration in native proteins and protein assemblies.* Proceedings of the National Academy of Sciences, 104(50):19819–19824.

Ferreiro, D. U., Hegler, J. A., Komives, E. A., y Wolynes, P. G. (2011). *On the role of frustration in the energy landscapes of allosteric proteins.* Proceedings of the National Academy of Sciences, 108(9):3499–3503.

Ferreiro, D. U. y Komives, E. A. (2010). *Molecular mechanisms of system control of NF-κB signaling by IκBα.* Biochemistry, 49(8):1560–1567.

Ferreiro, D. U., Komives, E. A., y Wolynes, P. G. (2014). *Frustration in biomolecules.* Quarterly reviews of biophysics, 47(4):285–363.

Ferreiro, D. U., Komives, E. A., y Wolynes, P. G. (2018). *Frustration, function and folding.* Current opinion in structural biology, 48:68–73.

Ferreiro, D. U., Walczak, A. M., Komives, E. A., y Wolynes, P. G. (2008). *The energy landscapes of repeat-containing proteins: topology, cooperativity, and the folding funnels of one-dimensional architectures.* PLoS computational biology, 4(5):e1000070.

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y.-l., y Postlethwait, J. (1999). *Preservation of duplicate genes by complementary, degenerative mutations.* Genetics, 151(4):1531–1545.

Forrer, P., Stumpp, M. T., Binz, H. K., y Plückthun, A. (2003). *A novel strategy to design binding molecules harnessing the modular nature of repeat proteins.* FEBS letters, 539(1-3):2–6.

Freiberger, M. I., Guzovsky, A. B., Wolynes, P. G., Parra, R. G., y Ferreiro, D. U. (2019). *Local frustration around enzyme active sites.* Proceedings of the National Academy of Sciences, 116(10):4037–4043.

Fuxreiter, M. (2018). *Fuzziness in protein interactions—A historical perspective.* Journal of molecular biology, 430(16):2278–2287.

Galaz-Davison, P., Román, E. A., y Ramírez-Sarmiento, C. A. (2021). *The N-terminal domain of RfaH plays an active role in protein fold-switching.* PLoS Computational Biology, 17(9):e1008882.

Galea, C. A., Nourse, A., Wang, Y., Sivakolundu, S. G., Heller, W. T., y Kriwacki, R. W. (2008). *Role of intrinsic flexibility in signal transduction mediated by the cell cycle regulator, p27Kip1.* Journal of molecular biology, 376(3):827–838.

Galpern, E. A., Freiberger, M. I., y Ferreiro, D. U. (2020). *Large Ankyrin repeat proteins are formed with similar and energetically favorable units.* PLoS One, 15(6):e0233865.

Galpern, E. A., Marchi, J., Mora, T., Walczak, A. M., y Ferreiro, D. U. (2022). *Evolution and folding of repeat proteins.* Proceedings of the National Academy of Sciences, 119(31):e2204131119.

Gianni, S., Camilloni, C., Giri, R., Toto, A., Bonetti, D., Morrone, A., Sormanni, P., Brunori, M., y Vendruscolo, M. (2014). *Understanding the frustration arising from the competition between function, misfolding, and aggregation in a globular protein.* Proceedings of the National Academy of Sciences, 111(39):14141–14146.

Giardina, B., Brunori, M., Antonini, E., y Tentori, L. (1978). *Properties of hemoglobin G. Ferrara (β57 (E1) Asn$\rightarrow$ Lys).* Biochimica et Biophysica Acta (BBA)-Protein Structure, 534(1):1–6.

Gilbert, W. (1978). *Why genes in pieces?* Nature, 271(5645):501–501.

Gógl, G., Alexa, A., Kiss, B., Katona, G., Kovács, M., Bodor, A., Reményi, A., y Nyitray, L. (2016). *Structural Basis of Ribosomal S6 Kinase 1 (RSK1) Inhibition by S100B Protein: MODULATION OF THE EXTRACELLULAR SIGNAL-REGULATED KINASE (ERK) SIGNALING CASCADE IN A CALCIUM-DEPENDENT WAY*–.* Journal of Biological Chemistry, 291(1):11–27.

Hadži, S., Mernik, A., Podlipnik, Č., Loris, R., y Lah, J. (2017). *The thermodynamic basis of the fuzzy interaction of an intrinsically disordered protein.* Angewandte Chemie International Edition, 56(46):14494–14497.

Haldane, J. B. (1932). *The causes of evolution*, volume 5. Princeton University Press.

Hardison, R. C. (2012). *Evolution of hemoglobin and its genes.* Cold Spring Harbor perspectives in medicine, 2(12):a011627.

Harren, T., Matter, H., Hessler, G., Rarey, M., y Grebner, C. (2022). *Interpretation of structure–activity relationships in real-world drug design data sets using explainable artificial intelligence.* Journal of Chemical Information and Modeling, 62(3):447–462.

Horvath, A., Miskei, M., Ambrus, V., Vendruscolo, M., y Fuxreiter, M. (2020). *Sequence-based prediction of protein binding mode landscapes.* PLoS computational biology, 16(5):e1007864.

Huang, J., Zhao, X., Yu, H., Ouyang, Y., Wang, L., y Zhang, Q. (2009). *The ankyrin repeat gene family in rice: genome-wide identification, classification and expression profiling.* Plant molecular biology, 71(3):207–226.

Huang, Y., Niu, B., Gao, Y., Fu, L., y Li, W. (2010). *CD-HIT Suite: a web server for clustering and comparing biological sequences.* Bioinformatics, 26(5):680–682.

Hughes, A. L. (1994). *The evolution of functionally novel proteins after gene duplication.* Proceedings of the Royal Society of London. Series B: Biological Sciences, 256(1346):119–124.

Innan, H. y Kondrashov, F. (2010). *The evolution of gene duplications: classifying and distinguishing between models.* Nature Reviews Genetics, 11(2):97–108.

Islam, Z., Nagampalli, R. S. K., Fatima, M. T., y Ashraf, G. M. (2018). *New paradigm in ankyrin repeats: Beyond protein-protein interaction module.* International journal of biological macromolecules, 109:1164–1173.

Itano, H. A. (1957). *The human hemoglobins: their properties and genetic control.* En *Advances in Protein Chemistry*, volume 12, pages 215–268. Elsevier.

Ivankov, D. N., Garbuzynskiy, S. O., Alm, E., Plaxco, K. W., Baker, D., y Finkelstein, A. V. (2003). *Contact order revisited: influence of protein size on the folding rate.* Protein science, 12(9):2057–2062.

Jacob, M. y Harrison, S. (1998). *Structure of an IκBα/NF-κB complex.* Cell, 95(6):749–758.

Javadi, Y. y Itzhaki, L. S. (2013). *Tandem-repeat proteins: regularity plus modularity equals design-ability.* Current opinion in structural biology, 23(4):622–631.

Jones, W. K., Yu-Lee, L., Clift, S. M., Brown, T. L., y Rosen, J. (1985). *The rat casein multigene family. Fine structure and evolution of the beta-casein gene.* Journal of Biological Chemistry, 260(11):7042–7050.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Tunyasuvunakool, K., Ronneberger, O., Bates, R., Žídek, A., Bridgland, A., et al. (2020). *AlphaFold 2.* In Fourteenth Critical Assessment of Techniques for Protein Structure Prediction (Abstract Book.

Kajava, A. V. (2012). *Tandem repeats in proteins: from sequence to structure.* Journal of structural biology, 179(3):279–288.

Kanz, C., Aldebert, P., Althorpe, N., Baker, W., Baldwin, A., Bates, K., Browne, P., van den Broek, A., Castro, M., Cochrane, G., et al. (2005). *The EMBL nucleotide sequence database.* Nucleic acids research, 33(suppl_1):D29–D33.

Kavanaugh, J. S., Rogers, P. H., Arnone, A., Hui, H. L., Wierzba, A., DeYoung, A., Kwiatkowski, L. D., Noble, R. W., Juszczak, L. J., Peterson, E. S., et al. (2005). *Intersubunit interactions associated with Tyr42α stabilize the quaternary-T tetramer but are not major quaternary constraints in deoxyhemoglobin.* Biochemistry, 44(10):3806–3820.

Keren, H., Lev-Maor, G., y Ast, G. (2010). *Alternative splicing and evolution: diversification, exon definition and function.* Nature Reviews Genetics, 11(5):345–355.

Kobe, B. y Kajava, A. V. (2001). *The leucine-rich repeat as a protein recognition motif.* Current opinion in structural biology, 11(6):725–732.

Kolkman, J. A. y Stemmer, W. P. (2001). *Directed evolution of proteins by exon shuffling.* Nature biotechnology, 19(5):423–428.

Korennykh, A. V., Egea, P. F., Korostelev, A. A., Finer-Moore, J., Zhang, C., Shokat, K. M., Stroud, R. M., y Walter, P. (2009). *The unfolded protein response signals through high-order assembly of Ire1.* Nature, 457(7230):687–693.

Kriventseva, E. V., Koch, I., Apweiler, R., Vingron, M., Bork, P., Gelfand, M. S., y Sunyaev, S. (2003). *Increase of functional diversity by alternative splicing.* Trends in Genetics, 19(3):124–128.

Kumar, S., Clarke, D., y Gerstein, M. (2013). *Localized structural frustration for evaluating the impact of sequence variants.* Nucleic acids research, 44(21):gkw927.

Kumar, S., Rosenberg, J. M., Bouzida, D., Swendsen, R. H., y Kollman, P. A. (1992). *The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method.* Journal of computational chemistry, 13(8):1011–1021.

Leff, S. E., Rosenfeld, M. G., y Evans, R. M. (1986). *Complex transcriptional units: diversity in gene expression by alternative RNA processing.* Annual review of biochemistry, 55(1):1091–1117.

Levinthal, C. (1968). *Are there pathways for protein folding?* Journal de chimie physique, 65:44–45.

Light, S. y Elofsson, A. (2013). *The impact of splicing on protein domain architecture.* Current opinion in structural biology, 23(3):451–458.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. (2023). *Evolutionary-scale prediction of atomic-level protein structure with a language model.* Science, 379(6637):1123–1130.

Lindstrom, I. y Dogan, J. (2018). *Dynamics, Conformational Entropy, and Frustration in Protein–Protein Interactions Involving an Intrinsically Disordered Protein Domain.* ACS chemical biology, 13(5):1218–1227.

Lowe, A. R. (2007). *Rational redesign of the folding pathway of a modular protein.* Proceedings of the National Academy of Sciences, 104(8):2679–2684.

Lu, W., Bueno, C., Schafer, N. P., Moller, J., Jin, S., Chen, X., Chen, M., Gu, X., Davtyan, A., de Pablo, J. J., et al. (2021). *OpenAWSEM with Open3SPN2: A fast, flexible, and accessible framework for large-scale coarse-grained biomolecular simulations.* PLoS computational biology, 17(2):e1008308.

Lukhele, S., Bah, A., Lin, H., Sonenberg, N., y Forman-Kay, J. D. (2013). *Interaction of the eukaryotic initiation factor 4E with 4E-BP2 at a dynamic bipartite interface.* Structure, 21(12):2186–2196.

Lynch, M. y Conery, J. S. (2000). *The evolutionary fate and consequences of duplicate genes.* science, 290(5494):1151–1155.

Main, E. R., Jackson, S. E., y Regan, L. (2003a). *The folding and design of repeat proteins: reaching a consensus.* Current opinion in structural biology, 13(4):482–489.

Main, E. R., Lowe, A. R., Mochrie, S. G., Jackson, S. E., y Regan, L. (2005). *A recurring theme in protein engineering: the design, stability and folding of repeat proteins.* Current opinion in structural biology, 15(4):464–471.

Main, E. R., Xiong, Y., Cocco, M. J., D'Andrea, L., y Regan, L. (2003b). *Design of stable α-helical arrays from an idealized TPR motif.* Structure, 11(5):497–508.

Marcotte, E. M., Pellegrini, M., Yeates, T. O., y Eisenberg, D. (1999). *A census of protein repeats.* Journal of molecular biology, 293(1):151–160.

Mello, C. C. y Barrick, D. (2004). *An experimentally determined protein folding energy landscape.* Proceedings of the National Academy of Sciences, 101(39):14102–14107.

Mello, C. C., Bradley, C. M., Tripp, K. W., y Barrick, D. (2005). *Experimental characterization of the folding kinetics of the notch ankyrin domain.* Journal of molecular biology, 352(2):266–281.

Miskei, M., Horvath, A., Vendruscolo, M., y Fuxreiter, M. (2020). *Sequence-based prediction of fuzzy protein interactions.* Journal of molecular biology, 432(7):2289–2303.

Mollan, T. L., Yu, X., Weiss, M. J., y Olson, J. S. (2010). *The role of alpha-hemoglobin stabilizing protein in redox chemistry, denaturation, and hemoglobin assembly.* Antioxidants & redox signaling, 12(2):219–231.

Moore, J. A., Pullon, B. M., Wang, D., y Brennan, S. O. (2021). *Hb Waikato [α127 (H10) Lys→ Gln; HBA1: c.382A > C]: A Novel High Oxygen Affinity Variant.* Hemoglobin, 45(1):41–45.

Moore, M. J., Query, C. C., y Sharp, P. A. (1993). *Splicing of precursors to mRNAs by the spliceosome.* Cold Spring Harbor Monograph Series, 24:303–303.

Morcos, F., Schafer, N. P., Cheng, R. R., Onuchic, J. N., y Wolynes, P. G. (2014). *Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection.* Proceedings of the National Academy of Sciences, 111(34):12408–12413.

Mosavi, L. K., Cammett, T. J., Desrosiers, D. C., y Peng, Z.-y. (2004). *The ankyrin repeat as molecular architecture for protein recognition.* Protein science, 13(6):1435–1448.

Mosavi, L. K., Minor, D. L., y Peng, Z.-y. (2002a). *Consensus-derived structural determinants of the ankyrin repeat motif.* Proceedings of the National Academy of Sciences, 99(25):16029–16034.

Mosavi, L. K., Williams, S., y Peng, Z.-y. (2002b). *Equilibrium folding and stability of myotrophin: a model ankyrin repeat protein.* Journal of molecular biology, 320(2):165–170.

Muller, H. (1935). *The origination of chromatin deficiencies as minute deletions subject to insertion elsewhere.* Genetica, 17(3):237–252.

Ohno, S. (2013). *Evolution by gene duplication.* Springer Science & Business Media.

Onuchic, J. N. y Wolynes, P. G. (2004). *Theory of protein folding.* Current opinion in structural biology, 14(1):70–75.

Papoian, G. A., Ulander, J., Eastwood, M. P., Luthey-Schulten, Z., y Wolynes, P. G. (2004). *Water in protein structure prediction.* Proceedings of the National Academy of Sciences, 101(10):3352–3357.

Pâques, F., Leung, W.-Y., y Haber, J. E. (1998). *Expansions and contractions in a tandem repeat induced by double-strand break repair.* Molecular and cellular biology, 18(4):2045–2054.

Parra, R. G., Espada, R., Verstraete, N., y Ferreiro, D. U. (2015). *Structural and energetic characterization of the ankyrin repeat protein family.* PLoS computational biology, 11(12):e1004659.

Parra, R. G., Schafer, N. P., Radusky, L. G., Tsai, M.-Y., Guzovsky, A. B., Wolynes, P. G., y Ferreiro, D. U. (2016). *Protein Frustratometer 2: a tool to localize energetic frustration in protein molecules, now with electrostatics.* Nucleic acids research, 44(W1):W356–W360.

Patthy, L. (1987). *Intron-dependent evolution: preferred types of exons and introns.* FEBS letters, 214(1):1–7.

Patthy, L. (1996). *Exon shuffling and other ways of module exchange.* Matrix biology, 15(5):301–310.

Patthy, L. (2009). *Protein evolution.* John Wiley & Sons.

Perutz, M. F. (1970). *Stereochemistry of cooperative effects in haemoglobin: haem–haem interaction and the problem of allostery.* Nature, 228(5273):726–734.

Plückthun, A. (2015). *Designed ankyrin repeat proteins (DARPins): binding proteins for research, diagnostics, and therapy.* Annual review of pharmacology and toxicology, 55:489–511.

Potoyan, D. A., Zheng, W., Komives, E. A., y Wolynes, P. G. (2016). *Molecular stripping in the NF-κB/IκB/DNA genetic regulatory network.* Proceedings of the National Academy of Sciences, 113(1):110–115.

Ramírez-Sarmiento, C. A., Noel, J. K., Valenzuela, S. L., y Artsimovitch, I. (2015). *Interdomain contacts control native state switching of RfaH on a dual-funneled landscape.* PLoS computational biology, 11(7):e1004379.

Rauer, C., Sen, N., Waman, V. P., Abbasian, M., y Orengo, C. A. (2021). *Computational approaches to predict protein functional families and functional sites.* Current Opinion in Structural Biology, 70:108–122.

Rausch, A. O., Freiberger, M. I., Leonetti, C. O., Luna, D. M., Radusky, L. G., Wolynes, P. G., Ferreiro, D. U., y Parra, R. G. (2021). *FrustratometeR: an R-package to compute local frustration in protein structures, point mutants and MD simulations.* Bioinformatics, 37(18):3038–3040.

Rausell, A., Juan, D., Pazos, F., y Valencia, A. (2010). *Protein interactions and ligand binding: from protein subfamilies to functional specificity.* Proceedings of the National Academy of Sciences, 107(5):1995–2000.

Reeves, P. (1993). *Evolution of Salmonella O antigen variation by interspecific gene transfer on a large scale.* Trends in Genetics, 9(1):17–22.

Rosenfeld, M. G., Lin, C. R., Amara, S. G., Stolarsky, L., Roos, B. A., Ong, E. S., y Evans, R. M. (1982). *Calcitonin mRNA polymorphism: peptide switching associated with alternative RNA splicing events.* Proceedings of the National Academy of Sciences, 79(6):1717–1721.

Saks, M. E., Sampson, J. R., y Abelson, J. (1998). *Evolution of a transfer RNA gene through a point mutation in the anticodon.* Science, 279(5357):1665–1670.

Sánchez, I. E., Galpern, E. A., Garibaldi, M. M., y Ferreiro, D. U. (2022). *Molecular information theory meets protein folding.* The Journal of Physical Chemistry B, 126(43):8655–8668.

Sanna, M. T., Giardina, B., Scatena, R., Pellegrini, M., Olianas, A., Manca, L., Masala, B., Castagnola, M., y Corda, M. (1994). *Functional alterations in adult and fetal hemoglobin Sassari Asp-alpha 126 (H9)–¿ His. The role of alpha 1 alpha 2 contact*. Journal of Biological Chemistry, 269(28):18338–18342.

Sayers, E. W., Cavanaugh, M., Clark, K., Ostell, J., Pruitt, K. D., y Karsch-Mizrachi, I. (2019). *GenBank*. Nucleic acids research, 47(D1):D94–D99.

Schneider, G. y Clark, D. E. (2019). *Automated de novo drug design: are we nearly there yet?* Angewandte Chemie International Edition, 58(32):10792–10803.

Schneider, P., Walters, W. P., Plowright, A. T., Sieroka, N., Listgarten, J., Goodnow Jr, R. A., Fisher, J., Jansen, J. M., Duca, J. S., Rush, T. S., et al. (2020). *Rethinking drug design in the artificial intelligence era*. Nature Reviews Drug Discovery, 19(5):353–364.

Schroeder, W. A., Shelton, J. R., Shelton, J. B., Cormick, J., y Jones, R. T. (1963). *The amino acid sequence of the γ chain of human fetal hemoglobin*. Biochemistry, 2(5):992–1008.

Schüler, A. y Bornberg-Bauer, E. (2016). *Evolution of protein domain repeats in Metazoa*. Molecular biology and evolution, 33(12):3170–3182.

Sedgwick, S. G. y Smerdon, S. J. (1999). *The ankyrin repeat: a diversity of interactions on a common structural framework*. Trends in biochemical sciences, 24(8):311–316.

Shaanan, B. (1983). *Structure of human oxyhaemoglobin at 2· 1resolution*. Journal of molecular biology, 171(1):31–59.

Sharma, R., Raduly, Z., Miskei, M., y Fuxreiter, M. (2015). *Fuzzy complexes: Specific binding without complete folding*. FEBS letters, 589(19):2533–2542.

Sharp, P. A. (2005). *The discovery of split genes and RNA splicing*. Trends in biochemical sciences, 30(6):279–281.

Shi, D., Svetlov, D., Abagyan, R., y Artsimovitch, I. (2017). *Flipping states: a few key residues decide the winning conformation of the only universally conserved transcription factor*. Nucleic acids research, 45(15):8835–8843.

Sievers, F. y Higgins, D. G. (2014). *Clustal Omega, accurate alignment of very large numbers of sequences*. Multiple sequence alignment methods, pages 105–116.

Stamos, J. L., Chu, M. L.-H., Enos, M. D., Shah, N., y Weis, W. I. (2014). *Structural basis of GSK-3 inhibition by N-terminal phosphorylation and by the Wnt receptor LRP6*. Elife, 3:e01998.

Stelzl, L. S., Mavridou, D. A., Saridakis, E., Gonzalez, D., Baldwin, A. J., Ferguson, S. J., Sansom, M. S., y Redfield, C. (2020). *Local frustration determines loop opening during the catalytic cycle of an oxidoreductase*. Elife, 9:e54661.

Street, T. O., Rose, G. D., y Barrick, D. (2006). *The role of introns in repeat protein gene formation*. Journal of molecular biology, 360(2):258–266.

Stumpp, M. T., Forrer, P., Binz, H. K., y Plückthun, A. (2003). *Designing repeat proteins: modular leucine-rich repeat protein libraries based on the mammalian ribonuclease inhibitor family*. Journal of molecular biology, 332(2):471–487.

Tang, K. S., Fersht, A. R., y Itzhaki, L. S. (2003). *Sequential unfolding of ankyrin repeats in tumor suppressor p16*. Structure, 11(1):67–73.

Tang, K. S., Guralnick, B. J., Wang, W. K., Fersht, A. R., y Itzhaki, L. S. (1999). *Stability and folding of the tumour suppressor protein p16*. Journal of molecular biology, 285(4):1869–1886.

Tevelev, A., Byeon, I.-J. L., Selby, T., Ericson, K., Kim, H.-J., Kraynov, V., y Tsai, M.-D. (1996). *Tumor suppressor p16INK4A: structural characterization of wild-type and mutant proteins by NMR and circular dichroism*. Biochemistry, 35(29):9475–9487.

Tomar, S. K., Knauer, S. H., NandyMazumdar, M., Rösch, P., y Artsimovitch, I. (2013). *Interdomain contacts control folding of transcription factor RfaH*. Nucleic acids research, 41(22):10077–10085.

Tompa, P. (2002). *Intrinsically unstructured proteins*. Trends in biochemical sciences, 27(10):527–533.

Tompa, P. (2005). *The interplay between structure and function in intrinsically unstructured proteins*. FEBS letters, 579(15):3346–3354.

Tompa, P. y Fuxreiter, M. (2008). *Fuzzy complexes: polymorphism and structural disorder in protein–protein interactions*. Trends in biochemical sciences, 33(1):2–8.

Toto, A., Camilloni, C., Giri, R., Brunori, M., Vendruscolo, M., y Gianni, S. (2016). *Molecular recognition by templated folding of an intrinsically disordered protein*. Scientific reports, 6(1):1–9.

Traag, V. A., Waltman, L., y Van Eck, N. J. (2019). *From Louvain to Leiden: guaranteeing well-connected communities*. Scientific reports, 9(1):5233.

Truhlar, S. M., Mathes, E., Cervantes, C. F., Ghosh, G., y Komives, E. A. (2008). *Pre-folding IκBα alters control of NF-κB signaling*. Journal of molecular biology, 380(1):67–82.

Urvoas, A., Guellouz, A., Valerio-Lepiniec, M., Graille, M., Durand, D., Desravines, D. C., van Tilbeurgh, H., Desmadril, M., y Minard, P. (2010). *Design, production and molecular structure of a new family of artificial alpha-helicoidal repeat proteins (αRep) based on thermostable HEAT-like repeats*. Journal of molecular biology, 404(2):307–327.

Varadamsetty, G., Tremmel, D., Hansen, S., Parmeggiani, F., y Plückthun, A. (2012). *Designed Armadillo repeat proteins: library generation, characterization and selection of peptide binders with high specificity*. Journal of molecular biology, 424(1-2):68–87.

Wagih, O. (2017). *ggseqlogo: a versatile R package for drawing sequence logos*. Bioinformatics, 33(22):3645–3647.

Webb, B. y Sali, A. (2017). *Protein structure modeling with MODELLER*. En *Functional genomics*, pages 39–54. Springer.

Wolynes, P. G. (1996). *Symmetry and the energy landscapes of biomolecules*. Proceedings of the National Academy of Sciences, 93(25):14249–14255.

Wright, P. E. y Dyson, H. J. (2009). *Linking folding and binding*. Current opinion in structural biology, 19(1):31–38.

Xu, G., Guo, C., Shan, H., y Kong, H. (2012). *Divergence of duplicate genes in exon–intron structure*. Proceedings of the National Academy of Sciences, 109(4):1187–1192.

XU, Z., YANG, J., LIU, H., y HUANG, W. (2020). *Protein complex identification algorithm based on XGboost and topological structural information*. Journal of Computer Applications, 40(5):1510.

Yamada, Y., Liau, G., Mudryj, M., Obici, S., y de Crombrugghe, B. (1984). *Conservation of the sizes for one but not another class of exons in two chick collagen genes*. Nature, 310(5975):333–337.

Yang, J., Roy, A., y Zhang, Y. (2012). *BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions*. Nucleic acids research, 41(D1):D1096–D1103.

Zhavoronkov, A., Ivanenkov, Y. A., Aliper, A., Veselov, M. S., Aladinskiy, V. A., Aladinskaya, A. V., Terentiev, V. A., Polykovskiy, D. A., Kuznetsov, M. D., Asadulaev, A., et al. (2019). *Deep learning enables rapid identification of potent DDR1 kinase inhibitors*. Nature biotechnology, 37(9):1038–1040.

Zhaxybayeva, O. y Gogarten, J. P. (2004). *Cladogenesis, coalescence and the evolution of the three domains of life*. TRENDS in Genetics, 20(4):182–187.

Zuber, P. K., Artsimovitch, I., NandyMazumdar, M., Liu, Z., Nedialkov, Y., Schweimer, K., Rösch, P., y Knauer, S. H. (2018). *The universally-conserved transcription factor RfaH is recruited to a hairpin structure of the non-template DNA strand*. Elife, 7:e36349.

Zuber, P. K., Schweimer, K., Rösch, P., Artsimovitch, I., y Knauer, S. H. (2019). *Reversible fold-switching controls the functional cycle of the antitermination factor RfaH*. Nature communications, 10(1):1–13.

Zweifel, M. E. y Barrick, D. (2001). *Studies of the ankyrin repeats of the Drosophila melanogaster Notch receptor. 2. Solution stability and cooperativity of unfolding*. Biochemistry, 40(48):14357–14367.