Tesis Doctoral

# Estructuras, simetrías y paisajes energéticos en la familia de proteínas con repeticiones de Ankirina

Parra, Rodrigo Gonzalo

2016-03-23


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

# Estructuras, simetrías y paisajes energéticos en la familia de proteínas con repeticiones de Ankirina

**Lic. Rodrigo Gonzalo Parra**

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

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, Marzo de 2016

Fecha de defensa: 23/03/2016

# Índice general

Resumen . . . . . 3

Abstract . . . . . 5

Agradecimientos . . . . . 7

Publicaciones . . . . . 9

**1. Introducción** **11**

1.1. Plegado de Proteínas y Paisajes Energéticos . . . . . 11

1.2. Proteínas Repetitivas . . . . . 16

1.3. La familia de Proteínas con Repeticiones de Ankirina . . . . . 19

1.3.1. Diseño de ANKs por consenso . . . . . 22

1.4. Objetivos . . . . . 26

**2. Teselado Proteico: Simetrías y Periodicidades en Estructuras de Proteínas** **27**

2.1. Alineamientos estructurales y teselas . . . . . 30

2.2. Modelo Homogéneo . . . . . 32

2.3. Teselado de Proteínas del Mundo Real . . . . . 36

2.4. Teselado de Proteínas Repetitivas del Tipo Solenoide . . . . . 40

2.5. Teselado de Proteínas Repetitivas con Topología Cerrada y del tipo Cuentas de Perla . . . . . 44

2.6. Teselado de Proteínas Globulares . . . . . 48

2.7. Teselado de Oligómeros . . . . . 50

2.8. Aportes del método de teselado para mejorar la detección de unidades repetitivas en proteínas . . . . . 53

2.9. Conclusiones del capítulo . . . . . 58

**3. Análisis Estructural y Energético de la Familia ANK** **67**

3.1. Análisis generales de las estructuras ANK . . . . . 68

3.2. Organización de los Arreglos Repetitivos en ANKs . . . . . 72

3.3. Definición Consistente de las Repeticiones de Ankirina . . . . . 78

3.4. Anotación y caracterización de las repeticiones . . . . . 82

3.5. Descripción Energética de las ANKs . . . . . 86

3.6. Interacciones Proteína-Proteína en los arreglos repetitivos de ANKs . . . . . 96

3.7. Mejores modelos de secuencia . . . . . 100

3.8. Conclusiones del capítulo . . . . . 103

**4. Explorando el paisaje energético de las proteínas ANK** **115**

4.1. Introducción . . . . . 115

4.2. Topología proteica y plegado . . . . . 116

4.3. 3 repeticiones . . . . . 118

4.4. 4 repeticiones . . . . . 124

4.5. 5 repeticiones . . . . . 131

4.6. 6 repeticiones . . . . . 137

4.7. Plegabilidad Jerárquica, ¿Un proxy hacia la dinámica del plegado proteico? . . 140

4.8. Conclusiones del capítulo . . . . . 142

**5. Métodos** **149**

5.1. Construcción de una base de datos para ANKs: . . . . . 149

5.2. Frustración Energética Local . . . . . 154

5.3. Contenido de Información . . . . . 156

5.4. Coordenada Qw . . . . . 157

5.5. Simulaciones de Dinámica Molecular del tipo grano grueso . . . . . 157

5.6. Modelo AMH-G$\bar{o}$ No aditivo . . . . . 158

5.7. Modelo Cinético Discreto . . . . . 160

**6. Conclusiones Generales** **163**

# Estructuras, Simetrías y Paisajes Energéticos en la Familia de Proteínas con Repeticiones de Ankirina

## Resumen

Las proteínas repetitivas están compuestas por repeticiones en tándem de motivos estructurales. Estas unidades interaccionan entre sí de forma tal que coalescen en arquitecturas del tipo solenoidal o toroidal. Por su simplicidad topológica constituyen modelos útiles para el estudio del plegado de proteínas. Esta tesis se centra en el estudio de los miembros de la familia de proteínas con repeticiones de Ankirina (ANKs) que contienen entre 3 y 34 copias de un motivo de 33 residuos de largo. Siendo principalmente descriptas como mediadoras de interacciones proteína-proteína, las ANKs poseen una versatilidad para el reconocimiento de otras moléculas comparable a la de los anticuerpos. A pesar de su simplicidad estructural, las proteínas repetitivas presentan varios desafíos inherentes a su no globularidad. Las repeticiones pueden ser muy degeneradas en sus secuencias dificultando la detección y definición de límites entre ellas, siendo este un problema aún no resuelto en este campo de estudio. Dado que las estructuras están mucho más conservadas que las secuencias, hemos desarrollado un método, llamado "Teselado proteico" capaz de detectar repeticiones estructurales de forma objetiva, rigurosa y eficiente. Dicho algoritmo es independiente de la secuencia de las moléculas analizadas y es además generalizable a todos los tipos de proteínas repetitivas. Su aplicación a distintas clases de proteínas nos permitió caracterizar sus periodicidades y espectro de simetrías. Hemos aplicado nuestro algoritmo a todos los miembros de la familia ANK. Haciendo uso de nociones de la teoría de paisajes energéticos y con la hipótesis de que las repeticiones constituyen unidades de plegado, fuimos capaces de definir el largo y la fase de las repeticiones en todos los casos. Este estudio representa el primer caso reportado en que una estrategia de anotación de las unidades repetitivas consistente es aplicada a lo largo de toda una familia, lo cual permite realizar análisis comparativos al nivel de las repeticiones individuales. Los análisis subsiguientes muestran que las repeticiones ANKs pueden clasificarse en 3 tipos diferentes que corresponden a las repeticiones ubicadas en la región N-terminal, interna o C-terminal. Cada tipo de repetición muestra patrones de conservación en secuencia

y en su energía de plegado específicos de su tipo. Pudimos observar además que los niveles de conservación de la secuencia y de la energía de plegado se encuentran correlacionados de forma lineal y positiva, lo cual indica que aquellas secuencias parecidas al consenso que se obtiene a partir del alineamiento múltiple de secuencias son más plegables que aquellas que se alejan del mismo. Existe una red de interacciones conservadas y energéticamente favorables en todas las repeticiones internas de las ANKs que conectan aquellos residuos más conservados en secuencia y que por tanto constituyen un núcleo de estabilidad estructural de las mismas. Al analizar en cambio todos los elementos que no forman parte de la estructura canónica de las repeticiones (inserciones y deleciones), las regiones cercanas a las mismas, se encuentran enriquecidas en interacciones frustradas, es decir, son desfavorables para el plegado de las mismas. Lo anterior sugiere que la presencia de inserciones y deleciones indica adaptaciones funcionales de las proteínas en que se encuentran. Este enriquecimiento de interacciones frustradas se observa también en los sitios de interacción con otras proteínas. Finalmente hemos estudiado la dinámica de plegado de estas moléculas usando métodos basados en estructura. Hemos caracterizado los mecanismos de plegado de varios miembros de la familia ANK y relacionando los mismos con los patrones energéticos previamente descriptos a partir de los estados nativos. Nuestros resultados ofrecen nuevas perspectivas acerca del funcionamiento de las proteínas de la familia ANK y pueden ser de gran utilidad para aquellos interesados en el diseño de este tipo de moléculas para diferentes fines y el entendimiento biofísico de estas moléculas en forma general.

**Palabras clave:** proteínas con repeticiones de ankirina, simetrías, repeticiones, plegado proteico, frustración local, dinámicas moleculares.

# Structures, Symmetries and Energy Landscapes in the Ankyrin Repeat Protein Family


## Abstract

Repeat proteins are composed of specific structural motifs that are tandemly repeated. These units interact between each other leading to solenoidal or toroidal architectures. Given their topological simplicity these molecules constitute useful models to study the protein folding problem. This thesis is focused in studying the Ankyrin Repeat Protein Familiy (ANKs). Proteins from this family contain between 2 and 34 copies of a structural motif of 33 residues long. Being mainly described as protein-protein interactors, the ANKs have a high versatility to recognize other molecules, comparable to that of antibodies. Despite their structural simplicity, repeat proteins present several challenges that are inherent to their non-globularity. Repeats can be highly degenerated in their sequences difficulting their detection and the definition of limits in between them. Given that the structure is more conserved than the sequence, we have developed a method called "Protein tiling" that is able to detect structural repetitions in an objective, rigorous and efficient manner. This algorithm is independent from the sequences of the proteins being analyzed and also it is generalizable to other repeat protein families and types. Its application to different protein classes allowed us to characterize their periodicities and symmetry spectra. We have applied the Protein Tiling algorithm to all the members in the ANK family. Using notions from the energy landscapes theory and the hypothesis of repeats as protein folding units, we were able to define the length and phase for the repeating units in all the ANKs. This study represents the first reported case in which a consistent repeats annotation strategy is applied to a whole family which allows to perform comparative analysis of repeats as individual units. Subsequent analysis showed that ANK repeats can be classified into 3 different types that correspond to repetitions localized at the N-terminal, internal or C-terminal regions. Each repeat type shows specific sequence and energy conservation patterns. We observed that the conservation degree at the sequences and the energy patterns are linearly and positively correlated, indicating that those sequences that are similar to the consensus obtained from the multiple sequence alignment are more foldable than

those that have more differences with it. There exists a network of conserved interactions along all internal ANK repeats that connect those residues that are highly conserved in sequence and hence constitute a structural stability core for the overall structure. On the contrary, when analyzing those elements that do not belong to the canonical repeats structure (insertions and deletions), those regions that are close to them are enriched in frustrated interactions, i.e. where local folding is unfavorable. The latter suggests that the presence of insertions and deletions indicate functional adaptations of the proteins in which they are contained. This enrichment of frustrated interactions is also observed at the interaction sites of ANKs with other proteins. Finally we have studied the folding dynamics of several ANK members using state of the art computational methods. We have characterized the folding mechanisms of several members in the ANK family and related those with the energetic patterns previously described from the native states. Our results offer new insights into how ANK proteins function and can be of great value to those researchers that are interested in using these type of molecules for protein design with different aims and for the biophysical understanding of these molecules in general.

**Keywords:** ankyrin proteins, symmetry, repetitions, protein folding, local frustration, dynamic simulations.

# Agradecimientos

A mi director y maestro, el Dr. Diego Ferreiro, por haberme aceptado como su primer estudiante doctoral en su laboratorio. Gracias por enseñarme a hacer ciencia en el sentido más puro de la palabra. Siempre voy a recordar esa frase en mi primer año: "No te preocupes por los papers, si laburas bien y a conciencia, los papers vienen solos". Mi más grande admiración y gratitud no sólo por la formación académica que me ha transmitido, pero además por su calidad y gradeza humana.

Al Dr. Ignacio Sánchez, co-director del laboratorio por tantas enseñanzas, consejos, mails, ideas y su implacable búsqueda del orden para que vivamos lo más organizados y felices posible en el laboratorio.

A mis compañeros del laboratorio, pasados y presentes: Ezequiel Bos, Nico Palopoli, Juliana Glavina, Rocío Espada, Nina Verstraete, Brenda Guzovsky. Gracias por no ser simplemente un número más en un entorno de trabajo, sino siempre brindar la calidad humana necesaria para que la vida del laboratorio haya sido tan fresca y llevadera.

A la gran familia de los grupos QB6, QB65, QB10 y derivados. Por los almuerzos, los cafés, los mates, las charlas... la vida dentro de este micromundo de la ciencia en exactas.

A mis amigos. Porque vine a Buenos Aires a hacer el doctorado, pero en el camino me llevo algo infinitamente más valioso, los momentos compartidos y su amistad de cara al futuro. A los nefastos, los abuelos, los del coro, los del malbec, los entrerrianos, los sanluiseños, los de la vida... gracias por tanto rock y color.

A mi familia, por dejarme ir del nido, por siempre decir si, por siempre creer en mis decisiones y en mis sueños, por siempre estar orgullosos de lo que sea que hiciera, por hacerme quien soy en este mundo.

A la memoria de mi abuela Chicha, quien hace poco se fuera de viaje a la eternidad. Por ser la luz de mi vida, la caricia del alma de todos los días, mi brújula moral, mi freno a la adultez.

A Marcia, el amor de mi vida. Por existir y caminar a mi lado.

A la educación pública. Porque cada peso que directamente o indirectamente se invirtió en mi educación es un peso que no se destinó a otro argentino para resolver alguna de sus

necesidades. Eternamente agradecido a mi pueblo y espero devolverles algún día algo de lo tanto que me han brindado.

"El tiempo, es el bien más preciado de la humanidad porque no se recupera nunca más". Gracias a todos aquellos que dedicaron tantos momentos de sus vidas a compartirlos conmigo en este camino.

# Publicaciones

**En el marco de esta tesis:**

**1.** (2015) **Parra RG**, Espada R, Verstraete N, Ferreiro DU. Structural and Energetic Characterization of the Ankyrin Repeat Protein Family. PLOS Comp. Biol. DOI:10.1371/journal.pcbi.1004659

**2.** (2015) R. Espada*, **RG Parra***, MJ Sippl, T. Mora, AM. Walczak, DU Ferreiro. Repeat Proteins challenge the concept of structural domains. Biochem. Soc. Trans. Oct 09, 2015, 43 (5) 844-849; DOI: 10.1042/BST20150083. *Jointly first author.

**3.** (2014) T Di Domenico, E Potenza, I Walsh, **RG Parra**, M Giollo, G Minervini, A Ihsan, C Ferrari, AV. Kajava, SCE. Tosatto. "RepeatsDB: A database of tandem repeat protein annotations". Nucleic Acids Research Database Issue 2014. doi: 10.1093/nar/gkt1175

**4.** (2013) **RG Parra** , R Espada , IE. Sanchez , MJ. Sippl , and DU. Ferreiro. "Detecting repetitions and periodicities in proteins by tiling the structural space ." J. Phys. Chem. B. DOI: 10.1021/jp402105j. Publication Date (Web): 11 Jun 2013.

**5.** (2012) M. Jenik, **RG. Parra**, LR Radusky, A Turjanski, PG Wolynes, DU Ferreiro. "Protein Frustratometer : A Tool to Localize Energetic Frustration in Protein Molecules". Nucleic Acids Res. 2012 Jul;40(Web Server issue):W348-51. Epub 2012 May 29. doi: 10.1093/nar/gks447

**Fuera del marco de esta tesis:**

**1.** (2015) **Parra RG***, Rohr CO*, Koile D, Perez-Castro C and Yankilevich P. ÏNSECT 2.0: a web-server for genome-wide cis-regulatory modules prediction"Bioinformatics. 2015 Dec 12. pii: btv726.

**2.** (2015) R Espada, **RG Parra**, T Mora, AM Walczak, D Ferreiro. Capturing coevolutionary signals in repeat proteins. BMC Bioinformatics 2015 July 2;16:207. doi:10.1186/s12859-015-0648-3.

**3.** (2013) CO. Rohr*, **RG Parra*** , P Yankilevich and C Perez- Castro. "INSECT: In Silico SEarch for Co-ocurring Transcription factors." *Jointly first author.. Bioinformatics Oxford Journals (2013). doi: 10.1093/bioinformatics/

**Opiniones / En el marco de actividades en sociedades científicas:**

**1.** (2014) **RG Parra**, FL Simonetti, MA Hasenahuer, GJ Olguin-Arellana, AK Shanmugan. Highlights from the 1st ISCB Latin American Student Council Symposium 2014. BMC Bioinformatics. 2015, 16 (Suppl 8)A1. Doi:10.1186/1471-2105-16-S8-A1.

**2.**(2014) Mishra T, **Parra RG**, Abeel T. "The upside of failure: how regional student groups learn from their mistakes.". PLoS Comput Biol. 2014 Aug 7;10(8):e1003768. doi: 10.1371/journal.pcbi.1003768

**3.** (2015) K Wilkins, M Hassan, M Francescatto, J Jespersen, **RG Parra**, B Cuypers, D DeBlasio, A Junge, A Jigisha and F Rahman. Highlights from the eleventh ISCB Student Council Symposium 2015. Submitted BMC Bioinformatics.

**Aceptadas:**

**1. Parra RG**, Schafer N, Radusky L, Tsai MY, Guzovsky AB, Wolynes PG, Ferreiro DU. Protein Frustratometer 2 A tool to localize energetic frustration in protein molecules, now with electrostatics. Aceptada en Nucleic Acids Research Web Server Issue 2016.

**2.** Turjanski P, **Parra RG**, Espada R, Becher V, Ferreiro DU. Protein Repeats from First Principles. Aceptada en Scientific Reports.

# Capítulo 1

# Introducción

## 1.1. Plegado de Proteínas y Paisajes Energéticos

De las macromoléculas encontradas en la biosfera (ácidos nucleicos, lípidos, azúcares y proteínas) las proteínas son quizás las más fascinantes de ellas, no solo por la belleza y diversidad de sus estructuras, sino por la complejidad del comportamiento que emerge de la composición de su estructura primaria, es decir, de la composición y orden relativo de los elementos que las componen, unos 20 aminoácidos.

Es fascinante pensar en lo mucho que hemos avanzado en nuestro entendimiento del micro mundo de estas peculiares moléculas. Merecedor de 2 premios Nobel, Sanger fue una de las mentes más influyentes en nuestro avance en la comprensión de como funcionan las biomoléculas. En 1951, reportó por primera vez la secuencia precisa de aminoácidos que componían una proteína, específicamente la correspondiente a la insulina [Sanger and Tuppy, 1951]. Posteriormente, en 1977, hizo pública la secuencia del genoma del bacteriófago $\phi$-X174 [Sanger et al., 1977]. En poco más de medio siglo, gracias al aporte de innumerables científicos hemos descubierto el código que relaciona la información contenida en los genomas y cómo ésta es transferida desde éstos, en un lenguaje de 4 letras, a las proteínas, cuyo lenguaje consta de 20. Tal es la fuerza del descubrimiento de este código, que hoy en día existen bases de datos muy grandes en donde la información de los genomas de cientos de organismos es depositada, al tiempo de que las porciones que corresponden a los genes son asociadas de forma unívoca a las proteínas que éstos codifican. Gracias a la existencia de este código genético, podemos

saber de forma precisa, qué cambios ocurrirán en la secuencia de aminoácidos dado un cambio en la secuencia del ADN correspondiente al gen que la codifica.

Sin embargo, a pesar del gran avance que todo lo anterior implica, todavía hay un nivel de complejidad superior de estas moléculas que aún no comprendemos en su totalidad. Una vez codificada su secuencia de aminoácidos, una proteína es capaz de adoptar un conjunto de estructuras definidas que constituyen su llamado estado nativo. El proceso mediante el cual las proteínas adoptan una estructura espacial definida a partir de su estructura primaria, se conoce como "plegado proteico". Dicho proceso es robusto, de forma que cada vez que una proteína es sintetizada en el interior de una célula (bajo las mismas condiciones fisiológicas) esta adoptará el mismo conjunto de estructuras, correspondiente a su ensamble nativo. En 1969, Cyrus Levinthal señaló que, debido al gran número de grados de libertad en una cadena polipeptídica desplegada, la molécula tiene un número astronómico de posibles conformaciones. Si una proteína de 100 residuos buscara su estado nativo probando conformaciones el azar, necesitaría de un tiempo mayor a la edad del universo para alcanzar a la conformación correcta. El hecho de que la mayoría de las proteínas se plieguen en el orden de segundos o minutos, hace necesario que haya algún tipo de heurística que baje considerablemente la dimensionalidad del problema y permita que las proteínas se plieguen en tiempos compatibles con los tiempos en que funcionan los sistemas biológicos de los que forman parte [Levinthal, 1968].

La clave de todo este problema es comprender que las proteínas son sistemas evolucionados y como tales, tiene propiedades particulares respecto de los polímeros que se pueden generar mediante combinaciones aleatorias de los aminoácidos naturales. Si se analizara el universo de posibilidades de generar polímeros aleatorios de un largo determinado, se observaría que la mayoría de ellos no son capaces de adoptar una estructura definida. Dentro de dicho espacio, sólo una fracción reducida de los mismos sería capaz de adoptar una estructura definida después de algún tiempo (tan grande como se quiera), denominados polímeros termodinámicamente plegables. Sólo una fracción ínfima de estos últimos sería capaz de adoptar una estructura en tiempos biológicamente significativos, denominados polímeros cinéticamente plegables a los cuales corresponden las proteínas (Fig. 1.1).

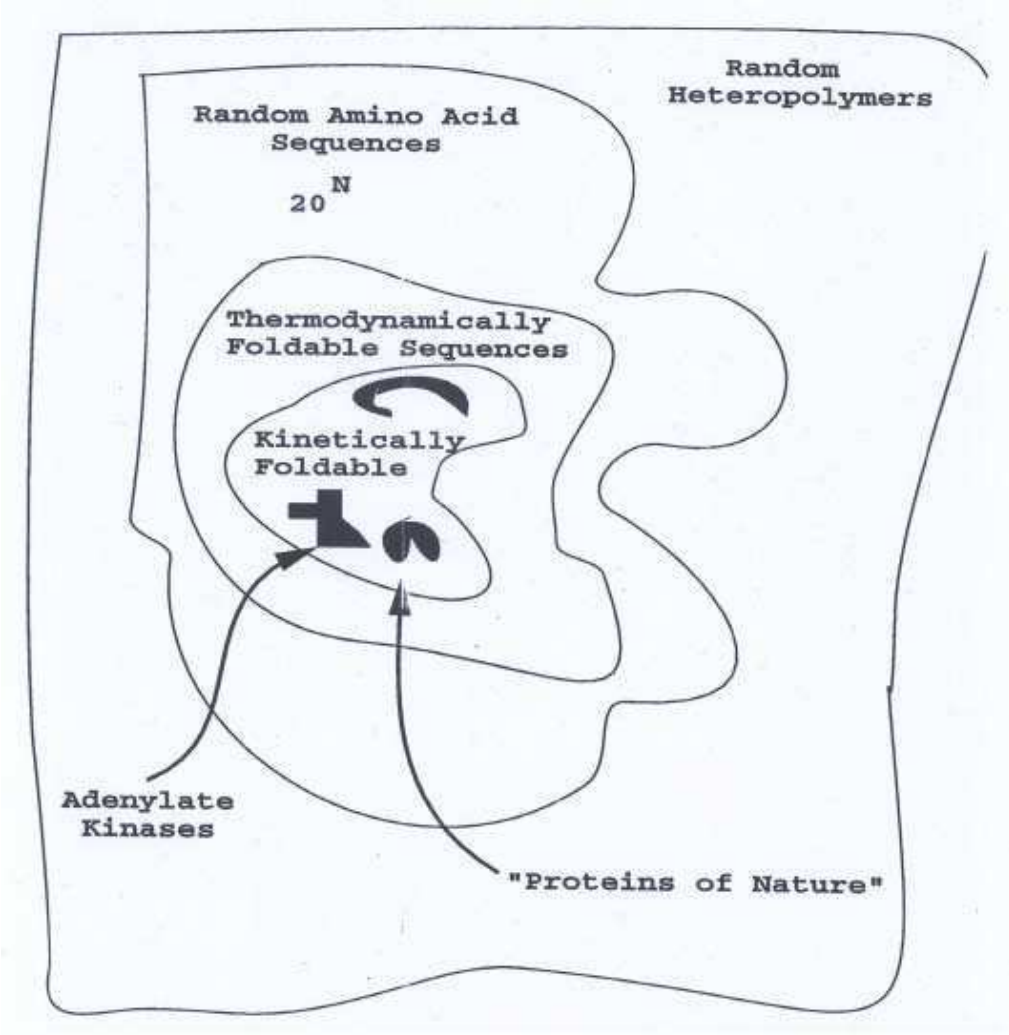


**Figura 1.1:** Descripción esquemática del espacio de secuencias de polímeros aleatorios. Las diferentes regiones se muestran sólo a modo ilustrativo y no están en escalas que representen las fracciones reales que ocupan en el espacio de secuencias.

En 1987, Peter Wolynes presentó la teoría de "Paisajes Energéticos" en la que explica que las proteínas son sistemas evolucionados de forma que las interacciones presentes en el estado nativo son mucho más favorables que las interacciones aleatorias durante el proceso de plegado. De esta forma, cada vez que se establece una interacción nativa en la cadena polipeptídica, la energía baja más de lo que lo haría en el caso de formarse una interacción aleatoria. Como consecuencia de esta cooperatividad entre las interacciones nativas, la forma global del paisaje energético de las proteínas naturales tiene la forma de un embudo corrugado (Fig. 1.2) en donde existe un fuerte sesgo energético hacia el estado nativo.

Una proteína no se pliega probando interacciones al azar sino que lo hace minimizando sus conflictos energéticos bajando por el paisaje energético hacia la base en donde converge a un conjunto de estructuras de mínima energía que constituyen el "estado nativo". Esto se conoce como "Principio de mínima frustración". Este principio no excluye que puedan existir conflictos energéticos remanentes en el ensamble nativo y por el contrario, es posible encontrar residuos que están en conflicto energético con la estructura en la que se encuentran, es decir, están energéticamente frustrados. Estos conflictos, han sido seleccionados evolutivamente y se ha descripto que la frustración que se mantiene en los estados nativos de proteínas tiene implicancias funcionales y ayuda a que la proteína pueda visitar los distintos estados que conforman el ensamble nativo o funcional [Ferreiro et al., 2014]. Un concepto muy importante que se deriva de todo lo dicho anteriormente es que las proteínas no están optimizadas para plegarse, sino que están optimizadas para funcionar. Si bien las secuencias de las proteínas modernas han sido seleccionadas de forma que poseen paisajes energéticos con un fuerte sesgo hacia un grupo de estructuras, ciertos conflictos energéticos específicos son mantenidos en lugares estratégicos de forma de ser útiles para que la molécula lleve a cabo su función. Si bien, en los últimos años se ha avanzado mucho en el campo de la predicción de la estructura de una proteína a partir de su secuencia, todavía mucho camino queda por recorrer en cuanto al diseño de una secuencia que pueda plegarse de una forma determinada y sea capaz de llevar a cabo una función determinada. ¿Cómo alterar el orden de aminoácidos de una proteína para modificar sus propiedades? ¿Cómo diseñar una proteína que se pliegue de una forma determinada y dónde ubicar los conflictos energéticos para que sea capaz de realizar una función? ¿Cómo

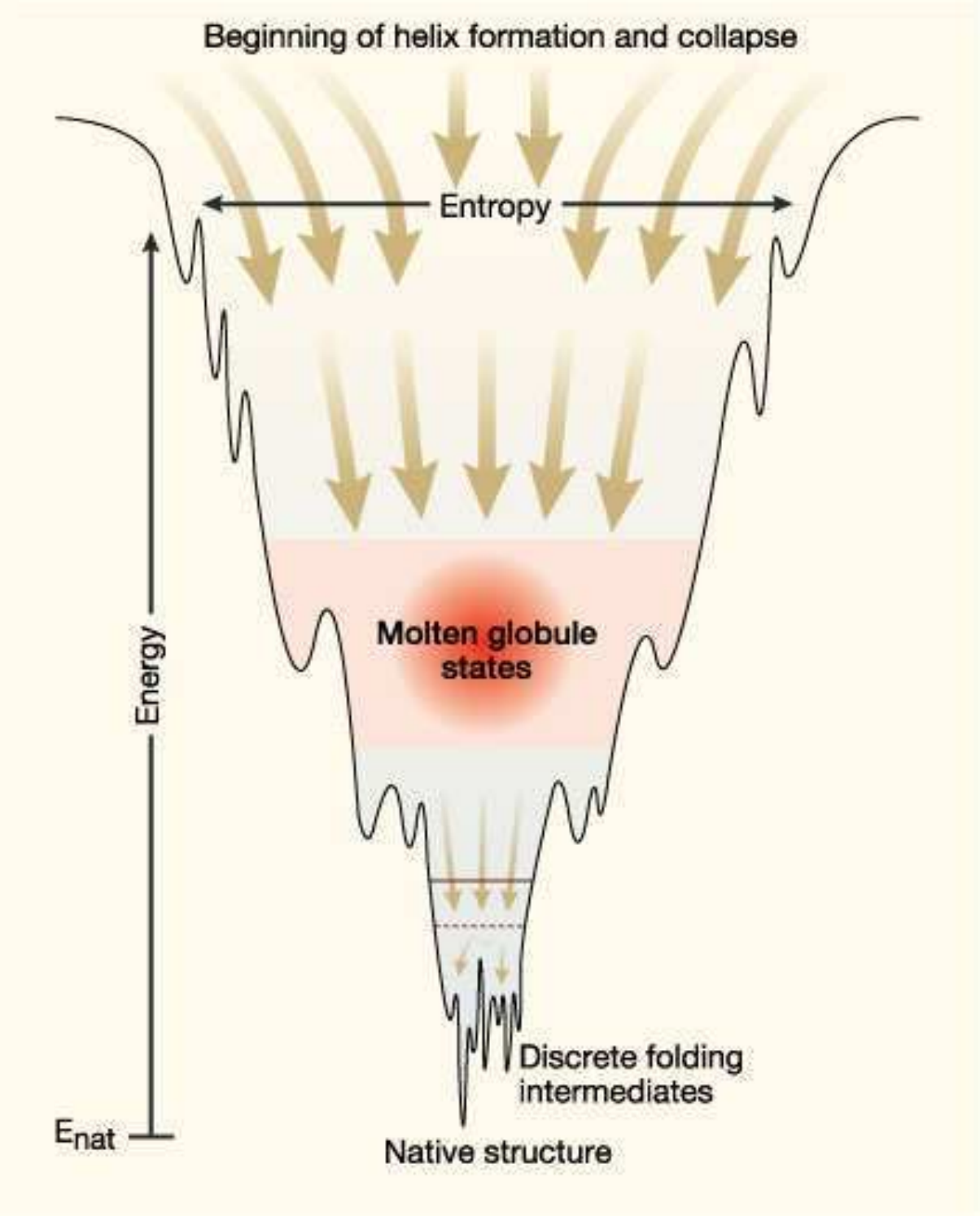


**Figura 1.2:** Esquema de un paisaje energético con forma de embudo que describe el proceso de plegado de una proteína natural. El ancho del embudo representa la entropía conformacional; la profundidad del embudo representa el cambio total en la energía entre el estado desplegado y el estado nativo. Esta figura fue extraída de [Liu et al., 2012] la cual es una modificación de la original que aparece en [Onuchic et al., 1995]

predecir el impacto que mutaciones puntuales en un gen tendrán en la estructura y la función de la proteína para la que éste codifica? Son algunas de las preguntas que aún debemos responder dentro del paradigma de "La secuencia codifica la estructura, la estructura codifica la dinámica y la función".

## 1.2. Proteínas Repetitivas

Una gran porción de proteínas naturales contienen motivos repetitivos en sus secuencias, los cuales corresponden con motivos estructurales y funcionales. Estas proteínas pueden clasificarse a grandes rasgos en 5 clases (Fig. 1.3) de acuerdo al largo de la unidad repetitiva [Kajava, 2012]. Aquellas proteínas con las repeticiones más pequeñas, menores a 5 residuos pueden formar agregados insolubles con estructuras cristalinas (clase I) o largas hélices enrolladas de estructuras fibrosas como el colágeno (clase II). En el extremo con repeticiones más largas podemos encontrar las proteínas de la clase "collares de cuentas" (*beads on a string*) con repeticiones de más de 50 residuos, que pueden formar dominios que pliegan de forma independiente (Clase V). En el medio de estos dos extremos encontramos aquellas repeticiones de entre 5-40 residuos que pueden formar estructuras abiertas del tipo solenoide (clase III) o estructuras del tipo toroidal cuya estructura se cierra dejando los extremos en proximidad (clase IV). Las clases III y IV constituyen los tipos más estudiados de proteínas repetitivas debido a la importancia funcional fundamental de muchos de los miembros que las componen. Cada clase dentro de esta clasificación se divide a su vez en subclases que representan familias o superfamilias, donde los miembros de las mismas se asumen relacionados evolutivamente.

Las estructuras de las proteínas del tipo solenoide están compuestas por repeticiones con largos entre $\sim$30 y $\sim$50 pudiéndose diferenciar por la composición de elementos de estructura secundaria de las mismas desde el tipo todo $\beta$ (por ejemplo las proteínas anti-congelamiento), pasando por aquellas con composición mixta $\alpha/\beta$ (por ejemplo, las proteínas de la familia *leucine rich*) hasta las proteínas del tipo todo $\alpha$ (por ejemplo las familias Armadillo y HEAT). Los solenoides se caracterizan por tener algunos de los dominios de plegado autónomo más grandes, en muchos casos con más de 500 residuos plegándose en una única estructura. Los toroides por su lado tienen su largo restringido debido a su naturaleza cerrada. Los casos más

conocidos de estructuras toroidales corresponden a los TIM-Barrels o $\beta$-Barrels cuyo número de unidades repetitivas es constante. Quizás un tipo de plegado toroidal más interesante corresponde a los $\beta$-Propellers que pueden mantener su estructura cerrada con una cantidad variable de repeticiones.

Una de las diferencias principales de las proteínas repetitivas con respecto de las globulares es que las interacciones entre los residuos que las componen se dan principalmente de forma local dentro de las repeticiones o entre repeticiones vecinas. Esta característica de localidad de las interacciones transforma a estas moléculas en un útil modelo para el estudio del problema del plegado de proteínas. La estructura modular de sus estructuras hace que sea más fácil cuantificar y localizar el efecto de perturbaciones locales en la secuencia y cómo éstas se propagan en las estructuras.

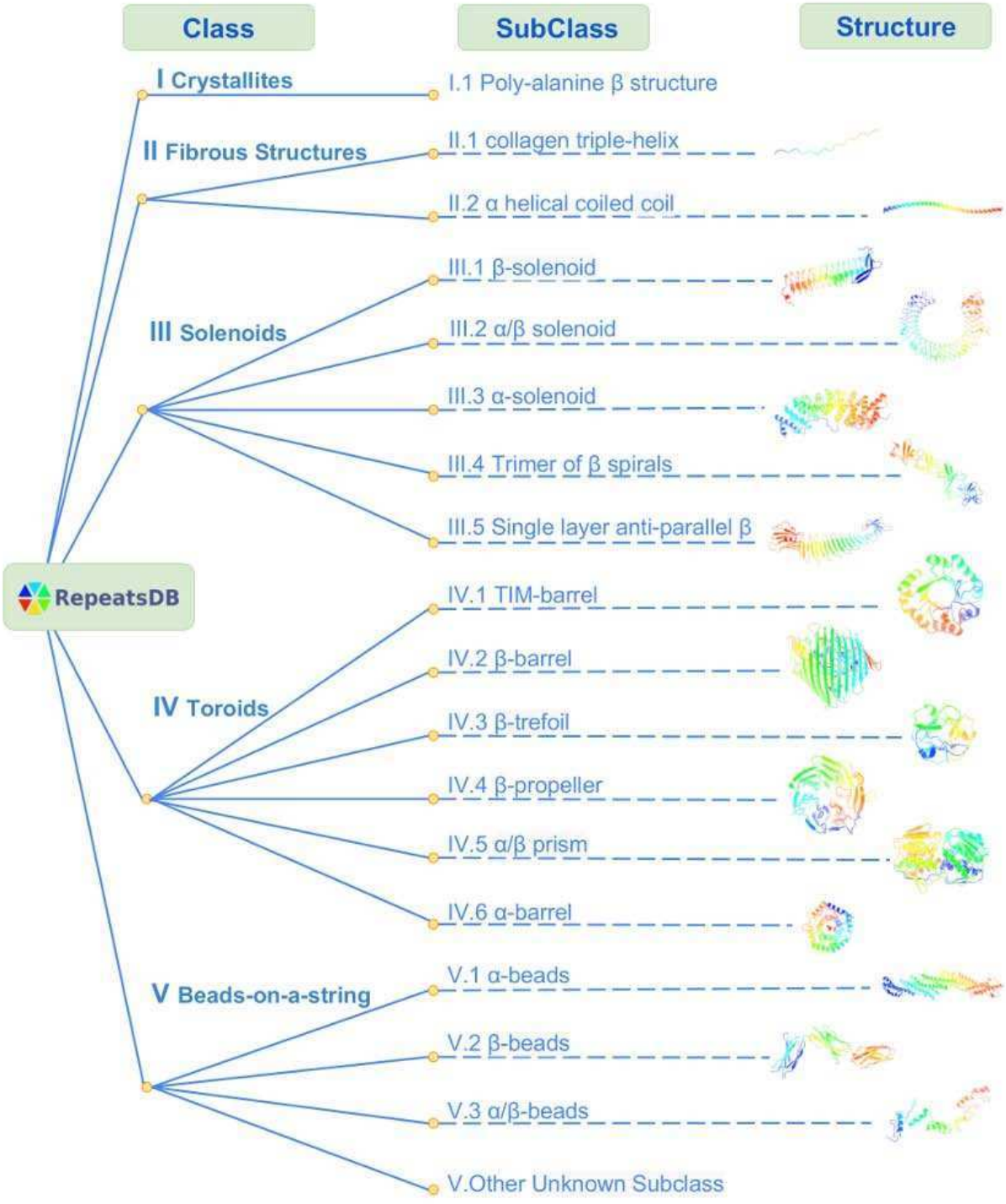


**Figura 1.3:** Clasificación de las proteínas repetitivas basada en la topología de la unidad repetitiva y la relación de simetría entre las diferentes repeticiones. Esta clasificación fue acuñada inicialmente por Kajava [Kajava, 2012], recientemente extendida en la publicación de la base de datos RepeatsDB, desarrollada en el grupo de Tosatto [Di Domenico et al., 2013]

Sin embargo, no fue hasta hace poco que este tipo de moléculas adquirieron notoriedad dentro del mundo proteico. Si bien constituyen un modelo interesante y simplificado desde el punto de vista estructural, plantean nuevos desafíos, dado que la mayoría de los métodos

bioinformáticos se han diseñado para ser aplicados en topologías del tipo globular. La simetría tanto a nivel primario como terciario sumado a la variabilidad en largos, de acuerdo a la cantidad de repeticiones en proteínas dentro de una misma familia, hacen que muchos métodos que son aplicados de forma rutinaria sobre familias globulares, deban ser adaptados para ser aplicados en proteínas repetitivas. Incluso lo que puede parecer la tarea más sencilla para el análisis y estudio de estas proteínas, que es definir donde empiezan y terminan las repeticiones, no es una tarea trivial. Esta dificultad se debe principalmente a que las secuencias que codifican para repeticiones de la misma clase, pueden ser muy divergentes de forma tal que la periodicidad observada a nivel estructural es indetectable al nivel de la secuencia de aminoácidos. La gran variabilidad secuencial que se observa en muchos casos hace que no exista actualmente un método estándar para detectarlas a partir de la secuencia y una gran cantidad de elementos repetitivos no son capaces de ser identificados o lo son de forma parcial.

## 1.3. La familia de Proteínas con Repeticiones de Ankirina

Las repeticiones de Ankirina fueron identificadas por primera vez en el regulador del ciclo celular Swi6/Cdc10 y en la proteína de señalización Notch en *Drosophila* [Breeden and Nasmyth, 1987]. Su nombre se debe a la proteína del citoesqueleto Ankirina, la cual contiene 24 copias de este tipo de repeticiones [Lux et al., 1990]. Desde entonces este tipo de proteínas han sido abundantemente descriptas. Las proteínas con repeticiones de Ankirina (ANKs) se encuentra en los tres súper reinos, incluyendo bacterias, arqueas y eucariotas como así también en numerosos tipos de virus. A pesar de su presencia a lo largo de todo el árbol de la vida, se ha descripto que este tipo de moléculas están particularmente enriquecidas en los organismos eucariotas. Los dominios modulares de proteínas como las ANKs, que pueden funcionar como andamiajes para las interacciones moleculares, están involucradas en numerosos caminos metabólicos que contribuyen a la gran complejidad de los organismos multicelulares [Marcotte et al., 1999]. En el caso de las ANKs presentes en bacterias y virus, se ha observado que en la mayoría de los casos, se encuentran enriquecidas en aquellos patógenos que se replican den-

tro de organismos eucariotas [Voth, 2011]. Se ha sugerido que la presencia de estas proteínas en dichos patógenos podría ser resultado de eventos de transferencia horizontal [Bork, 1993], aunque dada la alta divergencia de las secuencias de las mismas respecto de sus contrapartes eucariotas, no se descarta que eventos de evolución convergente pudieran haber ocurrido.

El número de repeticiones ANK por proteína puede variar. Si se analizan los números producto de hacer detecciones con los modelos ocultos de Markov (HMM) provenientes de las bases de datos de PFAM [Bateman et al., 2004] y SMART [Schultz et al., 2000], se observa que la cantidad de repeticiones puede ir desde 1 a 34 (en la proteína ORF EAA39756 del organismo *Giardia lamblia*), con una mayoría de proteínas que contienen 6 o menos unidades (Fig. 1.4). Los análisis en base a las detecciones realizadas por SMART sugieren que el número más común de repeticiones es dos, mientras que el número de repeticiones más común según las detecciones de PFAM, es tres [Mosavi et al., 2004]. Sin embargo, las repeticiones terminales son difíciles de detectar debido a la alta divergencia de secuencia que muestran respecto de las repeticiones internas. Esta variación en secuencia se debe principalmente a la divergencia de los residuos hidrofóbicos del marco canónico de los ANKs.

A lo largo de su historia evolutiva, las repeticiones terminales han ido reemplazando dichos residuos hidrofóbicos por residuos polares que facilitan la interacción con el solvente. Adicionalmente, las repeticiones terminales aparecen truncadas, lo cual representa otra dificultad para su detección. Por esta razón la cantidad de repeticiones en proteínas ANK puede ser mayor en una o dos unidades a los números mostrados en la Fig. 1.4.

Las repeticiones ANK se encuentran formando arreglos repetitivos únicos o combinados con dominios globulares dentro de la misma proteína. Cada repetición se pliega en una estructura que consiste de dos hélices antiparalelas seguidas por un $\beta$-hairpin o un loop largo. Las repeticiones consecutivas se apilan para formar un dominio con forma de L, que asemeja una mano en forma de copa donde los $\beta$-hairpins representarían los dedos (de ahí que también sean llamados *fingers*) y las hélices la palma. La estructura global de los dominios ANK es ligeramente curva, lo cual es evidente en las estructuras que contienen muchas repeticiones (ejemplo la proteína D34 que se observa en la Fig. 1.5).

Las proteínas ANK actúan principalmente como interactores proteicos y no se conocen

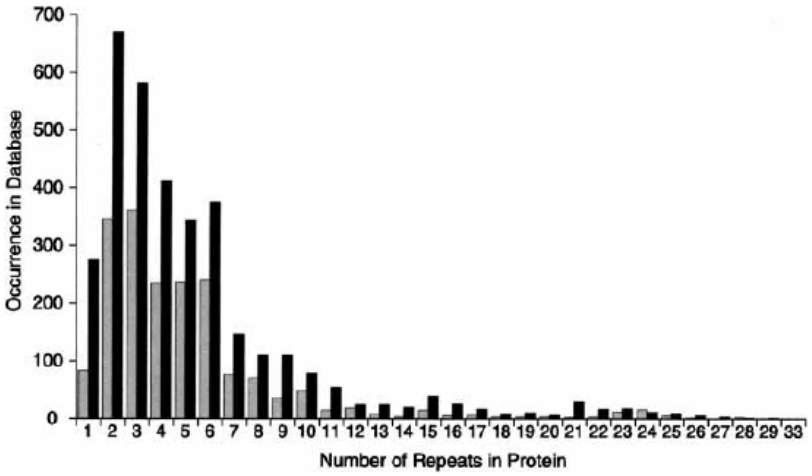


**Figura 1.4:** Distribución de número de repeticiones en ANKs según las predicciones usando los HMMs provenientes de las bases de datos de SMART (gris) [Schultz et al., 2000] y PFAM (negro) [Bateman et al., 2004]. Figura extraída de [Mosavi et al., 2002]

miembros de la familia con actividad enzimática [Sedgwick and Smerdon, 1999, Mosavi et al., 2002]. Debido a su versatilidad para reconocer estructuras de las más variadas formas, los dominios ANK se encuentran en proteínas con muy diversas funciones como señalización célula-célula, integridad del citoesqueleto, regulación de la transcripción y del ciclo celular, respuesta inflamatoria, desarrollo y varios fenómenos de transporte, entre otros. Ejemplos de proteínas que contienen dominios ANK son la familia de supresores tumorales INK4 p15, p16, p18, y p19, como también la proteína 53BP2, que es un regulador de la proteína supresora de tumores p53. Otra molécula conocida, que contiene repeticiones ANK es Notch, involucrada en múltiples decisiones del destino celular. Los factores de transcripción NF$\kappa$B que regulan la respuesta inflamatoria son inhibidos por I$\kappa$B que contiene 6-7 repeticiones ANK. Varios miembros de la familia de canales catiónicos TRPV, que funcionan como receptores sensibles al frío y calor o actúan como sensores mecánicos, contienen dominios ANK. En la Fig. 1.5 se observan las estructuras de varios miembros de la familia ANK.

La secuencia consenso de las repeticiones ANK contiene algunas firmas de aminoácidos que definen la forma de la repetición (Fig. 1.6). El motivo más prevalente es el TPLH que aparece en las posición 4-7 respecto de la fase definida por Peng (se denomina fase al aminoácido en que se considera que las repeticiones comienzan y terminan dado un arreglo de varias repeticiones) [Mosavi et al., 2002]. La Prolina en la posición 5 inicia un giro que es responsable de la forma de L de la estructura de los ANKs, que es estabilizada por puentes de hidrógeno entre las cadenas

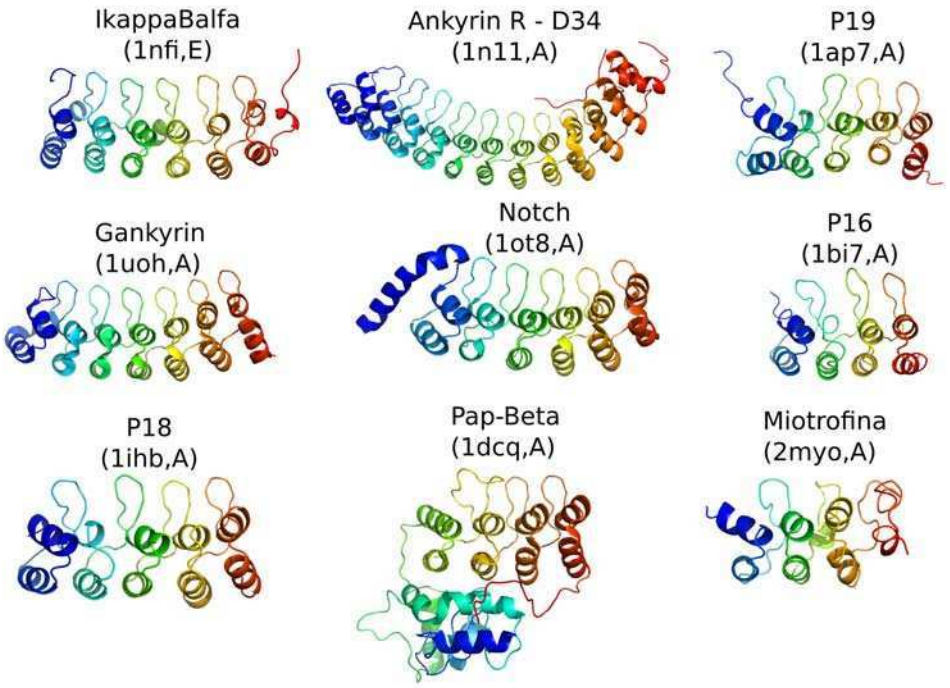


**Figura 1.5:** Representaciones como caricaturas de varios miembros de la familia ANK. Se uso el método de color arco iris para colorear los residuos. Se muestran los nombres comunes de cada molécula y entre paréntesis los identificadores de PDB con su respectiva cadena.

laterales de la Histidina en la posición 7 en la primer hélice y la Treonina de la posición 4 en la región hairpin/loop. Adicionalmente, las Glicinas conservadas en las posiciones 13 y 25 contribuyen a la terminación de las hélices, favoreciendo la transición a los loops intermedios y hairpin respectivamente. Los residuos no polares 9 y 18 (intra repetición) conjuntamente con las posiciones 8, 10, 17, 20 y 22 (inter repeticiones) y sumando las posiciones 6 y 21 conforman el núcleo hidrofóbico. Estos últimos son reemplazados generalmente por residuos polares en las repeticiones terminales.

### 1.3.1. Diseño de ANKs por consenso

La gran cantidad de secuencias conocidas para diferentes ANKs y su diversidad ilustra la versatilidad de este motivo. A lo largo de la evolución, las ANKs han mantenido la forma de L en su estructura que funciona en forma de andamiaje para interacciones proteína-proteína. Por dicha razón, el motivo ANK es muy atractivo como un modelo de diseño basado en consenso para la ingeniería y diseño de proteínas. Estrategias de diseño basado en consenso han

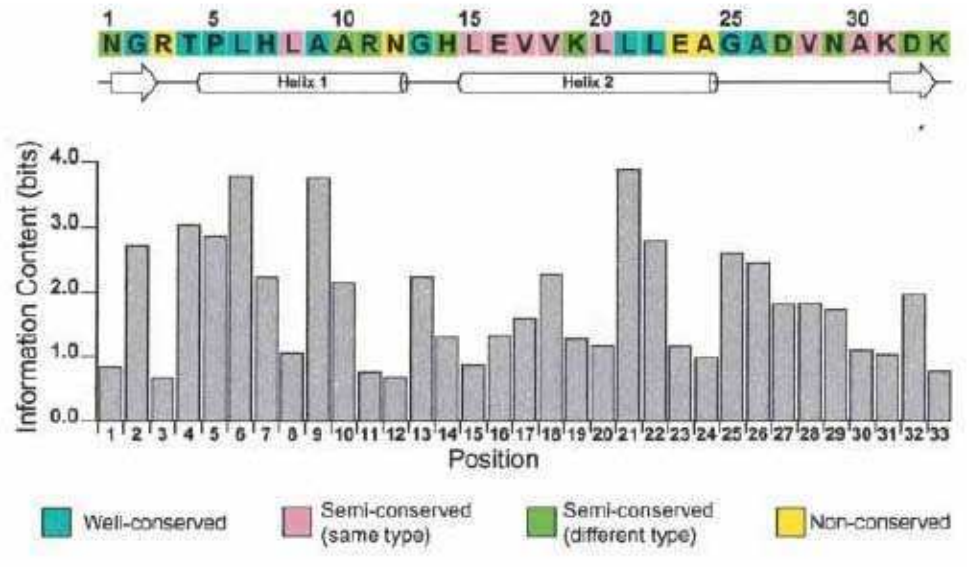


**Figura 1.6:** Secuencia consenso de una repetición ANK, según el grupo de Peng. Abajo, contenido de información medido a partir del alineamiento de secuencias de las repeticiones ANK, según Peng [Mosavi et al., 2002].

sido aplicadas previamente en varios tipos de proteínas con buenos resultados. Las proteínas del tipo "dedos de zinc" fueron de las primeras en ser diseñadas para unir secuencias de ADN específicas [Desjarlais and Berg, 1993]. El alineamiento de 13 Pythasas fúngicas fue usado para crear una Pythasa consenso con una actividad enzimática normal pero con una termoestabilidad significativamente mayor [Lehmann et al., 2000]. Otro ejemplo es el del motivo WW de 33 residuos de largo, cuya secuencia consenso fue determinada a partir de aproximadamente 60 dominios WW conocidos y que posteriormente estudios estructurales de alta resolución mostraron que este dominio WW prototípico era capaz de plegarse en forma correcta [Macias et al., 2000]

Varios grupos han sido exitosos en el diseño de ANKs por consenso [Mosavi et al., 2002,Binz et al., 2003, Tripp and Barrick, 2003, Main et al., 2003]. El grupo de Peng [Mosavi et al., 2002] usó el diseño por consenso para generar proteínas compuestas por 1-4 unidades repetitivas idénticas (1ANK, 2ANK, 3ANK y 4ANK) a partir de unas 4200 secuencias de repeticiones ANK derivadas de la base de datos de PFAM. La caracterización biofísica de estas moléculas mostró que aquellas compuestas por 3ANK y 4ANK eran monoméricas y capaces de plegarse correctamente, a diferencia de 2ANK que se encontraba plegada sólo de forma parcial y de 1ANK que se encontraba completamente desplegada. Estudios de desnaturalización térmica

de 3ANK y 4ANK revelaron además que estas moléculas diseñadas poseen termoestabilidades muy altas. Las estructuras de alta resolución obtenidas por difracción de rayos X, mostraron que ambas proteínas poseen estructuras bien empaquetadas y altamente regulares con una red de puentes de hidrógeno periódica. Estos resultados demostraron que las secuencias consenso derivadas de forma estadística contienen en sí mismas toda la información estructural necesaria para la arquitectura ANK, tanto para las interacciones internas de cada repetición, como de repeticiones vecinas. El grupo de Plückthun por su lado [Binz et al., 2003] generó su propio diseño consenso a partir de las secuencias disponibles en la base de datos de Smart, que en su momento correspondían a unas 229 repeticiones a lo que además incluyeron información de 10 estructuras de alta resolución correspondientes a diferentes ANKs. La estrategia además incluyó incorporar repeticiones N y C terminales adaptadas, a las que llamaron "capuchones" que fueron derivados de la estructura de la proteína GABPB. Adicionalmente, permitieron la incorporación de aminoácidos no consenso (excepto Glicina, Prolina o Cisteína) en las posiciones 2, 3, 5, 13, 14 y 33, correspondientes a la hélice más corta y la región del $\beta$-hairpin. Generaron variaciones aleatorias en las posiciones descriptas mediante la generación de librerías y seleccionaron 6 miembros con 4, 5 y 6 repeticiones que fueron caracterizadas biofísicamente, las cuales mostraron estar bien plegadas con alta estabilidad. La estructura del miembro con 5 repeticiones, llamado E3_5, fue dilucidada mediante cristalografía de rayos X, la cual mostró una estructura regular con las cadenas de los residuos bien empaquetadas [Kohl et al., 2003]. A pesar de las diferencias en el diseño, 3ANK, 4ANK y E3_5 son capaces de plegarse en forma correcta y con estabilidades mayores a cualquier otra proteína ANK previamente descripta. Estos resultados demuestran que los análisis estadísticos a partir de bases de datos de secuencias son una forma efectiva para diseñar marcos proteicos altamente estables que sirven como base para diseñar luego, proteínas para mediar interacciones proteína-proteína específicas. Varios grupos, con el de Plückthun a la cabeza han generado una gran variedad de moléculas de este tipo, capaces de reconocer y unirse a diversas proteínas [Pluckthun, 2015]. Sin embargo, ninguno de estos diseños es tan eficiente y versátil como sus contrapartes naturales a la hora de funcionar como interactores proteicos, principalmente debido a una falta de flexibilidad en el andamiaje estructural y aunque algunas estrategias ya han sido aplicadas para compensar esto [Schilling et al., 2014] todavía el diseño de interactores específicos se basa

principalmente en la aleatorización de posiciones de bajo contenido de información y posterior mejoramiento de las propiedades biofísicas de las moléculas generadas.

## 1.4. Objetivos

**Objetivo General:**

En este trabajo de tesis proponemos estudiar de forma general los miembros de la familia de proteínas con repeticiones de Ankirina (ANK) para profundizar el entendimiento de sus relaciones secuencia-estructura-dinámica-función. Analizaremos los patrones de conservación tanto de sus secuencias como de sus estructuras derivadas experimentalmente y evaluaremos el impacto de los mismos en la dinámica de plegado de varios miembros de la familia a partir de simulaciones computacionales. Debido a que las proteínas repetitivas son un caso especial dentro del mundo proteico que desafía la aplicabilidad de las herramientas bioinformáticas típicas, este trabajo alterna entre la aplicación de métodos estándar de la bioinformática y la biología computacional y el desarrollo de métodos propios. El estudio de esta familia de proteínas en particular permitirá ampliar los hallazgos aquí encontrados a otras topologías de proteínas repetitivas y a familias proteicas en general.

**Objetivos Particulares:**

**1)** Generar una base de datos de proteínas con repeticiones de Ankirina en donde se depositen los datos primarios referentes a secuencias, estructuras y datos termodinámicos experimentalmente derivados. Medidas secundarias derivadas tanto de las secuencias como de las estructuras también serán depositadas en la base de datos.

**2)** Desarrollo de un método y estrategia para detección consistente de repeticiones estructurales en proteínas.

**3)** Análisis de la conservación de motivos en secuencia, estructurales y energéticos a lo largo de los miembros de la familia ANK.

**4)** Análisis de la dinámica de miembros de la familia ANK mediante simulaciones computacionales del plegado del tipo grano grueso usando métodos basados en estructura.

# Capítulo 2

# Teselado Proteico: Simetrías y Periodicidades en Estructuras de Proteínas

De acuerdo a la teoría de paisajes energéticos, las proteínas son moléculas peculiares que poseen paisajes energéticos con forma de embudos, en contraste con polímeros aleatorios donde los paisajes energéticos son excesivamente rugosos [Bryngelson et al., 1995]. Debido a que los aminoácidos en las proteínas naturales parecen estar distribuidos de forma aleatoria [Weiss et al., 2000], tienen que existir correlaciones de alto nivel en las secuencias que conlleven a formas de plegado estables. La teoría de paisajes energéticos predice que es más fácil que una proteína posea un paisaje energético con forma de embudo si existen simetrías en las estructuras que conforman su estado nativo [Wolynes, 1996]. Que los paisajes energéticos posean una morfología del tipo embudo y la consecuente posibilidad de que la cadena polipeptídica pueda adquirir una estructura espacial estable, implica que patrones estructurales relativamente independientes entre sí pueden formarse en diferentes partes de la molécula, los cuales pueden coalescer hacia el ensamblado de estructuras de orden superior. Esto reduce enormemente el problema de la búsqueda conformacional planteada por Levinthal [Levinthal, 1968], organizando de forma repetitiva y eficiente bloques fundamentales, relativamente pequeños para la construcción de las estructuras globales, también denominados "foldones" [Panchenko et al., 1996]. La mera existencia de repeticiones o unidades fundamentales no garantiza que el

sistema sea simétrico, sino que además estas unidades deben organizarse en patrones de orden superior. Lo anterior significa que la existencia de periodicidades puede implicar cierto grado de simetría pero pueden existir repeticiones sin la necesidad de existencia de simetrías a un nivel global de la molécula. Es por esto que la detección de repeticiones y patrones repetitivos es un primer paso en el entendimiento de su ensamblaje en estructuras complejas. Estos mosaicos estructurales, simétricos y periódicos podrían ser parte de paisajes energéticos con múltiples embudos anidados que coalescen en un paisaje energético global, simplificando así el proceso del plegado en proteínas de esta clase [Wales, 1998, Ferreiro and Wolynes, 2008, Itoh and Sasai, 2009].

Múltiples algoritmos se han usado para caracterizar repeticiones en secuencias de proteínas [Luo and Nijveen, 2013, Kajava, 2012]. La mayoría de esos métodos están basados en el auto alineamiento de la estructura primaria, mientas que otros implementan análisis espectrales de características pseudo-químicas de los aminoácidos [Luo and Nijveen, 2013]. Debido a que la misma unidad estructural puede ser codificada por secuencias que parecen completamente no relacionadas entre sí, no es de sorprender que los métodos basados en secuencias fallen al inferir repeticiones estructurales cuando la similitud en secuencia es baja. En contraste a los métodos basados en secuencia, sólo existe un puñado de métodos para detectar repeticiones a partir de estructuras proteicas. Éstos usualmente buscan repeticiones mediante alineamientos de la estructura consigo misma [Shih and Hwang, 2004, Abraham et al., 2008]. Algunos métodos incluyen sofisticadas transformaciones de las matrices de alineamiento que mejoran la detección y caracterización de repeticiones estructurales [Murray et al., 2004, Taylor et al., 2002]. Otros métodos, basados en estrategias de aprendizaje computacional, pueden reconocer regiones repetitivas en estructuras del tipo solenoide con una gran velocidad de cómputo [Walsh et al., 2012]. Aunque múltiples familias con motivos repetitivos han sido descriptas [Marcotte et al., 1999, Kajava, 2012], todavía no existe una forma clara de definirlas de forma consistente y unificada [Schaper et al., 2012, Luo and Nijveen, 2013] incluso para parámetros básicos como el largo de la unidad repetitiva, la ubicación de las diferentes unidades repetitivas y cómo éstas se agrupan en patrones de orden superior.

En este capítulo describimos el desarrollo y aplicación de un método puramente estructural para la detección y análisis de patrones periódicos en estructuras de proteínas y definición de

unidades repetitivas. Las repeticiones son identificadas como motivos estructurales que al ser repetidos maximizan el cubrimiento del espacio estructural de la molécula de interés. Nuestro método muestra ser robusto en cuanto a la detección de unidades repetitivas de forma general para una gran variedad de proteínas pertenecientes a diferentes clases y familias, sin restricciones en cuanto al tipo de simetría entre las unidades repetitivas (rotacionales, traslacionales o mixtas). Además del diseño del algoritmo, hemos desarrollado conceptos y métodos para la detección y análisis consistentes de repeticiones en estructuras de proteínas. Haciendo uso de una herramienta de alineamiento estructural extremadamente robusta y rápida llamada TopMatch [Sippl and Wiederstein, 2008], como así también de métricas apropiadas [Sippl and Wiederstein, 2012], nuestro método analiza de forma exhaustiva que tan bueno es cada subfragmento para recomponer la estructura global de la que forma parte, mediante repeticiones de sí mismo. Esta evaluación se obtiene mediante el alineamiento estructural del fragmento, aplicando rotaciones y traslaciones apropiadas, para producir alineamientos sobre diferentes regiones de la estructura entera. El resultado es un método de teselado de las estructuras proteicas en términos de unidades básicas denominadas teselas (*tiles*). El proceso de teselado conlleva a una visualización intuitiva de las unidades repetitivas y como estas coalescen en estructuras de orden superior. Debido a que en los patrones de teselado pueden observarse de forma colectiva los puntajes de teselado de todos los posibles fragmentos de la estructura proteica al mismo tiempo, es interesante observar como pequeñas perturbaciones en repeticiones individuales se propagan a la simetría global de la molécula. Encontramos que algunas arquitecturas proteicas pueden ser descriptas como puramente periódicas (o quasi-periódicas) mientras que en otras se observan claras disrupciones en los patrones repetitivos. La principal ventaja de este método es que es totalmente independiente de la secuencia para encontrar los patrones de teselado y permite la detección y comparación de motivos estructurales recurrentes que pueden ser codificados por una variedad de elementos de secuencia, no necesariamente relacionados evolutivamente.

## 2.1. Alineamientos estructurales y teselas

Para poder caracterizar las repeticiones e identificar cuales son las teselas básicas en las estructuras proteicas, usamos la herramienta TopMatch [Sippl and Wiederstein, 2008, Sippl and Wiederstein, 2012]. Dado un par de estructuras, este algoritmo genera una lista exhaustiva de alineamientos, basados puramente en la información estructural de las moléculas, devolviendo además las correspondientes matrices de rotación y traslación para realizar la superposición. Estos alineamientos están ordenados de acuerdo a la maximización de la superposición de los átomos $C^{\alpha}$ equivalentes, lo cual se traduce en el puntaje S de TopMatch definido como $S = \sum_i^L e^{-r_i^2/\sigma^2}$ que provee una métrica para medir la similitud estructural [Sippl, 2008]. Para el cálculo de $S$, $L$ representa el largo del alineamiento entre las dos estructuras y $r_i$ es la distancia euclídea entre los átomos $C^{\alpha}$ equivalentes. Básicamente, $S$ es una función de $L$ y la desviación estructural de los fragmentos superpuestos estructuralmente, donde el factor de escala $\sigma$ determina cuanto se reduce $L$ en función de las desviaciones estructurales. En este trabajo, hemos usando un $\sigma = 6{,}35\mathring{A}$, siendo el valor reportado como óptimo previamente [Sippl and Wiederstein, 2012].

Las proteínas a veces contienen motivos estructurales recurrentes, que pueden ser considerados como repeticiones que contienen variaciones respecto de una unidad estructural básica. Para poder detectar este tipo de repeticiones estructurales, tratamos las estructuras como un mosaico y las descomponemos en unidades más pequeñas que denominamos *teselas* con la restricción de que deben ser estructuralmente similares entre sí. Una proteína no necesariamente está compuesta por una sola clase de teselas, ni tienen que éstas ser capaces de teselar la totalidad del espacio estructural. En cualquier caso, es posible identificar aquellas teselas que, cuando repetidas, maximizan la cobertura de la estructura mediante copias de sí mismas.

Dada una estructura, cada fragmento continuo de la cadena polipeptídica representa una posible tesela. Así, la búsqueda del largo óptimo de las mismas puede buscarse en un rango que va desde el largo $N$ de la proteína, hasta la unidad fundamental mínima, que sería sólo un aminoácido. Ya que las trazas de $C^{\alpha}$s de uno o unos pocos residuos son muy pequeñas para el análisis propuesto, careciendo de significado en cuanto a repeticiones estructurales, el largo mínimo en nuestro algoritmo fue fijado en 6 residuos. En una proteína de $N$ residuos de largo,

hay 1 tesela de largo $N$, 2 teselas de largo $N-1$, y así con un total de $N_T = \sum_{L=6}^{N}(N-L+1)$ teselas posibles en toda la estructura. Cada uno de esas teselas $T_i$ es luego usada como elemento de búsqueda (*query*) mediante TopMatch para identificar otras teselas $T_k$ que son estructuralmente similares a $T_i$. Cada match en el procedimiento de alineamiento es unívocamente identificado por su largo $L_{ik}$, su ubicación en la estructura mediante el centro del fragmento $Z_{ik}$, y el puntaje asociado al alineamiento de la tesela usada como elemento de búsqueda (*query*) con el fragmento en cuestión (*target*) $S_{ik}$. Los *matches* son luego ordenados por su valor $S_{ik}$, donde el alineamiento del *query* consigo mismo ($i \equiv k$) necesariamente representa el máximo puntaje, ya que el largo de su alineamiento es maximal y su desviación estructural igual a cero. Así $L_{ii} = S_{ii}$, es decir el puntaje obtenido a partir del alineamiento de una tesela consigo misma equivale al largo $L_{ii}$ del alineamiento. Del conjunto de datos total, ordenado segundo el puntaje $S$, extraemos aquel conjunto de fragmentos que maximizan la suma del puntaje de cobertura $C_i = \text{máx}(\sum_k S_{ik})$, donde dadas dos teselas $T_h$ y $T_k$ que se incluyan en dicha suma, no deben superponerse espacialmente sobre la estructura blanco. Esta suma define el parámetro $C_i$ de una tesela $T_i$ que fue usada para generar los alineamientos. Definimos entonces un puntaje asociado a cada tesela como: $\Theta_i = \frac{C_i - L_{ii}}{N - L_{ii}}$. Este puntaje representa la fracción del espacio estructural que puede ser cubierta por repeticiones de una tesela en particular. Cuando se considera la lista ordenada de hits, hay muchas formas de definir un conjunto de alineamientos no redundantes. El caso más restrictivo sería aquel en que sólo se incluyeran aquellas repeticiones $T_k$ para las cuales, la región alineada involucre a toda la tesela, es decir $L_{ik} \equiv L_{ii}$. Una variante más flexible es incluir todos los alineamientos donde $L_{ii}/2 < L_{ik} \leq L_{ii}$, que es cuando más de la mitad del largo de $T_k$ produce un alineamiento en $T_i$. En este último caso, también añadimos la restricción de que el primer y último residuo de cualesquieras dos teselas, $T_h$ y $T_k$, del conjunto aceptado para calcular el puntaje correspondiente de cobertura $\Theta_i$ no sean superponibles en el espacio estructural de la molécula evaluada.

## 2.2. Modelo Homogéneo

A los fines de explicar el algoritmo propuesto, hemos construido un modelo teórico para ejemplificar como funciona el proceso de teselado y que se espera del mismo. Hemos denominado a nuestro ejemplo como "modelo homogéneo" el cual corresponde a un modelo completamente simétrico en que cualquier subfragmento del mismo es alineable en múltiples regiones de la estructura total, conllevando a un cubrimiento maximal de la misma. Un modelo de estas características podría ser representado como una hélice $\alpha$ totalmente regular, o si se quiere, una línea recta en el espacio (Fig. 2.1).

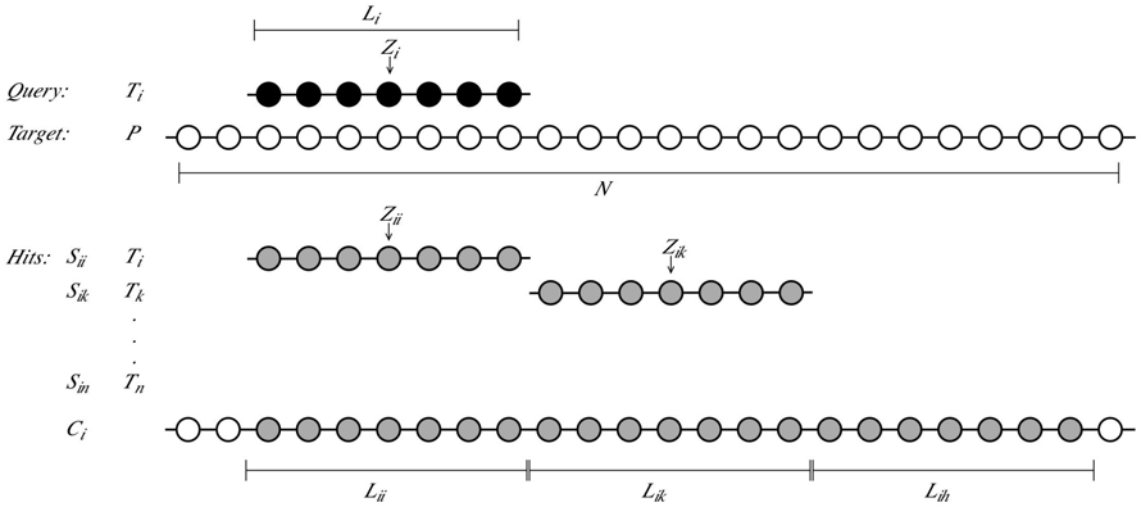


**Figura 2.1:** Teselado de un modelo homogéneo: Modelo esquemático de como se tesela una estructura a partir de alineamientos de fragmentos contra la estructura total

Cada tesela es caracterizada unívocamente por su largo $L_i$ y su centro respecto de la estructura total $Z_i$, satisfaciendo la siguiente ecuación:

$$\frac{L_i}{2} \leq Z_i \leq N - \frac{L_i}{2}. \tag{2.1}$$

Al superponer la tesela $T_i$ contra la estructura total se obtiene una serie de alineamientos que cumplen con la condición de no ser superponibles entre sí, sobre la estructura. Los puntajes de los alineamientos entre $T_i$ y la región $T_k$ tienen un valor máximo $S_{ik}$ igual a $L_{ii}$ en caso de que todos los residuos de ambas teselas sean superponibles tal que la desviación estructural sea mínima. Es útil recordar que el valor de $L_{ii}$, que es el largo de la tesela $T_i$ consigo misma, corresponde al largo de la tesela $T_i$. Si una tesela $T_k$ posee desviaciones respecto de otra $T_i$,

el valor de $S_{ik}$ será menor a $L_{ii}$ y por tanto, habrá residuos que no serán alineables.

El cubrimiento máximo se obtiene cuando las copias se arreglan de forma contigua en la estructura global, sin dejar espacios entre ellas (espacio subutilizado). En tal caso, los centros de las copias ($Z_{ik}$) son definidos de la forma:

$$Z_{ik} = Z_i + n \cdot L_i \qquad n \in Z. \tag{2.2}$$

En el caso en que sólo se acepten copias que alinean de forma completa ($S_{ik} = L_{ii}$) contra la estructura global, los valores de los centros de los fragmentos se encuentran restringidos a:

$$\frac{L_i}{2} \leq \quad Z_{ik} \quad \leq N - \frac{L_i}{2} \tag{2.3}$$

$$\frac{L_i}{2} \leq \quad Z_i + n \cdot L_i \quad \leq N - \frac{L_i}{2} \tag{2.4}$$

Esta expresión puede ser reacomodada para obtener:

$$\frac{1}{2} - \frac{Z_i}{L_i} \leq \quad n \quad \leq \frac{N}{L_i} - \frac{1}{2} - \frac{Z_i}{L_i} \tag{2.5}$$

Con límites correspondientes a:

$$n_{min} \quad = \quad \lceil \frac{1}{2} - \frac{Z_i}{L_i} \rceil \tag{2.6}$$

$$n_{max} \quad = \quad \lfloor \frac{N}{L_i} - \frac{1}{2} - \frac{Z_i}{L_i} \rfloor \tag{2.7}$$

El número total de copias no superponibles que se pueden ubicar a lo largo de la proteína se define como: $n_c = n_{max} - n_{min} + 1$ y el valor de cubrimiento correspondiente sería $C_i = n_c \cdot L_i$. Entonces, el valor de teselado $\Theta_i$ es:

$$\Theta_i = \frac{(n_c - 1)L_i}{N - L_i} \tag{2.8}$$

Con estas ecuaciones es posible recrear el patrón de teselado teórico de un modelo homogéneo de $N$ residuos de longitud. El patrón de teselado de una proteína de $N = 120$ residuos, de acuerdo al esquema de aceptación de copias de alineamientos enteros se muestra en la Fig. 2.2a.

Hasta aquí el modelo presentado tiene la restricción de que las copias aceptadas deben cumplir con la restricción de alinear la totalidad de sus residuos con la tesela evaluada. Una dificultad que surge de este modelo es que, en proteínas naturales, las repeticiones pueden variar en sus largos debido a la presencia de inserciones y/o deleciones y por tanto las mismas no serían aceptadas como repeticiones válidas. Además, diferentes repeticiones pueden presentar variaciones estructurales entre sí, provocando que residuos análogos no puedan ser alineados estructuralmente a pesar de estar presentes. Una extensión del modelo anterior para poder lidiar con estos inconvenientes es necesaria para poder aplicarse en proteínas naturales.

A continuación se describe como se generaliza el modelo homogéneo en caso de que se acepten copias parcialmente alineadas con la tesela evaluada, siendo el parámetro $\alpha$ la proporción máxima de tesela que se permite no alinear. La restricción que hemos usado es que al menos la mitad de los residuos que componen la tesela deben ser alineados con una copia para que esta última se considere válida, es decir $S_{ik} < L_{ii}$ con $L_{ii}/2 < L_k < L_{ii}$, $\alpha = 0{,}5$, dado el alineamiento de la tesela $T_i$ con su copia $T_k$. En primer lugar debemos calcular cuantos residuos se dejan descubiertos al teselar con copias enteras debido a una subutilización del espacio. La tesela que se localiza más a la izquierda, está centrada en $Z_i + n_{min} \cdot L_i$, y su inicio por tanto se localiza en $Z_i + n_{min} \cdot L_i - \frac{L_i}{2}$ que equivale también con el número de aminoácidos no cubiertos al principio de la proteína $C_{beg}$. De forma análoga, la tesela localizada más a la derecha se encuentra centrada en $Z_i + n_{max} \cdot L_i$ y la posición de su aminoácido final se localiza en $Z_i + n_{max} \cdot L_i + \frac{L_i}{2}$ y el número de aminoácidos no cubiertos entonces es $C_{end} = N - \left[Z_i + n_{max} \cdot L_i + \frac{L_i}{2}\right]$.

Si $C_{beg}$ y $C_{end}$ son mayores que $\alpha L_i$ cada tesela parcialmente alineada posee la siguiente

contribución al cubrimiento obtenido con copias enteras:

$$\chi(x) = \begin{cases} 0 & \text{if} \quad x < 0 \\ 1 & \text{if} \quad x \geq 0 \end{cases} \tag{2.9}$$

Y su correspondiente valor de teselado es:

$$\Theta_i = \frac{(n_c - 1) \cdot L_{ii} + C_{beg} \cdot \chi\left(C_{beg} - L_{ii}/2\right) + C_{end} \cdot \chi\left(C_{end} - L_{ii}/2\right)}{N - L_{ii}} \tag{2.10}$$

El patrón de teselado de una proteína modelo de $N = 120$ residuos se muestra en la Fig. 2.2b.

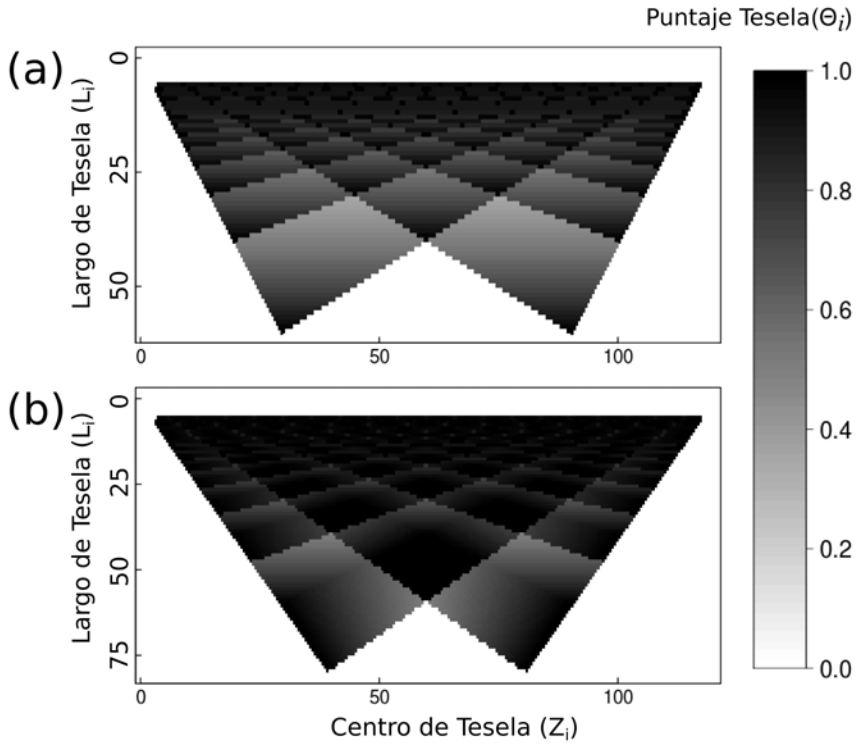


**Figura 2.2:** Teselado de un modelo homogéneo: Las teselas están ordenadas de acuerdo a su tamaño (eje vertical) y su centro (eje horizontal) expresados en unidades de aminoácidos. El puntaje de teselado $\Theta_i$ se muestra en escala de grises. El panel a) muestra el perfil de teselado obtenido cuando sólo copias enteras de las teselas son aceptadas y en el panel b) cuando copias parcialmente alineadas son aceptadas.

Los patrones correspondientes para cada esquema de aceptación de alineamientos entre la tesela y sus regiones análogas estructurales o copias, de la Fig. 2.2 se obtienen de aplicar las fórmulas descriptas anteriormente en un *script* en R, fijando el valor de $N = 120$

## 2.3. Teselado de Proteínas del Mundo Real

En la sección anterior mostramos a través de un modelo lo que se espera del método de teselado en el caso de ser aplicado sobre una estructura que está compuesta por repeticiones que son copias completamente alineables entre sí, o en caso en que por algún motivo, fracciones de las mismas no pueden ser alineadas.

Para ilustrar las propiedades básicas del teselado de estructuras proteicas, usamos la proteína 4ANK (pdb:1n0r,A con $N = 126$ residuos) que es un constructo sintético de repeticiones canónicas del tipo Ankirina [Mosavi et al., 2002]. Esta proteína tiene la particularidad de estar compuesta por 4 copias idénticas de la secuencia consenso (aunque la cuarta repetición posee un truncamiento de 6 residuos en su extremo C-terminal), obtenida del alineamiento de todas las repeticiones ANK conocidas hasta el momento de su publicación, dispuestas de forma contigua en su estructura primaria para luego obtener su estructura tridimensional. Como consecuencia de su completa periodicidad en secuencia, la estructura de 4ANK es muy regular y constituye un excelente ejemplo de aplicación intermedio entre el modelo homogéneo y proteínas naturales. La Fig. 2.3A muestra los puntajes de los mejores 15 hits para 3 fragmentos diferentes provenientes de la estructura, que fueron usados como tesela de búsqueda $T_i$. En todos los casos, la tesela que aparece más alta en el ranking es la correspondiente al auto-alineamiento contra sí misma ($i \equiv k$). En cada caso, los 2 hits subsiguientes con valores máximos corresponden a alineamientos con las otras 2 copias completas, ya que existe una cuarta repetición truncada en la estructura. Para las teselas subsiguientes en el *ranking*, el valor del teselado decae rápidamente.

A continuación, usamos el ranking para obtener un grupo de fragmentos, no superponibles entre sí, para poder cubrir la estructura de 4ANK mediante copias de la tesela $T_i$. Para cada posible tesela $T_i$ el valor de teselado $\Theta_i$ fue calculado, como se ha explicado anteriormente. El valor de $\Theta_i$ para teselas donde $L_{ii} > N/2$, es siempre cero, ya que no se pueden obtener repeticiones de esos fragmentos (Fig. 2.3B). La tesela más grande que puede ser ubicada dos veces es aquella donde $L_i = 57$ aminoácidos. Aquellas teselas anidadas dentro de teselas más largas, necesariamente tienen puntajes más bajos. En el patrón de teselado de 4ANK se pueden observar 2 o 3 repeticiones para $L_i = 33$, y 3 o 4 para $L_i = 24$ (Fig. 2.3B). Los puntos

más altos en dicha figura corresponden con aquellos fragmentos que ocurren más de una vez y para los cuales, cualquiera de sus extensiones ocurre menos veces, es decir, son *elementos maximales*. El rápido decrecimiento en $\Theta_i$ proviene de fragmentos que están anidados en las teselas maximales. Esto puede inferirse a partir del modelo homogéneo donde un grupo de teselas que ocurren $n$ veces obtienen puntajes $\Theta_i = (n-1)L_{ii}/(N - L_{ii})$.

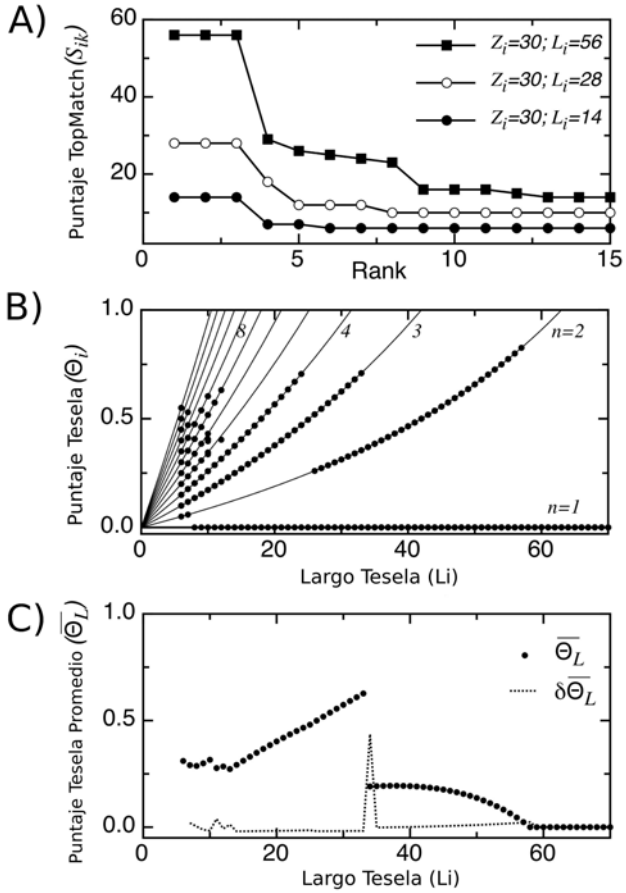


**Figura 2.3:** Cálculo del puntaje para las teselas: Se seleccionan fragmentos continuos de una proteína modelo como 4ANK (PDB:1n0r,A) y son alineados estructuralmente contra toda la proteína de la que provienen. Se genera una lista para cada fragmento, ordenada por el valor del parámetro $S_{ik}$ de TopMatch, donde tres casos se muestran en el panel a) $L_i$ representa el largo del fragmento en unidades de aminoácidos y $Z_i$ es el centro, de acuerdo a la numeración de los átomos $C^{\alpha}$ de los residuos contenidos en la estructura del archivo PDB de 4ANK. b) Distribuciones de los valores del puntaje $\Theta_i$ para cada largo de tesela ($L_i$). Cada punto en la imagen corresponde con el valor obtenido cuando se restringe el procedimiento a alineamientos perfectos ($S_{ik} = L_{ii}$). Las líneas corresponden a la expresión $\Theta_i = (n-1)L_i/(N - L_i)$ donde $N = 126$ y el número de copias de teselas que pueden repetirse corresponde a $n = 1, 2, 3..,12$ como se indica. c) Los puntos corresponden al promedio $\overline{\Theta_i}$ calculado para cada $L_i$. La línea de puntos representa la diferencia entre puntos consecutivos a lo que llamamos $\delta\overline{\Theta_i}$.

El hecho de que haya un número de teselas con puntajes similares pero de diferente largo $L_i$, implica que la arquitectura general de la proteína puede ser cubierta por un conjunto de teselas anidadas en donde las más largas pueden ser fragmentadas en teselas más pequeñas

que mantienen la característica de ser maximales. Así, la pregunta que surge es, ¿Cuál de esos largos conlleva a un máximo cubrimiento mediante el proceso de teselado? En el caso de proteínas reales, copias de teselas individuales generalmente exhiben variaciones estructurales con respecto a una unidad básica. Dichas variaciones reducen el puntaje $S$ de los motivos estructurales análogos alineados. La reducción relativa es en general más pronunciada para teselas pequeñas comparadas con aquellas más largas, que puede resultar en una disminución relativamente grande al puntaje de teselado $\Theta_i$. En resumen, si varias copias de una tesela pequeña tienen desviaciones estructurales significativas, entonces el puntaje $\Theta_i$ puede ser subóptimo respecto del obtenido por teselas de mayor tamaño.

Por lo anterior, es conveniente tomar el promedio $\overline{\Theta_L}$ sobre todos los puntajes $\Theta_i$ que corresponden a teselas con el mismo largo $L = L_{ii}$ (Fig. 2.3C). Es evidente a partir del ejemplo que el máximo ocurre en $\overline{\Theta}_L = 33$ residuos, indicando que las teselas de ese largo teselan la estructura de una forma más óptima respecto de aquellos con otros largos. Formalmente, el largo óptimo de tesela es obtenido como la raíz de la derivada de la función $d\overline{\Theta}_L/dL$, es decir que puede ser obtenido a partir de las diferencias finitas $\Delta\overline{\Theta}_L = \overline{\Theta}_L - \overline{\Theta}_{L-1}$. Una vez hecho esto, se identifica el largo óptimo $L$ y es posible seleccionar la tesela que maximiza el puntaje de teselado a dicho largo para obtener aquel fragmento $T_i$ que representa de mejor manera las repeticiones estructurales presentes en la molécula.

Dado que toda tesela $T_i$ se encuentra caracterizada por la posición relativa de su centro respecto de la estructura global $Z_i$ y su largo $L_i$, todo el conjunto de posibles teselaciones a partir de todos los fragmentos definibles en una estructura proteica es representable en 2 dimensiones en donde una tercera dimensión en escala de grises puede ser añadida representando el valor de la función $\Theta_i(L_i, Z_i)$ (Fig. 2.4). Esta representación muestra en que grado, cada tesela definible en la estructura, es capaz de teselar a esta última por medio de copias de sí misma. 4ANK (1n0r,A) es teselada de forma óptima por 2 repeticiones de un largo de 57 residuos, centrados en las posiciones 30 y 96. Al acortar el largo de las repeticiones, el puntaje decae hasta que se consideran fragmentos de 24 aminoácidos, donde el puntaje vuelve a ser máximo de forma local. Estas teselas de 24 residuos de largo, a su vez pueden descomponerse en unidades de 8 y 10 residuos. Estas unidades más pequeñas corresponden a las $\alpha$-hélices que son parte del motivo canónico de las ANKs (Fig. 2.4).

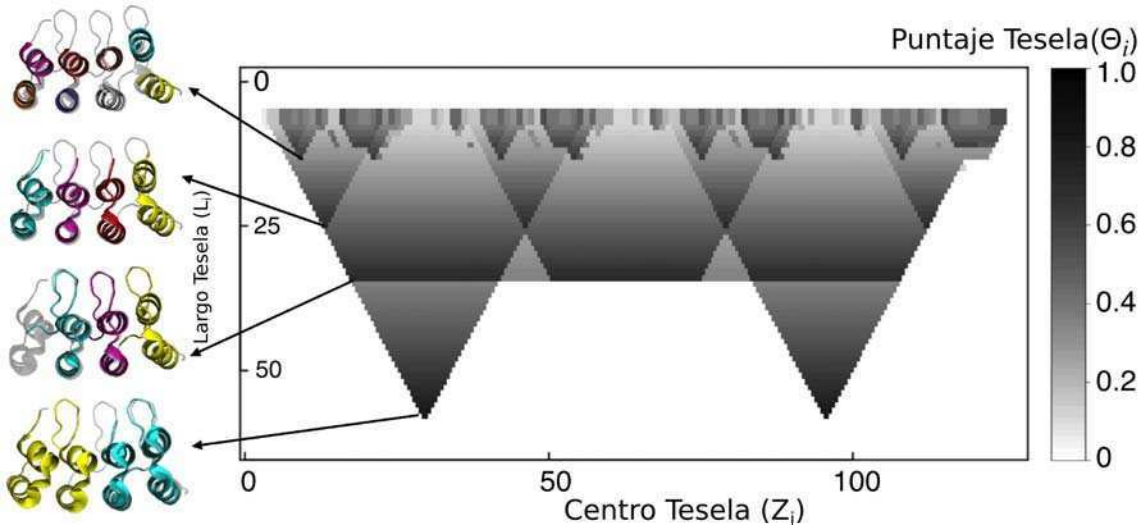


**Figura 2.4:** Teselado de una proteína altamente simétrica: Una proteína con repeticiones de Ankirina diseñada, 4ANK (pdb: 1n0r,A) con un largo de 126 residuos fue fragmentada en 7381 teselas diferentes. Estas se encuentran ordenadas de acuerdo a su tamaño (eje vertical) y su eje centro (eje horizontal) en unidades de aminoácidos. El valor de $\Theta_i$ para cada una de ellas se muestra en escala de grises. Las estructuras de la proteína y de los teselados respectivos a valores diferentes y representativos de $L_i$ se muestran en la izquierda. La estructura nativa se encuentra pintada en gris y sobre ella, superpuesta se encuentra la tesela elegida (en amarillo) como elemento de búsqueda y sus copias en cian, magenta, rojo, etc.

Al observar el patrón de teselado, es evidente que algo particular sucede para las teselas en donde $L_i = 33$. Cualquier tesela de este largo produce un teselado casi total de la estructura. Además, a ese largo las repeticiones que quedan definidas por las diferentes teselas están separadas por una distancia igual a su largo, es decir, se encuentran contiguas sin espaciamientos entre ellas. El hecho de que sin importar "la fase" de la tesela, el puntaje sea máximo, le da a la estructura una característica de onda (en el sentido de señal periódica) donde para un período fijo, cualquier fase permite reconstruir el patrón repetitivo, a excepción de ciertas teselas donde hay un decrecimiento en el puntaje de teselado producto de una condición de contorno, que produce una subutilización del espacio que puede ser cubierto. El largo característico en el ejemplo es igual a $L = 33$, donde la estructura puede ser maximalmente cubierta por casi cualquiera de las fases posibles $\phi = 0, 1, \ldots, L-1$.

Todas estas observaciones implican que aquellas teselas que cubren óptimamente la estructura repetitiva de una proteína poseen un puntaje promedio máximo para $\overline{\Theta_L}$ y además poseen un valor alto de $\Delta\overline{\Theta}_L$ (Fig. 2.3C). A partir del conjunto de teselas que contribuyen a $\overline{\Theta_L}$ definimos la mejor tesela $T_i$ como aquella que posee el puntaje máximo $\Theta_i(L_i, Z_i)$ respecto de las demás $T_k(L_i)$.

Como mencionamos anteriormente, 4ANK es un caso intermedio entre el modelo homogéneo y las proteínas que pueden encontrarse en la naturaleza. Las repeticiones en estructuras de proteínas son generalmente atribuidas a procesos de duplicación de ciertas regiones génicas correspondientes a secuencias particulares de aminoácidos. En general estas duplicaciones resultan en copias exactas del material duplicado. Al nivel de aminoácidos, la similitud entre las copias decae con el tiempo debido a la acumulación de sustituciones, inserciones y deleciones en las diferentes unidades. Las estructuras, en cambio, son más robustas ya que la similitud de las secuencias decae más rápido que la similitud al nivel estructural. Sin embargo, las inserciones, deleciones y demás variaciones también afectan la estructura tridimensional de las repeticiones individuales y por tanto, en proteínas naturales, las unidades repetitivas raramente son exactamente iguales y se encuentran generalmente afectadas por regiones no repetitivas entre ellas que han ido apareciendo y siendo retenidas evolutivamente. A continuación discutiremos los patrones de teselado obtenidos a partir de una variedad de estructuras proteicas correspondientes a diferentes tipos de proteínas naturales tanto repetitivas como globulares.

## 2.4. Teselado de Proteínas Repetitivas del Tipo Solenoide

Muchas proteínas naturales contienen repeticiones en tándem, compuestas por cadenas similares de aminoácidos. Estas moléculas son usualmente clasificadas en grupos, de acuerdo al largo de su unidad repetitiva fundamental. Las repeticiones cortas, de hasta 5 residuos, usualmente forman estructuras fibrilares como el colágeno o la proteína de la seda, mientras que repeticiones más largas que 50 residuos frecuentemente se pliegan como dominios globulares independientes [Kajava, 2012, Wolynes, 1997]. Hay una clase de proteínas repetitivas que se sitúa entre medio de los casos anteriores en donde el plegado de unidades vecinas se encuentra frecuentemente acoplado, conformando dominios que no son fáciles de definir [Espada et al., 2015]. Los defectos que se encuentran interrumpiendo la estructura regular en los arreglos repetitivos, aparte de dificultar la detección y anotación de los mismos, probablemente afectan también sus transiciones de plegado y función biológica. Por esta razón, aplicamos el procedimiento de teselado para definir las repeticiones en estructuras de proteínas naturales,

como así también la presencia de inserciones y deleciones de una forma puramente geométrica.

I$\kappa$B$\alpha$ es una proteína con repeticiones de Ankirina que inhibe la actividad del factor de transcripción NF-$\kappa$B [Ferreiro and Komives, 2010]. El procedimiento de teselado es capaz de identificar correctamente el largo correspondiente a la unidad canónica repetitiva que es igual a 33 residuos (Fig. 2.5a). Las repeticiones que se encuentran, contienen desviaciones respecto del largo canónico comprendiendo longitudes entre 30 y 39 residuos, indicando que no todas las repeticiones ANK son geométricamente equivalentes en el arreglo y contienen modificaciones estructurales. Aquellos fragmentos con puntajes máximos de teselado, pueden ubicar hasta 6 copias en la estructura, cubriendo cerca del 92 % de la estructura total (Tabla 2.1). Es aparente que la repetición ubicada en el extremo C-terminal se encuentra distorsionada respecto de las demás ya que los valores de $\Theta_i$ correspondientes a dicha región son menores. Esta observación sobre la repetición C-terminal concuerda con la incapacidad de los métodos basados en secuencia para detectarla. Si se analizan teselas que agrupan repeticiones consecutivas, es decir que tienen valores $L_i$ mayores que 33, se puede observar que son aquellas correspondientes a las repeticiones centrales las que maximizan el puntaje. Esto indica que las inserciones detectadas en teselas de largo 33, son responsables de distorsionar la simetría del arreglo a largos superiores. Tal vez no es coincidencia que en esta proteína se observa experimentalmente,*in vitro*, que su plegado consiste de 3 transiciones consecutivas, que en grandes rasgos, parecen coincidir con las regiones definidas al teselar la estructura con fragmentos de largo $L_i = 70$ [Ferreiro et al., 2007a, DeVries et al., 2011] .

La proteína aglutinante de germen de trigo (*wheat-germ agglutinin*, PDB ID: 1k7u,A) ha sido descripta como poseedora de 4 subdominios de tipo *hevein* [Muraki et al., 2002]. Al aplicar nuestro algoritmo de teselado, se detectan dos teselas de largo $L_i = 86$ residuos, como así también 4 repeticiones de largo $L_i = 43$, cubriendo en ambos casos la totalidad de la estructura (Fig. 2.5b). Al tomar el promedio de los valores $\Theta_i$ a cada valor de $L_i$ se puede ver una discontinuidad ocurriendo en un largo de 43, lo que constituye la frecuencia característica de esta estructura. A ese largo, casi todas las teselas son igualmente buenas para cubrir el espacio estructural con repeticiones de sí mismas. La disposición simétrica de las 4 mejores teselas a ese largo de fragmento hacen que la estructura sea altamente periódica produciendo que la fase de las repeticiones esté únicamente determinada por los límites de los extremos N

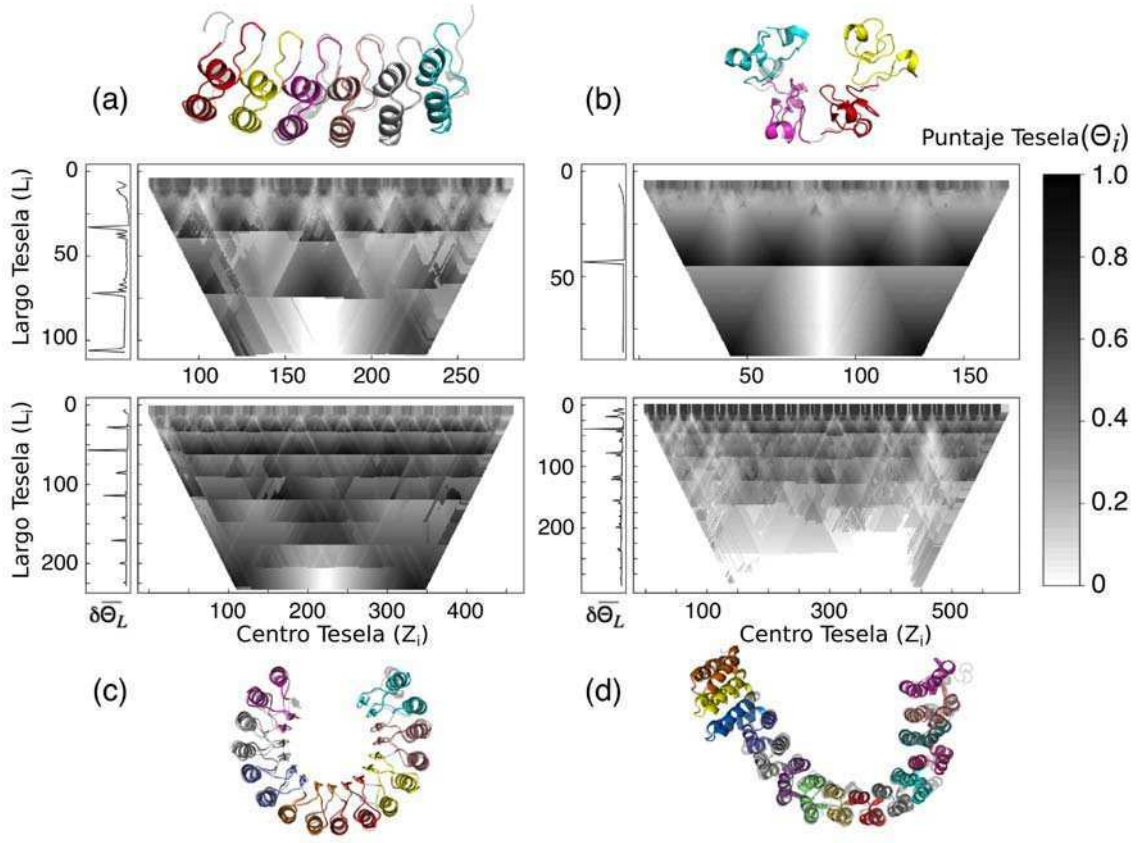


**Figura 2.5:** Teselado de proteínas repetitivas clásicas. El perfil de teselado se muestra en escala de grises, junto con el valor de $\delta\overline{\Theta_L}$ proyectado en la izquierda. Las estructuras de la proteína nativa y el teselado correspondiente al puntaje máximo a la frecuencia característica se muestran usando el mismo esquema de colores que la figura anterior. El largo ($L_i$) y el centro ($Z_i$) de la tesela seleccionada es: a) Ankyrin repeat: I$\kappa$B$\alpha$ (pdb:1nfi,E) $L_i = 33, Z_i = 191{,}5$ b) Hevein: proteína del germen de trigo (pdb:1k7u,A) $L_i = 43, Z_i = 150{,}5$ c) Leucine-rich: Porcine ribonuclease inhibitor (pdb:2bnh,A) $L_i = 57, Z_i = 139{,}5$ d) HEAT: PR65/A (pdb:1b3u,A) $L_i = 39, Z_i = 530{,}5$

y C-terminales.

El inhibidor de la ribonucleasa porcina (PDB ID: 2bnh,A) es una proteína con repeticiones del tipo ricas en Leucina (*leucine-rich*) en la que se han definido 16 repeticiones a partir de su secuencia. Aunque son muy similares al nivel de su estructura primaria, estas repeticiones no son estructuralmente equivalentes. Al aplicar el algoritmo de teselado, detectamos que hay en realidad 2 tipos diferentes de teselas de 28 y 29 aminoácidos de largo respectivamente (Fig. 2.5c). Se puede observar además, que estas repeticiones aparecen alternadas a lo largo de la estructura, produciendo un perfil escalonado a esos largos en el patrón de teselado (Fig. 2.6). Debido a que estas unidades están arregladas de forma simétrica, la estructura puede

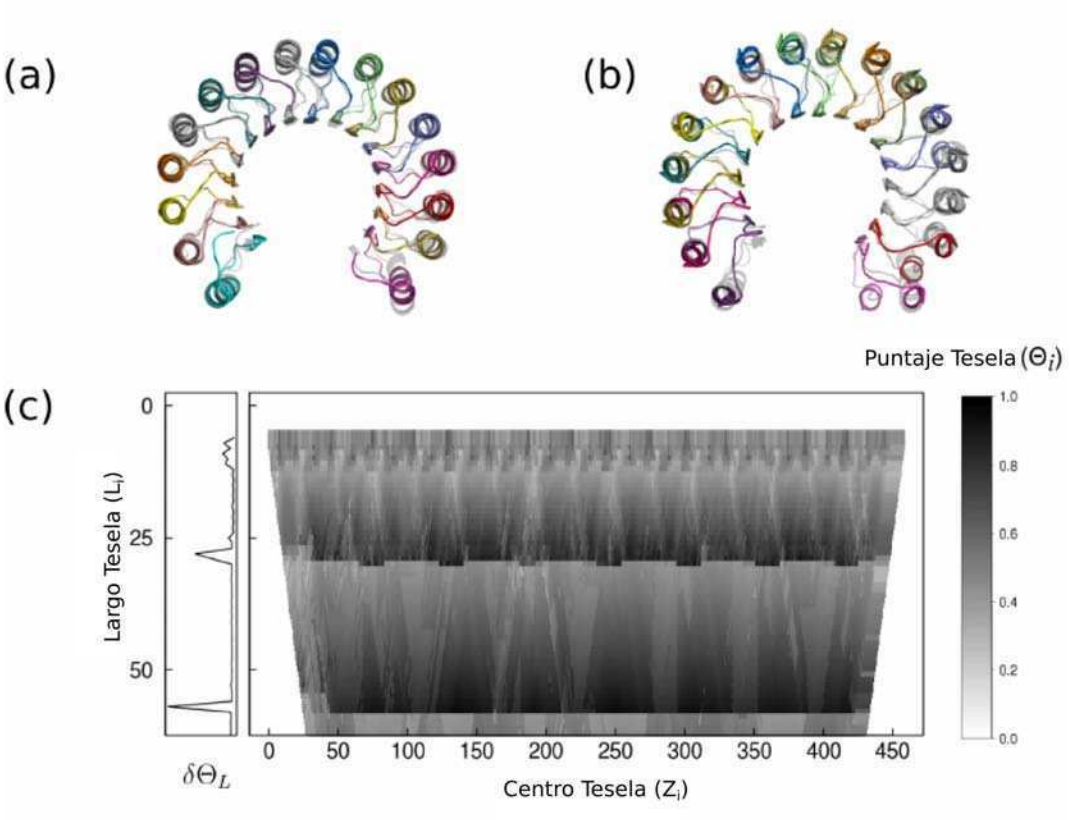


**Figura 2.6:** Acercamiento sobre el perfil de teselado de la proteína Porcine Ribonuclease Inhibitor (2bnh,A). El perfil de teselado se muestra en gris, junto a la gráfica que muestra los valores de la función $\delta\overline{\Theta_i}$ proyectada en la izquierda. Esta proteína tiene una frecuencia característica en $L_i$=57. Las teselas en ese largo están compuestas por otras teselas más pequeñas con $L_i$=28 y $L_i$=29, respectivamente que aparecen alternadas en la estructura proteica. Esto es evidente en el patrón con forma de sierra alrededor de largos $L_i \approx 30$ como así también en el pico secundario en $\delta\overline{\Theta_i}$ para esos largos. La estructura de la proteína y el teselado correspondiente se muestra en las figuras a) Teselado con $L_i$=28, $Z_i$ =388. c) Teselado con $L_i$=29, $Z_i$ =153.5.

componerse por fragmentos más grandes que incluyan pares de repeticiones de 28 y 29 residuos (Fig. 2.5c). Así, al largo $L_i = 57$ residuos, casi todo fragmento es tan bueno como los demás para explicar la estructura global mediante repeticiones de sí mismo. De esta forma, el largo de tesela que mejor explica la periodicidad en esta estructura es aquel que se compone por pares de repeticiones canónicas de diferentes largos. Anteriormente Haigis y colaboradores asignaron la repetición de 57 residuos de largo como la unidad evolutiva de esta proteína, mediante el análisis de los límites de exones en los transcriptos primarios [Haigis et al., 2002].

La subunidad de andamiaje de la proteína fosfatasa 2A, PR65/A, es una proteína repetitiva muy grande compuesta por repeticiones del tipo HEAT (acrónimo formado por letras presentes en las proteínas **H**untingtin, **E**F3, PP2**A** y **T**OR1) [Groves et al., 1999]. Nuestro algoritmo de teselado detecta una tesela maximal de 39 residuos de largo que puede cubrir la estructura

global con 15 copias de sí misma, lo que coincide con la detección del patrón de secuencia del motivo HEAT (Fig. 2.5). Esta proteína exhibe una estructura super-helicoidal, aunque se pueden observar irregularidades en el arreglo repetitivo entre repeticiones, lo cual afecta el patrón de teselado a largos superiores, generando grupos de repeticiones consecutivas en dicho patrón. El empaquetamiento periódico de las repeticiones HEAT es interrumpido entre las repeticiones 3 y 4 ($Z_i = 117$) y entre las repeticiones 12 y 13 ($Z_i = 471$) [Groves et al., 1999]. Esto se refleja a valores $L_i$ superiores donde las teselas localizadas alrededor del residuo 300 poseen puntajes más altos, de forma consistente, indicando que las repeticiones centrales se encuentran ordenadas entre sí de una forma más simétrica que las de los extremos (Fig. 2.5d).

## 2.5. Teselado de Proteínas Repetitivas con Topología Cerrada y del tipo Cuentas de Perla

En contraste con las arquitecturas solenoidales que exhiben usualmente las proteínas repetitivas clásicas, algunas proteínas muestran simetrías rotacionales. A menudo, los extremos N y C-terminales están en contacto, produciendo un cierre en la estructura de una forma polihédrica. Otro tipo de proteínas, contienen repeticiones con largos suficientes para que las mismas se plieguen de forma independiente en estructuras globulares repetidas, llamadas proteínas con repeticiones "cuentas de perla" (*beads on a string*). En esta sección, mostramos cómo el algoritmo de teselado es capaz de identificar repeticiones estructurales en miembros correspondientes a las topologías más comunes de estos tipos.

El TIM barrel es una de los tipos de arquitecturas proteicas más comunes en enzimas monoméricas [Nagano et al., 2002]. Típicamente está descripto como una colección de motivos $\beta$-$\alpha$ unidos por loops variables que se cierran en forma de cilindro u hojas $\beta$ paralelas, rodeadas por una capa de hélices $\alpha$. Hay una conservación estructural relativamente alta entre las proteínas de este tipo aunque sus secuencias puedan parecer no relacionadas, abriendo la discusión acerca de la naturaleza evolutiva de las unidades repetitivas y sus arreglos (duplicación y divergencia o evolución convergente) [Soding et al., 2006]. Hemos aplicado el

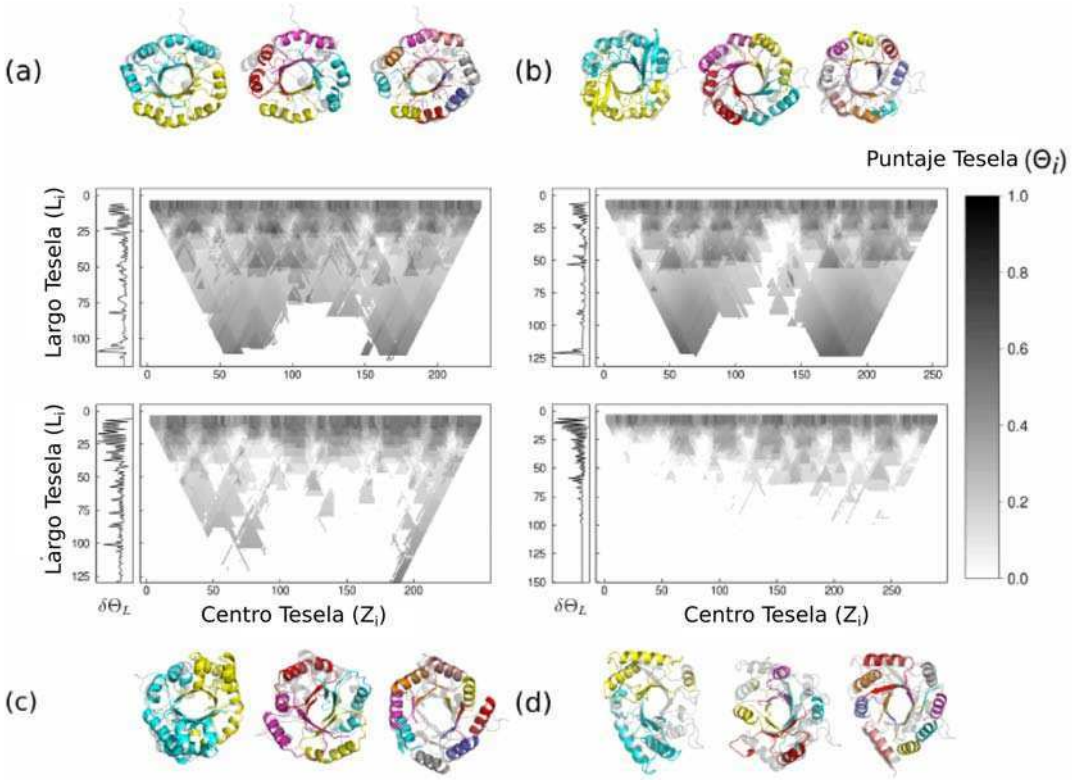


**Figura 2.7:** Teselado de diferentes TIM barrels. El perfil de teselado se muestra en escala de grises, junto con la función $\delta\overline{\Theta_i}$ proyectada sobre la izquierda. Las estructuras de las proteínas y sus patrones de teselado correspondientes a largos ($L_i$) y centros ($Z_i$) correspondientes con simetrías del tipo 2-fold, 4-fold y 8-fold se muestran en las figuras a) Ribulose-phosphate 3-epimerase (pdb:1rpx,A) 2-fold: $L_i = 108$, $Z_i = 63$, 4-fold $L_i = 46$, $Z_i = 92$, 8-fold $L_i$= 27, $Z_i = 105.5$ b) HisF (pdb:1thf,D) 2-fold: $L_i = 121$, $Z_i = 181.5$, 4-fold $L_i = 53$, $Z_i = 58.5$, 8-fold $L_i = 22$, $Z_i = 94$ c) Glycosomal, Triosephosphate isomerase (pdb:5tim,A) 2-fold: $L_i = 129$, $Z_i = 186.5$, 4-fold $L_i = 50$, $Z_i = 29$, 8-fold $L_i = 23$, $Z_i = 115.5$ d) Narbonin (pdb:1nar,A) 2-fold: $L_i = 97$, $Z_i = 116.5$, 4-fold $L_i = 43$, $Z_i = 180.5$, 8-fold $L_i = 25$, $Z_i = 251.5$

procedimiento de teselado sobre algunos de los casos más discutidos y para la mayoría detectamos señales de periodicidad para aquellas teselas que pueden repetirse 2, 4 y 8 veces (Tabla 2.1, Fig. 2.7). No todos los TIM barrels muestran la misma frecuencia característica. Algunas de las estructuras son mejor descriptas por teselas que corresponden a medio barril (2.8a), mientras otras poseen señales comparables a largos correspondientes tanto para medio barril como para un cuarto (Fig. 2.7). Los casos más irregulares tienen frecuencias características a largos de fragmentos muy pequeños, comparables con el tamaño de elementos de estructura secundaria simples (Tabla 2.1). Soding *et al* ya habían anotado desviaciones equivalentes en esta familia topológica basados en patrones de secuencia [Soding et al., 2006].

Un gran número de proteínas adopta la arquitectura del tipo $\beta$-propeller, siendo éste uno de los tipos de plegado más frecuentes en proteínas y al igual que el TIM-Barrel es considerado

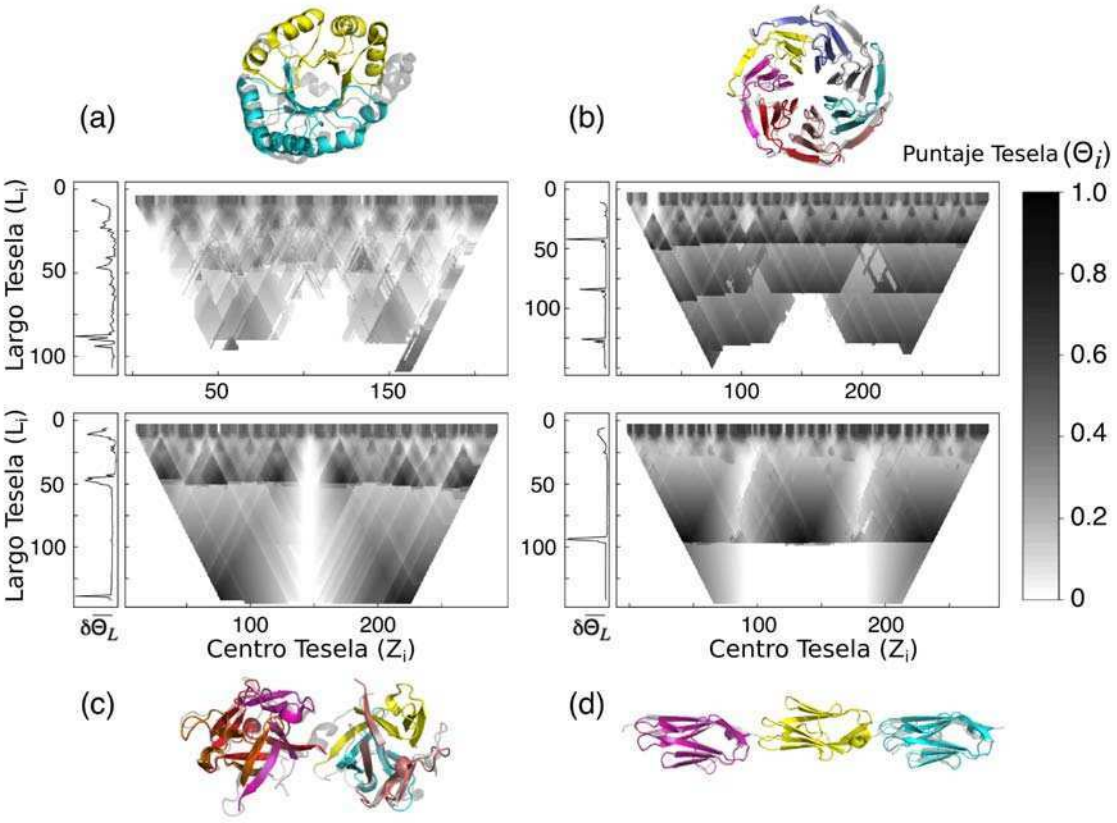


**Figura 2.8:** Teselado de proteínas repetitivas con topología repetitiva cerrada. El perfil de teselado se muestra en escala de grises, junto con los valores de $\delta\overline{\Theta_L}$ proyectados en la izquierda. Las estructuras de la proteína nativa y el teselado correspondiente al puntaje máximo a la frecuencia característica se muestran, usando el mismo esquema de colores que en la figura anterior. El largo ($L_i$) y el centro ($Z_i$) de la tesela seleccionada: a) TIM barrel: (pdb:1fq0,A) $L_i = 88, Z_i = 60$ b) $\beta$-propeller (pdb:3ow8,A) $L_i = 42, Z_i = 200$ c) Trefoil (pdb:1ybi,A) $L_i = 142, Z_i = 223$ d) Ig-repeats (pdb:2rik,A) $L_i = 94, Z_i = 140$

un superfold [Orengo et al., 1997]. Los propellers contienen un número variable de hojas $\beta$ antiparalelas ordenadas de forma radial, denominadas "aspas" (*blades*) [Fulop and Jones, 1999]. Nuestra estrategia nos permite identificar las aspas de forma correcta, en los casos testeados, observando propellers que contienen 4, 5, 6 y hasta 7 aspas (Fig. 2.9), incluso en presencia de un dominio no propeller en la misma cadena polipeptídica (Fig. 2.9d).

Una excepción interesante ocurre en la subclase de los propellers pertenecientes a la familia WD-40, donde las teselas maximales no se corresponden con las aspas (Fig. 2.8b). En el caso de los miembros de la familia WD-40, se detecta una frecuencia característica de $L_i = 42$ aminoácidos, con teselas que se repiten 7 veces, contribuyendo con 3 hojas $\beta$ para un aspa y una hoja $\beta$ para el aspa que se encuentra de forma contigua (Fig. 2.8b). Es notable que esa

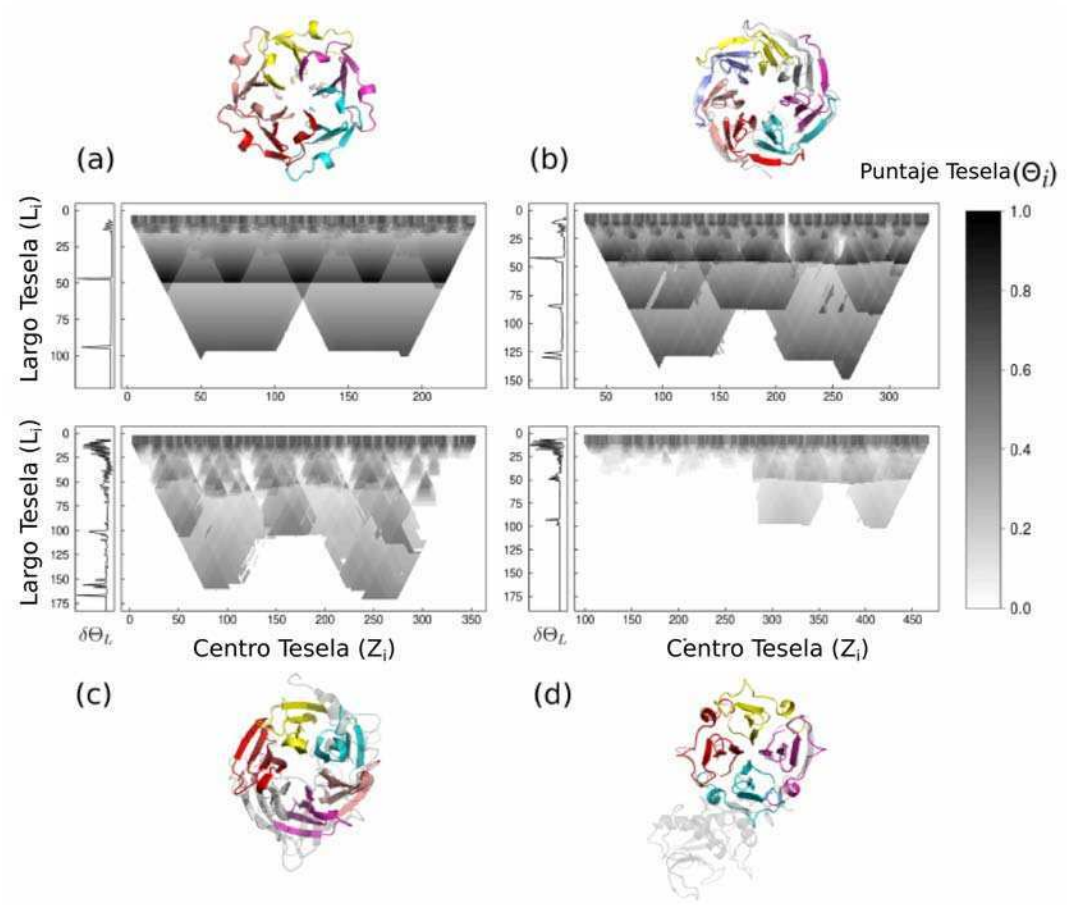


**Figura 2.9:** Teselado de diferentes $\beta$-propellers. El perfil de teselado se muestra en escala de grises, junto con la función $\delta\overline{\Theta_i}$ proyectada sobre la izquierda. Las estructuras de las proteínas y sus patrones de teselado correspondientes a largos ($L_i$) y centros ($Z_i$) son mostrados en las figuras a) Propeller de 5 aspas, Tachylectin-2 (pdb:1tl2,A), $L_i = 47$, $Z_i = 213.5$ b) Propeller de 7 aspas, WD repeat-containing protein-5 (pdb:3smr,A), $L_i = 42$, $Z_i = 138$ c) Propeller de 6 aspas, 3-phytase (pdb:3ams,A), $L_i = 45$, $Z_i = 251.5$ d) Propeller de 4 aspas, Interstitial collagenase (pdb:1fbl, A) $L_i = 49$, $Z_i = 399.5$

fase en particular fue originalmente descripta por otros investigadores cuando aún no había estructuras disponibles para miembros de la familia [Neer et al., 1994].

La proteína hematoglutinante HA33 de *Clostridium botulinum* es una proteína asociada a neurotoxinas que se pliega de forma tal que posee dos subdominios del tipo $\beta$-trefoil (Fig. 2.8c). La frecuencia catacterística ($L_i = 142$) nos lleva a dos fragmentos que tienen el máximo $\Theta_i$ y que corresponden a cada subdominio trefoil. La mejor fase al segundo pico ($L_i = 46$) corresponde a teselas que pueden ser ubicadas 3 veces en cada subdominio y que son compatibles con la unidad foil anotada para la arquitectura $\beta$-trefoil.

Por último, también hicimos un teselado sobre la estructura de la proteína titina como representante de la topología del tipo “beads on a string“ o cuentas de perla. Este tipo de

arquitectura se conforma por repeticiones con largos suficientes, de forma que pueden plegarse de forma independiente. La Fig. 2.8d muestra el resultado del teselado para un fragmento de la estructura de la proteína Titina, correspondiente a 3 repeticiones en tándem de tipo similar a inmunoglobulina (Ig). A $L_i = 94$ aminoácidos, la mejor fase coincide con los dominios Ig típicos. También se observa que otras fases también obtienen puntajes altos a ese largo de fragmento lo cual es un indicativo de la regularidad en el ordenamiento de las repeticiones.

## 2.6. Teselado de Proteínas Globulares

En algún nivel, todas las proteínas están formadas por repeticiones de aminoácidos. La simetría de las interacciones del *backbone* en las estructuras secundarias fueron claves para la propuesta de Pauling y Corey de que éstas emergían de la repetición regular de los enlaces peptídicos planares [Pauling et al., 1951,Pauling and Corey, 1951]. Se han propuesto motivos de estructura secundaria como candidatos para ser los bloques fundamentales para la construcción de proteínas globulares, en línea con el éxito de los métodos de predicción estructural basados en el ensamblaje de fragmentos [Moult et al., 2011,Hegler et al., 2009,Simons et al., 1997]. Ya que las repeticiones pueden ser encontradas de forma robusta mediante el proceso de teselado estructural, hemos explorado hasta qué punto estructuras proteicas sin repeticiones evidentes pueden ser vistas como una composición de teselas, ilustrando mediante algunos ejemplos clásicos.

Las proteínas $\beta\gamma$-crystallina incrementan el índice de refracción y mantienen la transparencia de los ojos en los vertebrados. Su estructura presenta un claro ejemplo de motivos estructurales coealesciendo en patrones de orden superior. El teselado de las estructuras de estas moléculas establece que la estructura puede ser bien descripta mediante 2 repeticiones de un barril-$\beta$ de 8 hojas con $L_i = 87$ residuos, centrados en las posiciones $Z_i = 44{,}5$ y $Z_i = 133{,}5$ (Fig. 2.10a). A su vez, cada uno de estos barriles puede ser descompuesto en 2 unidades de 40 residuos que corresponden con el motivo de llave griega, que puede ser a su vez descompuesto en 3 hojas$\beta$ de 10 residuos de largo. La frecuencia característica se encuentra en $L_i = 43$ aminoácidos, que corresponde con el motivo de llave griega a modo de repetición fundamental. Es aparente que las irregularidades presentes en la estructura hacen que la segunda y

cuarta llaves griegas tengan mayores valores de $\Theta_i$ que las demás y por ende diferentes largos maximales $L_i$.

Otro caso que hemos analizado corresponde a la Mioglobina en donde cerca del 70 % de su estructura puede ser descripta por 6 copias de un fragmento de 18 residuos de largo. Este fragmento corresponde a la hélice-$\alpha$ 'B', y constituye un fragmento maximal. El puntaje a mayores largos de fragmentos disminuye rápidamente (Fig. 2.10b). En este caso no somos capaces de detectar una frecuencia relevante correspondiente a fragmentos más largos que los segmentos $\alpha$-helicoidales, indicando que estos fragmentos no se repiten en una forma simétrica capaz de generar patrones de ordenes superiores, algo que sorprendió en su momento a Kendrew *et al* cuando resolvieron la estructura cristalográfica de esta molécula [Kendrew et al., 1958]. El teselado de esta estructura muestra que no es suficiente que existan fragmentos similares en la molécula para que esta sea simétrica, sino que los mismos deben relacionarse entre sí de una forma periódica para coalescer en una estructura simétrica.

La proteína verde fluorescente se pliega como un barril-$\beta$ con una hélice coaxial, con su fluorósforo localizado en la hélice central [Ormö et al., 1996]. Se identifican, mediante el teselado, fragmentos de largo $L_i = 15$ que pueden cubrir cerca del 70 % del espacio estructural con 11 repeticiones, que se corresponden con las hojas-$\beta$ (Fig. 2.10c). A mayores largos no hay fragmentos que muestren un nivel de periodicidad significativa.

La $\beta$-lactamasa de *Bacillus licheniformis* es un muy buen ejemplo de una estructura con topología $\alpha\beta$, compuesta de dos subdominios discontinuos [Santos et al., 2004]. Nuevamente, en este caso, no hay un largo de fragmento particular al que se pueda definir una frecuencia característica útil (Fig. 2.10d). El mejor teselado ocurre a un $L_i = 15$ donde el fragmento corresponde con una de las 10 hélices-$\alpha$ de la estructura, que al repetirse, puede cubrir cerca del 74 %.

## 2.7. Teselado de Oligómeros

La mayoría de las cadenas polipeptídicas de los organismos vivos, cuando están en su entorno natural, no se encuentran plegadas como monómeros esferoidales, sino que nuclean con otras para conformar complejos oligoméricos formados por dos o más subunidades. Más fre-

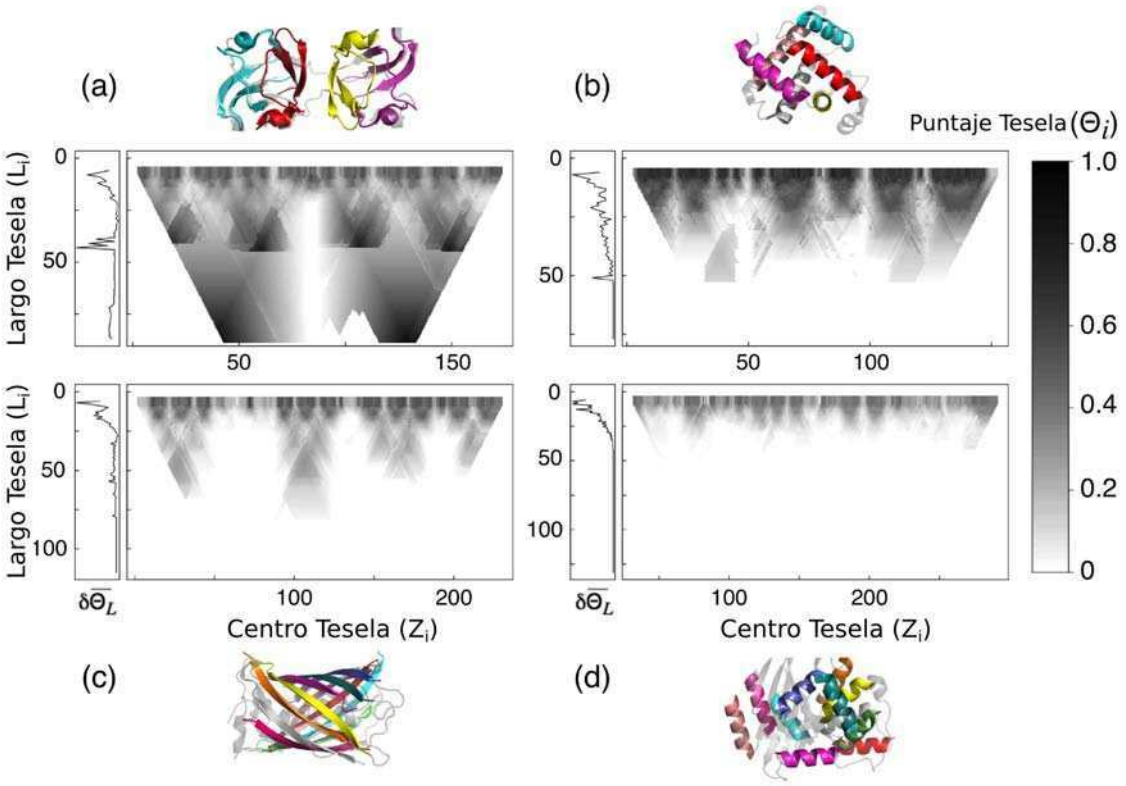


**Figura 2.10:** Teselado de proteínas globulares clásicas. El perfil de teselado se muestra en escala de grises, junto con los valores de $\delta\overline{\Theta_L}$ proyectados en la izquierda. Las estructuras de la proteína nativa y el teselado correspondiente al puntaje máximo a la frecuencia característica se muestran, usando el mismo esquema de colores que en la figura anterior. El largo ($L_i$) y el centro ($Z_i$) de la tesela seleccionada: a) $\beta\gamma$-crystallin (pdb:1h4a,X) $L_i = 43, Z_i = 149{,}5$ b) Myoglobin (pdb:1mbd,A) $L_i = 18, Z_i = 29$ c) Green Fluorescent Protein (pdb:1gfl,A) $L_i = 15, Z_i = 182{,}5$ d) $\beta$-Lactamase(pdb:4blm,A) $L_i = 15, Z_i = 185{,}5$

cuentemente se encuentran formando complejos homodiméricos, aunque los hetero-oligómeros no son poco comunes e incluso complejos formados por cientos o miles de unidades pueden encontrarse. Las bases simétricas de este fenómeno se han explorado desde incluso antes de que se resolvieran las primeras estructuras [Goodsell and Olson, 2000]. Un estudio reciente estima que cerca del 95 % de los complejos homodiméricos cristalizados son simétricos [Swapna et al., 2012], y se espera que pequeñas inserciones y deleciones puedan tener efectos profundos en la funcionalidad de la proteína, modulando la estabilidad oligomérica, la especificidad y la agregación [Hashimoto and Panchenko, 2010].

Para analizar los detalles de simetría en complejos multi-cadena, debemos primero definir los bloques elementales que los constituyen. Para explorar esto, hemos aplicado el mismo procedimiento de fragmentado y teselado descripto anteriormente pero ahora usando arreglos de unidades cuaternarias como blanco para calcular sus patrones de teselado. Si los monómeros

que pueden forman un homo-oligómero no pueden ser descompuestos en teselas con altos puntajes de teselabilidad, esperamos que la del mayor puntaje corresponda con la cadena monomérica. De hecho encontramos que éste es el caso para la mayoría de los oligómeros que hemos evaluado. Notamos, sin embargo, algunos casos interesantes en donde la subunidad puede descomponerse en teselas internas con altos puntajes.

La proteína E2c de unión a ADN de Papillomavirus es un muy buen modelo para estudiar la especificidad para reconocer secuencias [Sánchez et al., 2010]. Este dominio está compuesto por dos cadenas idénticas que forman un barril-$\beta$ que expone 4 hélices-$\alpha$. El procedimiento de teselado identifica un fragmento de 81 residuos como aquel que posee el máximo puntaje, el cual corresponde a la cadena monomérica (Fig. 2.11a). Sin embargo, estas pueden a su vez ser descompuestas en teselas de $L_i = 43$, capaces de cubrir cerca del 90 % del espacio estructural. La mejor tesela a esa frecuencia corresponde con el motivo $\beta\alpha\beta$ que se entrelaza en cada monómero y que con estos contribuye medio barril $\beta$ (Fig. 2.11a).

La hemoglobina es el ejemplo clásico de un arreglo cuaternario simétrico, un tetrámero de cadenas $\alpha_2\beta_2$. La Fig. 2.11b muestra un patrón de teselado regular en el que se pueden distinguir 4 regiones casi idénticas. Esto coincide con la identidad estructural reconocida entre las cadenas $\alpha$ y $\beta$. Como en el caso de la Mioglobina, no se observa una descomposición de la estructura en teselas más pequeñas con valor informativo aparente.

En algunas ocasiones, las estructuras proteicas revelan chances y necesidades de su historia evolutiva. La Fig. 2.11 muestra las estructuras de la subunidad $\beta$ de una DNA polimerasa III de una arquea (un homodímero, Fig. 2.11c), junto con el factor de procesividad de la ADN polimerasa $\delta$ de eucariotas (un homotrímero, Fig. 2.11d). El teselado de estos complejos cuaternarios identifica subunidades e identifica frecuencias características similares en los dos casos, $L_i = 128$ y $L_i = 132$, respectivamente. En ambos casos, las teselas elegidas a sus respectivos $L_i$ cubren cerca del 94 % de la estructura de los complejos. Es aparente que un clamp de ADN de este tipo puede construirse ya sea con 2 o 3 cadenas polipeptídicas, cada una conteniendo 3 o 2 teselas según el caso, que coalescen en una estructura con una simetría rotacional de orden 6 [Sippl and Wiederstein, 2012]. Esta tesela común puede ser luego descompuesta en 2 de largo $L_i = 65$ manteniendo prácticamente la misma capacidad de

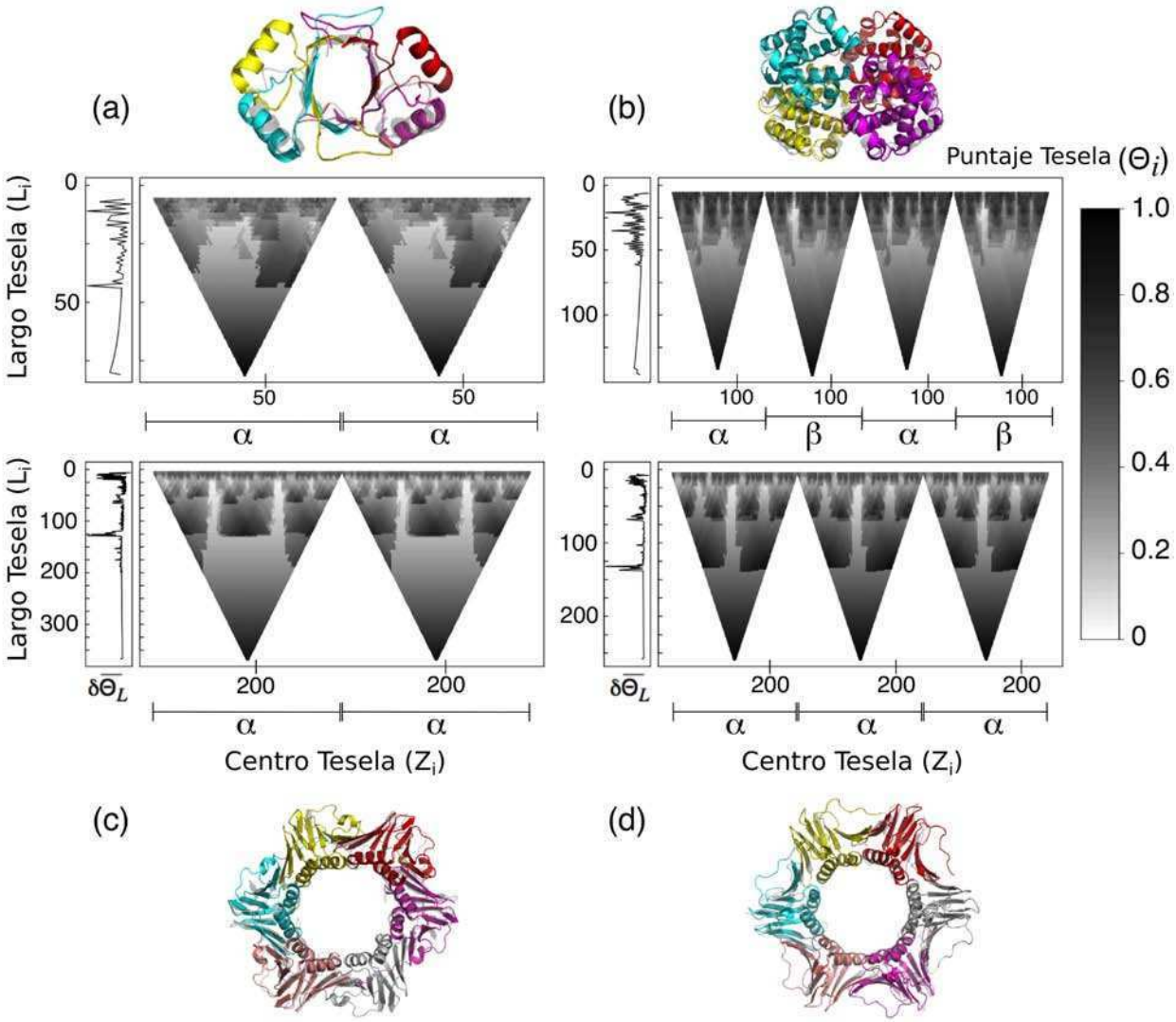


**Figura 2.11:** Teselado de complejos cuaternarios. El perfil de teselado se muestra en escala de grises, junto con los valores de $\delta\overline{\Theta_L}$ proyectados en la izquierda. Las estructuras de la proteína nativa y el teselado correspondiente al puntaje máximo a la frecuencia característica se muestran, usando el mismo esquema de colores que en la figura anterior. El largo ($L_i$) y el centro ($Z_i$) de la tesela seleccionado: a) Homodímero HPV-16 E2c (pdb:1r8p) $L_i = 43, Z_i = 58{,}5$ b) Deoxy-Hemoglobina (pdb:2hhb) $L_i = 141, Z_i = 71{,}5$ c) $\beta$-subunit of *Thermotoga maritima* DNA polymerase III (pdb:1vpk) $L_i = 128, Z_i = 297$ d) Factor de procesividad de la DNA polymerasa-$\delta$ de *Saccharomyces cerevisiae* (pdb:1plq) $L_i = 132, Z_i = 190$

cubrimiento que las teselas más grandes. Es interesante como estos fragmentos se entrelazan para formar los patrones de orden superior con una complementariedad que no se observa en otros fragmentos maximales de la estructura global.

## 2.8. Aportes del método de teselado para mejorar la detección de unidades repetitivas en proteínas

Ya hemos discutido anteriormente que es común que las repeticiones en proteínas repetitivas tengan un bajo porcentaje de identidad al compararlas entre sí. Esta divergencia es particularmente alta en las repeticiones de los extremos en el caso de proteínas del tipo solenoide, ya que estas se encuentran en un entorno con mayor accesibilidad al solvente que aquellas que se encuentran localizadas en medio del arreglo repetitivo. En el caso de las proteínas con simetría rotacional, los extremos N y C terminal, se encuentran en contacto. Esto, por un lado soluciona el problema de las diferencias en cuanto a la accesibilidad al solvente, pero al mismo tiempo, limita la cantidad de repeticiones que pueden contener las proteínas de una misma familia/clase. En el caso de los miembros de la familia WD-40, estas contienen 7 repeticiones por cada dominio, los cuales adoptan una arquitectura del tipo $\beta$-propeller. Usamos la familia WD-40 a fines de analizar el problema de la detección de repeticiones en una familia repetitiva, usando métodos basados en secuencia. Seleccionamos todos los miembros de la familia WD-40 con estructura conocida, mediante detección de unidades del tipo WD-40 en las secuencias de todo el *Protein Data Bank* (PDB). Dicha selección se realizó usando el modelo oculto de Markov para la familia WD-40 (WD40 HMM) presente en Pfam, usando el módulo hmmsearch de la suit HMMER [Eddy, 2001]. Se realizó además una reducción de redundancia del conjunto de datos, seleccionando una estructura por cada identificador de Uniprot presente dentro del conjunto de proteínas recuperadas del paso anterior. En total obtuvimos 85 estructuras de proteínas no redundantes.

Por un lado contamos cuántas unidades WD-40 fuimos capaces de detectar sobre las secuencias de dichas estructuras y por el otro, hicimos una detección visual del número de unidades repetitivas presentes en las estructuras. Esta detección visual, se realizó con 3 cu-

radores de forma independiente y se anotó el consenso (Luna D, Nazzi E, Parra RG (2015), resultados no publicados). Esta metodología es empleada por los curadores de la base de datos RepeatsDB [Di Domenico et al., 2013] tanto para clasificar proteínas repetitivas, contar las unidades repetitivas, como así también definir los límites entre las mismas. La metodología se justifica en la facilidad que tenemos los observadores humanos para generalizar un objeto y reconocer su patrón morfológico en otras estructuras relacionadas [Di Domenico et al., 2013]. Al comparar el número de detecciones en secuencia con el que surge de observar las estructuras, podemos observar que hay una cantidad considerable de repeticiones que no son detectadas en secuencia (Fig. 2.12A).

Al hacer detecciones con el módulo hmmsearch de Hmmer sobre las secuencias, no sólo se obtiene el número de dominios detectados, sino que además el programa informa entre que posiciones relativas al WD40 HMM se produjo la misma. Esto quiere decir que las detecciones pueden ser totales o parciales. Si bien existen casos en que medio dominio proteico puede estar presente en una proteína debido a truncaciones, en proteínas repetitivas la incompletitud en las detecciones se debe mayoritariamente a la alta divergencia en secuencia de las unidades repetitivas. Podemos observar en la Fig. 2.12B, la distribución de comienzos y finales de las detecciones de las unidades WD40 en las secuencias. Es fácil observar que para los comienzos una gran cantidad de detecciones comienzan más allá de la posición 1 respecto del WD40 HMM que representa a la familia. Así mismo podemos observar que los finales de las detecciones en una importante cantidad de casos no llegan hasta el final del WD40 HMM.

Observamos que al teselar dos estructuras con dominios WD-40 se obtienen valores diferentes en los máximos para la función $\delta\overline{\Theta_L}$ (Fig. 2.13). En el caso de la estructura de la subunidad $\beta$ de la proteína transductina (1a0r,B) se observa que los fragmentos que mejor teselan la proteína poseen 42 residuos de largo. Por el otro lado, para la estructura de la proteína de unión a daño de ADN (4e5z,B) observamos que el largo de fragmentos que mejor teselan la estructura es 11. Es claro, que el motivo estructural más importante en la estructura es de un largo mayor que el detectado por el máximo de la función $\delta\overline{\Theta_L}$. Estas diferencias en los largos de las repeticiones maximales tienen que ver principalmente con el grado de simetría presente en una y otra molécula, lo cual se relaciona con el nivel promedio de puntajes globales

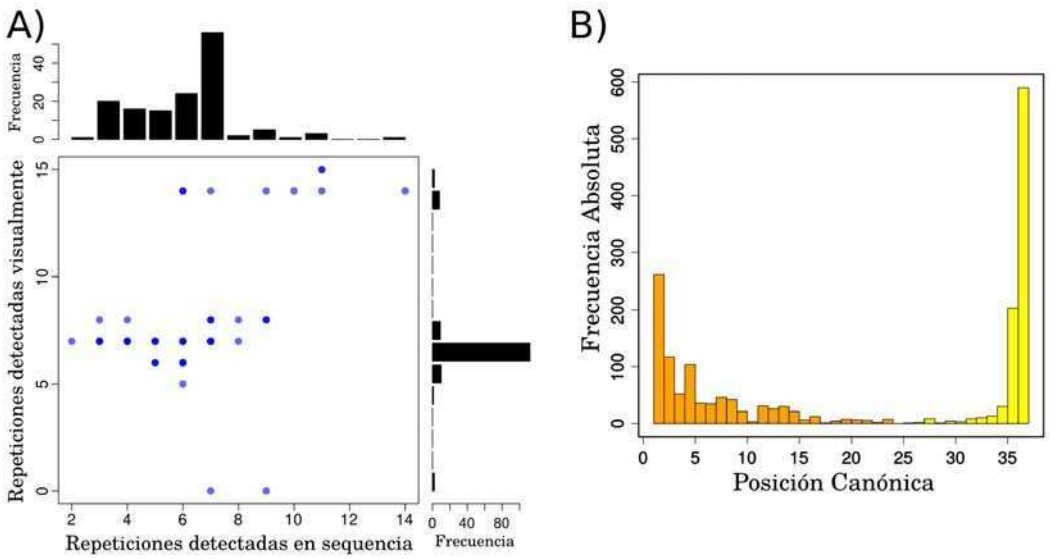


**Figura 2.12:** A) Número de repeticiones WD-40 que pueden ser detectadas de forma computacional mediante HMMER y el HMM para WD-40 presente en Pfam (eje x) y detectadas en forma visual por humanos (eje y). Se observa como para muchos casos, la cantidad de unidades detectadas por HMMER es menor a las detectadas en forma visual. B) Distribuciones de posiciones relativas al HMM de WD-40 en que se detectan los comienzos (naranja) y finales (amarillo) de las unidades WD-40 detectadas por HMMER. Puede observarse que una proporción significativa de repeticiones son detectadas de forma incompleta en su extremo inicial o terminal.

de los diferentes fragmentos en los diagramas de teselado (el nivel de negro es mayor en el patrón de teselado de 1a0r,B que en 4e5z,B como se puede ver en la Fig. 2.13). Debido a esto diseñamos una nueva estrategia para seleccionar las teselas que se correspondan con los motivos estructuralmente repetidos en moléculas que siendo de la misma familia repetitiva, presentan este tipo de discrepancias. Dicha estrategia consiste en primer lugar en encontrar el fragmento maximal no de forma global sino de acuerdo a cuantas copias no superponibles de sí mismo es capaz de superponer sobre la estructura. Así se seleccionan aquellos fragmentos que maximizan el cubrimiento mediante 2, 3, .., n copias. En este caso, usamos un n igual a 14 (que corresponde con aquellas moléculas que poseen 2 dominios WD-40 en la misma cadena).Podemos observar que el mejor cubrimiento se obtiene en ambos casos con fragmentos que pueden ubicar 7 copias de sí mismos en la estructura (Fig. 2.13). Al recuperar dichos fragmentos, encontramos que estos poseen 42 residuos de largo, lo cual se repite para todas las estructuras con dominios WD-40 que hemos analizado. Es interesante que dicho largo detectado para la repetición WD-40 difiere de los 39 residuos que posee el WD40 HMM en Pfam. Al ubicar los hits en la secuencia, podemos observar que los mismos se encuentran localizados con distancias de aproximadamente 2 residuos entre el final de una repetición y el inicio de la

siguiente, lo cual evidencia que el WD40 HMM, está definido de forma incompleta respecto del motivo estructural que podemos encontrar usando nuestra estrategia.

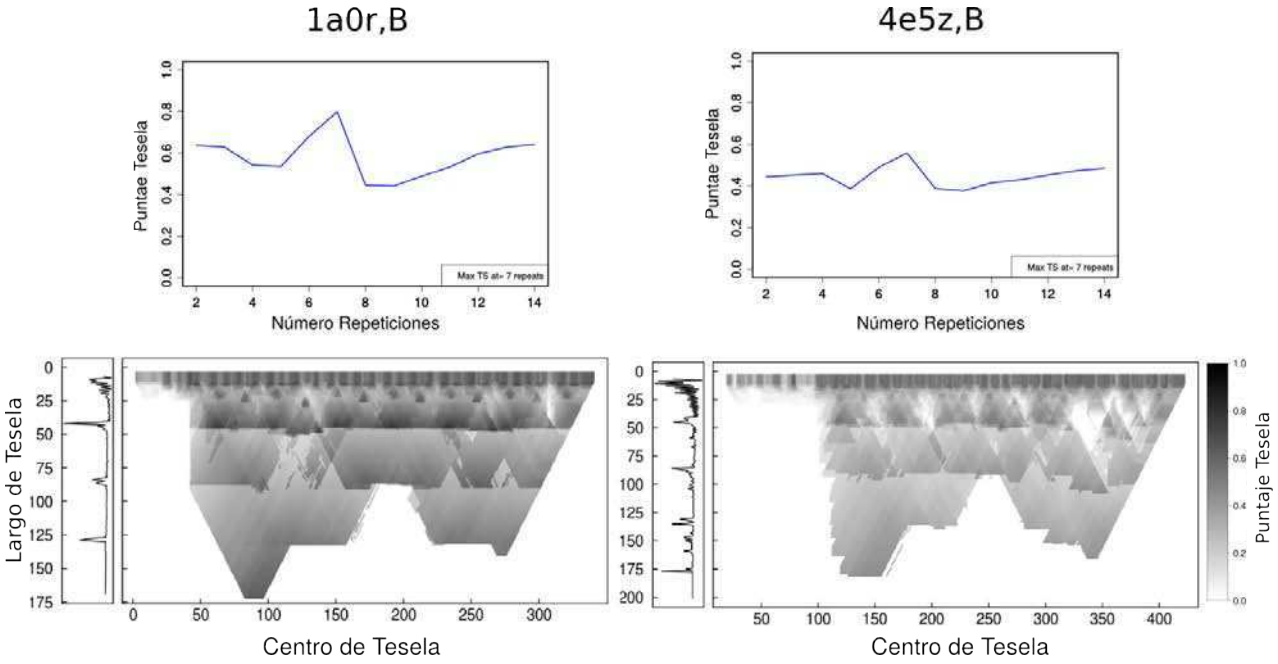


**Figura 2.13:** Teselado de dos estructuras proteicas con dominios del tipo WD-40 (Izquierda: 1a0r,B; Derecha: 4e5z,B). En los paneles superiores se observan los valores del puntaje de teselado para aquellas repeticiones maximales que pueden ubicar 2, 3, .., n veces sobre la estructura. Para ambas proteínas el máximo se encuentra en 7 repeticiones. En los paneles inferiores se observa el patrón global de teselado de las estructuras y en sus laterales los valores de la función $\delta\overline{\Theta_L}$.

Es notable observar que a pesar de la alta divergencia de las repeticiones a nivel de secuencia, el método de teselado es capaz de localizar las repeticiones estructurales tanto en estructuras con un alto grado de regularidad en cuanto al arreglo de las aspas como en aquellas en donde el ordenamiento es más irregular como se observa en la Fig. 2.14.

Es evidente que los dominios repetitivos presentan dificultades específicas para los métodos bioinformáticos clásicos en cuanto a la capacidad de definir donde las unidades repetitivas se encuentran localizadas. Tanto el largo de las repeticiones como la fase en que se encuentran definidas, son parámetros no trivialmente identificables. En el ejemplo de la Fig. 2.14 se puede observar que a pesar de que el método de teselado es capaz de encontrar repeticiones estructurales en los dominios WD-40, lo cuales poseen un largo promedio de 42 residuos, con desviaciones debido a inserciones o deleciones, la fase de los motivos encontrados no es consistente en ambas moléculas. Esto es evidente al localizar el motivo WD típico que se encuentra al final de las repeticiones, mientras en 1a0r,B este motivo se corresponde con los

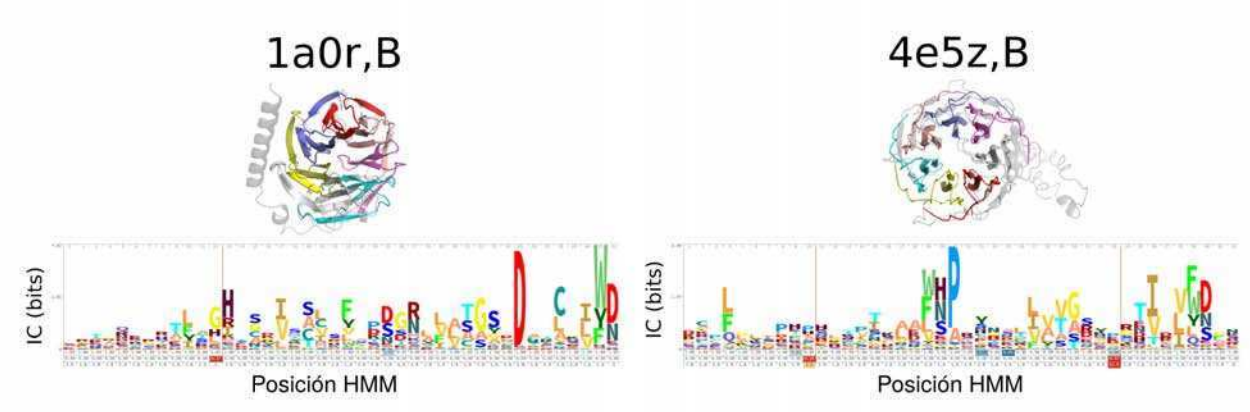


**Figura 2.14:** Teselado de dominios WD-40 (izquierda 1a0r,B, derecha 4e5z,B). Arriba: Estructuras teseladas con aquel fragmento que maximiza el cubrimiento de la estructura total con 7 copias de si mismo. Abajo: Logos de secuencia a partir de las repeticiones detectadas. Los logos de secuencia se generaron usando el skylign (www.skylign.org) a partir de modelos ocultos de Markov generados a partir de los alineamientos de secuencias obtenidos a partir de los alineamientos estructurales mediante el módulo hmmbuild de hmmer. Se observa la menor conservación en los motivos de secuencia en el caso de 4e5z,B respecto de 1a0r,B (altura máxima en bits de cada logo) lo cual es congruente con su menor grado de simetría a nivel estructural.

dos últimos residuos, en 4e5z,B se encuentra desplazado dos residuos hacia el extremo N-terminal. Esto también se observa para los casos analizados en la Tabla 2.1, en donde para miembros de una misma familia, los parámetros de largo y fase no son identificables de forma congruente teniendo en cuenta sólo aspectos estructurales. De todas formas, el proceso de teselado, por su naturaleza exhaustiva, permite que todos los fragmentos que componen una proteína sean evaluados en cuanto a su capacidad de ser verdaderas unidades repetitivas. Dado un conjunto de restricciones en cuanto a la fase requerida, o el largo representativo de las repeticiones, el método de teselado permitiría que eventualmente todos los miembros puedan ser descompuestos en repeticiones de una forma consistente. La anotación consistente de repeticiones a lo largo de miembros de una misma familia repetitiva, es fundamental para realizar estudios comparativos. Actualmente, más allá de todas las falencias descriptas, el único método que permite realizar anotaciones consistentes es HMMER, mediante el uso de HMMs representativos del motivo repetitivo. El resto de los algoritmos disponibles generan anotaciones de las repeticiones en donde la fase y largo de las unidades son dependientes de la secuencia que es analizada y no es posible imponer restricciones para los mismos. En el próximo capítulo mostraremos cómo usamos el método de teselado en conjunto con otras estrategias para generar por primera vez una anotación consistente de los miembros de una familia de proteínas repetitivas.

## 2.9. Conclusiones del capítulo

Es más fácil encontrar secuencias plegables con paisajes energéticos con forma de embudo si la estructura del estado nativo es simétrica [Wolynes, 1996]. Muchos investigadores aprecian la existencia de simetría como una característica emergente de la parsimonia de la naturaleza, que resulta de las formas de interacción limitadas entre un pequeño número de partes elementales que se ensamblan en estructuras de órdenes superiores [Goodsell and Olson, 2000, Wolynes, 1988, Wales, 2012, Denton et al., 2002]. Las simetrías inexactas presentes en las moléculas biológicas son algo aún más llamativo [Wolynes, 1996, Wolynes, 1988]. Sutiles aperiodicidades pueden dar lugar a grandes efectos biológicos [Schrödinger, 1944] importantes para la regulación fisiológica de los sistemas en estudio. Para poder detectar y caracterizar repeticiones en estructuras proteicas, hemos presentado un simple esquema basado en el análisis de la distribución de alineamientos estructurales subóptimos de fragmentos continuos. El procedimiento desarrollado identifica fragmentos maximales, que son aquellos que maximizan una función objetivo, pero que al ser extendidos rápidamente pierden dicha capacidad (Fig. 2.3B) al disminuir su número de ocurrencias en el ensamble de soluciones. Contando el número de ocurrencias de fragmentos no superponibles y teniendo una buena métrica para calcular el cubrimiento total de la estructura, definimos un puntaje de teselado, $\Theta$, que ordena a los diferentes fragmentos (no idénticos) de acuerdo a su capacidad para teselar la estructura global mediante copias de sí mismos. Encontramos que en la mayoría de los casos hay un largo particular de fragmento al cual el cubrimiento ganado por repeticiones es máximo, lo cual definimos como frecuencia característica. En algunos casos hay una colección discreta de fragmentos que permiten definir la mejor fase de forma unívoca. En esos casos la unidad repetitiva, el número de ocurrencias y sus límites pueden ser definidos de forma confiable (Tabla 2.1). En otros casos, hay muchas fases equivalentes a la frecuencia característica, señalando que esas estructuras pueden ser consideradas como casi periódicas donde la definición de una tesela básica puede permanecer arbitraria (Fig. 2.3, Tabla 2.1). Esto es común en el caso de las proteínas solenoides donde diferentes investigadores han definido la unidad repetitiva con diferentes fases y períodos [Schaper et al., 2012]. Incluir otra información más allá de la geometría estructural podría indicar si existe una fase *biológicamente* preferida. Esta información puede provenir de

diferentes aspectos como por ejemplo, la caracterización de sitios de inserción, la variabilidad en secuencias ortólogas, límites entre exones o mecanismos de plegado [Schafer et al., 2012]. Explicaremos una alternativa a este problema en el próximo capítulo.

Aquellas proteínas donde las repeticiones se empaquetan de forma simétrica entre sí pero no se trasladan a lo largo de un eje, pueden formar estructuras cerradas. El proceso de fragmentado y teselado puede ser aplicado sobre topologías con simetría rotacional como barriles, propellers y trefoils. Se pueden encontrar en dichas estructuras, diversos tipos de unidades repetitivas, organizadas jerárquicamente e incluso observar diferencias finas entre las mismas (Fig. 2.8, Tabla 2.1).

Si las teselas fundamentales están organizadas de forma simétrica, entonces tiene que haber teselas más grandes que agrupen a las primeras. Estas teselas de orden superior aparecen como picos máximos adicionales en la función $\Theta_{L_i}$ hacia largos $L_i$ mayores en comparación con las teselas básicas. Este anidado jerárquico de teselas puede ser capturado mediante un puntaje de teselación que se computa de la siguiente forma: Para cada largo $L_i$ de tesela se toma el máximo puntaje $\Theta_i$ y se calcula el promedio de los mismos sobre todo $L$. Este puntaje de *teselabilidad* ($\Xi$) vale 1.0 para el modelo homogéneo, se acerca a 1 para estructuras muy regulares como en el caso de hélices $\alpha$ y decae a valores cercanos a cero para estructuras no repetitivas. En la Fig. 2.15 se muestra una variedad de estructuras de proteínas ordenadas por sus respectivos valores de teselabilidad $\Xi$ (Tabla 2.1). El valor más grande de $\Xi$ se obtiene para una hélice $\alpha$ perteneciente a una estructura del tipo *coiled coil*. Cerca del valor máximo de la hélice se encuentran varias proteínas solenoidales. Aquellas proteínas más regulares dentro del grupo de las proteínas diseñadas se encuentran más arriba en el ranking que aquellas con mayor proporción de irregularidades estructurales. Estas estructuras son seguidas por aquellas proteínas repetitivas con una forma semejante a las globulares. Al final de esta escala encontramos aquellas proteínas que contienen dominios globulares típicos que no poseen periodicidades más allá de unos pocos residuos. Notamos que los miembros pertenecientes a una misma topología no se agrupan juntos en esta escala, sino que segregan de acuerdo al grado de irregularidades que presentan. Esto es una muestra de que el proceso de teselado no es afectado por el tipo de simetría presente en la estructura (Fig. 2.15), sino solo por la existencia de arreglos simétricos.

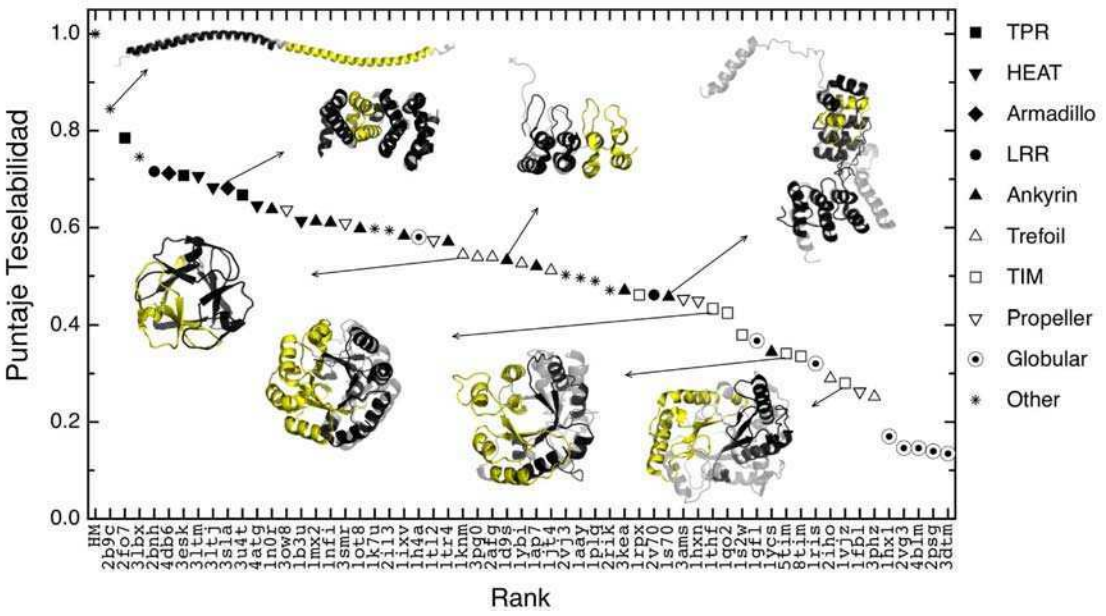


**Figura 2.15:** Teselabilidad de estructuras de proteínas. El procedimiento de teselado se aplicó a diversos representantes de proteínas repetitivas y globulares como también al modelo homogéneo y se ordenaron de acuerdo a su valor de teselabilidad $\Xi$. Ejemplos de las teselaciones se muestran con la tesela seleccionada coloreado en amarillo y las copias en color negro, superimpuestas sobre la estructura nativa en color gris. Símbolos llenos: Proteínas repetitivas solenoides. Símbolos vacíos: Proteínas repetitivas con topología cerrada.

El mismo procedimiento de teselado se puede aplicar al nivel de complejos proteicos, analizando los detalles de cómo copias de fragmentos entre cadenas cubren el espacio estructural. A este nivel, encontramos que las mejores teselas corresponden siempre con las cadenas monoméricas o con los clásicos dominios globulares dentro de ellas. Sin embargo, hay algunas excepciones interesantes en donde descomposiciones sucesivas se pueden generar a partir de las cadenas monoméricas, en teselas más pequeñas, que mantienen gran parte de la propiedad de teselabilidad global (Fig. 2.15). Algo interesante de testear en el futuro es como estas teselas geométricamente definidas, coinciden con las cadenas polipeptídicas, los dominios globulares, los límites entre exones, los foldones o diversos motivos estructurales.

Es tentador especular acerca de las consecuencias funcionales que la distribución simétrica de fragmentos similares puede tener a diferentes escalas. La teoría de paisajes energéticos sostiene que la utilización de subunidades organizadas de forma simétrica en formas similares, da lugar a estructuras con energías libres similares, permitiendo que coexistan múltiples embudos de plegado en el paisaje energético total [Levy et al., 2005], necesitando sólo pequeñas perturbaciones para producir un intercambio entre las mismas [Hegler et al., 2008].

Se ha visto que la presencia de simetría es clave en varios fenómenos como por ejemplo

la cooperatividad del plegado, la unión de diversos ligandos, la estabilidad termodinámica, entre otros [Wales, 2012, Goodsell and Olson, 2000]. La organización simétrica de un sistema es una forma fácil y quizás infranqueable, para que aparezca el alosterismo [Monod et al., 1965, Kuriyan and Eisenberg, 2007]. La ocurrencia de repeticiones con simetría puntual, da lugar a arreglos cerrados como barriles al nivel terciario y anillos al nivel cuaternario. La simetría helicoidal da lugar a solenoides al nivel terciario y corresponde con organizaciones de tipo tubular al nivel cuaternario. La nucleación y taponado (*capping*) de esos arreglos repetitivos es en general crítico para su comportamiento fisiológico tanto al nivel terciario como cuaternario. No es de sorprender entonces que el funcionamiento fisiológico y los estados patológicos sean resultado de la agregación de fragmentos similares como por ejemplo, en la dinámica del citoesqueleto [Wang and Wolynes, 2011], fenómenos epigenéticos [Jablonka and Raz, 2009], anemia falciforme [Pauling and Itano, 1949] y procesos relacionados con amiloides [Treusch et al., 2009].

La organización de las moléculas proteicas puede ser apreciada a múltiples niveles, desde los motivos de secuencia hasta la dinámica de interacción de miles de componentes [Wang and Wolynes, 2011]. Debido a que las contribuciones relevantes de las fuerzas físicas cambian a las diferentes escalas de tamaño y tiempo, las restricciones organizacionales en cada nivel también requieren cambiar, aunque algunos principios comunes podrían permanecer subyacentes. Creemos que los conceptos postulados por la teoría de paisajes energéticos puede servir de guía en dicha búsqueda [Frauenfelder et al., 1991, Frauenfelder, 2002, Zhuravlev and Papoian, 2010] y es a lo que apunta el resto del trabajo de esta tesis.

**Tabla 2.1:** Conjunto de Estructuras Teseladas

| **Proteína** | | | | **Tesela Seleccionada** ($T_i$) | | | **Teselado** | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **PdbID** | **Arquitectura** | $N^a$ | $\Xi^b$ | $L_i^c$ | $Z_i^d$ | $\Theta_i^e$ | $n_{T_i}^f$ | $C_i^g$ | $I_i^h$ | $NR_i^1$ |
| HM | toy model | 120 | 1 | 79* | 39.5* | 1 | 2 | 1 | 0 | 0 |
| 2b9c,A | coiled-coil | 136 | 0.8449 | 63* | 134.5* | 0.84 | 2 | 0.93 | 0 | 0.07 |
| 2fo7,A | TPR | 136 | 0.7851 | 34* | 52* | 1 | 2 | 1 | 0 | 0 |
| 3lbx,A | Spectrin | 140 | 0.7459 | 6 | 64 | 0.82 | 20 | 0.86 | 0.11 | 0.03 |
| 2bnh,A | Leucine | 456 | 0.7159 | 57* | 139.5* | 0.96 | 8 | 0.99 | 0 | 0.01 |
| 2j8k,A | $\beta$-Solenoid | 175 | 0.715 | 10* | 87* | 0.88 | 17 | 0.93 | 0.03 | 0.04 |
| 4db6,A | Armadillo | 197 | 0.7122 | 42* | 33* | 0.9 | 5 | 1 | 0 | 0 |
| 3esk,A | TPR | 128 | 0.7079 | 41* | 286.5* | 0.75 | 3 | 0.84 | 0 | 0.16 |
| 3ltm,A | Heat | 185 | 0.7063 | 31 | 62.5 | 0.97 | 6 | 1 | 0 | 0 |
| 3ltj,A | Heat | 191 | 0.683 | 31* | 94.5* | 0.93 | 6 | 0.97 | 0 | 0.03 |
| 1n11,A | Ank | 408 | 0.6817 | 33* | 510.5* | 0.9 | 12 | 0.95 | 0 | 0.05 |
| 3sla,A | Armadillo | 166 | 0.6817 | 42 | 204 | 0.82 | 4 | 0.98 | 0.02 | 0 |
| 3u4t,A | TPR | 258 | 0.6673 | 34 | 470 | 0.85 | 7 | 0.98 | 0 | 0.02 |
| 2xtw,A | $\beta$-Solenoid | 210 | 0.6639 | 10* | 32* | 0.82 | 19 | 0.88 | 0.05 | 0.07 |
| 1plq@1 | Quaternary | 774 | 0.6515 | 132* | 190,A*+ | 0.92 | 6 | 0.95 | 0.0129 | 0 |
| 4atg,A | Heat | 195 | 0.6457 | 87* | 277.5* | 0.41 | 2 | 0.83 | 0 | 0.17 |
| 1n0r,A | Ank | 126 | 0.6379 | 33* | 18.5* | 0.87 | 4 | 0.99 | 0 | 0.01 |
| 3ow8,A | $\beta$-propeller | 300 | 0.6368 | 42 | 200 | 0.92 | 7 | 0.99 | 0.01 | 0 |
| 1b3u,A | Heat | 588 | 0.6138 | 39 | 530.5 | 0.8 | 15 | 0.98 | 0.02 | 0 |
| 1ihb,A | Ank | 156 | 0.6137 | 33* | 79.5* | 0.86 | 5 | 0.99 | 0.01 | 0 |
| 1mx2,A | Ank | 156 | 0.6128 | 33* | 79.5* | 0.86 | 5 | 0.99 | 0.01 | 0 |
| 1k1a,A | Ank | 228 | 0.6126 | 33 | 241.5 | 0.86 | 7 | 0.96 | 0.04 | 0 |
| 1nfi,E | Ank | 213 | 0.6104 | 33 | 128.5 | 0.82 | 6 | 0.97 | 0 | 0.03 |
| 3smr,A | $\beta$-propeller | 304 | 0.6089 | 42 | 138 | 0.92 | 7 | 0.99 | 0 | 0.01 |
| 1awc,B | Ank | 153 | 0.6088 | 33* | 54.5* | 0.86 | 5 | 0.99 | 0 | 0.01 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 2rfm,A | Ank | 183 | 0.604 | 33* | 138.5* | 0.85 | 5 | 0.9 | 0 | 0.1 |
| 1ot8,A | Ank | 209 | 0.5979 | 33* | 91.5* | 0.84 | 6 | 0.93 | 0 | 0.06 |
| 1k7u,A | Hevein | 171 | 0.5976 | 43 | 150.5 | 0.98 | 4 | 1 | 0 | 0 |
| 2i13,A | Zn-Finger | 154 | 0.5943 | 28* | 75* | 0.87 | 5 | 0.91 | 0 | 0.09 |
| 1r8p@0 | Quaternary | 162 | 0.5869 | 43* | 58.5,A$^{*+}$ | 0.78 | 4 | 0.9 | 0.09 | 0 |
| 1vpk@1 | Quaternary | 734 | 0.5856 | 128* | 297,A$^{*+}$ | 0.78 | 6 | 0.93 | 0.065 | 0 |
| 1ixv,A | Ank | 229 | 0.5834 | 34 | 190 | 0.84 | 7 | 1 | 0 | 0 |
| 1blx,B | Ank | 160 | 0.5818 | 65* | 64.5* | 0.61 | 2 | 0.81 | 0 | 0.19 |
| 1h4a,X | $\beta\gamma$-crystallin | 174 | 0.5805 | 43 | 149.5 | 0.85 | 4 | 0.95 | 0.03 | 0.02 |
| 2hhb@1 | Quaternary | 574 | 0.5768 | 21* | 66.5,A$^{*+}$ | 0.76 | 24 | 0.84 | 0.11 | 0.038 |
| 1tl2,A | $\beta$-propeller | 235 | 0.5740 | 47* | 119.5* | 0.98 | 5 | 1 | 0 | 0 |
| 1tr4,A | Ank | 226 | 0.5708 | 33 | 55.5 | 0.85 | 7 | 0.98 | 0 | 0.02 |
| 1bu9,A | Ank | 168 | 0.5602 | 33* | 82.5* | 0.83 | 5 | 0.98 | 0 | 0.02 |
| 1knm,A | Trefoil | 129 | 0.5439 | 40* | 27* | 0.84 | 3 | 0.91 | 0.05 | 0.05 |
| 3pg0,A | Trefoil | 140 | 0.5392 | 47* | 25.5* | 0.99 | 3 | 1 | 0 | 0 |
| 2afg,A | Trefoil | 129 | 0.5392 | 41 | 30.5 | 0.84 | 3 | 0.98 | 0.02 | 0 |
| 1d9s,A | Ank | 130 | 0.5327 | 66 | 98 | 0.59 | 2 | 0.99 | 0.01 | 0 |
| 1mbd,A | Globin | 153 | 0.5269 | 7 | 12.5 | 0.79 | 19 | 0.86 | 0.11 | 0.02 |
| 1ybi,A | Trefoil | 284 | 0.5261 | 139 | 79.5 | 0.83 | 2 | 1 | 0 | 0 |
| 1ikn,D | Ank | 220 | 0.524 | 106 | 126 | 0.46 | 2 | 0.95 | 0.01 | 0.04 |
| 1ap7,A | Ank | 168 | 0.52 | 65* | 67.5* | 0.5 | 2 | 0.77 | 0 | 0.23 |
| 1jt4,A | Trefoil | 137 | 0.5117 | 41 | 30.5 | 0.75 | 3 | 0.92 | 0.01 | 0.07 |
| 2vj3,A | EGF-like | 120 | 0.5022 | 38* | 471* | 0.9 | 3 | 0.97 | 0 | 0.03 |
| 2pnn,A | Ank | 248 | 0.4737 | 6 | 129 | 0.63 | 31 | 0.75 | 0.25 | 0 |
| 1s4u,X | $\beta$-propeller | 390 | 0.4716 | 171* | 125.5* | 0.5 | 2 | 0.9 | 0 | 0.1 |
| 2rik,A | Ig-like | 280 | 0.4705 | 94 | 140 | 0.94 | 3 | 1 | 0 | 0 |
| 3kea,B | Ank | 282 | 0.4702 | 99* | 165.5* | 0.31 | 2 | 0.7 | 0 | 0.3 |
| 1rpx,A | TIM | 230 | 0.4609 | 109 | 171.5 | 0.58 | 2 | 0.94 | 0 | 0.06 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 2v70,A | Leucine | 210 | 0.4609 | 24 | 577 | 0.59 | 6 | 0.68 | 0 | 0.32 |
| 3jxi,A | Ank | 253 | 0.4607 | 87 | 249.5 | 0.42 | 3 | 0.97 | 0.02 | 0.01 |
| 1s70,B | Ank | 291 | 0.457 | 33 | 86.5 | 0.59 | 7 | 0.77 | 0.1 | 0.13 |
| 3ams,A | $\beta$-propeller | 352 | 0.4528 | 167* | 244.5* | 0.39 | 2 | 0.92 | 0 | 0.08 |
| 1hxn,A | $\beta$-propeller | 210 | 0.4487 | 97* | 379.5* | 0.53 | 2 | 0.87 | 0.06 | 0.08 |
| 1fq0,A | TIM | 213 | 0.4477 | 88 | 60 | 0.5 | 2 | 0.89 | 0.03 | 0.08 |
| 1thf,D | TIM | 253 | 0.4331 | 121* | 179.5* | 0.61 | 2 | 0.94 | 0 | 0.06 |
| 2a0n,A | TIM | 251 | 0.429 | 121* | 181.5* | 0.65 | 2 | 0.96 | 0 | 0.04 |
| 1qo2,A | TIM | 241 | 0.4243 | 119 | 60.5 | 0.68 | 2 | 1 | 0 | 0 |
| 2aja,A | Ank | 347 | 0.4 | 11 | 226.5 | 0.74 | 28 | 0.88 | 0.11 | 0.01 |
| 1dx5,K | EGF-like | 118 | 0.3936 | 9 | 456.5 | 0.55 | 10 | 0.75 | 0.23 | 0.02 |
| 3ehq,A | Ank | 182 | 0.3845 | 33 | 156.5 | 0.46 | 4 | 0.66 | 0 | 0.34 |
| 1s2w,A | TIM | 275 | 0.3797 | 124* | 66* | 0.24 | 2 | 0.87 | 0.04 | 0.1 |
| 2a62,A | Cadherin | 322 | 0.3793 | 109 | 167.5 | 0.7 | 3 | 0.98 | 0 | 0.02 |
| 1gfl,A | GFP | 230 | 0.3676 | 7 | 92.5 | 0.65 | 26 | 0.78 | 0.21 | 0.01 |
| 1ycs,B | Ank | 193 | 0.3443 | 33 | 400.5 | 0.44 | 4 | 0.65 | 0 | 0.35 |
| 5tim,A | TIM | 249 | 0.3412 | 9* | 151.5* | 0.53 | 19 | 0.69 | 0.24 | 0.07 |
| 8tim,A | TIM | 247 | 0.3356 | 9 | 151.5 | 0.53 | 18 | 0.66 | 0.29 | 0.06 |
| 1dcq,A | Ank | 276 | 0.2941 | 7 | 314.5 | 0.54 | 29 | 0.74 | 0.24 | 0.03 |
| 2iho,A | Trefoil | 292 | 0.2901 | 51* | 82.5* | 0.37 | 3 | 0.52 | 0 | 0.47 |
| 1vjz,A | TIM | 325 | 0.28 | 7 | 105.5 | 0.48 | 34 | 0.72 | 0.28 | 0 |
| 1nar,A | TIM | 289 | 0.2755 | 7 | 153.5 | 0.5 | 31 | 0.75 | 0.21 | 0.03 |
| 1fbl,A | $\beta$-propeller | 367 | 0.2625 | 8 | 391 | 0.56 | 32 | 0.7 | 0.28 | 0.02 |
| 3phz,A | Trefoil | 285 | 0.2522 | 9 | 62.5 | 0.52 | 20 | 0.62 | 0.38 | 0 |
| 1sw6,A | Ank | 301 | 0.2249 | 78* | 380* | 0.17 | 3 | 0.77 | 0.14 | 0.09 |
| 1hx1,A | Globular | 377 | 0.17 | 7 | 65.5 | 0.48 | 41 | 0.76 | 0.24 | 0.01 |
| 2vg3,A | Globular | 284 | 0.146 | 7 | 182.5 | 0.52 | 30 | 0.72 | 0.28 | 0 |
| 4blm,A | Globular | 261 | 0.1459 | 7 | 84.5 | 0.52 | 30 | 0.58 | 0.42 | 0 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 2psg,A | Globular | 326 | 0.1396 | 10* | 268* | 0.54 | 25 | 0.76 | 0.24 | 0 |
| 3dtm,A | Globular | 263 | 0.1345 | 7 | 73.5 | 0.52 | 26 | 0.69 | 0.31 | 0 |

[a]Largo de la Proteína. [b]Puntaje de Teselabilidad. [c]Largo de Tesela. [d]Centro de Tesela. [e]Puntaje de Tesela.

[f]Número de copias de la tesela. [g]Fracción de cubrimiento por parte de las copias de la tesela.

[h] Fracción de cubrimiento por inserciones entre las teselas. [i]Fracción de cubrimiento por parte de regiones no repetitivas.

(*) Proteínas para las cuales hay más de una tesela con un valor idéntico de $max\Theta_i$ a la frecuencia característica.

($^+$) Complejos proteicos, la letra después de $Z_i$ indica el identificador de cadena

# Capítulo 3

# Análisis Estructural y Energético de la Familia ANK

En el capítulo anterior hemos presentado el método de teselado de estructuras de proteínas que permite analizar la periodicidad de las mismas y encontrar dónde se localizan los elementos estructuralmente repetidos. El método de teselado es un método exhaustivo que evalúa que tan bueno es cualquier fragmento de la estructura global para cubrir la totalidad de la misma, mediante copias de sí mismo. Una vez evaluados todos los fragmentos, hemos definido estrategias que permiten seleccionar aquel que maximiza el cubrimiento. Por un lado, definimos una función que nos permite encontrar la frecuencia característica de la estructura, que en muchos casos corresponde con el largo reportado para la unidad repetitiva de la misma. En muchos casos, perturbaciones estructurales en las repeticiones producen que la frecuencia característica se encuentre corrida o en múltiplos del largo de la unidad repetitiva. Por ello, también puede hacerse la búsqueda de fragmentos que maximicen el cubrimiento dado un número de copias fijo, es decir, cual es el mejor fragmento que maximiza el cubrimiento con 2, 3, 4, ..., n copias de sí mismo. De esa forma pudimos analizar el caso particular de los TIM Barrels y evaluar sus simetrías del tipo doble, cuádruples u óctuples ya que no está claro aún cuál es la unidad repetitiva básica en esta arquitectura, lo cual tiene relación con la historia evolutiva de como la misma emergió como tal. Sea cual sea el caso, el método de teselado tiene la potencialidad para encontrar repeticiones estructurales pero lo primordial es poder definir a las mismas de una forma consistente para poder realizar estudios comparativos. Es necesario

establecer un protocolo que permita encontrar de forma consistente los parámetros básicos de las unidades repetitivas. Si hacemos una analogía del patrón repetitivo de una estructura con una señal periódica, podemos usar la siguiente nomenclatura para los parámetros de la repetición: Período para el largo de la unidad repetitiva y fase para definir el principio y el final de cada ciclo de la onda. Las repeticiones encontradas para estructuras de la misma familia repetitiva en la Tabla 2.1, en muchos casos se encuentran definidas en diferentes períodos o fases.

En este capítulo vamos a describir una estrategia para definir los parámetros de las repeticiones de forma consistente. Vamos a usar la familia de repeticiones de Ankirina como modelo de aplicación de nuestra metodología. Una vez definida la ubicación de las unidades estructuralmente repetidas, analizaremos los patrones de conservación tanto en secuencia como en la estructura de las mismas. Analizaremos los focos de inestabilidad/estabilidad energética y los relacionaremos con los patrones en las secuencias y analizaremos las implicancias funcionales de los mismos.

## 3.1. Análisis generales de las estructuras ANK

Existen 169 estructuras dentro del Protein Data Bank (PDB) en las cuales al menos una de las cadenas corresponde a una proteína que contiene el motivo estructural ANK. Al mapear las estructuras a sus correspondientes identificadores de Uniprot, podemos observar que éstas corresponden a un total de 54 proteínas diferentes entre las que se encuentran 44 proteínas naturales y 19 proteínas diseñadas (estos números fueron actualizados por última vez en Diciembre de 2015).

Las estructuras de ANKs en general, contienen entre 3 y 12 repeticiones (Tabla 3.1, todas las tablas se encuentran al final del capítulo) y una estructura global compatible con la del tipo $\alpha$ solenoide según la clasificación de Kajava [Kajava, 2012]. Si bien la mayoría de las estructuras están compuestas principalmente por repeticiones, algunas de ellas tienen además una región globular presente en la misma cadena.

Cuando se realizan análisis comparativos que involucran proteínas globulares, una gran ventaja es que en general, el largo promedio del dominio (o de los dominios) analizados

está altamente conservado, salvo por ocurrencia de inserciones/deleciones en algunos casos particulares. Las proteínas repetitivas presentan una complejidad adicional, debido a que no sólo proteínas de la misma familia son homólogas entre sí, pero además poseen una homología serial interna entre sus repeticiones. Esto no sería una complicación si no fuera porque la cantidad de repeticiones en una proteína es variable. Esta variabilidad en el largo del dominio repetitivo constituye un obstáculo para la comparación de estas moléculas, debido a que cuando se alinean dos proteínas con diferente número de repeticiones, no es claro qué regiones en ambas proteínas son las que se deben superponer, pudiendo haber una multiplicidad de alineamientos posibles válidos. Para clarificar esto, supongamos que tenemos dos proteínas de 4 y 6 repeticiones respectivamente, éstas pueden ser alineadas de 3 formas equivalentes (Fig. 3.1). Si bien podría elegirse aquel alineamiento que maximiza el porcentaje de identidad entre las moléculas, dado que un alineamiento se basa en la hipótesis de ancestro común, no sería correcto desechar los alineamientos subóptimos.

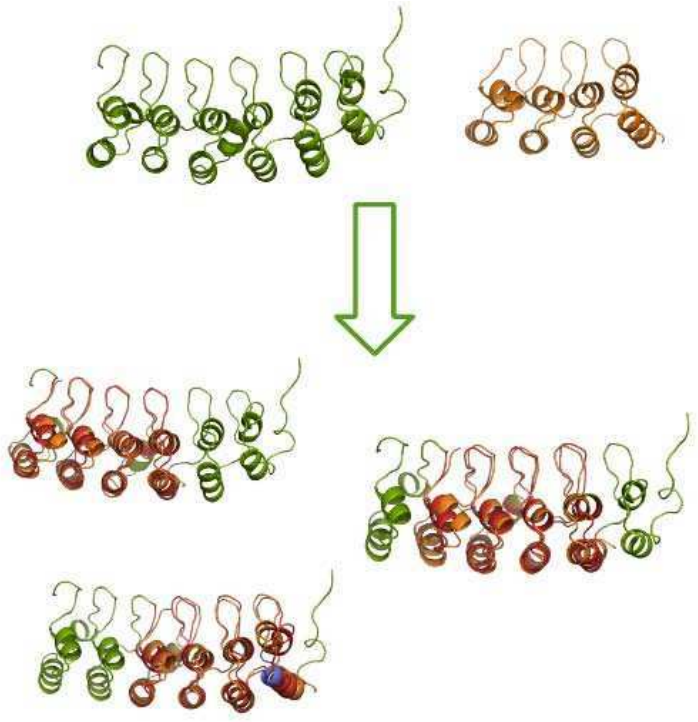

**Figura 3.1:** Esquematización de la multiplicidad de formas en que dos proteínas repetitivas con diferente cantidad de repeticiones pueden ser alineadas. En este ejemplo se muestran diferentes posibles alineamientos entre las proteínas I$\kappa$B-$\beta$ (1k3z,A) en verde y la proteína diseñada 4ANK (1n0r,A) en naranja. Las estructuras fueron alineadas usando el programa TopMatch.

Cuando se las observa de forma global, las ANKs poseen estructuras muy similares, sin embargo, sutiles diferencias al nivel de cada repetición individual pueden ser vistas. Estas

pequeñas perturbaciones son las que producen cambios en el ángulo de giro de la estructura súper helicoidal en ANKs, lo cual produce ciertas desviaciones de la forma solenoidal que usualmente se les atribuye. Si bien ya hemos puntualizado en el problema de alinear ANKs con diferente número de repeticiones, un primer acercamiento a la caracterización de las semejanzas y diferencias en sus estructuras es alinearlas de a pares de forma global y observar que tan parecidas son cuando se analiza el alineamiento que maximiza la superposición estructural. Para ello, usamos la herramienta de alineamientos estructurales TopMatch ( [Sippl and Wiederstein, 2008]), la cual implementa una métrica, $S$, para medir la similitud estructural entre dos moléculas. Cuando se observa el valor de $S$ entre dos estructuras alineadas, el valor entero de la función equivale al número de residuos que pueden ser correctamente alineados según la definición de la métrica. Se realizaron alineamientos de a pares de todas las estructuras del conjunto no redundante de ANKs y el alineamiento que maximiza el puntaje $S$ fue guardado para los análisis subsiguientes. Adicionalmente, a partir de un alineamiento estructural dado, el número de residuos alineados que además son idénticos en secuencia, es representado por el valor $I$. Los valores de los parámetros $S$ e $I$ fueron normalizados para tener en cuenta los largos diferentes de las moléculas comparadas. Si $la$ y $lb$ son los largos de las dos moléculas alineadas, sus contrapartes normalizadas son definidas como $relS = 2*Sab/(la+lb)$ (de acuerdo a su definición en [Sippl, 2008]) y $relI = 2*Sab*(I/100)/(la+lb)$ ambas definidas en el intervalo [0,1]. Podemos observar que los valores de $relI$ (media: 0.22 y sd=0.16) son bastante menores que los correspondientes valores de $relS$ (media: 0.7 y sd=0.17) para las moléculas alineadas (Fig. 3.2) lo cual es de esperar ya que en proteínas la estructura se conserva en mayor medida que la secuencia.

Los valores máximos de $relS$ y $relI$ son obtenidos al alinear proteínas diseñadas con valores medios de 0.85 (sd=0.1) y 0.6 (sd=0.17) respectivamente. En contraste, los alineamientos entre proteínas naturales tienen un rango mayor de variabilidad estructural con una distribución de $relS$ centrada en 0.55 (sd=0.16) mientras que $relI$ muestra valores un tanto menores con una distribución centrada en 0.17 (sd=0.14). Esta diferencia entre las distribuciones del $relI$ y $relS$ concuerda con la alta divergencia en secuencia que es descripta para ANKs en la bibliografía. Los alineamientos de a pares muestran que a pesar de tener una variabilidad en secuencia considerable, estructuras con valores de $relI$ tan bajos como 0.2 pueden tener valores de $relS$

mayores a 0.8.

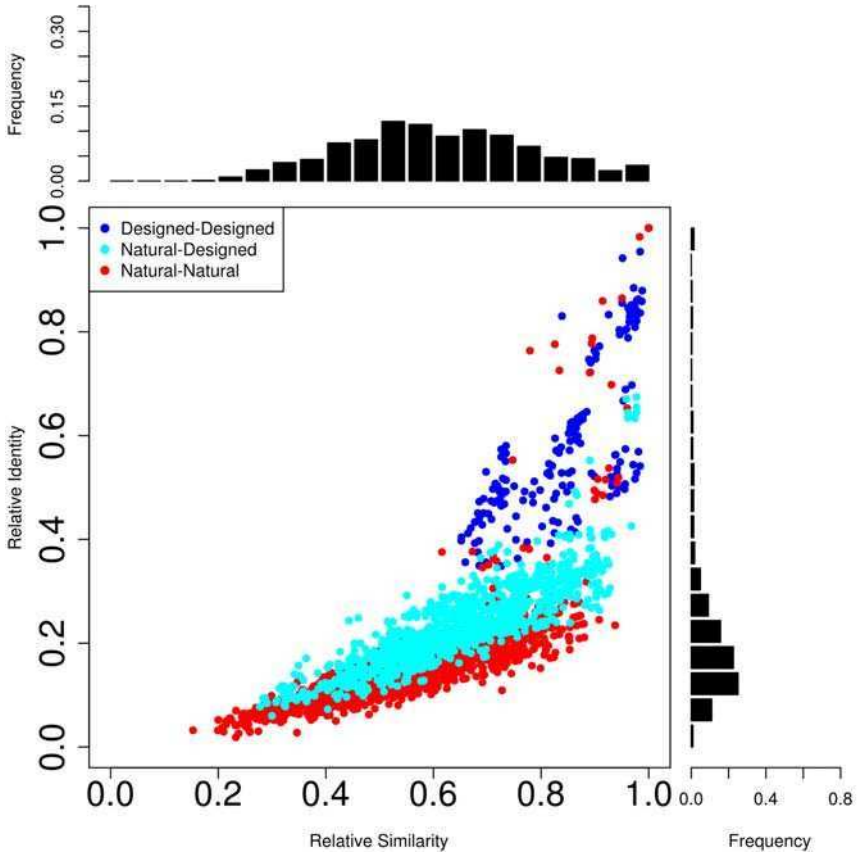


**Figura 3.2:** (A) Las proteínas del conjunto de datos no redundante fueron alineadas de a pares usando TopMatch. A partir de dichos alineamientos, la identidad de secuencias fue obtenida. Tanto el valor de similitud estructural $S$, como el valor de identidad de secuencia, derivado del alineamiento estructural $I$, fueron normalizados de acuerdo a los largos de las moléculas comparadas, obteniendo los parámetros $relS$ y $relI$. Las distribuciones de cada variable son mostradas en el panel superior y a la derecha, en color negro.

Para poder sacar más información acerca de qué características son capturadas en cada caso por los parámetros $relI$ y $relS$, realizamos análisis jerárquicos basados en ambas medidas (Fig. 3.3). Si se realiza un *clustering* basado en $relS$ (Fig. 3.3B), observamos que los grupos que se definen están compuestos, principalmente, por proteínas que contienen números similares de repeticiones. Se observa además la formación de un grupo compuesto por aquellas estructuras que no pertenecen a organismos eucariotas, aquellos que contienen dominios globulares asociados o pertenecen a una clase particular de canales iónicos (la familia TRPV) en cuyos miembros se observa la presencia de grandes desviaciones estructurales respecto de la estructura canónica de las repeticiones ANK [Parra et al., 2015]. A pesar de tener valores que implican una menor conservación en secuencia, respecto de la conservación a nivel estruc-

tural, el *clustering* basado en $relI$ (Fig. 3.3A), agrupa las proteínas principalmente de acuerdo a su ortología y paralogía. Las proteínas diseñadas se segregan como un grupo separado, evidenciando que las proteínas naturales no están, comúnmente, compuestas por repeticiones consenso.

Dado que claramente la estructura de las ANKs parece estar más conservada que su secuencia, decidimos analizar cuál es la región estructural más representativa de las moléculas en nuestro conjunto de estructuras. Para analizar esto, sobre el dendograma obtenido a partir del clustering jerárquico basado en el parámetro $relS$, colapsamos progresivamente las ramas desde la base hacia la raíz, de modo que en cada colapso, se guarda la región compartida por las estructuras colapsadas. Al iterar este proceso hasta llegar a la raíz, observamos que la subestructura común a todos los elementos del grupo, es una región compatible con las dos repeticiones internas de la proteína 4ANK, que corresponde a una proteína consenso, compuesta por cuatro repeticiones idénticas [Mosavi et al., 2002], siendo dicha región las más parecida simultáneamente a todas las proteínas del conjunto.

## 3.2. Organización de los Arreglos Repetitivos en ANKs

Las proteínas repetitivas se componen por colecciones de motivos análogos (potencialmente homólogos) organizados en tándem y que se encuentran relacionados entre sí por diferentes tipos de transformaciones pseudo-simétricas (rotacionales, traslacionales o mezcla de ambas [Kinoshita et al., 1999, Tripp and Barrick, 2003])) entre unidades vecinas. Las ANKs son usualmente descriptas como poseedoras de estructuras altamente regulares y con una arquitectura súper helicoidal [Mosavi et al., 2004]. Sin embargo, de forma similar a otras familias repetitivas, sutiles, pero importantes diferencias conformacionales existen entre las repeticiones. Estas diferencias codifican la compatibilidad entre repeticiones vecinas, afectando en última instancia la conformación global de la proteína y su patrón repetitivo [Ramisch et al., 2014]. Nosotros hemos mostrado que el puntaje de teselado [Parra et al., 2013] captura de alguna forma que tan colectivamente periódica es una estructura teniendo en cuenta las diferentes señales de periodicidad que muestra a diferentes niveles. Al analizar el patrón periódico de una estructura proteica, es posible encontrar un largo particular de fragmento

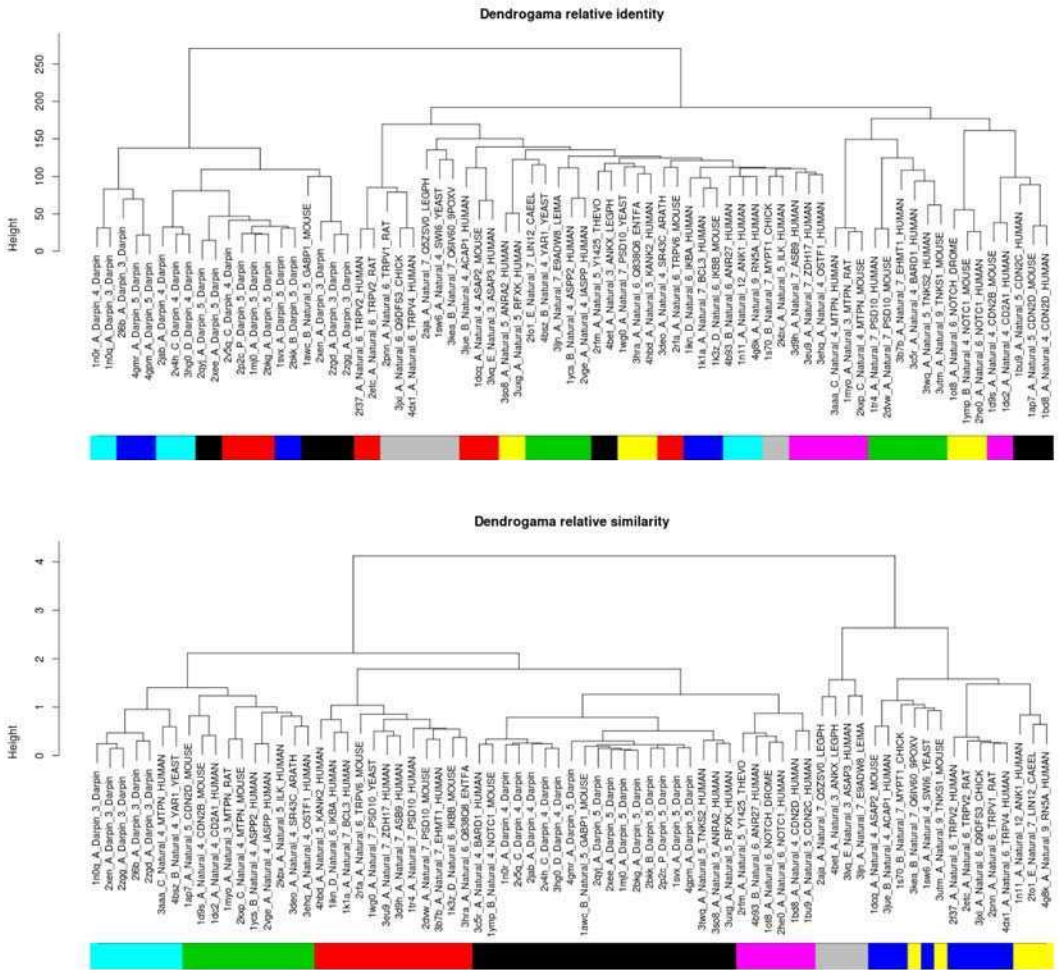


**Figura 3.3:** Clustering en ANKs: Dos métricas diferentes se usaron para construir dendogramas mediante la técnica de clustering aplicada sobre las ANKs. A) El valor de $relI$ se utilizó como métrica para construir el dendograma. Las ANKs en el mismo grupo se encuentran relacionadas por sus relaciones de ortología y paralogía B) El valor de $relS$ se utilizó para construir el dendograma. Las ANKs en el mismo grupo son comparables de acuerdo al número de repeticiones que las componen.

repetitivo en donde la proteína se muestra altamente periódica y los fragmentos de dicha longitud maximizan el cubrimiento de la estructura total mediante copias de sí mismos (son fragmentos maximales), respecto de fragmentos de largos mayores o menores. Vamos a llamar a esta longitud en particular de fragmento, frecuencia característica, por su analogía con el concepto de frecuencia fundamental en la descripción física de funciones periódicas. Así como la frecuencia fundamental se corresponde con la frecuencia de aquella componente en una señal periódica que más información aporta a la misma, al analizar las estructuras mediante el proceso de teselado, otras señales presentes en largos de fragmentos que son múltiplos de la frecuencia fundamental, también muestran altos grados de periodicidad (frecuencias armónicas). La Fig. 3.4 muestra el patrón de teselado de la región D34 de la proteína Ankirina R, que es

un componente del citoesqueleto mediando las interacciones entre las proteínas Espectrinas y Actinas [Michaely et al., 2002], compuesta por 12 repeticiones ANK. No sólo esta proteína es una de las estructuras con mayor número de repeticiones conocida hasta el momento, sino que además también es una de las que posee un mayor grado de simetría en cuanto a las relaciones geométricas entre las mismas. Podemos observar que la frecuencia característica de esta estructura se corresponde con fragmentos de 33 residuos de largo, dada la ubicación del pico de mayor amplitud en la función $\delta\overline{\Theta}$ en el panel lateral. Además se observa que hay otros picos presentes en la función $\delta\overline{\Theta}$ en largos correspondientes a múltiplos de 33 residuos. Esto es posible debido a que la estructura es lo suficientemente regular, para que al tomar fragmentos que contengan 2 o más repeticiones, dichos fragmentos siguen siendo buenos para teselar la estructura global, mediante copias de sí mismos. Dicha capacidad de teselado de fragmentos que se componen de múltiples repeticiones se vería reducida si entre repeticiones vecinas existieran inserciones, deleciones o perturbaciones estructurales que modifiquen la progresión simétrica en el arreglo repetitivo.

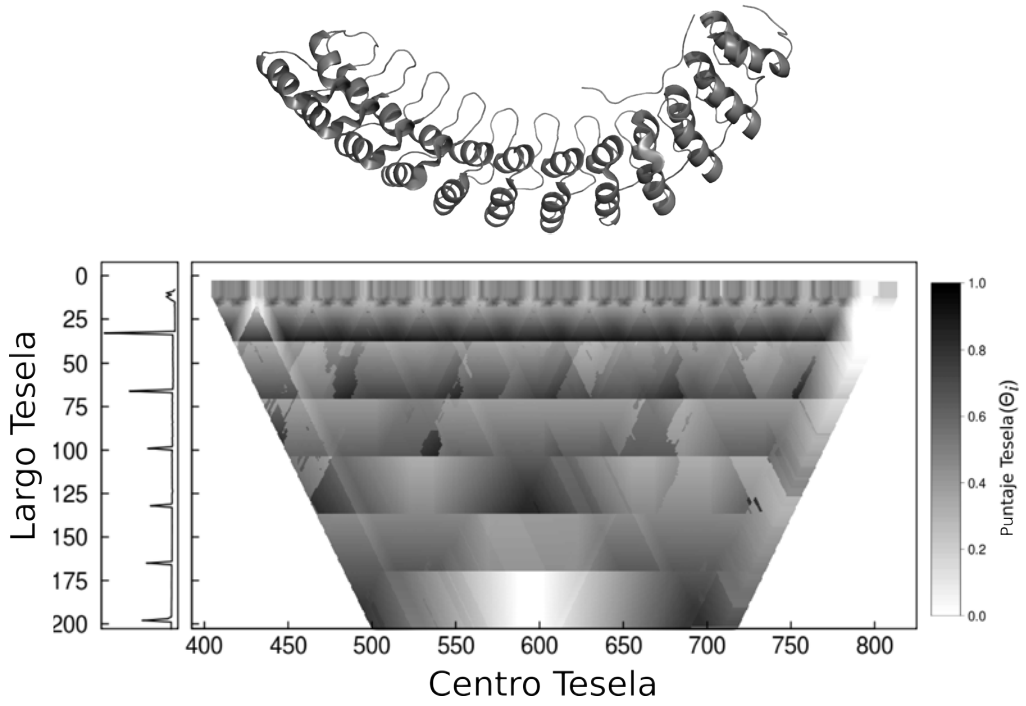


**Figura 3.4:** Patrón de teselado de la proteína D34 (1n11,A) que contiene 12 repeticiones ANK. Se puede observar que hay fragmentos que son maximales locales en largos de fragmentos que son aproximadamente múltiplos del largo del maximal global que corresponde a 33 residuos de largo. En el panel lateral se evidencian estas frecuencias características secundarias como picos de menor amplitud.

Cuando se utilizan fragmentos cuyo largo corresponde a la frecuencia característica para

realizar un teselado de la estructura global, es esperable que estos obtengan altos puntajes, incluso en presencia de perturbaciones en las repeticiones o entre las mismas. Por el otro lado, cuando se usan fragmentos de largos superiores a dicha frecuencia, los puntajes de teselado obtenidos en presencia de perturbaciones estructurales serán inferiores. Esta disminución de los puntajes de teselado se debe a que al tomar inserciones o regiones con deleciones como parte del fragmento a utilizar, al superimponer contra regiones análogas, estas no encontrarán su contraparte e impactarán de forma negativa en el cubrimiento. Las inserciones y deleciones modifican localmente la estructura de una proteína pudiendo modificar además las relaciones geométricas de simetría entre las diferentes repeticiones. Consecuentemente la presencia de inserciones y deleciones puede producir que regiones que son análogas en cada repetición no sean alineables al maximizar la superposición global.

El puntaje de teselado que hemos definido tomaría el valor de 1 en el caso de una molécula perfectamente periódica y un valor cada vez más cercano a 0, mientras menos periódica sea. Hemos mostrado anteriormente que las estructuras proteicas no toman los valores extremos de esta función, ya que no existen estructuras perfectamente simétricas y por el otro lado, incluso una proteína sin simetrías aparentes, es posible de ser teselada con fragmentos de determinados tamaños y elementos de estructura secundaria o súper secundaria simples [Parra et al., 2013]. Los puntajes de teselado para las estructuras de ANKs, del conjunto de datos no redundante, fueron calculados (Fig. 3.5A). A pesar de que estas proteínas comparten una arquitectura común, compuesta de motivos similares, las pequeñas perturbaciones entre y dentro de los mismos son propagadas hacia largos superiores disminuyendo la simetría global de la estructura. Las moléculas donde las transformaciones geométricas son visiblemente homogéneas entre repeticiones y que además no presentan inserciones o deleciones son aquellas que poseen los valores más altos de teselado (por ejemplo: 2he0,A; Fig. 3.5B; 1n11,A; 3so8,A). En contraste, aquellas proteínas que contienen una proporción considerable de estructura no repetitiva (1s70,B; 4bet,A) o que contienen numerosas perturbaciones estructurales (1sw6,A: Fig. 3.5C) se encuentran en el otro extremo de la curva de teselabilidad.

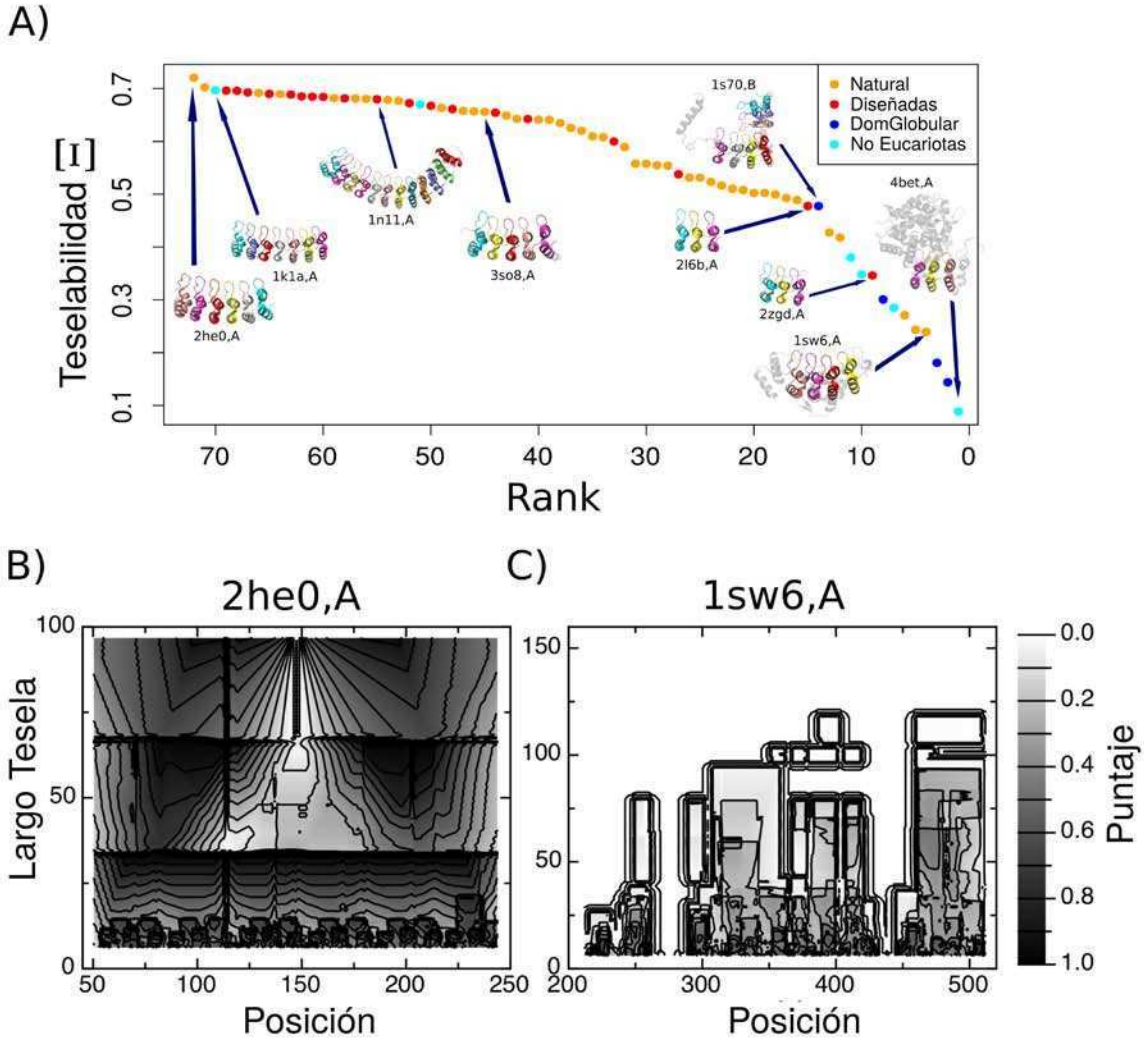


**Figura 3.5:** (A) Valores de teselabilidad para ANKs. Los valores más altos corresponden a las estructuras más regulares y por ello, más simétricas. Los valores más bajos corresponden con aquellas proteínas que contienen perturbaciones estructurales que rompen la propagación de simetrías a escalas superiores mediante la modificación del arreglo espacial de las unidades repetitivas. (B) Teselabilidad por residuo para la molécula 2he0,A que corresponde a la más teselable del conjunto de datos. (C) Tileabilidad por residuo para la molécula 1sw6,A que es una de las menos tileables debido a la ocurrencia de modificaciones estructurales en las repeticiones y la ocurrencia de inserciones entre ellas. Además en esa estructura hay una región en el N-terminal (desde el principio de la molécula hasta aproximadamente el residuo 260) que sirve como un capuchón no repetitivo lo cual le confiere una señal de tileabilidad alta, debido a la disposición espacial análoga de los elementos de estructura secundaria que componen respecto de la estructura de las repeticiones ANK.

Se ha descripto que las proteínas repetitivas usualmente necesitan versiones modificadas de sus repeticiones terminales para poder ser lo suficientemente solubles ( [Aksel et al., 2011]). Nuestro siguiente análisis consistió en separar las repeticiones estructurales detectadas anteriormente en tres grupos diferentes, las repeticiones N-terminales, internas y C-terminales, y analizamos sus propiedades de secuencia y estructura por separado. Análisis de la composición

de los diferentes grupos en términos de estructura secundaria por medio de la herramienta DSSP [Kabsch and Sander, 1983] muestran que existen diferencias observables entre los grupos. La primer hélice se encuentra usualmente extendida en el caso de las repeticiones localizadas en el extremo N-terminal mientras que en el caso de las ubicadas en el extremo C-terminal, la segunda hélice es la que se encuentra extendida (Fig. 3.6), dado que en la estructura canónica de las repeticiones ANK, las hélices son de diferentes largos, puede que estas extensiones sean importantes para la estabilidad de los mismos.

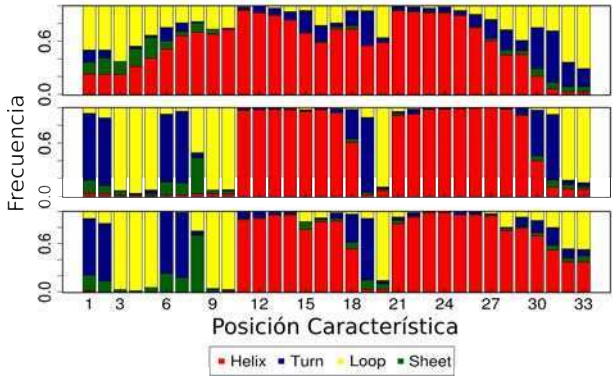


**Figura 3.6:** Perfiles de elementos de estructura secundaria (DSSP) para los diferentes tipos de repeticiones: Usamos el algoritmo de DSSP para definir que elementos de estructura secundaria se encuentran representados en cada posición canónica de las repeticiones ANK para los diferentes tipos de repeticiones. A) Repeticiones N-terminales B) Repeticiones internas C) Repeticiones C-terminales.

Por otro lado, para poder analizar que tan teselable es una estructura en diferente regiones de una estructura, hemos definido una versión del puntaje de teselado aplicable al nivel de residuo único. Este parámetro nos dice que tan frecuentemente un residuo es cubierto por copias de fragmentos usados en un proceso de teselado exhaustivo, es decir, usando todos los fragmentos posibles en cuanto a sus largos y fases, para evaluar su puntaje de teselado. Al hacer esto, somos capaces de encontrar en la estructura, cuáles son las regiones que contribuyen en mayor o menor medida a la teselabilidad global y que consecuentemente afectan a la simetría global de la molécula. Mientras que algunas ANKs son altamente periódicas, como 2he0,A, (Fig. 3.5 B), otras muestran señales de periodicidad leves en sus frecuencias características con varias regiones que son cubiertas de forma muy infrecuente durante el proceso de teselado como en el caso de 1sw6,A (Fig. 3.5 C). En este último ejemplo además, es posible observar la presencia de una región en el extremo N-terminal con una composición

de elementos de estructura secundaria que por su ordenamiento espacial, es estructuralmente alineable con una repetición ANK, sin ser una repetición de este tipo. Este tipo de arreglos de estructuras secundarias, que aparecen acoplados al arreglo repetitivo (presentes en otros casos no mostrados aquí) pueden funcionar a modo de capuchones no repetitivos en este tipo de moléculas proveyendo un nivel adicional de complejidad y versatilidad en las mismas que hasta ahora, no había sido descripto [Parra et al., 2015].

## 3.3. Definición Consistente de las Repeticiones de Ankirina

Aplicamos el método de teselado proteico para analizar las periodicidades y repeticiones en las estructuras de los miembros de nuestro conjunto no redundante de ANKs. Dada su naturaleza repetitiva, las estructuras de ANKs al ser analizadas mediante el método de teselado, deberían ser descomponibles en unidades minimales (teselas) que cuando son repetidas y alineadas espacialmente de manera conveniente contra la proteína de la que provienen, fueran capaces en conjunto de reconstruir la estructura global de la molécula. Como un resultado de esto, observamos que las ANKs son marcadamente periódicas cuando son teseladas con fragmentos de largo 33 (correspondientes al pico de máxima amplitud en el análisis maximal de los teselados en ANKs) y múltiplos de ese largo (Fig. 3.7A) lo cual está en concordancia con el largo que es aceptado en general para las repeticiones ANK en la literatura [Sedgwick and Smerdon, 1999, Tripp and Barrick, 2007]. Si suponemos una onda infinita, la misma tiene un período, pero su fase no es definible ya que no habría un punto inicial y toda fase sería equivalente para reconstruir la onda completa si un fragmento de largo igual al periodo se repitiese infinitamente. A pesar de que las proteínas son objetos finitos, la fase de la unidad repetitiva no es trivialmente definible ya que el punto de comienzo del patrón repetitivo se encuentra difuminado por la evolución de la molécula y la adaptación de su secuencia y estructura a su función y contexto biológicos. Mientras que un patrón periódico puede ser evidente a primera vista, definir los límites entre unidades repetitivas, o períodos, no es tan simple. Diferentes fases, han sido definidas para las repeticiones ANKs en estudio previos de acuerdo a diferentes

criterios. Michaely y Bennet [Michaely and Bennett, 1992] definieron la fase de los ANKs de forma tal que ésta fuera consistente con los límites entre exones e intrones en las secuencias génicas de estas proteínas y con el requerimiento de tener repeticiones completas en los extremos terminales. Esta es la misma fase que se adoptó para construir el modelo oculto de Markov que representa a la familia ANK en la base de datos de Pfam como así también para el diseño de ANKs consenso según el procedimiento de Mosavi y sus colaboradores [Mosavi et al., 2002]. Luego Sedgwick y Smerdon [Sedgwick and Smerdon, 1999] definieron las repeticiones de ANK de tal forma que se minimizara la presencia de elementos no conservados internos a la definición de la repetición. Esta fase fue adoptada por el grupo de Plückthun para diseñar las ANKs consenso, denominadas DARPins. Por su parte, Tripp y Barrick definieron un nuevo consenso [Tripp and Barrick, 2007] con una fase en donde los límites entre repeticiones adyacentes estuvieran localizadas en el medio del loop que las conecta. La elección de esta fase se debe a varias razones. En primer lugar, el loop es un lugar común para inserciones en ANKs. En segundo lugar, este loop tiene una conservación en secuencia relativamente más baja, que el resto de la estructura canónica de las repeticiones, además de una mayor exposición al solvente (y una menor densidad de empaquetamiento) sugiriendo que podría ser tolerante a pequeñas perturbaciones. También, el uso de este sitio evita inserciones dentro de las hélices que definen la repetición. Por último, la mayoría de los intrones en ANKs se encuentran en ese loop, indicando que el mismo podría ser un punto a partir del cual pueden ocurrir duplicaciones de este tipo de repeticiones.

Esta multiplicidad de fases definidas para las repeticiones en ANKs evidencia la falta de conocimiento en cuanto al origen evolutivo de las repeticiones, sus patrones de duplicación y su subsiguiente divergencia. Para nuestros estudios es importante elegir una fase que se encuentre definida de forma consistente en todas las repeticiones de la familia ANK para poder realizar anotaciones consistentes y análisis comparativos sobre las mismas. Anteriormente, fuimos capaces de mostrar que nuestro algoritmo de teselado proteico, era capaz de detectar unidades repetitivas que maximizaran el cubrimiento de la estructura total, por medio de repeticiones y alineamientos de un fragmento dado de la proteína. Sin embargo, lejos de ser exhaustivamente testeado sobre una familia en particular, en nuestro estudio inicial, el método fue aplicado sobre

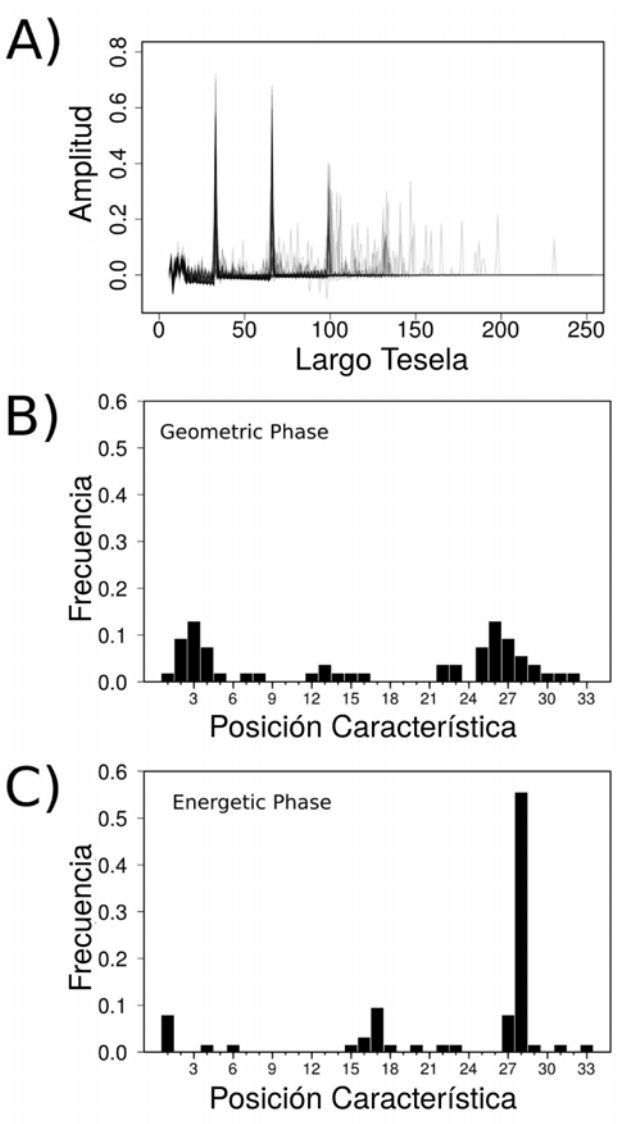


**Figura 3.7:** Definiendo la fase de las repeticiones de ANKs. (A) Funciones $\delta$ para las proteínas ANK. La mayoría de las proteínas ANK muestran un pico en sus funciones $\delta$ a un largo de fragmento de 33 residuos de largo. (B) Distribución de fases geométricas para el conjunto de estructuras ANK no redundantes derivadas del proceso de teselado. (C) Distribución de fases energéticas derivadas de la función de plegabilidad relativa.

un grupo acotado de estructuras repetitivas, representantes de las principales familias que a su vez, presentaban no sólo repeticiones del tipo solenoidal, donde la relación de simetría entre las unidades es principalmente traslacional, sino que también sobre estructuras con simetría rotacional o mixta [Parra et al., 2013]. En un intento de evaluar el método sobre una familia, se evaluaron los patrones de teselado de todos los fragmentos de 33 residuos de largo, sobre las estructuras de ANKs. Para cada caso, se seleccionó como fragmento representativo de la repetición básica, aquel que maximizara el valor de teselado y su fase fue definida de forma relativa al modelo oculto de Markov que representa a la repetición ANK (ANK Hmm) en la base de datos de Pfam (Pfam ID: PF00023) [Bateman et al., 2004]. Al hacer esto, se observa que en el 70 % de los casos, no fue posible obtener un fragmento único con un valor de teselado máximo, sino un conjunto de fragmentos que compartía dicho valor y por ende, la fase de la repetición no pudo ser definida por medio de este método geométrico de detección. Para el 30 % restante del grupo de datos, no se observó una fase claramente conservada para las repeticiones (Fig. 3.7B).

¿Hay alguna manera objetiva de definir la fase de las repeticiones ANKs? ¿Tiene alguna utilidad definir una fase conservada para toda la familia ANK? Importa en el sentido biológico la fase de las repeticiones o es sólo un requerimiento para poder realizar análisis comparativos como los que proponemos?

La estabilidad de las proteínas, y de los arreglos repetitivos, podría estar relacionada a la selección de una fase preferencial. En las proteínas repetitivas, cada unidad está compuesta por elementos estructurales similares que interactúan con sus vecinos cercanos. Como consecuencia de ello, el proceso de plegado de los mismos puede ser descripto como múltiples embudos que colapsan generando un paisaje energético con forma de embudo global ( [Ferreiro and Wolynes, 2008]), es decir, cada repetición podría funcionar como una subunidad de plegado. Siguiendo esta idea, tomamos todos los fragmentos de 33 residuos de largo y calculamos su valor de plegabilidad relativa (*relative foldability*) definido como $\Theta_r = \Delta E/(\delta E\sqrt{N})$ por Panchenko y Wolynes ( [Panchenko et al., 1996]). Para poder calcular $\Theta_r$, la energía que corresponde a la suma de todas las interacciones internas a un fragmento, medidas a partir de la función de energía AM/W (*Associative Memory Hamiltonian, Water Mediated*) [Papoian et al., 2004], es comparada a la energía media de un grupo N de estructuras ($\Delta$E) y su

varianza ($\delta$E). Las N estructuras que son usadas para calcular dichos valores corresponden a todos los posibles fragmentos del mismo largo que el fragmento que está siendo evaluado, que pueden ser definidos en la estructura global. El número de fragmentos en cada caso puede ser calculado como $N = L - l + 1$ donde $L$ es el largo de la estructura y $l$, el largo del fragmento. Finalmente, se obtiene una medida que representa que tan plegable es un fragmento respecto de todos los fragmentos de su mismo largo en la proteína. Aquellos fragmentos de 33 residuos de largo que maximizan el valor de plegabilidad en cada caso fueron seleccionados y sus fases obtenidas del mismo modo anteriormente mencionado, por comparación con el ANK Hmm de Pfam. Es posible observar que en este caso la ocurrencia de una fase conservada en la mayoría de los miembros del conjunto de datos, correspondiente a aquella fase definida a partir de la posición 28 respecto del modelo de Pfam. Aquellos casos que se desvían de esta fase corresponden principalmente a proteínas repetitivas diseñadas del tipo Darpin y a miembros de la familia de canales iónicos TRPV. Es interesante mencionar que aquellas proteínas diseñadas siguiendo el método ideado por Peng [Mosavi et al., 2002], en donde las covariaciones internas entre aminoácidos fueron respetadas, concuerdan con la fase derivada de los cálculos energéticos, compatible con las fases obtenidas para la mayoría de las proteínas naturales (Fig. 3.7D).

## 3.4. Anotación y caracterización de las repeticiones

Una vez que tanto el período como la fase de las repeticiones han sido unívocamente definidos, el siguiente paso en nuestro estudio es el de anotar las repeticiones en cada ANK de nuestro conjunto de datos no redundante (Tabla 3.1, todas las tablas se encuentran al final del capítulo). Para realizar esto, seleccionamos la primera repetición interna de la proteína diseñada 4ANK, correspondiente parcialmente a la región que se encuentra conservada a nivel estructural en todas las proteínas ANK (como se mostró en el análisis de *clustering* jerárquico basado en el parámetro $relS$ de similitud estructural). Este fragmento correspondiente a 4ANK fue definido de tal forma que su fase comenzara en la posición 28 respecto del modelo de ANKs en Pfam para así mantener la fase favorecida según la plegabilidad relativa en repeticiones de ANK. Aplicamos entonces una variación de nuestro método de teselado en donde en vez

de usar un fragmento interno a una estructura para realizar el teselado, en cambio se utiliza un fragmento de una estructura para teselar a una segunda (teselado cruzado, Fig. 3.8A). El fragmento definido a partir de 4ANK fue usado para teselar a todas las demás estructuras del grupo (incluyendo a sí misma) maximizando el cubrimiento de las mismas mediante copias de dicho fragmento.

Como resultado de este procedimiento, obtuvimos un conjunto de alineamientos de pares de secuencias, basados en el alineamiento estructural, entre el fragmento de 4ANK (1n0r,A) y las regiones que fueron alineadas en cada proteína blanco. En la Fig. 3.8 se ejemplifica el procedimiento, usando como ejemplo el teselado de la estructura correspondiente a 2rfa,A (Fig. 3.8A). Dado que el fragmento usado para el teselado cruzado, a partir de ahora llamado fragmento fuente, fue el mismo en todos los casos (se menciona como RefRepeat4ANK en la figura), los alineamientos obtenidos siempre contienen la secuencia del mismo en el par. Si se observa solamente la secuencia del fragmento fuente en los diferentes alineamientos de a pares, se observa que lo único que cambia entre ellos es la proporción de gaps y la ubicación de los mismos (secuencias superiores en cada par de secuencias). Dada esta característica, ideamos un método que fuera capaz de redistribuir los gaps en los diferentes alineamientos, respecto del fragmento fuente de forma que al final del proceso, todos estos fragmentos fueran equivalentes, respecto de las posiciones que contienen gaps. Durante ese procedimiento se aplicaron las siguientes condiciones: i) Cada vez que se modifica/inserta un gap en la secuencia del fragmento fuente, la misma modificación es aplicada al fragmento blanco. ii) se genera un alineamiento múltiple entre todos los fragmentos fuente, redistribuyendo los gaps hasta que todas las secuencias de los mismos queden equivalentes (solo pueden agregarse gaps y no quitarse ya que esto implicaría una deleción en el fragmento blanco asociado). Al final del proceso todas las secuencias blanco resultan alineadas entre ellas de forma indirecta y transiente dado el alineamiento de las secuencias fuente a las cuales están asociadas (Fig. 3.8C). Un punto crucial en este método es que en ningún momento se usa información de secuencia, generando un alineamiento de secuencias puramente guiado por los alineamientos estructurales. Los alineamientos de a pares, por su parte, no sólo sirven como paso intermedio para obtener el alineamiento múltiple, sino que además dado que el fragmento usado de

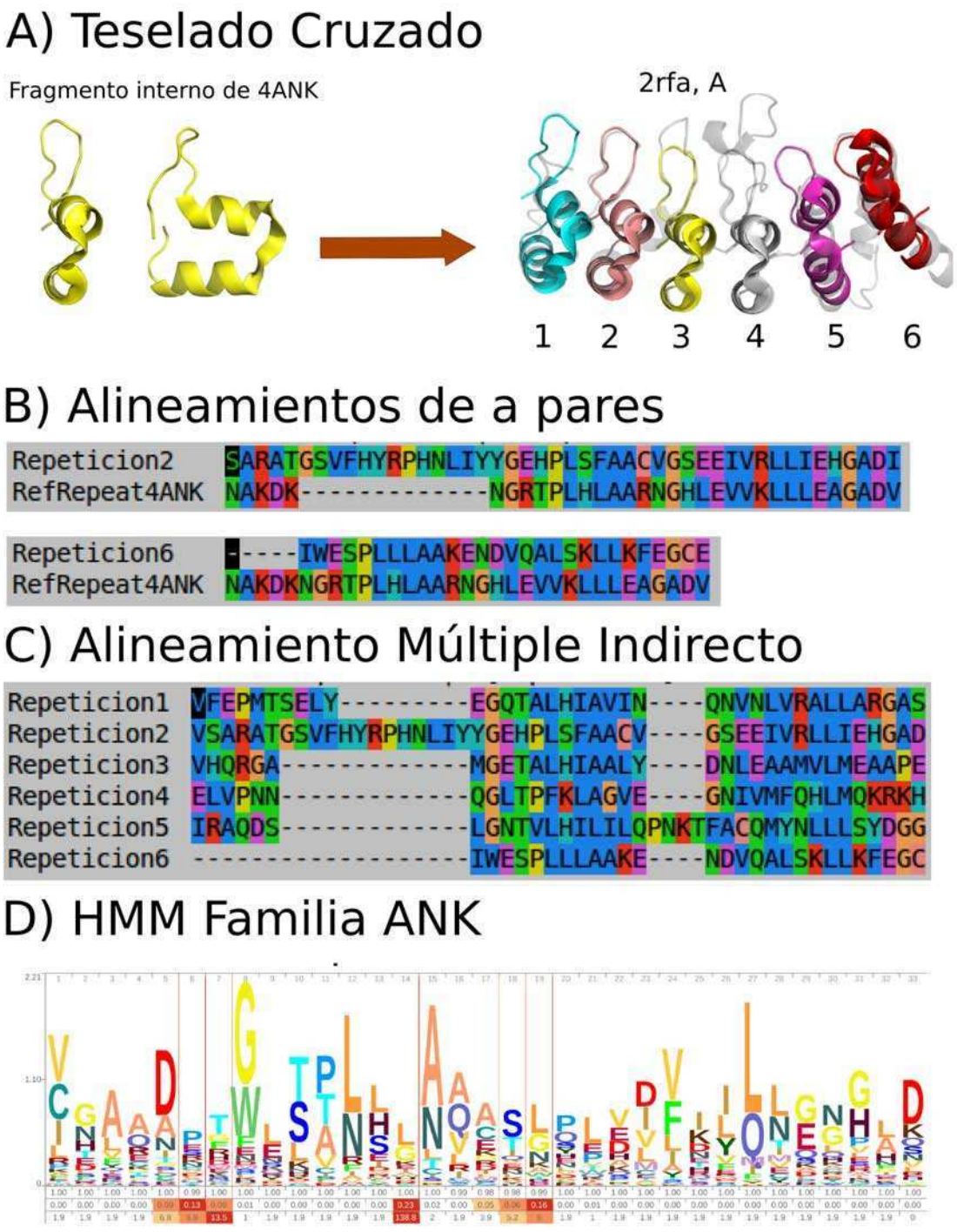


**Figura 3.8:** Esquematización del procedimiento de teselado cruzado y obtención de alineamientos múltiples de las secuencias de las repeticiones ANK.

4ANK, representa a una repetición canónica, en cada caso, pueden definirse las posiciones canónicas en las secuencias blanco, así como también las deleciones e inserciones presentes en las mismas (correspondientes a gaps presentes en las secuencias blanco o fuente en el alineamiento, respectivamente). Por ejemplo, observamos que en la repetición 2, detectada por este método en 2rfa,A, la secuencia correspondiente a GSVFHYRPHNLIY corresponde a una inserción respecto de la repetición usada como fuente. En el caso de la repetición 6 de la misma estructura, vemos que las primeras 5 posiciones, correspondientes al motivo canónico NAKDK, se encuentran ausentes en la secuencia blanco (Fig. 3.8D).

Cada una de las secuencias blanco obtenidas a partir del proceso de teselado proteico cruzado usando el fragmento de 4ANK, es considerada como una repetición en su estructura correspondiente. El largo de las repeticiones obtenidas varía entre 24 residuos a un máximo de 48 (Fig. 3.9A). Se observa que las distribuciones de largos difieren cuando se consideran las repeticiones internas o las que se encuentran en los extremos. Las repeticiones en los extremos son en mayor frecuencia más cortas, generalmente debido a la ausencia de la región $\beta$-hairpin que se compone a partir del inicio de una repetición y el final de una de las vecinas (Fig. 3.9B). Las inserciones y las deleciones no se encuentran homogéneamente distribuidas a lo largo de las repeticiones sino que en posiciones específicas según cada caso. Por un lado, se observa que las inserciones están principalmente localizadas en la región del $\beta$-hairpin y entre las dos $\alpha$-hélices, lo cual se intentaba evitar en la definición de fase de Tripp para las repeticiones de ANKs [Tripp and Barrick, 2007]. Por otro lado, las deleciones están particularmente ubicadas al final de la primer $\alpha$-hélice. A pesar de que el largo de las inserciones y deleciones puede ser variable (hasta 16 y 4 residuos, respectivamente), éste no correlaciona con la localización de las mismas dentro del marco canónico de 33 posiciones de las repeticiones (Fig. 3.9C y Fig. 3.9D). Al finalizar el proceso de anotación, cada residuo de cada estructura del grupo no redundante fue anotado como perteneciente a una posición canónica de una repetición específica, una inserción (también se lleva el registro de las posiciones canónicas que se encuentran delecionadas) o como parte de la región no repetitiva más allá del arreglo repetitivo.

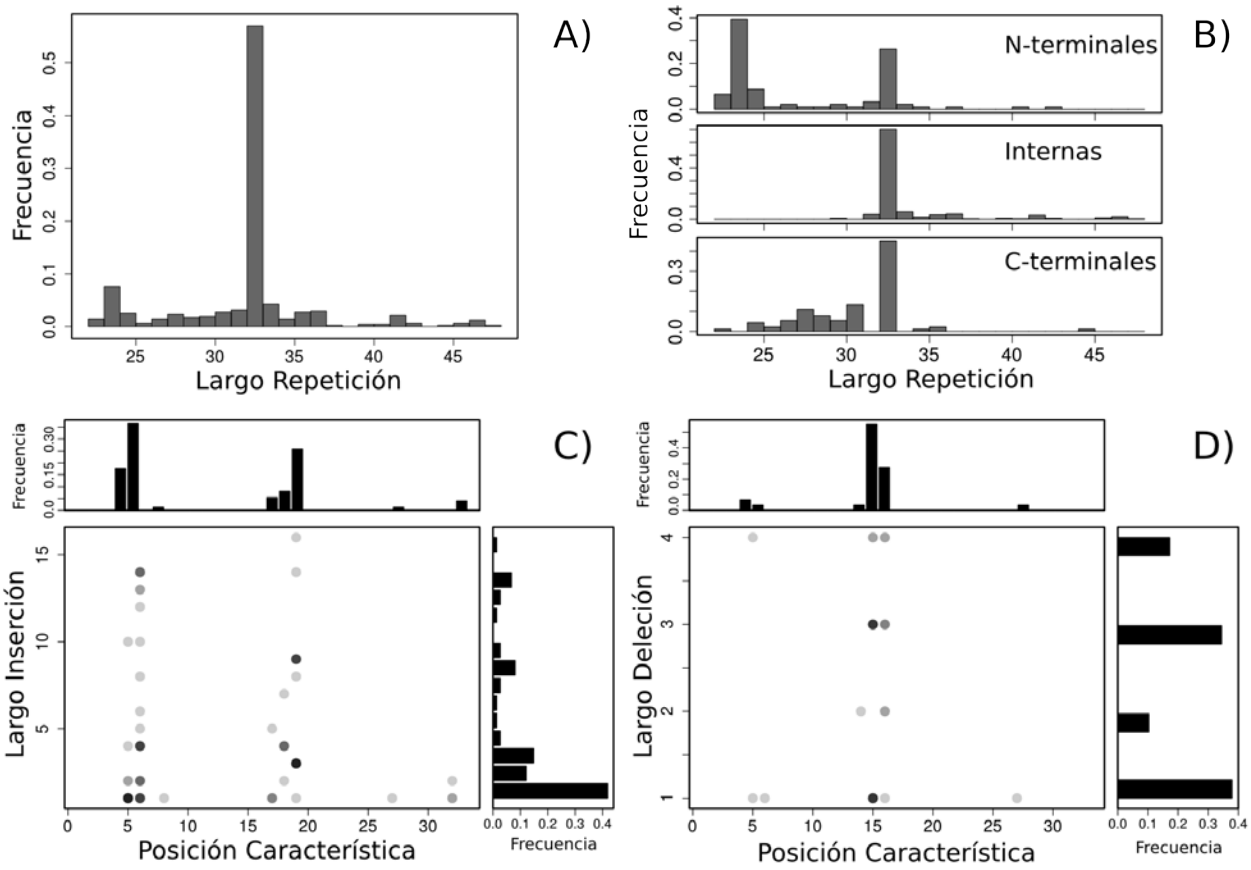


**Figura 3.9:** Estructura básica de las repeticiones ANK: (A) Distribución de largos de las repeticiones detectadas a partir de las estructuras. (B) Distribución de largos para los diferentes tipos de repeticiones de acuerdo a su ubicación en el arreglo repetitivo (C) Distribución de deleciones a lo largo de las posiciones canónicas de las repeticiones ANK relativas a la fase del HMM calculado a partir de las detecciones estructurales (eje x) y su largo (eje y). (D) Distribución de inserciones a lo largo de las posiciones canónicas en las repeticiones ANK relativas a la fase del HMM calculado a partir de las detecciones estructurales (eje x) y su largo (eje y).

## 3.5. Descripción Energética de las ANKs

Las proteínas ANK son típicamente descriptas como interactores proteicos, no habiendo miembros de la familia en donde se haya reportado algún tipo de actividad enzimática [Sedgwick and Smerdon, 1999]. Las proteínas son el resultado de dos fuerzas que muchas veces se contraponen llegando a un equilibrio. Por un lado se encuentra su habilidad de plegarse hacia una estructura estable y por el otro se encuentra la capacidad de llevar a cabo su función. En este sentido, los conflictos energéticos que puedan existir en la arquitectura de las ANKs, podrían estar relacionados con su habilidad de reconocer e interactuar con otras proteínas. Hemos usado la herramienta Frustratometer, desarrollada por nuestro grupo [Jenik et al., 2012] para poder encontrar dónde en las estructuras, éstos conflictos energéticos están localizados y cuantificarlos a partir del cálculo de los diferentes índices de frustración local [Ferreiro et al.,

2007b]. Según la metodología implementada en el Frustratometer (ver la sección de métodos para más detalles), las interacciones pueden clasificarse en 3 tipos: aquellas que están en conflicto con la estructura en que se encuentran, o altamente frustradas, aquellas que son neutras, o aquellas que están mínimamente frustradas, es decir, que son favorables para mantener la estructura de la que forman parte. Luego de calcular los patrones de frustración local sobre todos los miembros del grupo no redundante de ANKs, observamos que las interacciones altamente frustradas, no se encuentran distribuidas de forma aleatoria en sus estructuras. A continuación describiremos como se encuentra localizada y en que proporciones, la frustración en la arquitectura de las ANKs.

## Conservación de las secuencias y de los patrones de frustración local a lo largo de la estructura canónica de las repeticiones ANKs.

Hemos mostrado anteriormente que existen diferencias acerca de la composición, en términos de estructura secundaria, a lo largo de los diferentes tipos de repeticiones. Por esta razón, de la misma manera en que lo hicimos anteriormente, separamos el conjunto de repeticiones en 3 grupos, según sus ubicaciones en los arreglos repetitivos en repeticiones N-terminales, internas y C-terminales. Sobre estos grupos aplicamos el procedimiento de alineamiento indirecto basado en estructura de forma de obtener alineamientos de secuencias múltiples específicos para cada tipo de repetición. Una vez obtenidos los alineamientos, calculamos a partir de ellos sus correspondientes HMM a partir de los cuales se calcularon además sus logos de secuencia. Como era de esperarse, se pueden notar marcadas diferencias entre las señales de secuencia para cada tipo de repetición (Fig. 3.11A). Las repeticiones internas son las que son más similares al modelo que representa a la familia en Pfam y dentro de los 3 modelos obtenidos, es el único tipo de repetición en el que el motivo TPLH, importante para la estabilidad de los dominios repetitivos ANK [Guo et al., 2010], se encuentra altamente conservado. Por el contrario, el motivo TPLH parece estar completamente ausente en las repeticiones N-terminales y semi conservado en las repeticiones C-terminales. Otras posiciones que se encuentran altamente conservadas en el modelo de Pfam, como la Glicina en la posición 8, las Alaninas en las posiciones 15 y 16 o las Leucinas en las posiciones 27 y 28 se encuentran de igual forma,

altamente conservadas en el caso de las repeticiones internas, semi conservadas en las correspondientes al extremo C-terminal y menos conservadas en el caso N-terminal. Así mismo, el contenido de información de las secuencias ($ICSeq$, ver métodos para su definición) para el perfil completo, es más alto en el caso de las repeticiones internas (o también podría decirse que la entropía es más baja). El modelo de las repeticiones N-terminales por su lado, muestra el menor IC Seq, mientras que el de los C-terminales es un caso intermedio entre los otros dos. Usamos la herramienta HHAlign [Soding, 2005] para poder comparar estos 3 modelos, que muestran señales diferentes y poder estimar cual es la probabilidad de que los distintos modelos de repeticiones ANK, sean homólogos entre ellos. HHAlign calcula un puntaje y un e-valor para los dos HMMs en los que se está testeando la posible relación de homología (Fig. 3.10). El e-valor más alto (1.3e-7, puntaje=96.71) se obtuvo cuando se compararon los modelos N-terminal y el C-terminal lo cual estaría diciendo que estos dos modelos serían los menos probables de ser homólogos. E-valores significativamente menores se obtuvieron cuando se comparan los modelos de los extremos N-terminal y C-terminal contra el interno (3.7e-12, puntaje=98.58 y 3e-14, puntaje=98.99 respectivamente). Esto sugeriría que los e-valores mayores obtenidos al comparar los modelos de los terminales podrían ser producto de su alta divergencia en secuencia que estaría enmascarando su relación de homología la cual es evidente al comparar contra el modelo Interno que es cercano a ambos. Una de las posibles causas para esta divergencia tan alta entre los diferentes tipos, podría atribuirse, en parte, a la diferente exposición al solvente a la que se encuentran expuestos, dadas sus ubicaciones en los arreglos repetitivos [Bustamante et al., 2000].

Una vez generados los perfiles de secuencia para cada tipo de repetición realizamos un análisis de los mismos en cuanto a sus patrones energéticos evaluados a partir del cálculo de frustración local de las repeticiones estructurales correspondientes a las secuencias de cada modelo. El índice de frustración al nivel de residuo único (*single level frustration index*), fue calculado para cada posición canónica en cada repetición ANK [Ferreiro et al., 2007b]. Tres clases de residuos pueden definirse usando este índice: residuos mínimamente frustrados, neutros o altamente frustrados. Usando estas tres clasificaciones para el índice de frustración, calculamos el contenido de información basado en frustración ($FrustrationIC$, ver métodos

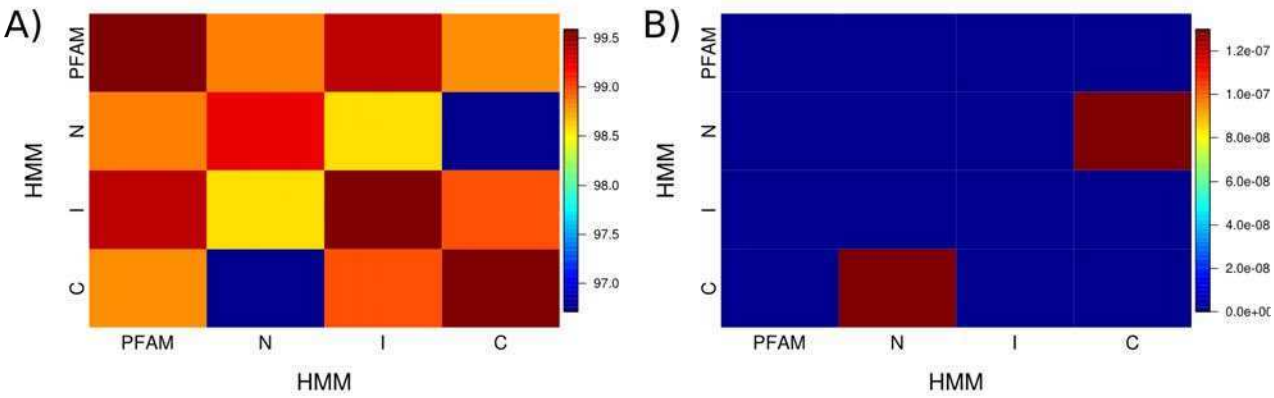


**Figura 3.10:** A) Análisis de homología entre los diferentes modelos de repeticiones ANK usando HHAlign. A) Valores del puntaje de HHAlign al analizar los diferentes modelos de a pares. Mientras mayor el puntaje de HHAlign, mayor es la probabilidad de que los dos modelos sean homólogos. Nótese que los valores de la diagonal son heterogéneos debido a que se muestran los puntajes crudos, en donde la comparación de un HMM contra sí mismo devuelve un valor dependiente del HMM y su contenido de información global. B) E-valores que calcula HHAlign dados los puntajes obtenidos para las comparaciones de a pares entre los modelos.

para su definición) para cada posición canónica para cada tipo de repetición (Fig. 3.11B). Al igual que el $SeqIC$, el $FrustrationIC$, es alto para el caso de las repeticiones internas y bajo para el caso de las repeticiones del extremo N-terminal y un caso intermedio para los del C-terminal. Es notable observar que no hay posiciones en ninguno de los 3 grupos, en donde el $FrustrationIC$ sea alto y el estado máximamente frustrado sea el estado más informativo, lo cual es un indicador de que no hay posiciones que se encuentren sistemáticamente frustradas en la arquitectura ANK. Esto último contrasta con análisis hechos en nuestro grupo para proteínas con actividad catalítica como las Lactamasas para las cuales observamos que determinados residuos catalíticos se encuentran altamente frustrados en forma conservada (Guzovsky y colaboradores, no publicado). Por el contrario, los residuos altamente frustrados parecen ser específicos de las proteínas en donde se encuentran y podrían estar directamente relacionados con su función. Otra observación notable es que aquellas posiciones que son permitidas para ser aleatorizadas mediante mutaciones, en las repeticiones internas de las Darpins diseñadas por Plückthun [Binz et al., 2003], que actúan como interactores proteína-proteína específicos, tienen bajos niveles de $FrustrationIC$ (posiciones marcadas con la letra P sobre las barras del modelo interno en la Fig. 3.11B). Sin embargo, hay otras posiciones con bajos valores de $FrustrationIC$ comparables a aquellos que se mutan en las Darpins. Particularmente

hay 5 residuos con esa característica, la posición 4 (respecto de la fase definida anteriormente), que para el modelo consenso de Peng (autor de otro tipo de Ankirinas diseñadas por consenso [Mosavi et al., 2002]), se corresponde con una K, ubicada en el $\beta$-hairpin, la posición 14 que se encuentra inmediatamente antes del motivo TPLH, y las posiciones 21, 22 y 29 que se encuentran ubicadas en el lado expuesto al solvente de la hélice 2.

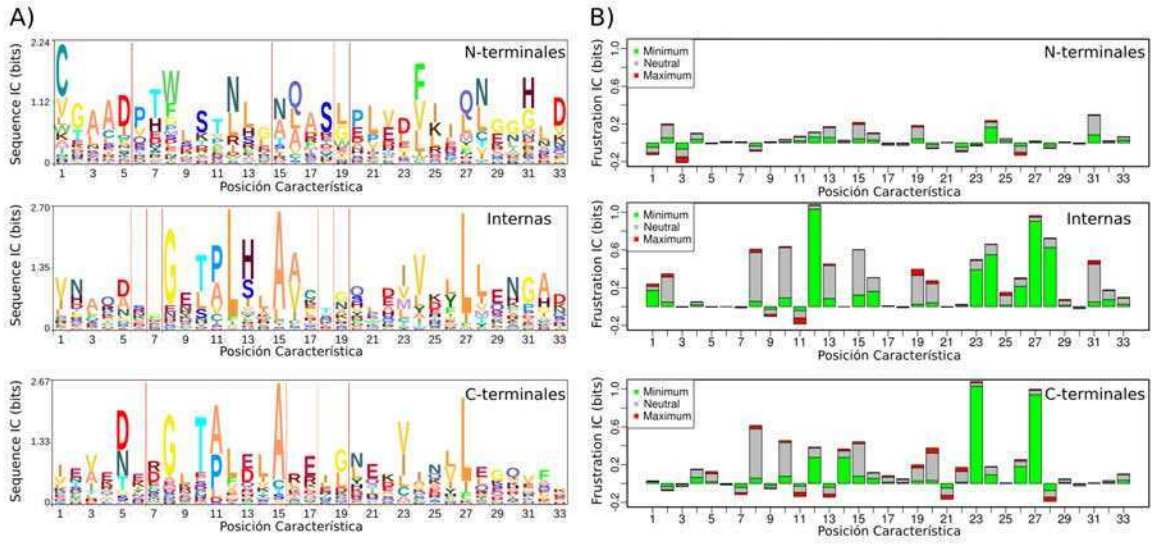


**Figura 3.11:** Conservación de la frustración en las repeticiones ANK. A) Logo de secuencias correspondiente al modelo oculto de Markov derivado del alineamiento estructural indirecto para las repeticiones N-terminales, Internas y C-terminales. B) Contenido de información para la conservación del estado de frustración de acuerdo al índice de frustración de residuo único para repeticiones N-terminales, Internas y C-terminales.

A pesar de ser calculados a partir de diferentes fuentes, los valores de $SeqIC$ y $FrustrationIC$, para las posiciones consenso en los diferentes tipos de repeticiones son comparables por estar en las mismas unidades (bits). Existe una correlación positiva y significativa entre éstas dos medidas para el caso de las repeticiones internas y C-terminales, siendo mayor el coeficiente de correlación para el primer caso. No se observó una correlación significativa para el caso de las repeticiones ubicadas en el extremo N-terminal (Fig. 3.12A). Esto podría significar, para los casos en que la correlación es significativa, mientras más conservado y similar al consenso es un aminoácido en una posición canónica, mayor es su contribución para la diferencia energética entre los estados plegado y desplegado de la estructura de la cual forma parte. Esto establece una conexión directa entre la conservación al nivel de la secuencia y la estabilidad estructural, que son medidas de formas independientes. Siguiendo esta lógica, los valores menores de $FrustrationIC$ obtenidos para las repeticiones de los extremos, significaría que

estos son menos estables, comparados con los internos. Esto se encuentra en concordancia con los experimentos en donde se ha mostrado que para varias proteínas repetitivas, las regiones terminales son las primeras en desplegarse. Se ha mostrado experimentalmente y computacionalmente que para el supresor de tumores P16, la repetición N-terminal es la primera en desplegarse [Tang et al., 2003, Interlandi et al., 2006], que también es el caso para la proteína D34 [Werbeck et al., 2008]. La proteína Notch por su parte, posee una repetición N-terminal que posee rasgos parciales de desorden intrínseco [Ehebauer et al., 2005]. La proteína Gankirina, por último, comienza a desplegarse por el extremo C-terminal cuando se encuentra aislada y cambia su mecanismo de desplegado cuando se encuentra en complejo, donde entonces es la región N-terminal, la primera en desplegarse [Settanni et al., 2013]. Por otro lado, la flexibilidad estructural de las repeticiones terminales de I$\kappa$B$\alpha$ es crucial la función de esta proteína. De las 6 repeticiones que componen el arreglo repetitivo, las repeticiones 1, 5 y 6 son conformacionalmente flexibles [Truhlar et al., 2006]. Se ha mostrado que las transiciones de plegado/desplegado en la sexta repetición son cruciales para regular la degradación mediada por proteosoma de I$\kappa$B$\alpha$ [Alvarez-Castelao and Castano, 2005]. Adicionalmente, la plasticidad conformacional de las repeticiones 5 y 6 junto con la secuencia PEST en el extremo C-terminal, es importante en el proceso de desmontado (*stripping*) de el complejo NF-$\kappa$B/ADN en el núcleo e inactivar su actividad transcripcional [Bergqvist et al., 2008]. Así, es común observar que las repeticiones terminales son menos estables que las internas. Esto puede tener un origen geométrico, ya que las repeticiones terminales tienen sólo una interfaz con repeticiones adyacentes y se ha observado que la formación de interfaces entre repeticiones es uno de los factores más estabilizantes en este tipo de moléculas. También se observa que las repeticiones terminales están enriquecidas en interacciones altamente frustradas lo cual también puede ser un indicio de esta inestabilidad haya sido seleccionada evolutivamente por su contribución a la actividad biológica.

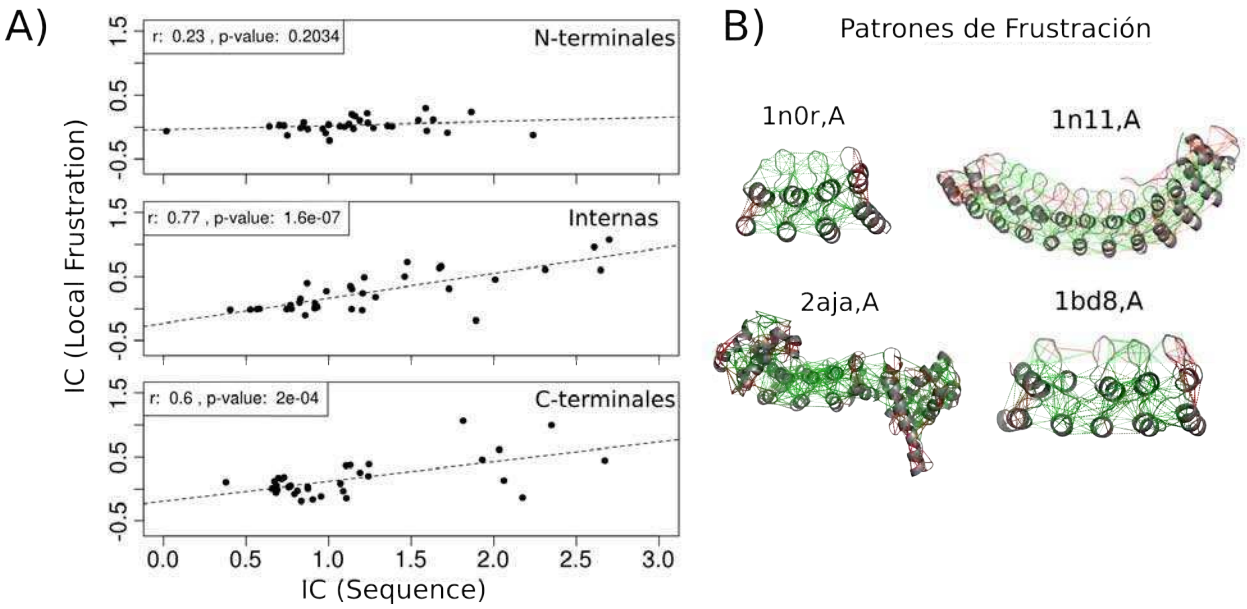


**Figura 3.12:** A) Relación entre el contenido de información para cada posición canónica en las repeticiones ANK medido a partir de la conservación de la secuencia de aminoácidos y el contenido de información medido según la conservación del estado de frustración. Existe una asociación positiva entre las dos variables con p-valores significativos (>0.05) para el caso de las repeticiones internas y las C-terminales. Esta asociación no es significativa en el caso de las repeticiones N-terminales. B) Indice de frustración configuracional calculado para algunos miembros de la familia ANK (PDBs: 1n0r,A; 1n11,A; 2aja,A y 1bd8,A ). Las líneas verdes corresponden a aquellas interacciones que son favorables para la mantención de la estructura mientras que las rojas corresponden a las interacciones que son desfavorables.

## Conservación de la frustración local a lo largo de los mapas de contacto canónicos en repeticiones ANK

En la sección anterior, calculamos los patrones de frustración local al nivel de los residuos en las estructuras ANK. Asimismo, la frustración local, puede medirse al nivel de contactos entre residuos. Al analizar residuos únicos, se usó la repetición de largo 33 para realizar los análisis. En el caso de los mapas de contacto, la unidad analizada corresponde a pares de repeticiones. La razón de usar los pares es para poder analizar tanto las interacciones que ocurren dentro de cada repetición, como así también, las interacciones que ocurren en las interfaces con los vecinos inmediatos. A los fines de analizar aquellos casos que sean comparables entre sí, se usaron sólo aquellos pares de contactos correspondientes a repeticiones internas dentro de los arreglos repetitivos. Una vez obtenidos todos los pares, se generaron los mapas de contactos de cada par de repeticiones, según los límites de distancias usados en el Frustratometer (ver sección métodos) para definir contactos entre residuos en proteínas.

Una vez calculado el mapa de contacto de cada par, se analizó la frecuencia de observar un contacto entre cada par posible de las 66 posiciones canónicas que los conforman (Fig. 3.13A, matriz triangular inferior). Para cada interacción canónica se calculó el valor de frustración configuracional. Posteriormente, una vez obtenida toda la distribución de valores para esa interacción en particular, en todo el conjunto de datos, se calculó el $FrustrationIC$, de la misma forma en que se calculó anteriormente para los residuos. Conjuntamente con el cálculo del $FrustrationIC$ para cada interacción canónica, también se registró en cada caso, cual de los 3 estados (altamente frustrado, neutro o mínimamente frustrado) fue el estado que más información aporta al valor total. El valor obtenido de $FrustrationIC$ fue pesado por su correspondiente valor de frecuencia relativa observado. Esta frecuencia indica cuán frecuentemente existe una interacción canónica dada en los pares de repeticiones dentro del conjunto de datos. Los valores obtenidos se dividieron según cuál sea el estado de frustración más informativo en cada interacción y las distribuciones para los 3 casos fueron calculadas. Los valores más altos de $FrustrationIC$ pesados por su frecuencia relativa, corresponden a contactos canónicos donde el estado mínimamente frustrado es el más conservado, mientras que aquellos en los que el estado altamente frustrado es el más informativo no presentan valores de $FrustrationIC$ pesado por su frecuencia relativa mayores que 0.5 (con un valor máximo teórico de $log_2(3) \simeq$=1.584 en el caso de un contacto totalmente conservado con una frecuencia relativa de 1). Para poder analizar en qué regiones de la estructura de las ANKs estas interacciones mínimamente frustradas con altos valores de $FrustrationIC$ pesado por su frecuencia relativa están localizadas, tomamos aquellas interacciones mínimamente frustradas con valores mayores que el valor más alto correspondiente a la distribución de las interacciones neutras (Tabla 3.1). Estas interacciones conectan principalmente aquellos residuos que están más conservados a nivel de secuencia (que poseen mayores valores de $SeqIC$) y que son importantes para la estabilidad interna de las repeticiones (Fig. 3.13B) como así también para la estabilización de las interfaces entre repeticiones vecinas, constituyendo una red de interacciones mínimamente frustradas (Fig. 3.13C). Esta red está compuesta por 15 interacciones que son internas a cada repetición y 8 interacciones entre repeticiones vecinas. El motivo TPLH está involucrado en varias de las interacciones. Es interesante, que algunas de estas interacciones conservadas se establecen entre residuos que se encuentran en las hélices y

residuos que componen el *beta*-hairpin. Una mayoría de estas interacciones involucran residuos hidrofóbicos (Tabla 3.1).

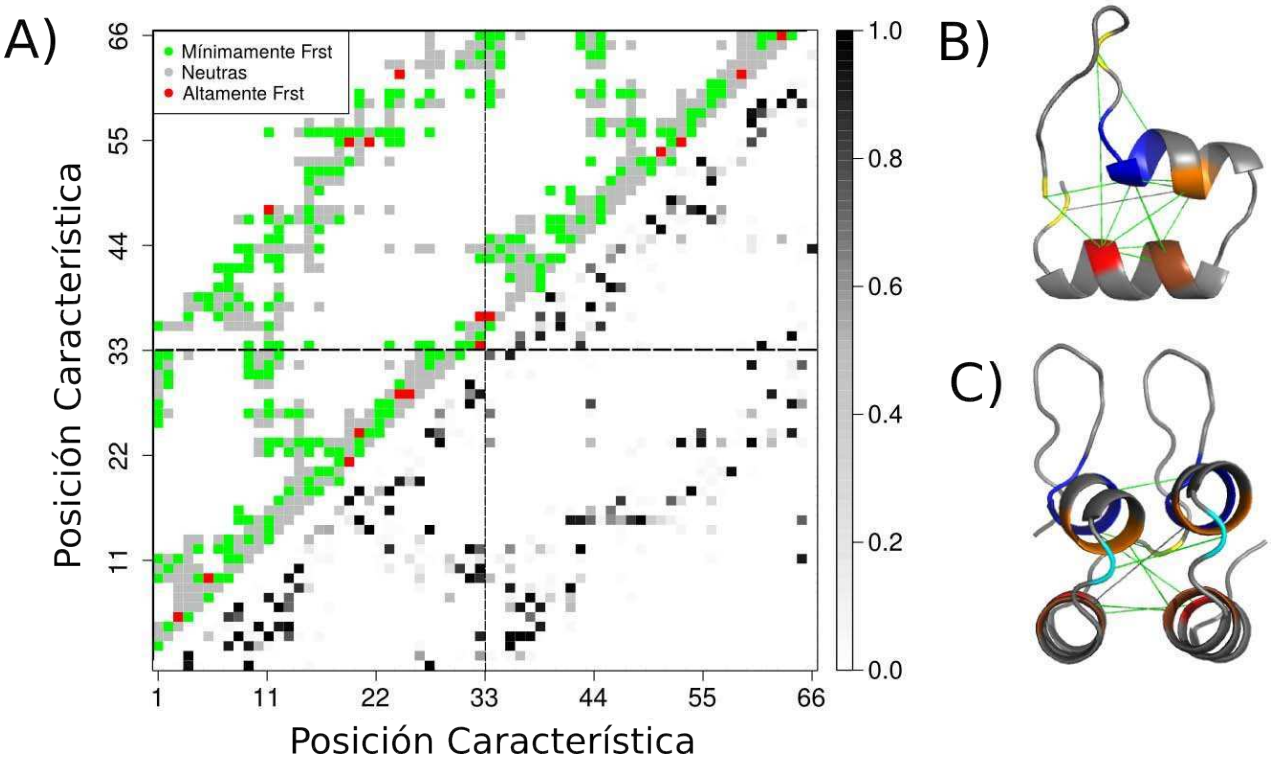


**Figura 3.13:** Conservación de frustración en mapas de contacto: A) Mapa de contacto para pares de repeticiones ANK. En la matriz triangular superior se muestra para cada contacto dentro del par, el estado más informativo según su contribución al $FrustrationIC$. Rojo representa que el estado más informativo corresponde al altamente frustrado, gris es para el estado neutro y verde para el caso en que el estado mínimamente frustrado es el más informativo. En la matriz triangular inferior se muestra en escala de grises la abundancia relativa de un contacto entre cada par de posiciones canónicas en el par de repeticiones. B) Conservación de interacciones intra repetición: Contactos dentro de las repeticiones ANK con un valor de contenido de información mayor al valor máximo dentro de las interacciones neutras. Residuos involucrados en esas interacciones conservadas se marcan según los siguientes colores: Amarillo: posición 8; G. Azul: posiciones 11-14;TPLH. Naranja: posiciones 15-16; AA. Rojo: posiciones 23-24; IV. C) Conservación de contactos en las interfaces entre pares de repeticiones

## Frustración local: El balance entre la estabilidad y la función

Los estados nativos de la mayoría de las proteínas son marginalmente estables, estando separados del estado desplegado por no más de 5-15 kcal $mol^{-1}$ [Dill, 1990]. A pesar de tener paisajes energéticos que han sido moldeados por la evolución para satisfacer el principio de mínima frustración, algunos conflictos energéticos son mantenidos en los ensambles nativos en donde se supone, son importantes para la función de estas moléculas [Ferreiro et al., 2007b,Ferreiro et al., 2011,Ferreiro et al., 2014]. Hemos calculado el índice de frustración configuracional para todo el conjunto de datos no redundante y se analizaron las distribuciones del mismo

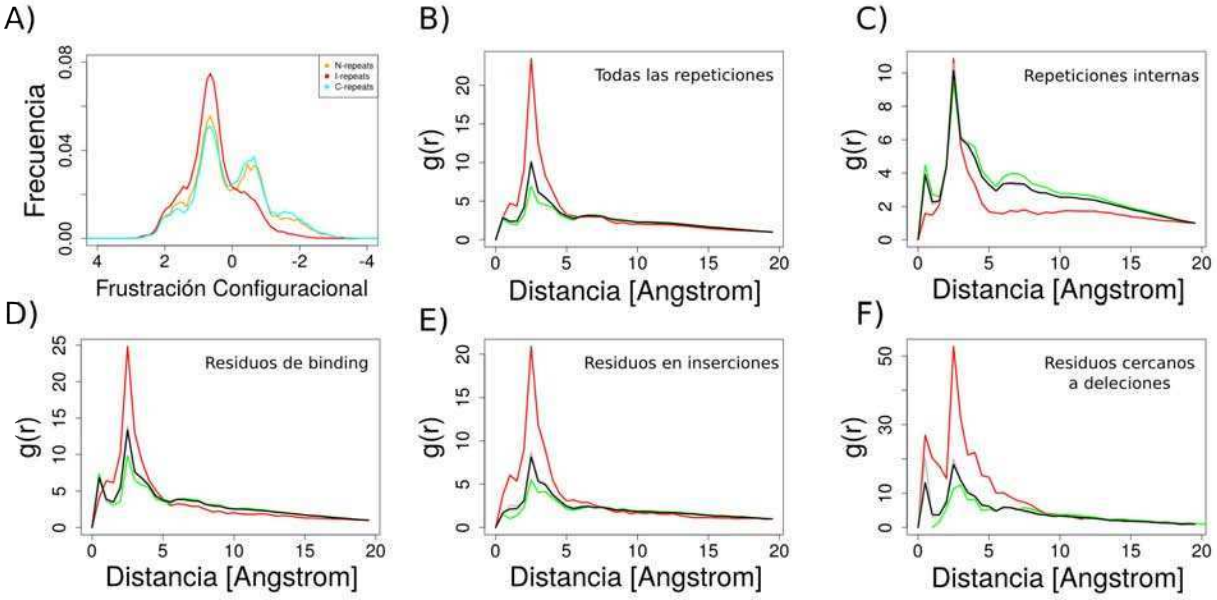


**Figura 3.14:** Distribución de frustración en las repeticiones ANK: (A) Indice de frustración configuracional calculado de forma separada para las repeticiones N-terminales (naranja), internas (rojas) y repeticiones C-terminales (cian), observamos que las repeticiones internas difieren en cuanto a sus distribuciones del índice de frustración mencionado respecto de las repeticiones terminales. (B) Función de distribución de pares g(r) calculada para aquellos residuos incluidos entre la primera y la última repetición detectadas a partir de las estructuras. (C) Función de distribución de pares g(r) calculada a partir de tomar sólo las repeticiones internas. (D) Función de distribución de pares g(r) calculada sobre los residuos que están en contacto con proteínas co-cristalizadas en complejos. (E) Función de distribución de pares g(r) calculada sobre los residuos que pertenecen a inserciones presentes en las repeticiones ANK. (F) Función de distribución de pares g(r) calculada sobre los residuos que se encuentran localizados inmediatamente antes y después de los puntos en donde se detectaron deleciones en las repeticiones ANK

para los 3 diferentes tipos de repeticiones (N-terminales, internas y C-terminales).

Se observan claras diferencias entre las repeticiones de los extremos y las internas (Fig. 3.14A). El índice de frustración se compone de dos términos, uno correspondiente a los contactos y otro a la interacción de los residuos con el solvente, siendo el último el que contribuye en mayor medida a las diferencias observadas. La función distribución de pares (*pair distribution function* o g(r)) se usó para calcular la ocurrencia de *clusters* de contactos de los diferentes niveles de frustración en las estructuras de ANKs. Cuando se analiza la función g(r), se puede ver que hay un enriquecimiento de *clusters* de interacciones frustradas cuando se analiza los arreglos repetitivos enteros (Fig. 3.14B) respecto del análisis de los arreglos sin tomar en cuenta las repeticiones de los extremos (Fig. 3.14C), lo cual indica que las repeticiones de los extremos son las responsables de este enriquecimiento observado en cuanto a interacciones altamente frustradas. Las repeticiones de los extremos tienen una mayor superficie expuesta al solvente comparado con las internas. El enriquecimiento en interacciones altamente frustradas

está en concordancia con lo anterior y aunque los perfiles de secuencia de las repeticiones terminales parecen haber divergido, probablemente en gran medida para poder hacer frente a esta mayor exposición, todavía hay varios conflictos energéticos en estas regiones respecto de la interacción de los residuos con el solvente que quedan sin ser resueltos. Tal vez estos conflictos fueron mantenidos evolutivamente para contribuir a la función biológica.

Sin embargo, las repeticiones terminales no son la única región de los arreglos repetitivos de ANKs que están enriquecidos en interacciones altamente frustradas, los sitios de unión a otras proteínas (Fig. 3.14D) (anotadas de acuerdo a los contactos entre cadenas, presentes en los co-cristales de ANKs con otras proteínas) como así también los residuos anotados como parte de inserciones (Fig. 3.14E), se encuentran enriquecidos en este tipo de interacciones. Por último, y no del todo esperable, también pudimos observar que los residuos que rodean aquellos lugares en donde se anotaron las posiciones canónicas deletadas en las repeticiones, también se encuentran enriquecidas en interacciones altamente frustradas (Fig. 3.14F).

## 3.6. Interacciones Proteína-Proteína en los arreglos repetitivos de ANKs

Para poder analizar cómo las ANKs interactúan con otras moléculas proteicas, seleccionamos aquellas que se encuentran cristalizadas en complejo. En total, hay 83 estructuras de ANKs co-cristalizadas, tomando en cuenta homoligómeros y heteroligómeros (basados en la asignación de unidades presentes en las anotaciones de los archivos PDB). De igual manera que para las cadenas simples, seleccionamos un grupo de complejos para definir un conjunto no redundante. Una entrada pertenece a este conjunto si no hay otro complejo que contenga una cadena con el mismo identificador de Uniprot o sí, en caso de existir, la misma está en complejo con un compañero diferente. Bajo estas condiciones, un total de 34 complejos son incluidos en el conjunto no redundante y usados para los posteriores análisis (Tabla 3.3).

Podemos observar que las proteínas del conjunto involucran en promedio alrededor del 20 % (sd=11 %) de sus residuos en unir a sus interactores. De estos residuos, 80 % (sd=0.24 %) corresponden a posiciones canónicas dentro de las repeticiones, 15 % (sd=0.19 %) corresponden

a inserciones ya sea dentro o entre las repeticiones y sólo un 5 % pueden ser mapeadas a regiones no repetitivas.

En la literatura se describe generalmente que los $\beta$-hairpins constituyen la principal región de las ANKs que median las interacciones con otras proteínas [Sedgwick and Smerdon, 1999]. En la Fig. 3.15 se muestra la distribución de densidad de contactos en los marcos canónicos de las repeticiones N-terminales, internas y C-terminales. Cuando se mapean las interacciones cristalográficas a los residuos canónicos de las repeticiones internas se observa un relativo enriquecimiento en los $\beta$-hairpins, como se describe en la bibliografía [Mosavi et al., 2002].

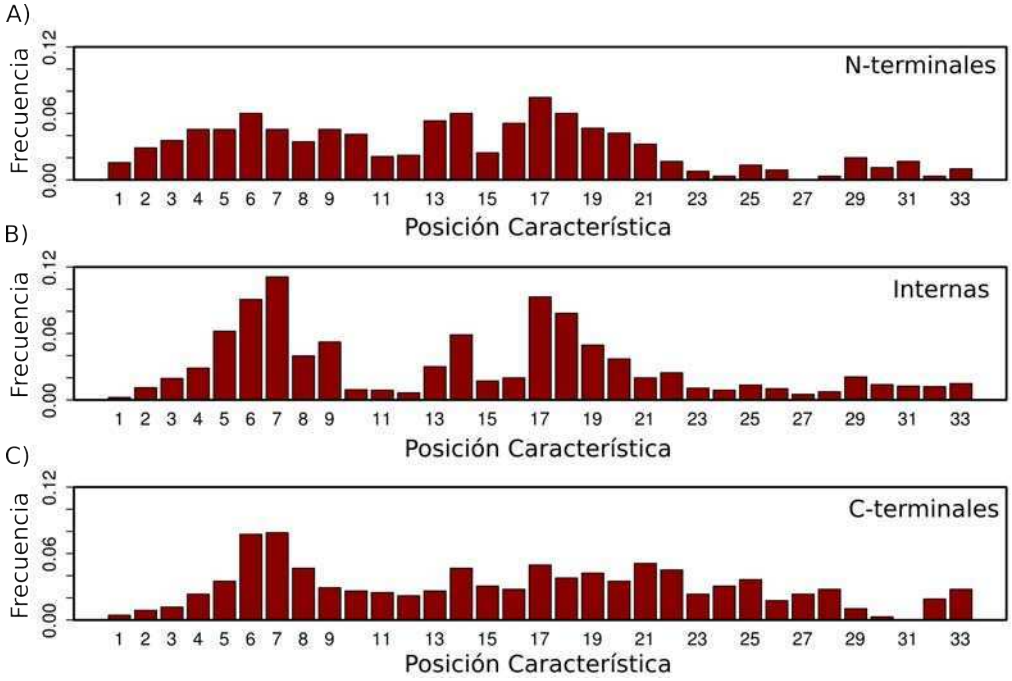


**Figura 3.15:** Perfiles de contactos de interacción proteína−proteína: Calculamos las abundancias relativas de contactos entre proteínas que afectan a cada posición canónica en los diferentes tipos de repeticiones A) Repeticiones N-terminales B) Repeticiones Internas C) Repeticiones C-terminales.

Sin embargo, la región inter hélices también se muestra altamente involucrada en la interacción proteína-proteína. En algún nivel, específica o inespecíficamente, todo el marco canónico de las repeticiones ANK, se ve involucrado en la interacción proteína-proteína. Para el caso de las repeticiones terminales se observa una distribución de contactos más homogénea. En el caso de la repetición N-terminal se observa una mayor densidad de contactos en su mitad N-terminal (residuos 1-21 que incluye la región de loop inicial y la primer hélice). En la repetición C-terminal se observa algo similar aunque hay una mayor densidad de contactos en los residuos 21-29, correspondientes a la segunda hélice.

Este mecanismo no conservado de las ANKs para unir otras proteínas se refleja en las difer-

entes formas que estas proteínas pueden interactuar con sus ligandos (Fig. 3.16A-D). Además de la falta de un motivo estructural conservado en la estructura canónica de las repeticiones para interactuar con sus ligandos, estas moléculas usualmente involucran regiones no repetitivas en la interacción como es el caso del complejo entre la Myosin Phosphatase targeting subunit 1 (MYPT1, PdbID=1s70) y la Ser/Thr Phosphatase-1 (delta) (Fig. 3.16A) donde una hélice se encuentra conectada al arreglo repetitivo a través de un loop ausente de estructura secundaria alfa o beta evidente, que interactúa con NF-$\kappa$B. Una proteína relacionada, I$\kappa$B$\alpha$, también contiene una región de loop que es usada para unir el heterodímero P50/P56 de NF-$\kappa$B, el cual se encuentra desordenado sin una estructura definida hasta que se produce el proceso de reconocimiento y unión. En el caso de I$\kappa$B$\beta$ todas las repeticiones se encuentran involucradas en la interacción con la otra molécula. En otras ANKs como MYPT1 o en el caso de la proteína YAR1 que une la proteína 40S ribosomal S3 (PdbID=4bsz) (Fig. 3.16C), la interfaz se encuentra distribuida de forma heterogénea a lo largo de las repeticiones. Las ANKs también pueden conformar homo-complejos que involucran diferentes tipos de interfaces donde uno de los ejemplos más interesantes es el del homodímero de Tankyrasa 1 (PdbID=3utm) (Fig. 3.16D). Esta molécula es capaz de formar un homodímero donde los monómeros se encuentran entrelazados en forma cruzada involucrando las repeticiones centrales de cada uno, que poseen hélices extendidas debido a inserciones. Previamente, mostramos que hay enriquecimiento de interacciones frustradas en sitios de unión a otras proteína. Para ir un paso más adelante en esto, comparamos el índice de frustración configuracional calculado tanto en el estado aislado de los monómeros ANK como en el estado en complejo (Fig. 3.16E). Las distribuciones de frustración para ambos estados son similares, compatible con paisajes energéticos mínimamente frustrados. Pero, aunque las distribuciones sean indistinguibles de forma global, la frustración local al analizar los contactos individuales muestran cambios en muchos casos. Comparamos el cambio de frustración para aquellos contactos que se encuentran involucrados en las interfaces de interacción con sus ligandos en los co-cristales disponibles. Observamos que si consideramos aquellos contactos con cambios de frustración mayores en valor absoluto que 0.3 (es decir todo aquello que se encuentra fuera de la tendencia observada alrededor de cero), la mayoría de ellos, $\tilde{6}0\,\%$, cambian hacia valores menores de frustración concentrándose en diferencias de alrededor de $\sim$1.5 unidades de frustración. Este cambio es

contribuido mayormente por el cambio de accesibilidad al solvente asociado al proceso de unión que es capturado por el termino de burial de la función de energía AMW, usada para calcular el índice de frustración. El $\tilde{4}0\,\%$ restante de los contactos que varían considerablemente, cambian en la dirección opuesta. Observamos que mientras gran parte de la frustración presente en los estados aislados es liberada en el estado en complejo, algunas interacciones altamente frustradas nuevas aparecen como producto de la formación de la interfaz. Es tentador especular que esta nueva frustración que aparece como consecuencia del proceso de oligomerización tiene consecuencias funcionales para transiciones conformacionales posteriores una vez que los complejos se encuentran ensamblados y no significa sólo un límite para la evolución en la minimización de la frustración total presente en el paisaje energético.

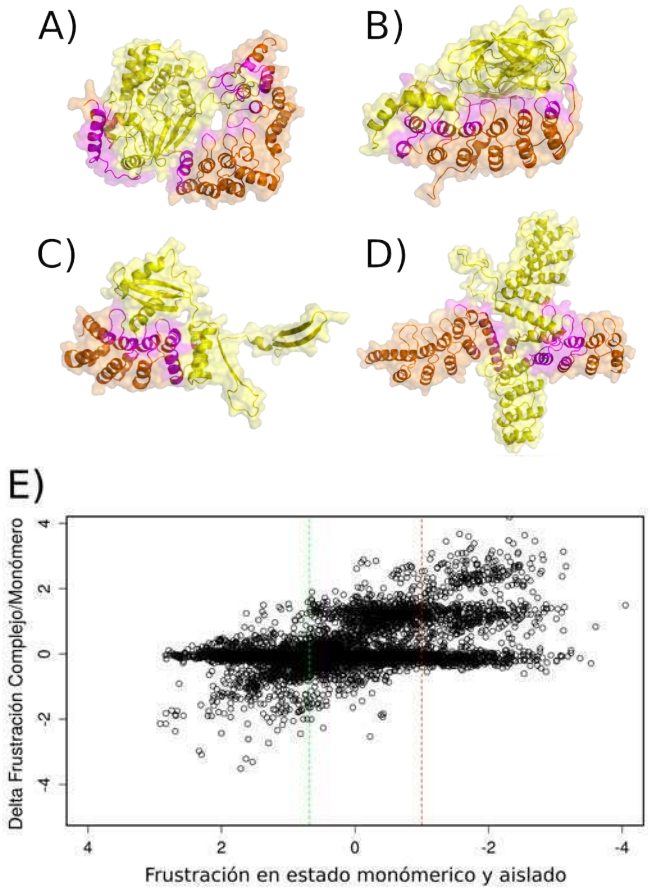


**Figura 3.16:** Complejos cuaternarios que involucran ANKs se muestran con la molécula ANK en color naranja y el resto en amarillo. Los residuos de la estructura ANK que están involucrados en la interfaz se muestran en color magenta. A) Myosin phosphatase targeting subunit 1 (MYPT1) en complejo con ser/thr phosphatase-1 (delta) (PdbID=1s70). B) Complejo de I$\kappa$Bb/NF-$\kappa$B p65x2 (PdbID=1k3z) C) YAR1 uniendo 40S ribosomal protein S3 (PdbID=4bsz) D) Tankyrase-1, (PdbID=3utm). E) Cambio en el índice de frustración configuracional entre el estado aislado y en complejo de las estructuras de proteínas ANK. En eje x muestra el valor de frustración configuracional para el estado monomérico mientras que el eje y muestra la diferencia de dicho índice entre el estado en complejo y el monomérico. Un valor positivo significa que el valor de frustración disminuyo en el estado en complejo respecto del monomérico y viceversa.

## 3.7. Mejores modelos de secuencia

La detección de repeticiones usando HMMs como los que se encuentran disponibles en las bases de datos como Pfam, son incapaces de detectar todas las repeticiones presentes en una secuencia, incluso cuando su presencia es obvia al observar la estructura correspondiente. Adicionalmente, para todas aquellas repeticiones que divergen más allá de cierto punto respecto de la secuencia consenso del alineamiento de secuencias, es común que las detecciones de las diferentes instancias sea incompleta. El principal problema es que estos modelos se construyen a partir de un alineamiento semilla que luego es sucesivamente mejorado mediante la inclusión de nuevas instancias detectadas usando los HMMs iniciales. Debido a que en general, todas las bases de datos del estilo Pfam, son pensadas para dominios globulares, las repeticiones tienen el problema de que necesitan ser cortadas primero para poder generar un alineamiento de buena calidad. Estos cortes y alineamientos se han hecho históricamente de forma ad-hoc, de forma que el espacio de secuencias que representan se encuentra incompleto y los alineamientos no son de buena calidad.

Como una consecuencia de esto, las bases de datos que poseen anotaciones de proteínas repetitivas contienen usualmente muchas instancias no anotadas debido a la gran divergencia en secuencia que caracteriza a estas moléculas. Nuestro objetivo en esta sección es mostrar como nuestros HMMs derivados a partir de la detección estructural de las repeticiones ANK (modelo C-terminal, interno y N-terminal) son útiles para mejorar la detección de las repeticiones al nivel de secuencia. Utilizamos el módulo hmmsearch de HMMER para buscar instancias correspondientes a los diferentes HMMs sobre las secuencias correspondientes a nuestro conjunto no redundante de ANKs que poseen estructuras. Para las búsquedas se usaron los HMMs de la siguiente manera a) El*HMM de Pfam* de forma individual b) El *HMM estructural general* derivado del alineamiento de todas las repeticiones detectadas con el método de teselado c) Los HMMs correspondientes a las repeticiones N-terminal, Internos y C-Terminal usados de forma individual con una posterior combinación de sus detecciones. Para cada caso comparamos la detección realizada con los diferentes HMMs con la detección hecha a partir del teselado cruzado usando la repetición interna de 4ANKs, como se describió anteriormente. Para cada búsqueda usando hmmsearch y los HMMs correspondientes calculamos el coeficiente

de correlación de Matthews (*Mathews Correlation Coefficient*, MCC) para evaluar la calidad de las detecciones cuya fórmula es: $MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$, pudiendo tomar valores entre -1 y 1. TP hace referencia a los verdaderos positivos, detecciones hechas por hmmsearch con un determinado HMM y que coincide con las detecciones consideradas verdaderas. TN hace referencia a los verdaderos negativos, regiones de las proteínas no detectadas por hmmsearch y que verdaderamente no son repeticiones. FP corresponde a los falsos positivos y son aquellas regiones detectadas por hmmsearch sin ser verdaderas repeticiones. Por último FN hace referencia a los falsos negativos y son aquellas repeticiones no detectadas como verdaderas pero que si lo son. Nuestra evaluación se hizo al nivel de residuos que forman parte de las repeticiones detectadas mediante el método de teselado. Aquellos residuos detectados por el método de teselado como parte de una repetición estructural se consideran los verdaderos casos positivos. Al realizar detecciones con hmmsearch, se obtienen los límites de las regiones detectadas como positivas y los FP, FN, TP y TN se definen acorde a como los mismos se condicen con los correspondientes a la detección estructural.

Al calcular los MCCs para los diferentes modelos encontramos que el HMM Estructural General (MCC=0.473) no produce mejores detecciones que el HMM de Pfam (MCC=0.508). Sin embargo, si realizamos detecciones usando los HMMs específicos para cada tipo de repetición observamos que el correspondiente a la repetición interna (MCC=0.586) es mejor que el HMM General y que el HMM de Pfam. Por su parte al usar los HMMs de las repeticiones N-terminal (MCC=-0.069) y C-terminal (MCC=0.407), estos no son mejores que los anteriores. Sin embargo, hay que tener en cuenta que estamos aplicando HMMs específicos para las repeticiones terminales para evaluar detecciones sobre todas las repeticiones. Si en cambio combinamos las detecciones individuales de los HMMs específicos por repetición observamos que la detección total de residuos detectados como pertenecientes a repeticiones totales es mejor que las predicciones de los demás HMMs (MCC=0.605). Es verdad que estamos usando los modelos derivados de las detecciones estructurales para realizar la detección de las repeticiones a partir de las cuales hemos construido los modelos y esto pone en duda si en verdad la mejora se debe a una mejor representación de las repeticiones ANK o a un artefacto de sesgo por autoconsistencia. Debido a que los casos correspondientes a las verdaderas repeticiones estructurales son pocos, no consideramos confiable hacer una división del conjunto de

datos para construcción de los modelos y evaluación de forma independiente. Debido a que el objetivo de esta tesis no se encuentra centrado en la detección de repeticiones en secuencia, consideramos la evaluación sistemática de las mejoras usando HMMs derivados de esta forma como parte de las perspectivas a futuro.

Sin embargo, es posible analizar algunos casos puntuales. En la bibliografía disponible se reporta que las ANKs poseen en general 5-6 repeticiones con un máximo de 60 repeticiones y un mínimo de 2. Sin embargo, no existen proteínas disponibles con más de 12 repeticiones cuya estructura sea conocida. Axton y sus colaboradores reportaron tiempo atrás [Axton et al., 1994] que una proteína llamada Plutonium perteneciente a *Drosophila*, era una ANK inusual que poseía únicamente 2 repeticiones. Si se utiliza hmmsearch para hacer una detección de repeticiones ANK sobre la secuencia de Plutonium usando el HMM de Pfam, en efecto sólo se encuentran 2 repeticiones, lo cual es además consistente con las anotaciones presentes en Uniprot para dicha proteína (UniprotID P42570) en donde las repeticiones son detectadas usando la herramienta REP [Andrade et al., 2000]. Al usar hmmsearch junto al HMM estructural general es posible detectar 4 regiones compatibles con el modelo aunque las repeticiones detectadas en el extremo N-terminal y C-terminal poseen bajos puntajes. Construimos un modelo por homología para la proteína Plutonium (Fig. 3.17) usando el servidor Phyre2 [Kelley et al., 2015] usando valores por defecto. El modelo recuperado, el cual uso la estructura 3kea,B como templado, es una estructura compatible con una ANK, la cual contiene 5 repeticiones ANK. Las predicciones de estructura secundaria hechas por el servidor sobre la secuencia, son compatibles con la estructura canónica de repeticiones ANK. Nuestra intención con este análisis es mostrar que los métodos tradicionales basados en secuencia para la detección de repeticiones presentan grandes problemas, no sólo por sus limitaciones técnicas sino además históricas debido a la materia prima a partir de la cual se generaron los modelos estadísticos que representan a las repeticiones. Dados nuestros análisis creemos que lo más lógico es que la proteína Plutonium tiene más de 2 repeticiones dado a que una estructura de menos de 3 repeticiones no ha sido observada y se estima que no sería estable [Mosavi et al., 2002] y es en realidad un problema de mala detección con los métodos tradicionales que ha sido reportada como poseedora de solo 2 repeticiones. Creemos que la metodología desarrollada por nosotros en este trabajo de tesis puede ser muy útil de manera general para

derivar mejores HMMs para las diferentes familias repetitivas y así mejorar las detecciones basadas en secuencias.

**Figura 3.17:** Modelo por homología de la proteína Plutonium de *Drosophila melanogaster*. En amarillo y naranja se muestran las regiones detectadas usando el HMM de Pfam. En rojo se muestran además las regiones detectadas de forma adicional por el HMM estructural general. La quinta repetición no es detectada por ninguno de los modelos

.

## 3.8. Conclusiones del capítulo

Para poder caracterizar a los miembros de una familia repetitiva como la familia ANK, es necesario ser capaces de generar una detección y anotación consistente de las repeticiones que los componen. La alta divergencia a nivel de secuencias entre las repeticiones ANK constituye un problema cuando se intenta anotarlas usando métodos basados en secuencia ya que en muchos casos las repeticiones son identificadas de forma incompleta o son completamente no detectadas. En este capítulo aplicamos el método de teselado [Parra et al., 2013] que desarrollamos y explicamos en el capítulo anterior para analizar las periodicidades en las proteínas ANK y encontrar sus repeticiones estructurales. Hemos combinado el método de teselado, que evalúa de forma exhaustiva que tan buenos son los fragmentos de una proteína para cubrir la estructura total mediante copias de sí mismos, con la función de plegabilidad

relativa [Panchenko et al., 1996] aplicada a dichos fragmentos para definir de forma consistente el largo y fase de las repeticiones ANK. La detección estructural de las repeticiones nos permitió detectar instancias que eran detectadas de forma incompleta usando métodos basados en secuencia, como así también otros métodos basados en estructura. De estos últimos AnkPred [Chakrabarty and Parekh, 2014] falla en detectar por ejemplo la sexta repetición de I$\kappa$B-$\alpha$ (PdbID: 1ikn,D) o las repeticiones internas de la proteína K1 del virus Vaccinia (PdbID: 3kea,B). El método Console [Hrabe and Godzik, 2014] por su parte, es capaz de detectar las 6 repeticiones presentes en I$\kappa$B-$\alpha$ y casi todas las repeticiones presentes en la proteína K1 pero no es posible fijar en el programa que fase deseamos para dicha detección y en cambio cada proteína tiene una fase dependiente de su propia estructura. Esta incapacidad para definir una fase determinada en la detección constituye un obstáculo para realizar estudios comparativos. Con nuestro procedimiento, todas las repeticiones de las estructuras presentes en nuestro conjunto no redundante (Tabla 3.1), fueron consistentemente anotadas, junto con sus correspondientes inserciones y deleciones. Los perfiles de secuencia obtenidos de esta manera, basados en alineamientos de secuencia guiados por los alineamientos estructurales, ofrecen una nueva perspectiva acerca de la divergencia en secuencia que pueden tolerar las repeticiones ANK. Aún más, estos perfiles pueden ser usados, en combinación con los ya existentes para mejorar la anotación y detección de las repeticiones al nivel de sus secuencias mejorando la cobertura de las mismas en bases de datos como Pfam y Uniprot. Además nuestras detecciones constituyen un aporte para bases de datos como RepeatsDB [Di Domenico et al., 2013] que nuclea las anotaciones y caracterizaciones de la mayor cantidad de proteínas repetitivas hasta el momento.

Hemos mostrado que la población de repeticiones ANK puede dividirse en 3 grupos, aquellas repeticiones situadas en el extremo N-terminal de los arreglos repetitivos, aquellas en el extremo C-terminal y aquellas que se encuentran en medio de las primeras dos, llamadas internas. Estos 3 tipos de repeticiones exhiben diferentes firmas tanto en sus secuencias como en la energética de sus estructuras y su composición de elementos de estructura secundaria. Las repeticiones internas son las que muestran los mayores niveles de conservación tanto a nivel de secuencias como de la energía estructural mientras que las repeticiones en los extremos N-terminales son las menos conservadas en esos aspectos. Si la conservación en el nivel

de secuencias es comparada con la conservación de los patrones de frustración local al nivel estructural, observamos que existe una correlación positiva y lineal entre ellas para las repeticiones internas y aquellas situadas en el extremo C-terminal mientras que dicha correlación no ocurre para aquellas localizadas en el extremo N-terminal. Esta correlación sugiere que mientras más similares a la secuencia consenso son las secuencias de las repeticiones esta será más plegable. Las mutaciones consenso han mostrado ser útiles para estabilizar proteínas [Steipe et al., 1994]. Mutaciones hacia secuencias consenso con efecto desestabilizante han sido observadas en aquellas posiciones que son altamente co-variantes o son invariantes y entonces pueden ocurrir correlaciones no detectables [Sullivan et al., 2012, Ferreiro et al., 2007a, Yang et al., 1995]. Hemos calculado cuales interacciones dentro del marco canónico de las estructuras de pares de repeticiones son las más conservadas y encontramos que hay un conjunto de residuos que se encuentran altamente conservados en secuencia que a su vez se encuentran conectados por una red de interacciones conservadas y mínimamente frustradas (Tabla 3.2). La mayoría de estas interacciones se establecen entre residuos hidrofóbicos. Interacciones no hidrofóbicas incluyen el motivo TPLH en interacción con el $\beta$-hairpin de su misma repetición o con interacciones con los loops de repeticiones vecinas. También se observan interacciones entre motivos TPLH adyacentes lo cual refuerza la importancia de este motivo en la estabilización intra e inter repeticiones ANK [Guo et al., 2010]

Algo muy sorprendente en las repeticiones ANK es que las interacciones estabilizantes dentro y entre repeticiones ANK se encuentran codificadas en la secuencia consenso de la repetición individual, la cual es compatible para interaccionar con copias de sí misma. Esto se evidencia en el éxito del diseño por consenso, apilando repeticiones idénticas lo cual resulta en estructuras plegables y altamente estables [Mosavi et al., 2002, Binz et al., 2003]. La cooperatividad de plegado de las proteínas repetitivas está altamente influenciada por las estabilidades intrínsecas de las diferentes repeticiones y sus interfaces. Ha sido mostrado computacionalmente que estas proteínas poseen un fino balance entre las energías de interacción intra e inter repeticiones que les permite desplegarse parcialmente bajo condiciones fisiológicas, lo cual sería un requerimiento para su función biológica [Ferreiro et al., 2005, Ferreiro et al., 2008]. Hemos mapeado las interacciones entre residuos canónicos que se encuentran más favorecidas energéticamente. Mapear cuales de esas interacciones dentro y entre repeticiones no se satis-

facen en arreglos naturales de repeticiones ANK nos ayudaría a desentrañar los determinantes de los comportamientos diferenciales en cuanto al plegado de diferentes proteínas ANK, que a pesar de tener el mismo número de repeticiones y ser muy similares al nivel estructural, poseen propiedades dinámicas sustancialmente diferentes.

Las repeticiones ANK pueden soportar una gran cantidad de modificaciones (inserciones o deleciones) en su marco canónico de 33 residuos de largo. Nuestros análisis sobre la energética de las interacciones que se encuentran próximas a dichas modificaciones muestran que existe un enriquecimiento de interacciones altamente frustradas alrededor de las mismas. Esto sugiere que las inserciones y deleciones que ocurren en las repeticiones ANK pueden tener consecuencias funcionales para la estructura en forma global, esculpiendo el paisaje energético de las mismas ya sea favoreciendo transiciones conformacionales o de forma indirecta la interacción con ligandos. Hemos visto además que los sitios de unión, identificados a partir de co-cristales muestran también un enriquecimiento en interacciones altamente frustradas, la cual es compensada al analizar los patrones de frustración de los complejos.

Las ANKs se encuentran adaptadas para llevar a cabo su principal función que es unir otras proteínas. Sus secuencias y estructuras pueden variar sustancialmente para maximizar sus propiedades de reconocimiento, introduciendo desviaciones estructurales considerables y mostrando incluso transiciones orden/desorden en muchos casos. Su modularidad les permite un ajuste exquisito al nivel de repeticiones individuales proveyéndoles propiedades dinámicas diferenciales a distintas regiones de los arreglos repetitivos. La presencia de interacciones altamente frustradas en los sitios de unión, inserciones y deleciones muestra que en muchos casos, la evolución parece haber mantenido estos conflictos energéticos que desestabilizan las estructuras pertenecientes a estas moléculas. Esto sería crítico para guiar el proceso de reconocimiento de los interactores liberando dicha frustración al establecer interacciones favorables una vez formado el complejo. Es notable que la distribución total de los valores de frustración local de las proteínas aisladas o en complejo, no muestran diferencias evidentes. Eso sugiere que si bien hay una proporción de interacciones que estaban altamente frustradas que se ven minimizadas, otras interacciones en diferentes partes de la molécula aumentan sus niveles de frustración de forma compensatoria lo cual puede ser necesario para posteriores transiciones conformacionales del complejo. Las ANKs combinan de forma estratégica la in-

troducción de perturbaciones estructurales en residuos claves dentro de la estructura canónica de las repeticiones, manteniendo otras invariantes. Esta calibración al nivel de la secuencia no sólo modifica la estructura global y modula la afinidad y especificidad para reconocer los ligandos, sino que también codifica complejos comportamientos dinámicos como la presencia de múltiples intermediarios de plegado o una mayor plasticidad conformacional que surge en consecuencia a dichas variaciones.

**Tabla 3.1:** Conjunto no redundante de estructuras pertenecientes a la familia ANK

| NumReps | Uniprot ID | PdbID | Largo Estructura | Organismo |
|---|---|---|---|---|
| 3 | DARPIN_NI1C_Mut4 | 2xen,A | 91 | Ninguno |
| 3 | DARPIN_3CA1A2N-OH | 2zgd,A | 106 | Ninguno |
| 3 | DARPIN_3CA1A2N | 2zgg,A | 88 | Ninguno |
| 3 | DARPIN_3ANK | 1n0q,A | 92 | Ninguno |
| 3 | P62775 | 1myo,A | 118 | Rattus norvegicus |
| 3 | Q8TDY4 | 3lvq,E | 257 | Homo sapiens |
| 3 | DARPIN_NRC | 2l6b,A | 106 | Ninguno |
| 3 | Q5ZXN6 | 4bet,A | 480 | Legionella pneumophila |
| 4 | DARPIN_H10_2_G3 | 2jab,A | 124 | Ninguno |
| 4 | DARPIN_1D5 | 2v4h,C | 125 | Ninguno |
| 4 | DARPIN_3H10 | 2v5q,C | 130 | Ninguno |
| 4 | DARPIN_4ANK | 1n0r,A | 126 | Ninguno |
| 4 | DARPIN_20 | 3hg0,D | 124 | Ninguno |
| 4 | P55273 | 1bd8,A | 156 | Homo sapiens |
| 4 | P55271 | 1d9s,A | 130 | Mus musculus |
| 4 | Q13625 | 1ycs,B | 193 | Homo sapiens |
| 4 | Q99728 | 3c5r,A | 122 | Homo sapiens |
| 4 | O22265 | 3deo,A | 183 | Arabidopsis thaliana |
| 4 | Q7SIG6 | 1dcq,A | 276 | Mus musculus |
| 4 | Q01705 | 1ymp,B | 135 | Mus musculus |
| 4 | Q8WUF5 | 2vge,A | 208 | Homo sapiens |
| 4 | P42771 | 1dc2,A | 156 | Homo sapiens |
| 4 | P62774 | 2kxp,C | 118 | Mus musculus |
| 4 | P58546 | 3aaa,C | 117 | Homo sapiens |
| 4 | Q15027 | 3jue,B | 298 | Homo sapiens |

| | | | | |
|---|---|---|---|---|
| 4 | P46683 | 4bsz,B | 150 | Saccharomyces cerevisiae |
| 5 | DARPIN_E3_5 | 1mj0,A | 156 | Ninguno |
| 5 | DARPIN_OFF7 | 1svx,A | 157 | Ninguno |
| 5 | DARPIN_E3_19 | 2bkg,A | 155 | Ninguno |
| 5 | DARPIN_AR_3A | 2bkk,B | 156 | Ninguno |
| 5 | DARPIN_AR_F8 | 2p2c,P | 158 | Ninguno |
| 5 | DARPIN_NI3C | 2qyj,A | 154 | Ninguno |
| 5 | DARPIN_NI3C_Mut5 | 2xee,A | 157 | Ninguno |
| 5 | Q60773 | 1ap7,A | 168 | Mus musculus |
| 5 | P42773 | 1bu9,A | 168 | Homo sapiens |
| 5 | P09959 | 1sw6,A | 301 | Saccharomyces cerevisiae |
| 5 | Q00420 | 1awc,B | 153 | Mus musculus |
| 5 | Q978J0 | 2rfm,A | 183 | Thermoplasma volcanium |
| 5 | Q92882 | 3ehq,A | 182 | Homo sapiens |
| 5 | Q13418 | 2kbx,A | 171 | Homo sapiens |
| 5 | Q9H2K2 | 3twq,A | 164 | Homo sapiens |
| 5 | Q9H9E1 | 3so8,A | 162 | Homo sapiens |
| 5 | O14593 | 3uxg,A | 163 | Homo sapiens |
| 5 | DARPIN_OR266 | 4gmr,A | 167 | Ninguno |
| 5 | DARPIN_OR264 | 4gpm,A | 157 | Ninguno |
| 5 | Q63ZY3 | 4hbd,A | 245 | Homo sapiens |
| 6 | P25963 | 1ikn,D | 221 | Homo sapiens |
| 6 | Q60778 | 1k3z,D | 258 | Mus musculus |
| 6 | P07207 | 1ot8,A | 238 | Drosophila melanogaster |
| 6 | P46531 | 2he0,A | 243 | Homo sapiens |
| 6 | O35433 | 2pnn,A | 248 | Rattus norvegicus |
| 6 | Q9DFS3 | 3jxi,A | 253 | Gallus gallus |
| 6 | Q9WUD2 | 2etc,A | 252 | Rattus norvegicus |

| | | | | |
|---|---|---|---|---|
| 6 | Q9Y5S1 | 2f37,A | 248 | Homo sapiens |
| 6 | Q91WD2 | 2rfa,A | 222 | Mus musculus |
| 6 | Q9HBA0 | 4dx1,A | 250 | Homo sapiens |
| 6 | Q838Q8 | 3hra,A | 199 | Enterococcus faecalis |
| 6 | Q96NW4 | 4b93,B | 231 | Homo sapiens |
| 7 | P50086 | 1wg0,A | 243 | Saccharomyces cerevisiae |
| 7 | P20749 | 1k1a,A | 228 | Homo sapiens |
| 7 | O75832 | 1tr4,A | 226 | Homo sapiens |
| 7 | Q9Z2X2 | 2dvw,A | 229 | Mus musculus |
| 7 | P14585 | 2fo1,E | 365 | Caenorhabditis elegans |
| 7 | Q9H9B1 | 3b7b,A | 234 | Homo sapiens |
| 7 | Q8IUH5 | 3eu9,A | 232 | Homo sapiens |
| 7 | Q90623 | 1s70,B | 291 | Gallus gallus |
| 7 | Q96DX5 | 3d9h,A | 235 | Homo sapiens |
| 7 | Q6IV60 | 3kea,B | 282 | Vaccinia virus |
| 7 | E9ADW8 | 3ljn,A | 304 | Leishmania major |
| 7 | Q5ZSV0 | 2aja,A | 347 | Legionella pneumophila |
| 7 | Q8K424 | 4n5q,B | 254 | Mus musculus |
| 9 | Q05823 | 4g8k,A | 307 | Homo sapiens |
| 10 | Q6PFX9 | 3utm,A | 320 | Mus musculus |
| 12 | P16157 | 1n11,A | 408 | Homo sapiens |

**Tabla 3.3:** Conjunto de datos no redundante de complejos ANK

| **PdbID** | **Estado Oligomérico** | **Composición Oligomérica** |
|---|---|---|
| 1BLX | 2 | heteroligómero |
| 1SW6 | 2 | homoligómero |
| 1YCS | 2 | heteroligómero |
| 2DVW | 2 | heteroligómero |

| | | |
|---|---|---|
| 1S70 | 2 | heteroligómero |
| 3IXE | 2 | heteroligómero |
| 3KEA | 2 | homoligómero |
| 1BI7 | 2 | heteroligómero |
| 3AAA | 2 | heteroligómero |
| 3AJI | 2 | heteroligómero |
| 3TWQ | 2 | homoligómero |
| 3UXG | 2 | heteroligómero |
| 2DZO | 2 | heteroligómero |
| 2HE0 | 2 | homoligómero |
| 1N0Q | 2 | homoligómero |
| 1SVX | 2 | heteroligómero |
| 2BKK | 2 | heteroligómero |
| 2V5Q | 2 | heteroligómero |
| 4BET | 2 | homoligómero |
| 1IKN | 3 | heteroligómero |
| 1G3N | 3 | heteroligómero |
| 1K3Z | 3 | heteroligómero |
| 3EU9 | 3 | homoligómero |
| 3UTM | 3 | heteroligómero |
| 2KXP | 3 | heteroligómero |
| 3ZKJ | 3 | heteroligómero |
| 2FO1 | 3 | heteroligómero |
| 1BI8 | 4 | heteroligómero |
| 1AWC | 4 | heteroligómero |
| 2V4H | 4 | heteroligómero |
| 3HG0 | 4 | heteroligómero |
| 2P2C | 6 | heteroligómero |

| | | |
|---|---|---|
| 4BSZ | 8 | heteroligómero |

**Tabla 3.2:** Red de contactos conservados y mínimamente frustrados en las repeticiones ANK

| **Intra repetición** | | | |
|---|---|---|---|
| **Posición i** | **Posición j** | **Aminoácido i** | **Aminoácido j** |
| 1 | **12** | N | **L** |
| 1 | 27 | N | L |
| 8 | **10** | G | **T** |
| 8 | **13** | G | **H** |
| **10** | 27 | **T** | L |
| **12** | 15 | **L** | A |
| **12** | 16 | **L** | A |
| **12** | 23 | **L** | V |
| **12** | 24 | **L** | V |
| **12** | 27 | **L** | L |
| 15 | 27 | A | L |
| 15 | 32 | A | A |
| 16 | 24 | A | V |
| 23 | 27 | V | L |
| 24 | 27 | V | L |
| **Entre repeticiones** | | | |
| **Repetición i** | **Repetición j** | **Aminoácido i** | **Aminoácido j** |
| **12** | 1 | **L** | V |
| 12 | 27 | L | L |
| **13** | **10** | **H** | **T** |
| 16 | 23 | A | V |
| 19 | 19 | G | G |
| 24 | 23 | V | V |
| 27 | 15 | L | A |
| 28 | 27 | L | L |

1

# Capítulo 4

# Explorando el paisaje energético de las proteínas ANK

## 4.1. Introducción

Hasta aquí hemos presentado métodos y estrategias para el análisis de periodicidades y repeticiones [Parra et al., 2013] de los miembros de la familia de proteínas con repeticiones de Ankirina (ANKs) y las hemos caracterizado estructural y energéticamente [Parra et al., 2015]. Presentamos por primera vez una estrategia objetiva y consistente para detectar las repeticiones de todas las proteínas ANK para las cuales se conocen sus estructuras de forma que pudimos realizar análisis comparativos sobre las mismas. Encontramos que las proteínas ANK están compuestas por 3 tipos de repeticiones diferentes, las situadas en el extremo N-terminal, las situadas en el extremo C-terminal y las internas. Estas moléculas tienen como principal función unir a otras proteínas sin que se conozca un mecanismo de reconocimiento conservado, lo cual es entendible dada la gran diversidad de moléculas con las que interactúan. Observamos que dada la estructura canónica de las repeticiones ANKs, todo aquello que se desvía de dicha canonicidad, se encuentra enriquecido en interacciones altamente frustradas, es decir, grupos de aminoácidos que se encuentran en conflicto con la estructura en la que están incluidos. Esta frustración local, se intuye como imprescindible para la función biológica de las ANKs debido a que sería fundamental para la excursión conformacional de estas moléculas dentro del ensamble de confórmeros que hacen a la función de las mismas, lo cual incluye

aquellas que son compatibles con la interacción con los ligandos apropiados.

En este capítulo presentaremos los resultados de aplicar métodos de simulación de dinámica molecular del tipo de grano grueso a varios miembros de la familia ANK, de diferentes largos y con una variedad de detalles estructurales que modifican la estructural global de las mismas. Los métodos aplicados corresponden a modelos del tipo G$\bar{o}$, que son métodos que evalúan la dinámica de una estructura proteica, basados en su topología, sin tener en cuenta su secuencia. Las simulaciones se realizaron usando la plataforma AWSEM-MD [Davtyan et al., 2012]. Aplicaremos diferentes tipos de análisis sobre las dinámicas corridas con modelos G$\bar{o}$ para extraer conclusiones acerca de los paisajes energéticos de varias proteínas ANK.

Para facilitar la notación y evitar redundancia, haremos referencia a las repeticiones en un arreglo de la forma *R1, R2, R3, .., Rn*; correspondiendo con las repeticiones *1, 2, 3, .., n* en el mismo.

## 4.2. Topología proteica y plegado

Como se ha comentado anteriormente, actualmente se entiende el plegado de proteínas bajo la teoría de paisajes energéticos en que una proteína posee un fuerte sesgo para plegarse hacia el estado nativo. Este sesgo es consecuencia de la cooperatividad existente entre las interacciones nativas dado que las secuencias de proteínas naturales cumplen con el principio de "mínima frustración". La morfología general del paisaje energético se asemeja a la de un embudo sobre la superficie del cual existen rugosidades debido a la frustración energética causada por conflictos entre los residuos que componen la cadena polipeptídica. Para poder plegarse en tiempo acotados y de forma robusta, la diferencia de energía entre los ensambles del estado nativo y el desplegado debe ser significativamente mayor que la rugosidad del embudo.

Sin embargo la rugosidad energética no es el único factor limitante para la plegabilidad de una secuencia en una estructura dada. Incluso si la rugosidad energética pudiera ser completamente removida del sistema, el paisaje energético no sería completamente liso. Tanto estudios teóricos [Wolynes, 1996, Nelson and Onuchic, 1998, Betancourt and Onuchic, 1995] como experimentales [Grantcharova et al., 1998, Martinez et al., 1998] indican que la estructura final de una proteína tiene un rol importante en la determinación de la plegabilidad de

la misma. De esta forma algunos motivos de plegado particulares podrían ser más diseñables que otros. El concepto de frustración topológica fue introducido para hacer referencia a la rugosidad del paisaje energético que es independiente de la frustración energética y depende sólo de la arquitectura de la proteína [Shea et al., 1999]. A pesar de que el nivel de frustración topológica puede ser modificado usando algunas estrategias [Plotkin and Onuchic, 2000], no puede eliminarse completamente reflejando que existe una dificultad intrínseca inherente a la plegabilidad de una estructura en particular. Además de seleccionar secuencias que poseen un bajo nivel de frustración energética, la evolución parece haber seleccionado aquellos motivos estructurales que minimizan la frustración topológica durante el proceso de plegado, filtrando aquellos motivos que serían muy difíciles de plegar [Betancourt and Onuchic, 1995, Wolynes, 1996].

Existen modelos para simular el plegado de una proteína en ausencia de frustración energética de forma de evaluar solamente el efecto de la topología en la dinámica del mismo. Estos modelos denominados de tipo G$\bar{o}$, debido a que se basan en el trabajo de modelos del tipo rejilla (*lattice*) ideados por Nobuhiro G$\bar{o}$ [Taketomi et al., 1975], han mostrado ser de gran utilidad para describir varios aspectos fundamentales del proceso de plegado [Oliveberg and Wolynes, 2005, Levy et al., 2005, Cho et al., 2008, Cho et al., 2009]. Estructuras pertenecientes a ensambles de estados de transición [Clementi et al., 2000], intermediarios de plegado [Shoemaker and Wolynes, 1999], mecanismos de dimerización [Levy et al., 2004] e intercambio de dominios [Yang et al., 2004] han sido predichos usando este tipo de modelos en los cuales se ha removido completamente la frustración energética siendo la información topológica de la estructura la única entrada.

Hemos aplicado un modelo del tipo G$\bar{o}$, llamado *Non-Additive AMH-G$\bar{o}$* sobre varios miembros de la familia ANK. Este modelo es un típico modelo libre de frustración energética en donde AMH hace referencia al *Associative Memory Hamiltonian*, desarrollado por el grupo del Prof. Peter Wolynes y *Non-Additive* hace referencia al uso del concepto de no-aditividad para modelar la interacción energética entre más de dos residuos [Eastwood and Wolynes, 2001]. Este modelo además es del tipo grano grueso (coarse-grained) ya que a diferencia de los modelos del tipo *all-atoms*, en que todos los átomos pesados son explícitamente modelados, aquí la unidad de simulación fundamental es el aminoácido, como una partícula esférica,

centrada en el carbono $\alpha$. Hemos simulado el proceso de plegado de diferentes ANKs tanto diseñadas como naturales y con diferentes números de repeticiones y describiremos como el número de repeticiones en los arreglos repetitivos, las asimetrías en las estructuras de los mismos y las perturbaciones estructurales introducidas por inserciones y deleciones afectan los mecanismos de plegado de estas proteínas. Hemos analizado un total de 11 proteínas con diferentes largos de arreglos repetitivos, desde un mínimo de 3 repeticiones y hasta un máximo de 6. La estructuras de las proteínas analizadas pueden observarse en la Fig. 4.1.

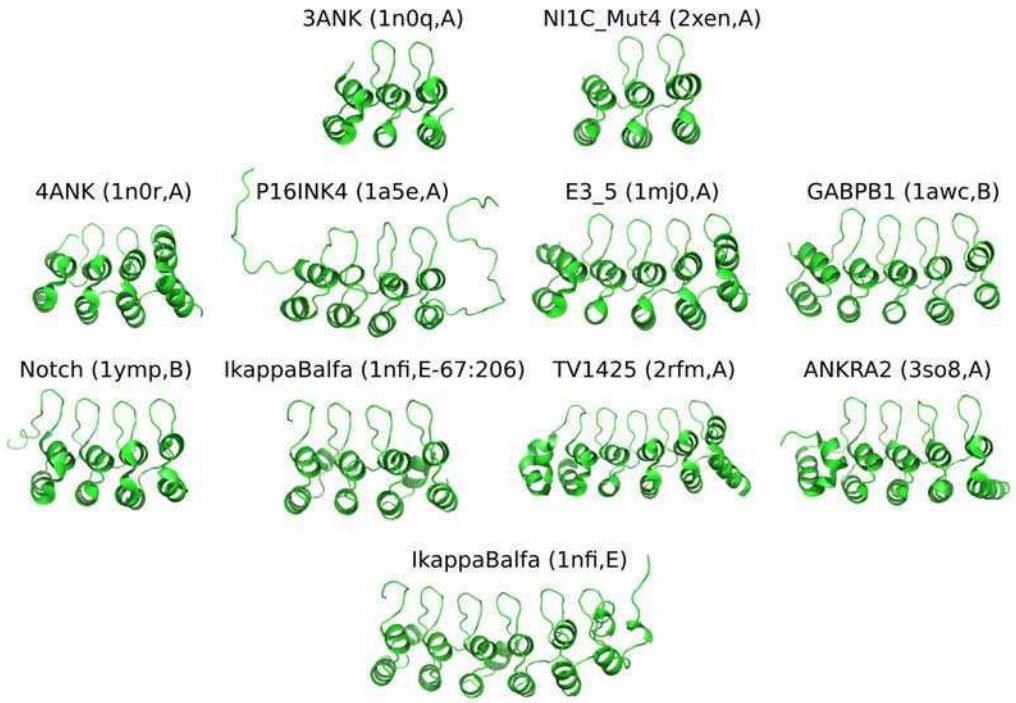


**Figura 4.1:** Estructuras representadas en forma de caricaturas de las proteínas analizadas en este capítulo. Se observan los nombres comunes usados para las mismas y en paréntesis los códigos PDB y la cadena de referencia.

## 4.3. 3 repeticiones

No existen hasta el momento estructuras de ANKs naturales que contengan 3 repeticiones. Existen en cambio estructuras correspondientes a ANKs diseñadas por diferentes grupos de investigación, siguiendo diferentes estrategias. Por un lado estudiamos la dinámica de plegado de 3ANK (1n0q,A) generada a partir de 3 repeticiones completamente consenso [Mosavi

et al., 2002]. Por otro lado también analizamos a $NI_1C$_Mut4 (2xen,A) una proteína diseñada siguiendo la metodología para generar DARPins del grupo de Plückthun, que consiste en una repetición consenso interna [Binz et al., 2003] y versiones modificadas del mismo para las repeticiones terminales de forma de aumentar la estabilidad del arreglo [Kramer et al., 2010].

Mediante simulaciones de dinámica molecular del tipo grano grueso se hizo una exploración del paisaje energético de estas proteínas usando como coordenada global de reacción el parámetro Qw, el cual ha sido descripto como óptimo para describir el proceso de plegado [Cho et al., 2006]. Estas simulaciones se realizaron a una temperatura constante igual a la temperatura de plegado ($T_f$) estimada en 598 °K para 4ANK y de 600 °K para $NI_1C$_Mut4. Es importante aclarar que las temperaturas expresadas aquí no se corresponden con temperaturas reales, ya que el solvente no es modelado de forma explícita y parámetros como el peso del sesgo de *Umbrella Sampling* o la magnitud del parámetro de no aditividad, modifican los valores de temperatura a los cuales se realizan las simulaciones. A partir de los resultados de las trayectorias de dinámica molecular obtenidas se calcularon los perfiles de energía libre de las proteínas en función de la coordenada Qw. Los detalles y protocolos de las simulaciones efectuadas se encuentran detallados en la sección de Métodos.

Al observar los valores de energía libre, medida en Kcal/mol, en función del valor de Qw, vemos en ambos casos que los perfiles calculados son compatibles con un mecanismo de plegado de 2 estados (Fig. 4.2A y Fig. 4.2B) en donde los estados plegados y despegados están separados por una barrera bien definida de 16.2 Kcal/mol en el caso de 4ANK y de 13.2 Kcal/mol en el caso de $NI_1C$_Mut4. En muchos casos es útil usar una segunda coordenada de reacción para separar aquellas conformaciones que pudieran tener valores semejantes de Qw, pero difieran en otros aspectos. Para ello, hemos elegido graficar los valores de energía libre tanto en función de Qw como del radio de giro (Rg) de las estructuras. En el caso de $NI_1C$_Mut4 (Fig. 4.2D) al usar Rg como segunda coordenada de reacción, la región correspondiente al estado plegado es más amplia que en el caso de 4ANK (Fig. 4.2C) en donde existe presencia de conformaciones de energía relativamente similares a la del estado nativo en donde la repetición C-terminal, se encuentra desplegada. Si bien en 4ANK no se observa un estado expandido tan estable como en el caso de $NI_1C$_Mut4, para ambas estructuras el estado plegado corresponde con estructuras con un Qw de aproximadamente 0.7 en donde hay una cierta perdida de

estructuración, respecto de la estructura cristalográfica usada como entrada. Se discutirá más en profundidad este punto en los análisis subsiguientes.

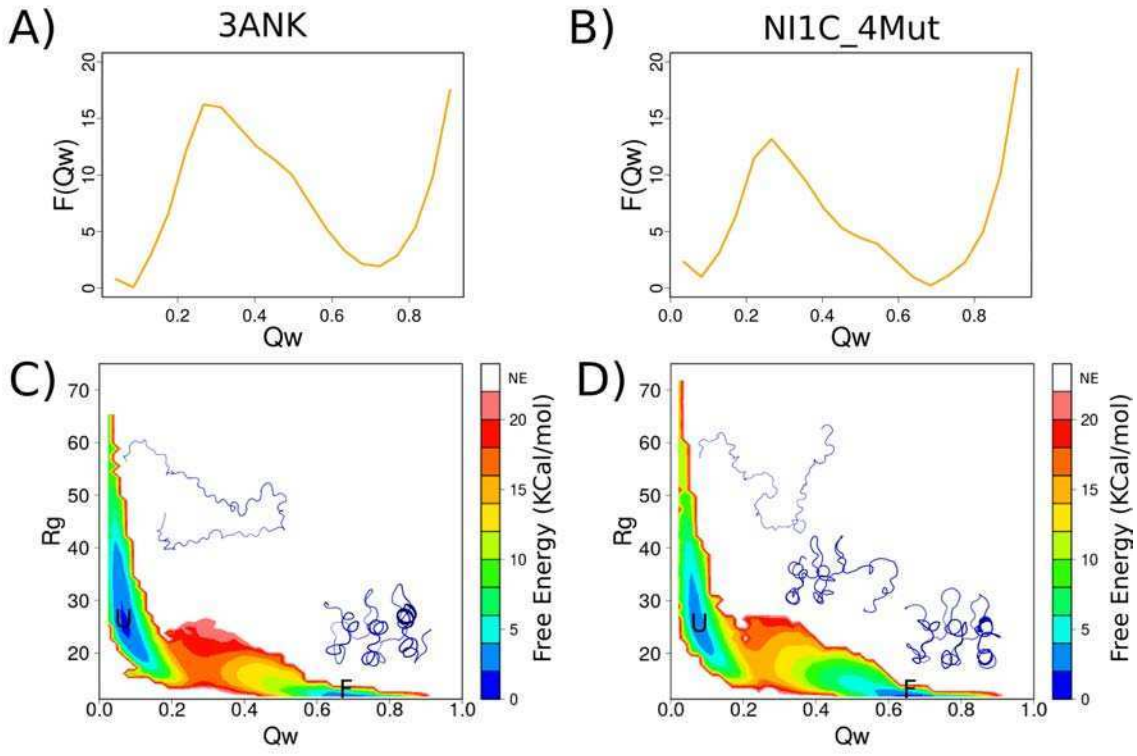


**Figura 4.2:** Curvas de Energía libre en función de Qw para A) 4ANK (1n0q,A) y B) $NI_1C$_Mut4(2xen,A). Perfiles de energía libre en 3 dimensiones en función de Qw y Rg para C) 4ANK y D) $NI_1C$_Mut4,A. Las regiones correspondientes con el estado plegado y desplegado se marcan en los gráficos con las letras F y U respectivamente

Para evaluar en detalle el mecanismo de plegado agrupamos todas las estructuras exploradas en pequeños intervalos de Qw. Dada una estructura que posee un valor de Qw específico puede evaluarse para cada residuo cuantas de las interacciones presentes en el estado nativo se encuentran formadas, calculando de esta forma un Qw local por residuo. En la Fig. 4.3 puede observarse para las estructuras de un mismo Qw global el valor promedio de Qw local por residuo. Se puede ver que tanto para 3ANK (Fig. 4.3A) como para $NI_1C$_Mut4 (Fig. 4.3B) los perfiles de plegado analizados de esta forma son muy similares. El mecanismo de plegado es claramente cooperativo, en donde vemos que tempranamente para estructuras con valores Qw de aproximadamente 0.3, comienzan a formarse las interacciones nativas de un gran cantidad de residuos correspondientes a R1 y R2, de forma cooperativa. Sin embargo, existen diferencias sutiles entre ambas moléculas. Los residuos pertenecientes a la región terminal de R3 en

$NI_1C$_Mut4 se pliegan en forma más tardía respecto de la región análoga en 3ANK. Si bien R1 comienza a plegarse en forma temprana se observa un evento de retroceso (formación de estructura nativa en etapas tempranas del plegado que luego se rompen para volverse a formar en etapas más tardías) entre valores de Qw de 0.5 y 0.6. Algo análogo sucede en $NI_1C$_Mut4 aunque menos acentuado.

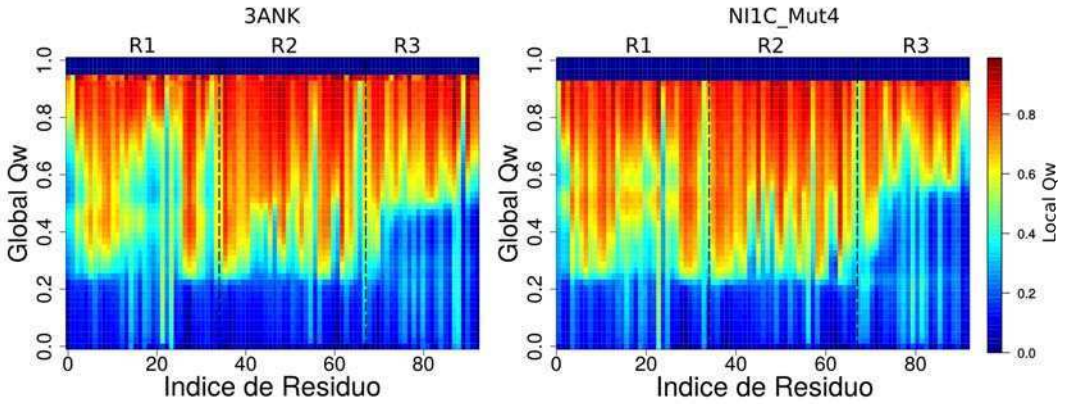


**Figura 4.3:** Mecanismos de plegado: En el eje x se observan las posiciones correspondientes a los diferentes residuos de las proteínas. En el eje y se observa la coordenada Qw global. El color indica el promedio del valor de Qw local para un residuo específico en todo el conjunto de estructuras de un determinado valor global de Qw. Las líneas de puntos marcan los límites entre repeticiones adyacentes.

Como ya hemos mencionado anteriormente, las repeticiones podrían considerarse como elementos de plegado o foldones. En los casos analizados observamos que las repeticiones definidas corresponden aproximadamente con las regiones que se pliegan de forma cooperativa, es decir como foldones (Fig. 4.3). Si se observa además el gráfico en la región correspondiente al valor de Qw en donde ocurre el estado de transición (TSE, Qw$\sim$0.3 en ambos casos), podemos ver que la nucleación del plegado incluye una parte mayoritaria de la repetición central y cerca de la mitad de la repetición N-terminal. Ya anteriormente se ha descripto que en la familia ANK, la nucleación de plegado corresponde con más de una repetición y menos de dos [Ferreiro et al., 2005].

Si consideramos cada foldón, como un micro-estado que corresponde a un mínimo local en el paisaje energético, podemos definir para cada uno de ellos 2 estados discretos: plegado o desplegado (1 o 0, respectivamente). Schafer y colaboradores, describieron recientemente un método para construir modelos cinéticos discretos del plegado de proteínas a partir de simulaciones de plegado [Schafer et al., 2012] y lo aplicaron sobre 4ANK. Como bien ellos discutieron en el artículo, definir cuales son los foldones en estructuras de proteínas repetitivas

no es algo trivial y se limitaron a aplicar el modelo usando las definiciones de foldones definidas por el grupo de Peng [Mosavi et al., 2002] al momento de sintetizar la proteína. El modelo cinético desarrollado por Schafer usa la definición de foldón como una región contigua de estructura primaria que se puede plegar de forma independiente. Esta visión es compatible con la definición acuñada por Panchenko y Wolynes [Panchenko et al., 1996], que requiere que la región correspondiente al foldón sea cinéticamente competente. Nosotros hemos usado nuestras repeticiones definidas mediante nuestro método de teselado como foldones para poder aplicar el método de Schafer de forma consistente en varios miembros de la familia ANK. Dada la definición de los foldones y que cada uno de ellos puede estar en un (micro)estado plegado (1) o desplegado (0). Una proteína que posee $N_f$ foldones tiene un total de $2^{N_f}$ macro estados posibles. Así, 4ANK posee un total de $2^4 = 16$ macro estados y se representan de la forma 0000 (todos los foldondes desplegados), 1111 (todos los foldones plegados) o 0011 (los primeros dos foldones desplegados y los últimos dos plegados), etc. Nótese, que durante la dinámica no necesariamente todos los macro estados posibles son explorados. La definición del estado plegado/desplegado de cada foldón se hace mediante la evaluación del valor Qw correspondiente a dicha región en donde si dicho valor supera un valor de Qw correspondiente a 0.6, se asigna el estado plegado y desplegado en caso contrario. A partir de los resultados de las trayectorias con muestreo usando *Umbrella Sampling* se calculan los valores de energía libre relativa para cada macro estado explorado usando el método MBAR [Shirts and Chodera, 2008]. Los valores son relativos, ya que siempre se asigna el valor 0.00 KTs a alguno de los macroestados y los valores de los demás se calculan de forma relativa al mismo.

Una vez definidos los foldones y los macroestados posibles de la estructura global, pueden calcularse las tasas de interconversión entre estos últimos. El modelo de Schafer postula que dos macroestados están conectados si para convertir uno en otro se necesita una única conversión de un foldón en estado 1 a 0, o viceversa. Detalles de cómo se calcula la interconversión entre macroestados se proveen en la sección de Métodos. En la Fig. 4.4A se muestran los resultados de aplicar el modelo cinético discreto de Schafer sobre 4ANK.

Se mencionarán a partir de acá los macroestados como combinaciones de 0 y 1, haciendo referencia al estado plegado y desplegado de las repeticiones de acuerdo a su posición en el

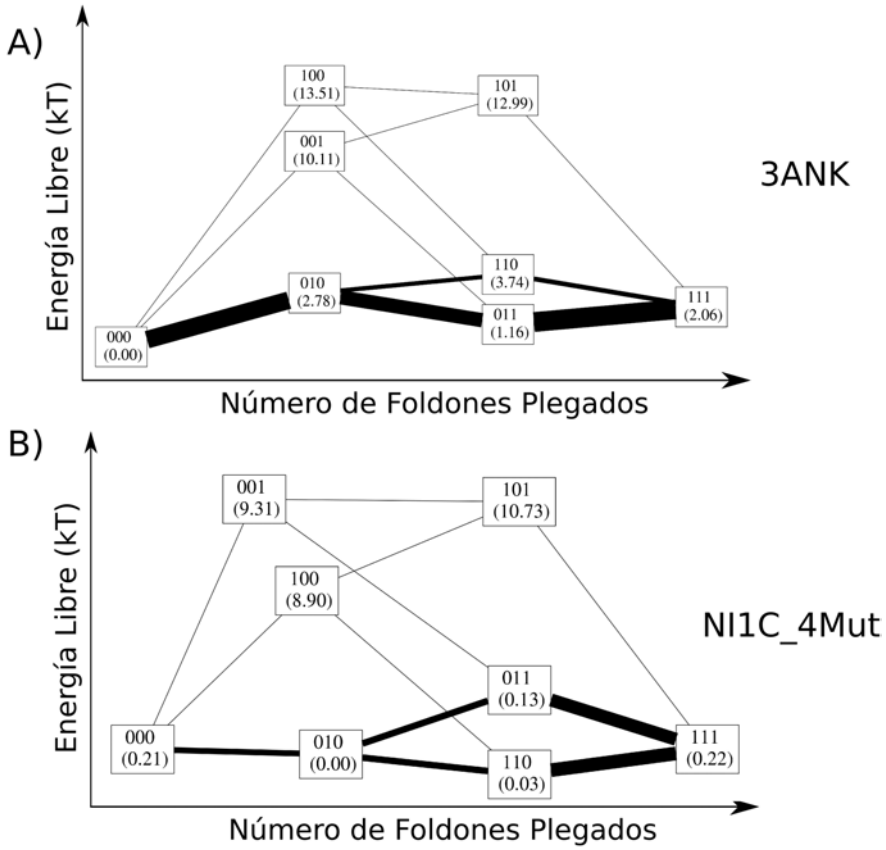


**Figura 4.4:** La coordenada vertical aproxima la energía libre de cada macroestado. Por motivos gráficos, para no superponer representaciones de macroestados, la relación exacta entre alturas de las cajas no se corresponde con el valor de energía libre, los mismos se muestran entre paréntesis debajo de las mismas. La coordenada horizontal aproxima la coordenada de reacción global (Qw). Se dibuja una línea entre cada par de macroestados conectados cuyo espesor es proporcional al flujo de transición entre los mismos.

arreglo. Se puede observar que el macroestado más estable es 000, es decir el que posee todas las repeticiones desplegadas. Luego de dicho estado, el más estable es el que posee la repetición N-terminal desplegada (011) al que se llega transicionando primero por 010. En cuanto a los valores de energía libre relativos, es interesante observar que de aquellos macroestados en donde existe un foldón desplegado, el más estable es aquel en donde la repetición N-terminal se encuentra desplegada, respecto del caso en que la repetición C-terminal lo está. Esto no es aparente en los análisis mostrados anteriormente (Fig. 4.3A). Las razones posiblemente sean el valor de corte usado para definir el estado plegado (Qw interno mayor a 0.6) y la presencia del evento de retroceso en el C-terminal. Hay que recordar que mientras en el análisis anterior, el Qw se calcula por residuo dado un valor global de Qw, en este análisis el valor de Qw se calcula por foldón. Puede ocurrir que mientras ciertos residuos se encuentran formando una alta proporción de interacciones nativas en una repetición a nivel global de la misma como en

el caso de la repetición N-terminal (Fig. 4.3A), esta no se considere plegada y por ello el estado 011 es más estable que el 110 (Fig. 4.4B). El caso en que la repetición interna se encuentra desplegada únicamente es un macroestado muy desfavorable. En la Fig. 4.4B podemos observar los resultados del mismo análisis sobre $NI_1C$_Mut4. En este caso el macroestado más estable es 010 seguido por 110.

Con estos análisis podemos observar que a pesar de comportarse mayoritariamente como proteínas con mecanismos de plegado de dos estados, hay más información que puede obtenerse acerca del proceso de plegado en general. Pudimos observar que ambas estructuras poseen en su estado plegado una desestructuración parcial de una de sus repeticiones terminales. En el caso de $NI_1C$_Mut4 la transición del estado plegado hacia aquel en que se despliega la repetición C-terminal o N-terminal parece ser muy favorecida. En 4ANK ocurre algo similar, pero con desestructuración de la repetición N-terminal.

## 4.4. 4 repeticiones

Hasta el momento los arreglos repetitivos naturales de ANKs más cortos conocidos consisten en 4 repeticiones. Así como en el caso de 3 repeticiones existen también proteínas repetitivas diseñadas con 4 repeticiones. En esta sección analizaremos los mecanismos de plegado de la proteína consenso 4ANK (1n0r,A) como también aquellos correspondientes a 3 proteínas naturales bastante conocidas: la proteína supresora de tumores P16INK4A (1a5e,A,*Homo Sapiens*), la región N-terminal de la proteína Notch (repeticiones 4-7, 1ymp,B, *Mus Musculus*) y por último un constructo ad-hoc de las primeras 4 repeticiones de la proteína I$\kappa$B$\alpha$ (1nfi,E, región 67-206, *Homo Sapiens*). Los valores de $T_f$ se estimaron en 585 °K para 4ANK, 680 °K P16INK4A, 600 °K para Notch y 635 °K para I$\kappa$B$\alpha$.

4ANK muestra la barrera energética más alta entre su estado desplegado y su estado plegado con un valor de Qw alrededor de 0.3 (Fig. 4.5A). P16INK4A (Fig. 4.5B) y Notch (Fig. 4.5C) en cambio, poseen barreras más bajas y un estado plegado con valores cercanos a Qw$\sim$0.6. El caso de P16INK4A es interesante ya que posee un estado de transición que abarca una mayor región del perfil de energía libre comparado con las demás proteínas. En el caso de I$\kappa$B$\alpha$ (Fig. 4.5D), podemos observar que es la que tiene la barrera más baja de todas y con

la particularidad de que en su perfil de energía libre, los estados estables corresponderían al estado desplegado y a una estructura con un Qw de alrededor de 0.4, no pudiendo estabilizarse el estado completamente plegado.

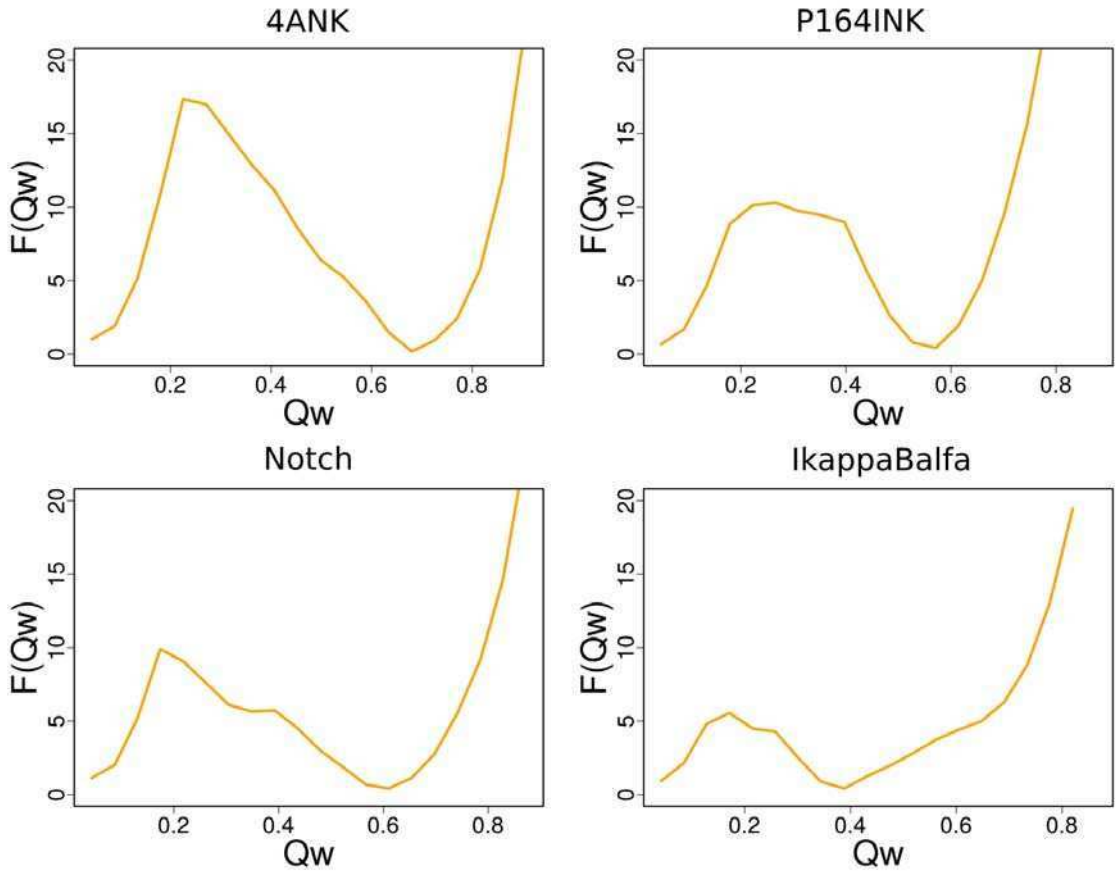


**Figura 4.5:** Perfiles de energía libre en 2 dimensiones en función de Qw para A) La proteína diseñada 4ANK. B) P164INK. C) Notch. D) I$\kappa$B$\alpha$ en su versión truncada que contiene las 4 primeras repeticiones.

Al usar Rg como segunda coordenada de reacción puede observarse que en el perfil energético de 4ANK (Fig. 4.6A) hay dos estados mínimos bien definidos en donde ocurre una rápida expansión en Rg de la estructura una vez superado el estado de transición ubicado en un valor de Qw de ∼0.3. Para P16INK4A (Fig. 4.6B) se observa en cambio una expansión en Rg más progresiva desde el estado plegado hacia el desplegado, con una barrera energética menor entre ellos. Se observa que el estado plegado se encuentra en un valor de Qw cerca a 0.55. Este relativamente bajo valor de Qw para el estado nativo se debe a que esta proteína posee regiones de loop en sus terminales, las cuales es difícil mantener en su conformación nativa a lo largo de la dinámica, incluso a bajas temperaturas. En el caso de Notch (Fig. 4.6C), se observa un estado plegado, en el que el extremo N-terminal no se encuentra en la misma

conformación que en el cristal, reduciendo su valor de Qw hasta ∼0.6 y el arreglo repetitivo con sus 4 repeticiones plegadas. Se observa un posible intermediario de relativa baja energía en un Qw de 0.35, que tiene un Rg mayor dado que se encuentra desplegada la repetición N-terminal. Por último en el caso de I$\kappa$B$\alpha$ (Fig. 4.6D) se puede observar, que el estado más estable luego del estado desplegado posee la repetición C-terminal desplegada. A Qw=0.2 se observan confórmeros en donde ambas repeticiones terminales se encuentran desplegadas.

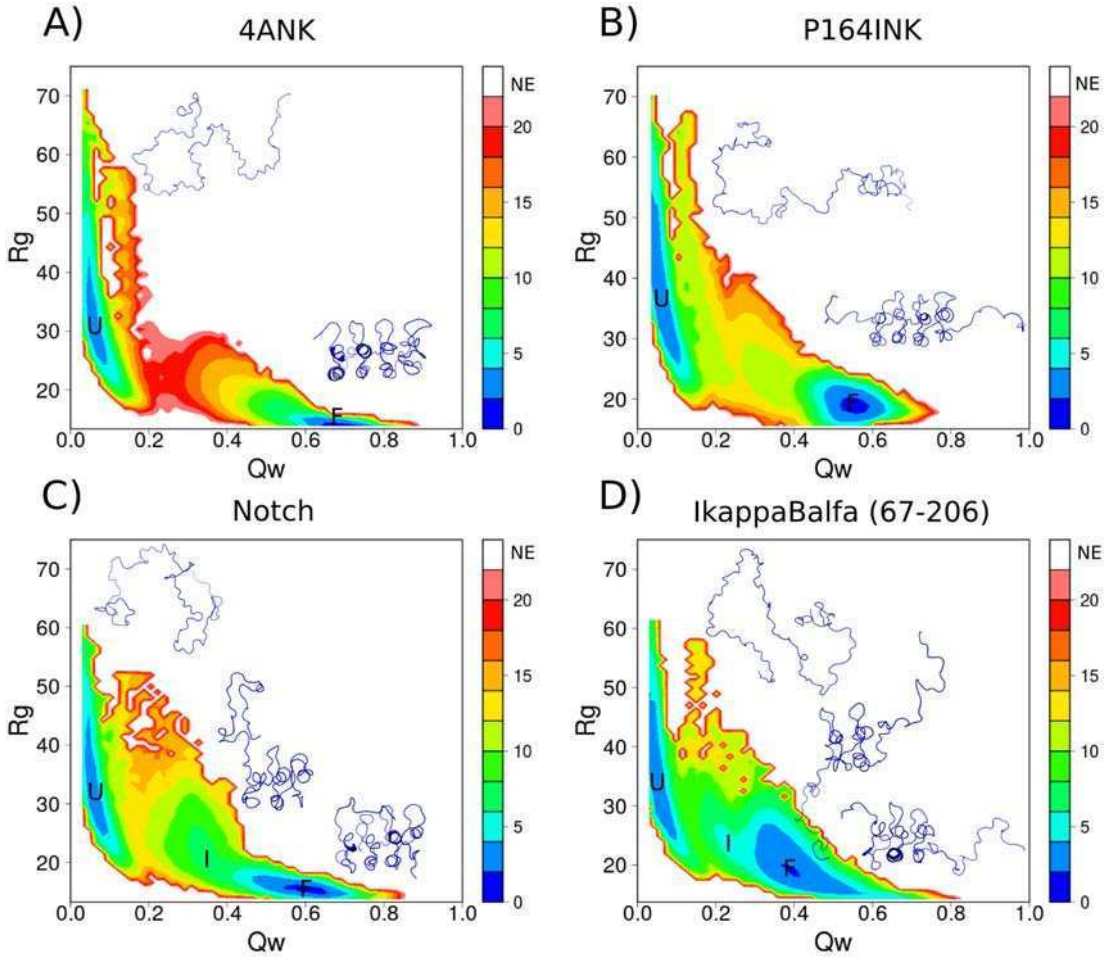


**Figura 4.6:** Perfiles de energía libre en 3 dimensiones en función de Qw y Rg para A) 4ANK. El estado plegado se marca con la letra F y el desplegado con la letra U. Se muestran ejemplos de las estructuras muestreadas para ambos estados. B) P164INK. El estado plegado se marca con la letra F y el desplegado con la letra U. Se muestran ejemplos de las estructuras muestreadas para ambos estados. C) Notch. El estado plegado se marca con la letra F, el desplegado con la letra U y el intermediario de plegado con la letra I. Se muestran ejemplos de las estructuras muestreadas para los tres estados. D) I$\kappa$B$\alpha$, en su versión truncada que contiene las primeras 4 repeticiones (aminoácidos 67 a 206). El estado plegado se marca con la letra F, el desplegado con la letra U y el intermediario de plegado con la letra I. Se muestran ejemplos de las estructuras muestreadas para los tres estados.

En cuanto al detalle de los mecanismos de plegado de estas proteína, al observar los valores de Qw para los diferentes residuos que componen las estructuras en función de los Qw

globales, podemos ver que los mecanismos difieren bastante. En el caso de 4ANK (Fig. 4.7A) se observa un plegado polarizado en sentido C-terminal a N-terminal siendo las repeticiones 3 y 4 y algunas regiones de la repetición 2 las primeras en plegarse de forma cooperativa. Es notable a medida que la estructura global adquiere una mayor proporción de interacciones nativas, localmente la repetición 4 decrece su Qw local para volver a plegarse en estadios más cercanos a la estructura totalmente plegada, lo cual se conoce como mecanismo de retroceso (*backtracking*). El retroceso se produce cuando se forma una estructura del tipo nativo la cual tiene que desplegarse y volver a replegar a medida que la estructura avanza hacia conformaciones más similares a las del estado nativo, lo cual es un indicio de frustración topológica en el paisaje energético. En el caso de P16INK4A (Fig. 4.7B) se observa, a diferencia del caso anterior, un mecanismo polarizado en el sentido N-terminal a C-terminal en donde se pliegan primero las repeticiones 1 y 2, comienza a plegarse la repetición 3 a valores Qw globales de $\sim$0.3 y posteriormente en valores de Qw $\sim$=.4 se pliega la última repetición. Dinámicas moleculares hechas anteriormente por el grupo de Caflisch [Interlandi et al., 2006] reportaron un mecanismo polarizado en el orden inverso. Cabe recordar que nuestros modelos sólo tienen en cuenta la topología de la estructura y es de esperar que haya diferencias con aquellos en que la influencia secuencia es explícitamente modelada.

En este tipo de análisis es más simple observar que las regiones no correspondientes al arreglo repetitivo en cada extremo, no adoptan conformaciones demasiado similares a la estructura nativa a lo largo de las trayectorias observadas. En el caso de Notch (Fig. 4.7C) se observa un mecanismo de plegado muy similar al de 4ANK, con algunas diferencias en la repetición N-terminal, que en el caso de Notch posee además una región helicoidal extra en el extremo N-terminal. En el caso de I$\kappa$B$\alpha$ (Fig. 4.7D) se observa que la nucleación de plegado ocurre en las repeticiones internas con una mayor contribución de la segunda repetición. Si bien la repetición N-terminal comienza a plegarse en estadios más tempranos (Qw=0.3), los valores de Qw local de los residuos que la componen se mantienen en valores intermedios ($\sim$0.6) adoptando altas proporciones de contactos nativos recién en valores de Qw cercano a 0.75. En contraste la repetición C-terminal comienza a plegarse tardíamente a partir de estructuras cuyo Qw es cercano a 0.6, pero lo hace de una manera más cooperativa. En la región analizada I$\kappa$B$\alpha$ posee dos inserciones, la primera entre las repeticiones 1 y 2 y la segunda entre

las repeticiones 3 y 4. Es interesante observar que la inserción entre las repeticiones 1-2 corresponde a una de las regiones que más tempranamente se pliegan en la estructura, mientras que por el contrario la inserción entre las repeticiones 3-4 adquiere estructuración nativa recién en valores globales de Qw superiores a 0.75. Probablemente dicha inserción esté involucrada en la perturbación estructural que imposibilita que la repetición C-terminal se encuentre plegada de forma estable en el estado plegado de menor energía.

Las nucleaciones en las proteínas analizadas son diferentes en todos los casos. 4ANK nuclea su plegado en la repetición R3 y parte de R4. P164INK comienza a plegarse a partir de R2 y parte de R1. Si bien Notch y I$\kappa$B$\alpha$ comienzan a plegarse a partir de las repeticiones centrales, Notch lo hace a partir de la totalidad de R3 y parte de R2 mientras que I$\kappa$B$\alpha$ lo hace al revés (Fig. 4.7).

Los modelos cinéticos discretos aplicados sobre este conjunto de proteínas nos ayudan a analizar la existencia o no de plegados acoplados entre las diferentes repeticiones. Se observa que para el caso de 4ANK (Fig. 4.8A), el estado más estable después del desplegado es 0110. A partir del estado completamente plegado (1111) es relativamente fácil una transición hacia el estado 0111 o 1110. Salvo el caso 0100 no se observa ningún estado estable en que una repetición se encuentra plegada sin tener otra repetición plegada en forma adyacente. Compatible con un mecanismo de plegado de 2 estados, 4ANK muestra altas constantes de transición entre los macroestados más estables, principalmente se observan altas constantes de transición hacia los estados en que uno o dos repeticiones terminales se despliegan, hecho conocido como deshilachado o "*fraying*". Para el caso de P164INK (Fig. 4.8B), el macroestado más favorable es 1111 con una favorable transición hacia el estado en que se despliega la repetición C-terminal (0111). A partir de este último estado lo más favorable es desplegar la repetición N-terminal. Al igual que en el caso anterior, antes de pasar al estado totalmente desplegado lo más favorable es desplegar la repetición 3 (transicionando al macroestado 0100). Si para estos dos primeros casos se observan los macroestados más estables para el caso de tener 4, 3, 2, 1, 0 repetición(es) plegada(s), se observa el mismo conjunto de macroestados para ambos (1111-0111-0110-0100-0000). Para Notch (Fig. 4.8C), el estado más estable es 0000 seguido por 0111 que corresponde al macroestado del intermediario descripto anteriormente

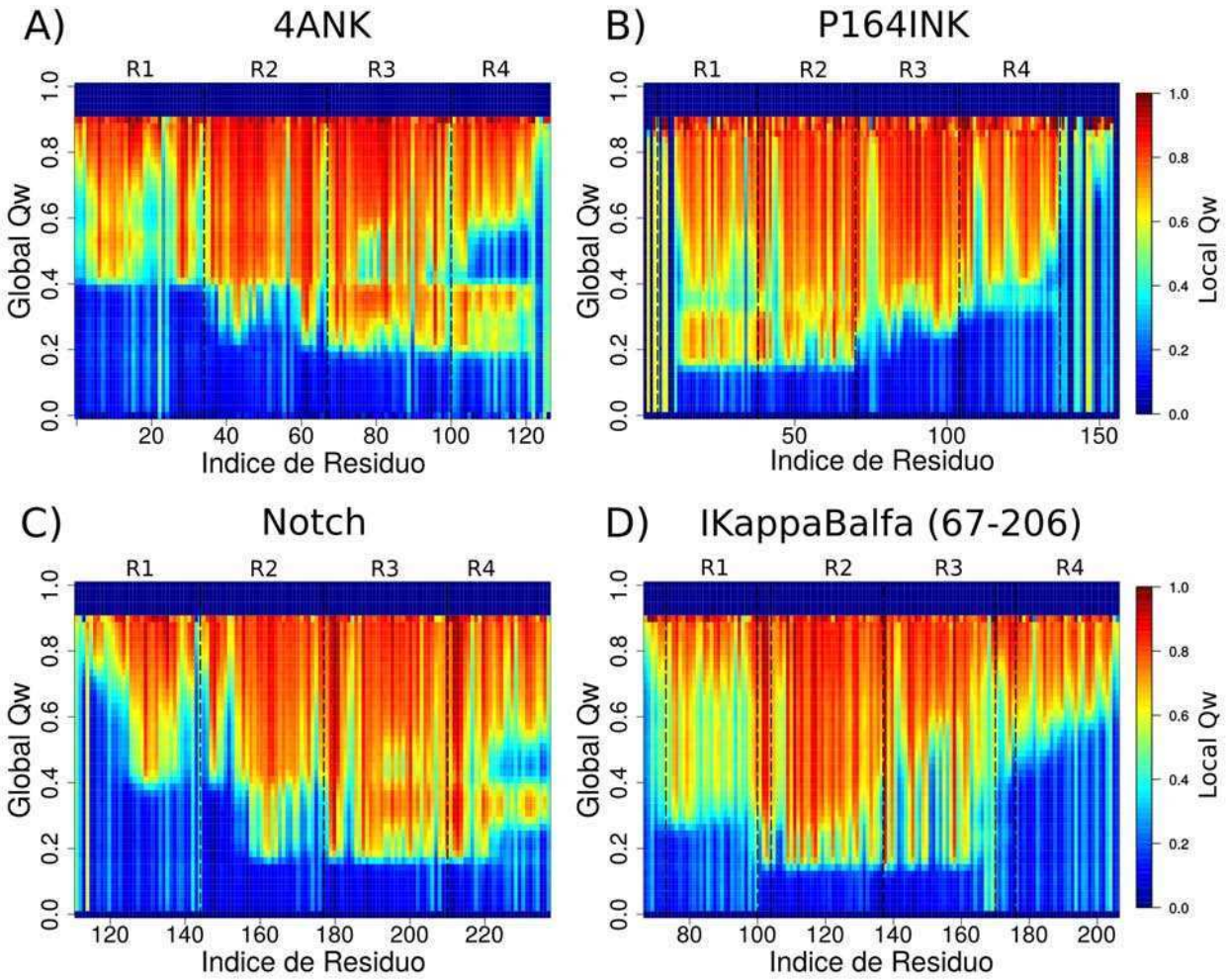


**Figura 4.7:** Mecanismos de plegado: En el eje x se observan las posiciones correspondientes a los diferentes residuos de las proteínas. En el eje y se observa la coordenada Qw global. El color indica el promedio del valor de Qw local para un residuo específico en todo el conjunto de estructuras de un determinado valor global de Qw. Las líneas de puntos marcan los límites entre repeticiones adyacentes.

(Fig. 4.6C). A partir de allí es probable transicionar a través de 0110 y 0010 hacia el estado desplegado. A diferencia de las otras dos proteínas y de I$\kappa$B$\alpha$, es el único caso en que es más favorable la transición 0110 $\longrightarrow$ 0010 $\longrightarrow$ 0000 que la correspondiente a 0110 $\longrightarrow$ 0100 $\longrightarrow$ 0000. El estado más estable en el caso de I$\kappa$B$\alpha$ (Fig. 4.8D) es aquel correspondiente a tener las dos repeticiones internas plegadas (0110) a partir del cual, se puede transicionar fácilmente hacia el estado 0100 antes de visitar el estado totalmente desplegado. Las altas constantes de transición observadas para 1111 $\longrightarrow$ 1110 $\longrightarrow$ 0110 o 1111 $\longrightarrow$ 0111 $\longrightarrow$ 0110 se deben a que es muy costoso entrópicamente mantener a la estructura en el estado completamente plegado como así también en los estados en que alguna de las repeticiones terminales se encuentra plegada.

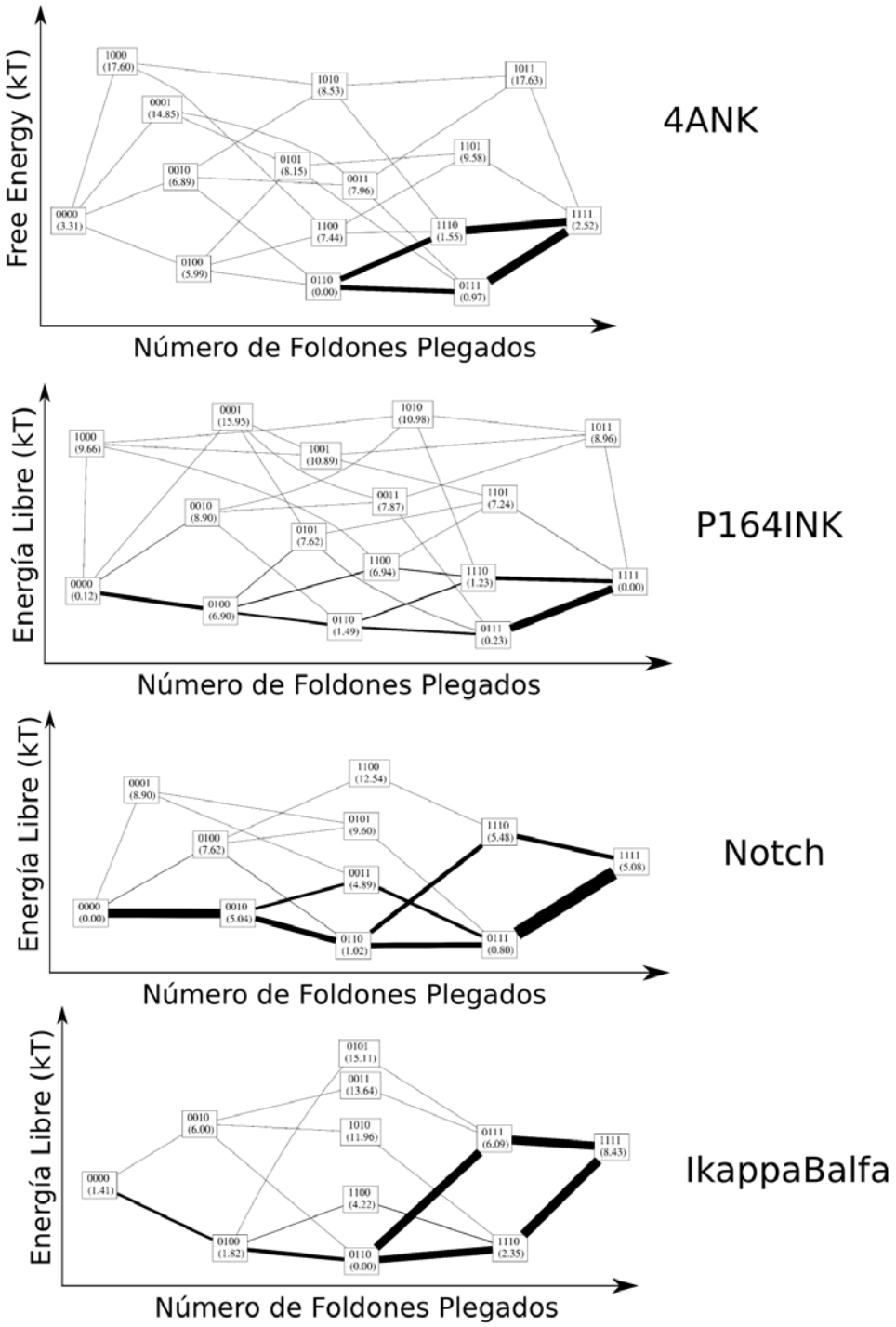


**Figura 4.8:** La coordenada vertical aproxima la energía libre de cada macroestado. Por motivos gráficos, para no superponer representaciones de macroestados, la relación exacta entre alturas de las cajas no se corresponde con el valor de energía libre, los mismos se muestran entre paréntesis debajo de las mismas. La coordenada horizontal aproxima la coordenada de reacción global (Qw). Se dibuja una línea entre cada par de macroestados conectados cuyo espesor es proporcional al flujo de transición entre los mismos.

En general en estas moléculas, no se observa en los modelos cinéticos discretos la ocurrencia de rutas paralelas excepto en el caso de desplegado de repeticiones terminales. Una excepción puede ser el caso de Notch (Fig. 4.8C) en que los estados 1110 y 0011 se encuentran conectados a través de constantes de transición altas y se encuentran en rutas diferentes. Sin embargo dado que dichos macroestados tienen altos valores de energía libre relativa, es de esperar que la población de rutas paralelas esté sesgada.

## 4.5. 5 repeticiones

En el caso de proteínas con 5 repeticiones hemos analizado el plegado de la proteína diseñada E3_5 (1mj0,A), GABPB1 (1awc,B, *Mus Musculus*), la proteína termófila TV1425 (2rfm,A, *Thermoplasma volcanium*) y ANKRA2 (3so8,A, *Homo Sapiens*). La $T_f$ estimada fue de 570°K para E3_5, 570 K para GABPB1, 612 K para TV1425 y 662 K para ANKRA2.

E3_5 (Fig. 4.9A) posee su estado plegado en un valor de Qw ∼0.7 y su perfil de energía libre es compatible con un mecanismo de plegado de 2 estados. GABPB1 (Fig. 4.9B) posee una barrera de energía similar a E3_5, aunque posee su estado plegado en un Qw cercano a 0.55 exhibiendo además una especie de hombro en la barrera a Qw=0.3, lo cual probablemente sea un intermediario de plegado. TV1425 (Fig. 4.9C) y ANKRA2 (Fig. 4.9D) poseen barreras energéticas similares a las dos anteriores. TV1425 tiene un estado plegado estable en Qw ∼0.55. El estado de transición en este caso es más amplio como se observa en la meseta en la región más alta de la barrera de energía. En el caso de ANKRA2 se observan estructuras estables con valores de Qw ∼ 0.45. A pesar de que las moléculas son topológicamente similares, los estados estables son diferentes.

Al observar los perfiles de energía usando como coordenada de reacción adicional Rg, podemos ver que E3_5 (Fig. 4.10A) tiene dos estados estables bien diferenciados y que siendo una proteína diseñada, la barrera energética que separa a estos estados es menor que en el caso de 3ANK, $NI_1C$_Mut4 y 4ANK, que tienen un menor número de repeticiones. El mecanismo de plegado parece ser menos cooperativo que en los casos de las demás proteínas diseñadas dado el aumento progresivo de Rg a medida que Qw decrece, seguramente acompañado de desestructuraciones de las repeticiones terminales. Los perfiles de energía libre para las demás

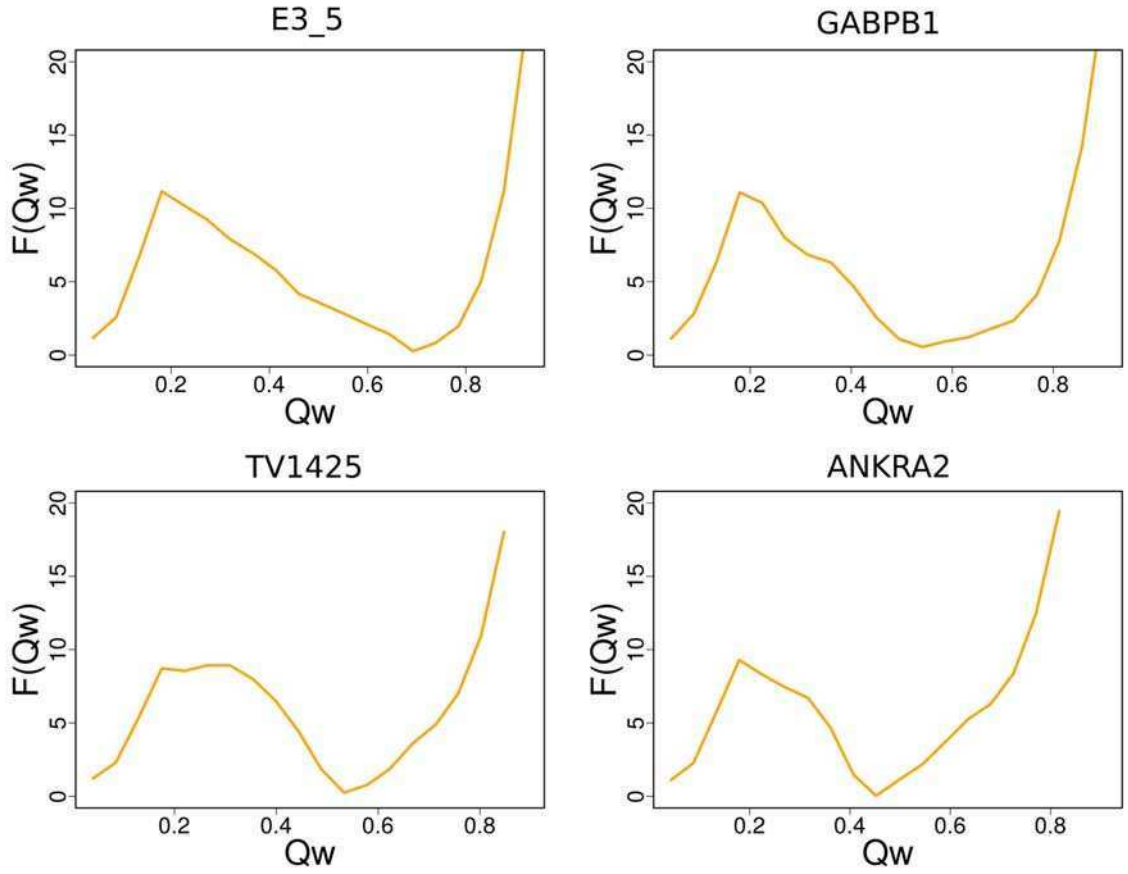


**Figura 4.9:** Perfiles de energía libre en 2 dimensiones en función de Qw para A) La proteína diseñada E3_5. B) GABPB1. C La proteína termófila TV1425. D)ANKRA2.

proteínas de este grupo son muy parecidos, con la diferencia que en el caso de GABPB1 (Fig. 4.10B), TV1425 (Fig. 4.10C) y ANKRA2 (Fig. 4.10D) el estado plegado se encuentra estabilizado en menores valores de Qw.

Si bien los perfiles de energía libre son bastante similares en los casos analizados, los análisis de Qw local a lo largo de las dinámicas evidencian amplias diferencias en los mecanismos de plegado. En el caso de E3_5 (Fig. 4.11A) se observa que las primeras repeticiones en plegarse son R3 y R4, seguidas por la repetición C-terminal, R5. Al igual que en el caso de 4ANK y Notch, la repetición R5 presenta retrocesos es su proceso de plegado, mostrando dos eventos de este tipo en Qw global $\sim$0.3 y 0.5. A esos mismos valores de Qw, se observa retroceso en el plegado de R4, aunque en menor medida. Las últimas repeticiones en plegarse son R2 y R1. En GABPB1 (Fig. 4.11B) son R3 y R4 las repeticiones que comienzan a plegarse de forma más temprana y de forma cooperativa. Lo mismo ocurre con R2, en que algunos residuos centrales establecen interacciones nativas y recién a partir de Qw $> 0.4$ se pliega de

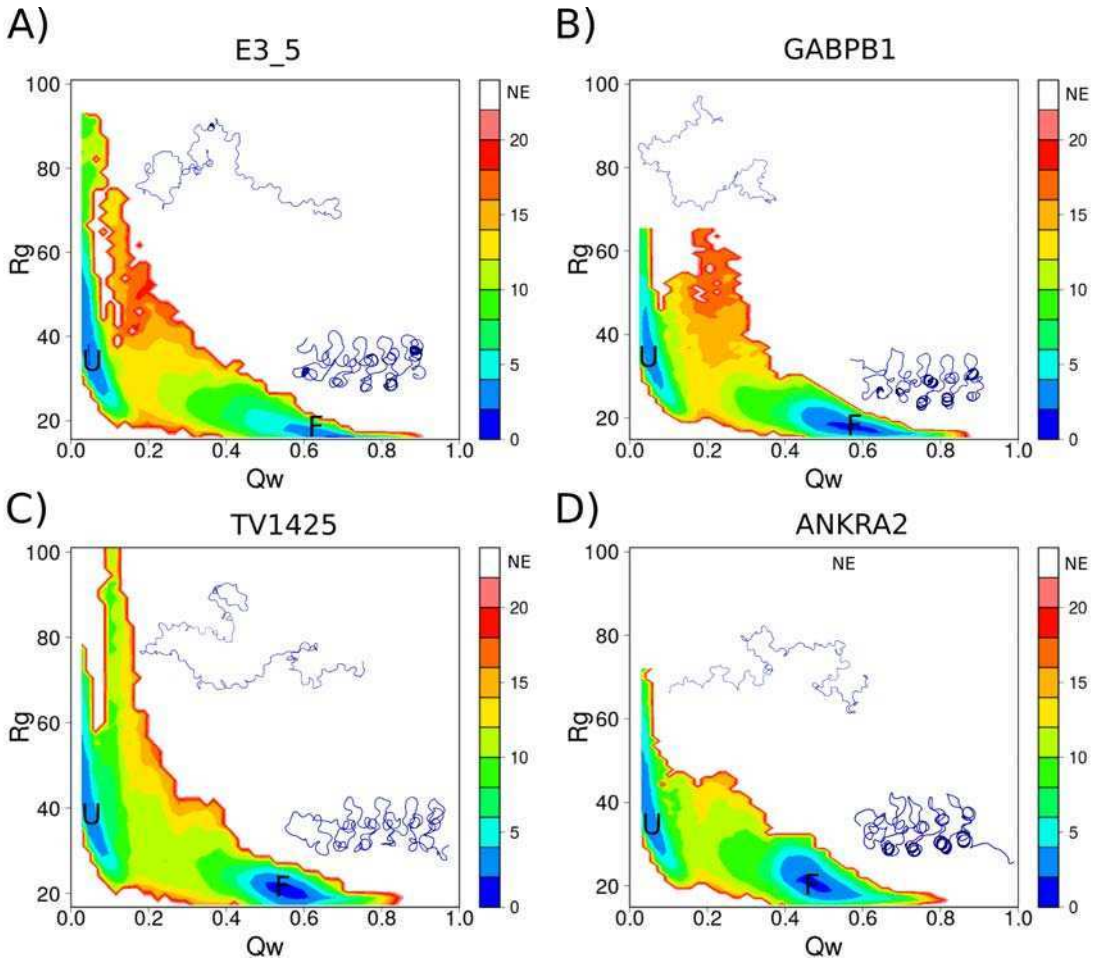


**Figura 4.10:** Perfiles de energía libre en 3 dimensiones en función de Qw y Rg para A) La proteína diseñada E3_5. El estado plegado se marca con la letra F y el desplegado con la letra U. Se muestran ejemplos de las estructuras muestreadas para ambos estados. B) GABPB1. El estado plegado se marca con la letra F y el desplegado con la letra U. Se muestran ejemplos de las estructuras muestreadas para ambos estados. C) La proteína termófila TV1425. El estado plegado se marca con la letra F y el desplegado con la letra U. Se muestran ejemplos de las estructuras muestreadas para ambos estados. D)ANKRA2. El estado plegado se marca con la letra F y el desplegado con la letra U. Se muestran ejemplos de las estructuras muestreadas para ambos estados.

forma completa, momento en que también se comienza a plegar R1. A diferencia de E3_5, la repetición C-terminal en este caso es la última en plegarse. TV1425 (Fig. 4.11C) posee un núcleo de plegado que involucra a R3, R4 y R5 observándose una polarización del plegado desde el extremo C-terminal al N-terminal, más fuerte que en el caso de E3_5. En este caso también se puede ver que la repetición C-terminal exhibe eventos de retroceso de plegado, aunque menos marcados que en E3_5. Es interesante notar que este proceso de retroceso involucra no sólo a R5 sino que también a R4. Ambas repeticiones muestran un patrón de plegado muy similar, probablemente conformando una única unidad de plegado. R2 y R1, en ese orden, se pliegan por último. Esta proteína tiene además una región helicoidal adicional

en su extremo N-terminal que es la última región en adoptar su estructura nativa en las simulaciones. ANKRA2 es la proteína con el mecanismo de plegado más diferente (Fig. 4.11D). El plegado comienza por las repeticiones R2, R3 y de forma parcial con R1. Las últimas en plegarse con las repeticiones R4 y R5, en ese orden, mostrando de forma global un plegado polarizado desde el extremo N-terminal al C-terminal. Algo para observar en todos los casos analizados es que las regiones que se encuentran sobre las líneas de puntos, en los límites entre las repeticiones (es decir, son inserciones), muestran en todos los casos altos valores de Qw locales, y comienzan a plegarse en etapas relativamente más tempranas que las regiones internas. Esto no se observó de forma tan marcada en proteínas con arreglos repetitivos más cortos. Puede que al tener una mayor cantidad de repeticiones, las interfaces entre las mismas comienzan a tener una mayor preponderancia en la estabilización secuencial de los elementos de plegado.

Al analizar los modelos cinéticos discretos podemos ver que para el caso de E3_5 (Fig. 4.12A) el estado más estable después del estado desplegado es aquel que contiene las dos repeticiones terminales desplegadas (01110) siendo la transición 11111 $\longrightarrow$ 01111 la más probable para llegar a este estado. Las constantes de transición para desplegar repeticiones adicionales son pequeñas y dichos estados son mucho más desfavorables en su energía libre que 01110 lo cual indica la rápida transición hacia el estado desplegado de una forma muy cooperativa, haciendo la siguiente transición 01100 $\longrightarrow$ 00100 $\longrightarrow$ 00000. En el caso de GABPB1 (Fig. 4.12B) también es muy desfavorable energéticamente tener todo el arreglo repetitivo plegado.

Desde allí las energías de desplegar cualquiera de las dos repeticiones terminales son parecidas. El estado más estable, como en casos anteriores es aquel que tiene ambas repeticiones terminales desplegadas. A partir de ese estado, la transición hacia el estado desplegado es 01100 $\longrightarrow$ 00100 $\longrightarrow$ 00000 al igual que E3_5. TV1425 (Fig. 4.12C) al igual que las demás tiene el estado 01110 como el más estable, siendo más favorable desplegar la repetición C-terminal en el paso intermedio desde el estado completamente plegado. Desde el estado 01110, desplegar repeticiones extra es muy desfavorable. ANKRA2 (Fig. 4.12D) presenta la mayor diferencia energética entre los estados 01110 y 11111, siendo este último muy inestable. A partir del estado con las repeticiones terminales desplegadas, lo más favorable es desplegar la

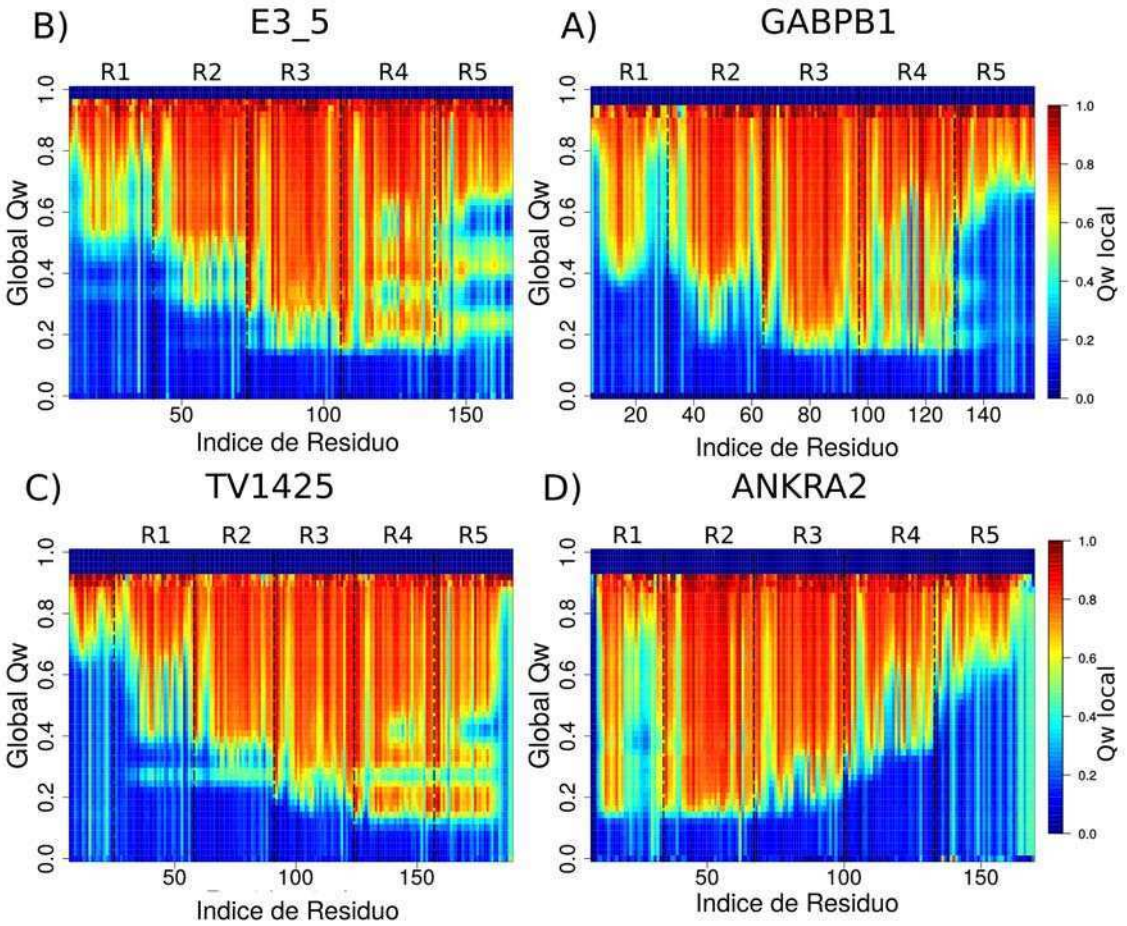


**Figura 4.11:** Mecanismos de plegado: En el eje x se observan las posiciones correspondientes a los diferentes residuos de las proteínas. En el eje y se observa la coordenada Qw global. El color indica el promedio del valor de Qw local para un residuo específico en todo el conjunto de estructuras de un determinado valor global de Qw. Las líneas de puntos marcan los límites entre repeticiones adyacentes.

repetición R4, y desde allí rápidamente se produce la transición en el orden 01000 $\longrightarrow$ 00000, ya que este último estado intermedio es muy desfavorable. Una vez más, a pesar de la alta simetría de estas proteínas observamos que los mecanismos están polarizados y no se observan rutas de plegado paralelas, salvo en el caso de repeticiones terminales.

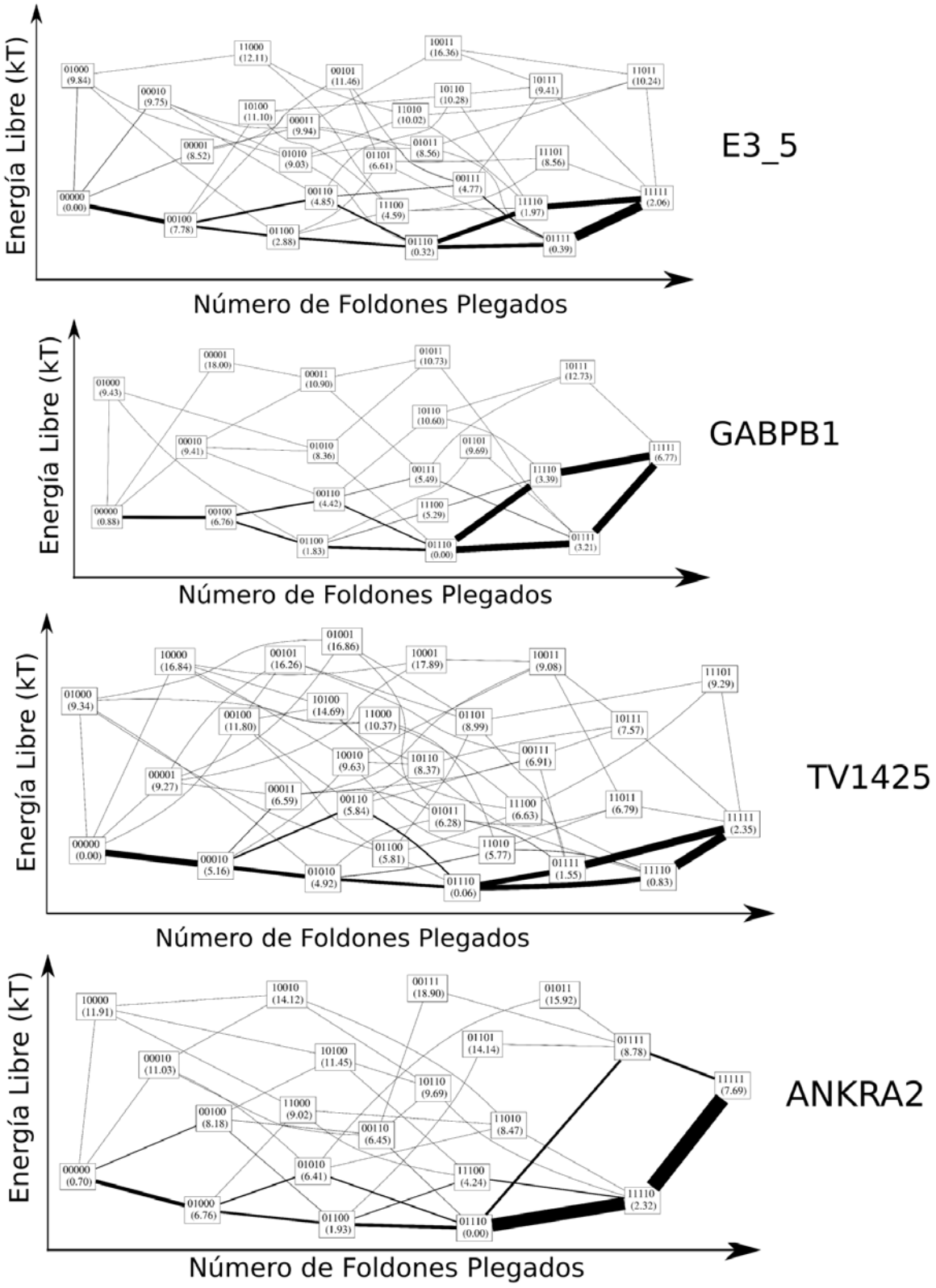


**Figura 4.12:** La coordenada vertical aproxima la energía libre de cada macroestado. Por motivos gráficos, para no superponer representaciones de macroestados, la relación exacta entre alturas de las cajas no se corresponde con el valor de energía libre, los mismos se muestran entre paréntesis debajo de las mismas. La coordenada horizontal aproxima la coordenada de reacción global (Qw). Se dibuja una línea entre cada par de macroestados conectados cuyo espesor es proporcional al flujo de transición entre los mismos.

## 4.6. 6 repeticiones

Por último hemos analizado el plegado de la versión completa de la proteína I$\kappa$B$\alpha$, para la cual previamente analizamos el plegado de la región que contiene las primeras 4 repeticiones. En este caso, la $T_f$ estimada es de 612 K. En el perfil de energía libre se observa una barrera energética a Qw 0.4 separando el estado plegado localizado en Qw $\sim$ 0.65 de un intermediario de plegado localizado en Qw $\sim$ 0.16 (Fig. 4.13A). Hay una segunda barrera energética menor que separa el intermediario de plegado y el estado desplegado (Qw $\sim$ 0.8). El intermediario de plegado localizado en Qw $\sim$ 0.16, corresponde a la estructura en donde las 3 primeras repeticiones se encuentran plegadas y el resto desplegadas, lo cual explica el aumento en Rg (Fig. 4.13B). El estado plegado está rodeado por una región de energía uniforme en la que se observan estructuras en donde la repetición N-terminal se encuentra desplegada (región E1). De forma similar, hay una región que rodea al estado intermedio, en donde podemos observar la existencia de conformaciones en donde el arreglo repetitivo se encuentra separado en dos debido a la desestructuración del $\beta$-hairpin que se encuentra en la interfaz entre las repeticiones 3 y 4 (región E2).

Al analizar los valores de Qw locales por residuo a lo largo de las trayectorias obtenidas por *Umbrella Sampling* a temperatura constante igual a la $T_f$ (Fig. 4.14), observamos que la estructura de I$\kappa$B$\alpha$ comienza a plegar principalmente por la región correspondiente a las repeticiones R2 y R3.

La repetición R1 comienza a plegarse en estadios similares a R2 y R3, pero en valores de Qw global en el rango 0.4 a 0.6 hay un evento de retroceso en que la repetición posee una fracción muy baja de interacciones nativas formadas. El desplegado de la repetición R1 ocurre en algún momento entre la región que abarca el estado expandido que rodea al estado plegado y hasta la barrera mayor que separa a estos estados del intermediario de plegado, localizado en Qw $\sim$0.2 en donde se observa la repetición R1 plegada y en cambio las repeticiones R4-R6 desplegadas. La desestructuración de R1, parece afectar o al menos ocurrir en simultáneo con una desestructuración leve de la repetición R2. Es notable que las inserciones presentes en la estructura se pliegan de forma más temprana que las regiones aledañas a las mismas. La inserción interna de R1 se pliega relativamente antes que el resto de la repetición. Así mismo,

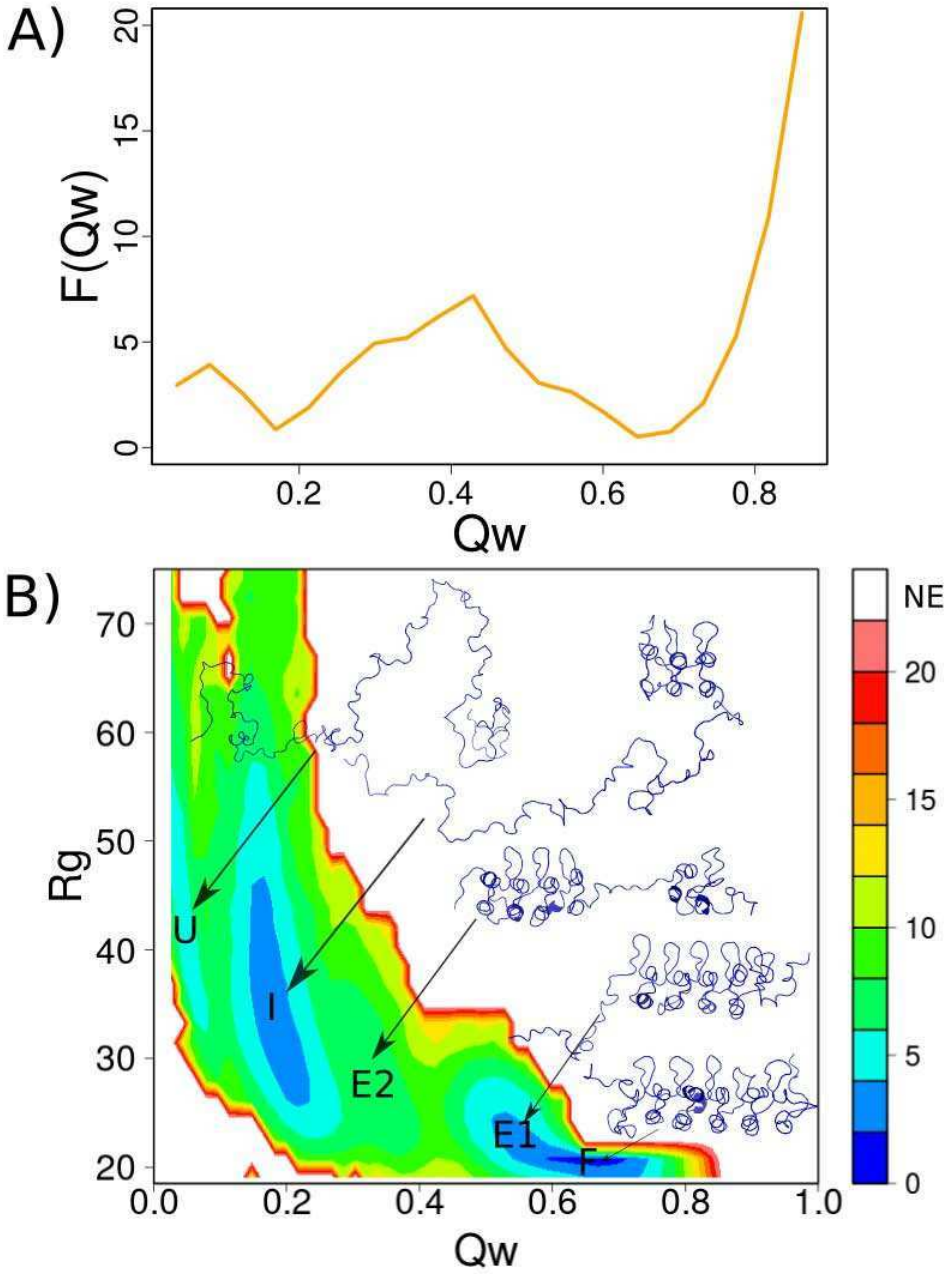


**Figura 4.13:** Perfiles de energía para la proteína I$\kappa$B$\alpha$ completa. A) Perfil de energía libre en 2 dimensiones usando Qw como coordenada de reacción. B) Perfil de energía libre en 3 dimensiones usando Qw y Rg como coordenadas de reacción. El estado plegado se marca con la letra F y el desplegado con la letra U. El intermediario de plegado se marca con la letra I. Se marcan además dos estados más, el estado E2, cercado al intermediario y el estado E1, cercano al estado plegado. Se muestran ejemplos de las estructuras muestreadas para todos los estados.

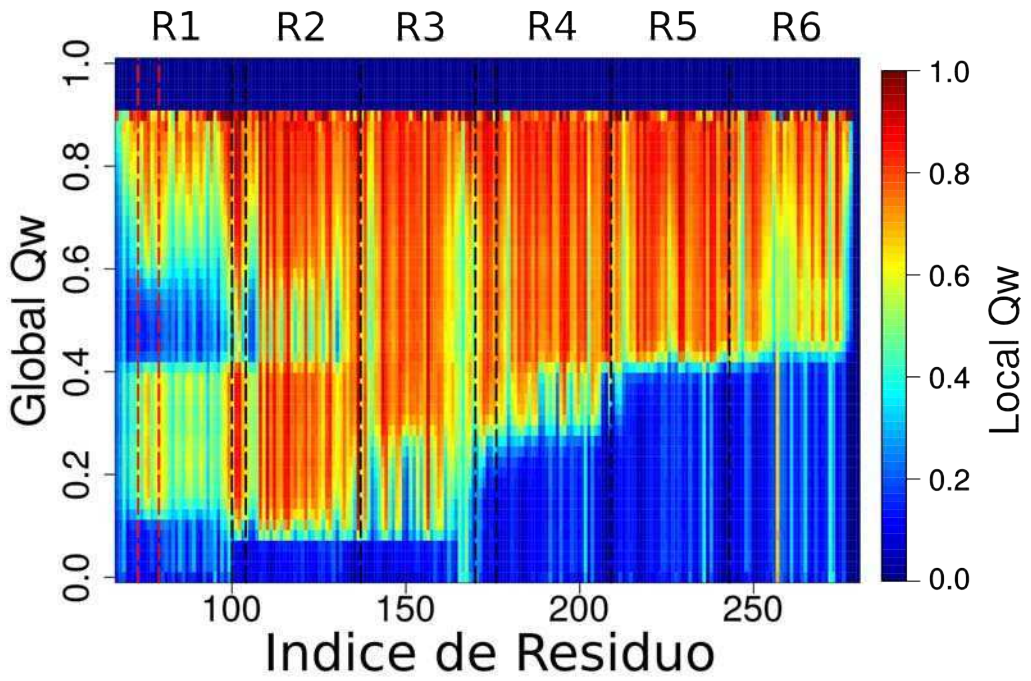


**Figura 4.14:** Mecanismos de plegado: En el eje x se observan las posiciones correspondientes a los diferentes residuos de las proteínas. En el eje y se observa la coordenada Qw global. El color indica el promedio del valor de Qw local para un residuo específico en todo el conjunto de estructuras de un determinado valor global de Qw. Las líneas de puntos marcan los límites entre repeticiones adyacentes.

las inserciones localizadas en las interfaces R1-R2 y R3-R4 establecen interacciones nativas de manera más temprana que los residuos circundantes. Además si se observa de forma global la Fig. 4.14 se puede separar cualitativamente el gráfico en 3 regiones (R1, R2-R3 y R4-R6) que se encuentran separadas por las 2 inserciones anteriormente mencionadas lo cual indicaría que los dominios de plegado se extienden más allá de repeticiones individuales. Se ha mostrado anteriormente como estas regiones en I$\kappa$B$\alpha$ pliegan como dominios independientes de plegado, que abarcan más de una repetición [Ferreiro et al., 2007a, DeVries et al., 2011].

Al aplicar el modelo cinético discreto sobre las simulaciones de I$\kappa$B$\alpha$ 4.15 observamos que el estado plegado (111111) está conectado al estado con la repetición N-terminal desplegada (011111) por medio de una constante de conversión alta, compatible con lo descripto anteriormente. Tal y como pasa con las demás estructuras analizadas, una vez desplegada una repetición terminal, lo más probable es que ocurra una transición hacia un estado con ambas repeticiones terminales desplegadas. Si se toman en cuenta las líneas más gruesas del diagrama, lo cual hace referencia a las constantes de transición con mayores valores, la próxima repetición en desplegarse sería R5 llevando la estructura al estado 011100. La transición al

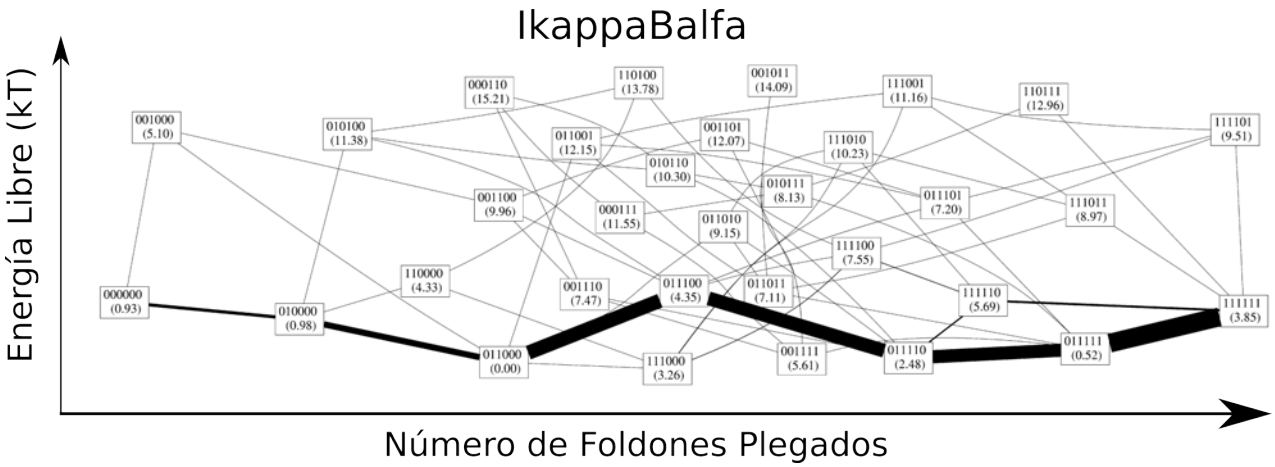


**Figura 4.15:** La coordenada vertical aproxima la energía libre de cada macroestado. Por motivos gráficos, para no superponer representaciones de macroestados, la relación exacta entre alturas de las cajas no se corresponde con el valor de energía libre, los mismos se muestran entre paréntesis debajo de las mismas. La coordenada horizontal aproxima la coordenada de reacción global (Qw). Se dibuja una línea entre cada par de macroestados conectados cuyo espesor es proporcional al flujo de transición entre los mismos.

estado desplegado se completaría desplegando R4 y R3 en forma consecutiva, siendo R2 la última repetición en desplegarse. Si en cambio observamos las estabilidades relativas de los diferentes macroestados, el que corresponde a 111000, es el más estable (aunque no cinéticamente competente) de aquellos en que se tienen 3 repeticiones plegadas y que corresponde a la estructura que observamos presente en la región del intermediario de plegado (Fig. 4.13B). A dicho macroestado se llega desplegando secuencialmente de la forma 111111 $\longrightarrow$ 111110 $\longrightarrow$ 111100 $\longrightarrow$ 111000. Esto representa una ruta alternativa a la anteriormente descripta que transiciona por el macroestado de 5 repeticiones plegadas en que la repetición desplegada es R6 en vez de R1 y puede ser la razón por la cual observamos el evento de retroceso en la región de R1 (Fig. 4.15). Ambas rutas convergen al macroestado 011000 a partir del cual transicionan de forma idéntica hacia el estado desplegado.

## 4.7. Plegabilidad Jerárquica, ¿Un proxy hacia la dinámica del plegado proteico?

La inherente simetría de las proteínas con repeticiones sugiere que las propiedades generales de plegado del dominio repetitivo o su separación en subdominios (la estabilidad y cooperatividad del arreglo) podrían derivarse a partir de una descripción microscópica del balance

energético dentro de cada elemento de plegado y su interacción con sus vecinos cercanos [Aksel and Barrick, 2009]. Debido al delicado balance energético en cada subunidad, variaciones sutiles en las interacciones en y entre repeticiones pueden generar cambios sustanciales en el paisaje energético [Ferreiro and Komives, 2007]. Dichas variaciones pueden desacoplar los elementos de plegado y como consecuencia especies parcialmente plegadas podrían poblarse, definiendo subdominios de plegado. Con suficiente información sobre la población de dichos estados es posible producir modelos cuantitativos de la distribución energética a lo largo de una proteína [Ferreiro et al., 2008, Schafer et al., 2012]. Las trampas cinéticas que dan lugar a la población de intermediarios pueden surgir de la mera topología de la molécula como hemos mostrado anteriormente, debido a empaquetamientos discontinuos, inserciones, deleciones o loops. Alternativamente, la población de intermediarios puede surgir como consecuencia de sitios enriquecidos en interacciones altamente frustradas [Ferreiro et al., 2014].

La Fig. 4.16 muestra los patrones de frustración local de dos ejemplos de proteínas ANK, una diseñada y una natural. Para ambas, los extremos de los arreglos repetitivos se encuentran enriquecidos en interacciones altamente frustradas. La proteína diseñada se encuentra densamente conectada por una red de interacciones mínimamente frustradas a diferencia de la contraparte natural que muestra parches de alta frustración, correspondientes con los sitios de interacción con otras proteínas [Ferreiro et al., 2014]. Dado que estas proteínas poseen arquitecturas elongadas y bajo la hipótesis de que las distintas repeticiones constituyen elementos de plegado, consideramos que es posible obtener un patrón de como se pueblan diferentes rutas de plegado visualizando la plegabilidad relativa de sus fragmentos [Panchenko et al., 1996, Tsai et al., 2000]. Para obtener dicho patrón, la energía según la función de energía de AWSEM [Schafer et al., 2014] de cada posible fragmento continuo es computada y comparada con la energía de fragmentos de su mismo largo en la proteína (mismo procedimiento usado anteriormente para determinar la fase de las repeticiones ANK). Al calcular la plegabilidad para todos los posibles fragmentos presentes en la estructura, se obtiene un patrón que nos informa por un lado a un largo determinado, cuáles regiones son más favorables para ser plegadas dada su arquitectura nativa. Pero además al observar tamaños inmediatamente más largos o más cortos, podemos intuir hacia que dirección conviene extender o reducir el largo de los fragmentos para mejorar o empeorar la plegabilidad resultante. Así al evaluar el

patrón global, pueden observarse diferentes regiones de alta y baja plegabilidad en donde el fragmento de plegabilidad máxima es representado en todos los casos por la estructura nativa global. Del mismo modo en que los patrones de teselado son útiles para observar los efectos que pequeñas perturbaciones pueden tener a mayores escalas, los patrones de plegabilidad relativa jerárquica son útiles para evaluar como se propaga el efecto de ciertas interacciones altamente frustradas hacia dominios o regiones completas dentro de la estructura.

En los casos presentados en la Fig. 4.16 es aparente que ambas proteínas pueden tener patrones de plegado complejos con múltiples rutas, por ejemplo en el caso de la diseñada se espera que pueble rutas paralelas nucleadas en las repeticiones centrales (Fig. 4.16, izquierda). En contraste, la plegabilidad de la proteína natural I$\kappa$B$\alpha$ parece indicar una polarización desde el fragmento N terminal hacia el C-terminal, en donde la repetición más cercana al extremo N-terminal sería la última en plegarse. Tal vez, no sea coincidencia que el plegado de esta proteína haya sido observado como polarizado de forma experimental, con un plegado acoplado al proceso de reconocimiento de la proteína con la que interactúa [Lamboy et al., 2013]. Así, las trampas que aparecen en el plegado de I$\kappa$B$\alpha$ deben tener una contribución de los efectos topológicos locales, como sugieren simulaciones recientes, en donde la presencia de desorden en regiones de algunas repeticiones inicia un efecto del tipo dominó desestabilizando regiones circundantes, en efecto mostrando rupturas en la simetría de la molécula a nivel primario [Sivanandan and Naganathan, 2013].

## 4.8. Conclusiones del capítulo

Los modelos basados en estructura han sido aplicados ampliamente en diferentes arquitecturas proteicas para analizar cómo la topología de las mismas afectan el proceso de plegado, pudiéndose describir propiedades biofísicas del mismo [Shoemaker and Wolynes, 1999, Clementi et al., 2000, Levy et al., 2004, Yang et al., 2004, Oliveberg and Wolynes, 2005, Levy et al., 2005, Cho et al., 2008, Cho et al., 2009]. Hemos simulado el proceso de plegado de estructuras correspondientes a miembros de la familia ANK con diferentes largos de arreglos repetitivos yendo desde un mínimo de 3 repeticiones para proteínas diseñadas hasta un máximo de 6 repeticiones en el caso de la proteína I$\kappa$B$\alpha$. Si bien la secuencia codifica la estructura prote-

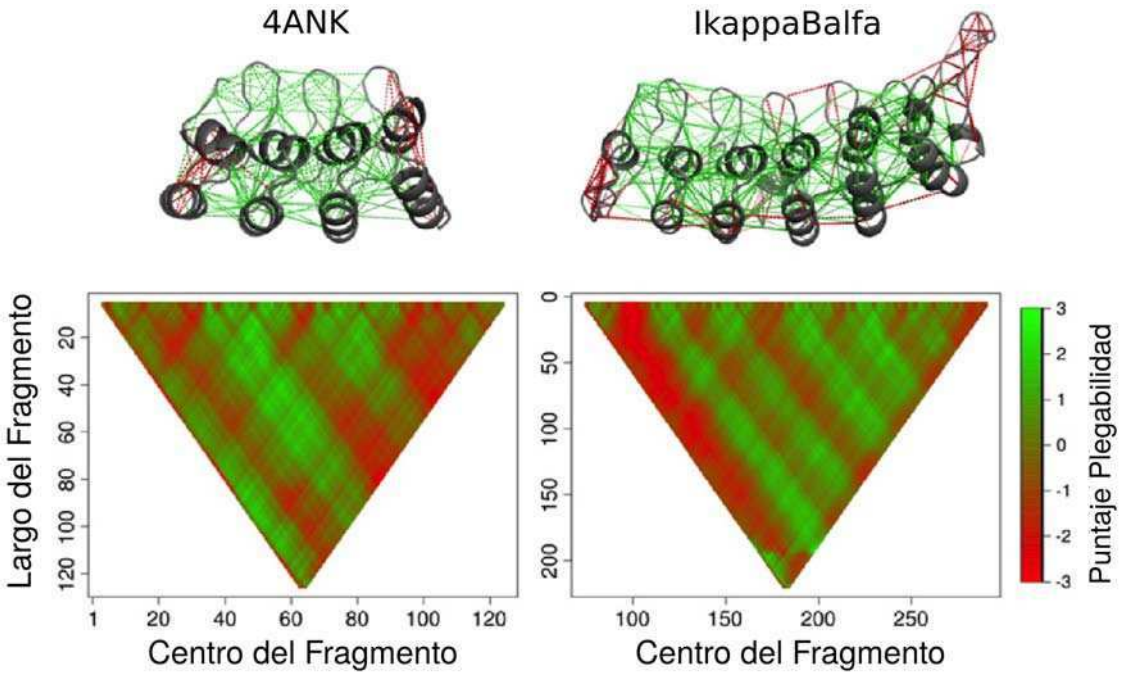


**Figura 4.16:** Frustración local y rutas de plegado: Los mapas representan la plegabilidad jerárquica relativa de cada posible fragmento continuo de un largo dado (eje y) y centro (eje x). La plegabilidad relativa se define como $\Theta_r = \Delta E/(\delta E\sqrt{N})$, donde $\Delta E$ es la diferencia entre la energía de un fragmento en particular con respecto a la energía media de todos los fragmentos de su mismo largo y $\delta E$ es el desvío estándar de dicha distribución calculada a partir de N fragmentos. En la parte superior se muestra una representación de la estructura de las proteínas 4ANK en la izquierda y de I$\kappa$B$\alpha$ en la derecha. Los patrones de frustración local se muestran sobre las estructuras en donde las líneas verdes corresponden a interacciones mínimamente frustradas y las líneas rojas a interacciones altamente frustradas.

ica, los modelos aquí aplicados no tienen en cuenta a la primera en los cálculos, sino que la dinámica sólo depende de la estructura inicial ofrecida como entrada. Todas las interacciones presentes en la estructura nativa se consideran favorables y las interacciones no nativas son consideradas desfavorables, eliminando así toda la contribución de la frustración energética local al plegado. Así, los diferentes eventos descriptos a lo largo de este capítulo están relacionados con las variaciones estructurales presentes en las diferentes estructuras.

En el caso de estructuras con 3 repeticiones, observamos que la correspondiente a 3ANK que es una proteína completamente consenso posee una barrera energética de 17.5 Kcal/mol en comparación con $NI_1C$_Mut4 que posee una barrera de 13 Kcal/mol y fue sintetizada usando repeticiones terminales adaptadas para mejorar sus propiedades de estabilidad usando como modelo las repeticiones terminales de la proteína GABPB1 [Binz et al., 2003]. Ambas proteínas muestran mecanismos de plegado compatibles con el modelo de 2 estados típico para proteínas globulares. Ambas muestran un evento de nucleación del plegado que involucra la

repetición central y parte de la repetición N-terminal. Se observa además que los terminales se deshilachan siendo más favorable desplegar una de las repeticiones terminales que mantener el arreglo entero plegado (Fig. 4.4).

En cuanto a proteínas con arreglos repetitivos compuestos por 4 repeticiones, hemos analizado 4 miembros de la familia. La proteína 4ANK, también una proteína completamente consenso es la que muestra la barrera energética más alta del grupo (17.34 Kcal/mol). P164INK con una barrera de 10.3 Kcal/mol posee regiones no repetitivas en sus extremos, lo cual se traduce en un estado plegado estabilizado en valores de Qw cercanos a 0.55, dado que es desfavorable energéticamente mantener regiones de loop en su conformación nativa a lo largo de la dinámica, debido a la falta de una gran densidad de contactos. Por su parte la versión truncada de Notch, con una barrera de 9.05 Kcal/mol, es la proteína del grupo que muestra en su perfil de energía libre lo que es muy probablemente un intermediario de plegado que posee su repetición N-terminal desplegada, la cual además posee en su extremo N-terminal una región helicoidal extra. I$\kappa$B$\alpha$ es la proteína con la menor barrera energética, para la cual el estado con todas sus repeticiones plegadas no se observa estable en su perfil energético. En cambio, el estado con mayor estructuración y estabilidad comparable a la del estado desplegado, corresponde a confórmeros en que la repetición C-terminal se encuentra desplegada. Observamos además que cercano al estado nativo hay confórmeros con valores de energía libre relativamente bajos en que ambas repeticiones terminales se encuentran desplegadas. De esta forma el alejamiento de un comportamiento 2 estados en todos los casos involucra estructuras en las que las regiones terminales se encuentran parcial o totalmente desplegadas. La nucleación del plegado ocurre involucrando una repetición y algunos residuos de una repetición vecina. A pesar de sus similitudes estructurales la nucleación no ocurre en regiones análogas en las moléculas analizadas. Mientras Notch y I$\kappa$B$\alpha$ comienzan a plegarse por las repeticiones centrales (de formas diferentes),4ANK tiene el núcleo de plegado desplazado hacia el extremos C-terminal y P164INK hacia el extremo N-terminal (Fig. 4.7).

En el caso de las proteínas analizadas con 5 repeticiones. En todos los casos vemos barreras energéticas más homogéneas con E3_5 mostrando la mayor barrera (11.16 Kcal/mol) y TV1425 la menor (8.92 Kcal/mol). Si bien todas muestran un estado nativo amplio, no se observan intermediarios o confórmeros cercanos con valores de energía libre menores a 5 Kcal/mol. Como

en los casos anteriores, estas proteínas exhiben distintos núcleos de plegado. E3_5 comienza a plegar a partir de sus repeticiones 3 y 4 al igual que GABPB1 aunque esta última involucra algunos residuos de R2. TV1425, en cambio, comienza su plegado a partir de sus repeticiones localizadas en el extremo C-terminal mientras que ANKRA2 lo hace a partir de su repeticiones en el extremo N-terminal (Fig. 4.11).

Por último I$\kappa$B$\alpha$ fue la única proteína analizada con 6 repeticiones mostrando el paisaje energético más complejo de todas las proteínas. Caracterizamos en las simulaciones que existe un intermediario estable que posee las 3 últimas repeticiones desplegadas. Además se observan 2 regiones, una cercana al estado plegado y otra cercana al estado intermediario donde se observan conformaciones con valores de energía libre cercanos a las 5 Kcal/mol. Las estructuras muestreadas cerca del estado nativo poseen la repetición N-terminal desplegada mientras que aquellas cercanas al estado intermediario poseen plegada dicha repetición, con una desestructuración parcial de la región C-terminal y perdida de las interacciones de la interfaz entre las repeticiones 4 y 5.

Todas estas proteínas son muy parecidas estructuralmente siendo sus mayores diferencias la presencia de inserciones/deleciones o regiones no repetitivas en sus extremos. Estas modificaciones, conjuntamente con las desviaciones estructurales producto de las variaciones en secuencia producen alteraciones en el empaquetamiento de repeticiones vecinas alterando el balance de las estabilidades tanto internas de cada repetición como la energética de la interacción entre repeticiones adyacentes. Esto puede dar lugar a la estabilización de intermediarios, lo cual se observó en algunos casos. Más allá de la presencia de intermediarios de plegado observables, estas diferencias estructurales pueden modificar las regiones que actúan como focos tempranos de plegado y generar que las diferentes repeticiones actúen como unidades independientes de plegado o que las distintas repeticiones se segreguen en grupos a lo largo del proceso de plegado. Para caracterizar esto realizamos cálculos de la fracción de interacciones nativas que se establecen a nivel de residuo a lo largo del plegado y también aplicamos un modelo discreto cinético para observar las estabilidades relativas de las distintas unidades de plegado (definidas como las repeticiones detectadas mediante el método de teselado) y las constantes de transición entre los diferentes macroestados de los arreglos repetitivos (definidos como la combinación de repeticiones en estados binarios equivalentes a que las mismas estén plegadas

o desplegadas). Lo más simple es comenzar describiendo los mecanismos observados para las proteínas diseñadas, que carecen de inserciones o regiones no repetitivas. Las proteínas de 3 repeticiones en ambos casos comienzan a plegarse principalmente por su repetición central, con una contribución parcial de la repetición N-terminal y una estructuración nativa tardía de la repetición C-terminal. Ambas proteínas muestran leves eventos de retroceso en su repetición N-terminal siendo este más acentuado en 3ANK que en $NI_1C$_Mut4. 4ANK por su parte, comienza a plegar sus 2 repeticiones centrales y su repetición C-terminal la cual muestra un evento claro de retroceso. 3ANK y 4ANK están diseñadas de la misma forma y sus repeticiones son completamente consenso, con una versión C-terminal más corta a la que le faltan los residuos correspondientes al $\beta$ hairpin. Sin embargo a pesar de que lo único que cambia es que 4ANK posee una repetición interna extra, sus mecanismos de plegado son muy diferentes. La proteína sintética más larga simulada fue E3_5 con 5 repeticiones. Esta proteína comienza a plegar por sus repeticiones 3 y 4 y posee un comportamiento en el C-terminal similar a 4ANK, sólo que con dos eventos de retroceso en vez de 1. Las proteínas naturales analizadas muestran comportamientos diversos, algunas comienzan a plegar por las regiones centrales como es el caso de Notch truncada, I$\kappa$B$\alpha$ truncada, GABPB1 mientras otras parecen tener un plegado polarizado desde un extremo al otro. P164INK, ANKRA2 y la versión completa de I$\kappa$B$\alpha$ parecen plegar de forma polarizada desde el extremo N-terminal hacia el C-terminal mientras que TV1425 lo hace en sentido contrario. En todos los casos en que se observa que una de las repeticiones terminales se pliega de forma temprana, se observan también eventos de retroceso en el mismo de forma más o menos acentuada. Es notable también que en todos los casos en que se observan estos eventos, al mismo valor de Qw global estos eventos coinciden con el comienzo de plegado de otras regiones lejanas en el arreglo repetitivo. En 3ANK y $NI_1C$_Mut4, estos eventos en la repetición N-terminal coinciden con el momento en que se comienza a plegar el C-terminal. En 4ANK, el evento de retroceso en la repetición C-terminal coincide con el valor de Qw global en que comienza a plegarse la repetición N-terminal lo cual sucede de forma similar en Notch truncada. En E3_5 el primer evento de retroceso coincide con el comiendo de plegado de la repetición 2 y el segundo evento de retroceso coincide con el momento en que comienza a plegarse la repetición N-terminal. Comportamientos similares se producen en TV1425 y la versión completa de I$\kappa$B$\alpha$. En esta última molécula se puede

observar en su perfil de energía libre la existencia de un intermediario y dos ensambles de conformaciones con relativamente baja energía. Mientras que en la región cercana al estado nativo se observa la repetición N-terminal desplegada, en la región del intermediario (y la región circundante) se observa dicha repetición plegada lo cual pudimos ver en el modelo cinético discreto de esta molécula que se corresponden a estados visitados a través de diferentes rutas que conectan el estado plegado con el intermediario. ¿Están los eventos de retroceso de repeticiones terminales relacionados a rutas paralelas en el paisaje energético? ¿Podrá ser que los eventos de retroceso se encuentren relacionados a interacciones de larga distancia no evidentes entre repeticiones alejadas y por tanto ayudar a que otras regiones del arreglo se plieguen? Análisis en los que se hace un seguimiento de los valores Qw de las que muestran eventos de retroceso y las que pliegan a los valores de Qw en que estos ocurren nos podrán dar información acerca de la concomitancia o no de estos eventos y sin son uno consecuencia del otro.

Es interesante notar que las repeticiones definidas a partir de los cristales mediante nuestro procedimiento de teselado, corresponden en gran medida con las regiones que se pliegan de forma cooperativa dentro de las dinámicas realizadas, lo que es compatible con el concepto de foldones definido por Panchenko [Panchenko et al., 1996].

Los modelos cinéticos muestran que las repeticiones terminales son en la mayoría de los casos proclives a encontrarse desplegadas en los macroestados más estables de forma global, lo cual podría ser explicable debido a que los mismos tienen una interfaz menos con repeticiones adyacentes respecto de los internos. Esto ha sido descripto para casos como en el 4ANK en donde la repetición C-terminal se desestructura en un proceso descripto como deshilachado (*fray* en inglés) [Mosavi et al., 2002]. Esto también se ha observado mediante estudios de NMR [Cortajarena et al., 2008], intercambio de hidrógeno [Main et al., 2005] y simulaciones con modelos de *ising* [Kajander et al., 2005] en miembros de la familia repetitiva TPR .

# Capítulo 5

# Métodos

## 5.1. Construcción de una base de datos para ANKs:

Trabajar con un conjunto de datos del cual se quiere extraer información pero que además incluye aportes de diferentes fuentes, requiere de un soporte tecnológico que ayude a su manejo a medida que el volumen de los datos aumenta.

A los fines de poder hacer análisis a nivel de familia, se construyó una base de datos relacional usando MySQL en la cual se incluyeron diferentes entidades y atributos relacionados a los miembros de la familia ANK que luego serían necesarios para realizar los estudios subsiguientes.

### Modelo Entidad-Relación o ER:

El modelo ER describe los datos como entidades, relaciones y atributos. A continuación se realiza una descripción acotada de los aspectos básicos para comprender la estructura de un modelo de entidad-relación con el único objetivo de dar fundamento conceptual a la estructura de una base de datos relacional.

**Entidades y sus atributos.** El objeto básico representado por el modelo ER es una entidad, que es una cosa del mundo real con una existencia independiente. Una entidad puede ser un objeto con una existencia física (por ejemplo, una persona en particular, un coche, una casa o un empleado) o puede ser un objeto con una existencia conceptual (por ejemplo, una empresa, un trabajo o un curso universitario). Cada entidad tiene atributos (propiedades

particulares que la describen). Por ejemplo, una entidad EMPLEADO se puede describir mediante el nombre, la edad, la dirección, el sueldo y el trabajo que desempeña. Una entidad en particular tendrá un valor para cada uno de sus atributos. Los valores de los atributos que describen cada entidad se convierten en la parte principal de los datos almacenados en la base de datos. En el modelo ER se dan varios tipos de atributos: simple o compuesto, mono valor o multivalor, y almacenado o derivado.

**Atributos clave de un tipo de entidad:**

Una restricción importante de las entidades de un tipo dado es la clave o restricción de unicidad de los atributos. Una entidad normalmente tiene un atributo cuyos valores son distintos para cada entidad individual del conjunto de entidades. Dicho atributo se denomina atributo clave, y sus valores se pueden utilizar para identificar cada entidad de forma unívoca.

**Relaciones entre entidades:**

Dados dos tipos de entidades, estas pueden estar asociadas definiendo una relación $r_i$. Dicha asociación incluye exactamente una entidad de cada tipo de entidad participante. Cada una de dichas instancias de relación $r_i$ representa el hecho de que las entidades que participan en $r_i$ están relacionadas de alguna forma en la situación correspondiente del mini mundo. Por ejemplo, considere un tipo de relación TRABAJA PARA que asocia una entidad EMPLEADO con una entidad DEPARTAMENTO.

La cardinalidad indica el número de entidades con las que puede estar relacionada una entidad dada. Existen 3 tipos de cardinalidades:

Uno a Uno: La entidad A se relaciona únicamente con la entidad B y viceversa. Uno a Varios: La entidad A se relaciona con cero o varias entidades en B, pero una entidad B se relaciona con una única entidad A. Varios a Varios: La entidad A se relaciona con cero o varias entidades en B y viceversa.

La Fig. 5.1 muestra un ejemplo de diagrama de entidad-relación típico.

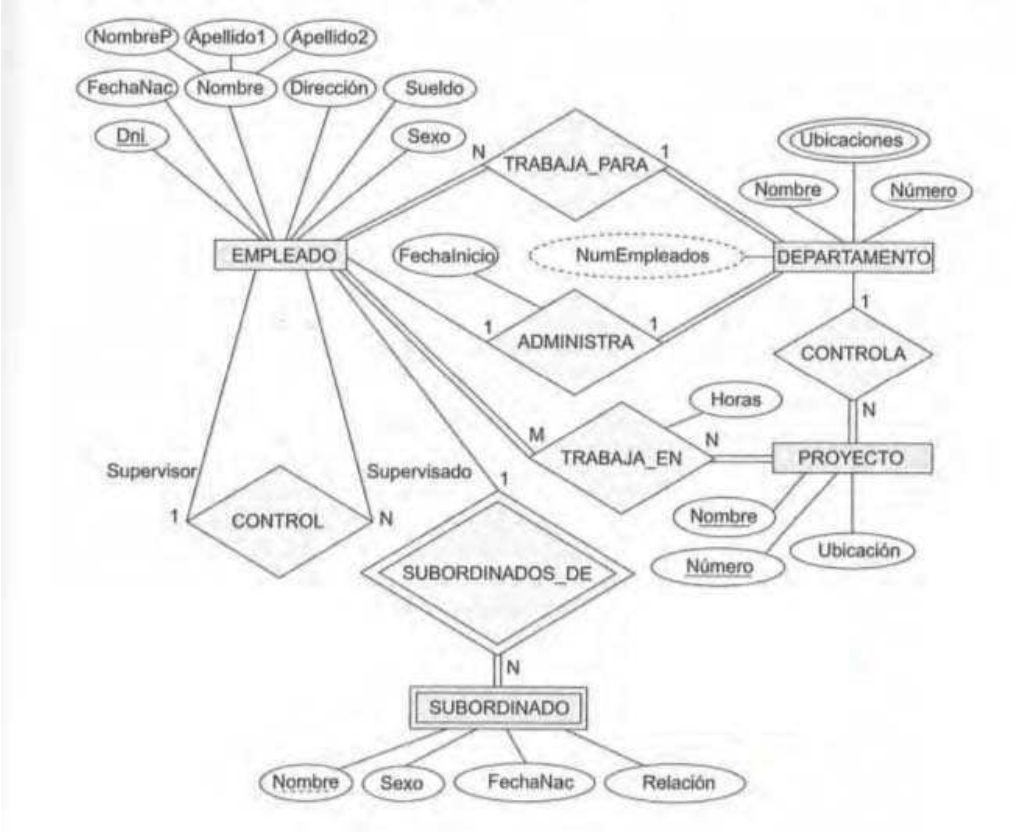


**Figura 5.1:** Ejemplo de diagrama de entidad relación donde pueden verse los diferentes elementos que los componen.

## Base de datos estructural de ANKs

Siguiendo los lineamientos descriptos anteriormente, diseñamos una estructura de base de datos para nuestra base de datos estructural de ANKs. Hemos generado un diagrama de Entidad-Relación (DER) siguiendo las buenas prácticas, haciendo especial énfasis en evitar la redundancia excesiva de datos entre tablas. La estructura de nuestro DER consiste de 23 tablas (Fig. 5.2).

**Descripción general de la DB:**

Se descargaron todas las estructuras del PDB (*Protein Data Bank*, www.pdb.org), en donde al menos una de las cadenas contiene repeticiones ANK. Los datos relevantes acerca de cada estructura se almacenaron en la tabla **Structure** y los datos de la publicación en que la estructura aparece reportada se almacenan en la tabla **Publication** si es que existe (algunas estructuras como las que son generadas por consorcios, como el *Protein Structure Initiative*,http://sbkb.org/about/about-psi, dedicados a la dilucidación de estructuras,

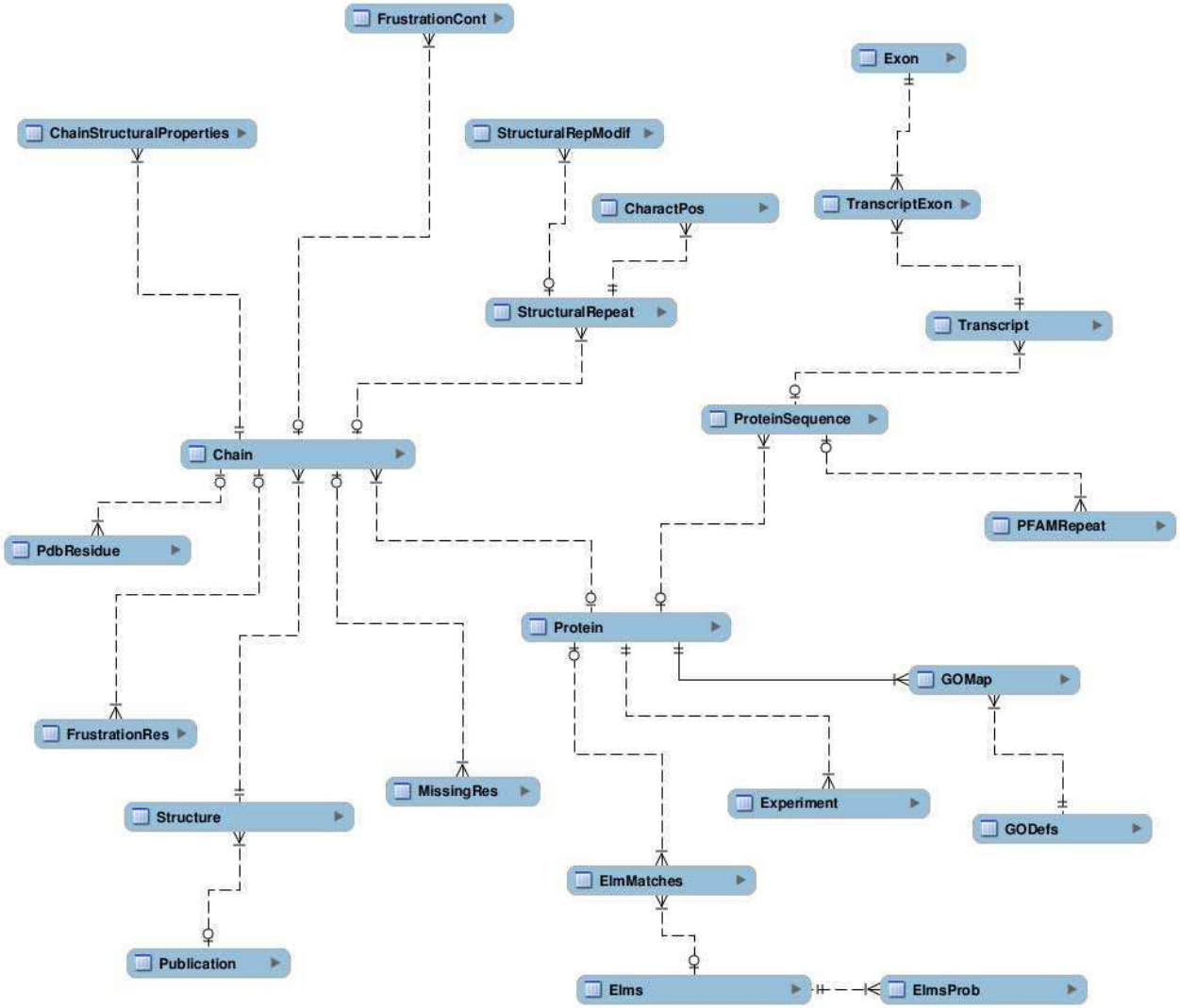


**Figura 5.2:** Diagrama de Entidad-Relación de nuestra base de datos estructural de ANKs

no poseen publicaciones asociadas). Cada estructura está compuesta por un número variable de cadenas que pueden corresponder a proteínas, péptidos o ácidos nucleicos. Cada instancia correspondiente a una cadena en una estructura, se almacena junto a los parámetros relevantes en la tabla **Chain**. La entidad Chain, contiene además otras tablas que se derivan de la misma a los fines de evitar redundancia debido a la aparición de parámetros complejos. Ejemplos de estas tablas derivadas son **ChainStructureProperties** que contiene los valores para características estructurales de las estructuras como el radio de giro, la accesibilidad al solvente y otras. Otro ejemplo son las tablas FrustrationRes y FrustrationCont, que debido a la cantidad de información que poseen y su relevancia para nuestros estudios, se constituyen como entidades propias. Además de las mencionadas también guardamos la información acerca de los residuos ausentes en la tabla MissingRes como así también un mapeo entre las posiciones de cada cadena almacenada en Chain y su correspondencia con los residuos en la entidad correspondiente a Uniprot dentro de la entidad **Protein**. Por último, tenemos a la tabla **StructuralRepeat** que corresponde a las repeticiones que hemos detectado a partir de un método estructural de identificación de estos elementos estructurales (explicado más adelante, en otro capítulo) del cual derivan dos tablas que corresponden a la descripción de las posiciones canónicas de cada repetición, sus propiedades y su correspondencia a la entidad en Uniprot almacenadas en la tabla **CharactPos** y a las regiones que no forman parte de la región repetitiva, sino que representan inserciones (intra o inter repeticiones) o deleciones en el arreglo repetitivo almacenadas en la tabla **StructuralRepModif**.

Como mencionamos, cada cadena esta relacionada casi siempre (exceptuando a los constructos artificiales que no poseen una entrada en Uniprot) a una instancia de la entidad Protein. Al igual que Chain, la tabla **Protein** constituye un nodo central en el DER ya que se relaciona con muchas otras tablas. La tabla **ProteinSequence** contiene todos los datos relativos a la secuencia de la proteína derivada de Uniprot y también a su vez se encuentra relacionada a la tabla Transcript, que contiene información de los transcriptos que se relacionan con dicha proteína en la base de datos de Ensembl (www.ensembl.org) que a su vez se relaciona a través de una tabla intermedia (**TranscriptExon**) con la tabla **Exon**. La tabla PFAMRepeats contiene las instancias de repeticiones detectadas en las proteínas correspondientes a la familia ANK, detectadas a partir del uso de la herramienta HMMER usando el

Modelo Oculto de Markov que representa al dominio ANK en la base de datos de PFAM. Cada instancia dentro de la tabla Protein además se encuentra relacionada a instancias en la tabla **GoDefs** a través de la tabla **GOMap** que representa el mapeo entre las instancias de proteínas y sus términos de ontología génica (http://geneontology.org/) relacionados, derivados de la base de datos de AmiGO (amigo.geneontology.org/). La tabla **Experiment** contiene información de ensayos experimentales relacionados a plegado de proteínas ANK del tipo *wildtype* así como también de mutantes y constructos, recopilados de bibliografía (curación manual) y de bases de datos públicas como ProTherm (http://www.abren.net/protherm/). Por último, la información referente a la ocurrencia de motivos lineales, detectados a partir de expresiones regulares presentes en la base de datos de motivos lineales eucariotas (ELM, elm.eu.org/) se encuentran almacenados en la tabla **Elms** a través de la tabla intermedia **ElmMatches** mientras que una serie de parámetros ajustados a la distribución de aminoácidos en la familia ANK, se encuentran almacenados en la tabla **ElmsProb**.

Una vez diseñado el DER, se generó una base de datos basada en el mismo usando un motor de MySQL. Al crearse la estructura, el paso siguiente fue ingresar los datos para cada entidad, con sus diferentes parámetros y datos derivados. La cantidad de información relacionada a cada estructura, hace necesario la automatización del proceso de ingreso de los datos y actualización de los mismos. Se generaron múltiples scripts en Perl que se encargan de descargar la información primaria de los sitios en internet correspondientes, procesar la información y ejecutar todos los programas necesarios para los cálculos y derivación de los datos secundarios. Como resultado final, se obtuvo un pipeline, para el ingreso de los datos que solo recibe como input, el ID de la estructura almacenada en la base de datos del Protein Data Bank y que como resultado ingresa todos los datos correspondientes a cada entidad correspondiente al DER de la base de datos en MySQL.

## 5.2. Frustración Energética Local

Las regiones de las proteínas que promueven el plegado proteico (el núcleo hidrofóbico) generalmente se encuentran bien empaquetadas y se espera que estén mínimamente frustradas. En contraste, las regiones funcionales en general se encuentran altamente frustradas.

Las proteínas naturales representan sistemas complejos altamente moldeados por la evolución [Frauenfelder, 2002]. Según la teoría de paisajes energéticos, las proteínas se pliegan minimizando sus conflictos energéticos, lo cual es posible gracias a la cooperatividad existente entre sus interacciones. Existe un fuerte sesgo energético para plegarse hacia estructuras similares al estado nativo. Esto se conoce como principio de mínima frustración [Bryngelson and Wolynes, 1987]. Este principio no implica que ciertos conflictos pueden persistir en los estados nativos de las proteínas. De hecho se ha mostrado a lo largo de estos años que la presencia de conflictos energéticos presentes en los ensambles nativos de las proteínas tienen consecuencias funcionales en múltiples aspectos facilitando el moviendo de la proteína en el espacio conformacional.

El Frustratometer es un método [Jenik et al., 2012] basado en la teoría de paisajes energéticos que implementa los métodos teóricos desarrollados para cuantificar el grado de frustración en estructuras proteicas [Ferreiro et al., 2007b]. Este método ha sido útil en el estudio de interfaces de interacción entre proteínas [Ferreiro et al., 2007b], transiciones alostéricas [Ferreiro et al., 2011], agregación y unión de ligandos [Gianni et al., 2014, Das and Plotkin, 2013], dinámica conformacional [Sutto et al., 2007, Fuglestad et al., 2013, Truong et al., 2013], ha sido relacionado con patrones evolutivos [Abriata et al., 2012, Tripathi et al., 2015, Galen et al., 2015] y polimorfismos en enfermedades [Dixit and Verkhivker, 2011].

El Frustratometer localiza la frustración local de la siguiente manera: Dados dos residuos que están en contacto en una estructura proteica, se generan 1000 señuelos (*decoys*) en los se modifican ciertos parámetros correspondientes a la interacción nativa. Haciendo uso del potencial AMW (*Associative Memory Hamiltonian optimized with Water-mediated interactions*, [Papoian et al., 2004]) se mide la energía de la interacción nativa y la de los señuelos y se calcula el índice de frustración como un *Z-score* $Fi = \frac{E_N - < E_D >}{\sigma D}$ donde $E_N$ corresponde a la energía nativa, $< E_D >$ a la media de la distribución de energía de los señuelos y $\sigma D$ a su desvío estándar.

La función AMW tiene en cuenta 3 términos: $H_{contacts}, H_{water}, H_{burial}$ que dependen de las identidades ($\lambda$) de los aminoácidos, las densidades ($\rho$) (relacionadas a la accesibilidad del solvente) y la distancia de interacción ($r_{ij}$). Existen 3 variantes para el índice de frustración local, basado en como se generan los señuelos.

**1) Mutacional:** Dados dos residuos en contacto, se generan los señuelos aleatorizando los valores $\lambda$.

**2) Configuracional:** Dados dos residuos en contacto, se generan los señuelos aleatorizando $\lambda$ y $r_{ij}$ y $\rho$.

**3) Residuo único:** Dado un residuo $i$, se calcula su frustración aleatorizando sólo la identidad $\lambda$ del mismo, manteniendo el resto de los parámetros nativos.

Los contactos se clasifican en mínimamente frustrados ($Fi >= 0{,}78$), neutros ($-1 < Fi < 0{,}78$) y altamente frustrados ($Fi <= -1$) para el caso de los índices mutacional y configuracional. Los residuos se clasifican en mínimamente frustrados ($Fi >= 0{,}58$), neutros ($-1 < Fi < 0{,}58$) y altamente frustrados ($Fi <= -1$) para el caso del índices de residuo único.

Los contactos se definen según cortes en distancias entre C-$\beta$ de pares de residuos. Se tienen en cuenta 3 tipos de contactos: 1) de corto alcance ($r_{ij} <= 6{,}5\mathring{A}$) 2) Largo alcance ($6{,}5\mathring{A} < r_{ij} < 9{,}5\mathring{A}$) 3) Mediados por agua (mismo corte de distancias que los de largo alcance, pero ambos residuos deben estar expuestos al solvente, sasa>0.05).

## 5.3. Contenido de Información

El contenido de información (CI) calculado a partir de alineamientos múltiples de secuencias, mide la conservación de los aminoácidos en posiciones específicas de éstos y es la medida que típicamente se muestra en el eje de ordenadas en los logos de secuencias. El CI puede calcularse para medir la conservación de los estados de cualquier variable discreta en un conjunto. En nuestro caso calculamos el CI tanto para medir la conservación de las identidades de los aminoácidos como para los estados del índice de frustración local. Se utilizó el procedimiento descripto por Schneider [Schneider et al., 1986] para calcular los valores de CI tanto de secuencias (IC Seq) como de frustración (IC Frst).

El contenido de información IC por posición o por contactos puede pensarse como la reducción de la incerteza respecto del máximo posible ($H_{max}$). La incerteza observada en un sistema, de acuerdo a la definición de Shannon [Shannon, 1997] es $H_{observed} = -\sum_{i=1}^{M} p_i log_2 p_i$, donde $p_i$ es la probabilidad de que el sistema se encuentre en el estado $i$. En este trabajo los estados

son las identidades de los 20 aminoácidos para ICSeq y los estados mínimo/neutra/altamente frustrado para ICFrst. Las probabilidades son normalizadas de forma que $\sum_{i=1}^{M} p_i = 1$, donde $M$ es el tamaño del alfabeto ($M = 20$ para ICSeq y $M = 3$ para ICFrst). El valor de CI se obtiene entonces de la forma $CI = H_{max} - H_{observed}$. En general se considera que $H_{max}$ se obtiene para una distribución equiprobable de los estados en donde $p_i^{max} = 1/M$ y $H_{max} = log_2(M)$. Si se sabe que la distribución de los estados es heterogénea se debe usar la frecuencia relativa de los estados para que $H_{max} = \sum p_{ib} log_2(p_{ib})$, donde $p_{ib}$ es la frecuencia para el estado $i$ de referencia, a diferencia de $p_i$ que es la frecuencia observada para el estado $i$ en los datos.

## 5.4. Coordenada Qw

Para seguir la evolución de las transiciones conformacionales a lo largo de las trayectorias de dinámica molecular se utilizo la coordenada Qw, que representa la fracción de interacciones nativas de una molécula respecto de la estructura experimental usada como entrada para la simulación.

$$Qw = \frac{1}{N_p} \sum_i \sum_j exp[\frac{-(r_{ij} - r_{ij}^u)^2}{2\sigma_{ij}^2}] \tag{5.1}$$

Los subíndices i y j representan las posiciones de los aminoácidos i y j en la cadena polipeptídica. $N_p$ representa el número total de pares (ï,j). $r_{ij}$ es la distancia entre los C$\alpha$ de los residuos i y j, $r_{ij}^u$ es la misma distancia pero medida en la estructura nativa de referencia determinada experimentalmente.

## 5.5. Simulaciones de Dinámica Molecular del tipo grano grueso

Para las dinámicas moleculares de tipo grano grueso se utilizó la plataforma AWSEM-MD. En particular se usó un modelo SBM llamado AMH-G$\bar{o}$ que tiene la particularidad de incluir un término de no aditividad que permite simular de forma más realista la cooperatividad entre

las interacciones nativas, sobre todo al muestrear los estados de transición .

Para la estimación de la $T_f$ se realizaron simulaciones de desnaturalización (*Melting*) que consiste en evolucionar el sistema llevándolo desde una temperatura baja (300 K) hasta temperaturas lo suficientemente altas para asegurar el desplegado del mismo (800 K). Las variaciones de temperatura se realizaron a lo largo de 10 millones de pasos entre los cuales el lapso de tiempo simulado es 3 femto segundos (fs). Durante este procedimiento debido a la cooperatividad entre las interacciones nativas, al observarla evolución de Qw en función de la temperatura, se observa una rápida transición entre el estado plegado y el desplegado en un corto rango de temperaturas. La $T_f$ corresponde al punto medio de la pendiente de transición entre ambos estados.

Una vez estimada la $T_f$ se realizan simulaciones a esta temperatura usando el método de muestreo de *Umbrella Sampling* usando como coordenada colectiva que toma valores entre 0 (totalmente desplegado) y 1 (totalmente plegado). El método de *Umbrella Sampling*, funciona anclando un potencial en un determinado valor de Qw, lo cual permite explorar aquellas estructuras que posean valores cercanos a dicho valor. En total dividimos el intervalo de Qw en 40 ventanas equiespaciadas en las cuales hemos anclado potenciales *umbrella*. Para cada intervalo de Qw se simula el sistema durante 10 millones de pasos cada 3 fs. Las 40 trayectorias obtenidas para los diferentes sesgos de *umbrella* son integradas mediante el método WHAM (*Weighted Histogram Analysis Method*) obteniéndose de esta forma los valores de Energía Libre proyectados en diferentes coordenadas como Qw o el Radio de Giro (Rg). Información general acerca de como realizar dinámicas de este tipo y de otras metodologías puede leerse en el artículo [Schafer et al., 2014].

## 5.6. Modelo AMH-G$\bar{o}$ No aditivo

Este modelo se encuentra descripto en [Eastwood and Wolynes, 2001] y se explican a continuación algunos aspectos básicos del mismo.

Este es un modelo de cadena explícita, de grano grueso, basado en estructura y no aditivo. El modelo considera 3 átomos por residuo (C$\alpha$, C$\beta$ y O, las posiciones de los átomos N y C' son calculadas asumiendo una geometría ideal para el *backbone*) para aliviar la carga com-

putacional y no tiene en cuenta de forma explícita las moléculas del solvente. Las interacciones atractivas corresponden únicamente a aquellas que se encuentran presentes en la estructura determinada de forma experimental de la molécula simulada (interacciones nativas). Todas las interacciones nativas son homogéneas, es decir, tienen la misma fuerza dentro del modelo independientemente de la identidad de los aminoácidos. Toda interacción no presente en la estructura nativa se considera no nativa y por tanto de carácter repulsivo. Esto hace que el paisaje energético se represente como un embudo perfecto sin rugosidad ya que se elimina la frustración energética del mismo. A pesar de que en la realidad las interacciones no nativas pueden ocurrir, su efecto principal es el de agregar una fuente adicional de fricción frenando la progresión de la molécula a través de los ensambles parcialmente plegados [Bryngelson and Wolynes, 1989, Wang et al., 1996].

La estructura general del Hamiltoniano de este modelo [Eastwood and Wolynes, 2001] está compuesta por un término que modela la geometría del *backbone* de la cadena polipeptídica y otro término que modela las interacciones entre los residuos de manera no aditiva (Ecuación 5.2).

$$H = H_{backbone} + H_{na} \tag{5.2}$$

El término $H_{backbone}$ incluye varios términos que aseguran que el *backbone* adopte conformaciones físicamente válidas.

La componente energética correspondiente a $H_{na}$ depende de términos de interacción del tipo Gaussiano para las interacciones nativas (Ecuación 5.3).

$$H_{na} = -\frac{1}{2}\sum i|E_i|^p \tag{5.3}$$

donde:

$$E_i = \sum_j \epsilon_{ij}(r_{ij}) = -\sum_j |\frac{\epsilon}{a}|^{\frac{1}{p}} \theta(r_c - r_{ij}^N)\gamma_{ij} \exp(-\frac{(r_{ij} - r_{ij}^N)^2}{2\sigma_{ij}^2}) \quad (5.4)$$

Los índices $i$ y $j$ iteran sobre todos los átomos C$\alpha$ y C$\beta$, $r_{ij}$ es la distancia entre los átomos $i$ y $j$. El exponente $p$ representa la fuerza de la no aditividad del modelo. Valores mayores de $p$ resulta en una mayor cooperatividad en la interacción de varios cuerpos y una mayor barrera en la transición del plegado. Para este estudio se utilizó un valor $p$=2 (nótese que un valor de $p$=1 correspondería a un modelo completamente aditivo). $r_c$ es un parámetro de corte que a partir de la función $\theta(r_c - r_{ij}^N)$ asegura que sólo interacciones entre sitios más cercanos que ese corte se encuentran presentes en la estructura nativa. Aquí se usó un valor de $r_c$=8$\mathring{A}$. Se usó el valor $\sigma_{ij} = |i-j|^{0,15}\mathring{A}$. Todas las interacciones se consideran homogéneas y por ello el peso para las mismas se fijo para todas ellas en $\gamma_{ij} = 1$. La unidad de energía se define como $\epsilon$ y se define en términos de la energía del estado nativo, excluyendo la componente de *backbone*, $\epsilon = H_{na}/4N$, donde N es el número de residuos. Esta equivalencia se asegura normalizando mediante la constante $a = \frac{1}{8N}\sum_i |\sum_j \gamma_{ij}\theta(r_c - r_{ij}^N)|^p$

## 5.7. Modelo Cinético Discreto

Es un método general para facilitar la interpretación de simulaciones computacionales del plegado proteico. Grupos de residuos son definidos como foldones los cuales son usados para mapear el espacio conformacional de las proteínas en un grupo de macroestados discretos. Las energías libres de los macroestados individuales son calculados así como también constantes de transición entre los mismos en el espacio conformacional. Este modelo se aplicó a partir de simulaciones sesgadas usando un modelo basado en estructura no aditivo. En nuestro trabajo en particular, los foldones se definieron como aquellas repeticiones detectadas de forma estructural por nuestro método de teselado. Se dejaron fuera del análisis las inserciones entre repeticiones o regiones no repetitivas en los extremos N y C-terminales.

Los macroestados se definen como una secuencia de símbolos que pueden tomar el valor de 1 o 0, que significan plegado o desplegado. Un macroestado particular de una proteína con

4 foldones podría ser 1010, en el cual los foldones 1 y 3 se encuentran en el estado plegado y los foldones 2 y 4 se encuentran en estado desplegado. Para determinar el estado plegado o desplegado de un foldón específico se calcula el valor de Qw correspondiente a las interacciones nativas del foldón, presentes en el estado nativo. Un foldón es considerado en su estado plegado si Qw≥0.6.

Una proteína que posee un número total de $N_f$ foldones tienen en total un espacio conformacional de $2^{N_f}$ macroestados posibles. Durante una simulación de plegado, no necesariamente todos los macroestados son observados. A partir de los resultados de las dinámicas realizadas, aplicando sesgos de muestreo de *Umbrella Sampling*, y las definiciones de foldones se calculan las energías libres relativas de los macroestados usando el método MBAR (*Multistate Bennet Acceptance Ratio* [Shirts and Chodera, 2008]).

En cuanto a la conectividad de los macroestados entre sí, el modelo plantea que dos macroestados pueden estar conectados sólo si para transformar uno en otro se requiere cambiar de estado un sólo foldón. La tasa de transición entre los macroestados i y j, $K_{ij}$ se calcula como se muestra en la Ecuación 5.5 donde $\Delta F = F_j - F_i$ es la diferencia de energía libre entre los macroestados i y j, $k_B$=0.001987 kcal/mol/K es la constante de Boltzman, $T$ es la temperatura absoluta y $k_0 = 1\mu s^{-1}$ es la tasa de transición universal que se asume en un modelo de plegado *downhill*. Al calcular todos los valores $K_{ij}$ se construye la matriz $\mathbf{K}$ en la cual los valores de la diagonal se definen de forma de conservar una probabilidad $K_{ii} = -\sum_{i \neq j} K_{ij}$, donde $K_{ij}$ refiere al elemento en la columna $i$gésima y la fila $j$ésima de la matriz $\mathbf{K}$.

$$K_{ij} = \begin{cases} k_0 & \text{if} \quad \Delta F_{ij} < 0 \\ k_0 \exp -\dfrac{\Delta F_{ij}}{k_B T} & \text{if} \quad \Delta F_{ij} \geq 0 \end{cases} \tag{5.5}$$

Más detalles del método y sus aplicaciones a otros sistemas se describen en el articulo de Schafer [Schafer et al., 2012].

# Capítulo 6

# Conclusiones Generales

Las proteínas son las macromoléculas que mayor diversidad presentan dentro de la biósfera. A pesar de ser codificadas de forma lineal en los genomas de los sistemas que integran, su traducción a partir de un alfabeto de 4 especies químicas a uno de 20 (y sus posteriores modificaciones) permite que estas moléculas adopten una enorme cantidad de formas que les confieren la habilidad de realizar múltiples funciones ya sea mecánicas, de andamiaje estructural, catalizar reacciones químicas y principalmente, procesar información dentro de los sistemas de los que forman parte [Bray et al., 1995]. A los fines de poder comprender a estas máquinas moleculares y en última instancia modificar, modular, apagar o activar su funcionamiento es imprescindible mejorar nuestro entendimiento de las relaciones secuencia-estructura-dinámica-función.

Las proteínas con motivos estructurales repetitivos representan un excelente modelo para estudiar las relaciones secuencia-estructura-dinámica-función debido a su bajo orden de contacto, que carecen de interacciones evidentes entre repeticiones que están más lejos que sus vecinas inmediatas. Las proteínas de la familia ANK por su parte, representan una de las familias repetitivas más abundantes [Mosavi et al., 2004] tanto en cuanto a experimentos de mesada húmeda cómo a cantidad de miembros con secuencias y estructuras depositadas en bases de datos públicas. Usando estas moléculas como andamiaje para el diseño de proteínas, usando secuencias consenso a partir de alineamientos múltiples, se han generado moléculas con alta afinidad y especificidad para interactuar con una diversa cantidad de proteínas con múltiples aplicaciones incluyendo terapias para tratar diferentes tipos de cáncer, Alzheimer,

HIV y suministro de drogas [Pluckthun, 2015]. Recientemente, usando diseño de novo se han generado proteínas repetitivas solenoides y no solenoides, que exploran una región del espacio de secuencias completamente no relacionada con las proteínas naturales conocidas avanzando un paso más en el diseño de proteínas respecto del diseño basado en secuencias consenso [Brunette et al., 2015, Doyle et al., 2015, Huang et al., 2016]. Estos avances abren todo un nuevo horizonte para el diseño de proteínas de este tipo, ya que algo deseable para muchos investigadores en el campo del diseño proteico, es usar como andamiaje, proteínas sin funciones biológicas conocidas. Los patrones de frustración sobre estructuras cristalográficas derivadas de estos estudios, muestran una gran densidad de interacciones mínimamente frustradas, lo cual asegura un correcto plegado de la molécula y una alta estabilidad. Al mismo tiempo se observa una muy baja proporción de interacciones altamente frustradas, al igual que sucede en las proteínas ANK diseñadas. A lo largo de este trabajo, hemos mostrado que las proteínas naturales exhiben parches de interacciones conflictivas con la estabilidad debido a que los mismos son importantes para la excursión conformacional de la molécula en su ensamble funcional. El próximo paso para el diseño racional de proteínas de este tipo, requiere entender como insertar conflictos energéticos en la estructura, sin comprometer de forma crítica la estabilidad y pudiendo introducir funciones útiles en un entorno celular.

Como contraparte de su simplicidad estructural, la periodicidad y simetría interna inherente de las proteínas repetitivas presentan desafíos metodológicos en cuanto a la aplicabilidad de las herramientas bioinformáticas típicamente usadas para el estudio de proteínas globulares. Cuando se manipulan datos correspondientes a proteínas globulares, las comparaciones se realizan, en general, al nivel de dominios. Bases de datos como Pfam [Bateman et al., 2004], SMART [Schultz et al., 2000], entre otras, han sido fundamentales para el estudio de las relaciones entre miembros de familias de este tipo ya que agrupan los dominios en familias, proveyendo alineamientos múltiples de sus secuencias y diferentes parámetros asociados a los mismos. Debido a la homología existente entre repeticiones dentro de una misma proteína y aquellas presentes en miembros de la misma familia y que los arreglos repetitivos pueden contener un número variable de repeticiones, comparar este tipo de proteínas presenta una complejidad adicional y dificulta la definición de los dominios estructurales [Espada et al., 2015].

La respuesta lógica a este inconveniente es hacer las comparaciones y estudios al nivel de las unidades básicas, las repeticiones. Sin embargo, segmentar las secuencias de estas proteínas en sus unidades repetitivas no es una tarea trivial debido a que usualmente exhiben una alta divergencia al límite en que la homología entre las mismas no puede ser detectada. La presencia de regiones no repetitivas, inserciones y deleciones también aporta a la dificultad para definir las unidades repetidas. Incluso la definición del largo de la unidad repetitiva o el marco a partir del cual se considera que éstas comienzan y terminan, es decir su fase, no son parámetros fácilmente derivables. Si se observan las entradas en Smart o Pfam para la familia ANK, nos encontramos con que existen incongruencias en cuanto al largo y fase de la unidad repetitiva (lo cual sucede para muchas familias de este tipo). Para las ANKs, podemos encontrar una multiplicidad de modelos que las representan en la base de datos de Pfam. Estos modelos, agrupados dentro de lo que en Pfam se denomina un Clan (Pfam Clan ID: CL0465), son 7 (Ank, Ank2, Ank3, Ank4, Ank5, DUF3420 y DUF3447), de los cuales 2 corresponden a los llamados dominios DUF (*Domains of Unknown Function*), para los cuales no se conoce una función asociada. Los 5 modelos restantes poseen largos diferentes, y Ank2, Ank4 y Ank5 contienen la frase "many copies" (varias copias), en su descripción, haciendo referencia a que se trata del motivo básico ANK repetido en tándem al menos dos veces. Por su parte, la base de datos de Smart, posee un modelo correspondiente al motivo ANK (Smart ID: SM00248) cuyo largo es de 30 residuos. Si bien el motivo TPLH en este modelo se encuentra en la misma posición relativa que en el modelo ANK de Pfam, estos dos difieren en su largo.

En cuanto a su detección, los métodos basados en secuencia, en su mayoría basados en descripciones estadísticas de las repeticiones, son subóptimos en la detección ya que no son capaces de detectar con precisión los límites de las mismas sobre todo de aquellas localizadas en los extremos de la cadena polipeptídica, debido a la alta divergencia de las secuencias. Pese a esta gran variabilidad, las estructuras de estas moléculas se encuentran bastante conservadas, siendo relativamente sencillo detectar las repeticiones de manera visual. Esta propiedad ha sido usada para la creación de la base de datos RepeatsDB [Di Domenico et al., 2013] que constituye el primer esfuerzo para clasificar este tipo de proteínas y anotar la localización de las repeticiones mediante el análisis visual por parte de curadores expertos. En los últimos años, diferentes métodos fueron apareciendo para realizar la detección de unidades repetiti-

vas al nivel estructural. Algunos de estos métodos como ANKPred [Chakrabarty and Parekh, 2014], están basados en el análisis de los mapas de contacto usando teoría de grafos, otros como Console [Hrabe and Godzik, 2014], en el análisis espectral de imágenes. Sin embargo la detección de ciertas repeticiones que divergen demasiado en sus mapas de contacto son invisibles a los métodos basados en grafos. Por otro lado, el método Console es bastante sensible para detectar repeticiones pero lo hace de manera que no es posible definir la fase en que la detección es realizada, asignando fases heterogéneas a diferentes miembros de una misma familia. Esto imposibilita la realización de estudios comparativos a lo largo de una familia completa. En este trabajo hemos desarrollado una metodología llamada *Teselado proteico* [Parra et al., 2013] para el análisis de periodicidades y la detección de repeticiones en estructuras proteicas. Este método analiza de forma exhaustiva que tan bueno es cualquier fragmento definible dentro de la estructura global, para cubrir la totalidad de la misma mediante copias de sí mismo. Dicho método nos ha permitido medir el nivel de repetitividad de una estructura proteica y definir unidades estructurales repetidas en las mismas. Mientras algunas proteínas son claramente periódicas, en otras se observa la presencia de claras disrupciones en las relaciones simétricas entre las unidades repetidas. Esta metodología puede aplicarse a cualquier tipo de proteína repetitiva o globular, ya sea en su estado monomérico o formando parte de un complejo cuaternario.

La aplicación del método de teselado y las diferentes funciones desarrolladas para analizar sus resultados generan detecciones automáticas de las repeticiones, seleccionando el período y fase de las mismas que maximizan el cubrimiento geométrico de la estructura global. Aunque en una mayoría de casos la fase de las repeticiones detectadas, en proteínas pertenecientes a la misma familia no es homogénea, la exhaustividad del método asegura que la consistencia en la definición de las fases puede alcanzarse si funciones adecuadas son sumadas al análisis. En nuestro caso hemos utilizado la función de *plegabilidad relativa* para encontrar dentro de las teselas presentes, a largos iguales a los de la frecuencia característica mayoritaria para los miembros de la familia ANK, aquellas que maximizan su capacidad de plegarse de forma autónoma. Al analizar la fase de las teselas recuperadas de esta forma, observamos que hay una fase mayoritariamente conservada en las proteínas naturales. Esta fase fue utilizada para la detección consistente de las repeticiones ANK en un conjunto de proteínas no redundantes.

Mediante la aplicación de un método de alineamiento indirecto de secuencias basado en estructura, se generaron alineamientos múltiples para todas las repeticiones detectadas y se asignaron posiciones análogas para todas ellas respecto de la repetición consenso. Debido a que las repeticiones internas y aquellas presentes, tanto en el extremo N-terminal como en el extremo C-terminal exhiben diferentes patrones energéticos consideramos analizar estos tres tipos de repeticiones por separado en todo el estudio subsiguiente. El análisis de los HMMs de los 3 tipos de repetición, indica la existencia de homología entre todos ellos, excepto entre el N-terminal y el C-terminal, lo cual probablemente se deba a una excesiva divergencia entre los mismos ya que se detecta homología entre ambos y la repetición interna.

A partir de los alineamientos se calcularon dos medidas de conservación usando la medida de *contenido de información (CI)*, la primera relacionada a las secuencias analizando la distribución de las identidades de los residuos en cada columna de los alineamientos y una segunda medida basada en las estructuras correspondientes a dichas secuencias, analizando la distribución del estado de frustración local [Ferreiro et al., 2007b]. Ambas medidas aunque provenientes de diferentes fuentes, se expresan en unidades de bits y por tanto pueden ser comparadas. Se observó que las mismas se encuentran positivamente correlacionadas para el caso de las repeticiones internas y las C-terminales lo cual indica que mientras más parecida es la secuencia de una repetición al consenso del alineamiento de secuencias, más favorablemente ésta se plegará en una estructura compatible con la estructura media de las repeticiones ANK analizadas. No se observó esta correlación para el caso de las repeticiones N-terminales. Se midió además el CI, basado en frustración local, para el mapa de contactos promedio de pares de repeticiones internas y encontramos que los contactos que maximizan dicho valor son energéticamente favorables (mínimamente frustradas) y forman una red de interacciones que conecta aquellos residuos que están más conservados en los alineamientos de secuencias. En contraste, encontramos que las regiones no repetitivas, las inserciones, las regiones circundantes a puntos de deleción, las repeticiones terminales y los sitios de unión a otros ligandos se encuentran enriquecidas en interacciones energéticamente desfavorables (altamente frustradas). Al analizar los cambios energéticos entre los monómeros ANK y los mismos formando un complejo cuaternario, vemos que los sitios de interacción entre las moléculas que en el estado monomérico se encuentran enriquecidos en interacciones frustradas, se desplazan

hacia valores neutros o mínimamente frustrados. No obstante, en otras regiones interacciones mínimamente frustradas se desplazan a valores de mayor frustración. Esto sugiere que existe una redistribución de los valores de frustración individuales de los contactos pero que la morfología global de la distribución de mantiene sin grandes cambios. Es tentativo especular que en el estado acomplejado, nuevas interacciones en conflicto son necesarias para modular transiciones conformacionales que le den plasticidad al complejo o para disociarlo.

Por último analizamos el proceso de plegado de varios miembros de la familia ANK por medio de simulaciones computacionales usando modelos basados en estructura. Simulamos el plegado de proteínas con 3, 4, 5 y 6 repeticiones. El núcleo de plegado en todos los casos analizados se corresponde con una repetición y una parte de alguna de las dos repeticiones adyacentes. Observamos que en muchos casos existen eventos de retroceso y que los mismos coinciden con el comienzo de plegado de regiones ubicadas en el extremo opuesto del arreglo repetitivo, quedando por analizar si ocurren en dinámicas en donde no se aplique un sesgo en la coordenada de reacción. Esta observación, que parece no ser casual, podría ser un indicio de interacciones indirectas a largo alcance entre repeticiones en extremos opuestos del arreglo. De hecho, los análisis de la dinámica correspondiente a I$\kappa$B*alpha* en su versión truncada que contiene las primeras 4 repeticiones y su versión completa, muestran que los mecanismos de plegado de la región común difieren. Dado que la versión truncada fue derivada de la estructura completa, las diferencias dinámicas en dicha región tiene que ser consecuencia de la presencia de las 2 repeticiones extra en el extremo C-terminal. Lo mismo se observa para el caso de 3ANK y 4ANK que difieren en la presencia de una repetición interna extra en la última, y muestran polarizaciones diferentes en sus mecanismos. Que el comportamiento dinámico de estas moléculas sea tan sensible a truncaciones de repeticiones individuales es una muestra más de su versatilidad.

A pesar de sus similitudes estructurales, las proteínas con igual cantidad de repeticiones mostraron, en muchos casos, comportamientos dinámicos muy diferentes. Las proteínas diseñadas fueron las que mostraron las barreras energéticas más altas entre su estado plegado y desplegado. Además en todos los casos, estas moléculas no muestran intermediarios de plegado. Otras proteínas en cambio exhiben especies conformacionales estables más allá de sus estados plegado y desplegado en las que un grupo de repeticiones se encuentran plegadas y otras

desplegadas. Más allá de la presencia o ausencia de intermediarios estables los análisis de interacciones nativas formadas al nivel de residuos nos permitió observar en qué estadios del plegado global se pliegan las diferentes repeticiones. Encontramos que diferentes proteínas muestran diferentes polarizaciones de su plegado. Mientras algunas comienzan a plegarse por sus repeticiones internas y desde ahí propagan el plegado hacia las demás, otras lo hacen de forma polarizada desde su extremo N-terminal o C-terminal. A partir del uso de modelos cinéticos discretos en que se asignaron las diferentes repeticiones como foldones, encontramos que no existen rutas paralelas de plegado obvias, excepto para el caso de las repeticiones terminales. En todos los casos, la estabilidad relativa de tener el arreglo completo en estado plegado fue menor que aquellos en que una o dos de las repeticiones terminales se encontraban en estado desplegado. Esto se refleja en los altos valores de las constantes de transición entre dichos estados, lo cual tiene su origen en la geometría de los arreglos y la falta de una interfaz de interacción en los terminales respecto de las repeticiones internas.

Hemos reportado el primer caso en que este tipo de estudios se realiza en una familia proteica del tipo repetitivo, no sólo presentando resultados novedosos acerca de la conservación de las firmas moleculares tanto al nivel de secuencias, estructuras y comportamientos dinámicos, sino que además hemos sentado las bases metodológicas para replicar los análisis en cualquier otra familia repetitiva. Nuestros aportes metodológicos y acerca de la biofísica de las proteínas ANK pueden ser recreados para otras topologías repetitivas y avanzar en su estudio de forma sistemática.

# Bibliografía


- Abraham, A. L., Rocha, E. P., and Pothier, J. (2008). Swelfe: a Detector of Internal Repeats in Sequences and Structures. *Bioinformatics*, 24(13):1536–1537.

- Abriata, L. A., Salverda, M. L., and Tomatis, P. E. (2012). Sequence–function–stability relationships in proteins from datasets of functionally annotated variants: The case of tem $\beta$-lactamases. *FEBS letters*, 586(19):3330–3335.

- Aksel, T. and Barrick, D. (2009). Analysis of repeat-protein folding using nearest-neighbor statistical mechanical models. *Meth. Enzymol.*, 455:95–125.

- Aksel, T., Majumdar, A., and Barrick, D. (2011). The contribution of entropy, enthalpy, and hydrophobic desolvation to cooperativity in repeat-protein folding. *Structure*, 19(3):349–360.

- Alvarez-Castelao, B. and Castano, J. G. (2005). Mechanism of direct degradation of IkappaBalpha by 20S proteasome. *FEBS Lett.*, 579(21):4797–4802.

- Andrade, M. A., Ponting, C. P., Gibson, T. J., and Bork, P. (2000). Homology-based method for identification of protein repeats using statistical significance estimates. *J. Mol. Biol.*, 298(3):521–537.

- Axton, J. M., Shamanski, F. L., Young, L. M., Henderson, D. S., Boyd, J. B., and Orr-Weaver, T. L. (1994). The inhibitor of DNA replication encoded by the Drosophila gene plutonium is a small, ankyrin repeat protein. *EMBO J.*, 13(2):462–470.

- Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., et al. (2004). The pfam protein families database. *Nucleic acids research*, 32(suppl 1):D138–D141.

- Bergqvist, S., Ghosh, G., and Komives, E. A. (2008). The IkappaBalpha/NF-kappaB complex has two hot spots, one at either end of the interface. *Protein Sci.*, 17(12):2051–2058.

- Betancourt, M. R. and Onuchic, J. N. (1995). Kinetics of proteinlike models: the energy landscape factors that determine folding. *The Journal of chemical physics*, 103(2):773–787.

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

- 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.

- Bray, D. et al. (1995). Protein molecules as computational elements in living cells. *Nature*, 376(6538):307–312.

- Breeden, L. and Nasmyth, K. (1987). Similarity between cell-cycle genes of budding yeast and fission yeast and the notch gene of drosophila.

- Brunette, T., Parmeggiani, F., Huang, P.-S., Bhabha, G., Ekiert, D. C., Tsutakawa, S. E., Hura, G. L., Tainer, J. A., and Baker, D. (2015). Exploring the repeat protein universe through computational protein design. *Nature*.

- Bryngelson, J. D., Onuchic, J. N., Socci, N. D., and Wolynes, P. G. (1995). Funnels, Pathways, and the Energy Landscape of Protein Folding: a Synthesis. *Proteins*, 21(3):167–195.

- Bryngelson, J. D. and Wolynes, P. G. (1987). Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. U.S.A.*, 84(21):7524–7528.

- Bryngelson, J. D. and Wolynes, P. G. (1989). Intermediates and barrier crossing in a random energy model (with applications to protein folding). *The Journal of Physical Chemistry*, 93(19):6902–6915.

- Bustamante, C. D., Townsend, J. P., and Hartl, D. L. (2000). Solvent accessibility and purifying selection within proteins of Escherichia coli and Salmonella enterica. *Mol. Biol. Evol.*, 17(2):301–308.

- Chakrabarty, B. and Parekh, N. (2014). Identifying tandem Ankyrin repeats in protein structures. *BMC Bioinformatics*, 15(1):6599.

- Cho, S. S., Levy, Y., and Wolynes, P. G. (2006). P versus q: Structural reaction coordinates capture protein folding on smooth landscapes. *Proceedings of the National Academy of Sciences of the United States of America*, 103(3):586–591.

- Cho, S. S., Levy, Y., and Wolynes, P. G. (2009). Quantitative criteria for native energetic heterogeneity influences in the prediction of protein folding kinetics. *Proceedings of the National Academy of Sciences*, 106(2):434–439.

- Cho, S. S., Weinkam, P., and Wolynes, P. G. (2008). Origins of barriers and barrierless folding in bbl. *Proceedings of the National Academy of Sciences*, 105(1):118–123.

- Clementi, C., Nymeyer, H., and Onuchic, J. N. (2000). Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? an investigation for small globular proteins. *Journal of molecular biology*, 298(5):937–953.

- Cortajarena, A. L., Mochrie, S. G., and Regan, L. (2008). Mapping the energy landscape of repeat proteins using nmr-detected hydrogen exchange. *Journal of molecular biology*, 379(3):617–626.

- Das, A. and 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., and Papoian, G. A. (2012). AWSEM-MD: Protein Structure Prediction Using Coarse-grained Physical Potentials and Bioinformatically Based Local Structure Biasing. *J Phys Chem B*, 116(29):8494–8503.

- Denton, M. J., Marshall, C. J., and Legge, M. (2002). The Protein Folds as Platonic Forms: New Support for the Pre-Darwinian Conception of Evolution by Natural Law. *J. Theor. Biol.*, 219(3):325–342.

- Desjarlais, J. R. and Berg, J. M. (1993). Use of a zinc-finger consensus sequence framework and specificity rules to design specific dna binding proteins. *Proceedings of the National Academy of Sciences*, 90(6):2256–2260.

- DeVries, I., Ferreiro, D. U., Sanchez, I. E., and Komives, E. A. (2011). Folding Kinetics of the Cooperatively Folded Subdomain of the IkappaBalpha Ankyrin Repeat Domain. *J. Mol. Biol.*, 408(1):163–176.

- Di Domenico, T., Potenza, E., Walsh, I., Parra, R. G., Giollo, M., Minervini, G., Piovesan, D., Ihsan, A., Ferrari, C., Kajava, A. V., et al. (2013). Repeatsdb: a database of tandem repeat protein structures. *Nucleic acids research*, page gkt1175.

- Dill, K. A. (1990). Dominant forces in protein folding. *Biochemistry*, 29(31):7133–7155.

- Dixit, A. and Verkhivker, G. M. (2011). The energy landscape analysis of cancer mutations in protein kinases. *PLoS One*, 6(10):e26071.

- Doyle, L., Hallinan, J., Bolduc, J., Parmeggiani, F., Baker, D., Stoddard, B. L., and Bradley, P. (2015). Rational design of $\alpha$-helical tandem repeat proteins with closed architectures. *Nature*, 528(7583):585–588.

- Eastwood, M. P. and Wolynes, P. G. (2001). Role of explicitly cooperative interactions in protein folding funnels: a simulation study. *The Journal of Chemical Physics*, 114(10):4702–4716.

- Eddy, S. R. (2001). {HMMER: Profile hidden Markov models for biological sequence analysis}.

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

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

- Ferreiro, D. U., Cervantes, C. F., Truhlar, S. M., Cho, S. S., Wolynes, P. G., and Komives, E. A. (2007a). Stabilizing IkappaBalpha by çonsensus"Design. *J. Mol. Biol.*, 365(4):1201–1216.

- Ferreiro, D. U., Cho, S. S., Komives, E. A., and Wolynes, P. G. (2005). The energy landscape of modular repeat proteins: topology determines folding mechanism in the ankyrin family. *J. Mol. Biol.*, 354(3):679–692.

- Ferreiro, D. U., Hegler, J. A., Komives, E. A., and Wolynes, P. G. (2007b). Localizing frustration in native proteins and protein assemblies. *Proc. Natl. Acad. Sci. U.S.A.*, 104(50):19819–19824.

- Ferreiro, D. U., Hegler, J. A., Komives, E. A., and Wolynes, P. G. (2011). On the role of frustration in the energy landscapes of allosteric proteins. *Proc. Natl. Acad. Sci. U.S.A.*, 108(9):3499–3503.

- Ferreiro, D. U. and Komives, E. A. (2007). The plastic landscape of repeat proteins. *Proc. Natl. Acad. Sci. U.S.A.*, 104(19):7735–7736.

- Ferreiro, D. U. and Komives, E. A. (2010). Molecular mechanisms of system control of nf-kappab signaling by ikappabalpha. *Biochemistry*, 49(8):1560–7.

- Ferreiro, D. U., Komives, E. A., and Wolynes, P. G. (2014). Frustration in biomolecules. *Q. Rev. Biophys.*, 47(4):285–363.

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

- Ferreiro, D. U. and Wolynes, P. G. (2008). The capillarity picture and the kinetics of one-dimensional protein folding. *Proc. Natl. Acad. Sci. U.S.A.*, 105(29):9853–9854.

- Frauenfelder, H. (2002). Proteins: Paradigms of complexity. *Proc. Natl. Acad. Sci. U.S.A.*, 99 Suppl 1:2479–80.

- Frauenfelder, H., Sligar, S. G., and Wolynes, P. G. (1991). The Energy Landscapes and Motions of Proteins. *Science*, 254(5038):1598–1603.

- Fuglestad, B., Gasper, P. M., McCammon, J. A., Markwick, P. R., and Komives, E. A. (2013). Correlated motions and residual frustration in thrombin. *The Journal of Physical Chemistry B*, 117(42):12857–12863.

- Fulop, V. and Jones, D. T. (1999). Beta Propellers: Structural Rigidity and Functional Diversity. *Curr. Opin. Struct. Biol.*, 9(6):715–721.

- Galen, S. C., Natarajan, C., Moriyama, H., Weber, R. E., Fago, A., Benham, P. M., Chavez, A. N., Cheviron, Z. A., Storz, J. F., and Witt, C. C. (2015). Contribution of a mutational hot spot to hemoglobin adaptation in high-altitude andean house wrens. *Proceedings of the National Academy of Sciences*, 112(45):13958–13963.

- Gianni, S., Camilloni, C., Giri, R., Toto, A., Bonetti, D., Morrone, A., Sormanni, P., Brunori, M., and 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.

- Goodsell, D. S. and Olson, A. J. (2000). Structural Symmetry and Protein Function. *Annu Rev Biophys Biomol Struct*, 29:105–153.

- Grantcharova, V. P., Riddle, D. S., Santiago, J. V., and Baker, D. (1998). Important role of hydrogen bonds in the structurally polarized transition state for folding of the src SH3 domain. *Nat. Struct. Biol.*, 5(8):714–720.

- Groves, M. R., Hanlon, N., Turowski, P., Hemmings, B. A., and Barford, D. (1999). The structure of the protein phosphatase 2a pr65/a subunit reveals the conformation of its 15 tandemly repeated heat motifs. *Cell*, 96(1):99–110.

- Guo, Y., Yuan, C., Tian, F., Huang, K., Weghorst, C. M., Tsai, M. D., and Li, J. (2010). Contributions of conserved TPLH tetrapeptides to the conformational stability of ankyrin repeat proteins. *J. Mol. Biol.*, 399(1):168–181.

- Haigis, M. C., Haag, E. S., and Raines, R. T. (2002). Evolution of Ribonuclease Inhibitor by Exon Duplication. *Mol. Biol. Evol.*, 19(6):959–963.

- Hashimoto, K. and Panchenko, A. R. (2010). Mechanisms of Protein Oligomerization, the Critical Role of Insertions and Deletions in Maintaining Different Oligomeric States. *Proc. Natl. Acad. Sci. U.S.A.*, 107(47):20352–20357.

- Hegler, J. A., Lätzer, J., Shehu, A., Clementi, C., and Wolynes, P. G. (2009). Restriction versus guidance in protein structure prediction. *Proc Natl Acad Sci U S A*, 106(36):15302–7.

- Hegler, J. A., Weinkam, P., and Wolynes, P. G. (2008). The Spectrum of Biomolecular States and Motions. *HFSP J*, 2(6):307–313.

- Hrabe, T. and Godzik, A. (2014). ConSole: using modularity of contact maps to locate solenoid domains in protein structures. *BMC Bioinformatics*, 15:119.

- Huang, P.-S., Feldmeier, K., Parmeggiani, F., Velasco, D. A. F., Höcker, B., and Baker, D. (2016). De novo design of a four-fold symmetric tim-barrel protein with atomic-level accuracy. *Nature chemical biology*, 12(1):29–34.

- Interlandi, G., Settanni, G., and Caflisch, A. (2006). Unfolding transition state and intermediates of the tumor suppressor p16INK4a investigated by molecular dynamics simulations. *Proteins*, 64(1):178–192.

- Itoh, K. and Sasai, M. (2009). Multidimensional Theory of Protein Folding. *J Chem Phys*, 130(14):145104.

- Jablonka, E. and Raz, G. (2009). Transgenerational epigenetic inheritance: Prevalence, mechanisms, and implications for the study of heredity and evolution. *Q Rev Biol*, 84(2):131–76.

- Jenik, M., Parra, R. G., Radusky, L. G., Turjanski, A., Wolynes, P. G., and Ferreiro, D. U. (2012). Protein frustratometer: a tool to localize energetic frustration in protein molecules. *Nucleic Acids Res.*, 40(Web Server issue):W348–351.

- Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637.

- Kajander, T., Cortajarena, A. L., Main, E. R., Mochrie, S. G., and Regan, L. (2005). A new folding paradigm for repeat proteins. *Journal of the American Chemical Society*, 127(29):10188–10190.

- Kajava, A. V. (2012). Tandem repeats in proteins: from sequence to structure. *J. Struct. Biol.*, 179(3):279–288.

- Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., and Sternberg, M. J. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc*, 10(6):845–858.

- Kendrew, J., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H., and Phillips, D. C. (1958). A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, 181(4610):662–6.

- Kinoshita, K., Kidera, A., and Go, N. (1999). Diversity of functions of proteins with internal symmetry in spatial arrangement of secondary structural elements. *Protein Sci.*, 8(6):1210–1217.

- Kohl, A., Binz, H. K., Forrer, P., Stumpp, M. T., Plückthun, A., and Grütter, M. G. (2003). Designed to be stable: crystal structure of a consensus ankyrin repeat protein. *Proceedings of the National Academy of Sciences*, 100(4):1700–1705.

- Kramer, M. A., Wetzel, S. K., Plückthun, A., Mittl, P. R., and Grütter, M. G. (2010). Structural determinants for improved stability of designed ankyrin repeat proteins with a redesigned c-capping module. *Journal of molecular biology*, 404(3):381–391.

- Kuriyan, J. and Eisenberg, D. (2007). The Origin of Protein Interactions and Allostery in Colocalization. *Nature*, 450(7172):983–990.

- Lamboy, J. A., Kim, H., Dembinski, H., Ha, T., and Komives, E. A. (2013). Single-molecule FRET reveals the native-state dynamics of the IkBa ankyrin repeat domain. *J. Mol. Biol.*, 425(14):2578–2590.

- Lehmann, M., Kostrewa, D., Wyss, M., Brugger, R., D'Arcy, A., Pasamontes, L., and van Loon, A. P. (2000). From dna sequence to improved functionality: using protein sequence comparisons to rapidly design a thermostable consensus phytase. *Protein Engineering*, 13(1):49–57.

- Levinthal, C. (1968). Are there pathways for protein folding. *J. Chim. phys*, 65(1):44–45.

- Levy, Y., Cho, S. S., Shen, T., Onuchic, J. N., and Wolynes, P. G. (2005). Symmetry and Frustration in Protein Energy Landscapes: a near Degeneracy Resolves the Rop Dimer-folding Mystery. *Proc. Natl. Acad. Sci. U.S.A.*, 102(7):2373–2378.

- Levy, Y., Wolynes, P. G., and Onuchic, J. N. (2004). Protein topology determines binding mechanism. *Proceedings of the National Academy of Sciences of the United States of America*, 101(2):511–516.

- Liu, S.-Q., Tan, D.-Y., Zhang, K.-Q., Ji, X.-L., Tao, Y., and Fu, Y.-X. (2012). *Protein folding, binding and energy landscape: A synthesis*. INTECH Open Access Publisher.

- Luo, H. and Nijveen, H. (2013). Understanding and identifying amino acid repeats. *Brief Bioinform.*

- Lux, S. E., John, K. M., and Bennett, V. (1990). Analysis of cdna for human erythrocyte ankyrin indicates a repeated structure with homology to tissue-differentiation and cell-cycle control proteins.

- Macias, M. J., Gervais, V., Civera, C., and Oschkinat, H. (2000). Structural analysis of ww domains and design of a ww prototype. *Nature Structural & Molecular Biology*, 7(5):375–379.

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

- Main, E. R., Stott, K., Jackson, S. E., and Regan, L. (2005). Local and long-range stability in tandemly arrayed tetratricopeptide repeats. *Proceedings of the National Academy of Sciences of the United States of America*, 102(16):5721–5726.

- Marcotte, E. M., Pellegrini, M., Yeates, T. O., and Eisenberg, D. (1999). A census of protein repeats. *J. Mol. Biol.*, 293(1):151–160.

- Martinez, J. C., Pisabarro, M. T., and Serrano, L. (1998). Obligatory steps in protein folding and the conformational diversity of the transition state. *Nat. Struct. Biol.*, 5(8):721–729.

- Michaely, P. and Bennett, V. (1992). The ANK repeat: a ubiquitous motif involved in macromolecular recognition. *Trends Cell Biol.*, 2(5):127–129.

- Michaely, P., Tomchick, D. R., Machius, M., and Anderson, R. G. (2002). Crystal structure of a 12 ANK repeat stack from human ankyrinR. *EMBO J.*, 21(23):6387–6396.

- Monod, J., Wyman, J., and Changeux, J. P. (1965). On the nature of allosteric transitions: a plausible model. *J. Mol. Biol.*, 12:88–118.

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

- Mosavi, L. K., Minor, D. L., and Peng, Z. Y. (2002). Consensus-derived Structural Determinants of the Ankyrin Repeat Motif. *Proc. Natl. Acad. Sci. U.S.A.*, 99(25):16029–16034.

- Moult, J., Fidelis, K., Kryshtafovych, A., and Tramontano, A. (2011). Critical assessment of methods of protein structure prediction (casp)–round ix. *Proteins*, 79 Suppl 10:1–5.

- Muraki, M., Ishimura, M., and Harata, K. (2002). Interactions of Wheat-germ Agglutinin with GlcNAc Beta 1,6Gal Sequence. *Biochim. Biophys. Acta*, 1569(1-3):10–20.

- Murray, K. B., Taylor, W. R., and Thornton, J. M. (2004). Toward the Detection and Validation of Repeats in Protein Structure. *Proteins*, 57(2):365–380.

- Nagano, N., Orengo, C. A., and Thornton, J. M. (2002). One fold with many functions: the evolutionary relationships between tim barrel families based on their sequences, structures and functions. *J Mol Biol*, 321(5):741–65.

- Neer, E. J., Schmidt, C. J., Nambudripad, R., and Smith, T. F. (1994). The ancient regulatory-protein family of wd-repeat proteins. *Nature*, 371(6495):297–300.

- Nelson, E. D. and Onuchic, J. N. (1998). Proposed mechanism for stability of proteins to evolutionary mutations. *Proc. Natl. Acad. Sci. U.S.A.*, 95(18):10682–10686.

- Oliveberg, M. and Wolynes, P. G. (2005). The Experimental Survey of Protein-folding Energy Landscapes. *Q. Rev. Biophys.*, 38(3):245–288.

- Onuchic, J. N., Wolynes, P. G., Luthey-Schulten, Z., and Socci, N. D. (1995). Toward an outline of the topography of a realistic protein-folding funnel. *Proceedings of the National Academy of Sciences*, 92(8):3626–3630.

- Orengo, C. A., Michie, A., Jones, S., Jones, D. T., Swindells, M., and Thornton, J. M. (1997). Cath–a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1109.

- Ormö, M., Cubitt, A. B., Kallio, K., Gross, L. A., Tsien, R. Y., and Remington, S. J. (1996). Crystal structure of the aequorea victoria green fluorescent protein. *Science*, 273(5280):1392–5.

- Panchenko, A. R., Luthey-Schulten, Z., and Wolynes, P. G. (1996). Foldons, protein structural modules, and exons. *Proc. Natl. Acad. Sci. U.S.A.*, 93(5):2008–2013.

- Papoian, G. A., Ulander, J., Eastwood, M. P., Luthey-Schulten, Z., and Wolynes, P. G. (2004). Water in protein structure prediction. *Proc. Natl. Acad. Sci. U.S.A.*, 101(10):3352–3357.

- Parra, R. G., Espada, R., Sanchez, I. E., Sippl, M. J., and Ferreiro, D. U. (2013). Detecting repetitions and periodicities in proteins by tiling the structural space. *J Phys Chem B*, 117(42):12887–12897.

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

- Pauling, L. and Corey, R. B. (1951). The pleated sheet, a new layer configuration of polypeptide chains. *Proc Natl Acad Sci U S A*, 37(5):251–6.

- Pauling, L., Corey, R. B., and Branson, H. R. (1951). The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci U S A*, 37(4):205–11.

- Pauling, L. and Itano, H. A. (1949). Sickle cell anemia a molecular disease. *Science*, 110(2865):543–8.

- Plotkin, S. S. and Onuchic, J. N. (2000). Investigation of routes and funnels in protein folding by free energy functional methods. *Proc. Natl. Acad. Sci. U.S.A.*, 97(12):6509–6514.

- Pluckthun, A. (2015). Designed Ankyrin Repeat Proteins (DARPins): Binding Proteins for Research, Diagnostics, and Therapy. *Annu. Rev. Pharmacol. Toxicol.*, 55:489–511.

- Ramisch, S., Weininger, U., Martinsson, J., Akke, M., and Andre, I. (2014). Computational design of a leucine-rich repeat protein with a predefined geometry. *Proc. Natl. Acad. Sci. U.S.A.*

- Sánchez, I. E., Ferreiro, D. U., Dellarole, M., and de Prat-Gay, G. (2010). Experimental snapshots of a protein-dna binding landscape. *Proc Natl Acad Sci U S A*, 107(17):7751–6.

- Sanger, F., Nicklen, S., and Coulson, A. R. (1977). Dna sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467.

- Sanger, F. and Tuppy, H. (1951). The amino-acid sequence in the phenylalanyl chain of insulin. 2. the investigation of peptides from enzymic hydrolysates. *Biochemical Journal*, 49(4):481.

- Santos, J., Gebhard, L. G., Risso, V. A., Ferreyra, R. G., Rossi, J. P. F. C., and Ermácora, M. R. (2004). Folding of an abridged beta-lactamase. *Biochemistry*, 43(6):1715–23.

- Schafer, N. P., Hoffman, R. M., Burger, A., Craig, P. O., Komives, E. A., and Wolynes, P. G. (2012). Discrete Kinetic Models from Funneled Energy Landscape Simulations. *PLoS ONE*, 7(12):e50635.

- Schafer, N. P., Kim, B. L., Zheng, W., and Wolynes, P. G. (2014). Learning To Fold Proteins Using Energy Landscape Theory. *Isr. J. Chem.*, 54(8-9):1311–1337.

- Schaper, E., Kajava, A. V., Hauser, A., and Anisimova, M. (2012). Repeat or not Repeat?–Statistical Validation of Tandem Repeat Prediction in Genomic Sequences. *Nucleic Acids Res.*, 40(20):10005–10017.

- Schilling, J., Schoppe, J., and Pluckthun, A. (2014). From DARPins to LoopDARPins: novel LoopDARPin design allows the selection of low picomolar binders in a single round of ribosome display. *J. Mol. Biol.*, 426(3):691–721.

- Schneider, T. D., Stormo, G. D., Gold, L., and Ehrenfeucht, A. (1986). Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, 188(3):415–431.

- Schrödinger, E. (1944). *What Is Life?* Cambridge University Press.

- Schultz, J., Copley, R. R., Doerks, T., Ponting, C. P., and Bork, P. (2000). Smart: a web-based tool for the study of genetically mobile domains. *Nucleic acids research*, 28(1):231–234.

- Sedgwick, S. G. and Smerdon, S. J. (1999). The ankyrin repeat: a diversity of interactions on a common structural framework. *Trends Biochem. Sci.*, 24(8):311–316.

- Settanni, G., Serquera, D., Marszalek, P. E., Paci, E., and Itzhaki, L. S. (2013). Effects of ligand binding on the mechanical properties of ankyrin repeat protein gankyrin. *PLoS Comput. Biol.*, 9(1):e1002864.

- Shannon, C. E. (1997). The mathematical theory of communication. 1963. *MD Comput*, 14(4):306–317.

- Shea, J. E., Onuchic, J. N., and Brooks, C. L. (1999). Exploring the origins of topological frustration: design of a minimally frustrated model of fragment B of protein A. *Proc. Natl. Acad. Sci. U.S.A.*, 96(22):12512–12517.

- Shih, E. S. and Hwang, M. J. (2004). Alternative Alignments from Comparison of Protein Structures. *Proteins*, 56(3):519–527.

- Shirts, M. R. and Chodera, J. D. (2008). Statistically optimal analysis of samples from multiple equilibrium states. *The Journal of chemical physics*, 129(12):124105.

- Shoemaker, B. A. and Wolynes, P. G. (1999). Exploring structures in protein folding funnels with free energy functionals: the denatured ensemble. *Journal of molecular biology*, 287(3):657–674.

- Simons, K. T., Kooperberg, C., Huang, E., and Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J Mol Biol*, 268(1):209–25.

- Sippl, M. J. (2008). On distance and similarity in fold space. *Bioinformatics*, 24(6):872–3.

- Sippl, M. J. and Wiederstein, M. (2008). A Note on Difficult Structure Alignment Problems. *Bioinformatics*, 24(3):426–427.

- Sippl, M. J. and Wiederstein, M. (2012). Detection of Spatial Correlations in Protein Structures and Molecular Complexes. *Structure*, 20(4):718–728.

- Sivanandan, S. and Naganathan, A. N. (2013). A disorder-induced domino-like destabilization mechanism governs the folding and functional dynamics of the repeat protein IkBa. *PLoS Comput. Biol.*, 9(12):e1003403.

- Soding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics*, 21(7):951–960.

- Soding, J., Remmert, M., and Biegert, A. (2006). HHrep: De Novo Protein Repeat Detection and the Origin of TIM Barrels. *Nucleic Acids Res.*, 34(Web Server issue):W137–142.

- Steipe, B., Schiller, B., Pluckthun, A., and Steinbacher, S. (1994). Sequence statistics reliably predict stabilizing mutations in a protein domain. *J. Mol. Biol.*, 240(3):188–192.

- Sullivan, B. J., Nguyen, T., Durani, V., Mathur, D., Rojas, S., Thomas, M., Syu, T., and Magliery, T. J. (2012). Stabilizing proteins from sequence statistics: the interplay of conservation and correlation in triosephosphate isomerase stability. *J. Mol. Biol.*, 420(4-5):384–399.

- Sutto, L., Lätzer, J., Hegler, J. A., Ferreiro, D. U., and Wolynes, P. G. (2007). Consequences of localized frustration for the folding mechanism of the im7 protein. *Proceedings of the National Academy of Sciences*, 104(50):19825–19830.

- Swapna, L. S., Srikeerthana, K., and Srinivasan, N. (2012). Extent of Structural Asymmetry in Homodimeric Proteins: Prevalence and Relevance. *PLoS ONE*, 7(5):e36688.

- Taketomi, H., Ueda, Y., and Gō, N. (1975). Studies on protein folding, unfolding and fluctuations by computer simulation. *International journal of peptide and protein research*, 7(6):445–459.

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

- Taylor, W. R., Heringa, J., Baud, F., and Flores, T. P. (2002). A fourier analysis of symmetry in protein structure. *Protein Eng*, 15(2):79–89.

- Treusch, S., Cyr, D. M., and Lindquist, S. (2009). Amyloid deposits: Protection against toxic protein species? *Cell Cycle*, 8(11):1668–74.

- Tripathi, S., Waxham, M. N., Cheung, M. S., and Liu, Y. (2015). Lessons in protein design from combined evolution and conformational dynamics. *Scientific reports*, 5.

- Tripp, K. W. and Barrick, D. (2003). Folding by consensus. *Structure*, 11(5):486–487.

- Tripp, K. W. and Barrick, D. (2007). Enhancing the stability and folding rate of a repeat protein through the addition of consensus repeats. *J. Mol. Biol.*, 365(4):1187–1200.

- Truhlar, S. M., Torpey, J. W., and Komives, E. A. (2006). Regions of IkappaBalpha that are critical for its inhibition of NF-kappaB.DNA interaction fold upon binding to NF-kappaB. *Proc. Natl. Acad. Sci. U.S.A.*, 103(50):18951–18956.

- Truong, H. H., Kim, B. L., Schafer, N. P., and Wolynes, P. G. (2013). Funneling and frustration in the energy landscapes of some designed and simplified proteins. *The Journal of chemical physics*, 139(12):121908.

- Tsai, C. J., Maizel, J. V., and Nussinov, R. (2000). Anatomy of protein structures: visualizing how a one-dimensional protein chain folds into a three-dimensional shape. *Proc. Natl. Acad. Sci. U.S.A.*, 97(22):12038–12043.

- Voth, D. (2011). Thanks for the repeat: Intracellular pathogens exploit a common eukaryotic domain. *Cellular logistics*, 1(4):128–132.

- Wales, D. J. (1998). Symmetry, near-symmetry and energetics. *Chem. Phys. Lett.*, 285(5-6):330–336.

- Wales, D. J. (2012). Decoding the energy landscape: Extracting structure, dynamics and thermodynamics. *Philos Transact A Math Phys Eng Sci*, 370(1969):2877–99.

- Walsh, I., Sirocco, F. G., Minervini, G., Di Domenico, T., Ferrari, C., and Tosatto, S. C. (2012). RAPHAEL: recognition, periodicity and insertion assignment of solenoid protein structures. *Bioinformatics*, 28(24):3257–3264.

- Wang, J., Saven, J. G., and Wolynes, P. G. (1996). Kinetics in a globally connected, correlated random energy model. *The Journal of chemical physics*, 105(24):11276–11284.

- Wang, S. and Wolynes, P. G. (2011). On the Spontaneous Collective Motion of Active Matter. *Proc. Natl. Acad. Sci. U.S.A.*, 108(37):15184–15189.

- Weiss, O., Jimenez-Montano, M. A., and Herzel, H. (2000). Information Content of Protein Sequences. *J. Theor. Biol.*, 206(3):379–386.

- Werbeck, N. D., Rowling, P. J., Chellamuthu, V. R., and Itzhaki, L. S. (2008). Shifting transition states in the unfolding of a large ankyrin repeat protein. *Proc. Natl. Acad. Sci. U.S.A.*, 105(29):9982–9987.

- Wolynes, P. G. (1988). Aperioidic crystals: Biology, chemistry and physics in a fugue with stretto. *AIP Conf. Proc.*, 180(39):39–65.

- Wolynes, P. G. (1996). Symmetry and the Energy Landscapes of Biomolecules. *Proc. Natl. Acad. Sci. U.S.A.*, 93(25):14249–14255.

- Wolynes, P. G. (1997). Folding funnels and energy landscapes of larger proteins within the capillarity approximation. *Proc Natl Acad Sci U S A*, 94(12):6170–5.

- Yang, R., Gombart, A. F., Serrano, M., and Koeffler, H. P. (1995). Mutational effects on the p16ink4a tumor suppressor protein. *Cancer research*, 55(12):2503–2506.

- Yang, S., Cho, S. S., Levy, Y., Cheung, M. S., Levine, H., Wolynes, P. G., and Onuchic, J. N. (2004). Domain swapping is a consequence of minimal frustration. *Proceedings of the National Academy of Sciences of the United States of America*, 101(38):13786–13791.

- Zhuravlev, P. I. and Papoian, G. A. (2010). Protein Functional Landscapes, Dynamics, Allostery: a Tortuous Path towards a Universal Theoretical Framework. *Q. Rev. Biophys.*, 43(3):295–332.