

Tesis Doctoral

Desarrollo de métodos estadísticos para el análisis de patrones biológicos en proteínas: aplicaciones a familias de proteínas repetitivas

Espada, Rocío

2017-04-07

Este documento forma parte de la colección de tesis doctorales y de maestría de la Biblioteca Central Dr. Luis Federico Leloir, disponible en digital.bl.fcen.uba.ar. Su utilización debe ser acompañada por la cita bibliográfica con reconocimiento de la fuente.

This document is part of the doctoral theses collection of the Central Library Dr. Luis Federico Leloir, available in digital.bl.fcen.uba.ar. It should be used accompanied by the corresponding citation acknowledging the source.

Cita tipo APA:

Espada, Rocío. (2017-04-07). Desarrollo de métodos estadísticos para el análisis de patrones biológicos en proteínas: aplicaciones a familias de proteínas repetitivas. Facultad de Ciencias Exactas y Naturales. Universidad de Buenos Aires.

Cita tipo Chicago:

Espada, Rocío. "Desarrollo de métodos estadísticos para el análisis de patrones biológicos en proteínas: aplicaciones a familias de proteínas repetitivas". Facultad de Ciencias Exactas y Naturales. Universidad de Buenos Aires. 2017-04-07.

EXACTAS UBA

Facultad de Ciencias Exactas y Naturales



UBA

Universidad de Buenos Aires



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

Desarrollo de métodos estadísticos para el análisis de patrones biológicos en proteínas: aplicaciones a familias de proteínas repetitivas

Tesis presentada para optar por el título de Doctor de la Universidad de Buenos Aires en el
área Química Biológica

Lic. Rocío Espada

Director de tesis: Dr. Diego U. Ferreiro
Consejera de estudios: Dra. Valeria Levi

Lugar de trabajo: Departamento de Química Biológica, FCEN - UBA / Instituto de Química Biológica - Ciencias Exactas y Naturales - UBA/CONICET

Ciudad Autónoma de Buenos Aires, 2017

Fecha de defensa 7 de Abril de 2017

Desarrollo de métodos estadísticos para el análisis de patrones biológicos en proteínas: aplicaciones a familias de proteínas repetitivas

Las proteínas son moléculas sumamente diversas, responsables de las más variadas funciones biológicas. Toda esta diversidad se encuentra codificada en las secuencias de aminoácidos que las constituyen. En este trabajo se aplican modelos cuantitativos, derivados de mecánica estadística, que permiten derivar propiedades generales (o macroscópicas) del plegado proteico a partir de su secuencia primaria (propiedades microscópicas). El trabajo se enfoca en el estudio de proteínas repetitivas, cuya simetría en secuencia y estructura, y propiedades derivadas de ellas, las han puesto en el centro del estudio del plegado proteico y del diseño de biomoléculas con aplicaciones biotecnológicas.

En este trabajo, la simetría en secuencia de las proteínas repetitivas han permitido reducir la dimensionalidad del problema, permitiendo un ajuste fino de los parámetros del modelo que no es posible en proteínas globulares de gran tamaño. Cada proteína (o estado) recibe una puntuación o energía que se encuentra relacionada con la estabilidad del plegado, e incluso permite estimar el efecto de mutaciones puntuales. Por otro lado, el modelo presentado es utilizado para generar ensambles de nuevas variantes que son indistinguibles respecto de las proteínas naturales mediante algoritmos de caracterización de secuencia actuales.

Los modelos desarrollados permiten derivar medidas de correlación entre sustituciones de aminoácidos en proteínas. Como es esperable, estas correlaciones reflejan distintas propiedades funcionales y evolutivas. Los análisis aquí desarrollados permiten distinguir las correlaciones originadas en la alta simetría de las secuencias de proteínas repetitivas y aquellas que evidencian contactos en la estructura nativa.

Palabras claves: proteínas repetitivas, plegado de proteínas, modelo de Potts, predicción de estructura, predicción de ΔG .

Statistical methods for the analyses of biological patterns in proteins: applications to repeat protein families

Proteins are extremely diverse molecules, responsible for the most varied biological functions. All this diversity is encoded in the amino acid sequences that constitute them. In this work, quantitative models, derived from statistical mechanics, that allow to derive general (or macroscopic) properties of the protein folding from its primary sequence (microscopic properties) are applied. This work focuses on the study of repetitive proteins, whose symmetry in sequence and structure, and properties derived from them, have placed them at the center of the study of protein folding and of the design of biomolecules with biotechnological applications.

In this work, the sequence symmetry of the repetitive proteins has allowed to reduce the dimensionality of the problem, permitting to adjust the parameters of the model, which is not possible in large globular proteins. Each protein (or state) receives a score or energy that is related to the stability of the folding, and even allows to estimate the effect of point mutations. On the other hand, the presented model is used to generate assemblies of new variants that are indistinguishable from natural proteins by current sequence characterization algorithms.

The developed models permit to derive measures of correlation between substitutions of amino acids in proteins. As expected, these correlations reflect different functional and evolutionary properties. The analyzes developed here enables to distinguish the correlations originated in the high symmetry of the repetitive protein sequences and those that show contacts in the native structure.

Keywords: repeat protein, protein folding, Potts model, structure prediction, ΔG prediction

Agradecimientos

Diego, gracias por embarcarte en este proyecto delirante, de impulsar la interdisciplina poniéndole el cuerpo, y en ello dedicarte incansablemente a que yo adquiriera una visión amplia, científica, cuantitativa y crítica de los problemas biológicos que encaramos. Y sobre todo, por el cariño.

Aleksandra y Thierry gracias por involucrarse de lleno en este proyecto y los aportes invaluable a mi formación científica. En particular a Aleksandra gracias por el apoyo constante, por ocuparte de mí y abrirme nuevas oportunidades.

A todo el LFP. Por el tiempo que se tomaron para discutir mi trabajo y valorarme. A Nacho, por no dejar que me conforme con mis respuestas. A Gonzalo, por el mutuo empuje, Brenda por la alegría renovada, y Juliana por los mates compartidos. A los excelentes científicos que me abrieron las puertas y me formaron en distintas perspectivas: Ariel Chernomoretz, Ernesto Roman, Veronica Becher, Pablo Trujanski. A todo el QB69 y amigos por compartir frustraciones y alegrías.

A las científicas que me ayudaron a lo largo de mi carrera y me inspiran con su labor a favor de la igualdad de género en la ciencia. Silvina Ponce Dawson, Luciana Bruno, Valeria Levi, y Vero de nuevo.

A mi familia por su apoyo incondicional. A mis viejos por siempre hacerme sentir querida. Por no dejar que me limite. Mi vieja, por enseñarme qué es ser una mujer fuerte y valiente. Mi viejo, por hablar siempre apasionadamente del conocimiento. A mis hermanos, mis compañeros en esta vida: Ramiro, por todas las nerdeadas, Martín, por compartir su visión artística de la vida, Uriel, por su sensibilidad y su fortaleza, y Joan, por enseñarme a querer ingenua y desinteresadamente. Me emociona cada nueva etapa que atravesamos juntos.

Emiliano, gracias por alentarme a superarme, por ser tan compañero, por las risas. Gracias por esta felicidad que somos.

A mis amigos, gracias por hacerme sentir fuerte. Los quiero infinito y los admiro profundamente a cada uno de ustedes. A las chicas de danza, por desconectarme y por la eterna discusión frío vs calor.

Este trabajo no sería posible sin el apoyo a la ciencia y la educación pública, gratuita

y de calidad que sostiene la sociedad argentina y que llevo como bandera. Es mi humilde deseo que este trabajo, y mis producciones científicas subsiguientes, aporten a reforzar estos valores.

Trabajos en el marco de esta tesis

Inferring repeat-protein energetics from evolutionary information

Espada, R., Parra, R. G., Mora, T., Walczak, A. M., y Ferreiro, D. U.

En revisión

Detailing protein landscapes under pressure

Espada, R., Sánchez, I. E., y Ferreiro, D. U.

Biophysical Journal, 111(11), 2339-2341.

2016

Protein repeats from first principles

Turjanski, P., Parra, R. G., Espada, R., Becher, V., y Ferreiro, D. U.

Scientific reports, 6.

2016

Structural and energetic characterization of the ankyrin repeat protein family

Parra, R. G., Espada, R., Verstraete, N., y Ferreiro, D. U.

PLoS Comput Biol, 11(12), e1004659.

2015

Repeat proteins challenge the concept of structural domains

Espada, R., Parra, R. G., Sippl, M. J., Mora, T., Walczak, A. M., y Ferreiro, D. U.

Biochemical Society Transactions, 43(5), 844-849.

2015

Capturing coevolutionary signals in repeat proteins

Espada, R., Parra, R. G., Mora, T., Walczak, A. M., y Ferreiro, D. U.

BMC bioinformatics, 16(1), 207.

2015

Detecting repetitions and periodicities in proteins by tiling the structural space

Parra, R. G., Espada, R., Sánchez, I. E., Sippl, M. J., y Ferreiro, D. U.

The Journal of Physical Chemistry B, 117(42), 12887-12897.

2013

Índice general

1. Introducción	1
1.1. Proceso de plegado de las proteínas	2
1.2. Proteínas repetitivas	8
1.3. Coevolución en secuencias de proteínas	14
1.4. Modelo estadístico	16
1.4.1. Información mutua	17
1.4.2. Modelo por máxima entropía	18
1.4.3. Gauge	20
1.4.4. Modelo alternativo para proteínas repetitivas	21
1.4.5. Aproximación de campo medio	22
1.4.6. Aproximación por Máxima Pseudo-Verosimilitud	24
1.5. Objetivos y orden de la tesis	25
2. Métodos	27
2.1. Representación de datos	28
2.1.1. Secuencia de una proteína	28
2.1.2. Alineamiento múltiple de secuencias	29
2.1.3. Alineamiento de secuencias por parejas	29
2.1.4. Secuencia consenso	30
2.1.5. Similitud e identidad de secuencia	31
2.1.6. Agrupamiento de secuencias	32
2.1.7. Mapa de contactos	32
2.2. Datos	33

2.2.1.	Alineamientos múltiples de secuencias	33
2.2.2.	Familias repetitivas	34
2.2.3.	Preprocesamiento de alineamientos múltiples de secuencia	34
2.2.4.	Estructuras tridimensionales	36
2.2.5.	Estabilidad de plegado	37
2.3.	Caracterización de AMS	37
2.3.1.	Corrección por sobrerepresentación filogenética	37
2.3.2.	Caracterización evolutiva de un AMS	39
2.4.	Modelos cuantitativos	41
2.4.1.	Correlación con Información mutua	41
2.4.2.	Modelo de Potts: Ajuste de parámetros	42
2.4.3.	Modelo de Potts: Aproximación de campo medio e Información Directa	45
2.4.4.	Modelo de Potts: aproximación por máxima pseudo-verosimilitud	47
2.4.5.	Calidad de predictor: curva ROC	48
3.	Modelo de Potts para proteínas repetitivas	49
3.1.	Convergencia del ajuste	51
3.2.	Energías de las secuencias naturales	55
3.3.	Simulación de nuevas secuencias	61
3.3.1.	Energías	61
3.3.2.	Secuencias	62
3.4.	Relación con estabilidad en el plegado	66
3.5.	Análisis de parámetros del campo	69
3.5.1.	Incidencia de los parámetros λ	73
3.6.	Evaluación de energía en estructuras particulares	75
4.	Información directa en proteínas repetitivas	77
4.1.	Información directa en proteínas repetitivas	78
4.1.1.	Posiciones coevolutivas	81
4.1.2.	Resultados para todas las familias de proteínas repetitivas	82
4.1.3.	Aproximación por máxima pseudo-verosimilitud	83

4.2. Predicción estructural	86
4.2.1. Eficacia del método para predecir contactos	92
4.3. Acoplamientos lejanos a lo largo de un conjunto de repeticiones	93
4.4. Cantidad de secuencias necesarias	95
4.5. Simulaciones moleculares de plegado	99
5. Conclusiones	101
Bibliografía	107
Apéndices	
A. Derivación del modelo estadístico	119
B. Código ajuste de parámetros de Modelo de Potts	137

Capítulo 1

Introducción

Las proteínas son las moléculas más versátiles dentro de los sistemas vivos. Participan en todos los procesos biológicos: funcionan como catalizadores, transporte de otras moléculas, aportan soporte mecánico, constituyen parte del sistema inmune, transmiten impulsos nerviosos, generan movimiento, están involucrados en el control de crecimiento y en los procesos de diferenciación. Algunos autores las han propuesto como computadoras moleculares, dada su capacidad de procesar información respecto de su entorno y generar una respuesta acorde [Bray, 1995].

Detrás de esta diversidad, sorprendentemente se encuentran principios comunes que definen a todas estas moléculas. Las proteínas son polímeros lineales contruidos a partir de 20 aminoácidos, cuyas propiedades fisicoquímicas definen a la proteína, sus capacidades funcionales, su estructura tridimensional, su localización. Cómo se encuentra codificada esta información es aún un problema abierto.

El orden de estos aminoácidos en cada proteína particular se denomina secuencia o estructura primaria. Estas moléculas pueden tener plegamientos locales (estructura secundaria) y uno global (estructura terciaria). En muchas proteínas, el plegado da lugar a un ensamble de estructuras necesarios para que la proteína lleve a cabo actividades biológicas específicas.

El ADN es considerado el material genético heredable que incluye, entre otras cosas, la información de las secuencias de aminoácidos de las proteínas. Según el código genético, tripletes específicos de bases nitrogenadas en el ADN codifican para un aminoácido en una proteína. Sin embargo, este material genético no es invariante en el tiempo, sino que es

modificado mediante diferentes procesos [Shapiro, 2013]. El mecanismo más simple son las mutaciones puntuales producidas por los errores de replicación del ADN en el proceso de división celular, que pueden implicar tasas de error de aproximadamente 10^{-5} , o disminuir a 10^{-9} en las células que tienen mecanismos de corrección. También existen mecanismos de reestructuración del ADN, donde segmentos de varias bases son cortados e incorporados en otras regiones, conocidos como transposones y dan origen, por ejemplo, a eventos de duplicación. Durante la reproducción sexual también existen recombinación entre segmentos de cromosomas entrecruzados. Un evento menos frecuente pero que ha tenido un impacto drástico es la simbiogénesis a los cuales se atribuye la incorporación de mitocondrias y plasmas. Finalmente, en bacterias se conocen mecanismos de transferencia horizontal mediante el cual dos células, en general bacterianas, intercambian material genético. Todos estos mecanismos generan cambios en el ADN que luego pueden tener implicancias a nivel fenotípico general, y en particular en las estructuras primarias de las proteínas expresadas por los organismos. Mirar los ensambles de proteínas actuales es observar variantes que han permitido la adaptación exitosa de sus portadores a las condiciones ambientales.

Este trabajo busca desarrollar métodos cuantitativos que permitan extraer información biológica a partir de estos catálogos. En este capítulo discutiré el marco y la relevancia del problema (sección 1.1), introduciré el sistema particular en que trabajé, las proteínas repetitivas (sección 1.2). Luego motivaré desde un marco teórico biológico los métodos cuantitativos a analizar (sección 1.3), y finalmente detallaré el modelo utilizado (modelo de Potts) y diferentes aproximaciones que he analizado (sección 1.4).

1.1. Proceso de plegado de las proteínas

Muchas proteínas tienen la capacidad de adoptar espontáneamente estructuras tridimensionales bien definidas y estables en tiempos cortos (desde microsegundos a unas pocas horas, según la longitud de la secuencia [Naganathan y Muñoz, 2005]), es decir, tienen la capacidad de *plegarse*. Los mecanismos moleculares que dan lugar a este proceso se encuentran, hace más de 50 años [Dill y MacCallum, 2012], en estudio bajo el nombre *el problema del plegado de las proteínas*. La funcionalidad de una proteína se encuentra estrechamente ligada a su

capacidad de plegado.

Entre los primeros trabajos sobre el plegado proteico, se destacan los estudios sobre la proteína ribonucleasa pancreática bovina realizados por Anfinsen [Anfinsen et al., 1961, Anfinsen, 1973]. Sus resultados permiten concluir que toda la información necesaria para que la proteína se pliegue se encuentra codificada en la secuencia de aminoácidos que la constituye. En la actualidad se conocen proteínas cuyos plegados se encuentran catalizados por chaperonas [Pauwels et al., 2007], e incluso proteínas que no se pliegan en estructuras bien definidas (llamadas *proteínas desordenadas*) [Uversky, 2011]. Sin embargo, los resultados de Anfinsen son extensibles a un amplio espectro de proteínas y decodificar la información de plegado a partir de la estructura primaria sigue siendo un desafío abierto.

La capacidad de una proteína de plegarse y llevar a cabo una tarea biológicamente relevante se encuentra relacionada a su éxito evolutivo. No cualquier secuencia de aminoácidos constituye una proteína ni se pliega. Un polímero del largo de una proteína, incluso una pequeña que contiene unos cientos de aminoácidos, puede tomar una infinidad de configuraciones estructurales distintas. Si una proteína transita al azar por todas ellas en busca de alcanzar la estructura plegada tardaría mucho más que los tiempos biológicamente relevantes. Esta observación se conoce como la paradoja de Levinthal [Levinthal, 1968]. Por lo tanto, debe existir un sesgo energético que guíe a la proteína hacia su estructura plegada.

En la actualidad, el marco teórico que describe el comportamiento de plegado proteico es la *teoría de paisajes energéticos* [Onuchic y Wolynes, 2004], la cual postula que el plegado de un polipéptido se encuentra sujeto a un potencial energético. En el caso de una proteína, el potencial toma la forma de un embudo (figura 1.1a) donde el estado plegado (también llamado nativo) es el mínimo, es decir que es ampliamente favorecido y se distingue de los estados desplegados [Sali et al., 1994]. La energía decrece a medida que se forman estructuras más similares al estado nativo, es decir a medida que se forman más contactos de la estructura nativa. El embudo presenta leves rugosidades que se relacionan con los caminos de plegado y estados de plegado intermedios que puede adquirir la proteína. En contraste, un polímero constituido al azar está guiado por un paisaje energético rugoso, donde muchos estados minimizan el potencial y las barreras energéticas que los separan son muy bajas, de manera que el polímero fluctúa por las distintas estructuras posibles sin hallar una configuración

estable.

La forma del potencial energético está determinado por las interacciones entre los aminoácidos que constituyen la secuencia, y con el ambiente que los rodea. Dada la alta dimensionalidad, muchas veces estas interacciones están en conflicto entre sí (*frustradas*). Según el *principio de mínima frustración*, las proteínas que han sido seleccionadas a través de los procesos evolutivos, minimizan los conflictos energéticos de las interacciones que definen su estructura nativa [Ferreiro et al., 2014]. Gracias a esto, las estructuras proteicas son robustas ante mutaciones puntuales. En cambio, las secuencias azarosas tienen contribuciones energéticas altamente frustradas, que desestabilizan la formación de estructuras tridimensionales.

Una componente entrópica toma protagonismo en el proceso de plegado: la cantidad de configuraciones que puede tomar una proteína desplegada supera ampliamente las que puede tomar una proteína en el estado nativo, es decir que al plegarse reduce su entropía. En el esquema del paisaje energético, la entropía se encuentra representada por el ancho del embudo. Es usual utilizar una representación que explicita las componentes entrópicas, la energía libre de Gibbs (G). En el caso de una proteína se observa que los estados plegado y desplegados son mínimos locales de G , con una diferencia de energía libre $\Delta G = G_{\text{desplegado}} - G_{\text{nativo}}$, y se encuentran separados por una barrera energética (figura 1.1b). Mientras la evolución da forma a la energía (fig. 1.1a), las mediciones experimentales de estabilidad dan información respecto de las variaciones de energía libre (fig. 1.1b). Las condiciones ambientales alteran el paisaje energético. Por ejemplo, al aumentar la temperatura o la concentración de desnaturalizantes el embudo energético adquiere mayores rugosidades que pueden servir de trampas cinéticas o termodinámicas al proceso de plegado. En la representación de Gibbs, se estabiliza el estado desplegado disminuyendo sus valores de energía libre (línea punteada en la figura 1.1b).

Diferentes mecanismos de plegado fueron propuestos [Fersht, 1997]. El modelo de *marco de referencia* propone que en primer lugar se forman las estructuras secundarias, y esto da la referencia necesaria para la formación de la estructura terciaria. El modelo de *colapso hidrofóbico* sostiene que las interacciones hidrofóbicas colapsan rápidamente a una estructura inicial que luego, sujeta a las restricciones impuestas, se acomoda hasta obtener la estruc-

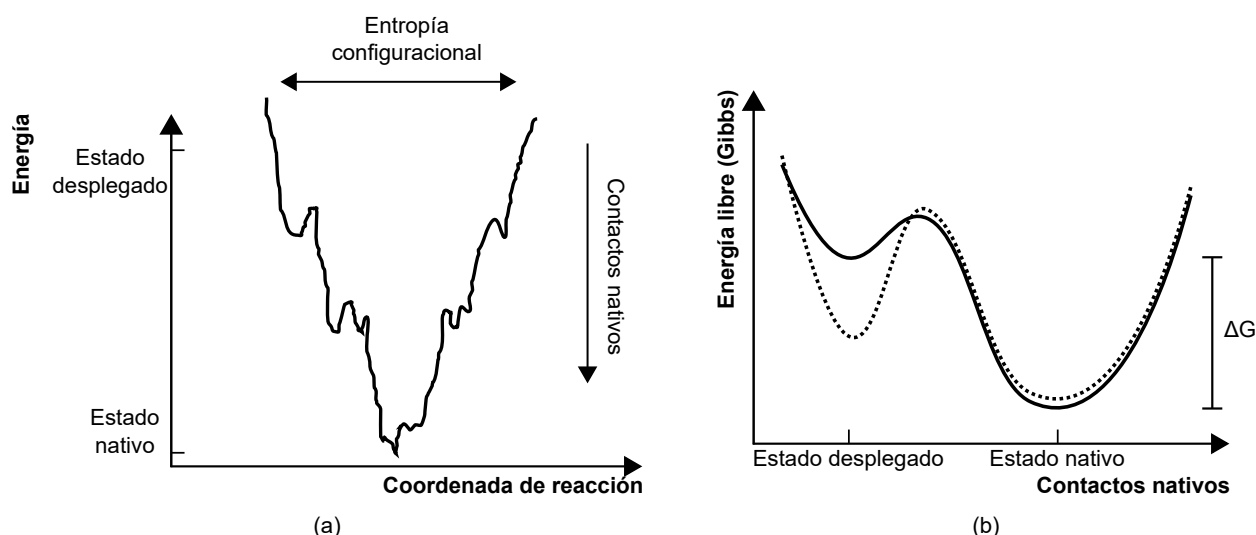


Figura 1.1: El plegado de proteínas naturales está guiado por un paisaje energético con forma de embudo. Un esquema unidimensional, en función de la coordenada de reacción, se presenta en (a). El estado nativo minimiza la energía del polipéptido. Esta disminución se encuentra guiada por la formación de contactos nativos. El estado desplegado tiene una alta entropía configuracional, es decir que el ensamble de estructuras desplegadas es mucho mayor que el de estructuras plegadas. En (b), representación de la energía libre de Gibbs. Los estados desplegado y nativo se encuentran separados por una barrera energética y tienen una diferencia de energía ΔG . Al incluir perturbaciones en el ambiente, por ejemplo agregar desnaturalizante o aumentar la temperatura, se estabiliza el estado desplegado (línea punteada), es decir se reduce ΔG . Figura basada en [Onuchic y Wolynes, 2004].

tura final. Al día de hoy, el mecanismo más aceptado y en acuerdo con la teoría de paisajes energéticos se denomina nucleación-condensación [Fersht, 1997]. Algunos residuos forman contactos, tanto de estructura secundaria como de largo alcance, y actúan como un núcleo a partir del cual el resto de la proteína se pliega. Sin embargo, la nucleación y la formación de toda la estructura se encuentran acoplados, siendo un proceso altamente cooperativo. Esta forma de plegarse se ha encontrado frecuentemente en las proteínas conocidas como de *dos estados*. Ellas transicionan entre dos configuraciones o estados macroscópicos, plegada o desplegada, sin presentar estados intermediarios (por ejemplo, plegados parciales). En estos casos la curva de plegado en función del tiempo para un ensamble de secuencias es exponencial. Un ejemplo puede verse en la figura 1.2a [Bryngelson et al., 1995]. Las condiciones ambientales pueden enlentecer o acelerar el proceso de plegado. De igual forma, el equilibrio alcanzado, es decir la proporción de secuencias que alcanzan el estado nativo a tiempos largos, puede variar. Esto equivale a variaciones en el ΔG de la representación en la figura 1.1b. Si se calcula la fracción plegada en el equilibrio con el ambiente para distintas

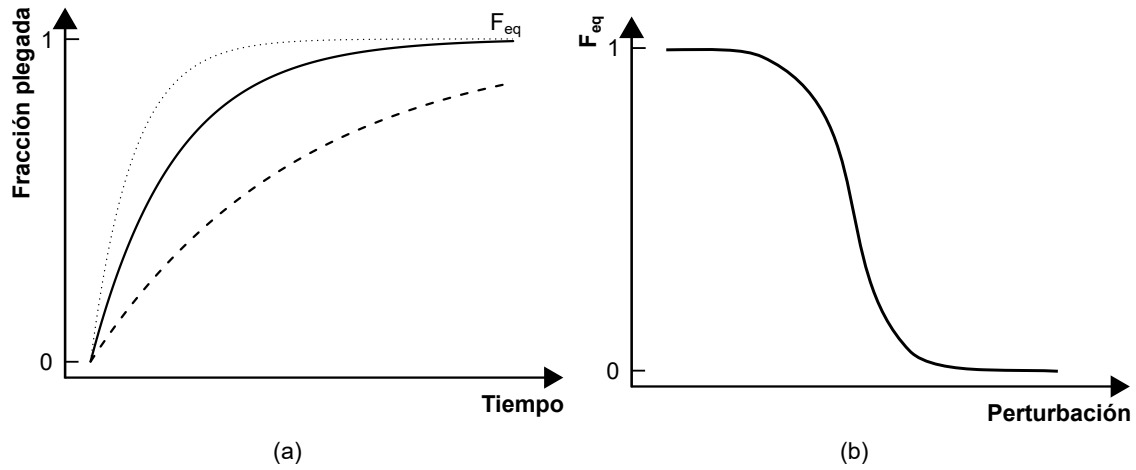


Figura 1.2: Esquema del proceso de plegado de un ensamble de proteínas. (a) Proporción de proteínas plegadas en función del tiempo. Diferentes condiciones del entorno pueden variar la velocidad del plegado, como esquematizan las líneas punteadas. La proporción de proteínas plegadas en el equilibrio, medida a tiempos largos, se denomina F_{eq} y puede ser menor a uno. En (b) se esquematiza el comportamiento de F_{eq} ante una perturbación de diferentes magnitudes. Típicamente temperatura o concentración de un químico desnaturalizante (por ejemplo, urea).

condiciones de una perturbación desnaturalizante, típicamente aumentos de temperatura o de concentración de un químico desnaturalizantes, se obtiene la curva 1.2b. La pendiente de la transición entre el estado nativo y el desplegado se relaciona con la cooperatividad del plegado.

Una amplia gama de enfermedades se encuentran asociadas a variaciones en el plegado de una especie proteica [Valastyan y Lindquist, 2014]. La primera informada fue la anemia de células falciformes. Los individuos con ella presentan una mutación puntual en la hemoglobina (de ácido glutámico a valina) que sufre cambios conformacionales en ambientes con bajo oxígeno. Estos derivan en la exposición de un parche hidrofóbico que permite la polimerización de las proteínas, y finalmente se manifiesta en la deformación de los glóbulos rojos. Actualmente, se conocen diferentes mecanismos por los cuales, si una especie proteica no es capaz de plegarse adecuadamente, emergen enfermedades en los individuos. Por ejemplo, en la fibrosis quística la proteína CFTR se pliega de forma incorrecta (debido a la delección de la fenilalanina en la posición 508) y por lo tanto es marcada para ser degradada. Otro mecanismo es la incorrecta localización de la proteína. Por ejemplo, alteraciones en $\alpha 1$ -antitripsina, una proteína de secreción, la retienen en el retículo endoplasmático. Dado que no se degrada, esta proteína se acumula y deriva en daños hepáticos. En algunos casos, la variante proteica

mutada es dominante llevando a mal funcionamiento incluso en individuos heterocigotos. Son ejemplos la epidermólisis ampollar simple o alteraciones en el factor de transcripción p53, relacionado a la proliferación del cáncer. En algunas especies proteicas, el plegado incorrecto la dota de una función tóxica. Un ejemplo es la estabilización de una estructura alternativa para la proteína APOE, que modifica su afinidad a los lípidos, interrumpiendo la función mitocondrial y evitando el crecimiento de las neuritas. Esta variante se encuentra en la mayoría de los pacientes con Alzheimer. Otro ejemplo, relacionado a procesos oncogénicos, es el SRC (Proto-oncogene tyrosine-protein kinase), cuya mutación en el sitio de fosforilación de auto-inhibición promueve la proliferación sin control de las células. Finalmente, variantes de proteínas mal plegadas toman la forma de fibras amiloides estables que generan agregados fibrosos e insolubles. Estos casos se encuentran asociados a desórdenes neurodegenerativos, incluyendo las enfermedades de Parkinson, Alzheimer y Huntington, y a las cataratas en los ojos.

La otra cara de este problema es el diseño inteligente de proteínas con fines tecnológicos, en particular en el desarrollo de drogas [Khoury et al., 2014]. Dada una estructura tridimensional, se busca determinar qué secuencia de aminoácidos se plegará de esa forma. Es un problema degenerado dado que muchas secuencias diferentes pueden plegarse en la misma estructura tridimensional. El poder de cómputo actual ha permitido el desarrollo de varias estrategias computacionales que buscan predecir estructura a partir de la información disponible. Entre ellos, se destacan I-Tasser [Zhang, 2008], HHpred-thread [Söding et al., 2005], RaptorX-YZ [Ma et al., 2013] y ROSETTA [Leaver-Fay et al., 2011]. Con distintos protocolos, buscan secuencias o segmentos homólogos en proteínas cuya estructura se conoce para predecir el plegado de una nueva variante. Estos métodos han permitido el desarrollo exitoso de aplicaciones terapéuticas [Khoury et al., 2014], que incluyen marcadores de cáncer de mama, candidatos a vacunas de HIV, inhibidores de la formación de fibras amiloides asociadas al Alzheimer, diseño de anticuerpos para identificar epítopes, e incluso el desarrollo de materiales proteicos. Sin embargo, las estimaciones *de novo* son todavía poco precisas, y muestran mejoras drásticas si se les provee de información estructural extra, por ejemplo incluir predicciones de estructura secundaria y de contactos entre residuos distantes [Kim et al., 2014]. En este trabajo abarco estrategias que permiten realizar predicciones de con-

tactos en el estado nativo a partir de la secuencia de aminoácidos, en el caso particular de proteínas repetitivas.

1.2. Proteínas repetitivas

En la figura 1.3 muestro ejemplos de estructuras de proteínas repetitivas, que probablemente delaten por qué son llamadas *repetitivas*. Estas proteínas se encuentran formadas por motivos estructurales de entre 20 y 40 aminoácidos, que se apilan entre ellos formando una estructura alargada, simil lineales. Definir unívocamente la repetición es aún un problema abierto [Parra et al., 2013]. En general, estas proteínas cuentan con un corazón hidrofóbico continuo y una superficie extendida accesible al solvente [Main et al., 2003a, Javadi y Itzhaki, 2013]. Los residuos cercanos en la estructura tridimensional se encuentran también cercanos en la secuencia, generalmente dentro de la misma repetición o en repeticiones adyacentes [Main et al., 2003a, Javadi y Itzhaki, 2013], a diferencia de las proteínas globulares. Es decir que se encuentran dominadas por interacciones de corto alcance. Entre los ejemplos más conocidos se encuentran las familias ANK, TPR, LRR, HEAT y ARM, que se encuentran en la figura 1.3. Otras proteínas con estructuras repetitivas, pero menos caracterizadas como proteínas repetitivas, se pueden ver en la figura 1.4.

Las repeticiones estructurales se encuentran reflejadas en las secuencias: las repeticiones de una misma familia presentan alta similitud en secuencia. Esto permite caracterizar las repeticiones por patrones de secuencia.

Las proteínas repetitivas se encuentran en todos los reinos y participan de un amplio espectro de procesos biológicos. Es posible que su éxito evolutivo se asocie a cierta facilidad de cambio: eventos de duplicación de repeticiones dan origen a proteínas largas y estables, basadas en módulos estructurales existentes, en vez de originar una nueva conformación estructural. Con un andamio en común, variaciones en la secuencia de aminoácidos les han permitido involucrarse en una gama de funciones muy diversa. Por ejemplo, representantes de la familia repetitiva TPR aparecen en la regulación del ciclo celular, en el control de la transcripción, en el transporte de proteínas, en la neurogénesis, y auxiliando el pegado de otras proteínas [Main et al., 2003a, Blatch y Lässle, 1999]. Representantes de la familia

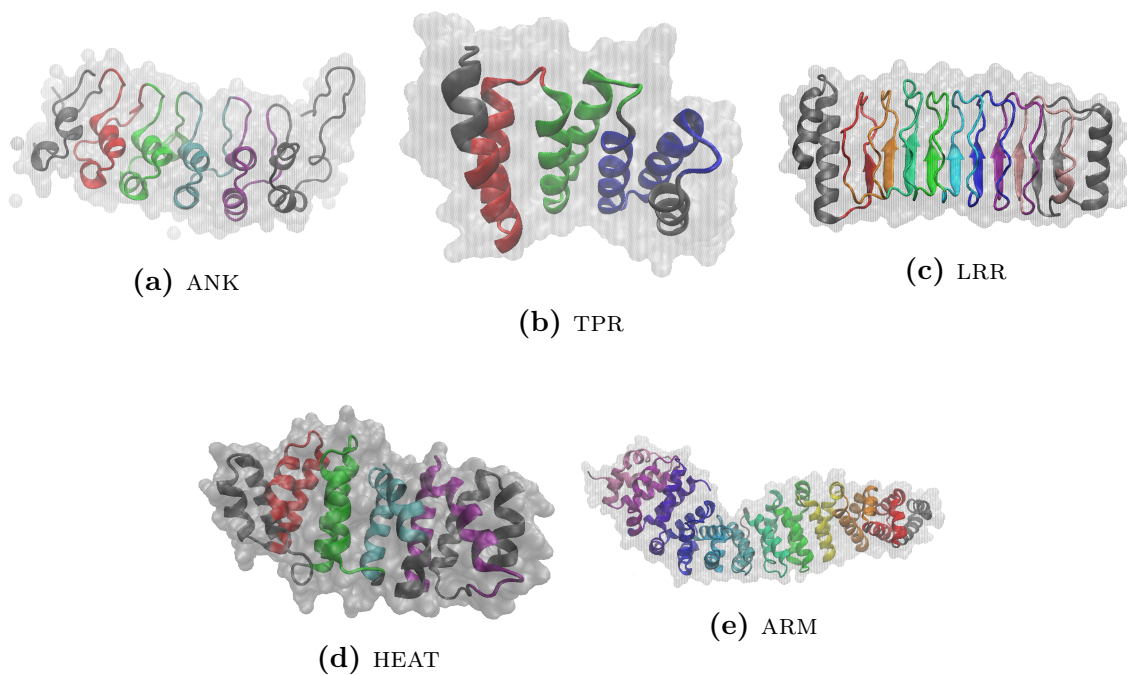


Figura 1.3: Ejemplos de estructuras de proteínas repetitivas de familias típicas.

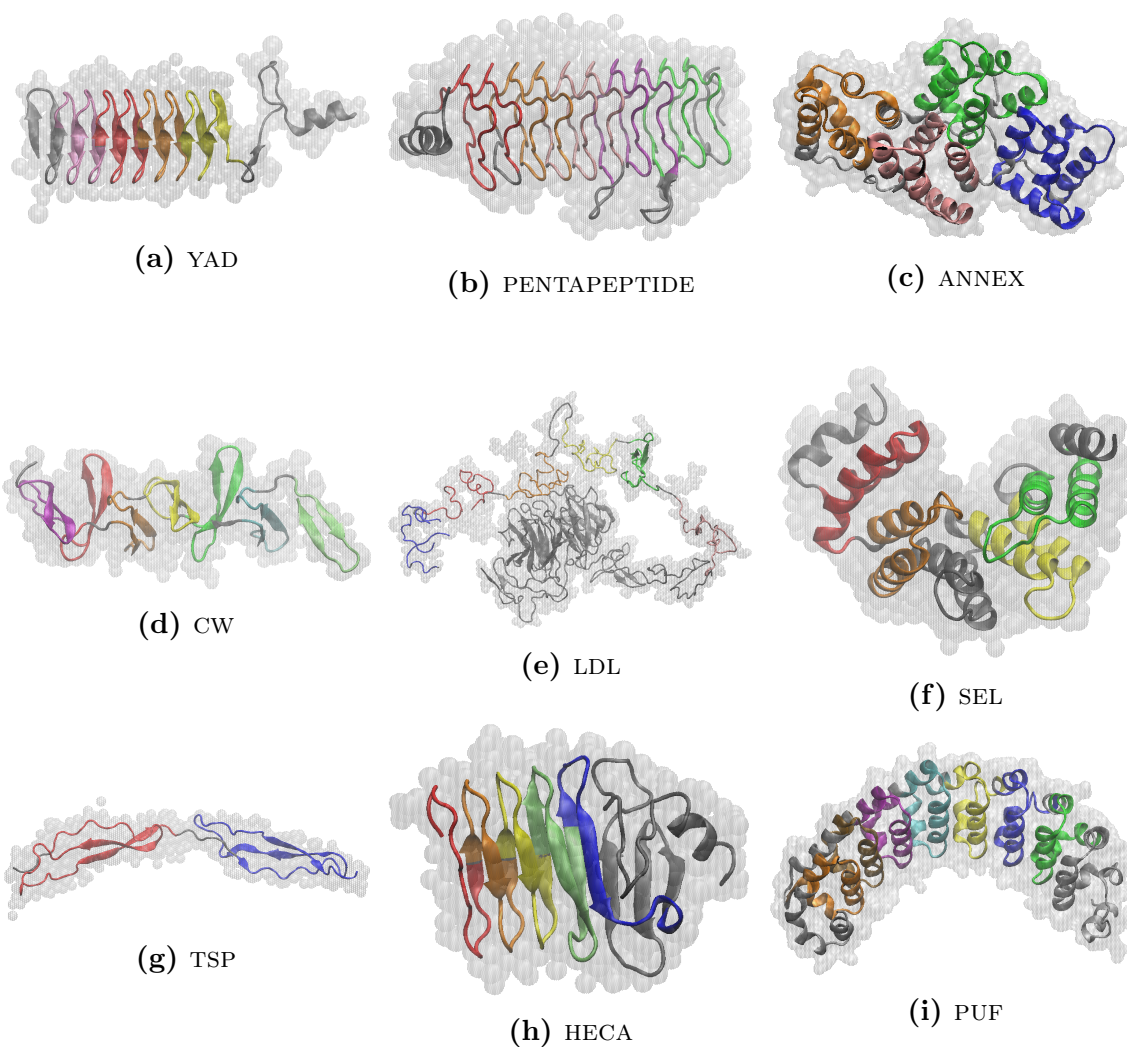


Figura 1.4: Ejemplos de estructuras de proteínas repetitivas de familias menos estudiadas.

ANK incluyen inhibidores de quinasas, reguladores transcripcionales, organizadores del citoesqueleto, reguladores del desarrollo y toxinas [Sedgwick y Smerdon, 1999]. Proteínas LRR participan del desarrollo temprano de mamíferos y desarrollo neuronal, en el tráfico celular, en la adherencia celular, en la polarización de la célula, en la regulación de expresión de genes, la señalización de la apoptosis y hasta incluye a la proteína que evita el congelamiento en la zanahoria [Kobe y Kajava, 2001]. Variantes de la familia ARM participan en la regulación transcripcional, en la adherencia celular, en la actividad de supresión de tumores y el transporte nucleocitoplásmico. [Parmeggiani et al., 2008]. Específicamente, la actividad molecular relacionada a las proteínas repetitivas es la mediación de interacciones proteína proteína [Sedgwick y Smerdon, 1999, Kobe y Kajava, 2001, Blatch y Lässle, 1999], y no presentan actividad enzimática.

Dado que los contactos son locales, uno puede pensar que el plegado de cada repetición es independiente de las demás. Sin embargo, esto no es así, sino que tienen un balance entre las señales locales de plegado y el comportamiento global de la cadena polipeptídica [Espada et al., 2015]. La mayoría de las proteínas repetitivas de un único dominio muestran, en mediciones en el equilibrio, un proceso de plegado cooperativo y de dos estados tanto para perturbaciones térmicas como químicas [Main et al., 2003a]. Un ejemplo puede verse en la figura 1.5. Más aún, las proteínas repetitivas se pliegan con tasas mucho más lentas que las esperadas por la cantidad de contactos entre residuos distantes en estructura primaria (*orden de contactos*).

Además de ser cooperativas, las proteínas repetitivas son modulares. Se han construido diferentes variantes proteicas donde repeticiones enteras fueron removidas o insertadas y la proteína puede plegarse a una conformación estable [Main et al., 2003a]. Sin embargo, una única repetición no es capaz de plegarse [Javadi y Itzhaki, 2013]. En TPR, la estructura mínima capaz de plegarse contiene una repetición y media [Main et al., 2003b], mientras que para ANK, contiene 3 repeticiones [Mosavi et al., 2002]. La cantidad de repeticiones es variable, y no es necesariamente un múltiplo de la mínima cantidad.

El estudio de los caminos de plegado también ha despertado interés en estas proteínas gracias a su estructura simétrica y elongada [Ferreiro et al., 2014, Kloss et al., 2008]. En algunas proteínas repetitivas se ha visto que tienen un plegado polarizado [Tang et al., 2003,

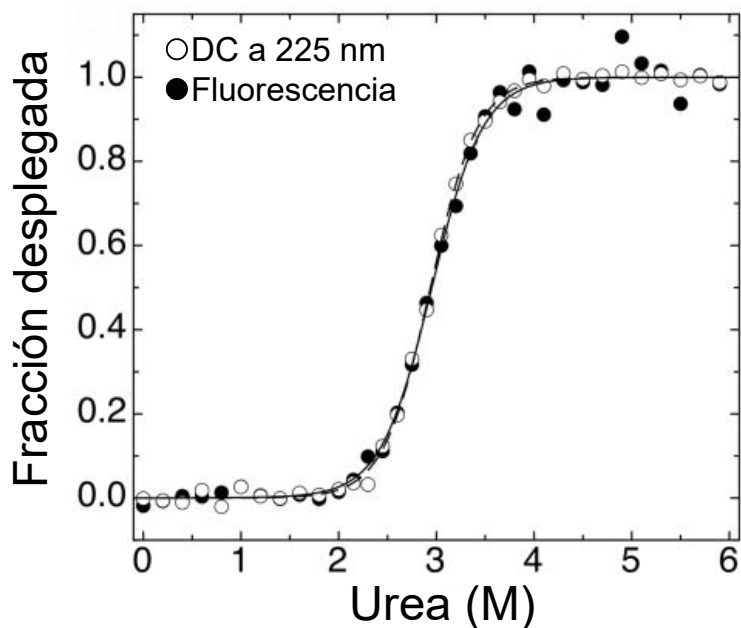


Figura 1.5: Mediciones experimentales sobre el proceso de desplegamiento en $i\kappa b\alpha$, una proteína de la familia ANK. Las mediciones se realizaron con dos métodos experimentales: dicroísmo circular (círculos vacíos) y fluorescencia (círculos llenos). Se observa que el proceso es compatible con uno de dos estados: en concentraciones bajas de urea toda la población proteica se encuentra plegada y en altas concentraciones de urea se encuentra desnaturalizada. El cambio es abrupto, indicando una alta cooperatividad de plegado. Imagen adaptada de [DeVries et al., 2011].

[Lowe y Itzhaki, 2007, Courtemanche y Barrick, 2008]. En experimentos cinéticos se han encontrado estados intermediarios en el plegado o desplegamiento. Estos intermediarios suelen estar sesgados hacia los extremos de la proteína, aunque también existen intermediarios con los extremos desplegados y las repeticiones centrales plegadas [Tang et al., 2003, Lowe y Itzhaki, 2007, Courtemanche y Barrick, 2008]. Todas estas observaciones son consistentes con un paisaje energético simétrico [Ferreiro y Wolynes, 2008], como el esquematizado en la figura 1.6. La similitud entre las repeticiones da lugar a diferentes configuraciones con energías libres similares. Esto origina rutas de plegado paralelas. Pequeñas perturbaciones en este paisaje, mediante mutaciones puntuales, producen mayores rugosidades que actúan como trampas locales que también han sido observadas experimentalmente. En la familia ANK se observó que las repeticiones situadas en los extremos de la cadena polipeptídica tienen las estructuras con mayores variaciones energéticas y secuencias más variables [Parra et al., 2015]. Esto posiblemente se asocia a que tienen mayor superficie de interacción con el solvente, dado que la interacción con repeticiones adyacentes suele ser estabilizante. Además, estas

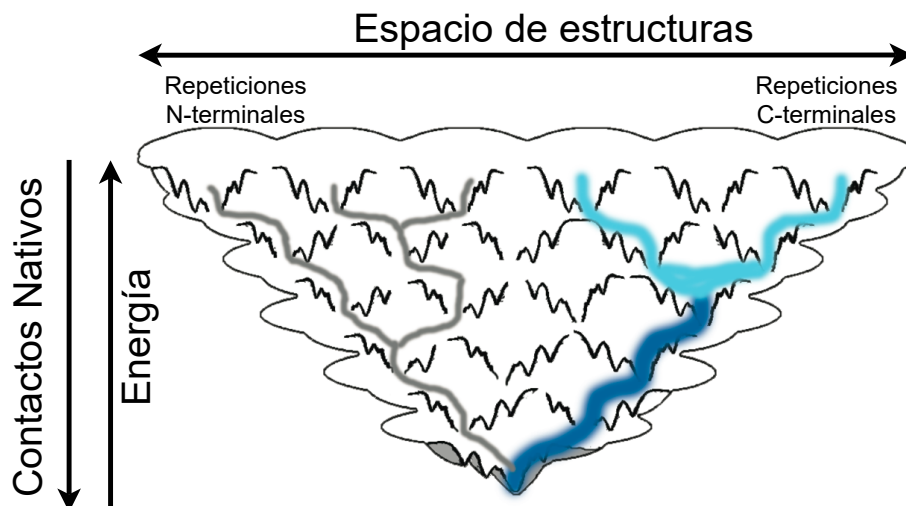


Figura 1.6: Esquema de paisaje energético de una proteína repetitiva. El paisaje tiene forma de embudo, y refleja la simetría estructural. Esta simetría da lugar a caminos paralelos de plegado (esquematisados en gris y azul), y configuraciones diferentes tienen energías similares. Adaptado de [Ferreiro y Wolynes, 2008].

rugosidades energéticas en los extremos posiblemente sean indicadores de su participación en las actividades biológicas específicas.

Mediar interacciones proteína proteína, constituir un andamio que puede adaptarse a diferentes interacciones y su modularidad son las propiedades responsables del interés por el diseño de proteínas repetitivas. En la bibliografía se pueden encontrar diseños de proteínas repetitivas que fueron experimentalmente validados. Una primera estrategia se basa en la búsqueda de la secuencia *consenso*. Esta se define como aquella secuencia que en cada posición tiene el aminoácido más frecuente entre las proteínas de la familia. Esta estrategia considera implícitamente que los residuos altamente conservados definen el plegado, mientras que los residuos más variables especifican la unión a otras moléculas. Bajo este esquema, se sintetizaron proteínas análogas a las familias ANK [Binz et al., 2003, Mosavi et al., 2002], TPR [Main et al., 2003b, Kajander et al., 2007], LRR [Stumpp et al., 2003], HEAT [Urvoas et al., 2010], y ARM [Parmeggiani et al., 2008, Varadamsetty et al., 2012], y proteínas TALE [Geiger-Schuller y Barrick, 2016] (ver debajo). En todos los casos, las estructuras que fueron resueltas mostraron una estructura tridimensional similar a sus contrapartes naturales. Construcciones de hasta 7 repeticiones presentan un plegado cooperativo y de dos estados, al igual que las proteínas naturales [Javadi y Itzhaki, 2013]. A medida que se aumenta la

cantidad de repeticiones, la estructura se vuelve más estable. En todos los casos, las proteínas diseñadas son más estables que las naturales de igual largo: hasta 20 kcal mol⁻¹ para diseñadas, contra 5 - 10 kcal mol⁻¹ para proteínas naturales [Javadi y Itzhaki, 2013]. Es posible que esto se deba a la remoción de interacciones desfavorables, y a la exclusión de inserciones y deleciones que proporcionan inhomogeneidades [Main et al., 2003a]. El diseño por consenso no fue exitoso en el dominio WW, una familia globular. Para poder diseñar proteínas sintéticas plegables en estructuras análogas a WW fue necesario recurrir al análisis de interacciones entre pares de residuos mediante el estudio de propiedades estadísticas de un conjunto de secuencias de la familia [Socolich et al., 2005].

Las proteínas diseñadas a partir del consenso funcionan como un andamiaje que asegura la plegabilidad, y las posiciones variables pueden ser seleccionadas para optimizar la unión a otras moléculas. Estos diseños alcanzaron aplicaciones concretas en diferentes campos. Las proteínas DARPins (variaciones de ANKS inicialmente diseñadas por consenso) han sido adaptadas para marcar enfermedades retinales, diagnósticos de cáncer, inhibición de la infección de HIV y fueron utilizadas para investigar señalización intracelular [Plückthun, 2015]. Las proteínas sintéticas α Rep fueron desarrolladas a partir de un subconjunto de proteínas HEAT [Urvoas et al., 2010, Valerio-Lepiniec et al., 2015]. Las proteínas α Rep son un andamio que puede optimizarse para unirse específicamente a otros sustratos. En particular, han servido como cofactores para cristalizar proteínas que no cristalizan solas y como marcadores en células vivas. Finalmente, las proteínas TALE (Efectores similares a Activadores de Transcripción, por sus siglas en inglés) han despertado interés por su capacidad de unirse específicamente al ADN. Estas proteínas están constituidas por repeticiones de 34 aminoácidos, donde los residuos de las posiciones 12 y 13 determinan específicamente a que base del ADN se unen. Usando esta especificidad, se han realizado construcciones proteicas unidas a nucleasas, denominadas TALEN, que permiten realizar ediciones de secuencias de ADN. También han sido modificadas para unirse al ADN y reprimir o activar transcripciones, para modificar el ADN mediante demetilación, y para estudiar las dinámicas de la cromatina codificando fluoróforos con alta especificidad de secuencia [Bogdanove y Voytas, 2011].

Otras familias proteicas, globulares, también presentan repeticiones, pero, a diferencia de las proteínas repetitivas, sus estructuras no son elongadas, sino que se cierran sobre

sí mismas. Son ejemplos las configuraciones hélices- β (que incluye a las familias WD40 y KELCH). Las estructuras de algunos ejemplos, que se estudian en este trabajo, se muestran en la figura 1.7. En particular, la familia WD40, al igual que las proteínas repetitivas, se ve involucrada en interacciones proteína proteína, y no se conocen variantes con actividad enzimática. Entre la diversidad de procesos biológicos en los que se ve involucrada se puede destacar la transcripción, la división celular, el ensamblado del citoesqueleto, la quimiotaxis y el procesamiento de ARN [Stirnemann et al., 2010].

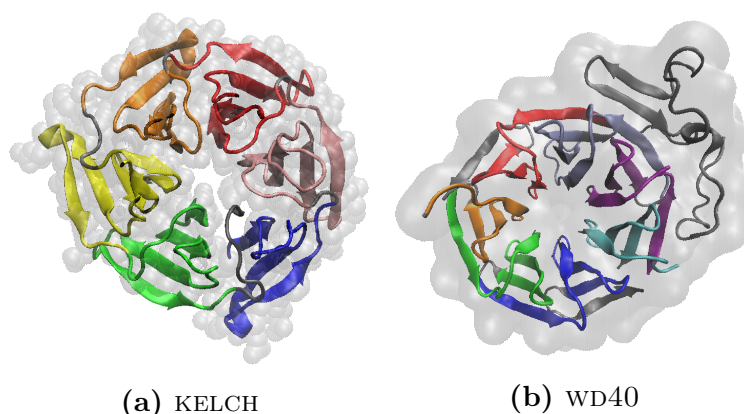


Figura 1.7: Ejemplos de estructuras de proteínas globulares con repeticiones.

1.3. Coevolución en secuencias de proteínas

A lo largo de tiempos evolutivos, las proteínas sufren modificaciones en sus secuencias que cambian sus propiedades fisicoquímicas, y por lo tanto su estabilidad de plegado o su capacidad de interacción con otras moléculas. Estas modificaciones pueden ser favorables, neutras, o deletéreas (desfavorables). En el último caso, cuando una mutación implica una pérdida en la estabilidad del plegado, esta puede ser compensada por la sustitución de otros aminoácidos que interactúen con él en la estructura plegada. Por lo tanto, se espera que los residuos que formen contactos en una estructura nativa tengan patrones de sustitución de aminoácidos restringidos, y por lo tanto correlacionados. Bajo esta premisa, los métodos de análisis de coevolución en secuencias de proteínas tienen como objetivo encontrar correlaciones entre mutaciones de aminoácidos a lo largo de las secuencias [De Juan et al., 2013]. Con la creciente disponibilidad de secuencias, incluso clasificadas según homología, este tipo de análisis ha tomado relevancia.

Entre las primeras medidas de coevolución se destaca la *información mutua* (IM), tomada de la teoría de información. IM fue diseñada para medir correlación entre dos magnitudes que pueden tomar q valores discretos no numéricos. En este área, se utilizó para medir la correlación de sustitución de aminoácidos en un par de posiciones de una proteína, y predecir contactos [Korber et al., 1993]. Inicialmente, esta medida era demasiado sensible a la conservación de la secuencia o a sesgos filogenéticos, que fueron superados por mejoras posteriores en los cálculos [De Juan et al., 2013]. Sin embargo, la capacidad predictiva de la IM fue ampliamente mejorada por nuevos métodos capaces de detectar correlaciones espúreas.

En Weigt et al. [2009] postularon la importancia de distinguir correlaciones en secuencias producidos por efectos *directos* de *indirectos*. Los efectos indirectos son aquellos cambios que puede producir una sustitución en otras regiones de la proteína sin estar necesariamente en contacto. Residuos conectados por efectos indirectos tienen valores altos de co-evolución y pueden estar conectados por interacciones en común con otros residuos. En el trabajo Burger y Van Nimwegen [2010] estimaron que en promedio dos residuos correlacionados pero no en contacto, tienen un camino de otros cinco residuos en contacto que los une [Burger y Van Nimwegen, 2010]. Además, estos cambios pueden estar asociados a la acumulación de varias interacciones pequeñas. Si bien [Weigt et al., 2009] versa sobre la identificación de contactos entre proteínas interactuantes, pronto se extendieron sus ideas al problema del plegado de proteínas. Las medidas de correlación directa mostraron alta eficiencia para predecir contactos en la estructura nativa de proteínas globulares [Morcos et al., 2011, Marks et al., 2011].

Bajo estas ideas, diferentes algoritmos fueron desarrollados para poder distinguir las correlaciones directas, y determinar con ellas residuos en contacto en la estructura nativa. Morcos et al. [2011] utilizaron una aproximación de campo medio para un modelo de Potts. Cocco et al. [2013] usaron la misma aproximación para el modelo de Hopfield-Potts, derivado de neurociencias. Ekeberg et al. [2013, 2014] desarrollaron la aproximación de maximización de la pseudo-verosimilitud. Andreatta et al. [2013] propusieron incorporar las matrices BLOSUM (descriptas en Métodos) en la cuantificación de las secuencias, y utilizar un algoritmo de *regularización de cuadrados mínimos*. Kamisetty et al. [2013] crearon GREMLIN, que además de las predicciones por máxima pseudo-verosimilitud incluye información *a priori*

de estructura basados en estructura secundaria y separación entre residuos en la secuencia. Sucesivamente, estos trabajos mejoraron las tasas de predicción estructural a la vez que redujeron el tiempo de cómputo y la cantidad de secuencias que se necesitan conocer.

Derivaciones de estos modelos se han extendido para predecir estructuras nativas utilizando simulaciones computacionales SBM (modelos basados en estructura) [Sutto et al., 2015, Sulkowska et al., 2012], interacciones entre proteínas [Cheng et al., 2014], cambios conformacionales en estructuras relacionado a las funciones proteicas [Jana et al., 2014], y caracterizar mutaciones puntuales [Figliuzzi et al., 2015, Hopf et al., 2015, Contini y Tiana, 2015].

En la siguiente sección describiré con detalle los métodos que hemos aplicado a proteínas repetitivas: información mutua, el modelo de Potts, la aproximación de campo medio y la aproximación de máxima pseudo-verosimilitud.

1.4. Modelo estadístico

El objetivo de esta sección es desarrollar un modelo que permita estudiar la probabilidad de ocurrencia de las secuencias proteicas, $P(\vec{s})$. Cada proteína es representada como una secuencia de aminoácidos $\vec{s} = a_1, \dots, a_n$, donde los a_i puede tomar $q = 21$ valores distintos: cada uno de los aminoácidos o un espacio vacío, es decir $a_i \in \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y, -\}$. En este trabajo tomaré un conjunto de M secuencia, todas del mismo largo L . Cada secuencia tiene asociada una probabilidad de ocurrencia $P(\vec{s})$, cuya expresión es desconocida. Sin embargo, gracias a datos experimentales colectados, es posible incorporar las siguientes restricciones a esta expresión: que se verifiquen las probabilidades marginales y conjuntas de ocurrencia de todos los aminoácidos (a_i y b_j), a lo largo de todas las posiciones de la secuencia (i y j). Estas restricciones se pueden formular como:

$$P_i(a_i) \equiv \sum_{a_k, k \neq i} P(\vec{s}) = f_i(a_i) \quad (1.1)$$

$$P_{ij}(a_i, b_j) \equiv \sum_{a_k, k \neq i, j} P(\vec{s}) = f_{ij}(a_i, b_j) \quad (1.2)$$

donde $P_i(a_i)$ y $P_{ij}(a_i, b_j)$ se refieren a los valores de probabilidad determinados por el modelo, mientras que $f_i(a_i)$ y $f_{ij}(a_i, b_j)$ son los valores medidos experimentalmente. La hipótesis más fuerte detrás de imponer estas restricciones es asumir que las correlaciones de a pares son condición necesaria y suficiente para describir los patrones de correlación de órdenes mayores [Levy et al., 2017].

En este trabajo, para derivar un modelo explícito de las probabilidades de ocurrencia de secuencias, utilizaré el método de maximización de entropía sujeto a las restricciones enumeradas. Este procedimiento, desarrollado en [Jaynes, 1957], es introducido en el apéndice A. En la sección 1.4.2 desarrollaré el modelo aplicado a proteínas, y en la sección 1.4.4 para proteínas repetitivas. Finalmente, describiré dos aproximaciones utilizadas: de campo medio (sección 1.4.5) y de máxima pseudo-verosimilitud (sección 1.4.6). Previamente, por completitud de las magnitudes a analizar, defino la información mutua (sección 1.4.1).

1.4.1. Información mutua

La información mutua (IM) es una medida de correlación entre dos posiciones. Se encuentra definida por la siguiente expresión:

$$\text{IM}_{ij} = \sum_{a_i}^q \sum_{b_j}^q f_{ij}(a_i, b_j) \log \left(\frac{f_{ij}(a_i, b_j)}{f_i(a_i)f_j(b_j)} \right) \quad (1.3)$$

donde $f_i(a_i)$ y $f_j(b_j)$ son las frecuencias marginales de ocurrencia del aminoácido a_i en la posición i y b_j en la posición j respectivamente, $f_{ij}(a_i, b_j)$ es la frecuencia de ocurrencia conjunta de los aminoácidos a_i en la posición i y b_j en la posición j , y la suma se realiza sobre los $q=21$ valores posibles de residuos.

En el caso límite en que ambas posiciones son completamente independientes, las frecuencias cumplen $f_{ij}(a_i, b_j) = f_i(a_i) \cdot f_j(b_j)$, por lo tanto el argumento del logaritmo es 1 y la suma da cero, es decir $\text{IM}_{ij} = 0$. En el caso contrario, cuando ambas posiciones se encuentran completamente correlacionadas, las frecuencias cumplen $f_{ij}(a_i, b_j) = f_i(a_i) = f_j(b_j)$ y

la expresión se reduce a la definición de entropía de una posición S_i :

$$\text{IM}_{ij} = \sum_{a_i}^q \sum_{b_j}^q f_i(a_i) \log \left(\frac{f_i(a_i)}{f_i(a_i)f_i(a_i)} \right) = - \sum_{a_i}^q f_i(a_i) \log (f_i(a_i)) \equiv S_i \quad (1.4)$$

Este es el máximo valor que puede tomar IM. Por este motivo, una normalización usual es dividir IM_{ij} por $\sqrt{S_i \cdot S_j}$, garantizando valores entre 0 y 1.

1.4.2. Modelo por máxima entropía

Este trabajo parte de modelar la probabilidad de ocurrencia $P(\vec{s})$ de acuerdo al procedimiento propuesto por Jaynes [1957], y que se encuentra detallado en el apéndice A. Como propuso Shannon [1948], la expresión de la probabilidad con menor incerteza es aquella que minimiza la expresión

$$H(P(\vec{s})) = -K \sum_i P(\vec{s}) \ln P(\vec{s}) \quad (1.5)$$

donde tomaré, sin pérdida de generalidad, $K = 1$. Además, $P(\vec{s})$ debe cumplir con las restricciones observadas, en este caso, las frecuencias de ocurrencia de los aminoácidos. Para esto, las frecuencias se expresan como los valores medios de una función partida:

$$f_i(a_i) = \sum_{\{\vec{s}\}} P(\vec{s}) \theta_i(a_i) \quad (1.6)$$

donde $f_i(a_i)$ es la frecuencia observada del aminoácido a_i en la posición i , y

$$\theta_i(a_i) = \begin{cases} 1 & \text{si } \vec{s} \text{ tiene } a_i \text{ en la posición } i \\ 0 & \text{sino} \end{cases} \quad (1.7)$$

Análogamente, para las frecuencias conjuntas:

$$f_{ij}(a_i, b_j) = \sum_{\{\vec{s}\}} P(\vec{s}) \theta_i(a_i) \theta_j(b_j) \quad (1.8)$$

La condición de normalización se impone sobre la probabilidad de ocurrencia de cada secuencia:

$$\sum_{\vec{s}} P(\vec{s}) = 1 \quad (1.9)$$

Utilizando multiplicadores de Lagrange la expresión para la probabilidad resulta (ver apéndice A):

$$P(\vec{s}) = \frac{1}{Z} \exp \left[- \sum_i^L h_i(a_i) - \sum_i^L \sum_{j>i}^L j_{ij}(a_i, b_j) \right] \quad (1.10)$$

donde L es el largo de las proteínas y Z es el factor de normalización:

$$Z = \sum_{\{\vec{s}\}} \exp \left[- \sum_i^L h_i(a_i) - \sum_i^L \sum_{j>i}^L j_{ij}(a_i, b_j) \right] \quad (1.11)$$

Estas expresiones remiten a un modelo de Potts [Wu, 1982], que permite enlazar las expresiones a la mecánica estadística. Se puede asociar el sistema a un ensamble canónico (en equilibrio térmico) [Pathria, 1972], cuyo Hamiltoniano es

$$H(\vec{s}) = \sum_i^L h_i(a_i) + \sum_i^L \sum_{j>i}^L j_{ij}(a_i, b_j) \quad (1.12)$$

donde los parámetros $h_i(a_i)$ son los campos locales y $j_{ij}(a_i, b_j)$ las energías de interacción. La distribución de probabilidades se encuentra determinada por la distribución de Boltzmann-Gibbs

$$P(\vec{s}) = \frac{1}{Z} e^{-\beta H(\vec{s})} \quad (1.13)$$

con $\beta = 1/k_B T$, k_B es la constante de Boltzmann y T la temperatura del sistema. La función de partición se relaciona con la energía libre de Helmholtz:

$$F = -k_B T \ln Z \quad (1.14)$$

Usualmente a lo largo de este trabajo tomaré $\beta = 1$. A partir de la función de partición se obtiene toda la información relevante del sistema:

$$\frac{\partial F}{\partial h_i(a_i)} = \frac{\partial(-\ln Z)}{\partial h_i(a_i)} = P_i(a_i) \quad (1.15)$$

$$\frac{\partial^2 F}{\partial h_i(a_i) \partial h_j(b_j)} = \frac{\partial^2(-\ln Z)}{\partial h_i(a_i) \partial h_j(b_j)} = -P_{ij}(a_i, b_j) + P_i(a_i)P_j(b_j) = -C_{ij}(a_i, b_j) \quad (1.16)$$

En la última igualdad definí la matriz de correlaciones conectadas $C_{ij}(a_i, b_j) = P_{ij}(a_i, b_j) - P_i(a_i)P_j(b_j)$. Nuevamente, las derivaciones correspondientes se pueden ver en el apéndice A.

1.4.3. Gauge

El modelo de la ecuación 1.12 tiene $Lq + \binom{L}{2}q^2$ parámetros, pero no son todos independientes entre sí. Esto se debe a que las condiciones impuestas para restringir el modelo, ecuaciones 1.1 y 1.2, no son independientes puesto que las probabilidades deben estar normalizadas:

$$\sum_{a_i=1}^q P_i(a_i) = 1 \quad (1.17)$$

$$\sum_{b_j=1}^q \sum_{a_i=1}^q P_{ij}(a_i, b_j) = 1 \quad (1.18)$$

La cantidad de parámetros independientes son $L(q-1) + \binom{L}{2}(q-1)^2$. Debido a esto, los valores de energía no se encuentran unívocamente definidos, aunque los valores relativos de energía sí lo están. Para fijar los valores absolutos de energía es necesario fijar un valor de referencia o *gauge*. En este caso, el gauge puede ser electo sin pérdida de generalidad. En este trabajo utilizaré dos posibilidades.

La primera es fijar todos los parámetros relacionados al espacio vacío a cero: $h_i(-) = 0$ $j_{ij}(a_i, -) = j_{ij}(-, a_i) = 0$. Este será utilizado en el resto de la introducción teórica.

La segunda opción es estandarizar los valores, pidiendo que

$$\sum_{a_i} h_i(a_i) = 0 \quad (1.19)$$

$$\sum_{a_i} \sum_{b_j} j_{ij}(a_i, b_j) = 0 \quad (1.20)$$

Esta elección la utilizaré en el capítulo 3.

En ambos casos, el número de parámetros libres se corresponden al número de restricciones independientes, garantizando una solución única al modelo derivado de la maximización de la entropía. Además, notar que los acoplamientos no tienen sentido en una única posición, como no lo tienen las probabilidades $P_{ii}(a_i, b_i)$. Por lo tanto $j_{ii}(a_i, b_i) = 0$.

1.4.4. Modelo alternativo para proteínas repetitivas

Las proteínas repetitivas presentan alta similitud en secuencia entre repeticiones sucesivas. En este trabajo propongo (y estudiaré en detalle en el capítulo 3), un modelo que explicita esta propiedad: agrego como restricción extra la frecuencia en que repeticiones consecutivas tienen una determinada identidad en secuencia. Esta frecuencia la anotaré f_{id} , y es introducida en completa analogía con las frecuencias marginales (ecuaciones 1.6 y 1.7):

$$f_{id} = \sum_{\{\vec{s}\}} P(\vec{s}) \phi(id) \quad (1.21)$$

con

$$\phi(id) = \begin{cases} 1 & \text{si la identidad entre las repeticiones es id} \\ 0 & \text{sino} \end{cases} \quad (1.22)$$

Al maximizar la entropía incluyendo esta restricción, se obtiene el campo:

$$H(\vec{s}) = \sum_i^L h_i(a_i) + \sum_i^L \sum_{j>i}^L j_{ij}(a_i, b_j) + \lambda_{id} \quad (1.23)$$

donde únicamente se suma el parámetro λ_{id} correspondiente a la identidad entre las repeticiones que constituyen la secuencia $\vec{s}$. Este modelo será estudiado en el capítulo 3 para secuencias conformadas por dos repeticiones consecutivas, y posteriormente generalizado para secuencias más largas.

1.4.5. Aproximación de campo medio

Analizar la función de partición 1.11 es intratable en la mayoría de los casos dado que el espacio de secuencias $\{\vec{s}\}$ sobre el que se computa la suma es demasiado grande. Por ejemplo, para secuencias de largo 5, la cantidad de variantes supera los 4 millones. En Morcos et al. [2011] los autores desarrollan una aproximación de campo medio que permite reducir el sistema de análisis y derivar una expresión de correlación entre posiciones de una proteína.

Para poder realizar la aproximación de campo medio es necesario explicitar que los acoplamientos entre residuos en la ecuación 1.12 son una perturbación a los campos locales. Para esto reescribo el campo como:

$$H = \sum_{i=1}^L h_i(a_i) + \alpha \sum_{i=1}^L \sum_{j=i}^L j_{ij}(a_i, a_j) \quad (1.24)$$

donde agregué el parámetro α , que permite interpolar entre el modelo de sitio independiente (con $\alpha = 0$) y el modelo original (con $\alpha = 1$). Como expliqué anteriormente, este campo permite definir una función de partición Z y una energía libre dada por $F = -\ln(Z)$. Utilizando la definición de la transformada de Legendre (ecuación A.70) defino el potencial de Gibbs:

$$G(\alpha) = \ln(Z) + \sum_i^L \sum_{a_i=1}^{q-1} h_i(a_i) P_i(a_i) \quad (1.25)$$

cuya primer derivada se relaciona con los parámetros de campo local:

$$h_i(a_i) = \frac{\partial G(\alpha)}{\partial P_i(a_i)} \quad (1.26)$$

Y la segunda derivada permite relacionarlos con la matriz de correlaciones

$$\frac{\partial h_i(a_i)}{\partial P_j(b_j)} = \frac{\partial^2 G(\alpha)}{\partial P_i(a_i) \partial P_j(b_j)} = \left[\frac{\partial^2 (-\ln Z)}{\partial h_i(a_i) \partial h_j(b_j)} \right]^{-1} = -(C_{ij}(a_i, b_j))^{-1} \quad (1.27)$$

En la segunda igualdad, usé la propiedad A.74 de las transformadas de Legendre, para pasar de G a $F = -\ln Z$, y en la última igualdad usé el resultado de la ecuación 1.16. A partir

de ella, y realizando una aproximación en torno de $\alpha \approx 0$, es posible despejar las siguientes expresiones

$$\frac{P_i(a_i)}{P_i(q)} = \exp \left\{ -h_i(a_i) - \sum_{j|j \neq i} \sum_{b_j=1}^{q-1} j_{ij}(a_i, b_j) P_j(b_j) \right\} \quad (1.28)$$

$$(C^{-1})_{ij}(a_i, b_j) = \begin{cases} -j_{ij}(a_i, b_j) & \text{si } i \neq j \\ \frac{\delta_{a_i, b_j}}{P_i(a_i)} + \frac{1}{P_i(q)} & \text{si } i = j \end{cases} \quad (1.29)$$

de las cuales es posible despejar los parámetros $j_{ij}(a_i, b_j)$ y $h_i(a_i)$ en función de datos conocidos si tomamos $P_i(a_i) = f_i(a_i)$.

Información directa

Para condensar toda esta información en un único parámetro, en Morcos et al. [2011] definieron la información directa como una medida de información mutua debida únicamente a acoplamientos directos. Para un par de posiciones aisladas definieron

$$P_{ij}^{dir}(a_i, b_j) = \frac{1}{Z_{ij}} \exp \left\{ -j_{ij}(a_i, b_j) + \tilde{h}_i(a_i) + \tilde{h}_j(b_j) \right\} \quad (1.30)$$

donde los campos locales son determinados, para cada par de posiciones, de manera que verifiquen las distribuciones marginales de aminoácidos:

$$f_i(a_i) = \sum_{b_j}^q P_{ij}^{dir}(a_i, b_j) \quad (1.31)$$

$$f_j(b_j) = \sum_{a_i}^q P_{ij}^{dir}(a_i, b_j) \quad (1.32)$$

Y Z_{ij} , la función de partición para las variantes en únicamente las posiciones i y j , se obtiene normalizando las probabilidades. La información directa, en analogía con la información

mutua, es definida como:

$$\text{ID}_{ij} = \sum_{a_i=1}^q \sum_{b_j=1}^q P_{ij}^{\text{dir}}(a_i, b_j) \ln \left(\frac{P_{ij}^{\text{dir}}(a_i, b_j)}{f_i(a_i)f_j(b_j)} \right) \quad (1.33)$$

Esta medida de correlación es estudiada en el capítulo 4 para proteínas repetitivas.

1.4.6. Aproximación por Máxima Pseudo-Verosimilitud

En [Ekeberg et al., 2013, 2014] aplicaron la estimación de los parámetros del Hamiltoniano de Potts (ecuación 1.12) mediante el procedimiento de maximización de la pseudo-verosimilitud (abreviado plmDCA por sus siglas en inglés *pseudo-likelihood maximization of Direct Coupling Analysis*). Este procedimiento se ha demostrado previamente que es consistente, es decir que los estimadores obtenidos llevan a los valores de probabilidad que se desea estimar a su valor verdadero a medida que la cantidad de observaciones tiende a infinito.

Para este procedimiento se analiza, en vez de la distribución de probabilidades de secuencias $P(\vec{s})$, la probabilidad *condicionada*. Esta se define como la probabilidad de que ocurra a_r en la posición r dado que ocurrieron $a_{|r}$ en las demás posiciones y puede expresarse como:

$$P(a_r = a_r^s | a_{|r} = a_{|r}^s) = \frac{\exp \left(h_r(a_r^s) + \sum_{j=1, j \neq r}^L j_{r,i}(a_r^s, b_j^s) \right)}{\sum_{a_r=1}^q \exp \left(h_r(a_r) + \sum_{j=1, j \neq r}^L j_{r,i}(a_r, b_j^s) \right)} \quad (1.34)$$

donde el supraíndice s refiere a una secuencia particular. Esta permite definir la *verosimilitud condicional* como:

$$g_r(h_r, j_r) = -\frac{1}{N} \sum_{s=1}^N \log \left(P(a_r = a_r^s | a_{|r} = a_{|r}^s) \right) \quad (1.35)$$

La suma de la máxima verosimilitud sobre todas las posiciones es denominada *pseudo-verosimilitud* :

$$l_{\text{pseudo}} = \sum_{r=1}^L g = -\frac{1}{N} \sum_{r=1}^L \sum_s \log \left(P(a_r = a_r^s | a_{|r} = a_{|r}^s) \right) \quad (1.36)$$

Al maximizar esta variable se obtiene un conjunto de parámetros para el modelo de Potts

que permiten estimar las probabilidad de ocurrencia de aminoácidos. Este procedimiento permite reducir los tiempos de cómputo al trabajar con las probabilidades conjuntas de las posiciones y no con expresiones de la probabilidad de ocurrencia de cada secuencia $P(\vec{s})$.

Para medir la correlación entre dos posiciones, se utilizan los valores de acoplamiento del modelo $j_{ij}(a_i, b_j)$ según la expresión:

$$||j_{ij}|| = \sqrt{\sum_{a_i}^q \sum_{b_j}^q (j_{ij}(a_i, b_j))^2} \quad (1.37)$$

Cuanto mayor sean los parámetros de acoplamiento (favorables o desfavorables) mayor será este valor, y por lo tanto más correlacionadas se encuentran las posiciones.

1.5. Objetivos y orden de la tesis

Este trabajo estudia las secuencias de familias de proteínas repetitivas con el objeto de encontrar, dentro de su variabilidad, huellas evolutivas que permitan inferir propiedades biológicamente relevantes.

En el capítulo 2 describo los detalles de los métodos utilizados y las herramientas bioinformáticas de uso común que he adoptado para el desarrollo del actual trabajo.

En el capítulo 3 se ajusta el modelo de Potts descrito en la sección 1.4.4 para tres familias repetitivas: ANK, TPR y LRR. Discutiré cómo se pueden utilizar los modelos correspondientes para evaluar la pertenencia de una secuencia a la familia proteica, y más aún, cómo puede ser utilizado como herramienta para el diseño de nuevas proteínas. Así mismo, el campo hallado puede utilizarse para evaluar la estabilidad de plegado de secuencias individuales.

En el capítulo 4, analizaré las medidas de correlación definidas en las ecuaciones 1.3 y 1.33. Analizaré las razones biológicas por las que surgen estas correlaciones, y en particular discutiré la utilización de estas medidas como predictores de estructura terciaria. Discutiré cómo la alta identidad de secuencia entre repeticiones se refleja en los valores de correlación. Propondré una evaluación de las medidas de correlación que permite discernir entre posiciones en contacto y posiciones relacionadas a la simetría de las proteínas repetitivas.

Finalmente, en el capítulo 5 resaltaré los principales resultados y daré las conclusiones

que se desprenden de este trabajo.

Capítulo 2

Métodos

En este capítulo repaso las herramientas y formatos utilizados en bioinformática para el estudio de grandes cantidades de secuencias de proteínas, que adopté para el presente trabajo. Este capítulo no busca ser una recopilación exhaustiva de los mismos, sino que se centra en las versiones utilizadas en este trabajo. Muchas de estas herramientas, presentan variantes.

Luego comento las fuentes dónde obtuve los datos: bases de datos públicas, que ofrecen secuencias ya catalogadas según familia proteica, y datos estructurales relacionados. Las secuencias conforman los datos a analizar, y los demás serán los resultados experimentales que se utilizarán para contrastar las predicciones de los modelos propuestos. Entrando a los objetivos y desarrollos del trabajo, explico qué propiedades consideré relevantes para caracterizar las familias proteicas, y cuales son los cuidados necesarios para evitar sesgos filogenéticos o humanos. Finalmente, especifico cómo realizar los cálculos de las medidas de correlación introducidas en el capítulo anterior, y cuyos resultados serán discutidos exhaustivamente en el desarrollo de los siguientes capítulos.

2.1. Representación de datos

2.1.1. Secuencia de una proteína

Dado que una proteína se encuentra conformada por un polipéptido lineal de aminoácidos, representar su secuencia es muy sencillo. En la tabla 2.1 se indican los nombres de los 20 aminoácidos junto a la abreviación de tres letras y la de una letra, estandarizados. Para representar una secuencia proteica se anota la sucesión de aminoácidos que la constituyen en el código de una letra, desde el N-terminal hacia el C-terminal de la cadena. En este trabajo, utilicé el formato de archivos fasta: una primer línea comienza con el símbolo > seguido por el nombre de la proteína. En las siguientes líneas se encuentra la secuencia de residuos, donde cada línea contiene a lo sumo 50 residuos. Si la proteína es más larga simplemente se utilizan los siguientes renglones.

Nombre	Código de 3 letras	Código de 1 letra
Alanina	Ala	A
Cisteína	Cys	C
Ácido aspártico	Asp	D
Ácido glutámico	Glu	E
Fenilalanina	Phe	F
Glicina	Gly	G
Histidina	His	H
Isoleucina	Iso	I
Lisina	Lys	K
Leucina	Leu	L
Meteonina	Met	M
Asparragina	Asn	N
Prolina	Pro	P
Glutamina	Gln	Q
Arginina	Arg	R
Serina	Ser	S
Treonina	Thr	T
Valina	Val	V
Triptófano	Trp	W
Tirosina	Tyr	Y

Tabla 2.1: Nomenclatura de aminoácidos

2.1.2. Alineamiento múltiple de secuencias

Para estudiar un conjunto de secuencias proteicas relacionadas entre sí, resulta conveniente ordenarlas de forma que se destaquen propiedades o segmentos en común. En un alineamiento múltiple de secuencias (AMS), cada secuencia proteica se escribe en una fila, y son alineadas entre ellas de manera que coinciden y se resaltan regiones similares. Para poder acomodar, se introduce el concepto de sitio vacío (o *gap* en inglés) que permite separar aminoácidos adyacentes, y se representa con un guión. En la figura 2.1, se muestra un ejemplo de cuatro secuencias antes y luego de alinear. Se resalta el patrón característico de esta familia proteica, GXTPLHLA, que se evidencia en las secuencias alineadas. Existen diferentes métodos computacionales para realizar un AMS y es un problema aún abierto; en este trabajo se utilizan los AMS provistos en la base de datos Pfam [Bateman et al., 2004].

Sin alinear		Alineadas	
>IKBA_HUMAN/108-142	NNLQQTPLHLAVITNQPEIAEALLGAGCDPELRDF	>IKBA_HUMAN/108-142	NNLQQTPLHLAVITNQPEIAEALLGAGCD---PELRDF-
>IKBA_HUMAN/143-175	RGNTPLHLACEQGCLASVGVLTQSCCTPHLSILKANY	>IKBA_HUMAN/143-175	--RGNTPLHLACEQGCLASVGVLTQSCCTPHLSILKANY
>IKBA_HUMAN/182-214	NGHTCLHLASIHGYLGIVELLVSLGADVNAQEP	>IKBA_HUMAN/182-214	--NGHTCLHLASIHGYLGIVELLVSLGAD----VNAQEP
>IKBA_HUMAN/215-248	CNGRTALHLAVDLQNPDLVSLLLKCGADVNRVTY	>IKBA_HUMAN/215-248	-CNGRTALHLAVDLQNPDLVSLLLKCGAD-----VNRVTY

Figura 2.1: Ejemplo de alineamiento múltiple de secuencias. A la izquierda, cuatro secuencias sin alinear. A la derecha, las mismas secuencias alineadas. Notar la utilización de sitios vacíos, representados por guiones, para aumentar las coincidencias entre secuencias. En gris, algunas regiones conservadas de la familia proteica que se evidencian en el alineamiento.

2.1.3. Alineamiento de secuencias por parejas

Si el mismo proceso de alineamiento se realiza únicamente entre dos secuencias se denomina *alineamiento por parejas*. Al maximizar la superposición de aminoácidos de únicamente dos secuencias, el alineamiento resulta más preciso que extraer dos secuencias alineadas de un AMS.

En este trabajo utilizaré alineamientos locales por parejas, obtenidos mediante el método Smith-Waterman propuesto en [Smith y Waterman, 1981] que consiste en:

1. Seleccionar matriz de sustitución (asigna costos al cambio de aminoácido en una secuencia) e imponer un valor a agregar espacios vacíos. Usualmente se utilizan las matrices de sustitución BLOSUM (sección 2.1.5) y un valor positivo para penalizar los sitios vacíos.

2. Construir una matriz donde una de las secuencias a alinear representa las filas y la otra, las columnas. En cada posición, la matriz toma el valor de acuerdo a los costos de sustitución de aminoácidos de las secuencias, de acuerdo a la matriz BLOSUM. Se comienza de arriba a la izquierda y avanzando hacia la diagonal inferior derecha.
3. Desde el sitio con mayor puntaje, se avanza en dirección inversa que el paso anterior por los valores más altos hasta alcanzar un cero. Este camino indica los aminoácidos que deben ser alineados.

Iterando este procedimiento se obtiene el mejor alineamiento local por parejas entre dos secuencias.

2.1.4. Secuencia consenso

Dado un conjunto de secuencias relacionadas entre sí, la secuencia consenso es aquella que contiene en cada posición al aminoácido más frecuente. Esta secuencia puede existir o no en la naturaleza, y puede o no plegarse.

La secuencia consenso de un AMS se calcula identificando en cada columna el aminoácido más frecuente. Un ejemplo simple puede verse en la figura 2.2. Si en una posición no hay una clara predominancia de un único aminoácido, sino que hay más de uno muy frecuente, todas las variantes con estos residuos son consideradas secuencias consenso. Si en una posición hay mucha variación, y por lo tanto los aminoácidos tienen frecuencias de ocurrencia bajas y similares entre sí, la posición se representa con una X y se dice que es *hipervariable*.

```

      1      10      20      30
>IKBA_HUMAN/108-142 NNLQQTPLHLAVITNQPEIAEALLGAGCD----PELRDF-
>IKBA_HUMAN/143-175 --RGNTPLHLACEQGCLASVGVLTSCTTPHLHSILKANY
>IKBA_HUMAN/182-214 --NGHTCLHLASIHGYLGIVELLVSLGAD-----VNAQEP
>IKBA_HUMAN/215-248 -CNGRTALHLAVDLQNPDLVSLLLKCGAD-----VNRVTY

>Consenso          --NGXTPLHLAVIXGXLXIVELLXXGAD-----VNRXXY

```

Figura 2.2: Ejemplo de secuencia consenso. Arriba, el AMS de secuencias. Los números en la parte superior indican la columna del alineamiento. El consenso (debajo) se obtiene a partir del aminoácido más frecuente en cada columna del AMS. Cuando no hay ningún aminoácido en esa calidad, se representa con una X, por ejemplo, en la columna 5.

2.1.5. Similitud e identidad de secuencia

La similitud e identidad de secuencia, como sus nombres lo indican, son medidas utilizadas para cuantificar cuán similares son dos proteínas a nivel de estructura primaria.

Para calcularlas, en primer lugar es necesario realizar un alineamiento entre ambas. La identidad entre secuencias (%Id) se define como el porcentaje de posiciones en el alineamiento en que ambas secuencias tienen exactamente el mismo aminoácido. Un ejemplo puede observarse en la figura 2.3.

```
>IKBA_HUMAN/108-142  NNLQQTPHLAVITNQPEIAEALLGAGCD----PELRDF-
>IKBA_HUMAN/143-175  --RGNTPLHLACEQGCLASVGLTQSCCTPHLHSILKANY

%Id = 8/40 100 = 20 %
```

Figura 2.3: Ejemplo de cálculo de porcentaje de identidad entre dos secuencias. Arriba se encuentran ambas secuencias alineadas, y en color se destacan aquellos aminoácidos coincidentes. Dado que son 8 y el largo total de las secuencias es 40, el %Id= 20.

La similitud indica cuál es el costo de mutación entre los aminoácidos de ambas secuencias, posición a posición. Estos costos se encuentran tabulados empíricamente, según la identidad de los aminoácidos en ambas secuencias, en las matrices llamadas BLOSUM (BLOCKs of amino acid SUBstitution Matrix) [Henikoff y Henikoff, 1992]. Un ejemplo de ellas se encuentra en la figura 2.4.

[illegible]

Figura 2.4: Matriz BLOSUM 62, tal como fue reportada en el trabajo original de [Henikoff y Henikoff, 1992] utilizada para evaluar costos en sustitución de aminoácidos en cálculos de similitud en secuencia.

2.1.6. Agrupamiento de secuencias

Algunas veces resulta conveniente agrupar los polipéptidos que son similares a nivel de estructura primaria, y distinguirlas de otras secuencias. La idea más inocente es tomar cada par de secuencias a analizar, realizar un alineamiento de a pares y medir la similitud o identidad entre ellas (*sim*). Esto permite obtener una medida de distancia entre cada par de secuencias: $d = 1 - \text{sim} / \text{máx}(\text{sim})$. Utilizando un método de agrupamiento o *clustering* jerárquico, es posible confeccionar un dendograma de las secuencias utilizadas y, eligiendo un corte, definir grupos.

En este trabajo utilicé el software CD-HIT [Li et al., 2001, 2002], que realiza un agrupamiento jerárquico, similar al proceso descrito, pero optimizando los tiempos de implementación. En primer lugar, en vez de realizar un alineamiento de a pares, identifica segmentos cortos de aminoácidos compartidos por ambas secuencias. De acuerdo a sus desarrolladores, dos secuencias de 100 aminoácidos cada una, con identidad entre ellas del 85 % comparten al menos 70 dipeptidos, 55 tripeptidos y 25 pentapeptidos [Li et al., 2002]. CD-HIT no realiza una matriz de distancia y un dendograma. Ordena las secuencias por su longitud, de mayor a menor, y recorre la lista. Verifica si cada secuencia es redundante con alguna anterior en la lista, es decir si la similitud es mayor al corte determinado. Si lo es, la incorpora en el grupo. Sino, forma un grupo nuevo [Li et al., 2001]. Este software optimiza significativamente los tiempos de cómputo.

En ambos casos, el resultado es un agrupamiento de acuerdo a la similitud en secuencia: cada grupo está conformado por secuencias altamente similares entre sí, y suficientemente disimiles respecto a las secuencias de otros grupos. El corte que indica cuán similares deben ser las secuencias que pertenecen a un mismo grupo es un parámetro que tomará valores entre 70 y 90 %Id, que indicaré en cada caso.

2.1.7. Mapa de contactos

Un mapa de contactos es una representación bidimensional de una estructura tridimensional conocida de una proteína. Para su construcción se mide la distancia euclídea entre cada par de aminoácidos en la estructura tridimensional. Se suele tomar la distancia entre

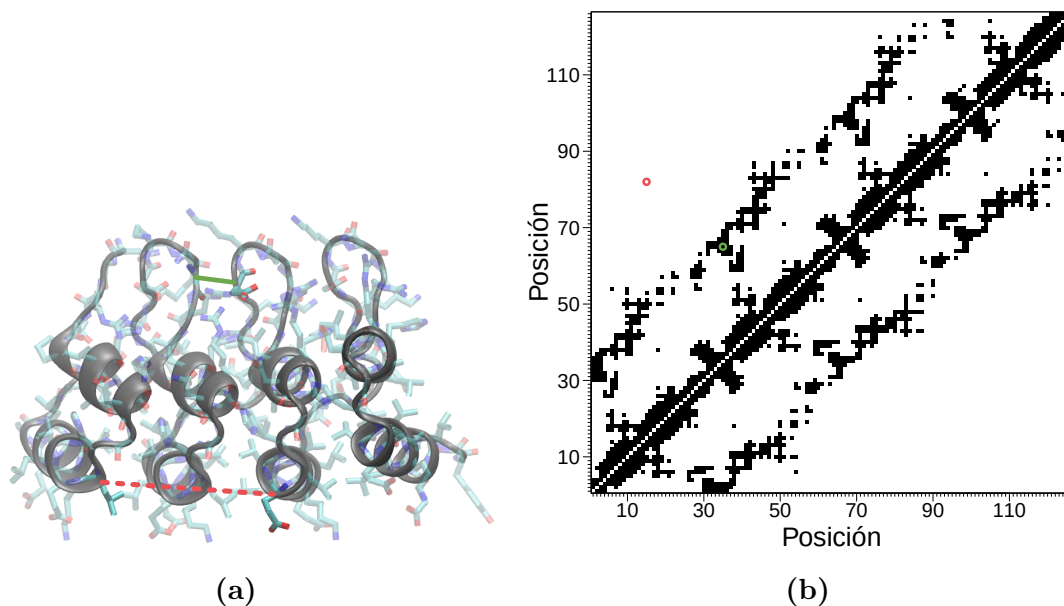


Figura 2.5: (a) Estructura resuelta experimentalmente de la proteína diseñada 4ANK (PDB: 1N0R). En verde represento la distancia entre dos residuos en contacto, y en rojo, la distancia entre dos residuos distantes. (b) Mapa de contactos de la estructura en (a). Los puntos verde y rojo se condicen con los pares de posiciones unidos por las rectas en (a).

los C_α , los C_β , o la mínima distancia entre todos los átomos de cada residuo. Luego se fija un corte entre estas distancias que distingue aquellos pares que están en contactos respecto de aquellos que no. Los valores de corte utilizados usualmente varían entre 5 Å. y 10 Å. Luego se representa en una matriz, donde cada fila y cada columna representan un residuo de la proteína, ordenados desde el N-terminal hacia el C-terminal, y se indica con un 1 si los aminoácidos están en contacto, o un 0 sino. Una representación gráfica puede verse en la figura 2.5.

En este trabajo, las distancias fueron calculadas entre C_β , y el corte para definir un contacto es 10 Å, salvo que se indique lo contrario.

2.2. Datos

2.2.1. Alineamientos múltiples de secuencias

Las secuencias de proteínas se obtuvieron de la base de datos pública Pfam [Bateman et al., 2004], disponible en <http://pfam.xfam.org/>.

Pfam clasifica y agrupa dominios de proteínas definidas a partir de sus secuencias prima-

rias. Para cada dominio, provee un alineamiento curado de secuencias al que llama *semilla*. Este lo utiliza para detectar el dominio en todas las secuencias de proteínas disponibles en otras bases de datos: el alineamiento *completo* incorpora todas las detecciones en la base de datos UniProt, y el alineamiento *NCBI* contiene las secuencias identificadas en la base de datos del Centro Nacional para la Información Biotecnológica de Estados Unidos (NCBI por sus siglas en inglés). Cada conjunto de secuencias pertenecientes a un mismo dominio proteico es denominado *entrada de Pfam*, y actualmente hay 16.295. A su vez, estas se encuentran agrupadas en clanes, relacionadas por similitud en secuencia y estructura, que actualmente suman 559 clanes.

Las proteínas repetitivas se encuentran clasificadas dentro de esta base de datos. Por ejemplo, la familia de proteínas ankyrina tiene un clan asociado *Ank*, que incluye 7 entradas: *Ank*, *Ank_2*, *Ank_3*, *Ank_4*, *Ank_5*, DUF3420 y DUF3447. Cada entrada contiene un alineamiento múltiple de secuencias de repeticiones ankyrinas, que puede coincidir con una o más repeticiones estructurales. En este trabajo, se seleccionó para cada familia de proteínas repetitivas una entrada cuyas secuencias coincidan con una única repetición estructural (o dos, cuando se indique), y se obtuvo el alineamiento NCBI. En la siguiente sección indicaré más detalles de las familias y alineamientos estudiados.

2.2.2. Familias repetitivas

En la tabla 2.2 se detallan las familias de proteínas repetitivas utilizadas en este trabajo, y algunas propiedades de los AMS.

2.2.3. Preprocesamiento de alineamientos múltiples de secuencia

Los alineamientos obtenidos de Pfam contienen secuencias de a una repetición, salvo excepciones. Para este trabajo, es de interés trabajar con secuencias de pares de repeticiones consecutivas. Además, el método de detección de dominios utilizado tiene menor sensibilidad hacia los extremos de las secuencias: muchas no tienen detectados los primeros y últimos residuos. Por lo tanto, los AMS fueron pre-procesados para completar los extremos de las secuencias, y para obtener AMS de pares de repeticiones.

Nombre	Identificador Pfam	Cantidad de secuencias	Cantidad de secuencias de pares de repeticiones	Largo una repetición
YAD (YADA HEAD)	PF05658	8054	9555	14
HECA	PF00353	38527	17200	18
CW	PF01473	18171	11359	19
LRR1	PF00560	96601	27764	24
LRR6	PF13516	142647	40696	24
HEAT	PF02985	5988	513	30
ANK	PF00023	139773	77730	33
TPR	PF00515	78812	18917	33-34 *
PUF	PF00806	7032	4523	33
LDL	PF00057	17417	10727	36
SEL	PF08238	42646	28502	36
WD40	PF00400	189768	88617	37
PENTAPEPTIDE	PF00805	17635	3430	39
ARM	PF00514	13312	7058	42
KELCH	PF01344	22556	12202	47
TSP	PF00090	16743	6256	51
ANNEX	PF00191	6752	4264	66

Tabla 2.2: Familias de proteínas analizadas. La tercer columna se refiere a la cantidad de secuencia contenidas en el AMS crudo, obtenido de Pfam, mientras que la cuarta a la cantidad de secuencias de pares de repeticiones construídas como se indica en el texto, y que se utilizan como entrada en los algoritmos de los siguientes capítulos.

* Debido a diferentes tratamientos de los AMS en los capítulos 3 y 4, la repetición de TPR definia en el primer caso tiene 34 aminoácidos (en concordancia con la definición de la bibliografía) y en el segundo 33.

Completar extremos de secuencias

Para cada secuencia de una repetición, se identificó la cadena polipeptídica entera. Se agregaron hacia los extremos del fragmento los aminoácidos adyacentes en la proteína completa. De esta manera, se disminuyó la cantidad de sitios vacíos en el AMS. Ver figura 2.6.

AMS de pares de repeticiones

Debido a la estructura elongada de las proteínas repetitivas, las interacciones y contactos entre residuos alcanzan únicamente a aminoácidos contenidos en una misma repetición o en repeticiones adyacentes. Esto sugiere que para los análisis aquí presentados es necesario estudiar más allá de una única repetición, y que es suficiente estudiar pares de repeticiones consecutivas. Por este motivo, se contruyó un AMS de pares de repeticiones a partir del AMS de repeticiones únicas obtenido en Pfam con el siguiente procedimiento:

```

>390354872/18-46
-DSTPLH-AA-SSNG-----...---HLEV-V-KDLIG-----QGADINR---
>119182253/479-508
---TPLH-LA-ASSS-----...---APAI-V-IALLT-----KAGADPTILN-
>123481649/799-821
---TPLQ-IA-VKYN-----...---RPQT-V-LYLLN-----KGA-----

>390354872/18-46
DDSTPLH-AA-SSNG-----...---HLEV-V-KDLIG-----QGADINRASN
>119182253/479-508
RCPTPLH-LA-ASSS-----...---APAI-V-IALLT-----KAGADPTILNG
>123481649/799-821
VDKTPLQ-IA-VKYN-----...---RPQT-V-LYLLN-----KGAKAGYIGE

```

Figura 2.6: Ejemplo de secuencias completadas hacia los extremos con los aminoácidos presentes en la proteína natural. Arriba las secuencias obtenidas de Pfam y debajo las completadas, con los aminoácidos agregados marcados en negrita. Los puntos suspensivos indican que estoy omitiendo parte de la secuencia.

1. Tomé las secuencias de las repeticiones de una misma proteína;
2. Las ordené según el número del aminoácido donde comienza la repetición;
3. Seleccioné pares consecutivos, pidiendo que la cantidad de residuos entre ambas repeticiones en la secuencia entera de la proteína sea menor a la mitad del largo de una repetición.

Así se obtiene un nuevo AMS que incluye, por ejemplo en el caso de una proteína con tres repeticiones consecutivas, una secuencia de las repeticiones 1 y 2, y otra secuencia de las repeticiones 2 y 3.

Finalmente, algunas posiciones de los AMS contienen mayormente espacios vacíos producidos durante la alineación a causa de inserciones en algunas pocas proteínas. En los métodos de este trabajo estas posiciones devuelven señales confusas y poco confiables [Contini y Tiana, 2015], a la vez que extienden los tiempos de cómputo. Siendo dispensables, han sido eliminadas aquellas posiciones de los alineamientos que contienen espacios vacíos en más de un 70 % de las secuencias del AMS.

2.2.4. Estructuras tridimensionales

Los datos estructurales de proteínas fueron obtenidos de la base de datos pública Protein Data Bank, disponible en [<http://pdb.org/>].

2.2.5. Estabilidad de plegado

Los datos de estabilidad de plegado fueron obtenidos y curados manualmente de la bibliografía disponible. Los trabajos analizados, fueron para la familia ANK [DeVries et al., 2011, Ferreira et al., 2007, Tang et al., 2003, Street et al., 2005], y para la familia LRR [Dao et al., 2015].

2.3. Caracterización de AMS

2.3.1. Corrección por sobrerrepresentación filogenética

Los AMS pueden tener secuencias sobrerrepresentadas por distintos motivos: sesgo muestral debido a relaciones filogenéticas entre especies, la secuenciación de proteínas de distintos linajes dentro de la misma especie, y debido a un sesgo en la selección de especies a ser secuenciadas. Para que el AMS represente adecuadamente las propiedades evolutivas de la familia proteica en estudio, es necesario corregir estos sesgos.

En este trabajo utilicé diferentes estrategias, basadas en la identidad en secuencia. Algunas consisten en asignar pesos a cada secuencia, que sean menores cuanto mayor representación en el AMS tiene la misma. Otra estrategia consiste en descartar aquellas secuencias que son muy similares entre sí, manteniendo sólo una representativa de cada grupo. En las siguientes subsecciones detallo estos mecanismos.

En cada caso, denomino p_s al peso de la secuencia s -ésima. La suma de los pesos de todas las secuencias será, en la mayoría de los casos, un estimado de la cantidad *efectiva* de secuencias que contiene el AMS luego de la corrección filogenética. La denominaré N_{ef} :

$$N_{ef} = \sum_s^N p_s \quad (2.1)$$

donde N es la cantidad de secuencias reales que tiene un AMS.

Pesos según agrupamiento de secuencias

El primer método consiste en identificar conjuntos de secuencias altamente similares entre ellas mediante un proceso de agrupamiento descrito en la sección 2.1.6. Se asigna a cada secuencia un peso inversamente proporcional al tamaño del grupo en el que se encuentra. Si una secuencia se encuentra en un grupo con otras n secuencias, se le asigna un peso $p_s = 1/n$. Una proteína con secuencia única tiene peso 1, mientras que proteínas redundantes, es decir con muchas otras similares, toman valores más pequeños. La cantidad efectiva de secuencias N_{ef} , es igual a la cantidad de grupos.

Pesos de Henikoff

Los pesos para secuencias en un AMS desarrollados por Henikoff y Henikoff [1994] buscan reducir la redundancia y enfatizar la diversidad utilizando la variabilidad observada en cada posición particular de un AMS particular, en vez de basarse en la similitud entre secuencias enteras.

Para computar estos pesos, en cada columna i del alineamiento se identifican cuántos residuos distintos hay representados (r^i), y cuántas secuencias lo contienen (s_r^i). Luego, una secuencia que contiene ese aminoácido en esa posición tendrá asignado, en esa columna, un peso igual a $p_s = 1/r^i s_r^i$. El peso total de la secuencia es la suma de los pesos asignados en cada columna.

La suma de los pesos de todas las secuencias es igual al largo del alineamiento, como puede mostrarse:

$$\sum_{p=1}^N \sum_{i=1}^L \frac{1}{r^i s_r^i} = \sum_{i=1}^L \sum_{p=1}^N \frac{1}{r^i s_r^i} = \sum_{i=1}^L 1 = L \quad (2.2)$$

donde p suma sobre las secuencias del AMS, i sobre las columnas, y $\sum_{p=1}^N \frac{1}{r^i s_r^i} = 1$ por la definición misma de los pesos. En este caso, no tiene sentido definir una cantidad efectiva de secuencias.

Reducción de cantidad de secuencias

Este método, en vez de asignar pesos reduce la cantidad de secuencias. Similar al primero, se identifican conjuntos de secuencias altamente similares entre ellas utiliza el software CD-HIT [Li et al., 2001, 2002], con un umbral c . Se escoge una secuencia representativa de cada grupo y se descartan las demás. Formalmente, y únicamente para ser congruente con la notación en la siguiente sección, puede indicarse que el peso es $p_s = 1$ si la secuencia fue seleccionada como representativa de un grupo, y $p_s = 0$ sino. En este caso, el número de secuencias efectivas N_{ef} será la cantidad de grupos obtenidos.

Este mecanismo reduce la variabilidad del conjunto de secuencias obtenidos. Por ejemplo, dadas las secuencias AAACKK y AAACKL, con un umbral de $c=5/6$, en el primer caso cada secuencia adquiere un peso $1/2$, mientras que en este método una se descarta (por ejemplo la segunda) y se pierde la información de contener una L en el último sitio.

2.3.2. Caracterización evolutiva de un AMS

Como discutí en la introducción, este trabajo busca caracterizar señales evolutivas en familias proteicas. Se conoce que en las proteínas repetitivas conviven dos mecanismos evolutivos: las mutaciones puntuales y las repeticiones de repeticiones enteras. En esta sección, describo qué medidas pueden realizarse sobre un AMS que se encuentren relacionados con estas señales evolutivas.

En primer lugar, las frecuencias de ocurrencia de cada aminoácido en cada posición, que denotaré $f_i(a_i)$, con i la posición y a_i el aminoácido. El espacio vacío es considerado equivalente a un aminoácido, por lo tanto a_i puede tomar 21 valores. La frecuencia no se calcula como la cantidad total de secuencias que tienen el aminoácido a_i en la posición i -ésima, sino como la suma de los pesos asignados p_s a cada una de estas secuencias. Luego, los valores de frecuencia son normalizados de manera que

$$\sum_{a_i} f_i(a_i) = 1. \quad (2.3)$$

En el caso de una posición completamente conservada $f_i(a_i) = 1$ para el aminoácido consenso,

y cero para los demás. Una posición que tiene igual representados todos los aminoácidos tiene $f_i(a_i) = \frac{1}{21} \forall a_i$. Para un AMS de L columnas habrá $21 \times L$ frecuencias marginales.

En segundo lugar, y con el fin de capturar correlaciones entre posiciones, se utilizan las frecuencias de ocurrencia conjunta de cada par de aminoácidos posibles en todos los pares de posiciones. Estos serán anotados $f_{ij}(a_i, b_j)$ donde i y j son las posiciones involucradas y a_i y b_j los aminoácidos en estas posiciones. Nuevamente, las frecuencias se computan sumando los pesos de las secuencias involucradas, y luego son normalizadas de manera que verifican

$$\sum_{a_i, b_j} f_{ij}(a_i, b_j) = 1. \quad (2.4)$$

Dos aminoácidos distintos no pueden ocurrir simultáneamente en la misma posición. Por lo tanto para una misma posición, si los aminoácidos no coinciden la frecuencia conjunta será nula, y sino equivaldrá a la frecuencia marginal de ocurrencia del aminoácido. Es decir:

$$f_{ii}(a_i, a'_i) = \delta_{a_i, a'_i} f_i(a_i). \quad (2.5)$$

donde δ_{a_i, a'_i} es la delta de Kronecker. Además, observar que:

$$f_{ij}(a_i, b_j) = f_{ji}(b_j, a_i) \quad (2.6)$$

Por lo tanto, para un AMS de L columnas habrá $\frac{21L}{2} (21L + 1)$ frecuencias conjuntas.

Las dos medidas previas responden al mecanismo evolutivo mediante mutaciones puntuales en las secuencias proteicas. La tercer propiedad que tuve en cuenta es propia de las proteínas repetitivas y está inspirada en la duplicación génica. Para cada secuencia en el alineamiento de pares de repeticiones adyacentes, calculé el %Id entre las dos repeticiones de una misma secuencia. En este caso, %Id sólo puede tomar $\frac{L}{2} + 1$ valores, y está acotado entre 0 y 100. Entonces es posible tomar la frecuencia de ocurrencia de cada valor de %Id, que denotaré f_{Id} . Esta es la tercer propiedad que caracteriza a los AMS.

2.4. Modelos cuantitativos

En esta sección se encuentran los procedimientos para los cálculos de medidas de co-evolución descritas en la sección 1.4. Al igual que en la introducción, comenzaré por la Información Mutua. Luego detallaré el procedimiento para obtener los parámetros del modelo de Potts (ecuaciones 1.12 y 1.23). Finalmente, describo el procedimiento para calcular la información directa (ecuación 1.33).

2.4.1. Correlación con Información mutua

La correlación está bien definida para dos tiras de elementos numéricos. Sin embargo, la extensión a otros datos no es inmediata. Con este fin, en el contexto de la teoría de la información se desarrolló la magnitud información mutua (IM), que aquí usamos para medir correlaciones entre posiciones de proteínas. Dadas las posiciones i y j de un AMS, se define la IM entre ellas según:

$$\text{IM}_{ij} = \sum_{a_i, b_j}^{21} f_{ij}(a_i, b_j) \ln \left(\frac{f_{ij}(a_i, b_j)}{f_i(a_i) f_j(b_j)} \right) \quad (2.7)$$

donde la doble suma es realizada sobre los 21 aminoácidos posibles. IM toma el mínimo valor posible, cero, si ambas posiciones son completamente independientes, pues $f_{ij}(a_i, b_j) = f_i(a_i) f_j(b_j)$ y por lo tanto

$$\text{IM}_{ij} = \sum_{a_i, b_j}^{21} f_{ij}(a_i, b_j) \ln(1) = 0 \quad (2.8)$$

En cambio, toma el máximo valor posible (igual a la entropía de la posición definida por $S_i = \sum_{a_i}^{21} f_i(a_i) \ln(f_i(a_i))$) cuando ambas posiciones se encuentran totalmente correlacionadas. Esto puede probarse fácilmente, dado que $f_{ij}(a_i, b_j) = f_i(a_i) \delta_{a_i, b_j} = f_j(b_j) \delta_{b_j, a_i}$. Entonces:

$$\text{IM}_{ij} = \sum_{a_i, b_j}^{21} f_i(a_i) \delta_{a_i, b_j} \ln \left(\frac{f_i(a_i) \delta_{a_i, b_j}}{f_i(a_i) f_j(b_j)} \right) = \sum_{a_i}^{21} f_i(a_i) \ln \left(\frac{1}{f_j(a_i)} \right) \quad (2.9)$$

$$= - \sum_{a_i}^{21} f_i(a_i) \ln(f_i(a_i)) \equiv S_i \quad (2.10)$$

Para poder comparar distintos pares de posiciones resulta conveniente normalizar la IM por la entropía de ambas posiciones:

$$\text{IM}_{ij}^n = \frac{\text{IM}_{ij}}{\sqrt{S_i \cdot S_j}} \quad (2.11)$$

donde el supraíndice n indica que es el valor normalizado. De esta forma, IM^n varía entre 0 (para dos posiciones independientes) y 1 (para dos posiciones completamente correlacionadas).

2.4.2. Modelo de Potts: Ajuste de parámetros

En esta sección se encuentran los métodos desarrollados para estimar los parámetros del modelo de Potts, detallados en las ecuaciones 1.12 y 1.23, y aquí llamadas *energía* o *campo energético*. El objetivo del ajuste es reproducir las variables macroscópicas $f_i(a_i)$, $f_{ij}(a_i, b_j)$ y, en el caso de la ecuación 1.23, f_{id} .

Como concepto general, el método es iterativo. En cada paso, se produce un conjunto de secuencias representativas del campo energético. Con ellas, se estiman las variables macroscópicas ($f_i(a_i)$, $f_{ij}(a_i, b_j)$ y f_{id}), y a partir de los errores en las estimaciones, se actualizan, proporcionalmente a estos errores, los valores de los parámetros.

El código fue escrito en C++ y se encuentra adjunto en el apéndice B.

Algoritmo de Montecarlo

El algoritmo de Montecarlo fue adaptado para generar un conjunto de secuencias compatibles con las propiedades macroscópicas de un campo energético dado. La adaptación realizada produce un conjunto de M secuencias de aminoácidos, cada una de largo L , asociadas a una energía ϵ .

El algoritmo se encuentra esquematizado en la tabla Algoritmo 1. Se inicia con una secuencia de L aminoácidos elegidos al azar $\vec{s}$, que tendrá asociada una energía ϵ_s (línea 1).

Se escoge aleatoriamente una mutación puntual en un sitio. La secuencia que contiene esta mutación $\vec{s}'$ tiene una nueva energía ϵ'_s (línea 4). Si esta nueva secuencia es favorable, es decir $\epsilon'_s - \epsilon_s < 0$, la mutación es aceptada (líneas 5 y 6). Sino, la mutación es aceptada con una probabilidad $e^{\epsilon'_s - \epsilon_s}$. Para implementar esta probabilidad, un número entre 0 y 1 es generado al azar (r). Si $r < e^{\epsilon'_s - \epsilon_s}$, la mutación es aceptada. Sino, es rechazada (líneas 7 a 12).

Cuando la mutación es aceptada, la secuencia original es reemplazada por la variante mutada. Sino, la variante mutada se descarta. En ambos casos, se repite el procedimiento.

Para producir un conjunto de secuencias independientes entre sí, es necesario que no haya memoria de una secuencia a la otra. Por este motivo, sólo se conserva una secuencia cada varias mutaciones probadas (t). En este trabajo, utilizo $t = 1000$, que es suficiente para cumplir la condición de pérdida de memoria entre secuencias sucesivas.

```

1   $\vec{s} = a_1 \dots a_L$  con energía  $\epsilon_s$ ;
2  for  $i$  in  $1:(Mt)$  do
3      Mutación al azar sobre  $\vec{s}$ :  $\vec{s}'$ , con energía  $\epsilon'_s$ ;
4       $\Delta E = \epsilon'_s - \epsilon_s$ ;
5      if  $\Delta E < 0$  then
6           $\vec{s} \leftarrow \vec{s}_{mut}$ ;
7      else
8           $r =$  número al azar entre 0 y 1;
9          if  $r < \exp(-\Delta E)$  then
10              $\vec{s} \leftarrow \vec{s}_{mut}$ ;
11         end
12     end
13     if  $i$  es múltiplo de  $t$  then
14         guarda  $\vec{s}$ 
15     end
16 end

```

Algoritmo 1: Esquema de adaptación del algoritmo de Montecarlo para producir un conjunto de secuencias compatibles con una expresión de energía.

Iteraciones de ajuste

En esta sección describo el proceso iterativo mediante el cual se ajustan los parámetros del modelo de Potts.

Se inician los parámetros arbitrariamente según:

$$h_i(a_i) = -\log(f_i(a_i)) \quad (2.12)$$

$$j_{ij}(a_i, b_j) = 0 \quad (2.13)$$

$$\lambda_{\text{id}} = 0 \quad (2.14)$$

Estos parámetros se utilizan en la expresión del campo (ecuación 1.12 o 1.23, según corresponda), para evaluar la energía de cada secuencia particular. Utilizando la adaptación del algoritmo de Montecarlo descrito, se produce un conjunto de secuencias $\{\vec{s}\}$. Sobre ellas, se miden las frecuencias marginales, conjuntas y de identidad entre vecinos, descriptas en la sección 2.1.2. Para distinguir estas magnitudes, calculadas sobre las secuencias simuladas y no sobre las secuencias naturales, las denomino respectivamente $p_i(a_i)$, $p_{ij}(a_i, b_j)$, y p_{id} . Los parámetros del campo son luego actualizados proporcionalmente a las diferencias entre las frecuencias naturales y las estimadas:

$$h_i^{t+1}(a_i) \leftarrow h_i^t(a_i) - \epsilon_1 (f_i(a_i) - p_i(a_i)) \quad (2.15)$$

$$j_{ij}^{t+1}(a_i, b_j) \leftarrow j_{ij}^t(a_i, b_j) - \epsilon_2 (f_{ij}(a_i, b_j) - p_{ij}(a_i, b_j)) \quad (2.16)$$

$$\lambda_{\text{id}}^{t+1} \leftarrow \lambda_{\text{id}}^t - \epsilon_3 (f_{\text{id}} - p_{\text{id}}) \quad (2.17)$$

Los parámetros ϵ_1 , ϵ_2 y ϵ_3 son los factores de proporcionalidad con los que se renuevan los parámetros. Si son muy chicos, los tiempos de cómputo se extienden demasiado. Si son demasiado grandes, las actualizaciones son demasiado gruesas y el algoritmo no converge. En este trabajo, luego de algunas pruebas preliminares, escogí $\epsilon_1 = 0,05$, $\epsilon_2 = 0,01$ y $\epsilon_3 = 10$.

Este proceso se repite hasta que el máximo error, es decir la máxima diferencia entre $p_i(a_i)$ y $f_i(a_i)$, $p_{ij}(a_i, b_j)$ y $f_{ij}(a_i, b_j)$, y p_{id} y f_{id} sea menor a 0.02.

Regularización L1

El proceso anterior devuelve un gran número de parámetros: $Lq h_i + \binom{L}{2} q^2 j_{ij}$, y $L/2 + 1 \lambda_{id}$. Para reducir la cantidad de parámetros $j_{ij}(a_i, b_j)$ utilizamos una regularización L1 que permite fijar a exactamente cero aquellos parámetros que no aportan significativamente a la estimación de las variables macroscópicas y, por lo tanto, oscilan en torno a cero durante el proceso iterativo.

Para esto, se reemplaza la ecuación 2.16 por:

$$j_{ij}^{t+1} \leftarrow \begin{cases} 0 & \text{si } j_{ij}^t = 0 \text{ y } |\Delta| < \gamma \\ \epsilon_2 [\Delta - \gamma \text{sgn}(\Delta)] & \text{si } j_{ij}^t = 0 \text{ y } |\Delta| > \gamma \\ j_{ij}^t + \epsilon_2 (\Delta - \gamma \text{sgn}(j_{ij}^t)) & \text{si } [j_{ij}^t + \epsilon_2 (\Delta - \gamma \text{sgn}(j_{ij}^t))] \cdot j_{ij}^t > 0 \\ 0 & \text{si } [j_{ij}^t + \epsilon_2 (\Delta - \gamma \text{sgn}(j_{ij}^t))] \cdot j_{ij}^t < 0 \end{cases} \quad (2.18)$$

donde $\Delta = f_{ij}(a_i, b_j) - p_{ij}(a_i, b_j)$, $\text{sgn}(x)$ es la función signo, es decir vale 1 si $x \geq 0$ y -1 si $x < 0$, y γ es el parámetro que indica la incidencia de la regularización. Las primeras dos líneas indican, que si el parámetro j_{ij} es cero, sólo se cambiará su valor si el error en la frecuencia asociada es mayor que γ . La última línea indica que si el valor de j_{ij} está oscilando alrededor de cero, es decir que si un valor de j_{ij} y su actualización tienen signos contrarios, el parámetro se fija a 0. En este trabajo utilicé $\gamma = 0,01$.

2.4.3. Modelo de Potts: Aproximación de campo medio e Información Directa

Los tiempos de cómputo para los métodos descriptos son largos, principalmente por su naturaleza iterativa y la utilización de simulaciones de Montecarlo en cada paso. Para lograr reducir estos tiempos se han desarrollado aproximaciones a la ecuación 1.12. Esto resulta en métodos más accesibles para la comunidad científica general.

Información directa

La información directa (ID) es una medida de correlación, basada en la IM pero derivada del Modelo de Potts. Para computarla, el primer paso es calcular los parámetros de

acoplamiento $j_{ij}(a_i, b_j)$. Para esto, se calcula la matriz de correlación conectada según

$$C_{ij}(a_i, b_j) = f_{ij}^p(a_i, b_j) - f_i^p(a_i) f_j^p(b_j) \quad (2.19)$$

El supraíndice p hace referencia a que las frecuencias computadas en la sección 2.3 fueron corregidas por pseudocuentas para evitar que la matriz sea singular. Este es un mecanismo frecuente para estimar probabilidades en análisis de secuencias, y se obtienen sencillamente mediante las siguientes expresiones:

$$f_i^p(a_i) = \left(\frac{\theta}{21} + f_i(a_i) \right) \frac{1}{\theta + N_{ef}} \quad (2.20)$$

$$f_{ij}^p(a_i, b_j) = \begin{cases} \left(\frac{\theta}{21^2} + f_{ij}(a_i, b_j) \right) \frac{1}{\theta + N_{ef}} & \text{si } i \neq j \\ f_{ij}(a_i, b_j) \frac{1}{\theta + N_{ef}} & \text{si } i = j \text{ y } a_i \neq b_j \\ \left(f_{ij}(a_i, b_j) + \frac{\theta}{21} \right) \frac{1}{\theta + N_{ef}} & \text{si } i = j \text{ y } a_i = b_j \end{cases} \quad (2.21)$$

donde θ es el parámetro de pseudocuentas y lo escogí igual a la cantidad efectiva de secuencias, $\theta = N_{ef}$.

Invirtiendo la matriz 2.19 se obtienen los valores de los acoplamientos bajo la aproximación de campo medio:

$$j_{ij}(a_i, b_j) = \begin{cases} -C_{ij}^{-1}(a_i, b_j) & \text{si } i \neq j \\ 0 & \text{sino} \end{cases} \quad (2.22)$$

donde $C_{ij}^{-1}(a_i, b_j)$ referencia al elemento (i, j, a_i, b_j) de la matriz inversa de 2.19.

A partir de estos parámetros, el siguiente paso es calcular para cada par de posiciones y aminoácidos posibles la probabilidad conjunta directa, definida por:

$$P_{ij}^{dir}(a_i, b_j) = \frac{1}{Z_{ij}} \exp \left(-j_{ij}(a_i, b_j) + \tilde{h}_i(a_i) + \tilde{h}_j(b_j) \right) \quad (2.23)$$

donde es necesario ajustar los parámetros $\tilde{h}_i(a_i)$ y $\tilde{h}_j(b_j)$ de manera que las sumas parciales sobre la probabilidad directa coincidan con las frecuencias marginales medidas en el AMS.

Para esto, utilizo un método iterativo de ajuste que describo a continuación.

1. Inicialmente se fijan todos parámetros igual a cero: $\tilde{h}_i(a_i) = \tilde{h}_j(b_j) = 0$.
2. Se calculan las frecuencias marginales computadas a partir de estos parámetros:
 - Cálculo de sumas parciales:

$$f_i^{est}(a_i) = \frac{1}{Z_{ij}} \sum_{b_j}^{21} \exp \left(-j_{ij}(a_i, b_j) + \tilde{h}_i(a_i) + \tilde{h}_j(b_j) \right) \quad (2.24)$$

$$f_j^{est}(b_j) = \frac{1}{Z_{ij}} \sum_{a_i}^{21} \exp \left(-j_{ij}(a_i, b_j) + \tilde{h}_i(a_i) + \tilde{h}_j(b_j) \right) \quad (2.25)$$

- Normalización dividiendo por la función de partición de dos posiciones:

$$Z_{ij} = \sum_i \sum_j \sum_{a_i} \sum_{b_j} \exp \left(-j_{ij}(a_i, b_j) + \tilde{h}_i(a_i) + \tilde{h}_j(b_j) \right)$$

3. Renovación de los parámetros de acuerdo al error en estimación en frecuencia:

$$\tilde{h}_i(a_i) \leftarrow \tilde{h}_i(a_i) + \epsilon(f_i(a_i) - f_i^{est}(a_i)) \quad (2.26)$$

$$\tilde{h}_j(b_j) \leftarrow \tilde{h}_j(b_j) + \epsilon(f_j(b_j) - f_j^{est}(b_j)) \quad (2.27)$$

4. Repetir pasos 2 y 3 hasta que el máximo error es menor o igual a 10^{-6}

Esto permite obtener un conjunto de parámetros $\tilde{h}_i(a_i), \tilde{h}_j(b_j)$ para cada par de posiciones i y j . Observar que no son únicos: por ejemplo el conjunto de parámetros $\tilde{h}_1(a_1)$ es distinto cuando se los calcula conjuntamente contra $j = 2, j = 3$, etc. Sin embargo, las probabilidades directas sí se encuentran unívocamente definidas. A partir de ellas se computa la ID sencillamente como:

$$ID_{ij} = \sum_{a_i}^{21} \sum_{b_j}^{21} P_{ij}^{dir}(a_i, b_j) \ln \left(\frac{P_{ij}^{dir}(a_i, b_j)}{f_i(a_i) f_j(b_j)} \right) \quad (2.28)$$

2.4.4. Modelo de Potts: aproximación por máxima pseudo-verosimilitud

Para realizar estos cálculos sobre los AMS de pares de repeticiones proteicas, utilicé el código disponible en <http://plmdca.csc.kth.se/> [Ekeberg et al., 2013, 2014].

2.4.5. Calidad de predictor: curva ROC

Una forma usual de medir la capacidad predictiva de un método es analizar la denominada curva ROC (Característica Operativa del Receptor, por su siglas en inglés). El objetivo es predecir una propiedad binaria. En este trabajo lo utilizaré para evaluar la capacidad de predecir contactos de diferentes métodos. Por lo tanto, el método indicará si un par de posiciones estará o no en contacto. De la misma forma, contaré con un conjunto de estructuras proteicas determinadas experimentalmente, y por lo tanto se conoce si los mismos pares de posiciones efectivamente están o no en contacto.

Dadas las predicciones del método y los datos conocidos a predecir (los contactos en estructuras conocidas), se calcula la *tasa de verdaderos positivos* como:

$$TVP = \frac{VP}{P} \quad (2.29)$$

donde VP es la cantidad de pares de posiciones predichas como contacto y que efectivamente lo son, y P es la cantidad de contactos total. La tasa de *falsos positivos* es:

$$FPR = \frac{FP}{N} \quad (2.30)$$

donde FP es la cantidad de pares de posiciones predichas como contacto y que no lo son, y N es la cantidad de pares de posiciones que no son contactos en la estructura conocida. La curva ROC consiste en graficar estos valores, VPR en el eje vertical y FPR en el eje horizontal. Cuanto más distantes y por encima de la identidad se encuentren, mejor es el predictor. Una forma de resumirlo en una única cantidad es medir el área debajo de la curva ROC. En caso de ser un predictor perfecto, el área será igual a uno. Si el predictor es incapaz de predecir, el gráfico igualará a la identidad y el área será 0.5.

Capítulo 3

Modelo de Potts para proteínas repetitivas

En este capítulo discuto la aplicación y el alcance de la variante del modelo de Potts aplicado a secuencias de proteínas repetitivas. Este modelo se encuentra descrito y derivado en detalle en la sección 1.4.4 y en el apéndice A.

Como expliqué en la introducción (capítulo 1), se cree que las proteínas repetitivas han sido sometidas a dos mecanismos evolutivos preponderantes: las mutaciones puntuales y las duplicaciones de repeticiones enteras. El primero se encuentra representado en el modelo por los términos de campos locales $h_i(a_i)$ y acoplamientos entre pares de residuos individuales $j_{ij}(a_i, b_j)$. El segundo, en un único parámetro que aglomera las restricciones de similitud entre residuos λ_{id} . El campo propuesto toma la forma:

$$E(\vec{s}) = - \left[\sum_i^L h_i(a_i) + \sum_i^L \sum_{j>i}^L j_{ij}(a_i, b_j) - \lambda_{\text{id}} \right] \quad (3.1)$$

Las proteínas repetitivas presentan una cuasi-simetría en su estructura primaria y terciaria, lo que permite reducir la dimensionalidad del problema. En vez de estudiar cadenas polipeptídicas completas analizo pares de repeticiones consecutivos y luego extendiendo los resultados a todo el arreglo. Es decir, que las sumas de la expresión 3.1 se realizan sobre las posiciones de las secuencias, con L el largo de dos repeticiones.

Este modelo fue aplicado en tres familias de proteínas repetitivas: ANK, TPR y LRR. Para ajustar los parámetros de una familia particular, se obtuvo el alineamiento de secuencias de repeticiones individuales, y se completaron las secuencias que contenían espacios vacíos en los extremos. A partir de él, se obtuvieron las secuencias de pares consecutivos de repeticiones (ver sección 2.2.3). Se eliminaron aquellas posiciones (columnas) que contienen espacio vacío en más del 80 % de las secuencias. Finalmente se redujo la redundancia de secuencias eliminando las secuencias repetidas con el procedimiento detallado en la sección 2.3.1. Sobre este AMS se calcularon las frecuencias marginales, conjuntas y de identidad entre repeticiones consecutivas, como se describe en la sección 2.3.2, y que son las variables a ajustar por el modelo.

El procedimiento para realizar el ajuste se encuentra detallado en la sección 2.4.2. Consiste en el método del gradiente (en inglés *steepest descent*), e incluye simulaciones de Montecarlo para recrear un conjunto de secuencias sujetas a las restricciones del modelo. Fue realizado en dos etapas. Primero se ajustó el modelo sin considerar las identidades entre repeticiones, es decir $\lambda_{id}=0$. Una vez que estos parámetros reproducen con cierta fidelidad las $f_i(a_i)$ y $f_{ij}(a_i, b_j)$, se incluyen los parámetros λ_{id} , permitiendo que varíen también los parámetros $h_i(a_i)$ y $j_{ij}(a_i, b_j)$ hasta que el campo reproduce además las f_{id} . Se incluyó una regularización L1 (sección 2.4.2) para reducir la cantidad de parámetros de interacción. Finalmente, se estandarizaron los parámetros según el gauge indicado en la sección 1.4.3, de manera que la suma de los parámetros locales es cero, e igual para la suma de los parámetros de acoplamiento.

Una vez ajustado el conjunto de parámetros, el campo permite asignar a cada secuencia un valor de energía. En la sección 3.2 analizo si estas energías son capaces de caracterizar a la familia proteica, más allá de las variables ajustadas. Como discutí en la introducción (sección 1.3), este campo espera capturar señales evolutivas. A su vez, se espera que las variantes proteicas exitosas en tiempos evolutivos mantengan su estabilidad de plegado. En la sección 3.4 estudio si los valores de energía del modelo propuesto tienen relación con la estabilidad del plegado de las proteínas naturales. Por otro lado, el campo ajustado permite, mediante un procedimiento de Montecarlo, generar ensambles de secuencias coincidentes con las frecuencias ajustadas. En la sección 3.3, analizo las similitudes de estas secuencias

con las secuencias naturales, y discuto la posibilidad de utilizar este método para proponer diseños de secuencias. Este método da un paso más que el diseño por consenso, validado en varios trabajos de proteínas repetitivas [Mosavi et al., 2002, Stumpp et al., 2003, Binz et al., 2003, Main et al., 2003b, Urvoas et al., 2010], que considera únicamente frecuencias locales. El campo aquí propuesto incluye correlaciones de segundo orden y un parámetro global para incidir sobre la repetitividad de las proteínas, permitiendo mayor variabilidad en las secuencias de diseño. Finalmente, en la sección 3.5 analizo los parámetros del modelo, su relación con las frecuencias modeladas y con propiedades estructurales de las familias proteicas analizadas.

3.1. Convergencia del ajuste

El procedimiento de ajuste es computacionalmente costoso. Como estimativo, los ajustes aquí presentados, para secuencias entre 42 y 68 residuos tomaron aproximadamente dos semanas cada uno. Este costo radica principalmente en la generación de un conjunto de secuencias representativas de todas las variantes posibles, cada vez que se actualizan los parámetros. En este trabajo, genero ensambles de 80000 secuencias. Para ejemplificar la calidad de las estimaciones con este número de secuencias, en la familia ANK si genero dos ensambles de 80000 secuencias bajo los mismos parámetros, y comparo las frecuencias obtenidas las diferencias son en promedio $\approx 10^{-4}$ y nunca mayores a 0.017. Además, 80000 secuencias supera ampliamente la cantidad de secuencias en los ensambles naturales analizados: 20153 secuencias ANK, 10020 TPR y 18839 LRR, luego de obtener las secuencias de pares de repeticiones y reducir la redundancia al 90 % de identidad.

Para tener un control del desarrollo del ajuste, guardé datos de las aproximaciones a las frecuencias cada 10, 20 o 100 pasos, según la corrida. Esto permite ver cómo evoluciona la exactitud del ajuste. En la figura 3.1, grafico para cada familia el máximo error cometido en la estimación de las frecuencias a medida que avanza el ajuste. Para definir un corte de máximo error permitido en el ajuste de las frecuencias, estimé cuán robustas son estas. Calculé las frecuencias en subconjuntos de la mitad de las secuencias naturales disponibles, y comparé los las frecuencias del conjunto entero. Las diferencias se utilizaron para fijar el

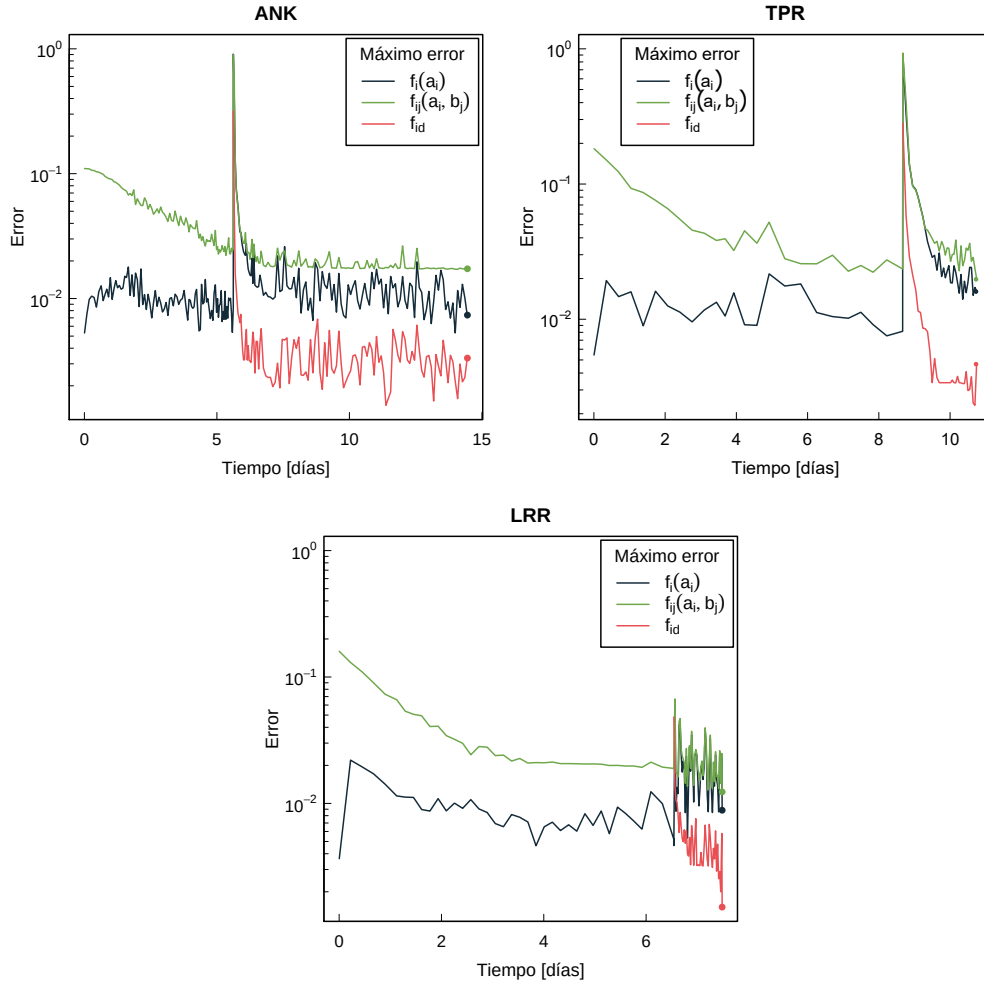


Figura 3.1: Evolución del error a lo largo de las iteraciones.

máximo error permitido: 0.02 en TPR y 0.015 en ANK. En el caso de LRR, tomé distintos valores para cada frecuencia: 0.01 para $f_i(a_i)$, 0.015 para $f_{ij}(a_i, b_j)$ y 0.002 para f_{id} . Dado que la cantidad de variables es muy grande, para poder visualizar las agrupé: muestro el máximo error entre todas las $f_i(a_i)$ en azul, el máximo error entre todas las $f_{ij}(a_i, b_j)$ en verde y el máximo error de las f_{id} en rojo. Estas últimas, naturalmente, se encuentran graficadas únicamente en los pasos donde fueron incluidas en el ajuste. En el eje horizontal represento el tiempo de cómputo hasta la iteración analizada. En el caso de la familia ANK, he guardado los resultados de 1 de cada 20 iteraciones. Para TPR y LRR, 1 cada 100 antes de incluir los parámetros λ_{id} , y 1 cada 20 luego. Los tiempos de cómputo por iteración son distintos para cada familia, ya que los alineamientos tienen distintos largos. La cantidad total de iteraciones fueron 4232 para ANK, 3112 para TPR y 4732 para LRR.

En las tres familias se observa un comportamiento similar. Inicialmente, las frecuencias

marginales (en azul) tienen un error bajo debido a las condiciones iniciales del algoritmo: todos los parámetros nulos excepto los $h_i(a_i) = \log(f_i(a_i))$. En los primeros pasos aumentan su error debido a que los parámetros de interacción cambian su valor. Al incorporar los parámetros λ_{id} aumenta abruptamente el error de las frecuencias marginales, pero en pocos pasos vuelve a reducirse. En las tres familias, incorporar los parámetros λ_{id} implica un aumento en el error de las frecuencias marginales, pero se mantienen por debajo del máximo error permitido en el algoritmo. En otras palabras, las frecuencias marginales parecen ser variables sencillas de ajustar, tanto con un modelo de un cuerpo, como incorporando las correlaciones de a pares e incluso con los parámetros asociados a la repetitividad.

Por su parte, las frecuencias conjuntas son las que contienen mayor error y parecen ser el factor limitante de la convergencia del método. En el caso de ANK, tres frecuencias conjuntas superan en todo momento el valor de corte fijado como máximo error. Estas son $f_{64,65}(-, -)$, $f_{64,66}(-, -)$ y $f_{65,66}(-, -)$, y toman valores un orden de magnitud menor de las ocurrencias naturales. Afortunadamente, esto no genera un error significativo en los ajustes de las frecuencias de aminoácidos reales (recordar que el sitio vacío es una incorporación artificial para alinear secuencias). En TPR y LRR, las $f_{ij}(a_i, b_j)$ alcanzan los valores de corte impuestos. Es llamativo que en la familia LRR la incorporación de los parámetros λ_{id} no genere un salto tan grande en las frecuencias conjuntas (como resulta en las otras familias), aunque sí produce mayores fluctuaciones a lo largo de las iteraciones respecto del ajuste sin los parámetros λ_{id} .

Finalmente, las frecuencias f_{id} tienen un error alto (la diferencia llega a alcanzar un valor de 0.8 cuando el rango de valores posibles es entre 0 y 1) si no se las ajusta explícitamente. Al incorporar los parámetros de ajuste λ_{id} rápidamente convergen y las frecuencias de identidades entre repeticiones adyacentes son reproducidas con una precisión alta.

Para complementar este análisis de la convergencia del método, en la figura 3.2 comparo las frecuencias naturales con los valores obtenidos mediante el proceso de ajuste. En los paneles izquierdos, grafico los valores de las frecuencias $f_i(a_i)$ y $f_{ij}(a_i, b_j)$ en las secuencias naturales (eje horizontal) y los valores estimados (eje vertical). En las tres familias proteicas, las frecuencias más altas se encuentran mejor estimadas (relativo a su valor absoluto), y

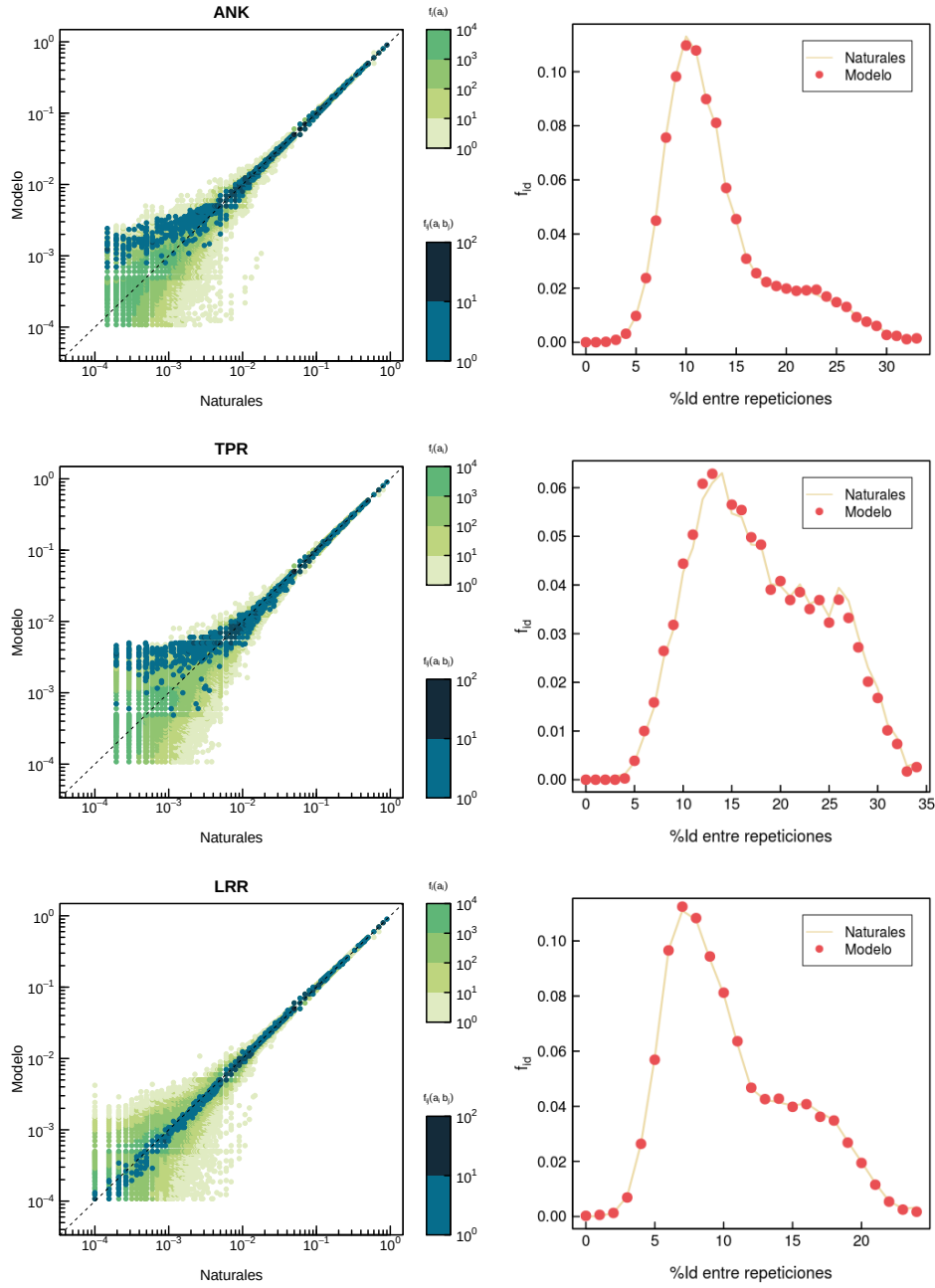


Figura 3.2: Errores en la estimación de frecuencias al finalizar el ajuste de los parámetros del campo (ecuación 3.1). En los paneles a la izquierda, muestro la comparación entre los valores de frecuencias $f_i(a_i)$ (en negro) y $f_{ij}(a_i, b_j)$ (en escala de verdes, en aumento de cantidad de puntos superpuestos). En el eje horizontal se encuentran los valores para las secuencias naturales, y en el vertical los valores ajustados. Ambos ejes se encuentran en escala logarítmica. En los paneles a la derecha, comparo la distribución f_{id} para las secuencias naturales en línea sólida, y los valores ajustados en puntos.

en los valores más bajos se encuentra mayor dispersión. La excepción son las frecuencias marginales de LRR que se encuentran bien estimadas en todo el rango de valores. En los paneles derechos de la figura, comparo las f_{id} de las secuencias naturales (línea sólida) y el modelo (puntos). En los tres casos, las distribuciones se reproducen fielmente.

Una vez ajustados los parámetros, paso a describir el estudio del campo resultante y las aplicaciones posibles.

3.2. Energías de las secuencias naturales

Como mencioné al inicio de este capítulo, una vez obtenidos los parámetros del modelo la ecuación permite asignar valores energéticos a cada secuencia individual mediante la ecuación 3.1.

Una primer pregunta es si las secuencias naturales, a partir de las cuales se obtuvieron los parámetros, minimizan el campo y cuánta variabilidad presentan. Para esto, evalué la energía de todas las secuencias naturales de pares de repeticiones y grafiqué la distribución (figura 3.3, curvas rojas). En la familia ANK (figura 3.3a), se observa una distribución bimodal. La mayoría de las secuencias tienen valores centrados en -100. Un segundo subconjunto de las secuencias naturales toma valores en torno de -150. Como discutiré más adelante, estas secuencias tienen repeticiones altamente similares entre sí. Un comportamiento análogo se observa en las familias TPR (figura 3.3b) y LRR (figura 3.3c).

Como control negativo y para fijar una referencia, generé secuencias de aminoácidos al azar que no representan proteínas, del largo de cada familia particular (66 aminoácidos para ANK, 68 para TPR y 48 para LRR). Las distribuciones de energías de estas secuencias *al azar* se representan con las curvas amarillas en la figura 3.3. En los tres casos, la distribución de energías se encuentra completamente separada de las secuencias naturales, con valores ligeramente por encima de cero. Estos datos permiten concluir que los valores de energía son capaces de distinguir secuencias pertenecientes a la familia proteica relacionada al campo.

¿Cuán influyentes son los parámetros de acoplamiento? Para analizar su relevancia analicé las energías sobre un conjunto control constituido por secuencias cuyas frecuencias marginales de aminoácidos representen las observadas en las secuencias naturales, pero las correlaciones

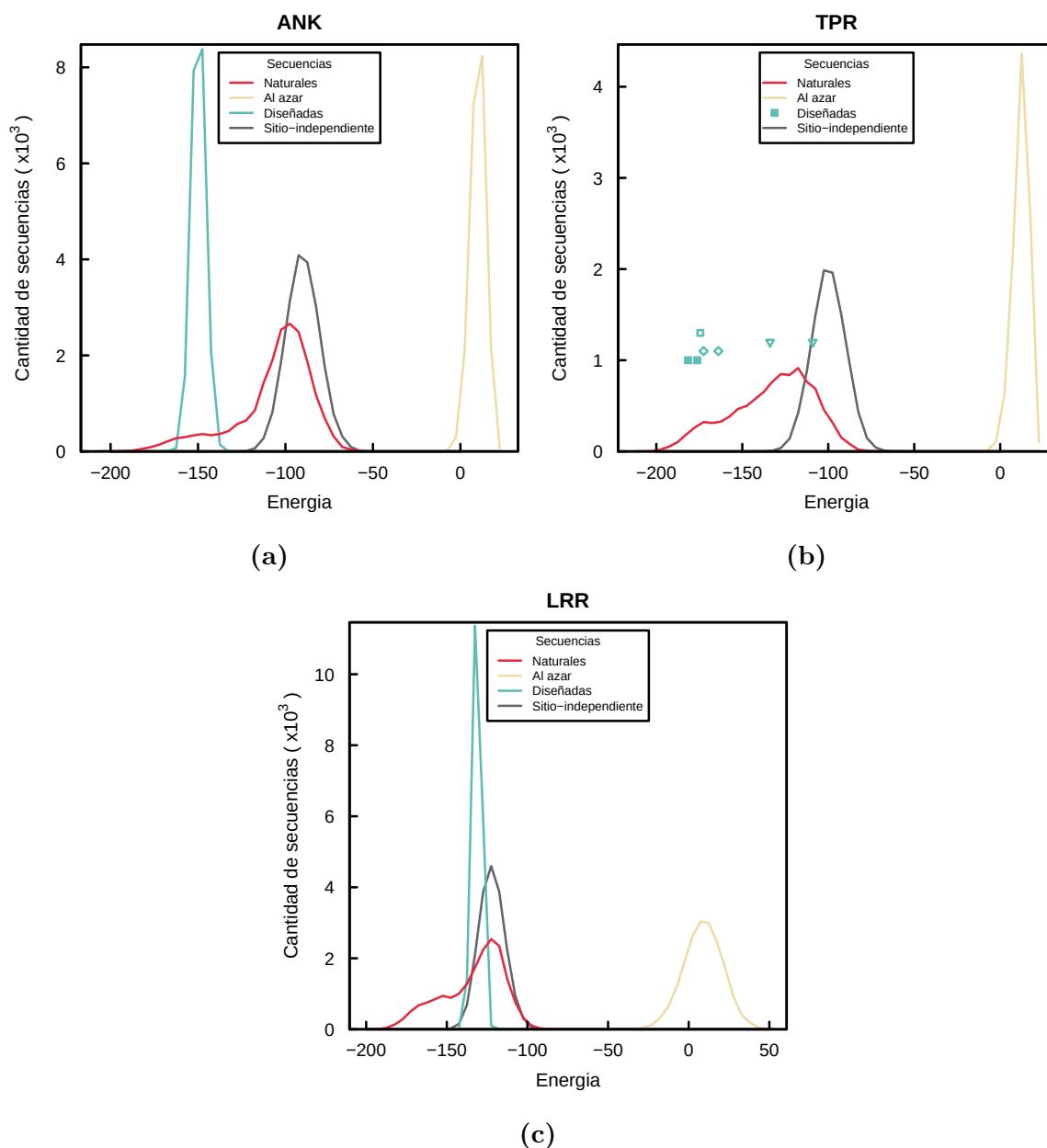


Figura 3.3: Distribución de energías para distintos conjuntos de secuencias. En **rojo** las secuencias naturales utilizadas para realizar el ajuste. En **amarillo**, tiras de aminoácidos al azar, que no representan proteínas. En **celeste** conjuntos de secuencias diseñadas por consenso. En **gris** conjuntos de secuencia del modelo posiciones independientes. De izquierda a derecha, para la familia ANK, TPR y LRR.

entre posiciones se pierdan. Este conjunto de secuencias se obtiene sencillamente mezclando los aminoácidos de cada columna del AMS de manera aleatoria e independientemente de las demás columnas. Me referiré a este conjunto de secuencias como sitio-independiente. Su distribución de energías se encuentra representada en gris en la figura 3.3. En las familias ANK y TPR se observa un corrimiento hacia los valores de energía más altos respecto de las energías de secuencias naturales, mientras que en LRR coincide con el pico más alto (en altura y en valores de energía) de las secuencias naturales. En todos los casos, las secuencias sitio-independientes no alcanzan las energías más bajas de las secuencias naturales, indicando la importancia de las correlaciones en la estabilización de las proteínas.

En el caso de LRR, la coincidencia de la distribución de las secuencias sitio-independiente con el mayor pico de las secuencias naturales puede interpretarse como una baja injerencia de los acoplamientos en un amplio conjunto de secuencias naturales. En la figura 3.4 muestro el valor de la suma de los acoplamientos contra los valores de energía local para cada secuencia natural LRR. Se ve que efectivamente las secuencias con energía mayor a -150 (línea punteada, coincidente con el extremo inferior de la distribución de energías de secuencias sitio independiente en la figura 3.3) tienen acoplamientos muy débiles, centrados en cero. Esto no

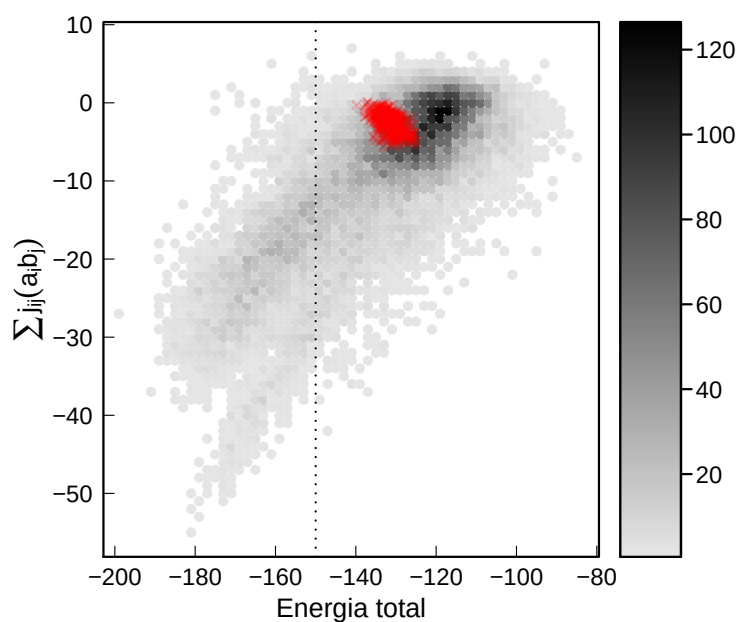


Figura 3.4: Para las secuencias naturales de la familia LRR, contribución de los términos de interacción ($\sum_{ij} j_{ij}(a_i, b_j)$) contra la energía total de la secuencia. La escala de colores indica la cantidad de puntos superpuestos. En rojo, conjunto de secuencias diseñadas por consenso.

ocurre en las otras familias proteicas. Es decir que este amplio conjunto de secuencias está dominado por los campos locales. Las secuencias con menor energía sí presentan campos de acoplamiento significativos respecto de la energía total. Estos resultados podrían estar relacionados con la calidad del ajuste en las frecuencias marginales mostrados en la figura 3.2.

Como un control positivo, analicé secuencias de proteínas diseñadas reportadas en la bibliografía y que han sido sintetizadas experimentalmente. El diseño fue realizado por consenso, es decir tomaron las secuencias disponibles, realizaron un alineamiento de repeticiones individuales y obtuvieron la secuencia consenso, aquella que en cada posición contiene al aminoácido más frecuente. En el caso de la familia ANK [Binz et al., 2003], la secuencia consenso, permutando cíclicamente las posiciones para que coincida con nuestra definición de repetición ANK, está dada por:

XGXTPLHLA**XX**GHLEIVEVLLK(HNY)GADVNA**XXDX**

donde las X rojas indican posiciones de alta variabilidad, y (HNY) indica una posición donde ocurren tanto Histidina, Asparagina y Tirosina. Permitiendo cualquier aminoácido en las posiciones de alta variabilidad, y teniendo en cuenta dos repeticiones consecutivas, esta expresión da lugar a $21^{12} \cdot 3^2 \approx 10^{16}$ secuencias distintas. Generé computacionalmente un subconjunto de estas posibilidades, 80000 secuencias en total. Su distribución de energías es la celeste en la figura 3.3, y coincide con las secuencias naturales de menor energía. Un punto a destacar es que experimentalmente se ha visto que estas proteínas diseñadas por consenso son extremadamente estables [Binz et al., 2003].

Para la familia TPR [Main et al., 2003b, Kajander et al., 2007] utilizaron sólo una secuencia consenso para la repetición: AEAWYNLGNAYYKQGDYDEAIEYYQKALELDPNN. La energía asignada por nuestro campo a un par de repeticiones de esta secuencia, se encuentra representada con un cuadrado celeste en la figura 3.3. Coincide con los valores más bajos de la distribución de energías de las secuencias naturales. Experimentalmente se ha observado que las versiones con 2 y 3 repeticiones de este diseño son altamente estables ($T_m=74$ y 83 °C, respectivamente). Trabajos posteriores usaron la misma secuencia con algunas mutaciones puntuales con el fin de optimizar propiedades de unión a otros ligandos [Cortajarena et al.,

2010, Krachler et al., 2010]. Las energías de estas variantes se encuentran representadas por cuadrados y rombos vacíos en la figura 3.3.

En esta familia también tuvo éxito recientemente otro diseño [Zhu et al., 2016]. Se basaron en una proteína no repetitiva, pero con similitud en secuencia a las repeticiones de TPR. Uniendo copias de la secuencia mediante loops análogos a los del consenso de TPR, obtuvieron una proteína con tres repeticiones. Las energías de estas variantes se encuentran representadas con un triángulo y, sorprendentemente, coinciden con la distribución de energías de secuencias de TPR naturales.

En el caso de la familia LRR [Stump et al., 2003], el consenso diseñado fue para un par de repeticiones consecutivas

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57
RLE1L1L11 2 D L T E A G 3 K D L A S V L R S N P S L R E L 4 L S 4 N K L G D A G V R L L L Q G L L D P G T

donde los números superiores indican la posición y los números en rojo representan variabilidad en composición según las siguientes proporciones:

Código	D	E	G	H	K	L	M	N	Q	R	S	T	V	W	Y
1	11	9			13			25	5	5	15			2	14
2								52			18	26			
3						48	28						24		
4	6		9	14				37			22	12			

En el presente trabajo utilizo otra definición de repetición, de 24 aminoácidos y sin diferencias entre repeticiones consecutivas. Por lo tanto adapté la definición anterior para repeticiones consenso a:

54 55 56 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49
P G T RLE1L1L11 2 D L T E A G 3 K D L A S N P S L R E L 4 L S 4 N K L G D A G V R L L L

donde los números superiores se refieren a la numeración de la expresión anterior, y los números rojos siguen el mismo código. El conjunto de secuencias generadas bajo este esquema adquieren la distribución de energías graficada en celeste en el tercer panel de la figura 3.3. Nuevamente se encuentra superpuesta con la distribución de energías naturales, pero en este caso no con los valores más bajo (asociados previamente a secuencias con acoplamientos fuertes), sino en el extremo inferior de las secuencias sitio-independientes, también solapadas

con mayor pico de la distribución de secuencias naturales. Es decir que en este caso, las secuencias diseñadas por consenso coinciden con las secuencias naturales que aparentan guiarse por los motivos locales de secuencia y no tienen fuertes correlaciones entre sus residuos. En la figura 3.4, en rojo, se observa que los campos locales de estas secuencias diseñadas son cercanos a cero. Coincidentemente, los valores de estabilidad de estas proteínas diseñadas, medidos experimentalmente, no son extremos.

Estos resultados sugieren que la energía acá definida podría estar relacionada a valores de estabilidad de plegado. De ser así, el campo desarrollado provee una herramienta para predicción de estabilidad de proteínas. Además, sugiere que las restricciones evolutivas que mantienen la capacidad de plegado pueden ser capturadas por correlaciones de dos cuerpos. En la sección 3.4 profundizaré este estudio.

Finalmente, utilizando la familia ANK realicé un control para verificar que el modelo no esté sobre entrenado al conjunto de secuencias analizado. Para esto, utilicé el 50 % de las secuencias disponibles para ajustar los parámetros del modelo (conjunto entrenamiento), y el 50 % las aparté como control independiente. En la figura 3.5, se puede ver que las distribuciones de energías para estos dos conjuntos de secuencia coinciden, descartando la posibilidad de sobre entrenamiento del modelo.

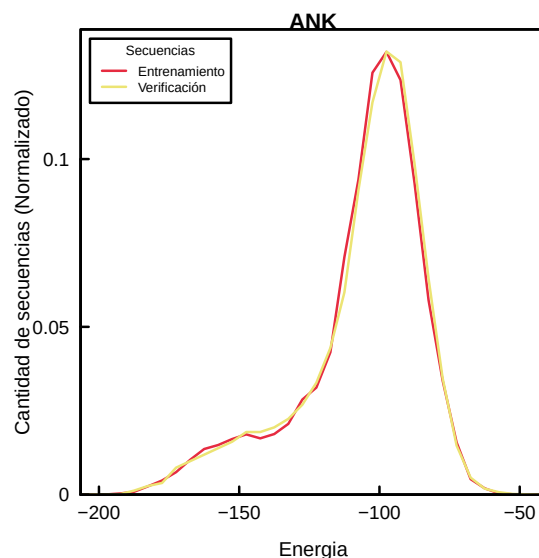


Figura 3.5: Distribución de energías para secuencias naturales de la familia ANK. Los parámetros del modelo se obtuvieron a partir de las medidas de frecuencias del conjunto de secuencias entrenamiento (rojo). Luego, se evaluaron las energías sobre un conjunto independiente y disjunto de secuencias ANK, (distribución amarilla).

3.3. Simulación de nuevas secuencias

El campo de la ecuación 3.1 también puede ser utilizado para generar un ensamble de secuencias nuevas, distintas de las naturales, que cumpla con las restricciones impuestas al modelo, $f_i(a_i)$, $f_{ij}(a_i, b_j)$ y f_{id} . Una vez obtenidos los parámetros, mediante una simulación de Monte Carlo (sección 2.4.2) produje computacionalmente un conjunto de 80000 nuevas secuencias para cada familia, a las que me referiré como *secuencias simuladas*. En esta sección analizo las secuencias simuladas, y las comparo con las secuencias naturales.

3.3.1. Energías

En primer lugar, estudio la distribución de energías de las secuencias simuladas. Análogamente a la sección anterior, en la figura 3.6 muestro sus distribuciones de energías (en azul) en comparación a las energías de las secuencias naturales (en rojo). Las distribuciones se encuentran normalizadas para hacer posible la comparación dado que hay más secuencias simuladas que naturales.

En las tres familias estudiadas las distribuciones de secuencias simuladas reproducen las naturales. En el caso de ANK las distribuciones son prácticamente idénticas, no únicamente el pico más alto de la distribución sino que también se ve reproducido el segmento de menor energía. En la familia TPR se encuentra la mayor desviación entre la distribución de energías de las secuencias simuladas y la distribución de energías de secuencias naturales. Incluye todo

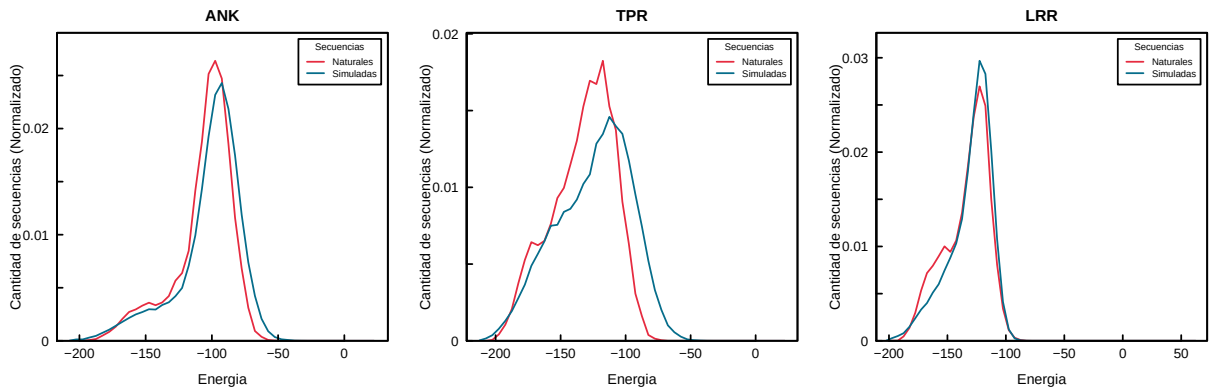


Figura 3.6: Distribución de energías para distintos conjuntos de secuencias. En **rojo** las secuencias naturales utilizadas para realizar el ajuste. En **azul**, secuencias simuladas bajo las restricciones del campo aprendido. De izquierda a derecha, para la familia ANK, TPR y LRR.

el rango de valores de las secuencias naturales, pero se extiende hacia los valores de energía más altos, hacia el extremo más “azaroso”. Finalmente, las LRR se ven reproducidas con mucha precisión, aunque subestimando ligeramente la población de secuencias con energía más baja (como discutí previamente, asociadas con los parámetros de acoplamiento, en esta familia).

En primer lugar, que el campo ajustado permita obtener secuencias con la misma distribución de energías que las secuencias naturales (utilizadas para obtener los parámetros), son una señal de buena convergencia del método. Pero aún más interesante es la posibilidad de usar este campo como método de diseño de nuevas secuencias, candidatas a ser variantes proteicas plegables. Este método da un paso más respecto de los diseños por consenso, que únicamente toman las abundancias locales de aminoácidos. El campo propuesto considera las interacciones de a pares y el sesgo de repetitividad en el parámetro λ_{id} , dando lugar a mayor variabilidad en las secuencias de diseño.

3.3.2. Secuencias

¿Cuán similares son las secuencias simuladas a las naturales? En la figura 3.7 muestro diez secuencias ANK naturales, y 10 secuencias simuladas con el campo asociado a esta familia, elegidas aleatoriamente. Al observarlas no se encuentra ninguna propiedad que permita distinguirlas sencillamente. Para verificar si estos conjuntos son indistinguibles a partir de su

	1								
221091624/299-361	KKDTALHIAA	RAGYLDIVKT	LVSIGANVNI	CSATDSTPLH	LAVINGDKDM	VEYLLEHNAK	VNVYDH		
46120346/264-327	SKFSPLYVAI	RTGQAAMVEL	LAKYGADIDH	LEDTDOTPLI	VAAQSGKDIV	VQSLIDMGAD	LDGKDN		
363895250/148-1550	DGTTPLILAA	RDLDDLIVKD	LIKASCKVNN	VDAQGKSALH	WAAAVNSHEV	TSELCKNGAK	KDMQDD		
391325291/133-197	TGWTPLMWAC	YRGHLRIARF	LLNGAFTNV	HDAYHMSPLI	WASGRGYLEI	VDTLIRHNAK	VNAGDK		
123508647/325-388	EGNLAIHMAA	LSNOVKNTFY	FLDHGFDINE	RNNEGVITPLH	IAVKYHSNDV	ADFLIQRGAN	INSKDL		
123433995/377-442	YGKTALHYAA	ENNDKEIVEV	LISHGANINE	KDKNGVKAIC	IAARHNSGET	AEVLISHGAN	INEKDK		
348668745/2700-2763	TGLTALMSAV	KNGHDECVEA	LLSGGVDPNA	GLPNGITPLH	LAKGYGQPKC	AQLLVEHGAC	LDAKTO		
154418410/439-502	RGETSLHIAF	QNYFKEITFE	FISHGANINE	KNOYGETVLH	IASDKNNKEM	LEYLITHGSD	VNAKDR		
123471510/586-650	KGETPLHNAT	ENNCIEITFE	FLAHGANINE	KNYNDETPLQ	KAESKNNIEV	VEFLLSHVAN	ANQONK		
123479371/129-191	YKKTALHYAA	EYNCKETTKL	LISHGANINE	KDRFGKAALH	YAAEYNCIEI	VELLISHGAN	INEKDG		
s_72675	QGENNLHEAV	HNGRFQIVFF	LLSKDONVNA	ASNNGRITPLH	LSADGKFESV	SELLVNNNDAD	INARDG		
s_42555	DGETSLHIAS	QNGHVQGVKY	LVSQGVLENE	RTNDGEASLH	IASFNGHLQV	VKYLVPKQVE	LN-ATN		
s_19490	-GRITPLHLAA	SEGNLDIVEI	LLEAKADVVDH	LNTGGLTPLH	LAASNMMHTHI	VKLLLAHGAD	AEAMTS		
s_27565	TGRITPLHVAA	RQGYKDLIKL	LINKKAD-EE	KDKKGLTPLH	VASKDGNKDI	VELLISHGAD	INKKDA		
s_6390	DGYNAFGEAI	KCHNDDYQKE	LLMTGAAANV	LCNIOVTAFH	EAAINNSIGV	TKILIEACAG	FTLRDD		
s_50055	DGKTPLMNAA	EKGHLMIVLL	LIAHGAEPYA	KNGSGKHPLI	LATENGSAV	IKLLIEKQTE	VWAGTI		
s_73173	DGWTPLHWAC	QKGHEACVNL	LVELGANVEA	TNEDGDTPLH	WACRYGREKC	VNLLLEQCAN	VEALNP		
s_57326	RGMTSLWVAA	SAGSLESVQY	IILRDVDVKA	NAHNGITATI	LAASMGHTTI	MRYLAKAGAN	ETVR-E		
s_19838	EEHAPLHIAV	FLNAPDCASL	LLSYMANPNK	KNKCGVSAIH	YAVIAQFYDC	VEELISVGVN	EDLEDD		
s_69513	EGRTPLHWAC	RSGNEIIVEM	LLOANADPDI	KNIEGRITPLH	TACDAGSHDV	VKMMLQYGAN	VDVLDD		

Figura 3.7: Diez secuencias naturales de la familia ANK (primeras) y diez secuencias simuladas a partir del modelo 3.1 (últimas).

estructura primaria, recurrí a métodos bioinformáticos de análisis de secuencia: tomé aleatoriamente 100 secuencias de proteínas naturales, y 100 representativas de las simuladas. La elección de 100 secuencias responde a que el proceso es computacionalmente costoso. Realicé un alineamiento de a pares según el método Smith-Waterman (sección 2.1.3) entre las 200 secuencias y calculé la similitud (s) entre cada par según la matrix BLOSUM 62 (figura 2.4). Esta magnitud la traduje a una medida de distancia como $d = 1 - s$ y la utilicé para realizar un dendograma, que muestro en la figura 3.9. Con asteriscos rojos indico las ramas correspondientes a secuencias naturales, y en azul las simuladas. En los tres dendogramas ambas especies se intercalan, y no hay un ordenamiento que permita separar las secuencias naturales de las simuladas. Resulta evidente que no es posible, utilizando únicamente la estructura primaria, distinguir estos dos conjuntos de secuencias.

¿Cómo se relacionan las energías con la identidad entre repeticiones adyacentes? ¿Aque-
llas secuencias constituídas por repeticiones similares toman energías más bajas? En [Jorda et al., 2010] ha sido propuesto que la alta similitud de secuencia entre repeticiones es principalmente evidencia de un evento de duplicación reciente, y no una restricción para favorecer la estabilidad de la estructura proteica. En la figura 3.8 grafico la energía según la expresión 3.1 de cada secuencia contra el porcentaje de identidad entre las dos repeticiones que lo constituyen. Dado que muchas secuencias coinciden, con una escala de colores en gris indico cuantos puntos se superponen. En las familias ANK y TPR se observa claramente una relación lineal entre ambas magnitudes: aquellas secuencias con repeticiones similares tienen energía

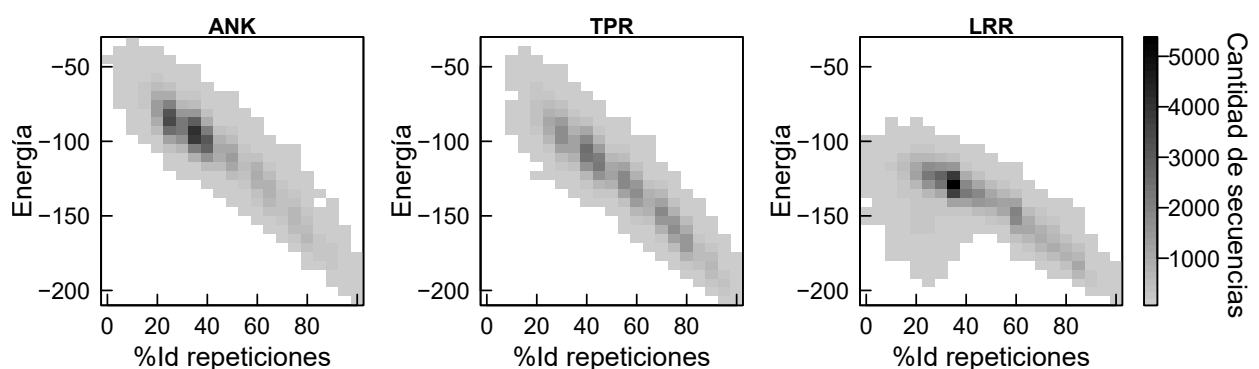


Figura 3.8: Relación entre la energía de las secuencias simuladas y el porcentaje de identidad entre repeticiones adyacentes. De izquierda a derecha, secuencias de la familia ANK, TPR y LRR.

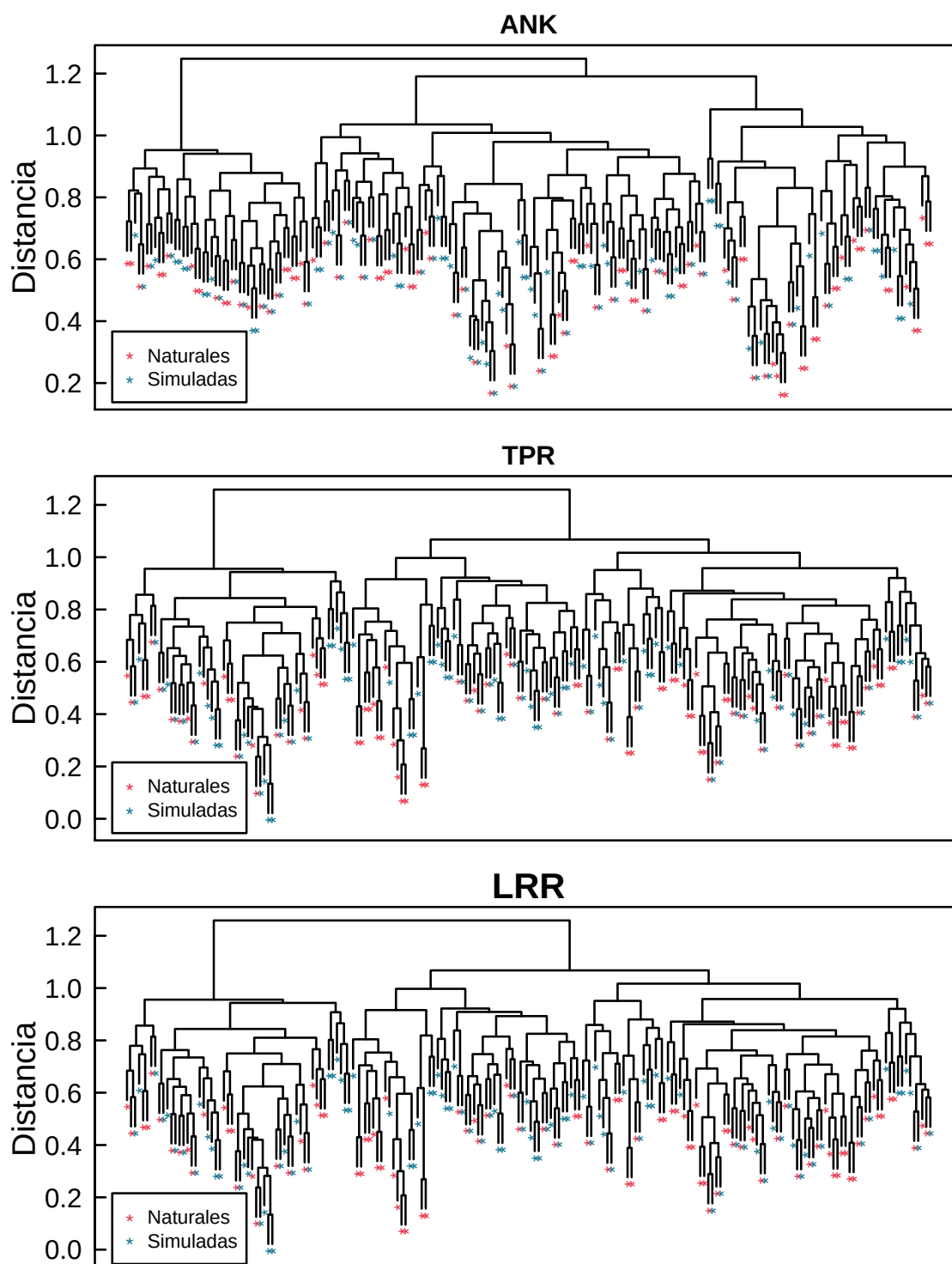


Figura 3.9: Dendrograma a partir de la similitud de secuencia de proteínas naturales (asteriscos rojos) y generadas (asteriscos azules). De arriba hacia abajo, para la familia ANK, TPR y LRR.

más baja. Dado que el campo aquí estudiado se basa en limitaciones en escalas de tiempo evolutivas, y como discutiré más adelante refleja restricciones estructurales y de estabilidad, me atrevo a aventurar que la mayor similitud entre secuencias tiene efectos estabilizantes, e incluso vincularse a la modulación de los caminos de plegado [Rowling et al., 2015, Ferreiro et al., 2008]. En el caso de LRR se observa el mismo comportamiento general, pero hay un subconjunto de secuencias con baja energía y poca similitud entre las repeticiones adyacentes. Estas variantes tienen un alto aporte de los parámetros de acoplamiento en su energía total.

Finalmente, ¿cómo se relacionan las secuencias simuladas con el consenso? Una primera intuición indica que aquellas secuencias más estables, es decir con energía más baja, deberían ser más cercanas al consenso. En la figura 3.10 grafico la energía de cada secuencia contra el %Id respecto de la secuencia consenso de la familia. En ella, se observa que los valores de energía no contienen ningún sesgo respecto de la similitud en secuencia con el consenso. Más aún, en todos los casos el porcentaje de identidad con la secuencia consenso es bajo: menor a 30 % en las familias ANK y TPR, y menor a 40 % en LRR. Estos resultados sugieren que existen diferentes secuencias que pueden ser estables, minimizar el campo obtenido, y no tienen que guardar relación con las secuencia consenso. Esto abre un abanico de posibilidades de secuencias estables que pueden ser exploradas más allá de las proteínas diseñadas por consenso.

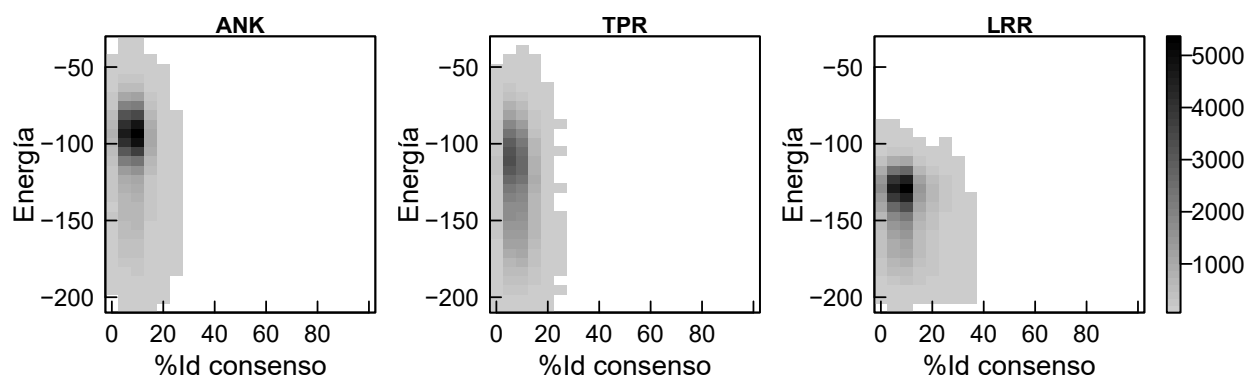


Figura 3.10: Relación entre la energía de las secuencias simuladas y el porcentaje de identidad entre las secuencias y el consenso. De izquierda a derecha, secuencias de la familia ANK, TPR y LRR.

3.4. Relación con estabilidad en el plegado

Como describí anteriormente, las secuencias ANK diseñadas por consenso son extremadamente estables de acuerdo a mediciones experimentales y tienen energías muy bajas en el método computacional. En esta sección profundizo la búsqueda de la relación entre estabilidad de plegado y energía bajo el campo desarrollado en este trabajo.

Experimentalmente es posible sintetizar variantes de proteínas con reemplazos de un único aminoácido respecto de la proteína salvaje. Estos reemplazos, denominados mutaciones puntuales, pueden traer aparejados cambios en las propiedades de plegado. De la bibliografía disponible recolecté los valores de variación en la energía libre de plegado (ΔG) de mutaciones puntuales en distintas proteínas repetitivas. Un valor positivo indica que la mutación estabiliza el estado plegado de la proteína, y un valor negativo que la desestabiliza.

Por otro lado, calculé la variación en la energía según la ecuación 3.1 de las secuencias mutadas respecto de las secuencias salvajes. Si bien esta expresión está definida para pares de repeticiones y no para proteínas completas, es sencillo extender la definición a un arreglo de repeticiones usando la estructura elongada de las proteínas repetitivas, que únicamente presentan interacciones entre repeticiones adyacentes. De nuestra expresión podemos asignar una energía interna de cada repetición según

$$E_{interna}(\vec{s}_{rep}) = - \sum_{i=1}^{L/2} h_i(a_i) - \sum_{i=1}^{L/2} \sum_{j>i}^{L/2} j_{ij}(a_i, b_j) \quad (3.2)$$

y una energía de interacción entre repeticiones adyacentes:

$$E_{interaccion}(\vec{s}_{rep}) = - \sum_{i=1}^{L/2} \sum_{j=L/2+1}^L j_{ij}(a_i, b_j) + \lambda_{id} \quad (3.3)$$

Para asignar una energía a una proteína basta con sumar los términos de energía interna para cada repetición y los términos de interacción entre todos los pares de repeticiones adyacentes.

Las comparaciones entre los valores de diferencia de energía ($E_{mutada} - E_{salvaje}$) y de ΔG

experimentales ($\Delta\Delta G$) entre las secuencias salvajes y las mutadas se encuentran en la figura 3.11. En el panel (a) muestro los resultados para las proteínas ANK I κ B α [DeVries et al., 2011, Ferreira et al., 2007], Notch [Street et al., 2005] y p16 [Tang et al., 2003]. Los valores de ΔG fueron determinados con distintas técnicas experimentales, y cada una de ellas devuelve un valor distinto, que no se superponen dentro del error experimental. Por ese motivo, es difícil obtener una relación precisa frente a los datos teóricos. Al realizar un ajuste lineal obtuve $R^2 \approx 0,61$. Sin embargo, de las 152 mutaciones analizadas, 124 (82 %) son predichas favorables cuando la mutación estabiliza la estructura plegada, y viceversa. La predicción que más desvía es una mutación de Serina a Prolina en la proteína Notch. Dado que la Prolina es un aminoácido disruptivo respecto de la estructura secundaria, decidí descartarla al realizar el ajuste.

¿Cuánto mejora la predicción al utilizar un modelo que incluye términos de acoplamiento respecto de un modelo consenso? En el panel 3.11.b repito el análisis para la familia ANK, utilizando como energía (eje vertical) únicamente información local de cada posición, según

$$E_{consenso} = - \sum_{i=1}^L \log(f_i(a_i)) \quad (3.4)$$

Esta expresión no incluye información respecto de correlaciones entre residuos ni de la repetitividad de las proteínas. La correlación con los datos experimentales es menor ($R^2 \approx 0,55$). Estos resultados llevan a considerar que las susceptibilidades locales a ciertos aminoácidos juega un papel importante, pero no se puede desestimar el rol de las interacciones entre residuos.

En los paneles 3.11.c y 3.11.d reitero los gráficos para la proteína pp32 [Dao et al., 2015], que contiene seis repeticiones LRR. Nuevamente, distintos métodos experimentales reportan diferentes valores de ΔG . Ambos modelos, el que considera únicamente información local (ecuación 3.4) y el que incluye correlaciones de dos cuerpos y repeticiones (ecuación 3.1) tienen un desempeño pobre ($R^2 \approx 0,21$ y $R^2 \approx 0,15$ respectivamente) aunque la mayoría de las mutaciones son predichas y reportadas como desestabilizantes (con el modelo local 28 (78 %) de las mutaciones estudiadas, y con el modelo completo 30 (83 %)). Como ya discutí previamente, un conjunto de las secuencias de esta familia está dominado por los

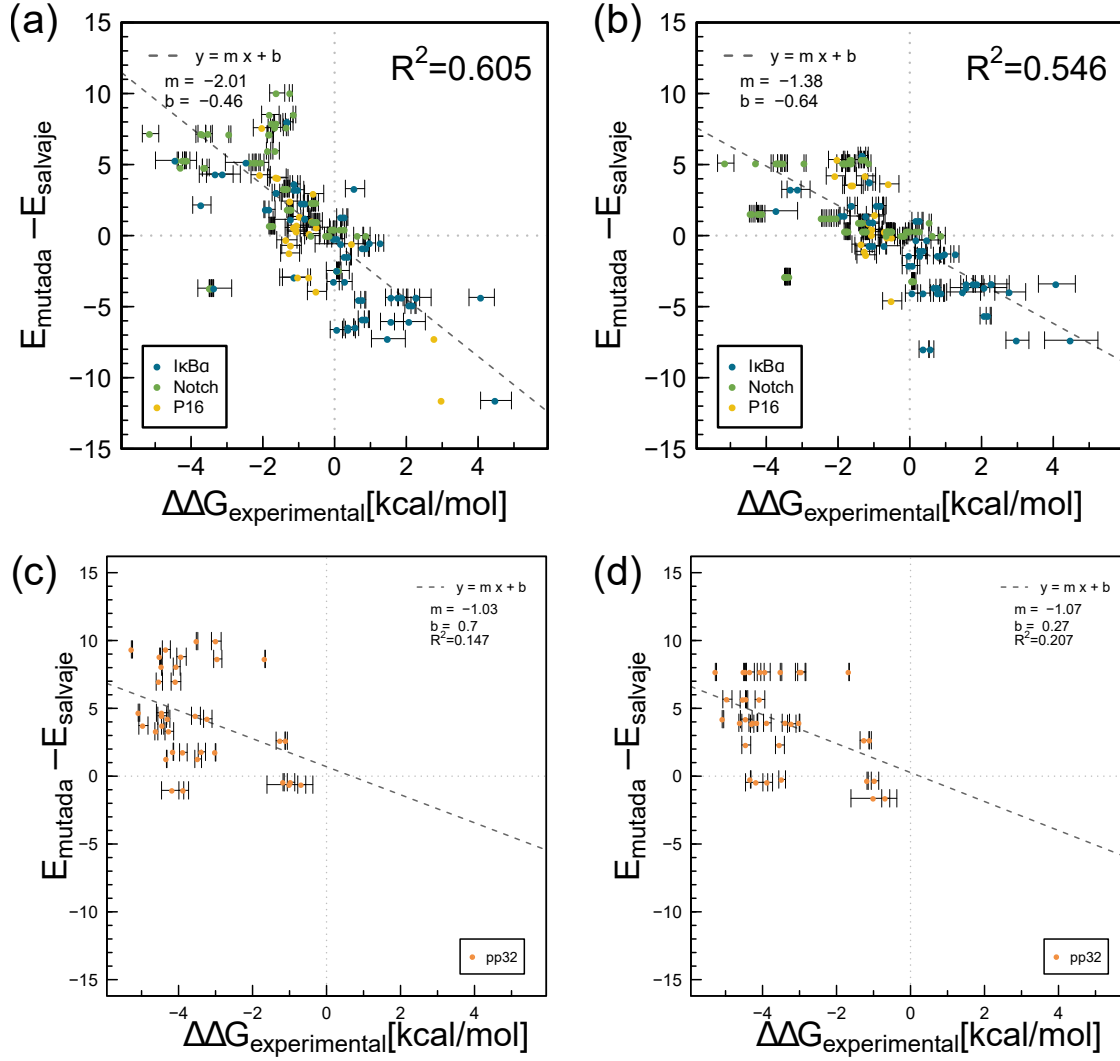


Figura 3.11: Análisis de variantes proteicas con mutaciones puntuales. (a) Cada punto indica una proteína ANK sometida a una, dos o tres mutaciones puntuales. La abscisa es el cambio en la energía libre de plegado ($\Delta\Delta G$) medida experimentalmente, y el eje vertical indica el cambio en la energía definida en este trabajo. Los colores refieren a las tres proteínas estudiadas, $\text{I}\kappa\text{B}\alpha$, Notch y p16. El ajuste lineal fue realizado excluyendo las mutaciones de Serina a Prolina. (b) Idem a (a), utilizando como energía predicha el modelo sitio-independiente. (c) Idem a (a), para la familia LRR. (d) Idem a (b) para la familia LRR.

términos locales, y los acoplamientos o efectos por el término de similitud entre repeticiones no parecen implicar una mejora sustancial.

3.5. Análisis de parámetros del campo

En esta sección analizo los valores de los parámetros de los modelos obtenidos mediante el proceso de ajuste, y sus relaciones con propiedades químicas, físicas y evolutivas de las familias proteicas.

En primer lugar, voy a enfocarme en los campos locales $h_i(a_i)$. En la figura 3.12 se encuentran graficados para cada posición (eje horizontal) y residuo posible (eje vertical). El código de colores indica el valor de $h_i(a_i)$, en tonalidades verdes para los valores negativos (desfavorables) y rojos para los valores positivos (favorables). Es decir que en rojo se encuentran marcadas las combinaciones más probables si obviamos las interacciones. Si, por ejemplo, miramos entre la posición 2 y 7 de la familia ANK, notamos que los valores más altos corresponden a la insignia GXTPLH de esta familia. Dado que estos campos están definidos para dos repeticiones consecutivas, los valores de los parámetros para la primer mitad de las posiciones es igual a los valores de la última mitad. En las tres familias, los $h_i(a_i)$ barren valores positivos y negativos: en ANK, 766 negativos y 620 positivos, en TPR 877 y 551, y en LRR 561 y 492. En el caso de LRR, el rango de valores alcanzados se ve extendido. Curiosamente, el rango de los parámetros de interacción $j_{ij}(a_i, b_j)$ es similar al de las otras dos familias (-0.79 a 1.30 para LRR, -0.73 a 1.39 para TPR y -0.57 a 1.02 para ANK).

Es esperable que los campos locales $h_i(a_i)$ den cuenta de cuán propenso es un sitio a tener determinado aminoácido, y por lo tanto que se encuentren relacionados con las frecuencias marginales. En la figura 3.13 muestro la relación entre estas magnitudes, y se observa que los valores ajustados se encuentran desviados respecto de la condición inicial $h_i(a_i) = \log(f_i(a_i))$. Por lo tanto, los términos de acoplamiento están alterando significativamente las distribuciones de aminoácidos e influyendo en las secuencias simuladas.

El número de parámetros de interacción $j_{ij}(a_i, b_j)$, incluso utilizando la regularización L1 para fijar a cero algunos de ellos, es demasiado grande para poder visualizarlos cómodamente, como puede apreciarse en la tabla 3.1.

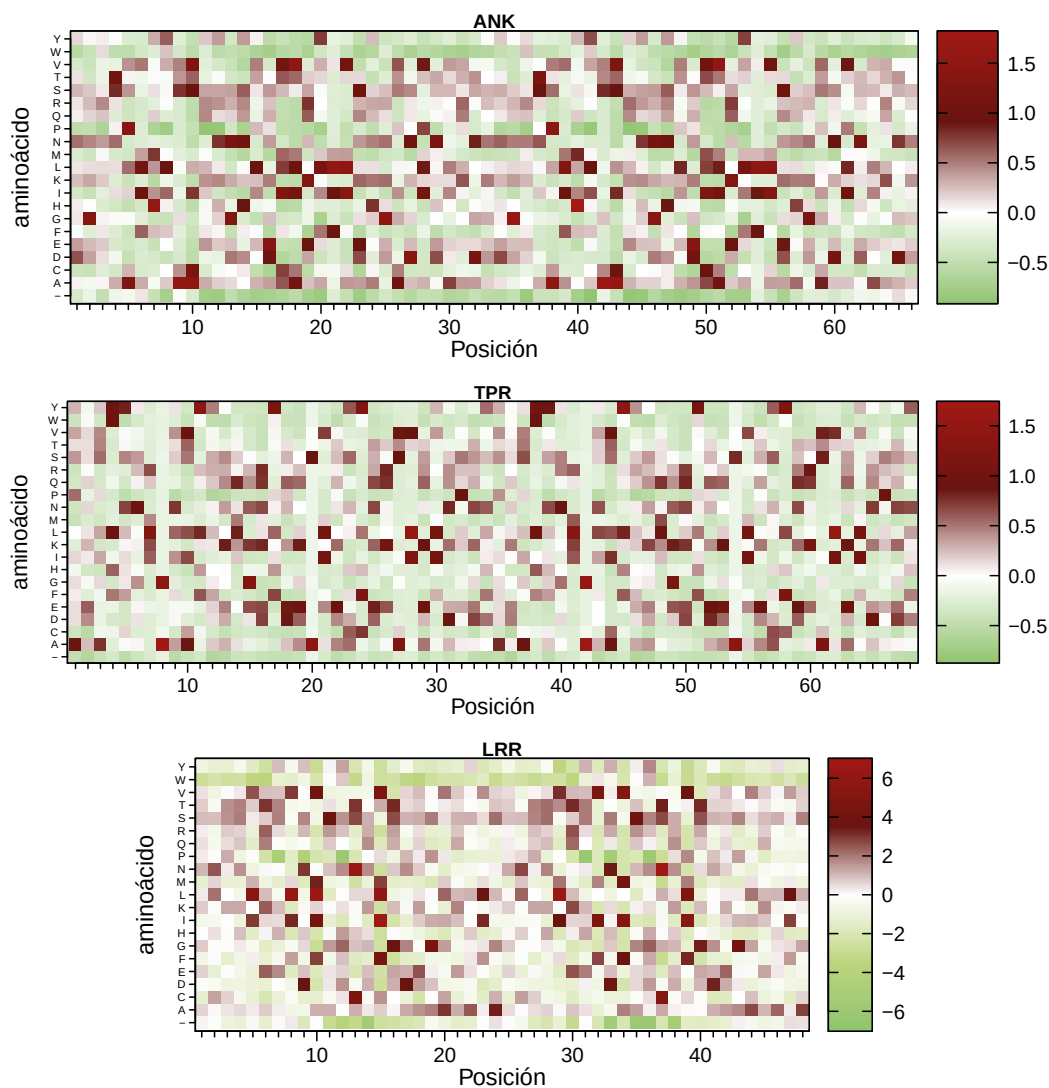


Figura 3.12: Parámetros $h_i(a_i)$ representados matricialmente. El eje horizontal indica la posición (i), el vertical el aminoácido (a_i) y el código de color $h_i(a_i)$. De arriba hacia abajo, para ANK, TPR y LRR.

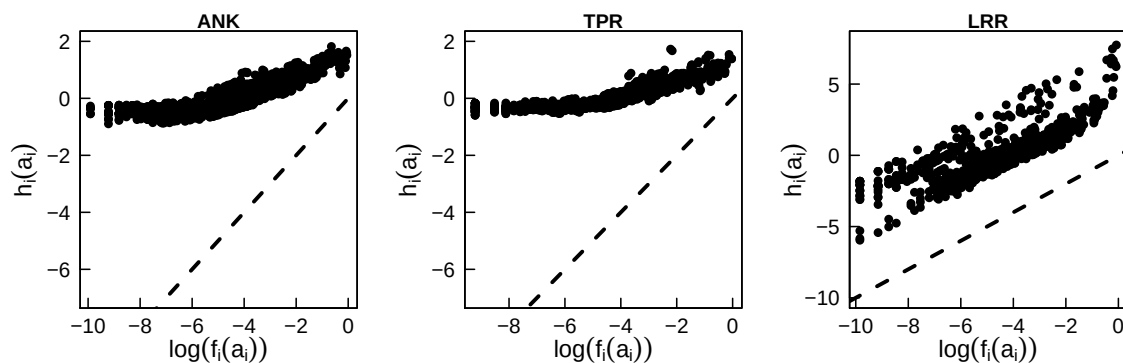


Figura 3.13: Relación de los parámetros $h_i(a_i)$ con las frecuencias marginales $f_i(a_i)$. Las líneas punteadas representan la identidad. De izquierda a derecha, para ANK, TPR y LRR.

Familia	Cantidad de $j_{ij}(a_i, b_j)$	Cantidad de $j_{ij}(a_i, b_j) \neq 0$
ANK	945,945	83,617
TPR	1,004,598	149,664
LRR	497,448	43,296

Tabla 3.1: Cantidad de parámetros de interacción para cada familia, determinados por el ancho del AMS, y la cantidad no nula al incluir el proceso de regularización L1.

Asimismo, se observa que la regularización L1 redujo un orden de magnitud la cantidad de parámetros de acoplamiento. En la figura 3.14 grafico los valores $j_{ij}(a_i, b_j)$ en una corrida sin la regularización L1, y en una con ella, para la familia ANK. Se observa que ningún parámetro sufre grandes cambios. Además, todos aquellos que fueron fijados a cero por la regularización, marcados en rojo, alcanzaban valores muy pequeños sin la regularización.

Como discutí previamente en la introducción, los análisis de coevolución en proteínas buscan detectar posiciones con sustituciones de aminoácidos altamente correlacionadas, asumiendo que estas correlaciones se dan por restricciones al nivel de la estructura terciaria de la proteína. Por lo tanto, es lógico pensar que los parámetros $j_{ij}(a_i, b_j)$, que canalizan los acoplamientos del modelo, estén relacionados con la estructura tridimensional. En particular que aquellos pares que se encuentran en contacto en la estructura tridimensional tengan valores extremos de $j_{ij}(a_i, b_j)$. Para verificar esto, en la figura 3.15 grafico el mapa de contactos

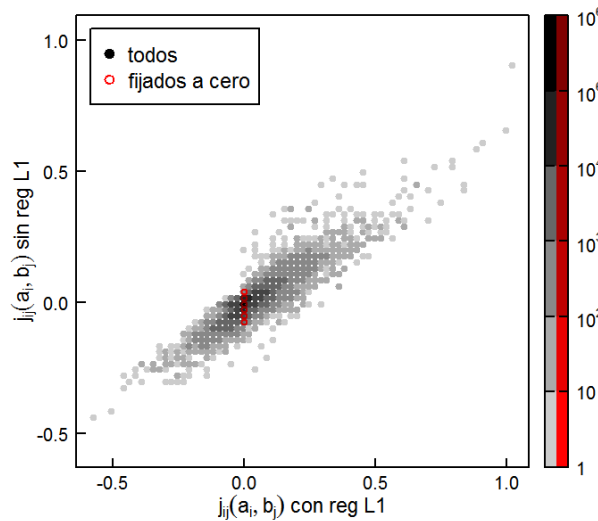


Figura 3.14: Comparación de los valores de los parámetros de acoplamiento $j_{ij}(a_i, b_j)$ obtenidos aplicando la regularización L1 (eje horizontal) y sin aplicarla (eje vertical). En rojo se destacan aquellos puntos que son fijados a exactamente cero en el procedimiento de regularización.

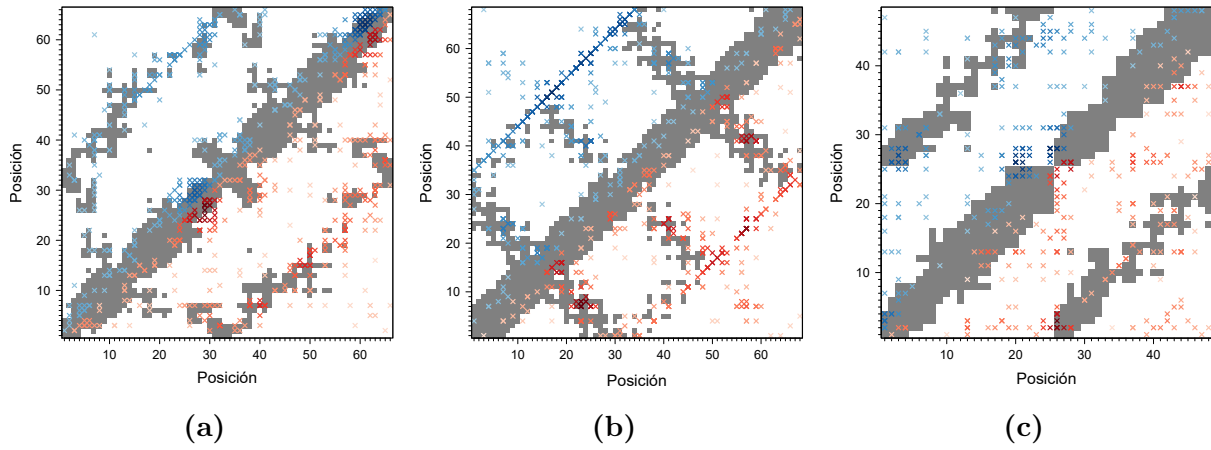


Figura 3.15: Valores extremos de $j_{ij}(a_i, b_j)$ sobre mapas de contacto, para la familia (a) ANK, (b) TPR, y (c) LRR.

(sección 2.1.7) de pares de repeticiones representativos de las familias ANK, TPR y LRR. Sobre ellos, marco en azul las posiciones de los $j_{ij}(a_i, b_j)$ más altos (400 pares para ANK y TPR, y 200 para LRR). Dado que un par de posiciones puede tener más de un $j_{ij}(a_i, b_j)$, con distintas combinaciones de aminoácidos, utilizo una escala de colores donde más oscuro indica más puntos superpuestos. Estos puntos tienen dos comportamientos: la mayoría de ellos coinciden con posiciones en contacto (para ANK 315 (79 %), para TPR 168 (42 %), y para LRR 97 (49 %)). Otro conjunto coincide con las posiciones equivalentes en las repeticiones sucesivas. Por ejemplo en ANK, cuya repetición tiene largo 33, son altos los valores entre la posición 2 y 35, 16 y 49 etc. Es decir que esta simetría inherente de las proteínas repetitivas, y que es capturada de forma global en los parámetros λ_{id} , también se ve reflejada en los valores de correlación, distinguiendo las identidades de los aminoácidos causantes de esta repetitividad. Análogamente, en escala de rojo se encuentran marcados los $j_{ij}(a_i, b_j)$ con valores más bajos, es decir, con mayor exclusión entre posiciones. La mayoría de ellos vuelve a encontrarse en las posiciones en contacto. Esto es coherente con la propuesta de coevolución: dos posiciones en contacto necesitan tener opciones de aminoácidos compatibles favorablemente entre sí, pero también pueden tener combinaciones inviables para la estructura, y esto reflejan los valores bajos (de rechazo) de $j_{ij}(a_i, b_j)$, por ejemplo el par 3 y 32 en la familia ANK.

3.5.1. Incidencia de los parámetros λ

¿Qué efecto tiene la inclusión de los parámetros λ_{id} en el modelo? El objetivo explícito de incluirlos es ajustar el campo final para que reproduzca la frecuencia de identidad entre repeticiones f_{id} . En la figura 3.17 se observa cómo cambian las precisiones de los ajustes cuando no se los incluye. Si se compara con los resultados de la figura 3.2 se observa que la dispersión de las frecuencias de ocurrencia de aminoácidos (marginales y conjuntas) es menor, en especial para aquellos valores de frecuencias bajas. Sin embargo, las distribuciones f_{id} no coinciden entre los valores del modelo y los correspondientes a secuencias naturales. En las tres familias se observa que la f_{id} del modelo es una distribución unimodal (a diferencias de las distribuciones naturales) cuyo centro coincide aproximadamente con el pico más alto de la distribución f_{id} de secuencias naturales.

A nivel de distribuciones energéticas también se observan cambios. En la figura 3.16 muestro las distribuciones de energías de conjuntos de secuencias, al evaluar con el campo sin los parámetros λ_{id} . Las secuencias con menor energía se encuentran subestimadas. Además, las falsas secuencias de aminoácidos, generadas con equiprobabilidad de ocurrencia de cada aminoácido, tienen mayor dispersión. Las secuencias diseñadas, por su parte, tienen un comportamiento análogo al campo anterior.

En conclusión, la incorporación de los parámetros λ_{id} permiten mejorar la descripción de la simetría en secuencia de las proteínas repetitivas sin actuar en detrimento de las

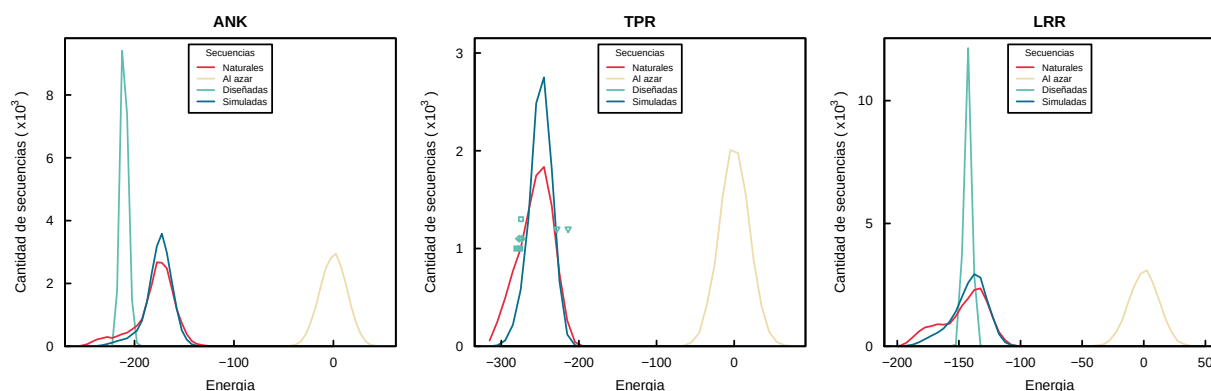


Figura 3.16: Distribución de energías utilizando el campo sin parámetros λ_{id} para distintos conjuntos de secuencias. En **rojo** las secuencias naturales utilizadas para realizar el ajuste. En **amarillo**, tiras de aminoácidos al azar, que no representan proteínas. En **celeste** conjuntos de secuencias diseñadas por consenso. En **azul**, secuencias simuladas bajo las restricciones del campo aprendido. De izquierda a derecha, para la familia ANK, TPR y LRR.

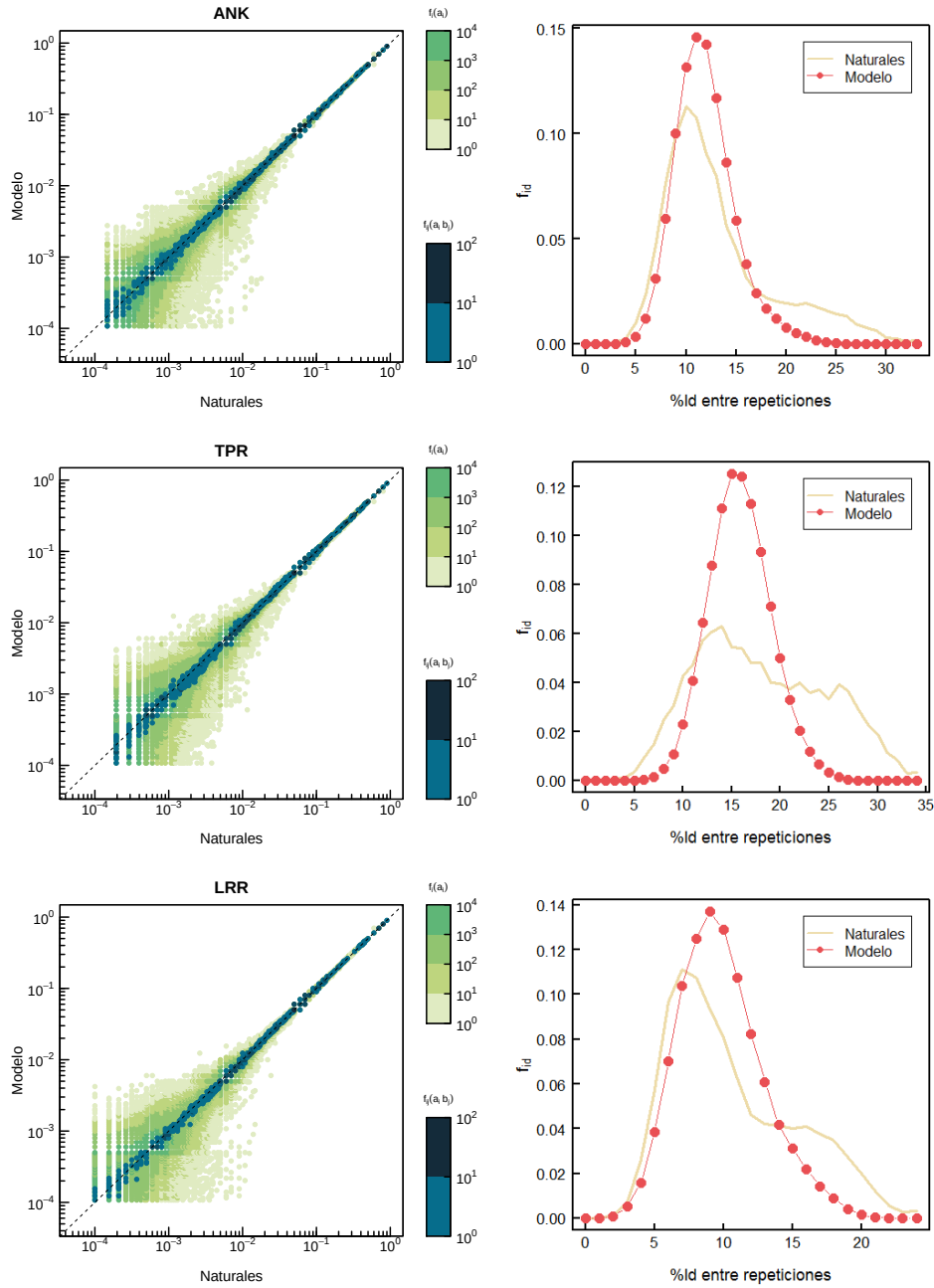


Figura 3.17: Errores en las frecuencias al ajustar con el modelo de Potts que no incluye los parámetros λ_{id} . En los paneles a la izquierda, muestro la comparación entre los valores de frecuencias $f_i(a_i)$ y $f_{ij}(a_i, b_j)$. En el eje horizontal se encuentran los valores para las secuencias naturales, y en el vertical los valores ajustados. Ambos ejes se encuentran en escala logarítmica. En los paneles a la derecha, comparo la distribución f_{id} para las secuencias naturales en línea sólida, y los valores ajustados en puntos.

caracterizaciones energéticas de las secuencias naturales de aminoácidos, y optimizando la simulación de secuencias y su distribución de energías.

3.6. Evaluación de energía en estructuras particulares

Resulta de interés analizar el comportamiento del campo en arreglos completos de repeticiones en proteínas naturales. Dado que muchas veces las repeticiones se encuentran mal detectadas a nivel de secuencia, evalué el campo en algunas proteínas cuyas estructuras son conocidas.

Obtuve en PDB [<http://pdb.org/>] las estructuras asociadas a la familia ANK de Pfam (PF00023). Por inspección, identifiqué las cadenas relevantes y la cantidad de repeticiones en cada una de ellas. Seleccioné aquellas que tienen 4 (3C5R,A; y 4A64,B), 6 (1K3Z,D; 1NFI,E; 3NBN,B; 4DX2,B; y 3JXI,A) o 7 (1K1A,A; 2FO1,E; 3J5P,A; y 4OT9,A) repeticiones. Incorporé además la estructura de la proteína diseñada de cuatro repeticiones 1N0R. Para cada conjunto (según cantidad de repeticiones) realicé un alineamiento estructural utilizando VMD [Humphrey et al., 1996], y luego curé manualmente el alineamiento a nivel de secuencia. Identifiqué las repeticiones y las equivalencias entre las posiciones de este alineamiento con las del AMS utilizado para ajustar los parámetros del modelo.

Calculé los valores de energía a partir de la expresión 3.1 y los parámetros aprendidos según la sección 2.4.2. En la figura 3.18 muestro los valores de energía total, distinguiendo

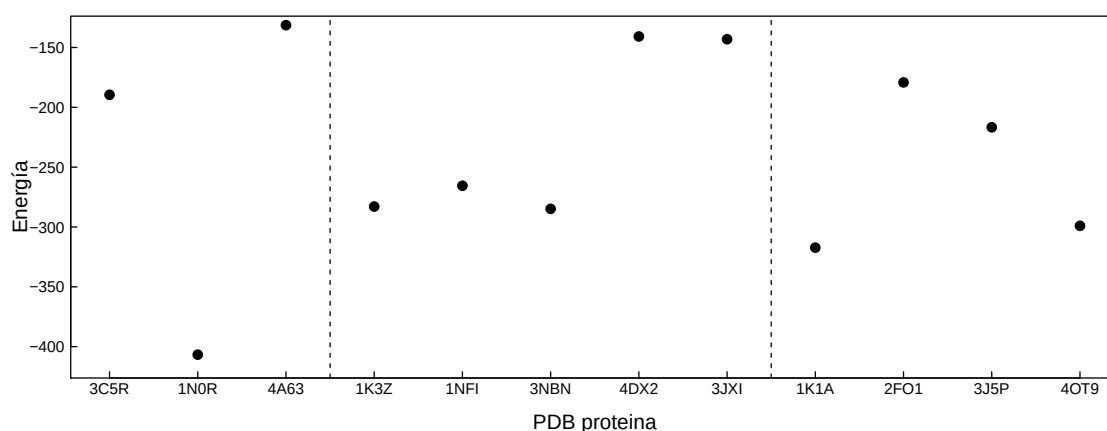


Figura 3.18: Evaluación de energía en proteínas cuya estructura se encuentra resuelta experimentalmente. Las líneas verticales separan las proteínas de 4 (a la izquierda), 6 (centro) y 7 (a la derecha) repeticiones.

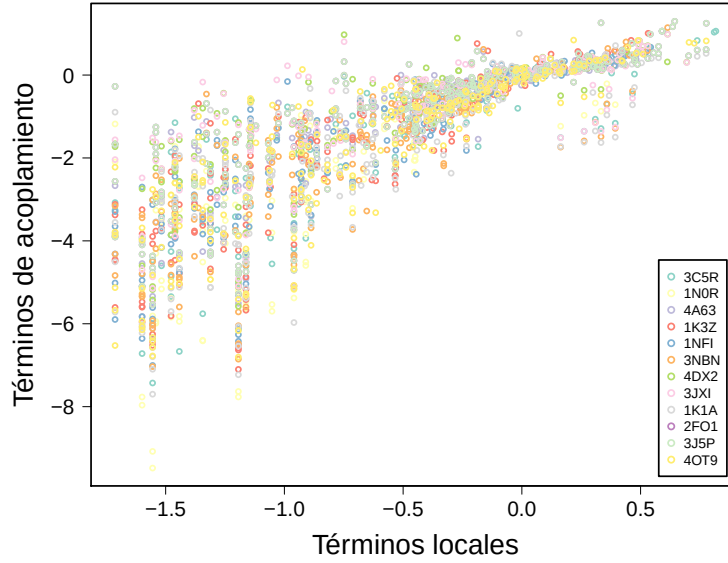


Figura 3.19: Comparación de términos locales y de acoplamiento para secuencias con estructura tridimensional resuelta de la familia ANK.

según la cantidad de repeticiones. Es notable que, a pesar de sumar distinta cantidad de términos, los rangos de valores de energía no dependen de la cantidad de repeticiones. Se destaca por su baja energía la proteína diseñada por consenso y de sólo 4 repeticiones, 1N0R. Entre las proteínas de 6 y 7 repeticiones, se encuentran variantes alrededor de -300, y para todos los largos, secuencias con -150 aproximadamente. Las unidades de energía son arbitrarias, pero en acuerdo con los valores utilizados para pares de repeticiones en las secciones anteriores.

En estos casos particulares, es interesante observar cómo se relacionan los términos locales y los de acoplamiento. En la figura 3.19 grafico $-h_i(a_i)$ en el eje horizontal y $-\sum_j j_{ij}(a_i, b_j)$ en el vertical, es decir que en ambos casos un término menor es más favorable en el campo de secuencias. Para todas las proteínas analizadas, se encuentra una tendencia que vincula ambos términos para cada posición. Aquellas posiciones que tienen un término alto de campo local muestran además los valores más altos en la suma de los acoplamientos. Las desviaciones fuertes que se observan para $-h_i(a_i)$ entre 0 y 0,5 corresponden a sitios con espacios vacíos producidos al alinear. Es decir que en esta familia, aquellos aminoácidos desfavorecidos localmente tienen además un rechazo energético agregado por los acoplamientos. Los aminoácidos más favorecidos a nivel local ($-h_i(a_i) < -0.5$) presentan mayor variabilidad en los acoplamientos asociados; es decir, dependen del entorno.

Capítulo 4

Información directa en proteínas repetitivas

En este capítulo discuto las aproximaciones de campo medio y maximización de la pseudo-verosimilitud, y su aplicación a proteínas repetitivas para el estudio de señales coevolutivas. Estos métodos estiman la correlación entre las sustituciones de aminoácidos que ocurren en cada par de posiciones de una familia proteica. Usualmente se considera que si un par de residuos se encuentran correlacionados, se encuentran vinculados estructural o funcionalmente.

Estos métodos tienen la ventaja de ser sensiblemente más veloces que el aprendizaje de parámetros iterativos (capítulo 3). Como ejemplo, para el AMS de ANK que tiene aproximadamente 75000 secuencias, cada una de largo 66, toman entre 20 y 30 minutos.

En los últimos años han sido utilizados para predecir contactos en proteínas globulares [Morcos et al., 2011, Marks et al., 2011, Ekeberg et al., 2013], y con ellos la estructura nativa [Kamisetty et al., 2013, Marks et al., 2012], e incluso diferentes conformaciones [Morcos et al., 2013]. También han mostrado eficiencia en detectar superficies de interacción entre proteínas interactuantes [Weigt et al., 2009, Dos Santos et al., 2015], y residuos relevantes en la interacción entre proteínas en procesos de señalización en bacterias [Cheng et al., 2014].

Al igual que en el capítulo anterior, estudié pares de repeticiones en vez de arreglos enteros, fundada en que la estructura elongada de esta proteínas produce interacciones únicamente entre residuos dentro de una misma repetición o de repeticiones adyacentes. Esto

permite reducir los tiempos de cómputo y estudiar simultáneamente proteínas de la misma familia proteica pero con distintos largos. Posteriormente es posible extender los resultados a los arreglos enteros en proteínas naturales. Como se detalló en la sección 2.2.3, para cada familia proteica se obtuvo el AMS de repeticiones individuales. A partir de él, se construyeron las secuencias de pares consecutivos de repeticiones. Se eliminaron aquellas posiciones (columnas) que contienen espacio vacío en más del 80 % de las secuencias.

Para reducir la redundancia de secuencias, sin perder la variabilidad, se pesaron las secuencias utilizando el método de Henikoff y Henikoff (ec. 2.2). Este procedimiento asigna a cada secuencia un peso que tiene en cuenta cuán frecuente es cada aminoácido de la secuencia en cada posición del alineamiento. A diferencia de métodos basados en el agrupamiento de secuencias por similitud de secuencia, los pesos pueden adquirir valores continuos. Una vez obtenidos los pesos de cada secuencia, se calcularon las frecuencias $f_i(a_i)$ y $f_{ij}(a_i, b_j)$. Dado que estos métodos necesitan que la matriz de correlación $C_{ij}(a_i, b_j) = f_{ij}(a_i, b_j) - f_i(a_i)f_j(b_j)$ no sea singular, se utilizaron *pseudo*-cuentas y se redefinieron las frecuencias de acuerdo a las ecuaciones 2.20 y 2.21. Luego, utilizando el procedimiento descrito en la sección 2.4.3, calculé la Información Directa entre cada par de posiciones del AMS de cada familia. Además se calcularon las correlaciones entre posiciones utilizando el método de máxima pseudo-verosimilitud, descrito en la sección 1.4.6.

En la sección 4.1 mostraré que la aplicación a proteínas repetitivas se ve interferida por la simetría a nivel de secuencias de estas familias. Luego, en la sección 4.2 mostraré una corrección posible que permite distinguir la señal evolutiva relacionada a esta simetría de aquella proveniente por restricciones estructurales, permitiendo utilizar estos métodos para la predicción de contactos en la estructura nativa. En la sección 4.3 estudiaré cómo cambia la señal al considerar repeticiones más lejanas. Finalmente, estudiaré en la sección 4.4 la cantidad necesaria de secuencias para obtener resultados robustos.

4.1. Información directa en proteínas repetitivas

En el panel A de la figura 4.1 muestro el resultado de calcular la información directa (ID) sobre el AMS de un par de repeticiones ANK. La ID se representa en una matriz, donde

las filas y columnas representan las posiciones en el AMS, y el elemento i, j la ID entre las posiciones i y j . Esta matriz es simétrica, es decir que es equivalente mirar la mitad superior o inferior de la matriz respecto de la diagonal. Para la familia ANK, el largo de las repeticiones de esta familia es 33: los valores entre las filas y columnas 1 a 33, o 34 a 66, corresponden a las interacciones entre residuos pertenecientes a la misma repetición, mientras que los pares entre las columnas 1 a 33 y filas 34 a 66 (o viceversa), corresponden a interacciones entre repeticiones consecutivas.

Dado que la cantidad de secuencias analizadas es finita, algunos valores de ID podrían deberse a correlaciones espúreas en pares de posiciones muy conservadas. Para descartar esta posibilidad, analicé el caso de sitio independiente: generé un alineamiento del mismo tamaño que conserva las frecuencias de ocurrencia local en cada posición pero rompe las correlaciones entre posiciones. Para lograrlo, mezclé independientemente cada columna del AMS. Al calcular la ID de este nuevo alineamiento se obtiene la matriz de la figura 4.1B, donde se observa que los valores son mayoritariamente cercanos a cero, aunque algunos valores son significativos. Finalmente, resto a la matriz de ID cruda la matriz sobre el alineamiento de posiciones independientes, y se obtiene la representación de la figura 4.1C. En ella se observa que hay señal co-evolutiva tanto entre residuos de la misma repetición como entre residuos de repeticiones adyacentes. En la siguiente sección estudiaré con qué propiedades de las proteínas se relacionan los valores más altos de ID.

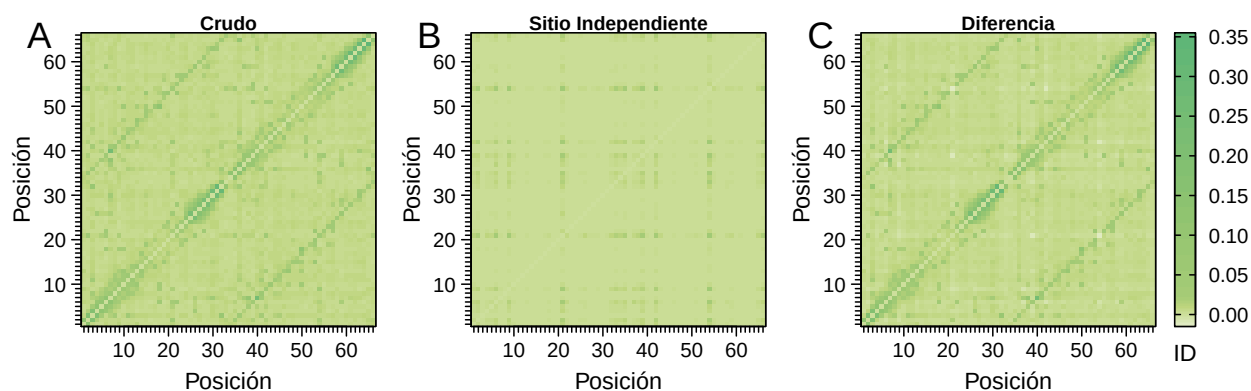


Figura 4.1: Matrices de ID para la familia ANK. Los ejes indican las posiciones del AMS, y el color la información directa entre ellos, según la escala a la derecha. Las matrices son simétricas, es decir que los datos en el triángulo inferior y superior son redundantes. A) ID calculada sobre el AMS de pares de repeticiones ANK. B) ID sobre un AMS de secuencias sitio independiente. C) Diferencia entre las dos matrices anteriores, representando la ID del alineamiento sin sesgos causados por señales locales.

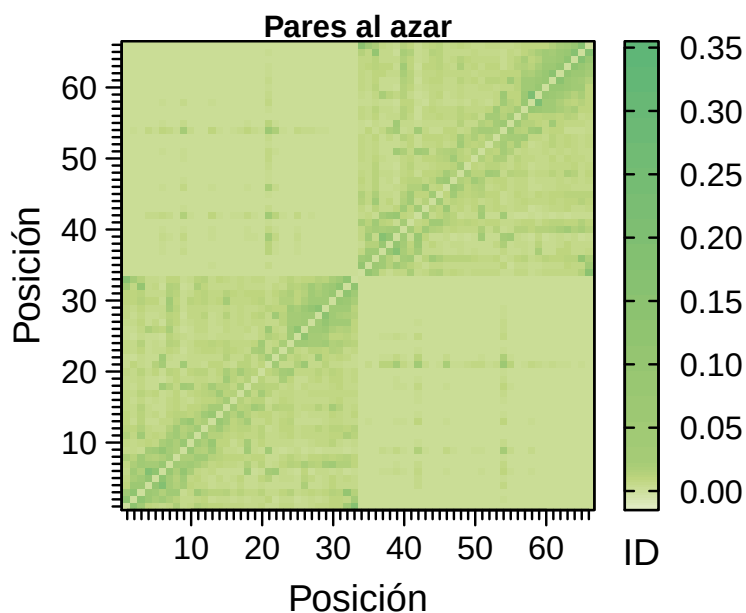


Figura 4.2: Información directa para pares de repeticiones ANK no adyacentes entre sí.

¿Las correlaciones entre repeticiones se dan entre repeticiones adyacentes en proteínas naturales, o son módulos intercambiables, con señal co-evolutiva únicamente por su alta similitud en secuencia? Para analizar esto construí computacionalmente un AMS de pares de repeticiones ANK no consecutivos entre sí. Cada secuencia tiene dos repeticiones naturales ANK, pero cada una pertenece a una proteína distinta. La ID de esta AMS se encuentra en la figura 4.2. En ella queda evidenciado que para este alineamiento las correlaciones entre repeticiones se desvanecen, es decir que son las repeticiones adyacentes las que se encuentran fuertemente relacionadas entre sí. Esta observación indica que las repeticiones no son módulos independientes, y coincide con el proceso de plegado de las proteínas repetitivas en el cual las repeticiones no se pliegan independientemente de sus vecinos sino que lo hacen en forma cooperativa [Main et al., 2003a]. Además coincide con la dificultad de definir unívocamente las repeticiones [Parra et al., 2013]. Por su parte, los valores dentro de las repeticiones no cambian significativamente dado que estas posiciones no fueron mezcladas relativamente entre ellas.

4.1.1. Posiciones coevolutivas

Me referiré a las posiciones en las secuencias con valores altos de ID como correlacionadas o coevolutivas. Dado que la ID, al igual que otras medidas de coevolución, suele relacionarse con propiedades estructurales de las proteínas, es usual representar los puntos de ID más altos sobre un mapa de contactos (sección 2.4.2), para verificar si coinciden o no con posiciones interactuantes en la estructura nativa.

Sin embargo, no hay ningún criterio fundamental que permita determinar la cantidad de pares de posiciones significativas. En la figura 4.3A se encuentran los valores de ID ordenados según su rango, es decir de menor a mayor. A su lado, en el panel 4.3B, se encuentra el histograma de los valores de ID. No hay ningún corte claro que permita diferenciar los valores *altos o significativos* de los *bajos*. Por lo tanto, es usual elegir arbitrariamente un corte. En esta figura muestro dos posibles candidatos: 0.01 (verde claro) y 0.02 (verde oscuro). En el panel 4.3C indico los valores de predicciones según estos cortes, respetando el mismo código de colores, sobre un mapa de contactos de un par de repeticiones ANK (PDB: 1N0R) en escala blanca (no contacto) y gris (contacto).

Como sugerí en la sección anterior, los pares de posiciones correlacionados involucran tanto residuos en repeticiones adyacentes como posiciones dentro de una misma repetición, como reflejo de su estructura elongada, donde repeticiones consecutivas se encuentran en contacto. Al analizar la figura 4.3C se observa que muchos pares de posiciones altamente correlacionados se dan entre posiciones cercanas en la cadena polipeptídica, menos de 5 aminoácidos (ceranos a la diagonal de la figura). Residuos tan cercanos suelen estar involucrados en la formación de la estructura secundaria. Entre los pares de posiciones con alto nivel de coevolución que se encuentran más alejados en la cadena polipeptídica, podemos distinguir varios pares en contacto, y un conjunto que se encuentran en posiciones equivalentes de repeticiones adyacentes, es decir posición 1 y 34 (que es la primer posición de la segunda repetición), la posición 2 y 35 (segunda de la segunda repetición), etc. Es decir que las medidas de coevolución en la familia repetitiva ANK captan restricciones impuestas por la estructura tridimensional, y correlaciones asociadas a su inherente simetría traslacional (asociada a procesos de duplicación).

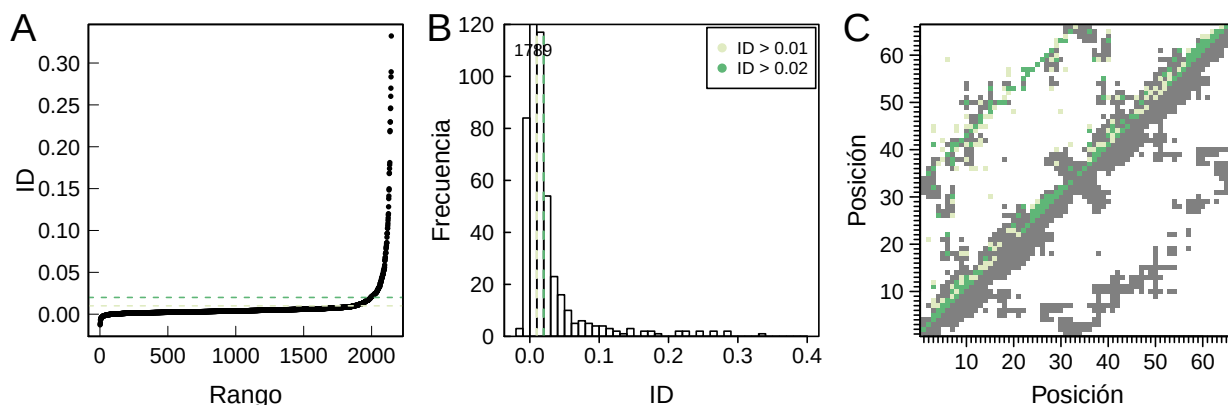


Figura 4.3: ¿Cómo elegir los valores de información directa significativos? Para los valores del AMS de ANK: A) valores de ID ordenados según el rango, de menor a mayor. La línea punteada en verde claro indica $ID=0.01$, y en verde oscuro $ID=0.02$. B) Histograma de los valores de ID. La barra más alta alcanza el valor de 1789. Las líneas punteadas indican los mismos cortes que el panel anterior. C) Mapa de contactos: gris indica que las posiciones indicadas por los ejes se encuentran en contacto en la estructura nativa, y blanco que no. Únicamente sobre la parte superior, en verde oscuro las posiciones del AMS donde $ID > 0.02$ y en verde claro donde $ID > 0.01$ ($y < 0.02$).

4.1.2. Resultados para todas las familias de proteínas repetitivas

El mismo procedimiento fue aplicado sobre distintas familias repetitivas: calculé para cada una la ID sobre un AMS de pares de repeticiones, y proyecté los valores más altos sobre un mapa de contactos representativo. Los resultados se encuentran en la figura 4.5. En primer lugar se observa que, análogamente a la familia ANK, todas las familias presentan alta correlación entre posiciones cercanas en secuencia (a menos de 5 aminoácidos en la cadena polipeptídica). Respecto de las posiciones más distantes se ven distintos comportamientos: la mayoría de las proteínas tienen posiciones altamente correlacionadas que son contactos o que son posiciones equivalentes en las repeticiones adyacentes. Este es el caso de TPR, TSP, CW, SEL, HEAT, WD40. Lo mismo sucede para las familias YAD, LRR, HECA y PENTAPEPTIDE, pero en este caso, las posiciones equivalentes de repeticiones adyacentes también son contactos de la estructura nativa. Luego se observan ANNEX, KELCH y LDL con alta coincidencia entre los ID más altos y los contactos, y finalmente ARM y PUF que no aparentan relación entre estas magnitudes.

Es de interés separar aquellas posiciones correlacionadas por ser contactos de aquellas correlacionadas por ser posiciones equivalentes, en particular para utilizarlo en predicciones de estructura. Esto no puede hacerse simplemente eliminando las correlaciones del segundo

caso dado que en muchas ocasiones son también los contactos que estabilizan la estructura proteica. En la sección 4.2 describiré el método que desarrollé para distinguir correlaciones debido a contactos en estructura.

4.1.3. Aproximación por máxima pseudo-verosimilitud

Otro método disponible para medir correlación entre posiciones es el cálculo de la máxima pseudo-verosimilitud, que describí en la sección 1.4.6. En el trabajo donde fue presentado [Ekeberg et al., 2013] indicaba mejor capacidad para predecir contactos en proteínas globulares. Utilicé la versión en Matlab disponible en <http://plmdca.csc.kth.se/> para realizar las estimaciones de correlación con este método sobre las familias de proteínas repetitivas bajo estudio. Los resultados pueden verse en la figura 4.4 para la familia ANK, junto a las predicciones de ID para simplificar la comparación, y en la figura 4.6 para las demás familias. Son, en la mayoría de los casos, muy similares a los obtenidos con ID. Se destaca una coincidencia levemente mayor con contactos en la familia ARM, respecto de los cálculos de ID.

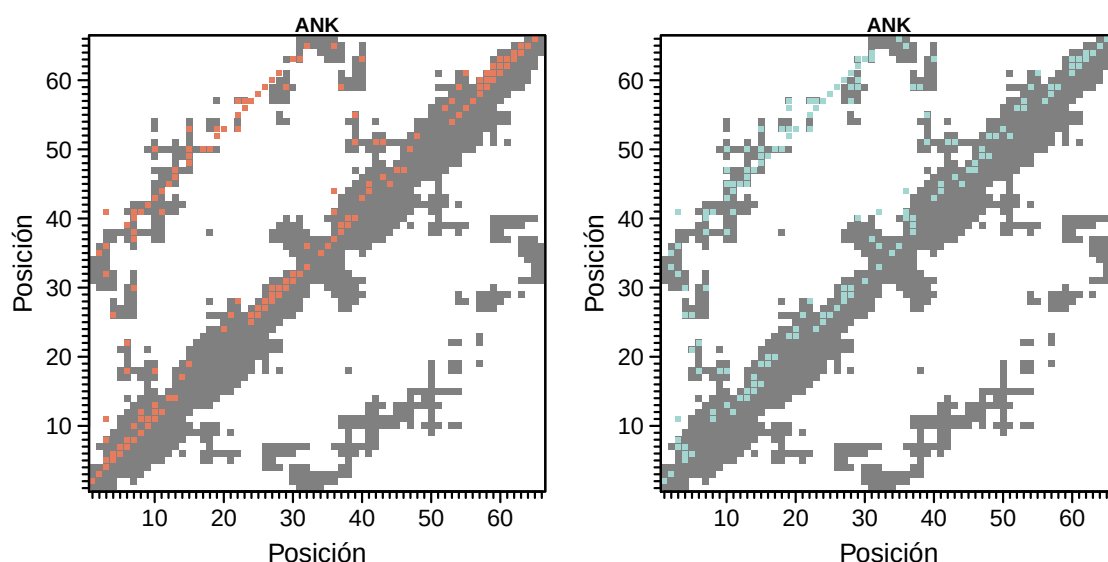


Figura 4.4: Comparación de posiciones con alta coevolución de acuerdo a ID (izquierda) y plmDCA (derecha).

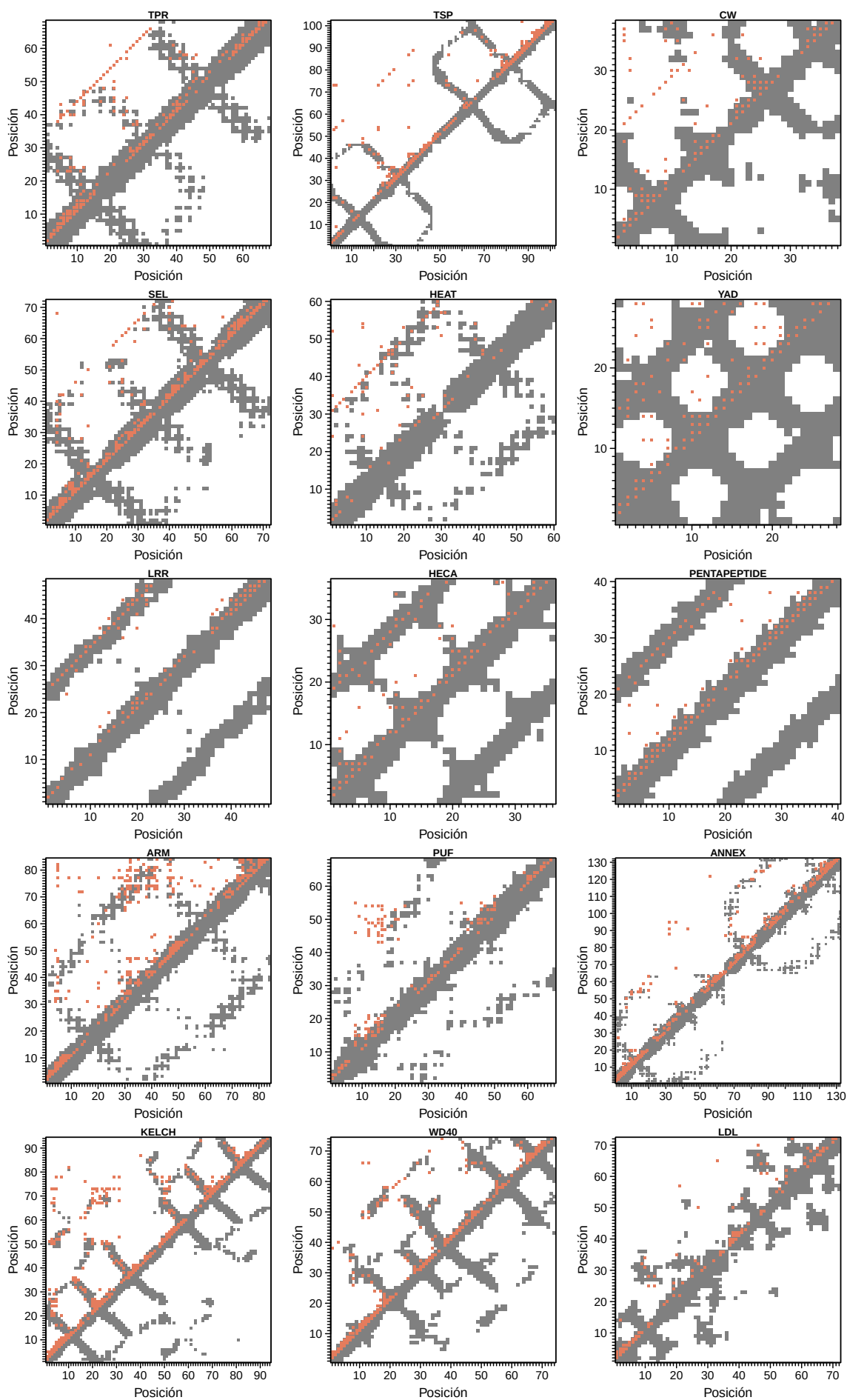


Figura 4.5: Mapas de contacto (blanco y gris) y valores más altos (naranja) de ID para familias de proteínas repetitivas.

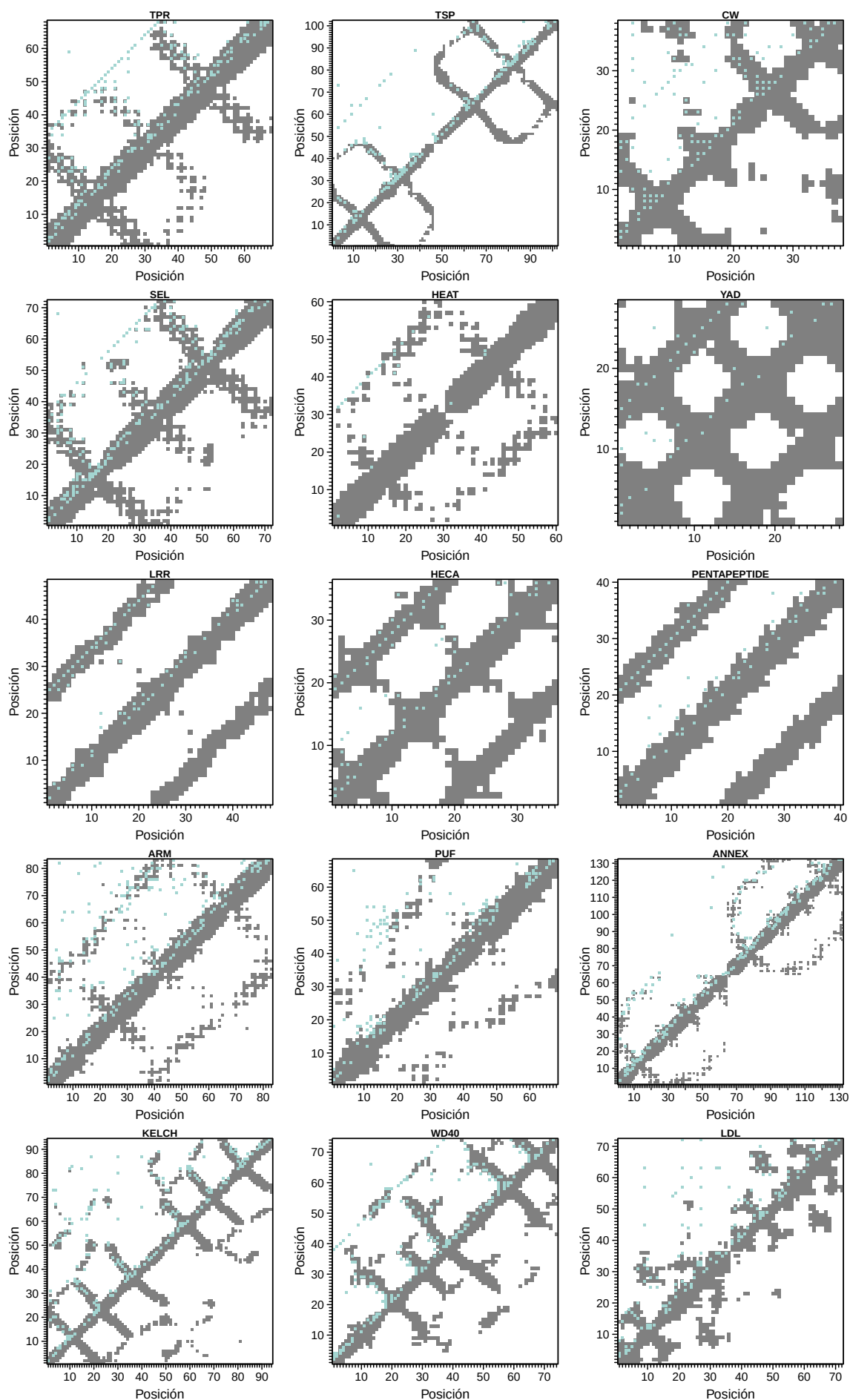


Figura 4.6: Mapas de contacto (blanco y gris) y valores más altos (celeste) de correlación calculados por máxima pseudo-verosimilitud para familias de proteínas repetitivas.

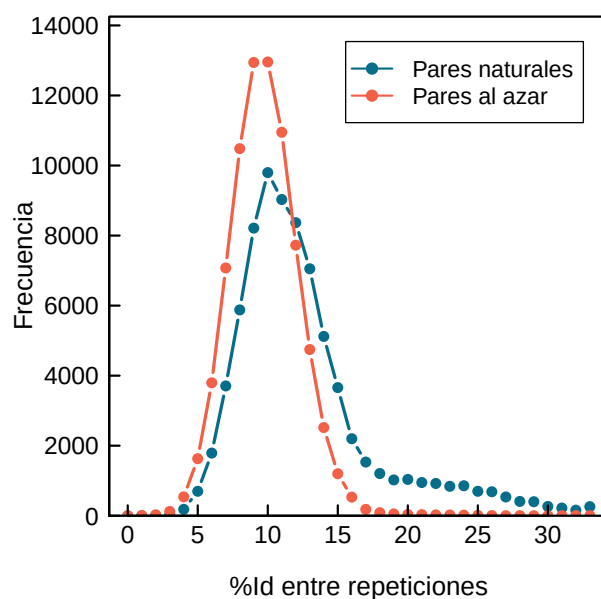


Figura 4.7: Histogramas de identidad en secuencia entre repeticiones consecutivas para la familia ANK. En azul la distribución para pares de repeticiones naturales. En naranja, para repeticiones naturales pero no consecutivas entre sí.

4.2. Predicción estructural

En la sección anterior se observó que los análisis de coevolución en secuencias de proteínas repetitivas resaltan las posiciones equivalentes de repeticiones adyacentes. Si las repeticiones consecutivas fueran exactamente iguales, los valores de ID entre estas posiciones deberían ser máximos dado que mostrarían coevolución perfecta. De igual manera, las submatrices de ID que abarca posiciones entre las repeticiones debería ser idéntica a la submatriz de pares de posiciones internos a cada repetición. Entonces, la identidad entre unidades repetidas debería ser tomada en cuenta al evaluar correlaciones entre repeticiones. Uno podría verse tentado a simplemente ignorar estas posiciones equivalentes entre repeticiones adyacentes, sin embargo esto es incorrecto dado que algunas veces se encuentran en contacto en la estructura tridimensional y son fundamentales para estabilizar el plegado nativo.

En la figura 4.7, para la familia ANK se encuentran la distribución para secuencias naturales (en azul) y para secuencias constituidas por dos repeticiones (cada una de una proteína natural) no adyacentes entre sí, que denominaré pares al azar (en naranja). Se observan diferencias entre ellas: las secuencias naturales tienen una media más grande y una cola hacia la derecha más poblada. A partir de ellas diseñé una corrección para los pesos asignados a cada

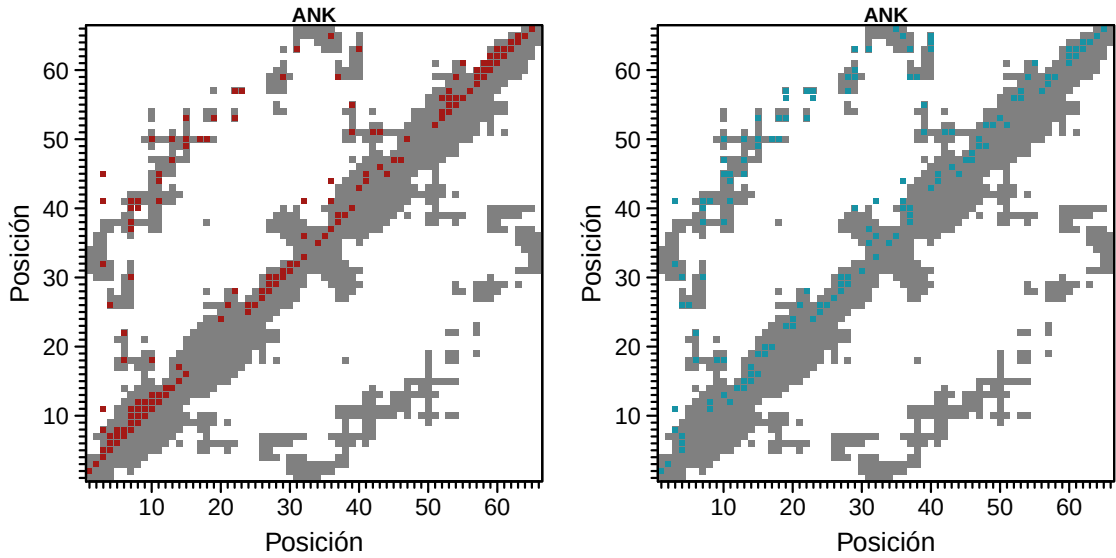


Figura 4.8: ID_{id} (izquierda) y $plmDCA_{id}$ (derecha) para la familia ANK representado sobre un mapa de contactos. El corte para identificar valores altos de coevolución es el mismo que utilizado en ID y plmDCA respectivamente (figura 4.4).

secuencia: una secuencia natural cuyas repeticiones tienen $\%Id=x$ (altura de la curva azul), adquiere un peso dado por la proporción de secuencias de pares al azar con ese mismo $\%Id$ (curva naranja), sobre la proporción de secuencias naturales con $\%Id=x$ (curva azul). Notar que estas proporciones se encuentran unívocamente definidas pues $\%Id$ sólo puede tomar valores discretos $n/33$, con $n=0,1,2,\dots,33$ (en el caso de ANK). Este valor multiplica al peso de Henikoff asignado previamente. Luego, se procede al igual que antes a calcular las frecuencias y la ID. Cuando incluya esta corrección, me referiré a la información directa como ID_{id} .

Los resultados para la información directa corregida ID_{id} , para la familia ANK se muestran en la figura 4.8. El corte para identificar valores relevantes de coevolución es el mismo que el utilizado en ID sin corregir. Se observa una clara mejora en la coincidencia de los valores altamente correlacionados con las posiciones en contacto en la estructura nativa. En particular varios puntos correspondientes a posiciones equivalentes en repeticiones adyacentes que no son contactos y tienen valores altos de ID tienen valores bajos de ID_{id} . Al mismo tiempo, puntos en las mismas posiciones que efectivamente se encuentran en contacto en la estructura nativa mantuvieron sus altos valores de correlación al incorporar la corrección.

El mismo análisis fue realizado sobre las demás familias de proteínas. En cada una de ellas se respetó el valor de corte para identificar pares de posiciones correlacionados que se

utilizó en ID sin corregir. Las distribuciones de porcentaje de identidad entre repeticiones adyacentes se pueden observar en la figura 4.9, y la ID_{id} en la figura 4.10. Para las familias TPR, SEL, WD40, TSP y CW las distribuciones de identidad entre repeticiones adyacentes naturales tienen desvíos hacia los valores más altos con respecto a la distribución de pares al azar. Asimismo, al incorporar la corrección por esta identidad a los cálculos de ID, las predicciones estructurales mejoran. El caso de HEAT tiene un comportamiento similar, pero presenta mucho ruido en sus distribuciones de identidad entre repeticiones. Más adelante mostraré que la cantidad de secuencias disponibles no es suficiente para realizar estas medidas. En otras familias se puede ver que no son afectadas negativamente. En el caso de HECA, PENTAPEPTIDE, LRR y YAD las posiciones equivalentes entre repeticiones se encuentran en contacto en la estructura tridimensional nativa. Si bien en las primeras dos las distribuciones de identidad entre repeticiones son iguales en las secuencias naturales y en los pares aleatorios, y en las otras dos familias son distintas, aplicar la corrección sobre el ID no modifica los pares detectados con mayor coevolución. De la misma manera, las familias ANNEX, KELCH y LDL que no presentaban un sesgo en las posiciones equivalentes entre repeticiones, y tampoco presentan variaciones en las distribuciones de identidad entre las repeticiones, mantienen los valores más altos de coevolución coincidentes con contactos. Finalmente las familias ARM y PUF no presentan coincidencias entre los pares de alta coevolución y los contactos, ni en ID ni en la versión corregida ID_{id} . En particular, las distribuciones de identidad entre repeticiones adyacentes y repeticiones aleatorias coinciden. En conclusión, la corrección aquí presentada es capaz de separar los altos valores de coevolución que se deben a residuos en contacto de aquellos altamente correlacionados por la repetitividad de estas proteínas. En ningún caso muestra actuar en detrimento de la predicción de contactos en el plegado nativo.

Es sencillo incorporar la misma corrección al método por máxima pseudo-verosimilitud ($plmDCA_{id}$). La corrección se realiza sobre las frecuencias marginales y conjuntas que sirven como datos en el procedimiento. Los resultados se muestran en la figura 4.11. Nuevamente, el valor de corte utilizado para identificar pares de posiciones correlacionados es idéntico al utilizado en la versión sin corregir, $plmDCA$. Las mejoras son equivalentes a las observadas en los cálculos de ID_{id} .

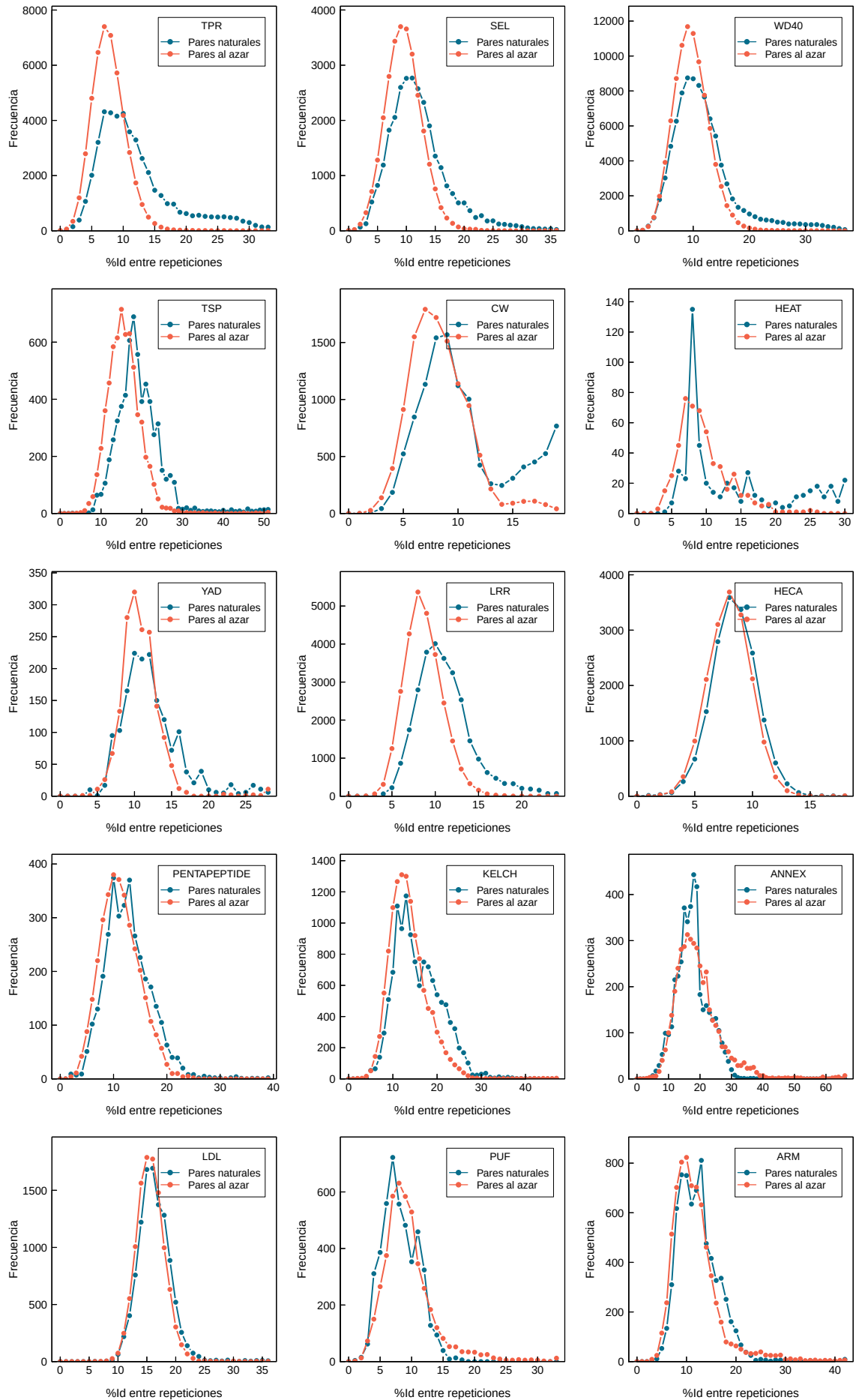


Figura 4.9: Histogramas de identidad en secuencia entre repeticiones consecutivas para varias familias repetitivas. En azul la distribución para pares de repeticiones naturales. En naranja, para repeticiones naturales pero no consecutivas entre sí.

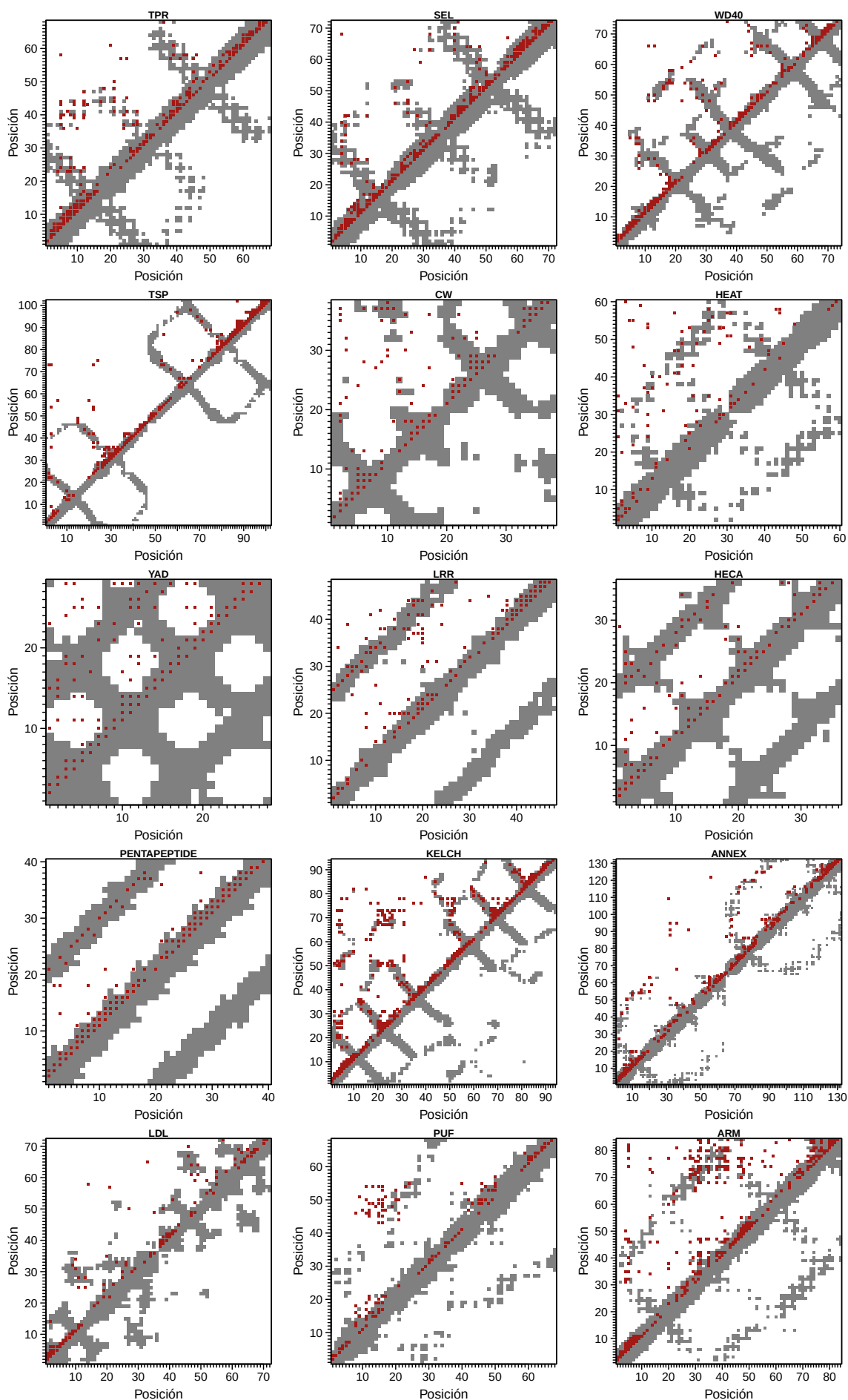


Figura 4.10: Mapas de contacto (blanco y gris) y valores más altos (rojo) de ID_{id} para familias de proteínas repetitivas.

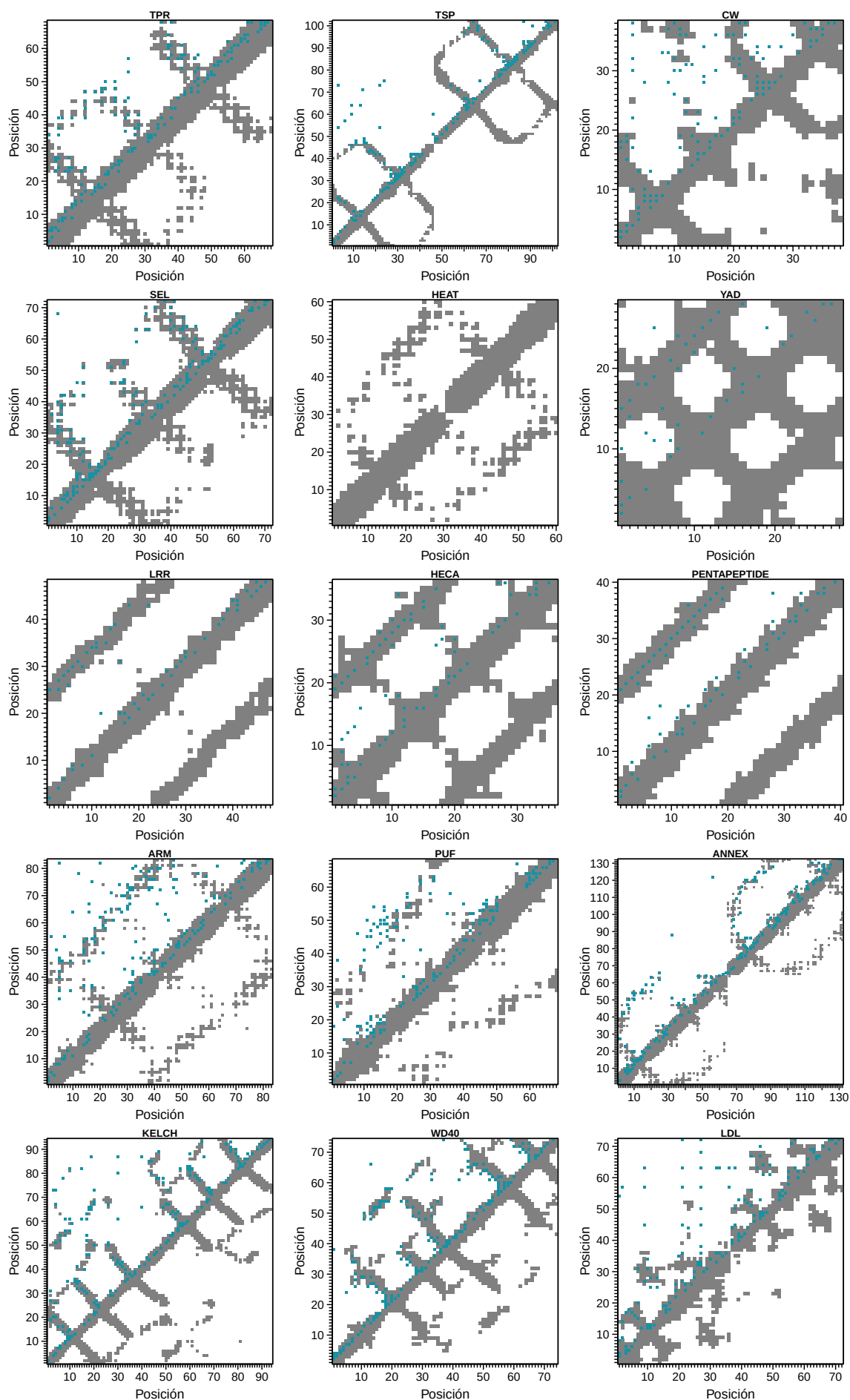


Figura 4.11: Mapas de contacto (blanco y gris) y valores más altos (azul) de $plmDCA_{id}$ para familias de proteínas repetitivas.

4.2.1. Eficacia del método para predecir contactos

Para analizar los métodos estudiados como predictores de contactos, analizamos las curvas ROC (definidas en la sección 2.4.5). Dado que no se sabe si los valores de correlación son comparables entre distintas familias proteicas, las cuantificaciones de la calidad de los métodos suele hacerse en referencia al largo de la proteína en estudio. En este análisis tomo el mismo criterio, estudiando las predicciones de acuerdo al largo de la repetición L de cada familia.

Para cada familia de proteínas consideré un subconjunto de todos los pares de posiciones posible, que consiste en los $5L$ pares con mayor valor de correlación. Entre ellos habrá una cantidad de contactos P y una cantidad de no-contactos N . Sobre este subconjunto, calculé las tasas de falsos positivos y verdaderos positivo tomando como predicciones los M valores más altos, con M igual a $L/10$, $L/5$, $L/3$, $L/2$, L , $3/2L$, $2L$, $2,5L$, $3L$, $3,5L$, $4L$, $4,5L$ y $5L$, redondeando en los casos que las fracciones no den números enteros. A partir de estas tasas, grafiqué las curvas ROC sobre todas las familias proteicas estudiadas simultáneamente para los distintos métodos estudiados: ID, IM, plmDCA y las correcciones por identidad entre repeticiones, ID_{id} , IM_{id} , y $plmDCA_{id}$ que muestro en la figura 4.12. También indico los valores del área bajo la curva, un indicador del éxito del predictor que va entre 0.5 (mal predictor) y 1 (excelente predictor).

En esta figura se observa que en proteínas repetitivas, ID es un mejor predictor de contactos que IM, al igual que fue mostrado en proteínas globulares. La corrección por identidad entre repeticiones mejora levemente las capacidades de predicción de ID y plmDCA. Esto está en acuerdo con los resultados informados en las secciones anteriores. Dado que la corrección ataca los errores en las posiciones equivalentes de repeticiones sucesivas, es esperable que las cuantificaciones en capacidad predictiva mejoren sólo levemente.

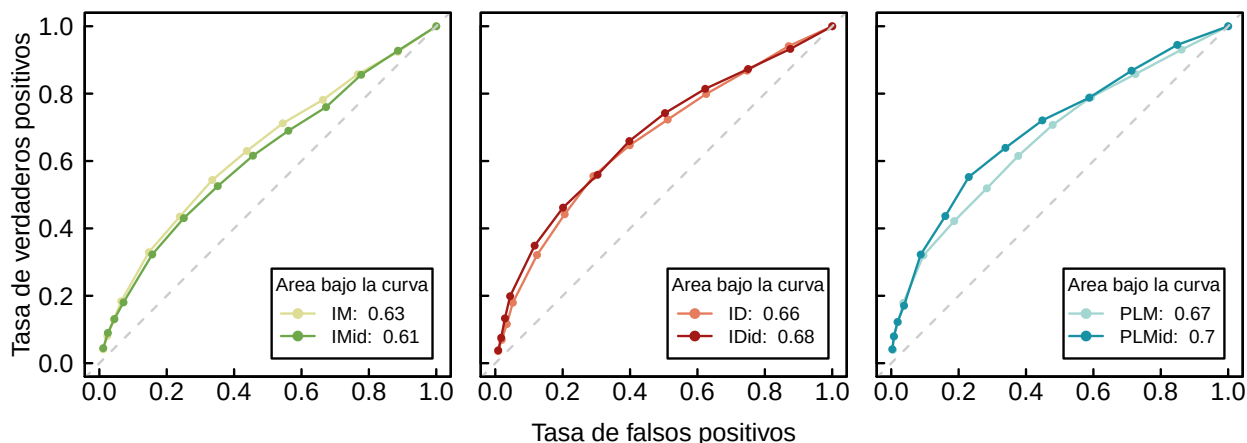


Figura 4.12: Curvas ROC sobre todas las familias estudiadas simultáneamente, para los métodos de análisis de correlación como predictor de contactos. De izquierda a derecha, IM, ID, y plmDCA. En colores más oscuros, las versiones corregidas por identidad de repeticiones consecutivas. En las leyendas, se indican los valores de área bajo la curva.

4.3. Acoplamientos lejanos a lo largo de un conjunto de repeticiones

El plegado de dominios repetitivos suele involucrar un proceso cooperativo de formación de estructura en una escala que excede a la primer repetición vecina [Aksel y Barrick, 2009]. El plegado de algunas regiones nuclea el plegado de los segmentos continuos, permitiendo tratar las dinámicas de plegado de proteínas repetitivas como cuasi-unidimensionales [Ferreiro y Wolynes, 2008]. Naturalmente surge la pregunta, ¿cómo cambian los acoplamientos evolutivos entre repeticiones a medida que aumenta la distancia entre ellas?

Para estudiar matrices de coevolución entre repeticiones lejanas, analicé por separado los conjuntos de repeticiones que se encuentran a una cantidad definida (n) de repeticiones en las secuencias naturales. Esto permite aumentar la cantidad de secuencias utilizadas para el análisis. Por ejemplo, en la 4.13 muestro un mapa de contactos de tres repeticiones consecutivas ANK (pdb:1N0R). Realicé un AMS de pares de repeticiones consecutivas y otro de pares de repeticiones que sean segundos vecinos en las cadenas polipeptídicas. En cada caso calculé las matrices de coevolución ID, cuyos pares de posiciones con valores más altos se encuentran marcados en naranja. Luego apliqué la corrección por identidad de secuencia, por separado para repeticiones consecutivas y para segundos vecinos. Los pares de posiciones con ID_{id} más altos se encuentran marcados en rojo. Con la corrección las señales simétricas se

atenúan y que surgen los verdaderos valores de coevolución. Se observa que hay menos puntos correlacionados entre segundos vecinos, consistente con la estructura elongada de estas proteínas en las que sólo se observan interacciones evidentes entre repeticiones consecutivas.

En principio, esta corrección se puede aplicar a arreglos de repeticiones tan largos como se desee, utilizando alineamientos de pares de repeticiones como hice para segundos vecinos. Sin embargo, la cantidad de secuencias disponibles se reduce a medida que aumenta el n requerido, y la calidad de los resultados empeora (sección 4.4). En la familia ANK hay disponibles muchas secuencias y pude extender este análisis a arreglos de hasta 17 repeticiones. En la figura 4.13, muestro la cantidad de pares de posiciones con valores de coevolución altos entre repeticiones (no entre posiciones internas a una repetición). Se observa que a medida que la separación entre repeticiones aumenta, tanto ID como ID_{id} entre repeticiones cae significativamente. Los residuos distantes en la cadena polipeptídica que interactúan son menos frecuentes, y esto se ve reflejado en los acoplamientos de coevolución. Mientras que ID_{id} cae hasta prácticamente cero para $n > 4$, una fracción de ID se mantiene no nula para repeticiones lejanas, indicando que la señal causada por la simetría en secuencia se

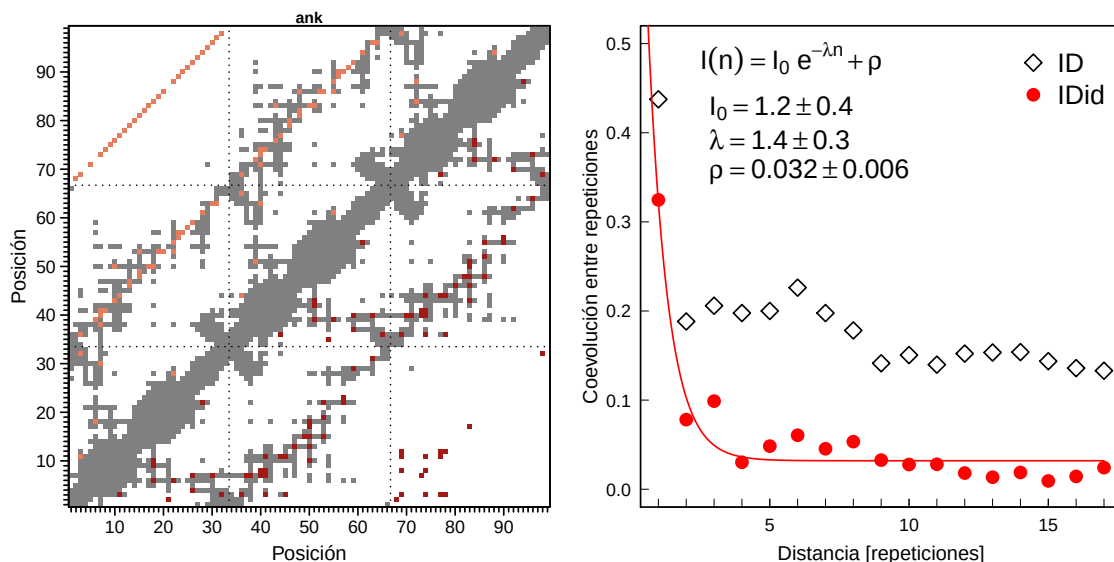


Figura 4.13: Izquierda. Mapa de contactos (gris) para tres repeticiones de la proteína diseñada 1N0R, de plegado ANK. En la parte superior, en naranja, los puntos de mayor coevolución de acuerdo a ID, utilizados. En la parte inferior, en rojo oscuro, los puntos de mayor coevolución de acuerdo a ID_{id}. A la derecha, para la familia ANK, proporción de puntos con altos valores de coevolución que radican en posiciones pertenecientes a distintas repeticiones según la distancia en secuencia de estas repeticiones. En diamantes negros de acuerdo a ID, y en puntos rojos según ID_{id}. El ajuste indicado corresponde a los valores de ID_{id}.

mantiene incluso para repeticiones distantes. El número de interacciones entre repeticiones cae exponencialmente con la separación entre las repeticiones, con un largo medio de $\approx 1,4$ repeticiones, sugiriendo que la interacción de coevolución para la familia ANK es de $\approx 1,5$ unidades. Trabajos anteriores sugirieron este valor de distancia característica para correlaciones entre repeticiones mediante el análisis de mecánica estadística sobre de experimentos de plegado [Street y Barrick, 2009, Wetzel et al., 2008] y simulaciones de plegado [Ferreiro et al., 2005].

4.4. Cantidad de secuencias necesarias

Para realizar cálculos robustos de ID es necesario tener suficientes secuencias efectivas (sección 2.3.1) para aproximar correctamente las distribuciones de frecuencia $f_i(a_i)$ y $f_{ij}(a_i, b_j)$. Dado que no existe un principio que indique cuál es el número de secuencias necesarias, busqué determinarlo de manera empírica.

Para la familia ANK tomé subconjuntos de secuencias removiendo recursivamente la mitad de las secuencias del AMS original de pares de repeticiones, y repetí los cálculos de ID_{id} sobre cada subconjunto. A modo de ejemplo, en la figura 4.14 se muestra la matriz de ID_{id} para todas las secuencias y para el subconjunto con 146 secuencias: los valores de ID_{id} se ven alterados, la señal se torna indistinguible del ruido de fondo y por lo tanto la medida de

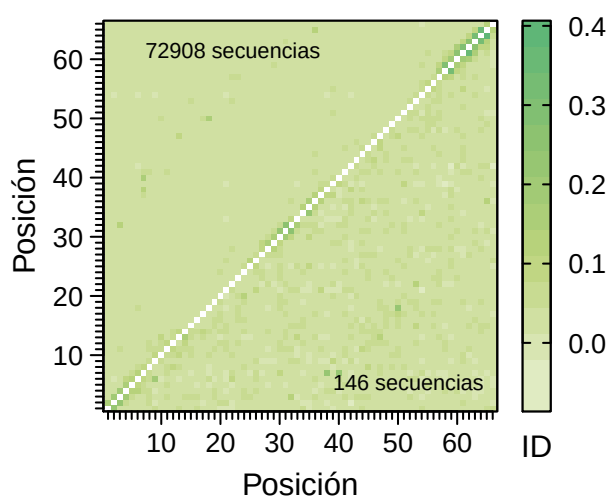


Figura 4.14: Comparación de la matriz de ID_{id} utilizando todas las secuencias disponibles (parte superior) o únicamente 146 (parte inferior).

correlación ya no es confiable. En la figura 4.15 se encuentran los cambios de ID_{id} para algunos pares de posiciones particulares cuando se reduce el número de secuencias en los cálculos, para esquematizar su dependencia. Cuando la cantidad de secuencias es grande, el valor de ID_{id} se mantiene estable, pero cuando son pocas comienza a tener grandes oscilaciones entre un subconjunto y otro. Una manera de cuantificar este comportamiento es poner una cota a la diferencia entre el valor obtenido con un subconjunto de secuencias (ID_{id}^s) y el obtenido al utilizar todas las secuencias disponibles (ID_{id}). Esta cota la escogí arbitrariamente como un 1 % de la variabilidad entre todas las posiciones al utilizar todas las secuencias. Es decir:

$$|ID_{id,ij}^s - ID_{id,ij}| < 0,01 \cdot (\text{máx}(ID_{id}) - \text{mín}(ID_{id})) \quad (4.1)$$

En la figura 4.15, esta cota se encuentra marcada por la franja gris. Para cuantificar la convergencia de manera global, es decir todos los pares de posiciones simultáneamente, evalué cuántos pares de posiciones convergen con el criterio de la ecuación 4.1 para cada subconjunto. Los resultados para la familia ANK pueden verse en la figura 4.16. Cuando se utilizan menos de mil secuencias, la convergencia es muy pobre. Como es de esperar, a medida que se incorporan más secuencias al análisis, la cantidad de puntos que adquieren su valor final aumenta, y para más de 10000 secuencias la convergencia es prácticamente completa. En particular, para 6500 secuencias el 90 % de valores se encuentran convergidos, y la curva alcanza un comportamiento de saturación. En la misma figura, en color celeste, muestro la

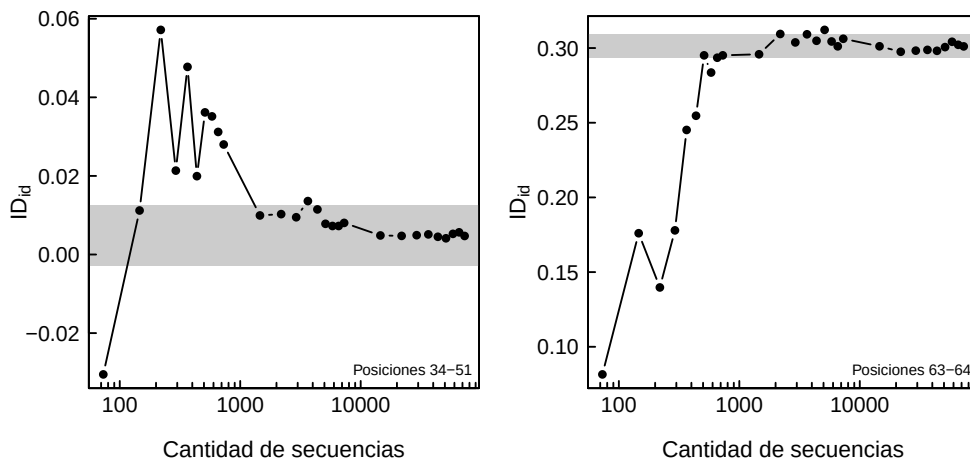


Figura 4.15: Ejemplos de comportamiento del valor de ID_{ID} según la cantidad de secuencias utilizadas.

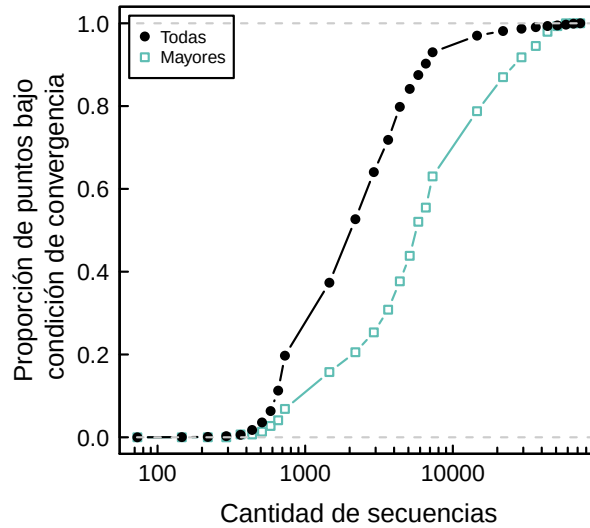


Figura 4.16: Curva empírica de convergencia de acuerdo a la cantidad de secuencias utilizadas para la familia ANK.

curva de convergencia para los valores más altos de ID_{id} (mayores a 0.02). Llamativamente, estos muestran una curva retrasada, necesitando más secuencias para llegar a sus valores finales.

Este mismo procedimiento fue aplicado en las demás familias. Los resultados se encuentran en la figura 4.17. Varias familias reproducen el comportamiento de ANK: TPR, LRR, SEL, KELCH, WD40, HECA, con una curva que satura al utilizar aproximadamente 3500 secuencias. La curva de los valores más altos de ID_{id} se encuentra retrasada, indicando que estos valores convergen con un mayor número de secuencias analizadas. Otras familias, ANNEX, PUF, ARM, CW, LDL, PENTAPEPTIDE, TSP y YAD, si bien la curva no converge sino que crece linealmente, al utilizar más de 3000 secuencias más del 80 % de los puntos alcanzaron el valor de saturación. Por otro lado, ARM, PUF, CW y LDL. En la mayoría de estos casos, los puntos con ID_{ID} alto sigue la misma tendencia que todo el conjunto de los valores. Finalmente HEAT evidencia que la cantidad de secuencias no es suficiente para determinar robustamente la matriz de ID_{id} . Al considerar la mitad del total de las secuencias, menos del 20 % de los puntos conserva el valor de ID_{ID} . De esta forma, desestimamos conclusiones que se puedan extraer de este alineamiento con los métodos aquí presentados.

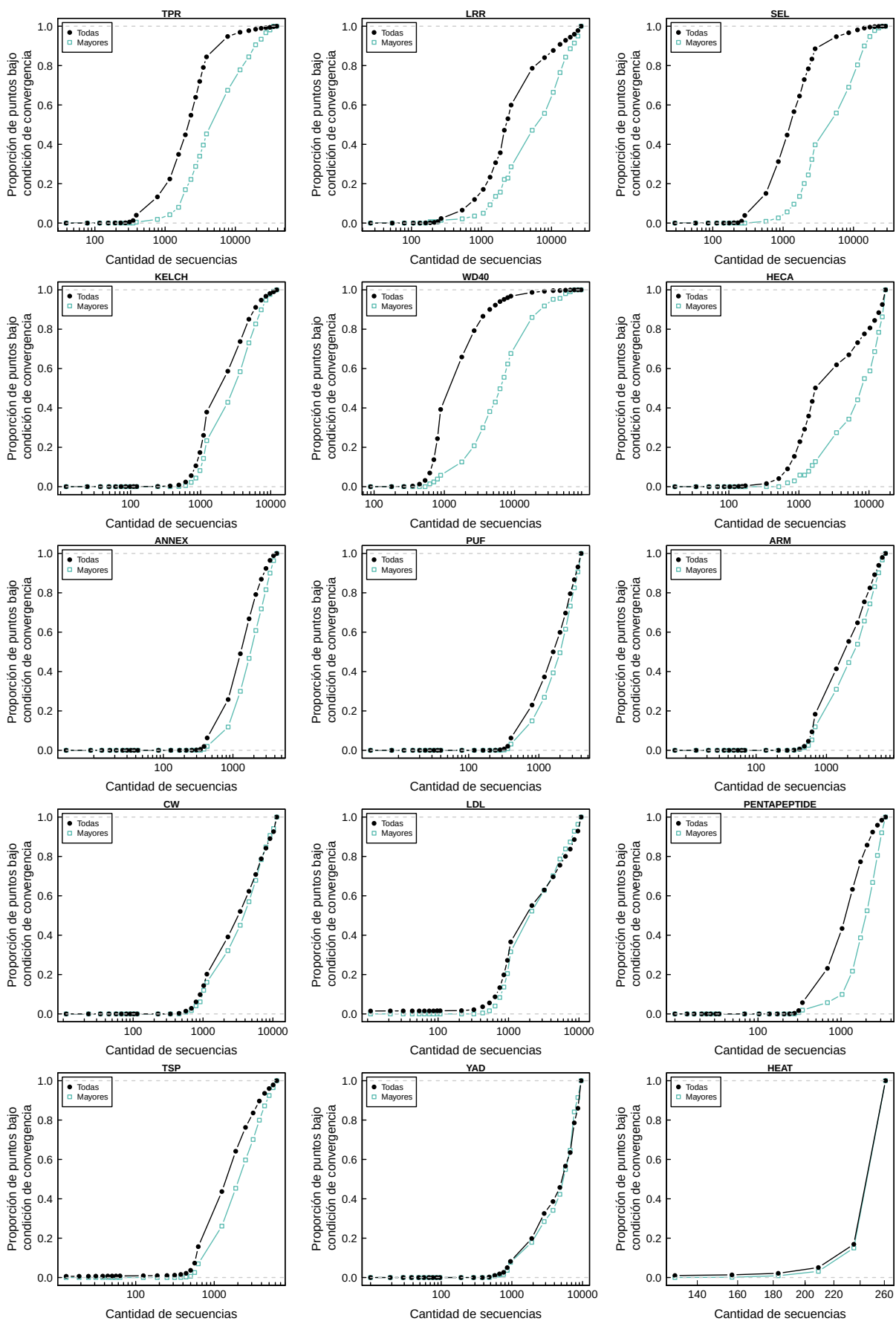


Figura 4.17: Curva empírica de convergencia para otras familias de proteínas.

4.5. Simulaciones moleculares de plegado

Los métodos de coevolución en proteínas son utilizados para predecir contactos en el plegado nativo, y con ellos la estructura tridimensional de la proteína. Para discutir sobre la necesidad del uso de la corrección en las medidas de ID, realizo estudios computacionales de plegado en la proteína 1N0R, proteína diseñada y que toma el plegado de las proteínas ANK.

Las simulaciones fueron realizadas con el software gromacs [Abraham et al., 2015], y el protocolo se basa en los propuestos por Sulkowska et al. [2012]. Estas simulaciones se conocen como grano grueso dado que cada aminoácido es representado por una única esfera centrada en el carbono α , y en contraposición a las simulaciones de *todos los átomos*, donde se incluyen todos los átomos que conforman los aminoácidos a costa de aumentar los tiempos computacionales. El algoritmo aplicado permite incluir tendencias en la formación de contactos entre pares de residuos, formadores de la estructura terciaria. Se realizaron dos simulaciones. Como predicciones de contactos se utilizaron en un caso los pares de residuos con mayor ID, y en otro caso los que tienen mayor ID_{id} , que se pueden comparar con un mapa de contactos en la figura 4.18A. Se observa que los valores de coevolución sin corregir por la simetría de las proteínas repetitivas, muestran muchas predicciones entre residuos distantes, más allá de las repeticiones adyacentes, y que no tienen correlato en el mapa de contactos. En los paneles B y C de la figura 4.18 muestro un cuadro de cada simulación, es decir una predicción de estructura utilizando los valores de ID_{id} (en rojo, en el panel B), y usando ID (en azul, en el panel C). En ambos paneles en gris muestro la estructura obtenida experimentalmente, reportada en la bibliografía [Mosavi et al., 2002]. Al utilizar los valores de coevolución sin corregir, las estructuras locales son similares a las presentes en la proteína natural, pero globalmente no toma la forma alargada, cuasi-unidimensional, sino que se quiebra sobre sí misma tomando un aspecto de proteína globular. En cambio, al utilizar los valores corregidos, ID_{id} , la estructura elongada se reproduce con alta fidelidad. Este resultado apoya la importancia del uso de la corrección por similitud en secuencia en el caso de proteínas repetitivas.

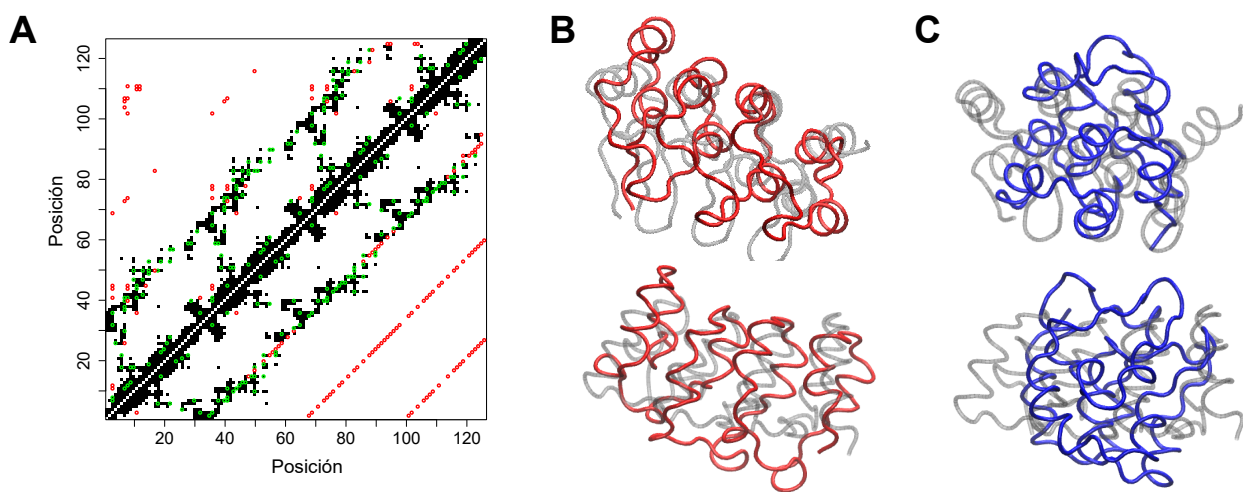


Figura 4.18: Simulaciones de plegado en la proteína diseñada 1N0R. (A) En blanco y negro, mapa de contactos, con corte en 8Å. En la parte superior, pares de posiciones con alta coevolución de acuerdo a ID_{id} , en verde si coinciden con un contacto y en rojo si no. En la parte inferior, pares de posiciones de alta coevolución de acuerdo con ID. (B) En gris, estructura experimental. En rojo, resultado de la simulación usando como predicción de contactos los puntos con alto ID_{id} . (C) En gris, estructura experimental. En azul resultados de la simulación utilizando como predicción de contacto los puntos con alto ID.

Capítulo 5

Conclusiones

Las proteínas son las moléculas más versátiles dentro de los sistemas vivos. Participan en todos los procesos biológicos. Gracias a su capacidad de procesar información respecto de su entorno y actuar en respuesta, han sido interpretadas como computadoras moleculares [Bray, 1995]. Esto despierta el interés en comprender los principios que las rigen, y la posibilidad de modificarlas y adaptarlas con fines ingenieriles, biomédicos y tecnológicos. La sucesión lineal de aminoácidos que constituye una proteína es la guía del proceso de plegado mediante el cual la molécula adopta estructuras tridimensionales, y en última instancia la responsable de las actividades biológicas que puede realizar. Sin embargo, la alta complejidad de estos sistemas dificulta nuestra comprensión de cómo la información relevante se encuentra codificada en la secuencia que constituye cada molécula.

A lo largo de tiempos evolutivos, las proteínas sufren modificaciones en sus secuencias que cambian sus propiedades fisicoquímicas, y por lo tanto su estabilidad de plegado o su capacidad de interacción con otras moléculas. Sus secuencias llevan en sí las huellas del proceso evolutivo. Las proteínas que se encuentran hoy en la naturaleza son un catálogo de variantes viables. Este trabajo se basa en las ideas de co-evolución en proteínas. Las modificaciones que sufren las secuencias de proteínas tienen consecuencias en el plegado. Dada que en la estructura plegada los residuos interactúan entre sí, la sustitución de uno de ellos puede implicar conflictos energéticos en los demás. Esto puede llevar a la inviabilidad de la variante, o a forzar sustituciones que permitan compatibilidad. En última instancia, estos acoplamientos causados por interacciones en la proteína funcional podrían verse como

correlaciones en las sustituciones dentro del catálogo actual de proteínas. Bajo esta premisa, los métodos de análisis de coevolución en secuencias de proteínas tienen como objetivo encontrar correlaciones entre mutaciones de aminoácidos a lo largo de las secuencias [De Juan et al., 2013]. Con la creciente disponibilidad de secuencias este tipo de análisis ha tomado relevancia.

Este trabajo se enfoca en familias de proteínas repetitivas. Estas proteínas se encuentran formadas por motivos estructurales de entre 20 y 40 aminoácidos que se repiten y al plegarse se apilan, formando estructuras elongadas que contienen contactos únicamente a corto alcance. Estas repeticiones presentan cuasi simetrías tanto a nivel de estructura primaria como de estructura terciaria. Estas proteínas se encuentran en todos los reinos y participan de un amplio espectro de procesos biológicos. Es posible que su éxito evolutivo se asocie a cierta facilidad de cambio: eventos de duplicación de repeticiones dan origen a proteínas largas y estables, basadas en módulos estructurales existentes, en vez de originar una nueva conformación estructural. Con un andamio en común, variaciones en la secuencia de aminoácidos les han permitido involucrarse en una gama de funciones muy diversa. Han sido reportadas mayormente como proteínas de interacción.

Las proteínas repetitivas han sido estudiadas para ser diseñadas. Un modelo que tuvo éxito es el diseño por consenso, donde cada repetición se construye de acuerdo a los motivos en secuencia más abundantes, y algunas posiciones que muestran alta variabilidad en las secuencias naturales tienen permitido variar para optimizar su unión a otras moléculas. Sin embargo este diseño es restrictivo, dejando variar sólo algunos residuos. La alta estabilidad de plegado de las variantes obtenidas, por encima de las proteínas naturales, sugiere que mayor variabilidad en las secuencias es posible con el fin de adaptarlas a diversas interacciones. En este sentido, los modelos de coevolución buscan encontrar límites más laxos, e igualmente efectivos, para el diseño proteico.

En el capítulo 3 presenté un modelo estadístico acorde a las ideas de co-evolución para describir los detalles energéticos de las proteínas repetitivas. Consiste en una generalización del modelo de Potts que incluye términos locales y de interacciones de a pares de residuos, y un término extra que capta las correlaciones de mayor orden, y que responde a la alta

similitud en secuencia entre repeticiones adyacentes. El modelo se encuentra restringido por características de las familias proteicas: las frecuencias de ocurrencia de aminoácidos, las frecuencias de co-ocurrencia y la identidad entre repeticiones adyacentes dentro de las proteínas naturales. El modelo es derivado únicamente a partir de las secuencias de aminoácidos. El ajuste de los parámetros es computacionalmente costoso, incluso en las proteínas repetitivas cuya simetría posibilita reducir la dimensionalidad del problema. Sin embargo, una vez obtenidos los parámetros, el campo puede utilizarse para asignar energías a secuencias individuales de manera fácil y rápida.

Estos modelos fueron desarrollados y estudiados en profundidad para tres familias de proteínas repetitivas: ANK, TPR y LRR. Los valores de energía para secuencias naturales permiten distinguirlas de tiras aleatorias de aminoácidos. Además, proteínas diseñadas de las familias ANK y TPR, que han sido reportadas ser extremadamente estables, obtienen valores de energía solapadas con la distribución de proteínas naturales, en el extremo de energías más bajas (más favorables energéticamente). Esto sugiere que la energía aquí descrita podría estar relacionada con la estabilidad del plegado. Reforzando esta idea, luego analicé los cambios de energía en secuencias de proteínas naturales que han sido sometidas a mutaciones puntuales de uno, dos o tres residuos. Estos cambios energéticos se corresponden con las variaciones en la estabilidad del plegado medidas experimentalmente.

El modelo desarrollado puede ser utilizado para generar nuevas secuencias mediante simulaciones de Montecarlo, que acuerden con las restricciones de frecuencias observadas en la familia proteica. Este nuevo conjunto de secuencias es estadísticamente indistinguible del ensamble de secuencias de proteínas naturales, sugiriendo que este protocolo puede ser utilizado para el diseño de nuevas proteínas repetitivas.

El diseño de proteínas repetitivas por consenso ha mostrado éxito en el desarrollo de estructuras plegables y estables [Binz et al., 2003, Mosavi et al., 2002, Main et al., 2003b, Kajander et al., 2007, Stumpp et al., 2003, Urvoas et al., 2010, Parmeggiani et al., 2008, Varadamsetty et al., 2012, Geiger-Schuller y Barrick, 2016]. Sin embargo esta forma de diseñar limita la variabilidad de las secuencias, permitiendo únicamente mutar aminoácidos en posiciones específicas. El conjunto de los resultados en el capítulo 3 sugieren que el modelo de Potts implementado puede ser utilizado para el diseño de proteínas dado que

captura propiedades estructurales y de estabilidad de las mismas. Debido a que tiene en cuenta las relaciones entre residuos a lo largo de la proteína, permite explorar secuencias que se apartan del consenso, incluso en posiciones muy conservadas, y por lo tanto obtener mayor variabilidad y flexibilidad.

En el capítulo 4 estudié las aproximaciones de campo medio y maximización de la pseudo-verosimilitud del campo para proteínas repetitivas. Estas aproximaciones permiten obtener medidas de correlación (ID y plmDCA) entre posiciones en la secuencia, ignorando la naturaleza del aminoácido. Estos modelos mostraron ser buenos predictores de contactos y superficies de interacción en proteínas globulares.

En el caso de proteínas repetitivas se encuentra señal alta de correlación en tres casos: en primer lugar entre residuos cercanos en la cadena polipeptídica, a una longitud relacionada con la formación de estructura secundaria (menos de 5 aminoácidos de distancia). En segundo lugar entre posiciones equivalentes en repeticiones consecutivas. Si una repetición tiene largo L , se encuentra señal entre los pares de posiciones i e $i + L$. Estos valores tienen su origen en la naturaleza repetitiva de estas proteínas. Si repeticiones consecutivas fueran idénticas, los pares de posiciones señalados tendrían covariación perfecta. Finalmente, los demás valores altos de correlación se relacionan con propiedades de la estructura terciaria de las proteínas, en particular con los contactos. Sin embargo, los últimos dos conjuntos no son disjuntos: pares de posiciones i e $i + L$ muchas veces se encuentran en contacto en la estructura tridimensional. Por ejemplo, en la familia PENTAPEPTIDE todos los pares $i-i + L$ son verdaderos contactos, por lo que las interacciones simétricas no pueden ser ignoradas. Por el contrario, los pares de repeticiones TPR en los que la mayoría de estos pares de posiciones $i-i + L$ no están en contacto a pesar de tener altos valores de correlación. Poder distinguir entre estas dos causalidades puede ser de interés, por ejemplo, para usar las medidas de correlación como predictores de contactos y, en última instancia, de estructura tridimensional.

Desarrollé un método de repesado de secuencias que permite distinguir entre estas señales. El método consiste en comparar las similitudes en secuencia de repeticiones adyacentes en proteínas naturales con las similitudes entre repeticiones en diferentes proteínas, y restar relevancia a aquellas secuencias naturales cuya similitud se aparta de las repeticiones no adya-

centes. En la práctica esto se traduce a restar la relevancia de las secuencias con porcentaje de identidad entre repeticiones adyacentes mayor al $\approx 15\%$. Como resultado, los nuevos cálculos de correlación permitieron distinguir los pares de posiciones asociados únicamente por la simetría de las proteínas repetitivas de aquellos que están en contacto. Esto desemboca en una mejora de las predicciones de contacto, que son más significativas al considerar arreglos más largos. La importancia de tener en cuenta esta corrección quedó plasmada en una simulación computacional de plegado para una proteína de la familia repetitiva ANK, utilizando como pares de contactos predichos los puntos con altos valores de co-evolución. Utilizar la corrección en estas predicciones permitió pasar de una predicción estructural doblada sobre sí misma, análoga a una proteína globular, a una estructura elongada, característica de las proteínas repetitivas.

La corrección desarrollada, aquí aplicada en los métodos aproximación de campo medio y maximización de la pseudo-verosimilitud, puede ser extendida a otros métodos de medición de correlaciones entre sustituciones de aminoácidos. Estos métodos son estadísticos y requieren de una número grande de secuencias disponibles, según la familia estudiada (su longitud, topología y ontología). Una cuantificación empírica para determinar la cantidad de secuencias en las familias estudiadas fue realizada. En la mayoría de las familias el $\sim 90\%$ de los pares de residuos se puede establecer con confianza con ~ 3000 secuencias. Las interacciones de mayor valor son detectadas con confianza incluso para muestreos escasos.

Las proteínas repetitivas se pliegan de forma cooperativa [Aksel y Barrick, 2009]. La nucleación del plegado en alguna región facilita el plegado de los segmentos contiguos, lo que permite un tratamiento cuasi unidimensional de la dinámica [Ferreiro y Wolynes, 2008]. Las medidas de co-evolución permitieron estudiar las escalas características de las variaciones de secuencia en la familia ANK. Estas caen exponencialmente con la distancia entre las repeticiones analizadas. La longitud característica de estas correlaciones es $\sim 1,4$ unidades repetidas, valor notablemente cercano a la inferida por análisis de mecánica estadística sobre de experimentos de plegado [Street y Barrick, 2009, Wetzel et al., 2008] y simulaciones de plegado [Ferreiro et al., 2005]. Estas predicciones se basan en la aproximación de covarianzas de largo alcance a partir de conjuntos de interacciones entre repeticiones de a pares, permitiendo la aplicación del procedimiento para estructuras arbitrariamente grandes para las

cuales un cálculo exacto sería computacionalmente prohibitivo.

Ya en los años '60 Anfinsen [Anfinsen et al., 1961] postuló su principio que indica que toda la información del plegado proteico (y en consecuencia de la maquinaria celular) está codificada en las secuencias de aminoácidos que conforman cada proteína. En los últimos años, con los avances en técnicas experimentales que permitieron la secuenciación de miles de proteínas, este principio se resignifica. Podemos estudiar enormes conjuntos de secuencias, analizar las huellas evolutivas estampadas en ellas, y aprender como han tomado forma a lo largo de la historia. A su vez, permite inferir cuales son las restricciones presentes en cada secuencia que dotan de propiedades biológicamente relevantes (estructura, actividad, capacidad de interaccionar, funcionalidad) a cada molécula. El trabajo desarrollado en esta tesis propone mecanismos de estudio para los desafíos que conllevan las posibilidades actuales, para extraer información sobre estas propiedades únicamente a partir de las estructuras primarias. Por otro lado, espero que las herramientas concretas desarrolladas en este trabajo aporten al diseño inteligente de proteínas para generar construcciones de moléculas más variables y maleables, que puedan ser adaptadas a las aplicaciones más diversas, al igual que han hecho las proteínas en la naturaleza.

Bibliografía

Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C Smith, Berk Hess, y Erik Lindahl. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1:19–25, 2015.

Tural Aksel y Doug Barrick. Analysis of repeat-protein folding using nearest-neighbor statistical mechanical models. *Methods in enzymology*, 455:95–125, 2009.

Massimo Andreatta, Santiago Laplagne, Shuai Cheng Li, y Stephen Smale. Prediction of residue-residue contacts from protein families using similarity kernels and least squares regularization. *arXiv preprint arXiv:1311.1301*, 2013.

Christian B Anfinsen. Studies on the principles that govern the folding of protein chains. *Science*, 181(4096):223–230, 1973.

Christian B Anfinsen, Edgar Haber, Michael Sela, y FH White. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proceedings of the National Academy of Sciences*, 47(9):1309–1314, 1961.

Alex Bateman, Lachlan Coin, Richard Durbin, Robert D Finn, Volker Hollich, Sam Griffiths-Jones, Ajay Khanna, Mhairi Marshall, Simon Moxon, Erik LL Sonnhammer, et al. The pfam protein families database. *Nucleic acids research*, 32(suppl 1):D138–D141, 2004.

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

- Gregory L Blatch y Michael Lässle. The tetratricopeptide repeat: a structural motif mediating protein-protein interactions. *Bioessays*, 21(11):932–939, 1999.
- Adam J Bogdanove y Daniel F Voytas. Tal effectors: customizable proteins for dna targeting. *Science*, 333(6051):1843–1846, 2011.
- Dennis Bray. Protein molecules as computational elements in living cells. *Nature*, 376(6538):307, 1995.
- Joseph D Bryngelson, José Nelson Onuchic, Nicholas D Socci, y Peter G Wolynes. Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins: Structure, Function, and Bioinformatics*, 21(3):167–195, 1995.
- Lukas Burger y Erik Van Nimwegen. Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS Comput Biol*, 6(1):e1000633, 2010.
- Ryan R Cheng, Faruck Morcos, Herbert Levine, y José N Onuchic. Toward rationally re-designing bacterial two-component signaling systems using coevolutionary information. *Proceedings of the National Academy of Sciences*, 111(5):E563–E571, 2014.
- Simona Cocco, Remi Monasson, y Martin Weigt. From principal component to direct coupling analysis of coevolution in proteins: low-eigenvalue modes are needed for structure prediction. *PLoS Comput Biol*, 9(8):e1003176, 2013.
- A Contini y G Tiana. A many-body term improves the accuracy of effective potentials based on protein coevolutionary data. *The Journal of chemical physics*, 143(2):07B608_1, 2015.
- Aitziber L Cortajarena, Jimin Wang, y Lynne Regan. Crystal structure of a designed tetratricopeptide repeat module in complex with its peptide ligand. *FEBS journal*, 277(4):1058–1066, 2010.
- Naomi Courtemanche y Doug Barrick. The leucine-rich repeat domain of internalin b folds along a polarized n-terminal pathway. *Structure*, 16(5):705–714, 2008.

- Thuy Phuong Dao, Ananya Majumdar, y Doug Barrick. Highly polarized c-terminal transition state of the leucine-rich repeat domain of pp32 is governed by local stability. *Proceedings of the National Academy of Sciences*, 112(18):E2298–E2306, 2015.
- David De Juan, Florencio Pazos, y Alfonso Valencia. Emerging methods in protein co-evolution. *Nature Reviews Genetics*, 14(4):249–261, 2013.
- Ingrid DeVries, Diego U Ferreiro, Ignacio E Sánchez, y Elizabeth A Komives. Folding kinetics of the cooperatively folded subdomain of the $\kappa\text{b}\alpha$ ankyrin repeat domain. *Journal of molecular biology*, 408(1):163–176, 2011.
- Ken A Dill y Justin L MacCallum. The protein-folding problem, 50 years on. *Science*, 338(6110):1042–1046, 2012.
- Ricardo N Dos Santos, Faruck Morcos, Biman Jana, Adriano D Andricopulo, y José N Onuchic. Dimeric interactions and complex formation using direct coevolutionary couplings. *Scientific reports*, 5:13652, 2015.
- Magnus Ekeberg, Cecilia Lövkvist, Yueheng Lan, Martin Weigt, y Erik Aurell. Improved contact prediction in proteins: using pseudolikelihoods to infer potts models. *Physical Review E*, 87(1):012707, 2013.
- Magnus Ekeberg, Tuomo Hartonen, y Erik Aurell. Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. *Journal of Computational Physics*, 276:341–356, 2014.
- Rocío Espada, R Gonzalo Parra, Manfred J Sippl, Thierry Mora, Aleksandra M Walczak, y Diego U Ferreiro. Repeat proteins challenge the concept of structural domains. *Biochemical Society Transactions*, 43(5):844–849, 2015.
- Diego U Ferreiro y Peter G Wolynes. The capillarity picture and the kinetics of one-dimensional protein folding. *Proceedings of the National Academy of Sciences*, 105(29):9853–9854, 2008.

- Diego U Ferreira, Samuel S Cho, Elizabeth A Komives, y Peter G Wolynes. The energy landscape of modular repeat proteins: topology determines folding mechanism in the ankyrin family. *Journal of molecular biology*, 354(3):679–692, 2005.
- Diego U Ferreira, Carla F Cervantes, Stephanie ME Truhlar, Samuel S Cho, Peter G Wolynes, y Elizabeth A Komives. Stabilizing $\kappa\text{b}\alpha$ by “consensus” design. *Journal of molecular biology*, 365(4):1201–1216, 2007.
- Diego U Ferreira, Aleksandra M Walczak, Elizabeth A Komives, y Peter G Wolynes. The energy landscapes of repeat-containing proteins: topology, cooperativity, and the folding funnels of one-dimensional architectures. *PLoS Comput Biol*, 4(5):e1000070, 2008.
- Diego U Ferreira, Elizabeth A Komives, y Peter G Wolynes. Frustration in biomolecules. *Quarterly reviews of biophysics*, 47(04):285–363, 2014.
- Alan R Fersht. Nucleation mechanisms in protein folding. *Current opinion in structural biology*, 7(1):3–9, 1997.
- Matteo Figliuzzi, Hervé Jacquier, Alexander Schug, Oliver Tenaillon, y Martin Weigt. Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase tem-1. *Molecular biology and evolution*, page msv211, 2015.
- Kathryn Geiger-Schuller y Doug Barrick. Broken tales: Transcription activator-like effectors populate partly folded states. *Biophysical Journal*, 111(11):2395–2403, 2016.
- Steven Henikoff y Jorja G Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 1992.
- Steven Henikoff y Jorja G Henikoff. Position-based sequence weights. *Journal of molecular biology*, 243(4):574–578, 1994.
- Thomas A Hopf, John B Ingraham, Frank J Poelwijk, Michael Springer, Chris Sander, y Debora S Marks. Quantification of the effect of mutations using a global probability model of natural sequence variation. *arXiv preprint arXiv:1510.04612*, 2015.

<http://pdb.org/>. Protein data base.

- William Humphrey, Andrew Dalke, y Klaus Schulten. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.
- Biman Jana, Faruck Morcos, y José N Onuchic. From structure to function: the convergence of structure based models and co-evolutionary information. *Physical Chemistry Chemical Physics*, 16(14):6496–6507, 2014.
- Yalda Javadi y Laura S Itzhaki. Tandem-repeat proteins: regularity plus modularity equals design-ability. *Current opinion in structural biology*, 23(4):622–631, 2013.
- Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.
- Julien Jorda, Bin Xue, Vladimir N Uversky, y Andrey V Kajava. Protein tandem repeats—the more perfect, the less structured. *FEBS journal*, 277(12):2673–2682, 2010.
- Tommi Kajander, Aitziber L Cortajarena, Simon Mochrie, y Lynne Regan. Structure and stability of designed tpr protein superhelices: unusual crystal packing and implications for natural tpr proteins. *Acta Crystallographica Section D: Biological Crystallography*, 63(7):800–811, 2007.
- Hetunandan Kamisetty, Sergey Ovchinnikov, y David Baker. Assessing the utility of coevolution-based residue–residue contact predictions in a sequence-and structure-rich era. *Proceedings of the National Academy of Sciences*, 110(39):15674–15679, 2013.
- George A Khoury, James Smadbeck, Chris A Kieslich, y Christodoulos A Floudas. Protein folding and de novo protein design for biotechnological applications. *Trends in biotechnology*, 32(2):99–109, 2014.
- David E Kim, Frank DiMaio, Ray Yu-Ruei Wang, Yifan Song, y David Baker. One contact for every twelve residues allows robust and accurate topology-level protein structure modeling. *Proteins: Structure, Function, and Bioinformatics*, 82(S2):208–218, 2014.
- Ellen Kloss, Naomi Courtemanche, y Doug Barrick. Repeat-protein folding: new insights into origins of cooperativity, stability, and topology. *Archives of biochemistry and biophysics*, 469(1):83–99, 2008.

- Bostjan Kobe y Andrey V Kajava. The leucine-rich repeat as a protein recognition motif. *Current opinion in structural biology*, 11(6):725–732, 2001.
- BT Korber, Robert M Farber, David H Wolpert, y Alan S Lapedes. Covariation of mutations in the v3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. *Proceedings of the National Academy of Sciences*, 90(15):7176–7180, 1993.
- Anne Marie Krachler, Amit Sharma, y Colin Kleanthous. Self-association of tpr domains: Lessons learned from a designed, consensus-based tpr oligomer. *Proteins: Structure, Function, and Bioinformatics*, 78(9):2131–2143, 2010.
- Andrew Leaver-Fay, Michael Tyka, Steven M Lewis, Oliver F Lange, James Thompson, Ron Jacak, Kristian Kaufman, P Douglas Renfrew, Colin A Smith, Will Sheffler, et al. Rosetta3: an object-oriented software suite for the simulation and design of macromolecules. *Methods in enzymology*, 487:545, 2011.
- Cyrus Levinthal. Are there pathways for protein folding. *J. Chim. phys*, 65(1):44–45, 1968.
- Ronald M Levy, Allan Haldane, y William F Flynn. Potts hamiltonian models of protein co-variation, free energy landscapes, and evolutionary fitness. *Current Opinion in Structural Biology*, 43:55–62, 2017.
- Weizhong Li, Lukasz Jaroszewski, y Adam Godzik. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, 17(3):282–283, 2001.
- Weizhong Li, Lukasz Jaroszewski, y Adam Godzik. Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics*, 18(1):77–82, 2002.
- Alan R Lowe y Laura S Itzhaki. Rational redesign of the folding pathway of a modular protein. *Proceedings of the National Academy of Sciences*, 104(8):2679–2684, 2007.
- Jianzhu Ma, Sheng Wang, Feng Zhao, y Jinbo Xu. Protein threading using context-specific alignment potential. *Bioinformatics*, 29(13):i257–i265, 2013.

- Ewan RG Main, Sophie E Jackson, y Lynne Regan. The folding and design of repeat proteins: reaching a consensus. *Current opinion in structural biology*, 13(4):482–489, 2003a.
- Ewan RG Main, Yong Xiong, Melanie J Cocco, Luca D’Andrea, y Lynne Regan. Design of stable α -helical arrays from an idealized tpr motif. *Structure*, 11(5):497–508, 2003b.
- Debora S Marks, Lucy J Colwell, Robert Sheridan, Thomas A Hopf, Andrea Pagnani, Riccardo Zecchina, y Chris Sander. Protein 3d structure computed from evolutionary sequence variation. *PloS one*, 6(12):e28766, 2011.
- Debora S Marks, Thomas A Hopf, y Chris Sander. Protein structure prediction from sequence variation. *Nature biotechnology*, 30(11):1072–1080, 2012.
- Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S Marks, Chris Sander, Riccardo Zecchina, José N Onuchic, Terence Hwa, y Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301, 2011.
- Faruck Morcos, Biman Jana, Terence Hwa, y José N Onuchic. Coevolutionary signals across protein lineages help capture multiple protein conformations. *Proceedings of the National Academy of Sciences*, 110(51):20533–20538, 2013.
- Leila K Mosavi, Daniel L Minor, y Zheng-yu Peng. Consensus-derived structural determinants of the ankyrin repeat motif. *Proceedings of the National Academy of Sciences*, 99(25):16029–16034, 2002.
- Athi N Naganathan y Victor Muñoz. Scaling of folding times with protein size. *Journal of the American Chemical Society*, 127(2):480–481, 2005.
- José Nelson Onuchic y Peter G Wolynes. Theory of protein folding. *Current opinion in structural biology*, 14(1):70–75, 2004.
- Fabio Parmeggiani, Riccardo Pellarin, Anders Peter Larsen, Gautham Varadamsetty, Michael T Stump, Oliver Zerbe, Amedeo Caflisch, y Andreas Plückthun. Designed armadillo repeat proteins as general peptide-binding scaffolds: consensus design and computational

- optimization of the hydrophobic core. *Journal of molecular biology*, 376(5):1282–1304, 2008.
- R Gonzalo Parra, Rocío Espada, Ignacio E Sánchez, Manfred J Sippl, y Diego U Ferreiro. Detecting repetitions and periodicities in proteins by tiling the structural space. *The Journal of Physical Chemistry B*, 117(42):12887–12897, 2013.
- R Gonzalo Parra, Rocío Espada, Nina Verstraete, y Diego U Ferreiro. Structural and energetic characterization of the ankyrin repeat protein family. *PLoS Comput Biol*, 11(12):e1004659, 2015.
- Raj Kumar Pathria. *Statistical mechanics*. Elsevier, 1972.
- Kris Pauwels, Inge Van Molle, Jan Tommassen, y Patrick Van Gelder. Chaperoning anfinzen: the steric foldases. *Molecular microbiology*, 64(4):917–922, 2007.
- Andreas Plückthun. Designed ankyrin repeat proteins (darpins): binding proteins for research, diagnostics, and therapy. *Annual review of pharmacology and toxicology*, 55:489–511, 2015.
- Pamela JE Rowling, Elin M Sivertsson, Albert Perez-Riba, Ewan RG Main, y Laura S Itzhaki. Dissecting and reprogramming the folding and assembly of tandem-repeat proteins. *Biochemical Society Transactions*, 43(5):881–888, 2015.
- Andrej Sali, Eugene Shakhnovich, y Martin Karplus. How does a protein fold? *nature*, 369:248–250, 1994.
- Steven G Sedgwick y Stephen J Smerdon. The ankyrin repeat: a diversity of interactions on a common structural framework. *Trends in biochemical sciences*, 24(8):311–316, 1999.
- Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.
- James A Shapiro. How life changes itself: The read-write (rw) genome. *Physics of Life reviews*, 10(3):287–323, 2013.

- Temple F Smith y Michael S Waterman. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, 1981.
- Michael Socolich, Steve W Lockless, William P Russ, Heather Lee, Kevin H Gardner, y Rama Ranganathan. Evolutionary information for specifying a protein fold. *Nature*, 437(7058):512–518, 2005.
- Johannes Söding, Andreas Biegert, y Andrei N Lupas. The hhpred interactive server for protein homology detection and structure prediction. *Nucleic acids research*, 33(suppl 2):W244–W248, 2005.
- Christian U Stirnimann, Evangelia Petsalaki, Robert B Russell, y Christoph W Müller. Wd40 proteins propel cellular networks. *Trends in biochemical sciences*, 35(10):565–574, 2010.
- Timothy O Street y Doug Barrick. Predicting repeat protein folding kinetics from an experimentally determined folding energy landscape. *Protein Science*, 18(1):58–68, 2009.
- Timothy O Street, Christina Marchetti Bradley, y Doug Barrick. An improved experimental system for determining small folding entropy changes resulting from proline to alanine substitutions. *Protein science*, 14(9):2429–2435, 2005.
- Michael T Stumpp, Patrik Forrer, H Kaspar Binz, y Andreas Plückthun. Designing repeat proteins: modular leucine-rich repeat protein libraries based on the mammalian ribonuclease inhibitor family. *Journal of molecular biology*, 332(2):471–487, 2003.
- Joanna I Sułkowska, Faruck Morcos, Martin Weigt, Terence Hwa, y José N Onuchic. Genomics-aided structure prediction. *Proceedings of the National Academy of Sciences*, 109(26):10340–10345, 2012.
- Ludovico Sutto, Simone Marsili, Alfonso Valencia, y Francesco Luigi Gervasio. From residue coevolution to protein conformational ensembles and functional dynamics. *Proceedings of the National Academy of Sciences*, 112(44):13567–13572, 2015.
- Kit S Tang, Alan R Fersht, y Laura S Itzhaki. Sequential unfolding of ankyrin repeats in tumor suppressor p16. *Structure*, 11(1):67–73, 2003.

- Agathe Urvoas, Asma Guellouz, Marie Valerio-Lepiniec, Marc Graille, Dominique Durand, Danielle C Desravines, Herman van Tilbeurgh, Michel Desmadril, y Philippe Minard. Design, production and molecular structure of a new family of artificial alpha-helical repeat proteins (α rep) based on thermostable heat-like repeats. *Journal of molecular biology*, 404(2):307–327, 2010.
- Vladimir N Uversky. Intrinsically disordered proteins from a to z. *The international journal of biochemistry & cell biology*, 43(8):1090–1103, 2011.
- Julie S Valastyan y Susan Lindquist. Mechanisms of protein-folding diseases at a glance. *Disease Models and Mechanisms*, 7(1):9–14, 2014.
- Marie Valerio-Lepiniec, Agathe Urvoas, Anne Chevrel, Asma Guellouz, Yann Ferrandez, Agnès Mesneau, Ines Li De La Sierra-Gallay, Magali Aumont-Nicaise, Michel Desmadril, Herman Van Tilbeurgh, et al. The α rep artificial repeat protein scaffold: a new tool for crystallization and live cell applications. *Biochemical Society Transactions*, 43(5):819–824, 2015.
- Gautham Varadamsetty, Dirk Tremmel, Simon Hansen, Fabio Parmeggiani, y Andreas Plückthun. Designed armadillo repeat proteins: library generation, characterization and selection of peptide binders with high specificity. *Journal of molecular biology*, 424(1):68–87, 2012.
- Martin Weigt, Robert A White, Hendrik Szurmant, James A Hoch, y Terence Hwa. Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences*, 106(1):67–72, 2009.
- Svava K Wetzel, Giovanni Settanni, Manca Kenig, H Kaspar Binz, y Andreas Plückthun. Folding and unfolding mechanism of highly stable full-consensus ankyrin repeat proteins. *Journal of molecular biology*, 376(1):241–257, 2008.
- Fa-Yueh Wu. The potts model. *Reviews of modern physics*, 54(1):235, 1982.
- Yang Zhang. I-tasser server for protein 3d structure prediction. *BMC bioinformatics*, 9(1):40, 2008.

Hongbo Zhu, Edgardo Sepulveda, Marcus D Hartmann, Manjunatha Kogenaru, Astrid Ursinus, Eva Sulz, Reinhard Albrecht, Murray Coles, Jörg Martin, y Andrei N Lupas. Origin of a folded repeat protein from an intrinsically disordered ancestor. *eLife*, 5:e16761, 2016.

Royce KP Zia, Edward F Redish, y Susan R McKay. Making sense of the legendre transform. *American Journal of Physics*, 77(7):614–622, 2009.

Apéndice A

Derivación del modelo estadístico

Modelo estadístico

El objetivo es modelar la probabilidad de ocurrencia de las secuencias proteicas, $P(\vec{s})$. Cada proteína es representada como una secuencia de aminoácidos $\vec{s} = a_1, \dots, a_n$, donde los a_i puede tomar $q = 21$ valores distintos: cada uno de los aminoácidos o un espacio vacío, es decir $a_i \in \{A, C, D, E, F, G, H, I, K, L, M, N, O, P, Q, R, S, T, V, W, Y, -\}$. El modelo se basa en un conjunto de M secuencias, todas del mismo largo L . Cada secuencias tiene asociada una probabilidad de ocurrencia $P(\vec{s})$, cuya expresión es desconocida. El modelo debe cumplir con las siguientes restricciones obtenidas de secuencias reales: las probabilidades marginales y conjuntas de ocurrencia de todos los aminoácidos (a_i y b_j), a lo largo de todas las posiciones de la secuencia (i y j). Estas restricciones se pueden formular como:

$$P_i(a_i) \equiv \sum_{a_k, k \neq i} P(\vec{s}) = f_i(a_i) \quad (\text{A.1})$$

$$P_{ij}(a_i, b_j) \equiv \sum_{a_k, k \neq i, j} P(\vec{s}) = f_{ij}(a_i, b_j) \quad (\text{A.2})$$

donde $P_i(a_i)$ y $P_{ij}(a_i, b_j)$ se refieren a los valores de probabilidad determinados por el modelo, mientras que $f_i(a_i)$ y $f_{ij}(a_i, b_j)$ son los valores medidos experimentalmente.

Para obtener un modelo explícito de las probabilidades de ocurrencia de secuencias, utilicé el método de maximización de entropía sujeto a las restricciones enumeradas. Este procedimiento fue desarrollado en [Jaynes, 1957] y lo describo a continuación.

Antecedentes teóricos: Estimación de máxima entropía

Es un problema usual querer determinar la expresión explícita de una distribución de probabilidades, restringida por observables experimentales, con el menor sesgo posible. La metodología para resolver esta problemática fue determinada por Jaynes [1957]. Primero reproduzco la metodología teórica general, para luego aplicarla al problema de secuencias proteicas.

Dada una magnitud discreta x_i , con $i = 1...n$ es de interés conocer la probabilidad de ocurrencia de cada posible valor p_i . La información disponible es el valor esperado de una magnitud $g(x)$, relacionada con las probabilidades según

$$\langle g(x) \rangle = \sum_{i=1}^n p_i g(x_i) \quad (\text{A.3})$$

y por supuesto la condición de normalización

$$\sum_{i=1}^n p_i = 1 \quad (\text{A.4})$$

En su trabajo, Jaynes [1957] mostró que la expresión para la probabilidad está dada por

$$p_i = e^{-\lambda - \mu g(x_i)} \quad (\text{A.5})$$

donde las constantes están determinadas por:

$$Z = \sum_{i=1}^n e^{-\frac{\mu g(x_i)}{K}} \quad (\text{A.6})$$

$$\lambda = K \ln(Z) \quad (\text{A.7})$$

$$\langle g(x) \rangle = \frac{1}{Z} \sum_{i=1}^n \left[e^{-\frac{\mu g(x_i)}{K}} \right] g(x_i) \quad (\text{A.8})$$

con K una constante arbitraria que en este trabajo se tomó igual a 1.

Demostración

Para encontrar esta expresión, Jaynes se basó en la incerteza para la distribución de probabilidades encontrada por Shannon [1948]:

$$H(p_1, \dots, p_n) = -K \sum_i p_i \ln p_i \quad (\text{A.9})$$

que, dada su similitud con la ecuación de Boltzmann utilizada en mecánica estadística, la nombra equivalente a la entropía de la distribución de probabilidades p_i . En su trabajo, Jaynes [1957] propone aplicar el método de multiplicadores de Lagrange para maximizar la expresión de la entropía sujeto a las restricciones de las ecuaciones A.3 y A.4. La cantidad a maximizar es:

$$I(p_1, \dots, p_n, \mu, \lambda) = H(p_1, \dots, p_n) + \mu \left[\langle g(x) \rangle - \sum_{i=1}^n p_i g(x_i) \right] + \lambda \left[1 - \sum_{i=1}^n p_i \right] \quad (\text{A.10})$$

Maximizar esta expresión implica pedir que

$$\nabla I = 0 \quad (\text{A.11})$$

lo que deriva en particular a la ecuación:

$$\frac{\partial I}{\partial p_i} = 0 \quad (\text{A.12})$$

$$-K (\ln p_i + 1) - \mu g(x_i) - \lambda = 0 \quad (\text{A.13})$$

$$p_i = e^{-\frac{\mu g(x_i) + \lambda}{K} - 1} \quad (\text{A.14})$$

$$p_i = e^{-\frac{\lambda}{K}} e^{-\frac{\mu g(x_i)}{K}} e^{-1} \quad (\text{A.15})$$

$$p_i = e^{-\frac{\lambda_0}{K}} e^{-\frac{\mu g(x_i)}{K}} \quad (\text{A.16})$$

En el último paso definí, sin pérdida de generalidad, $\lambda_0 = \lambda + K$. Reemplazando la última expresión en la condición de normalización (ecuación A.4) se obtiene:

$$\sum_{i=1}^n e^{-\frac{\lambda_0}{K}} e^{-\frac{\mu g(x_i)}{K}} = 1 \quad (\text{A.17})$$

$$e^{-\frac{\lambda_0}{K}} \sum_{i=1}^n e^{-\frac{\mu g(x_i)}{K}} = 1 \quad (\text{A.18})$$

$$e^{-\frac{\lambda_0}{K}} = \frac{1}{\sum_{i=1}^n e^{-\frac{\mu g(x_i)}{K}}} \quad (\text{A.19})$$

$$\lambda_0 = -K \ln \left(\frac{1}{\sum_{i=1}^n e^{-\frac{\mu g(x_i)}{K}}} \right) \quad (\text{A.20})$$

$$\lambda_0 = K \ln \left(\sum_{i=1}^n e^{-\frac{\mu g(x_i)}{K}} \right) \quad (\text{A.21})$$

Si defino

$$Z = \sum_{i=1}^n e^{-\frac{\mu g(x_i)}{K}} \quad (\text{A.22})$$

se obtiene finalmente

$$\lambda_0 = K \ln(Z) \quad (\text{A.23})$$

Por otro lado, reemplazando en la ecuación A.3,

$$\langle g(x) \rangle = \sum_{i=1}^n \left[e^{-\frac{\lambda_0}{K}} e^{-\frac{\mu g(x_i)}{K}} \right] g(x_i) \quad (\text{A.24})$$

$$= e^{-\frac{\lambda_0}{K}} \sum_{i=1}^n g(x_i) e^{-\frac{\mu g(x_i)}{K}} \quad (\text{A.25})$$

Utilizando la ecuación A.23 para sustituir λ_0 ,

$$\langle f(x) \rangle = e^{-\frac{K \ln(Z)}{K}} \sum_{i=1}^n \left[e^{-\frac{\mu g(x_i)}{K}} \right] g(x_i) \quad (\text{A.26})$$

$$= e^{-\ln(Z)} \sum_{i=1}^n \left[e^{-\frac{\mu g(x_i)}{K}} \right] g(x_i) \quad (\text{A.27})$$

$$= \frac{1}{Z} \sum_{i=1}^n \left[e^{-\frac{\mu g(x_i)}{K}} \right] g(x_i) \quad (\text{A.28})$$

$$(\text{A.29})$$

que se puede reescribir como

$$\langle g(x) \rangle = -\frac{\partial \ln Z}{\partial \mu} \quad (\text{A.30})$$

■

Estas expresiones pueden ser generalizadas para un conjunto de funciones:

$$\langle g_r(x) \rangle = \sum_i p_i g_r(x_i) \quad (\text{A.31})$$

obteniendo la función de partición

$$Z(\lambda_1, \dots, \lambda_r) = \sum_i e^{-[\lambda_1 g_1(x_i) + \dots + \lambda_r g_r(x_i)]}. \quad (\text{A.32})$$

La distribución de probabilidades asociada es

$$p_i = e^{-[\lambda_0 + \lambda_1 g_i(x_i) + \dots + \lambda_m g_m(x_i)]} \quad (\text{A.33})$$

donde las constantes se encuentran determinadas por

$$\langle g_r(x) \rangle = -\frac{\partial}{\partial \lambda_r} \ln Z \quad (\text{A.34})$$

$$\lambda_0 = \ln Z \quad (\text{A.35})$$

Modelo para proteínas

En el caso de las proteínas, la variable es cada secuencia posible de aminoácidos. Los observables utilizados para restringir el modelo son las frecuencias de ocurrencia de los aminoácidos. Para esto, las frecuencias se expresan como los valores medio de una función partida:

$$\langle f_i(a_i) \rangle = \sum_{\{\vec{s}\}} P(\vec{s}) \theta_i(a_i) \quad (\text{A.36})$$

donde $\langle f_i(a_i) \rangle$ es la frecuencia esperada del aminoácido a_i en la posición i , y

$$\theta_i(a_i) = \begin{cases} 1 & \text{si } \vec{s} \text{ tiene } a_i \text{ en } i \\ 0 & \text{sino} \end{cases} \quad (\text{A.37})$$

Análogamente, para las frecuencias conjuntas:

$$\langle f_{ij}(a_i, b_j) \rangle = \sum_{\{\vec{s}\}} P(\vec{s}) \theta_i(a_i) \theta_j(b_j) \quad (\text{A.38})$$

La condición de normalización se impone sobre la probabilidad de ocurrencia de cada secuencia:

$$\sum_{\vec{s}} P(\vec{s}) = 1 \quad (\text{A.39})$$

Finalmente, la expresión a maximizar es

$$\begin{aligned}
I = & - \sum_{\{\vec{s}\}} P(\vec{s}) \ln P(\vec{s}) + \lambda_0 \left(\sum_{\vec{s}} P(\vec{s}) - 1 \right) + \sum_i^L \sum_{a_i}^q \mu_{ia_i} \left(\sum_{\{\vec{s}\}} P(\vec{s}) \theta_i(a_i) - \langle f_i(a_i) \rangle \right) \\
& + \sum_i^L \sum_{j \neq i}^L \sum_{a_i}^q \sum_{b_j}^q \gamma_{ija_i b_j} \left(\sum_{\{\vec{s}\}} P(\vec{s}) \theta_i(a_i) \theta_j(b_j) - \langle f_{ij}(a_i, b_j) \rangle \right) \quad (\text{A.40})
\end{aligned}$$

donde explicité el largo de las proteínas L , y la cantidad de estados posibles $q = 21$. Además, excluí los parámetros de interacción de una posición contra sígo misma, excluyendo en el último término las sumas de $j = i$. Al maximizar, es decir derivar e igualar a cero, la condición relevante resulta

$$\begin{aligned}
\frac{\partial I}{\partial P(\vec{s}')} = & - \sum_{\{\vec{s}\}} \left(\frac{dP(\vec{s})}{dP(\vec{s}')} + \frac{dP(\vec{s})}{dP(\vec{s}')} \ln P(\vec{s}) \right) + \lambda_0 \sum_{\{\vec{s}\}} \frac{dP(\vec{s})}{dP(\vec{s}')} \\
& + \sum_i^L \sum_{a_i}^q \mu_{ia_i} \sum_{\{\vec{s}\}} \frac{dP(\vec{s})}{dP(\vec{s}')} \theta_i(a_i) \\
& + \sum_i^L \sum_{j \neq i}^L \sum_{a_i}^q \sum_{b_j}^q \gamma_{ija_i b_j} \sum_{\{\vec{s}\}} \frac{dP(\vec{s})}{dP(\vec{s}')} \theta_i(a_i) \theta_j(b_j) = 0 \quad (\text{A.41})
\end{aligned}$$

Para resolver, utilizo que $\frac{dP(\vec{s})}{dP(\vec{s}')} = \delta_{\vec{s}, \vec{s}'}$, con δ la delta de Kronecker; es decir que la derivada es nula a menos que $\vec{s}$ coincida con $\vec{s}'$. Por lo tanto, las sumatorias sobre $\vec{s}$ se reducen:

$$\begin{aligned}
\frac{\partial I}{\partial P(\vec{s}')} = & - \left(1 + \ln P(\vec{s}') \right) + \lambda_0 \\
& + \sum_i^L \sum_{a_i}^q \mu_{ia_i} \delta_{\vec{s}, \vec{s}'} \theta_i(a_i) \\
& + \sum_i^L \sum_{j \neq i}^L \sum_{a_i}^q \sum_{b_j}^q \gamma_{ija_i b_j} \delta_{\vec{s}, \vec{s}'} \theta_i(a_i) \theta_j(b_j) = 0 \quad (\text{A.42})
\end{aligned}$$

Debido a la definición de la función θ , en las sumatorias sobre los aminoácidos (u estados),

sobreviven únicamente los términos asociados a los aminoácidos es la secuencia $\vec{s'}$

$$\frac{\partial I}{\partial P(\vec{s'})} = - \left(1 + \ln P(\vec{s'}) \right) + \lambda_0 + \sum_i^L \mu_{ia_i} \theta_i(a_i) + \sum_i^L \sum_{j \neq i}^L \gamma_{ija_i b_j} \theta_i(a_i) \theta_j(a_j) = 0 \quad (\text{A.43})$$

donde mantuve las θ para explicitar que las sumas se realizan únicamente sobre los aminoácidos presentes en la secuencia $\vec{s'}$.

En analogía con las ecuaciones A.16 y A.33, defino $h_i(a_i) = \mu_{ia_i} \theta_i(a_i)$ y $j_{ij}(a_i, b_j) = 2\gamma_{ija_i b_j} \theta_i(a_i) \theta_j(a_j)$, cambio la notación $\vec{s'}$ por $\vec{s}$ y obtengo la siguiente expresión para la probabilidad de cada secuencia:

$$P(\vec{s}) = \frac{1}{Z} \exp \left[- \sum_i^L h_i(a_i) - \sum_i^L \sum_{j > i}^L j_{ij}(a_i, b_j) \right] \quad (\text{A.44})$$

donde Z es el factor de normalización, análogamente a A.22 y A.32:

$$Z = \sum_{\{\vec{s}\}} \exp \left[- \sum_i^L h_i(a_i) - \sum_i^L \sum_{j > i}^L j_{ij}(a_i, b_j) \right] \quad (\text{A.45})$$

Estas expresiones remiten a un modelo de Potts [Wu, 1982], que permite enlazar las expresiones a la mecánica estadística. Se puede asociar el sistema a un ensamble canónico (en equilibrio térmico) [Pathria, 1972], cuyo Hamiltoniano es

$$H(\vec{s}) = \sum_i^L h_i(a_i) + \sum_i^L \sum_{j > i}^L j_{ij}(a_i, b_j) \quad (\text{A.46})$$

donde los parámetros $h_i(a_i)$ son los campos locales y $j_{ij}(a_i, b_j)$ las energías de interacción. La distribución de probabilidades se encuentra entonces determinada por la distribución de Boltzmann-Gibbs

$$P(\vec{s}) = \frac{1}{Z} e^{-\beta H(\vec{s})} \quad (\text{A.47})$$

con $\beta = 1/k_B T$, k_B es la constante de Boltzmann y T la temperatura del sistema. La función

de partición se relaciona con la energía libre de Helmholtz:

$$F = -k_B T \ln Z \quad (\text{A.48})$$

Usualmente a lo largo de este trabajo tomé $\beta = 1$. A partir de la función de partición se obtiene toda la información relevante del sistema:

$$\frac{\partial F}{\partial h_i(a_i)} = \frac{\partial(-\ln Z)}{\partial h_i(a_i)} = P_i(a_i) \quad (\text{A.49})$$

$$\frac{\partial^2 F}{\partial h_i(a_i) \partial h_j(b_j)} = \frac{\partial^2(-\ln Z)}{\partial h_i(a_i) \partial h_j(b_j)} = -P_{ij}(a_i, b_j) + P_i(a_i)P_j(b_j) = -C_{ij}(a_i, b_j) \quad (\text{A.50})$$

En la última igualdad definí la matriz de correlaciones conectadas $C_{ij}(a_i, b_j) = P_{ij}(a_i, b_j) - P_i(a_i)P_j(b_j)$.

Demostración ecuaciones A.49 y A.50

Para la primer ecuación uso la regla de la cadena para derivar funciones compuestas:

$$\frac{\partial \ln Z}{\partial h_i(a_i)} = \frac{1}{Z} \frac{\partial Z}{\partial h_i(a_i)} \quad (\text{A.51})$$

$$= \frac{1}{Z} \frac{\partial}{\partial h_i(a_i)} \left[\sum_{\vec{s}} e^{-\beta H(\vec{s})} \right] \quad (\text{A.52})$$

$$= \frac{1}{Z} \sum_{\vec{s}} \frac{\partial e^{-\beta H(\vec{s})}}{\partial h_i(a_i)} \quad (\text{A.53})$$

$$= \frac{1}{Z} \sum_{\vec{s}} -\beta e^{-\beta H(\vec{s})} \frac{\partial H(\vec{s})}{\partial h_i(a_i)} \quad (\text{A.54})$$

Esta derivada da 1 si el estado $\vec{s}$ tiene a_i en la posición i -ésima, y 0 sino. Por lo tanto, la suma se ve reducida únicamente a estos estados:

$$\frac{\partial \ln Z}{\partial h_i(a_i)} = -\frac{\beta}{Z} \sum_{a_k, k \neq i} e^{-\beta H(\vec{s})} \quad (\text{A.55})$$

$$= -\beta P_i(a_i) \quad (\text{A.56})$$

en la última igualdad usé la definición de la ecuación A.1.

Para la siguiente ecuación uso la expresión A.56 y la regla de la derivación del cociente:

$$\frac{\partial^2 \ln Z}{\partial h_i(a_i) \partial h_j(b_j)} = \frac{\partial}{\partial h_j(b_j)} \left[-\frac{\beta}{Z} \sum_{a_k, k \neq i} e^{-\beta H(\vec{s})} \right] \quad (\text{A.57})$$

$$= -\frac{\beta}{Z^2} \left[Z \frac{\partial}{\partial h_j(b_j)} \left(\sum_{a_k, k \neq i} e^{-\beta H(\vec{s})} \right) - \sum_{a_k, k \neq i} e^{-\beta H(\vec{s})} \frac{\partial Z}{\partial h_j(b_j)} \right] \quad (\text{A.58})$$

$$= -\frac{\beta}{Z^2} \left[Z \sum_{a_k, k \neq i} -\beta e^{-\beta H(\vec{s})} \frac{\partial H(\vec{s})}{\partial h_j(b_j)} - \sum_{a_k, k \neq i} e^{-\beta H(\vec{s})} \frac{\partial Z}{\partial h_j(b_j)} \right] \quad (\text{A.59})$$

La derivada del segundo término es análogo al de la ecuación A.51, cambiando el nombre de la variable $h_i(a_i)$ a $h_j(b_j)$. Por lo tanto, junto a la expresión A.55 es posible despejar $\frac{\partial Z}{\partial h_j(b_j)} = -\beta Z P_j(b_j)$. La derivada del primer término conlleva el mismo razonamiento que el paso entre las ecuaciones A.54 y A.56, y por lo tanto reduce la suma a los términos que contienen simultáneamente a_i y b_j . Incorporando estos razonamientos se obtiene:

$$\frac{\partial^2 \ln Z}{\partial h_i(a_i) \partial h_j(b_j)} = -\frac{\beta}{Z^2} \left[Z \sum_{a_k, k \neq i, j} -\beta e^{-\beta H(\vec{s})} - \sum_{a_k, k \neq i} e^{-\beta H(\vec{s})} (-\beta) Z P_j(a_j) \right] \quad (\text{A.60})$$

$$= \frac{\beta^2}{Z} \left[\underbrace{\sum_{a_k, k \neq i, j} e^{-\beta H(\vec{s})}}_{Z P_{ij}(a_i, b_j)} - \underbrace{\sum_{a_k, k \neq i} e^{-\beta H(\vec{s})} P_j(a_j)}_{Z P_i(a_i)} \right] \quad (\text{A.61})$$

donde los valores indicados debajo de las llaves se deben a las definiciones de las ecuaciones A.1 y A.2. Acomodando la expresión finalmente se obtiene:

$$\frac{\partial^2 \ln Z}{\partial h_i(a_i) \partial h_j(b_j)} = \beta^2 (P_{ij}(a_i, b_j) - P_i(a_i) P_j(a_j)) \quad (\text{A.62})$$

Tomando $\beta = 1$ en las ecuaciones A.55 y A.62 se obtienen las expresiones A.49 y A.50. ■

Gauge

El modelo de la ecuación A.46 tiene $Lq + \binom{L}{2} q^2$ parámetros, pero no son todos independientes entre sí. Esto se debe a que las condiciones impuestas para restringir el modelo, ecuaciones A.1 y A.2, no son independientes dado que las probabilidades deben estar normalizadas:

$$\sum_{a_i=1}^q P_i(a_i) = 1 \quad (\text{A.63})$$

$$\sum_{b_j=1}^q \sum_{a_i=1}^q P_{ij}(a_i, b_j) = 1 \quad (\text{A.64})$$

La cantidad de parámetros independientes son $L(q-1) + \binom{L}{2}(q-1)^2$. Elijo, sin perder generalidad, fijar todos los parámetros relacionados al espacio vacío a cero: $h_i(1) = 0$ $j_{ij}(a_i, q) = j_{ij}(q, a_i) = 0$. El número de parámetros libres se corresponden al número de restricciones independientes, garantizando una solución única al modelo derivado de la ma-

ximización de la entropía.

Además, notar que los acoplamientos no tienen sentido en una única posición, como no lo tienen las probabilidades $P_{ii}(a_i, b_i)$. Por lo tanto $j_{ii}(a_i, b_i) = 0$.

Modelo alternativo para proteínas repetitivas

Las proteínas repetitivas presentan alta similitud en secuencia entre repeticiones sucesivas. En este trabajo utilicé un modelo que considere esta propiedad. La restricción impuesta es la frecuencia en que estas repeticiones tienen una determinada identidad en secuencia (f_{id}). Esta propiedad es introducida en completa analogía con las frecuencias marginales (ecuaciones A.36 y A.37):

$$f_{\text{id}} = \sum_{\{\vec{s}\}} P(\vec{s}) \phi(\text{id}) \quad (\text{A.65})$$

con

$$\phi(\text{id}) = \begin{cases} 1 & \text{si la identidad entre las repeticiones es id} \\ 0 & \text{sino} \end{cases} \quad (\text{A.66})$$

La expresión a maximizar (ecuación A.40) incorpora los términos

$$I + \lambda_{\text{id}} \left(\sum_{\{\vec{s}\}} P(\vec{s}) \phi(\text{id}) - f_{\text{id}} \right) \quad (\text{A.67})$$

Reproduciendo la maximización se obtiene el campo:

$$H(\vec{s}) = \sum_i^L h_i(a_i) + \sum_i^L \sum_{j>i}^L j_{ij}(a_i, b_j) + \lambda_{\text{id}} \quad (\text{A.68})$$

donde únicamente se suma el parámetro λ_{id} correspondiente a la identidad entre las repeticiones que constituyen la secuencia $\vec{s}$.

Aproximación de campo medio

Analizar la función de partición A.45 es intratable en la mayoría de los casos dado que el espacio de secuencias $\{\vec{s}\}$ sobre el que se computa la suma es demasiado grande. Por ejemplo, para secuencias de largo 5, la cantidad de variantes supera los 4 millones. En Morcos et al. [2011] los autores desarrollan una aproximación de campo medio que permite reducir el sistema de análisis y derivar una expresión de correlación entre posiciones de una proteína.

Para desarrollarla, primero explicitaré conceptos generales de la Transformada de Legendre, en la que se basa la aproximación.

Transformada de Legendre [Zia et al., 2009]

Definición 1D Dada una función $f : X \rightarrow \mathfrak{R}$, donde $X \in \mathfrak{R}$, $f(x)$ es estrictamente convexa y suave, la transformada de Legendre es aquella función que verifica

$$\frac{df}{dx} = \left(\frac{df^*}{dy}\right)^{-1} \quad (\text{A.69})$$

es decir, que las derivadas de $f(x)$ coincidan con las inversas de las derivadas de la función transformada $f^*(y)$. Puede hallarse sencillamente con la expresión

$$f^*(y) = y \cdot x(y) - f(x(y)) \quad (\text{A.70})$$

donde

$$y = \frac{df(x)}{dx} \quad (\text{A.71})$$

Propiedad I La transformada de Legendre es su propia inversa. Es decir que si f^* es la transformada de $f(x)$, entonces

$$(f^*)^* = f(x) \quad (\text{A.72})$$

Lo que implica que

$$x = \frac{df^*(y)}{dy} \quad (\text{A.73})$$

Propiedad II Las segundas derivadas verifican

$$\frac{d^2 f}{dx^2} \frac{d^2 f^*}{dy^2} = 1 \quad (\text{A.74})$$

Extensión a más dimensiones Las expresiones anteriores se pueden generalizar a funciones en más dimensiones. Si $f : X \rightarrow \mathfrak{R}$, donde $X \in \mathfrak{R}^n$, la definición A.69 se puede generalizar a

$$Df = (Df^*)^{-1} \quad (\text{A.75})$$

Y la expresión A.70 se adapta a

$$f^*(\vec{y}) = \langle \vec{x} \cdot \vec{y} \rangle - f(\vec{x}(\vec{y})) \quad (\text{A.76})$$

con

$$x = Df^*(y) \quad (\text{A.77})$$

$$y = Df(x) \quad (\text{A.78})$$

Finalmente, la propiedad A.74 se generaliza a:

$$\sum_m \partial_l \partial_m f^*(\vec{s}) \cdot \partial_m \partial_n f(x) = \delta_{ln} \quad (\text{A.79})$$

con δ_{ln} la matriz identidad.

Aproximación

Para poder realizar la aproximación de campo medio es necesario explicitar que los acoplamientos entre residuos en la ecuación A.46 son una perturbación a los campos locales.

Para esto reescribo el campo como:

$$H = \sum_{i=1}^L h_i(a_i) + \alpha \sum_{i=1}^L \sum_{j=i}^L j_{ij}(a_i, a_j) \quad (\text{A.80})$$

donde agregué el parámetro α , que permite interpolar entre el modelo de sitio independiente (con $\alpha = 0$) y el modelo original (con $\alpha = 1$). Como expliqué anteriormente, este campo permite definir una función de partición Z y una energía libre dada por $F = -\ln(Z)$. Utilizando la definición de la transformada de Legendre, ecuación A.70, defino el potencial de Gibbs:

$$G(\alpha) = \ln(Z) + \sum_i^L \sum_{a_i=1}^{q-1} h_i(a_i) P_i(a_i) \quad (\text{A.81})$$

Mediante la ecuación A.73 la transformada de Legendre devuelve la relación:

$$h_i(a_i) = \frac{\partial G(\alpha)}{\partial P_i(a_i)} \quad (\text{A.82})$$

Y derivando una vez más

$$\frac{\partial h_i(a_i)}{\partial P_j(b_j)} = \frac{\partial^2 G(\alpha)}{\partial P_i(a_i) \partial P_j(b_j)} = \left[\frac{\partial^2 (-\ln Z)}{\partial h_i(a_i) \partial h_j(b_j)} \right]^{-1} = -(C_{ij}(a_i, b_j))^{-1} \quad (\text{A.83})$$

En la segunda igualdad, usé la propiedad A.74 de las transformadas de Legendre, para pasar de G a $F = -\ln Z$, y en la última usé el resultado de la ecuación A.50.

Aproximación para $\alpha \approx 0$

Si se desarrolla el potencial de Gibbs para acoplamientos débiles, es decir para $\alpha \approx 0$, a primer orden se obtiene la expresión:

$$G(\alpha) = G(0) + \left. \frac{dG(\alpha)}{d\alpha} \right|_{\alpha=0} \alpha + O(\alpha^2) \quad (\text{A.84})$$

El **orden cero** corresponde al caso de sitio independiente. Utilizando la definición de los potenciales, se obtiene que el potencial de Gibbs para sitios desacoplados iguala a la

entropía:

$$F = U - TS \quad (\text{A.85})$$

$$G = -F + \sum_{i=1}^L \sum_{a_i=1}^{q-1} h_i(a_i) P_i(a_i) = -U + TS + \sum_{i=1}^L \sum_{a_i=1}^{q-1} h_i(a_i) P_i(a_i) \quad (\text{A.86})$$

Dado que el último término es equivalente a la energía media del sistema desacoplado, tomando $T=1$, se obtiene $G = S$. Por el otro lado, para un modelo sitio independiente, la entropía puede escribirse sencillamente como:

$$G(\alpha = 0) = S = - \sum_{i=1}^L \sum_{a_i=1}^q P_i(a_i) \ln P_i(a_i) \quad (\text{A.87})$$

que puede reescribirse, separando los términos asociados al aminoácido (o estado) q , como:

$$G(\alpha = 0) = S = - \sum_{i=1}^L \sum_{a_i=1}^{q-1} P_i(a_i) \ln P_i(a_i) - \sum_{i=1}^L \left[1 - \sum_{a_i=1}^{q-1} P_i(a_i) \right] \ln \left(1 - \sum_{a_i=1}^{q-1} P_i(a_i) \right) \quad (\text{A.88})$$

El cálculo del **primer orden**:

$$\begin{aligned} \frac{dG(\alpha)}{d\alpha} &= \frac{d}{d\alpha} \ln Z(\alpha) + \sum_{i=1}^L \sum_{a_i=1}^{q-1} \frac{dh_i(a_i)}{d\alpha} P_i(a_i) \\ &= - \sum_{\{\vec{s}\}} \left[\sum_i \sum_{j>i} j_{ij}(a_i, b_j) + \sum_i \frac{dh_i(a_i)}{d\alpha} \right] \frac{e^{-H(\alpha)}}{Z(\alpha)} - \sum_{i=1}^L \sum_{a_i=1}^{q-1} \frac{dh_i(a_i)}{d\alpha} P_i(a_i) \\ &= - \sum_{\{\vec{s}\}} \sum_i \sum_{j>i} j_{ij}(a_i, b_j) \frac{e^{-H(\alpha)}}{Z(\alpha)} + \sum_{\{\vec{s}\}} \sum_i \frac{dh_i(a_i)}{d\alpha} \frac{e^{-H(\alpha)}}{Z(\alpha)} - \underbrace{\sum_{i=1}^L \sum_{a_i=1}^{q-1} \sum_{\{\vec{s}|a_k, k \neq i\}} \frac{dh_i(a_i)}{d\alpha} \frac{e^{-H(\alpha)}}{Z(\alpha)}}_{\sum_{\vec{s}}} \\ &= - \sum_{\{\vec{s}\}} \sum_i \sum_{j>i} j_{ij}(a_i, b_j) \frac{e^{-H(\alpha)}}{Z(\alpha)} + \sum_{\{\vec{s}\}} \sum_i \frac{dh_i(a_i)}{d\alpha} \frac{e^{-H(\alpha)}}{Z(\alpha)} - \sum_{i=1}^L \sum_{\vec{s}} \frac{dh_i(a_i)}{d\alpha} \frac{e^{-H(\alpha)}}{Z(\alpha)} \\ &= \left\langle \sum_i \sum_{j>i} j_{ij}(a_i, b_j) \right\rangle_{\alpha} \end{aligned}$$

Y nuevamente, al evaluar en $\alpha = 0$ nos remitimos al caso sitio independiente, en el cual el

valor medio de los parámetros de acoplamiento está dado por:

$$\left. \frac{dG(\alpha)}{d\alpha} \right|_{\alpha=0} = - \sum_i^L \sum_{j>i}^L \sum_{a_i}^{q-1} \sum_{b_j}^{q-1} j_{ij}(a_i, b_j) P_i(a_i) P_j(b_j) \quad (\text{A.89})$$

Incorporando las ecuaciones A.88 y A.89 en A.84, obtenemos:

$$G(\alpha) = - \sum_{i=1}^L \sum_{a_i=1}^{q-1} P_i(a_i) \ln P_i(a_i) \quad (\text{A.90})$$

$$- \sum_{i=1}^L \left[1 - \sum_{a_i=1}^{q-1} P_i(a_i) \right] \ln \left[1 - \sum_{a_i=1}^{q-1} P_i(a_i) \right] \quad (\text{A.91})$$

$$- \left(\sum_i^L \sum_{j>i}^L \sum_{a_i}^{q-1} \sum_{b_j}^{q-1} j_{ij}(a_i, b_j) P_i(a_i) P_j(b_j) \right) \alpha + O(\alpha^2) \quad (\text{A.92})$$

Derivando esta expresión una vez es posible evaluar la ecuación A.82.

$$\begin{aligned} \frac{dG}{dP_{i'}(a'_i)} &= - \sum_{i=1}^L \sum_{a_i=1}^{q-1} \frac{dP_i(a_i)}{dP_{i'}(a'_i)} \ln P_i(a_i) - P_i(a_i) \frac{1}{P_i(a_i)} \frac{dP_i(a_i)}{dP_{i'}(a'_i)} \\ &- \sum_{i=1}^L (-1) \sum_{a_i=1}^{q-1} \frac{dP_i(a_i)}{dP_{i'}(a'_i)} \ln \left[1 - \sum_{a_i=1}^{q-1} P_i(a_i) \right] - \sum_{i=1}^L \left[1 - \sum_{a_i=1}^{q-1} P_i(a_i) \right] \frac{\frac{\sum_{a_i=1}^{q-1} dP_i(a_i)}{dP_{i'}(a'_i)}}{\left[1 - \sum_{a_i=1}^{q-1} P_i(a_i) \right]} \\ &- \left(\sum_i^L \sum_{j>i}^L \sum_{a_i}^{q-1} \sum_{b_j}^{q-1} j_{ij}(a_i, b_j) P_j(b_j) \frac{dP_i(a_i)}{dP_{i'}(a'_i)} \right) \alpha \end{aligned} \quad (\text{A.93})$$

Dado que $\frac{dP_i(a_i)}{dP_{i'}(a'_i)} = \delta_{i,i'} \delta_{a_i, a'_i}$, las sumatorias con estos términos se ven reducidas:

$$\frac{dG}{dP_{i'}(a'_i)} = - \ln P_{i'}(a'_i) - 1 + \ln \left[1 - \sum_{a'_i=1}^{q-1} P'_i(a'_i) \right] + 1 - \left(\sum_{j=1}^L \sum_{b_j}^{q-1} j_{i'j}(a'_i, b_j) P_j(b_j) \right) \alpha \quad (\text{A.94})$$

Reemplazando la expresión entre corchetes por $P_{i'}(q)$, tomando $\alpha = 1$, cambiando las expresiones primadas a sin primar (cambio de nomenclatura) e igualando a $h_i(a_i)$ de acuerdo a la ecuación A.82, se obtiene

$$\frac{dG}{dP_i(a_i)} = - \ln P_i(a_i) + \ln [P_i(q)] - \sum_{j=1}^L \sum_{b_j}^{q-1} j_{ij}(a_i, b_j) P_j(b_j) = h_i(a_i) \quad (\text{A.95})$$

De la cual se puede despejar la expresión

$$\frac{P_i(a_i)}{P_i(q)} = \exp \left\{ -h_i(a_i) - \sum_{j|j \neq i} \sum_{b_j=1}^{q-1} j_{ij}(a_i, b_j) P_j(b_j) \right\} \quad (\text{A.96})$$

Derivando una vez más y utilizando la ecuación A.83, se puede encontrar una relación entre los parámetros de acoplamiento y la matriz de correlaciones conectadas. Partiendo de la ecuación A.94, si $j \neq i$ se obtiene

$$\frac{d^2 G}{dP_i(a_i) dP_{j'}(b'_j)} = - \sum_{j=1}^L \sum_{b_j}^{q-1} j_{ij}(a_i, b_j) \frac{dP_j(b_j)}{dP_{j'}(b'_j)} = -j_{ij'}(a_i, b'_j) \quad (\text{A.97})$$

Si $j = i$,

$$\begin{aligned} \frac{d^2 G}{dP_i(a_i) dP_{j'}(b'_j)} &= \frac{d^2 G}{dP_i(a_i) dP_i(b'_i)} \\ &= -\frac{1}{P_i(a_i)} \frac{dP_i(a_i)}{dP_i(b'_i)} + \frac{1}{1 - \sum_{a_i=1}^{q-1} P_i(a_i)} \sum_{a_i=1}^{q-1} -\frac{dP_i(a_i)}{dP_i(b'_i)} - \sum_{i=1}^L \sum_{b_i}^{q-1} j_{ii}(a_i, b_i) \frac{P_i(b_i)}{P_i(b'_i)} \end{aligned} \quad (\text{A.98})$$

Los j_{ii} son nulos por definición. Reemplazando $\frac{dP_i(a_i)}{dP_i(b'_i)} = \delta_{a_i, b'_i}$:

$$\frac{d^2 G}{dP_i(a_i) dP_{j'}(b'_j)} = -\frac{\delta_{a_i, b'_j}}{P_i(a_i)} - \frac{1}{P_i(q)} \quad (\text{A.99})$$

Por lo tanto, de las ecuaciones A.97 y A.99 junto a la ecuación A.83 se obtiene

$$(C^{-1})_{ij}(a_i, b_j) = \begin{cases} -j_{ij}(a_i, b_j) & \text{si } i \neq j \\ \frac{\delta_{a_i, b_j}}{P_i(a_i)} + \frac{1}{P_i(q)} & \text{si } i = j \end{cases} \quad (\text{A.100})$$

Información directa

Para condensar toda esta información en un único parámetro, en el trabajo [Morcos et al., 2011] definieron la información directa como una medida de información mutua debida únicamente a acoplamientos directos. Para un par de posiciones aisladas definieron

$$P_{ij}^{dir}(a_i, b_j) = \frac{1}{Z_{ij}} \exp \left\{ -j_{ij}(a_i, b_j) + \tilde{h}_i(a_i) + \tilde{h}_j(b_j) \right\} \quad (\text{A.101})$$

donde los campos locales son determinados, para cada par de posiciones, de manera que verifiquen las distribuciones marginales de aminoácidos:

$$f_i(a_i) = \sum_{b_j}^q P_{ij}^{dir}(a_i, b_j) \quad (\text{A.102})$$

$$f_j(b_j) = \sum_{a_i}^q P_{ij}^{dir}(a_i, b_j) \quad (\text{A.103})$$

Y Z_{ij} se obtiene normalizando. La información directa, en analogía con la información mutua, fue definida como:

$$DI_{ij} = \sum_{a_i=1}^q \sum_{b_j=1}^q P_{ij}^{dir}(a_i, b_j) \ln \left(\frac{P_{ij}^{dir}(a_i, b_j)}{f_i(a_i)f_j(b_j)} \right) \quad (\text{A.104})$$

Aproximación por Máxima Pseudo-Verosimilitud

En esta aproximación se busca ajustar los parámetros del modelo de Potts pero en vez de estudiar la distribución de probabilidades de secuencias, se analiza la probabilidad condicionada. Se define la probabilidad de que ocurra a_r en la posición r dado que ocurrieron $a_{|r}$ en las demás posiciones según:

$$P(a_r = a_r^s | a_{|r} = a_{|r}^s) = \frac{\exp \left(h_r(a_r^s) + \sum_{j=1, j \neq r}^L j_{r,i}(a_r^s, b_j^s) \right)}{\sum_{a_r=1}^q \exp \left(h_r(a_r) + \sum_{j=1, j \neq r}^L j_{r,i}(a_r, b_j^s) \right)} \quad (\text{A.105})$$

Se define la verosimilitud condicional a partir de ella como:

$$g_r(h_r, j_r) = -\frac{1}{N} \sum_{s=1}^N \log (P(a_r = a_r^s | a_{|r} = a_{|r}^s)) \quad (\text{A.106})$$

La máxima verosimilitud sobre las probabilidades condicionales, para todas las posiciones simultáneamente, es denominada pseudo similitud y definida por:

$$l_{\text{pseudo}} = \sum_{r=1}^L g = -\frac{1}{N} \sum_{r=1}^L \sum_s^N \log (P(a_r = a_r^s | a_{|r} = a_{|r}^s)) \quad (\text{A.107})$$

Y es la variable a ser maximizada. Finalmente, la correlación entre dos posiciones se calcula como el promedio de los parámetros de interacción obtenidos:

$$\|j_{ij}\| = \sqrt{\sum_{a_i, b_j}^q (j_{ij}(a_i, b_j))^2} \quad (\text{A.108})$$

Apéndice B

Código ajuste de parámetros de Modelo de Potts

```
// Compilar con g++ -std=gnu++11 -o nombre nombre.cpp

#include <iostream>
#include <omp.h>
#include <stdio.h>
#include <stdlib.h>
#include <string>
#include <fstream>
#include <vector>
#include <math.h>
#include <ctime>
#include <random>

using namespace std;

// Funciones para generar numeros al azar
std::random_device rd;
std::mt19937 rng(rd());
std::uniform_int_distribution<int> uni20(0,20);
std::uniform_int_distribution<int> uni65(0,65);

int main()
{
/* 1- CARGAR DATOS E INICIALIZAR PARAMETROS */
```

```

/* Parametros generales | General parameters */
int SavingDataSteps=20;          // Saving All Data every these N of Steps
ofstream errorFile("Error.txt");

const int naa=21; // n of amino acids
const int npos=66; // length sequences
const int naanpos=naa*npos; // columns/rows of the matrix Pij
const int npid=34; // number of ID Lagrange Multipliers
const int npos2=npn/2;

const double epsilonSingle=0.05; // epsilon to renew parameters hi
const double epsilonJoint=0.01; // epsilon to renew parameters jij
const double gamma=0.001; // Parameter for L1 regularization
const double epsilonID=10; // epsilon to renew parameters lambda identity
const double maxErrorPermitted=0.015; // maximum error allowed between pij-fij and pi-fi
const int ncomb=npn*(npos+1)/2-npos; // number of possible combinations of positions (1-
int PosPos[ncomb][2]; // all possible combinations of positions
int a=0; //auxiliar counter
for(int i=0;i<npos;i++){
for(int j=i;j<npos;j++){
if(i!=j){
PosPos[a][0]=i;
PosPos[a][1]=j;
a+=1;
}
}
}

/* Read experimental frequencies data */
double f1Data[naanpos];
ifstream f1File("frecuenciasmarginales_train.txt");
for (int j=0; j<naanpos; j++)
{
f1File >> f1Data[j];
}
f1File.close();

static double f2Data[naanpos][naanpos];
ifstream f2File("frecuenciasconjuntas_train.txt");
for(int fila=0;fila<naanpos;fila++)
{
for(int col=0;col<naanpos;col++)
{
f2File >> f2Data[fila][col];
}
}
f2File.close();

double pid_input[npid];

```

```

ifstream pidFile("pid_data.txt");
for(int fila=0;fila<npid;fila++){
pidFile >> pid_input[fila];
}
pidFile.close();

/* 2- COMIENZO DE ITERACION DE PARAMETROS */

/* MC parameters */
int FirstSeqSaved=1000; // first sequence I save in the MC procedure (to miss initi
int SaveSequencesRate=1000; // Every how many iterations should I consider the seque
int TotalSequences=80000; // How many sequences should I consider in the MC to calc
int TotalIterations=(TotalSequences+FirstSeqSaved)*SaveSequencesRate; // How many it

/* Potts parameters matrices (Initially hi = log (fi) and Jij=0) */

static double jij[naanpos][naanpos];
ifstream j0File("Final_jij_v1.txt");
for(int i=0;i<naanpos;i++){
for(int j=0;j<naanpos;j++){
j0File >> jij[i][j];
}
}
j0File.close();

static double hi[naanpos];
ifstream h0File("Final_h_v1.txt");
for(int i=0;i<naanpos;i++){
h0File >> hi[i];
}
h0File.close();

static double lambda_id[npid];
for(int i=0;i<npid;i++){lambda_id[i]=0;}

// Definiciones de variables a usar durante el while
static double pij[naanpos][naanpos];
static double p1[naanpos];
static double pid_dis[npid];
double energy;

double error=2.00;
int counter=0;
bool writeFiles=false;
clock_t begin = clock();

std::ofstream sequencesFile;

```

```

std::ofstream energyFile;

while(error > maxErrorPermitted){
counter+=1;

writeFiles= (counter%SavingDataSteps==1);

if(writeFiles){
cout << counter << endl;
//File to save sequences:
std::string prefixSequenceFile="sequences";
std::string fullnameSequences=prefixSequenceFile.append(std::to_string(counter));
sequencesFile.open(fullnameSequences);

//File to save sequences' energies
std::string prefixEnergyFile="Energy";
std::string fullnameEnergy=prefixEnergyFile.append(std::to_string(counter));
energyFile.open(fullnameEnergy);
}

/* a) Recalculate Pij and Pi with the new parameters */
// To do this I generate a large set of sequences with MC

// reinicio valores de pij a cero
for(int i=0;i<naanpos;i++){
for(int j=0;j<naanpos;j++){
pij[i][j]=0;
}
}

// reinicio valores de pid_dis a cero
for(int i=0;i<npid;i++){pid_dis[i]=0.00;}

//Random initial sequence:
vector<int> sequence;
for(int i=0; i<npos; i++)
{
int random_integer = uni20(rng);
sequence.push_back(random_integer);
}

// Its pId
int pid_seq=0;
for(int j=0;j<npos2;j++){
if(sequence[j]==sequence[j+npos2]){pid_seq+=1;}
}

// It's energy
energy=0.00;

```

```

for(int i=0; i<npos; i++)
{
energy+= -hi[naa*i+sequence[i]];
if(i!=(npos-1)){
for(int j=(i+1); j<npos; j++)
{
energy+= -jij[naa*(i)+sequence[i]][naa*(j)+sequence[j]];
}
}
}
energy+=lambda_id[pid_seq];

// Do random mutations to recalculate Pij with new hi and jij parameters
// Hago mutaciones y recorro el paisaje de secuencias para calcular los Pij
int seqsavecounter=0;
int sign;

while(seqsavecounter<TotalIterations)
{
// Elijo mutacion al azar | random mutation
int mpos= uni65(rng);
int mres= uni20(rng);
int oldRes=sequence[mpos];
int pid_seq_nuevo=pid_seq;

if(mpos<npos2){
if(oldRes == sequence[mpos+npos2] & mres != sequence[mpos+npos2]){pid_seq_nuevo-=1;}
if(oldRes != sequence[mpos+npos2] & mres == sequence[mpos+npos2]){pid_seq_nuevo+=1;}
}else{
if(oldRes == sequence[mpos-npos2] & mres != sequence[mpos-npos2]){pid_seq_nuevo-=1;}
if(oldRes != sequence[mpos-npos2] & mres == sequence[mpos-npos2]){pid_seq_nuevo+=1;}
}

// Calculo la variacion de energia con esta mutacion | change of energy
double EiNewLocal= -hi[naa*mpos+mres];
double EiNewInteraction=0.00;
double EiNewPID=lambda_id[pid_seq_nuevo];
double EiOldLocal= -hi[naa*mpos+oldRes];
double EiOldInteraction=0.00;
double EiOldPID=lambda_id[pid_seq];

for(int j=0; j<npos; j++)
{
if(j!=mpos){
EiNewInteraction+= -jij[ naa*(mpos)+mres][naa*(j)+sequence[j]];
EiOldInteraction+= -jij[ naa*(mpos)+oldRes][naa*(j)+sequence[j]];
}
}
}

```



```

double dE=EiNewLocal+EiNewInteraction+EiNewPID-EiOldLocal-EiOldInteraction-EiOldPID;

// Elijo si me quedo o no con la mutacion | decide weater I keep the mutation
if(dE < 0)
{
sequence[mpos]=mres;
energy+=dE;
pid_seq=pid_seq_nuevo;
}else{
double r = ((double) rand() / (RAND_MAX));
double expE=exp(-dE);
if(r<expE){
sequence[mpos]=mres;
energy+=dE;
pid_seq=pid_seq_nuevo;
}
}
seqsavecounter+=1;

// Once every SaveSequenceRate, I refresh Pij values |
// Renuevo los valores de Pij con los valores de una secuencia cada SaveSequencesRate
int aux=seqsavecounter%SaveSequencesRate;
if(aux==1 & seqsavecounter>=(FirstSeqSaved*SaveSequencesRate))
{
// Actualizo Pij con los valores de esta secuencia
for(int j=0;j<ncomb; j++){
int index1=PosPos[j][0]*naa+sequence[PosPos[j][0]];
int index2=PosPos[j][1]*naa+sequence[PosPos[j][1]];
pij[index1][index2]+=1;
if(index1!=index2){
pij[index2][index1]+=1;
}
}
}

//Actualizo la distribucion de porcentaje de Identidad
pid_dis[pid_seq]+=1;

if(sequencesFile.is_open()){
// Write sequence and it's energy into files
for (int j=0; j<npos; j++)
{
sequencesFile << sequence[j] << " ";
}
sequencesFile << endl;
energyFile << energy << endl;
}
}
}

```

```

if(sequencesFile.is_open()){
sequencesFile.close();
energyFile.close();
}

// Normalization of Pij | Normalizo los Pij
for(int index1=0;index1<naanpos; index1++){
for(int index2=0;index2<naanpos;index2++){
pij[index1][index2]=pij[index1][index2]/TotalSequences;
}
}

// Calculate Pi values as partial sums of Pij | Calculo los nuevos Pi como suma parcial
//reinicio valores de p1 a cero
for(int indA=0;indA<naanpos;indA++){p1[indA]=0;}

// For position 1 | Los de la posicion 1
for(int indA=0;indA<naa;indA++){
for(int indx=naa;indx<(2*naa);indx++){
p1[indA]+=pij[indA][indx];
}
}
// For other positions | Los de las demas posiciones
for(int indA=naa;indA<naanpos;indA++){
for(int indx=0;indx<naa;indx++){
p1[indA]+=pij[indx][indA];
}
}

/* b) Compare Pij and Pi (from the model) to fij and fi (from data) */
double maxPij=0.00;
double m;
for(int fila=0;fila<naanpos;fila++)
{
int n1=fila/naa;
for(int col=0;col<naanpos;col++)
{
int n2=col/naa;
if(n1!=n2){
m=abs(f2Data[fila][col] - pij[fila][col]);
if(m>maxPij){maxPij=m;
}
}
}
}
for(int i=0;i<naanpos;i++)
{
m=abs(f1Data[i]-p1[i]);

```

```

if(m>maxPij){maxPij=m;
}
}

for(int i=0;i<npid;i++)
{
m=abs(pid_dis[i]/TotalSequences - pid_input[i]);
if(m>maxPij){maxPij=m;}
}
error=maxPij;

if(writeFiles){
// Parameters hi and Jij before re-calculating them
std::string prefixHI="ParamH";
std::string fullnameHI=prefixHI.append(std::to_string(counter));
ofstream hiFile(fullnameHI);
for (int j=0; j<naanpos; j++)
{
hiFile << hi[j] << "\t";
if((j+1)%naa==0){hiFile << endl;}
}
hiFile.close();
cout <<"    ParamHI ";

std::string prefixJij="ParamJ";
std::string fullnameJIJ=prefixJij.append(std::to_string(counter));
ofstream jijFile(fullnameJIJ);
for(int fila=0;fila<naanpos;fila++)
{
for(int col=0;col<naanpos;col++)
{
jijFile << jij[fila][col] << "\t";
}
jijFile<< endl;
}
jijFile.close();
cout <<"    ParamJij "<< endl;

// Parameters lambda_pid
std::string prefixID="ParamID";
std::string fullnameID=prefixID.append(std::to_string(counter));
ofstream IDFile(fullnameID);
for (int j=0; j<npid; j++)
{
IDFile << lambda_id[j] << "\t";
}

```

```

IDFile.close();
cout <<"    ParamHI ";

//Pij model distribution
std::string prefixPIJN="Pij";
std::string fullnamePIJN=prefixPIJN.append(std::to_string(counter));
ofstream PijFile(fullnamePIJN);
for(int fila=0;fila<naanpos;fila++)
{
for(int col=0;col<naanpos;col++)
{
PijFile << pij[fila][col] << "\t";
}
PijFile<< endl;
}
PijFile.close();
cout <<"    PijDistrib ";

//Pi model distribution
std::string prefixPI="Pi";
std::string fullnamePI=prefixPI.append(std::to_string(counter));
ofstream PiFile(fullnamePI);
for (int j=0; j<naanpos; j++)
{
PiFile << p1[j] << "\t";
if((j+1)%naa==0){PiFile << endl;}
}
PiFile.close();
cout <<"    PiDISTR ";

errorFile << error << endl;

// PID distribution
std::string prefixPID="Pid";
std::string fullnamePID=prefixPID.append(std::to_string(counter));
ofstream PiDFile(fullnamePID);
for (int j=0; j<npid; j++)
{
PiDFile << pid_dis[j] << "\t";
}
PiDFile.close();

// Mido cuanto tarda
clock_t end = clock();
double elapsed_secs = double(end - begin) / CLOCKS_PER_SEC;
cout << "... Tiempo: " << elapsed_secs << endl;
clock_t begin = clock();

```



```

/* 3- GUARDO DATOS FINALES */

// Parameters hi and Jij
ofstream hiFinalFile("Final_hi.txt");
for (int j=0; j<naapos; j++)
{
hiFinalFile << hi[j] << "\t";
if((j+1)%naa==0){hiFinalFile << endl;}
}
hiFinalFile.close();

ofstream jijFinalFile("Final_jij.txt");
for(int fila=0;fila<naapos;fila++)
{
for(int col=0;col<naapos;col++)
{
jijFinalFile << jij[fila][col] << "\t";
}
jijFinalFile<< endl;
}
jijFinalFile.close();

//Pij model distribution
ofstream PijFinalFile("Final_Pij.txt");
for(int fila=0;fila<naapos;fila++)
{
for(int col=0;col<naapos;col++)
{
PijFinalFile << pij[fila][col] << "\t";
}
PijFinalFile<< endl;
}
PijFinalFile.close();

//Pi model distribution
ofstream PiFinalFile("Final_Pi.txt");
for (int j=0; j<naapos; j++)
{
PiFinalFile << p1[j] << "\t";
if((j+1)%naa==0){PiFinalFile << endl;}
}
PiFinalFile.close();

errorFile << error << endl;
errorFile.close();

// Parameters lamba_pid

```

```

ofstream IDFinalFile("Final_lambdaID.txt");
for (int j=0; j<npid; j++)
{
IDFinalFile << lambda_id[j] << "\t";
}
IDFinalFile.close();

// Pid distribution
ofstream PIDFinalFile("Final_ID.txt");
for (int j=0; j<npid; j++)
{
PIDFinalFile << pid_dis[j] << "\t";
}
PIDFinalFile.close();

}

```