RIDAA
Repositorio Institucional
Digital de Acceso Abierto de la
Universidad Nacional de Quilmes

Universidad
Nacional
de Quilmes

Marchetti, Julia

Desarrollo de herramientas bioinformáticas basadas en métodos evolutivos y estructurales para el estudio de proteínas

Esta obra está bajo una Licencia Creative Commons Argentina.
Atribución - No Comercial 2.5
https://creativecommons.org/licenses/by-nc/2.5/ar/

Documento descargado de RIDAA-UNQ Repositorio Institucional Digital de Acceso Abierto de la Universidad Nacional de Quilmes de la Universidad Nacional de Quilmes

Cita recomendada:
Marchetti, J. (2021). Desarrollo de herramientas bioinformáticas basadas en métodos evolutivos y estructurales para el estudio de proteínas. (Tesis de doctorado). Universidad Nacional de Quilmes, Bernal, Argentina. Disponible en RIDAA-UNQ Repositorio Institucional Digital de Acceso Abierto de la Universidad Nacional de Quilmes http://ridaa.unq.edu.ar/handle/20.500.11807/3000

Puede encontrar éste y otros documentos en: https://ridaa.unq.edu.ar

Universidad Nacional de Quilmes
Roque Sáenz Peña 352 // Bernal
Buenos Aires // Argentina
t.: (+41 11) 4365 7100
f.: (+54 11) 4365 7101
info@unq.edu.ar

Julia Marchetti, Repositorio Institucional Digital de Acceso Abierto,
Marzo de 2021, pp. 177,
http://ridaa.unq.edu.ar,
Universidad Nacional de Quilmes, Secretaría de Posgrado,
Doctorado en Ciencia y Tecnología

Desarrollo de herramientas bioinformáticas basadas en métodos evolutivos y estructurales para el estudio de proteínas

TESIS DOCTORAL

Julia Marchetti

julimarchetti@gmail.com

Resumen

A lo largo de los años se hizo cada vez más evidente la necesidad de contar con el conocimiento de la estructura tridimensional de una proteína, ya que permite entender numerosos aspectos de la misma. La conservación de la estructura proteica durante la evolución reafirma, a nivel molecular, la importancia de la relación estructura-función, concepto fundacional en Biología.

El presente trabajo se centra en el estudio de un conjunto específico de proteínas llamadas Proteínas Intrínsecamente Desordenadas (***I**ntrinsically **D**isordered **P**rotein* IDPs por sus siglas en inglés y llamadas así de aquí en adelante). Estas proteínas presentan características estructurales y dinámicas únicas convirtiéndolas en un desafío para la Biología. Se trata de proteínas 'recientes' en el mundo de la Biología Estructural históricamente gobernada por la aparente abundancia de las paradigmáticas proteínas globulares que cuentan con una estructura definida y son paradigmas de la palabra 'proteínas' en cualquier libro de texto de Bioquímica.

¿Cuán desestructuradas son las IDPs? ¿Sigue siendo válido el concepto de la relación estructura-función? ¿La ausencia de estructura tiene alguna consecuencia esencial en el proceso de sustitución de los aminoácidos? Si bien las IDPs pueden no tener estructura, no escapan a otros principios fundacionales como el enunciado por Theodosius Dobzhansky: '*nada tiene sentido en biología si no es a la luz de la evolución' (Dobzhansky, 1973)* y es por ello que en este trabajo de tesis nos enfocamos en dilucidar la relación entre la ausencia de estructura y el proceso de evolución molecular de las IDPs.

En el capítulo 1 de esta tesis se dan nociones introductorias que son importantes para el desarrollo de la tesis tales como el concepto del ensemble(1) nativo, plegado de proteínas, diversidad conformacional y una breve introducción sobre la biología de las proteínas intrínsecamente desordenadas. Luego se desarrollan conceptos introductorios de evolución molecular, modelos de evolución y métodos de evaluación de modelos.
En el capítulo 2 nos enfocamos en encontrar posibles restricciones estructurales que restringen la divergencia de la secuencia en las IDPs.
En el capítulo 3 mostramos el impacto diferencial de las velocidades de evolución sitio-específicas para regiones ordenadas y regiones desordenadas de las IDPs y su relación con características estructurales de las mismas.
En el capítulo 4 estudiamos la importancia de determinadas conformaciones en el ensemble de las IDPs en cuanto a definir el estado nativo, y de ahí su importancia en la función biológica de la proteína.
En el capítulo 5 contiene las conclusiones generales de este trabajo y perspectivas a futuro.
Palabras clave: proteínas intrínsecamente desordenadas - evolución - ensembles conformacionales - diversidad conformacional.

(1) En este trabajo se usa la palabra ensemble en inglés ya que es el término utilizado en la literatura y además porque su traducción al español (conjunto o ensamble) no representa fidedignamente el concepto que se quiere ilustrar

Repositorio Institucional Digital de Acceso Abierto, Universidad Nacional de Quilmes

Desarrollo de herramientas bioinformáticas basadas en métodos evolutivos y estructurales para el estudio de proteínas

Tesis para optar al título de Doctor en Ciencia y Tecnología de la Universidad Nacional de Quilmes

Julia Marchetti

Director: Dr. Gustavo Parisi

Co-directora: Dra. María Silvina Fornasari

Consejero: Dr. Facundo Temprana

Febrero - 2021

Publicaciones

- En el marco de esta tesis doctoral:

  Palopoli, N.*, **Marchetti, J**.*, Monzon, A.M., Zea, D.J., Tosatto, S.C.E., Fornasari, M.S., Parisi, G., 2021. Intrinsically disordered protein ensembles shape evolutionary rates revealing conformational patterns. J. Mol. Biol. 433, 166751. doi:10.1016/j.jmb.2020.166751. *Ambos autores contribuyeron en la misma extensión en la elaboración del trabajo.

  Marchetti, **J.**, Monzon, A.M., Tosatto, S.C.E., Parisi, G., Fornasari, M.S., 2019. Ensembles from Ordered and Disordered Proteins Reveal Similar Structural Constraints during Evolution. J. Mol. Biol. 431, 1298–1307. doi:10.1016/j.jmb.2019.01.031

- Fuera del marco de la tesis doctoral como parte del proyecto de extensión del grupo de investigación:

  Velez Rueda, A.J., Benítez, G.I., **Marchetti, J**., Hasenahuer, M.A., Fornasari, M.S., Palopoli, N., Parisi, G., 2019. Bioinformatics calls the school: Use of smartphones to introduce Python for bioinformatics in high schools. PLoS Comput. Biol. 15, e1006473. doi:10.1371/journal.pcbi.1006473.

Agradecimientos

A Gustavo y a Silvina, por su dirección y su apoyo siempre que lo necesité.

A Nicolás por su paciencia infinita y estímulo constante.

A todo el SBG, que creció enormemente y que cada día se consolida más como grupo de investigación de alto impacto. Gracias por su calidad humana y su acompañamiento.

A mis amigxs, todxs, lxs de ahora, lxs de antes, lxs de siempre. Gracias por estar y por la infinidad de momentos hermosos compartidos.

A Walde por su compañía y cariño incondicional (y por varias de las imágenes mostradas en este trabajo).

A Guido, por enseñarme todos los días a ver la vida de otra manera.

A mamá y a papá, por amarme tanto y darme absolutamente todo, sin ellos la elaboración de este trabajo hubiera sido imposible.

A quien lea este trabajo,
Gracias.

Pedro Pietroboni fue un trabajador rural que apenas sabía leer y escribir, generaciones más tarde su bisnieta presenta esta tesis doctoral.
A la educación, libre, pública, gratuita y laica.

Índice general

Título 3

Resumen General 3

Publicaciones 5

Agradecimientos 6

1 Introducción 18

1.1 Relación entre la estructura y función proteica 18

1.1.1 El estado nativo: una breve reseña histórica 19

1.1.2 Plegamiento de proteínas: teoría del paisaje energético 27

1.1.3 El estado nativo hoy: el ensemble conformacional 33

1.1.4 Cuantificación de la diversidad conformacional y distribución de movilidad 34

1.1.5 Proteínas Intrínsecamente Desordenadas (IDPs) 39

1.2 Evolución de Proteínas 48

1.2.1 Evolución molecular 48

1.2.2 El rol de la estructura proteica durante el curso de la evolución 53

1.2.3 El Patrón de sustitución y los modelos de evolución 55

1.2.4 Modelos de evolución irrestrictos 59

1.2.5 Modelos de evolución restrictos 61

1.2.6 Evaluación de modelos de evolución 66

Objetivos 69

Objetivos generales 69

Objetivos específicos 69

2 Estudio de las restricciones estructurales en proteínas intrínsecamente desordenadas 72

2.1 Resumen 72

2.2 Introducción 73

2.3 Resultados 76

2.3.1 Descripción del conjunto de datos 76

2.3.2 Determinación de restricciones estructurales a partir de los contactos físicos 80

2.3.3 Determinación de restricciones estructurales por medio de métodos evolutivos. 84

2.3.4 Sitios restringidos en proteínas desordenadas 89

2.4 Discusión 94

2.5 Métodos específicos del capítulo 95

2.5.1 Construcción de los conjuntos de datos 95

2.5.2 Estimación de contactos terciarios 96

2.5.3 Estimación de sitios estructuralmente restringidos SC 96

3 Las velocidades de evolución de IDPs son modeladas por los ensembles conformacionales. 103

3.1 Resumen 103

3.2 Introducción 103

3.3 Resultados 106

3.3.1 Los perfiles de velocidad permiten la caracterización de la organización estructural de las IDPs 106

3.3.2 Efectos de los contactos entre residuos en las velocidades de evolución. 112

3.3.3 Relación entre el desorden interno y el desorden terminal y las velocidades de evolución 122

3.4 Discusión 125

3.5 Métodos específicos del capítulo 126

3.5.1 Construcción del conjunto de datos 126

3.5.2 Estimación de las velocidades de evolución 129

3.5.3 Información estructural en confórmeros y ensembles 129

3.5.4 Obtención de los perfiles de velocidad y patrones conformacionales 130

3.6 Disponibilidad de datos y código. 131

4 Estudio y evaluación de la información estructural en ensembles conformacionales 133

4.1 Evaluación de la información estructural en ensembles mediante el modelo de evolución SCPE. 133

4.1.1 Resultados 134

4.1.2 Metodología 135

4.1.3 Discusión 136

4.2 Evaluación de la información estructural en ensembles mediante velocidades de evolución. 136

4.2.1 Resultados 137

4.2.2 Metodología 143

4.2.3 Discusión 144

5 Conclusiones generales y perspectivas a futuro 146

6 Listado de abreviaturas utilizados en esta tesis 150

7 Apéndice A 152

8 Bibliografía 157

Índice de figuras

Figura 1.1 Presentación de la estructura de la lisozima por David Phillips en el Royal Institution, Londres, Noviembre de 1965. 19
Figura 1.2 Representación esquemática del modelo llave cerradura. 20
Figura 1.3 Representación del estado nativo formado por una única conformación. 21
Figura 1.4 Paisaje de energía libre de una proteína. 22
Figura 1.5 Representación esquemática del modelo de ajuste inducido. 23
Figura 1.6 Esquema del modelo de pre-equilibrio. 25
Figura 1.7 Representación de la estructura de la Mioglobina como si fuera un castillo. 26
Figura 1.8 Superficie de energía potencial - Teoría del golfista ciego 28
Figura 1.9 Camino de plegado (Folding pathway) según Levinthal. 29
Figura 1.10. Embudo de plegado (Folding funnel) - cooperativo. 29
Figura 1.11 Embudo de plegado (Folding Funnel). 30
Figura 1.12 Plegado proteico y el paisaje energético. 31
Figura 1.13 Esquema representativo del proceso de population shift. 32
Figura 1.14 Esquemas de las formas T y R de Hemoglobina. 33
Figura 1.15 Esquema de una superposición estructural de cuerpo rígido. 35
Figura 1.16 Ejemplos de diversidad conformacional en proteínas. 37
Figura 1.17 Distribución de la diversidad conformacional máxima para una proteína. . 38
Figura 1.18 Impacto en la Biología Estructural por el descubrimiento de las IDPs. 40
Figura 1.19 Esquema de la evolución del paradigma de relación estructura-función. .. 41
Figura 1.20. Representación de IDPs. 42
Figura 1.21 El paisaje de energía libre y los ensembles conformacionales de IDRs o IDPs. 43
Figura 1.22 Funcionalidad de las IDPs. 45
Figura 1.23 Esquema de relación entre proteínas ortólogas y parálogas. 49
Figura 1.24 Ejemplo de un alineamiento múltiple de secuencias de aminoácidos (MSA). 51
Figura 1.25 Relación entre la divergencia estructural y la divergencia secuencial. 54
Figura 1.26 Ejemplos de matrices de sustitución basadas en proteínas. 60

Figura 1.27 Esquema general del algoritmo empleado para el SCPE. 63
Figura 1.28. Perfil energético de la proteína Lpxa. 65
Figura 1.29 Representación del significado del likelihood y la manera de computarlo. 67
Figura 2.1 Diversidad conformacional en función de la flexibilidad. 77
Figura 2.2 Distribución del porcentaje de desorden en IDPs y proteínas globulares. ... 79
Figura 2.3 Porcentaje de posiciones con contactos entre residuos por confórmero. 81
Figura 2.4 Representación de distintos confórmeros de un ensemble proteico. 82
Figura 2.5 Esquema del cálculo de sitios de contactos. 83
Figura 2.6 Porcentaje de posiciones con contactos entre residuos acumulado a lo largo del ensemble. 84
Figura 2.7 Distribución de sitios estructuralmente restringidos (SC) por confórmero. ... 87
Figura 2.8 Esquema de determinación de sitios estructuralmente restringidos (SC) evolutivamente a lo largo del ensemble. 87
Figura 2.9 Distribución del número de sitios restringidos estructuralmente (SC) acumulados a lo largo del ensemble. 88
Figura 2.10 Representación del ensemble conformacional conformacional de IDP. 90
Figura 2.11 Distribución de sitios estructuralmente restringidos (SC) en IDPs 91
Figura 2.12 Se muestra la distribución de sitios estructuralmente restringidos (SC) en ensembles de IDPs. 92
Figura 2.13 Porcentaje de conformaciones con contactos terciarios en sitios sitios estructuralmente restringidos (SC). 93
Figura 2.14 Esquema de flujo de trabajo. 97
Figura 2.15 Esquema de flujo de comparación de modelos para una determinada posición. 99
Figura 2.16 Esquema de definición de sitio sitios estructuralmente restringidos (SC). 101
Figura 3.1 Representación de un ensemble IDPs. 107
Figura 3.2 Boxplots representando la distribución de las velocidades de evolución normalizadas por posición. 108
Figura 3.3 Perfiles de velocidad normalizada y de RMSF. 110
Figura 3.4 Esquema de cálculo del promedio de contactos a lo largo del ensemble. . 113

Figura 3.5 Velocidad de evolución normalizada en función del promedio de contactos. 113
Figura 3.6 Velocidad de evolución normalizada en función de información de contactos. 114
Figura 3.7 Esquema de cálculo de robustez de contacto para un ensemble. 115
Figura 3.8 Distribución de la robustez de contactos. 116
Figura 3.9 Distribución de las velocidades de evolución normalizadas en función de la robustez de contactos. 117
Figura 3.10 Velocidad de evolución normalizada en función de la robustez de contacto. 118
Figura 3.11 Distribución de las velocidades de evolución normalizadas en función de la distancia al residuo ordenado más cercano. 119
Figura 3.12 Distribución de RMSF normalizados en función de la distancia al residuo ordenado más cercano. 120
Figura 3.13 Distribución del promedio de contactos en función de la distancia al residuo ordenado más cercano. 120
Figura 3.14 Distribución de la proporción de contactos. 121
Figura 3.15 Esquema de definición de desorden terminal y de desorden interno. 123
Figura 3.16 Velocidad de evolución normalizada y promedio de contactos para desorden para distintos tipos de desorden. 124
Figura 3.17 Representación de IDPs presentes en el conjunto de datos. 128
Figura 3.18 Perfiles individuales de velocidad de evolución normalizada y RMSF normalizado. 131
Figura 4.1 Distribución de mínimo número de confórmeros. 134
Figura 4.2 Conformaciones necesarias para la obtención de la mejor correlación. 140
Figura 4.3 Representación de ensembles proteicos para cada grupo de correlación. 141
Figura 7.1 Boxplots representando la distribución de las velocidades de evolución normalizadas por posición. 152
Figura 7.2. Velocidad de evolución normalizada en función del promedio de contactos. 153

Figura 7.3 Distribución de las velocidades de evolución normalizadas en función de la robustez de contactos. 153
Figura 7.4 Velocidad de evolución normalizada en función de la robustez de contacto. 154
Figura 7.5 Distribución de las velocidades de evolución normalizadas en función de la distancia al residuo ordenado más cercano. 154
Figura 7.6 Distribución del promedio de contactos en función de la distancia al residuo ordenado más cercano..... 155

Índice de ecuaciones

Ecuación 1: Cálculo de RMSD. 35
Ecuación 2: Cálculo de $RMSD_{100}$. 36
Ecuación 3: Expresión matemática de la matriz de intercambiabilidad P(t). 56
Ecuación 4: Expresión matemática de la matriz de intercambiabilidad P(t) mediante expansión del exponencial como serie..... 56
Ecuación 5: Reversibilidad de la matriz Q. 56
Ecuación 6: Expresión para la matriz de velocidad Q_{ii}. 57
Ecuación 7: Expresión matemática de la matriz (Q)..... 57
Ecuación 8: Expresión matemático del score utilizado por el SCPE. 64
Ecuación 9: Ecuación de Boltzman para el cálculo de energía de interacción entre dos residuos..... 64
Ecuación 10: Cálculo del AIC. 98
Ecuación 11: Cálculo de ΔAIC. 99
Ecuación 12: Fórmula de obtención de combinaciones. 135

Índice de tablas

Tabla 1: Valores de los promedios de correlación obtenidos para confórmeros individuales o para un subconjunto de confórmeros. 139
Tabla 2: Valores de RMSD promedio, RMSD máximo y RG promedio para las mejores correlaciones obtenidas con la robustez de contactos o el promedio. 143

1 Introducción

1.1 Relación entre la estructura y función proteica

Históricamente se ha entendido a la relación estructura-función de las proteínas de manera tal que la estructura de una proteína es aquella que le permite a la proteína cumplir su función biológica en su entorno fisiológico. La condición necesaria y suficiente para que una proteína tenga estructura es que cuente con al menos tres niveles estructurales: el primer nivel está establecido por la secuencia aminoacídica (considerada estructura primaria). El segundo nivel está establecido por la estructura secundaria que está definida como un arreglo local estabilizado por los puentes de hidrógeno del esqueleto covalente de la proteína (Pauling et al., 1951). La estructura terciaria corresponde a la organización en el espacio de los elementos de estructura secundaria, o el plegamiento completo de la cadena polipeptídica y es estabilizada por interacciones no covalentes y puentes disulfuro. En esta tesis se utilizará el término contactos terciarios, para hacer referencia a interacciones no covalentes entre los átomos pertenecientes a los residuos de los aminoácidos.

El estado nativo, bajo la visión clásica, es una estructura que adquiere una conformación tridimensional única y de mínima energía libre que le permite a la proteína llevar a cabo su función en el entorno donde se encuentra (Figura 1.1). En este sentido, esa única estructura se conoce como el estado nativo de la proteína y de allí el paradigma bien establecido de una secuencia, una estructura, una función. Esta visión clásica de la relación estructura-función de las proteínas es consecuencia de un sesgo histórico en las proteínas que se estudiaron (mayormente globulares) y los métodos experimentales con los que fueron estudiadas (mayormente cristalografía de rayos X). Sin embargo, este concepto fue discutido y revisado extensamente en los últimos años debido al descubrimiento de nuevas proteínas y al desarrollo de nuevas técnicas de estudio. El objetivo de esta sección es examinar el recorrido histórico que tuvo el concepto del estado nativo, dar breves generalidades sobre la estructura de proteínas y una descripción general de las Proteínas Intrínsecamente Desordenadas (en inglés

Intrinsically Disordered Proteins, IDPs por sus siglas en inglés, acrónimo que se usará desde este punto en adelante).

Figura 1.1 Presentación de la estructura de la lisozima por David Phillips en el Royal Institution, Londres, Noviembre de 1965.

Para representar la idea de estado nativo, Phillips contrastó la estructura globular de la lisozima (recientemente obtenida por cristalografía de rayos X por él mismo y colaboradores (BLAKE et al., 1965)) hacia la derecha de Philips sobre el escritorio, con la estructura desplegada de la misma colgando del techo. Además, durante la charla comparó la estructura de la lisozima con una "molécula" inorgánica, el cloruro de sodio, también representado con un modelo sobre el escritorio hacia la izquierda de Phillips. Como nota de color, la estructura desplegada (10mts de largo en escala de 2cm=1 angstrom) de la lisozima contenía luces rojas y azules para representar los aminoácidos cargados. Tomado de The early history of lysozyme Louise N. Johnson nature structural biology • volume 5 number 11 • november 1998.

1.1.1 El estado nativo: una breve reseña histórica

En 1984 Emil Fischer (Fischer, 1894) propone un modelo para explicar la alta especificidad que muestran las enzimas por sus sustratos: el modelo de llave cerradura. En este modelo tanto la proteína como su sustrato, poseen lo que se denominó *complementariedad geométrica*. El sustrato hace las veces de 'llave' que encaja de

manera perfecta con la 'cerradura' (proteína, más específicamente el sitio activo), por lo que la estructura funcional de una proteína debía ser única para cumplir con su alta especificidad catalítica (Figura 1.2).

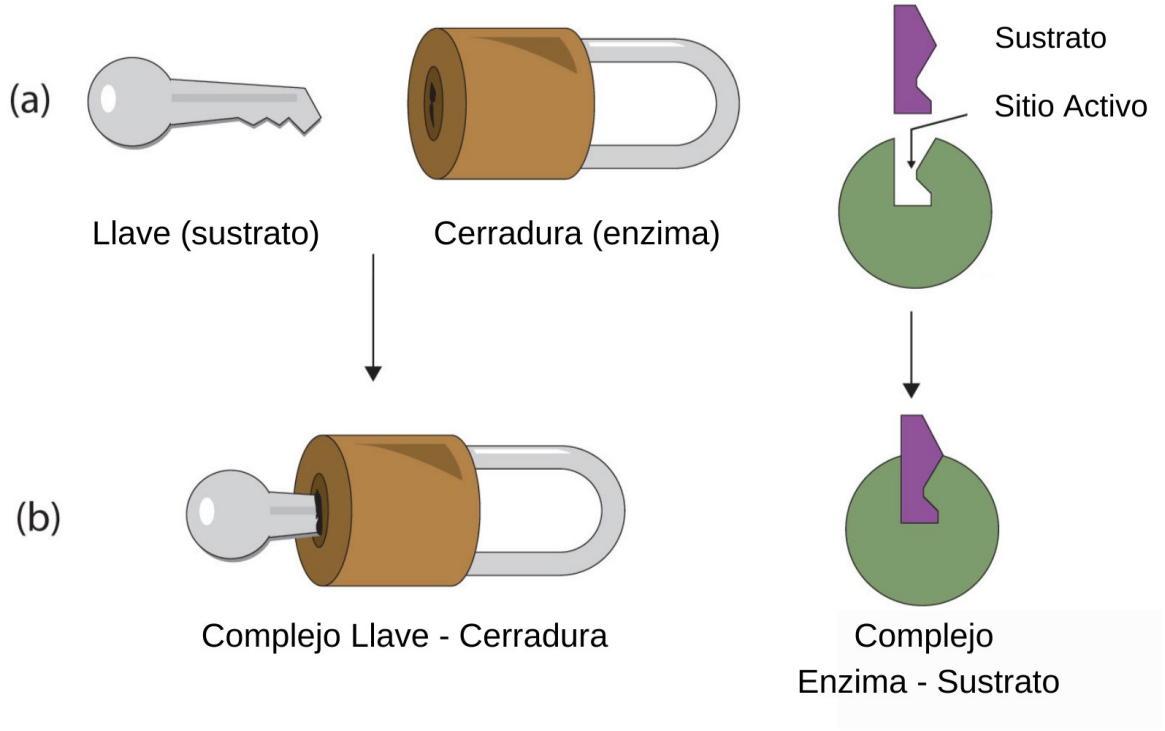


Figura 1.2 Representación esquemática del modelo llave cerradura.

El sustrato encaja de manera perfecta en el sitio activo de la proteína, como si se tratara de una llave que debe unirse correctamente a una cerradura para poder abrirla. La complementariedad geométrica entre sustrato (llave) y la proteína (cerradura) es condición necesaria para la correcta unión y posterior catálisis. Figura tomada y adaptada de https://saylordotorg.github.io/text_the-basics-of-general-organic-and-biological-chemistry/s21-06-enzyme-action.html

Posteriormente en 1936 Pauling y Mirsky (Mirsky and Pauling, 1936), enuncian que el estado nativo de una proteína consiste de una cadena polipeptídica que se pliega dando lugar a una configuración única, definida y estática y es a esta configuración a la cual se le atribuyen las características particulares de la proteína. Esa única estructura que determina el estado nativo de una proteína se corresponde con un mínimo de energía

libre. Si se asume esta hipótesis, es posible representar un diagrama de energía libre tal como se puede observar en Figura 1.3 en donde existe un único mínimo de energía libre que se corresponde a la única estructura del estado nativo.

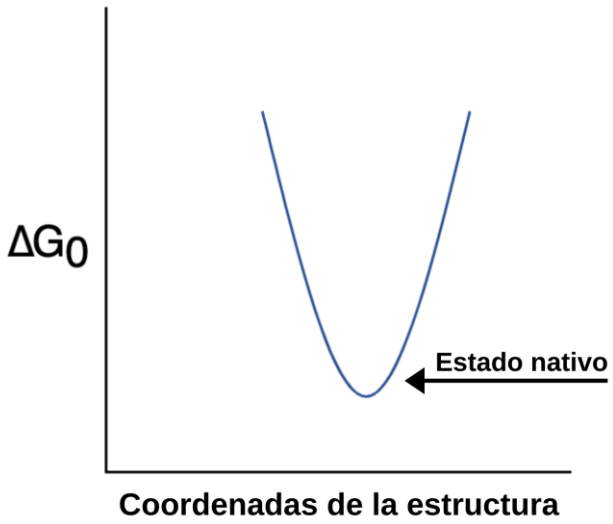


Figura 1.3 Representación del estado nativo formado por una única conformación.

El mínimo de energía libre corresponde a la estructura del estado nativo. Imagen tomada y adaptada de (James and Tawfik, 2003).

El problema con esta visión es que no permitía explicar determinados fenómenos observados experimentalmente como la unión de anticuerpos (que tenían casi la misma secuencia) a una gran variedad de antígenos. Pauling en 1940 (Pauling, 1940) propone que las proteínas debían existir como un conjunto de distintas estructuras en las que se podía plegar la cadena polipeptídica y que al menos una de ellas se corresponde con la estructura funcionalmente activa.

En 1950 Karush (Karush, 1950) encuentra que la albúmina muestra distintas afinidades por una gran cantidad de moléculas pequeñas no polares que a su vez presentaban una diversidad enorme de formas. La explicación que propuso para este hecho fue que el sitio de unión a ligando en la albúmina debía presentar variaciones estructurales, es decir que no tenía una configuración única. Karush se refirió a este fenómeno como

adaptabilidad configuracional y lo atribuye al hecho que el sitio activo podía asumir distintas configuraciones que presentan energía similar y que se encuentran en equilibrio, lo que permitía la unión a distintos tipos de sustratos. Con esta incorporación, es posible entender al paisaje de energía libre de la proteína como uno que contiene múltiples mínimos de energía donde cada uno se corresponde con una configuración proteica determinada (Figura 1.4).

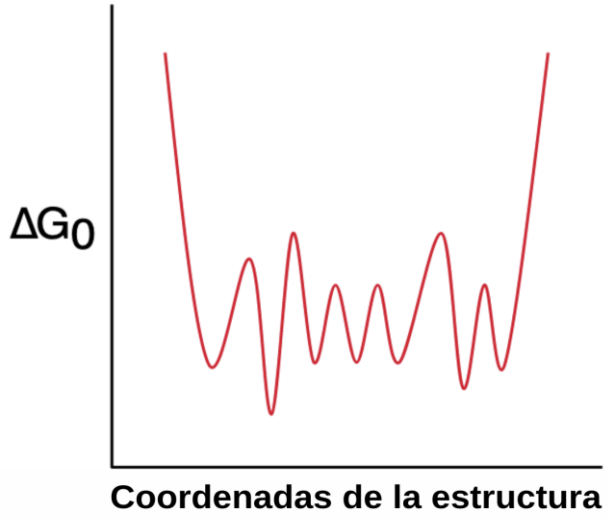


Figura 1.4 Paisaje de energía libre de una proteína.

Figura tomada y adaptada de (James and Tawfik, 2003).

Si bien los trabajos de Karush y Pauling sobre la flexibilidad del estado nativo fueron pioneros, ciertamente fueron ignorados o simplemente desconocidos en los espacios académicos. Sin embargo, no era desconocido que muchas proteínas o enzimas mostraban cierta promiscuidad de sustratos, esto es que una misma proteína podía interactuar o catalizar sustratos ligeramente diferentes. En 1958 Koshland (Koshland, 1958) propone el modelo de ajuste-inducido (Figura 1.5) en donde se establece que las estructuras de las proteínas son 'flexibles' y que por lo tanto el sitio activo de una enzima podría 'ajustar' su forma para la correcta unión a un determinado sustrato. A diferencia de lo propuesto en el modelo de llave cerradura, en el modelo de ajuste inducido ya no es necesaria la preexistencia de complementariedad geométrica entre sustrato y

proteína. Lo que Karush observó como *adaptabilidad configuracional*, Koshland lo define como ajuste inducido.

El modelo de ajuste inducido se basa en las siguientes premisas que Koshland propone en su trabajo:

(a) Para que exista una acción enzimática precisa tiene que existir 'orientación' determinada del sitio activo

(b) El sustrato podría alterar las relaciones tridimensionales entre los aminoácidos que forman el sitio activo. Es decir que la presencia de sustrato 'induce' un cambio de orientación y por lo tanto fisicoquímico del sitio activo.

(c) Ese cambio en el entorno del sitio activo provoca la correcta orientación de los residuos catalíticos y como consecuencia la reacción enzimática podría llevarse a cabo.

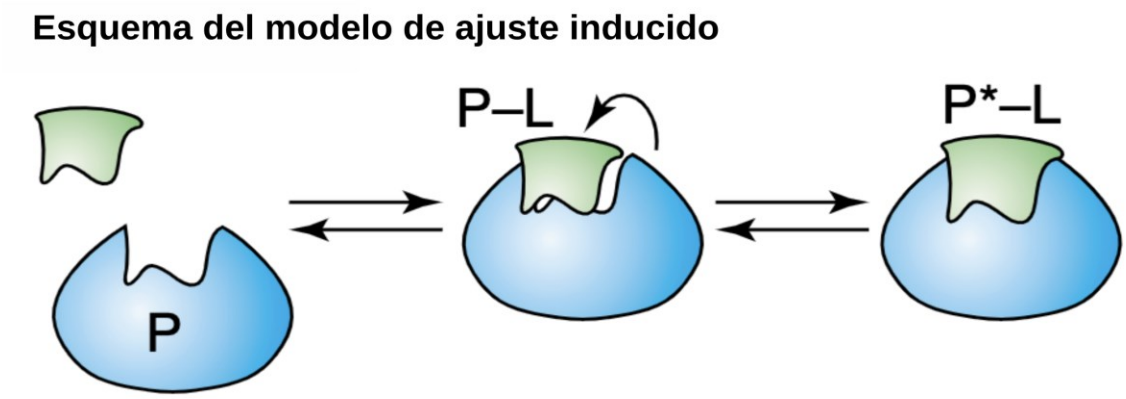


Figura 1.5 Representación esquemática del modelo de ajuste inducido.

P representa a la proteína sin unión al ligando lo que constituiría su estado nativo único o de mínima energía. P-L estado de transición en que el ligando L induce un cambio en el sitio activo de la proteína P. P-L complejo proteína-ligando. Figura extraída y adaptada de (James and Tawfik, 2003).*

Con este modelo fue posible explicar fenómenos observados tales como inhibición competitiva y la promiscuidad enzimática (comprendida como la posibilidad de que una misma enzima puede catalizar reacciones sobre sustratos de distintas formas y tamaños). Koshland en su propuesta no deja de lado la necesidad de la especificidad

entre el sustrato y la enzima a través del sitio activo, pero introduce términos referidos a la dinámica y la flexibilidad de la proteína hasta el momento no habían sido tenidos en cuenta. La visión 'clásica' que describe al estado nativo de las proteínas como una estructura estática, comienza a ser dejada de lado para dar lugar a una interpretación más rigurosa en donde la estructura tridimensional de las proteínas tiene características dinámicas que son clave para la función biológica.

Es interesante notar, que si bien los trabajos de Pauling, Karush y Koshland (anteriormente citados) sentaron las bases para la interpretación actual del estado nativo, en ese momento se utilizaba la expresión 'configuraciones' para referirse a cada una de las estructuras alternativas que podía adoptar una proteína. En la actualidad el término 'configuraciones' fue reemplazado por el término técnico 'conformaciones' y será el utilizado de aquí en adelante en esta tesis. Brevemente, las conformaciones son las distintas estructuras que puede tomar una molécula por rotación de sus enlaces simples mientras que las configuraciones son las distintas estructuras que puede tomar una molécula por ruptura y formación de enlaces covalentes (cómo clásico ejemplo los isómeros cis y trans del 1,2 dicloeteno).

En 1965 Monod, Wyman y Changeux (Monod et al., 1965) proponen un nuevo modelo de relación estructura-función en proteínas denominado modelo de pre-equilibrio. Aquí la hipótesis es que las distintas conformaciones existen incluso antes de la presencia de ligando. Ese equilibrio puede ser desplazado hacia un lado u otro como cualquier otro equilibrio químico haciendo que una determinada conformación se encuentre en mayor proporción al interaccionar con un ligando específico (Figura 1.6).

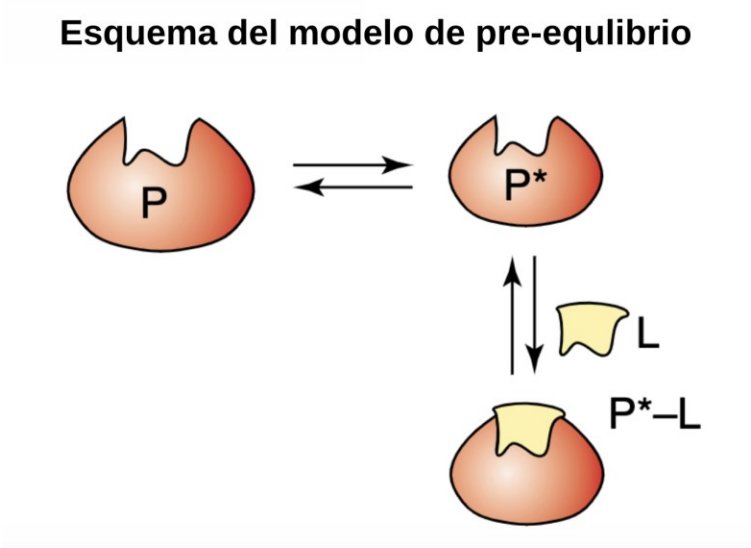


Figura 1.6 Esquema del modelo de pre-equilibrio.

Este modelo sostiene que existen conformaciones P (forma libre de la proteína) y P (confórmero de unión a ligando) que se encuentran en equilibrio y cuya existencia es previa a la presencia del ligando que solo puede unirse a la forma P*. La presencia del ligando desplaza el equilibrio de la forma P* al complejo (P*-L) (James and Tawfik, 2003).*

Tanto el modelo de ajuste inducido propuesto por Koshland como el de pre-equilibrio de Monod, incorporaron la dinámica estructural de las proteínas como factor determinante para comprender la función biológica. La mioglobina es un ejemplo sencillo para ilustrar la importancia de la dinámica proteica en la función biológica (Kendrew et al., 1958) ya que se trata de una proteína encargada de unir oxígeno en los músculos, gracias a la presencia de un grupo hemo. Si se entiende que la función principal de esta proteína es la de captar y liberar oxígeno, es claro que esta debe ser capaz de alterar su constante de disociación por el oxígeno, fenómeno solamente posible bajo la existencia de cambios conformacionales. En su trabajo “Myoglobin: The hydrogen atom of biology and a paradigm of complexity”, Frauenfelder un histórico investigador de la estructura y dinámica de la mioglobina, sugirió que deberíamos entender a las proteínas como castillos (Figura 1.7). Como sabemos los castillos están rodeados de una fosa (que la podemos pensar como la superficie de la proteína) y en su interior tienen patios (que serían las cavidades de las proteínas donde en general se ubican sitios de unión o sitios catalíticos). A los patios del castillo se accede por portones (en el caso de proteínas

estarían representados por túneles o bolsillos que son conductos que regulan el tránsito hacia y desde el interior/exterior de la proteína). La entrada y salida de los ligandos en una proteína están entonces regulados por la “apertura” o “cierre” de compuertas involucrando necesariamente cambios conformacionales.

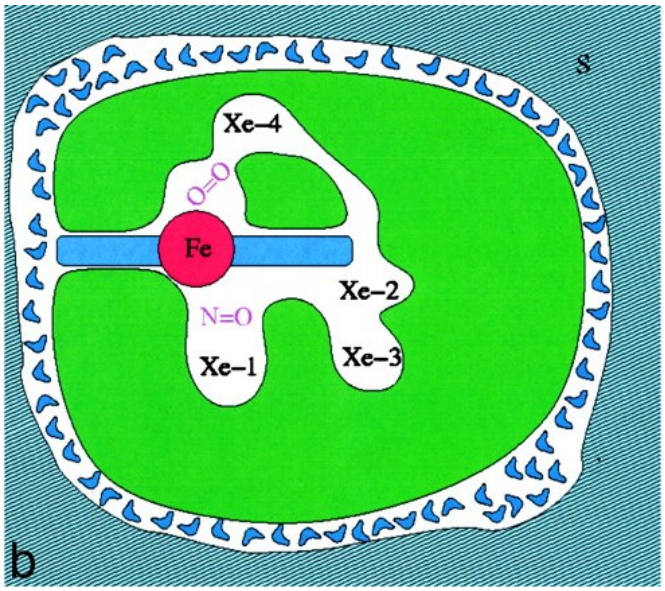


Figura 1.7 Representación de la estructura de la Mioglobina como si fuera un castillo.

La superficie de la proteína emula al foso del castillo mientras que sus cavidades (indicadas con Xe) a sus patios interiores. La barra azul con el símbolo del hierro indica el grupo hemo. A tales cavidades se accede por túneles que pueden estar abiertos o cerrados según sean los cambios conformacionales de la proteína. La proporción de confórmeros cerrados/abiertos a una determinada temperatura define la constante de afinidad de la proteína por el ligando. Figura extraída de (Frauenfelder et al., 2003).

De este breve apartado histórico sobre el concepto de estado nativo de las proteínas, se puede concluir que todos los modelos fueron concebidos para poder explicar diversos mecanismos de unión proteína/ligando e influencias de las condiciones fisicoquímicas del entorno. Hoy en día el modelo de ajuste inducido y el pre-equilibrio siguen siendo utilizados para explicar dichos fenómenos (Di Cera, 2020). Más recientemente se está reexaminando aún más el término de estructura nativa de una proteína, sobre todo debido a las particularidades de las proteínas desordenadas y las proteínas que sufren procesos de separación de fases. En particular, las proteínas que participan en la formación de organelas sin membrana (Shin and Brangwynne, 2017), parecen tener

funciones emergentes no estrictamente derivadas de una relación secuencia-estructura-función, ya que dichos agregados son no-estequiométricos haciendo vago el término “estado nativo”. Nuevamente la ausencia de estructura definida por parte de las IDPs agrega inestabilidad en la idea de estado nativo que sujeto a nuevos descubrimientos parece no converger conceptualmente (Pancsa et al., 2019).

1.1.2 Plegamiento de proteínas: teoría del paisaje energético

En la sección previa se mencionaron distintos modelos desarrollados para explicar cómo está definido el estado nativo de una proteína, pero ¿cómo es en términos estructurales el estado funcional de cada proteína? En este sentido también es válido preguntarse cómo una cadena de aminoácidos adopta ese estado funcional. Si se tiene una proteína de 100 aminoácidos donde cada uno puede adoptar 2 conformaciones distintas (se sabe que este número en la realidad es mucho mayor, pero se utiliza a los efectos de ilustrar) esta proteína tendría una cantidad de $1.3 * 10^{30}$ conformaciones distintas a las que explorar. Si se asume que la proteína se va a plegar de manera espontánea en un proceso al azar en el cual debe explorar las $1.3 * 10^{30}$ distintas conformaciones por ensayo y error y en donde que tarda 10 nanosegundos (este tiempo es una aproximación razonable del orden de una reacción química de transferencia de un protón) en explorar cada conformación posible, la proteína tardaría en plegarse correctamente en un proceso al azar en unos $1,3 \times 10^{22}$ segundos, es decir 400 billones de años. Esta es una versión simplificada (si se tiene en cuenta que las proteínas suelen estar formadas por entre 300 y 700 aminoácidos y que las conformaciones posibles en cada enlace son muy dispares y más que simplemente 2). Sin embargo, en la naturaleza las proteínas se pliegan en el orden de los micro o milisegundos y hasta unos pocos segundos las más grandes Esta inconsistencia se conoce como la paradoja de Levinthal (Levinthal, 1968).

Si se asume que todas las conformaciones son exploradas previo a encontrar a la o las que corresponden al estado nativo, y que la energía de la proteína en dicho estado es mucho menor que en el resto de las conformaciones, entonces la superficie potencial de

una proteína que se pliega podría representarse como en la Figura 1.8. Esta interpretación del proceso de plegado fue referida como la *Teoría del golfista ciego*, por analogía a la dificultad que tendría un jugador de golf ciego de encontrar el hoyo en un campo de golf plano.

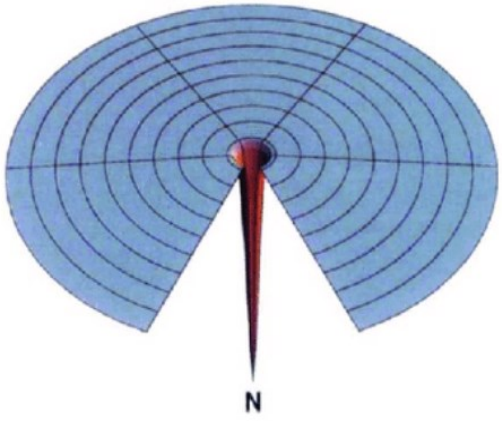


Figura 1.8 Superficie de energía potencial - Teoría del golfista ciego

*Superficie de energía potencial de una proteína con un espacio conformacional enorme y un estado nativo de energía mucho menor que el resto de las conformaciones. El punto **N** marca el estado nativo. Cada punto de la superficie describe un confórmero de la proteína.*

Sin embargo, las proteínas no exploran todas las conformaciones, sino que deben explorar un espacio reducido de conformaciones (o caminos) hasta llegar al estado nativo. Levinthal llama a este conjunto de caminos como *folding pathways*, es decir caminos que condujeran a la proteína a la conformación del estado nativo (Figura 1.9).

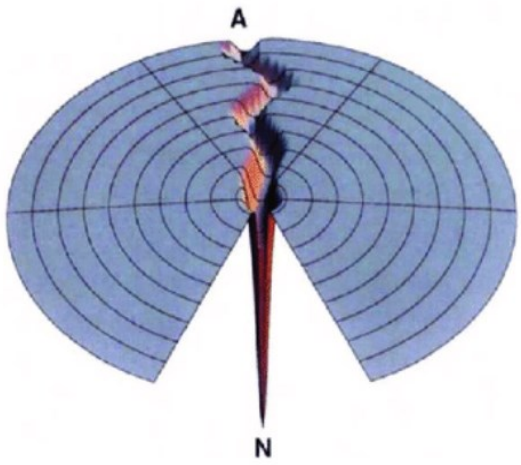


Figura 1.9 Camino de plegado (Folding pathway) según Levinthal.

***A** representa la estructura de partida que sigue una única ruta de plegamiento, con sus intermediarios, hasta la estructura nativa **N**.*

La siguiente pregunta a responder fue si este camino o *pathway* era único (Harrison and Durbin, 1985). En el trabajo (Bryngelson and Wolynes, 1989), proponen que en vez de existir caminos (*pathways*) únicos, debían existir caminos alternativos para llegar al estado nativo, ahora el proceso podría pensarse como un embudo de plegado (Bryngelson y Wolynes lo denominaron *folding funnel*) (Figura 1.10).

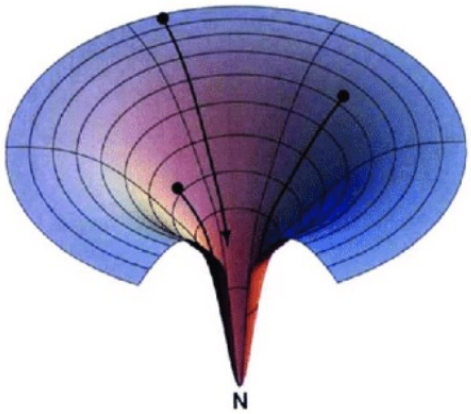


Figura 1.10. Embudo de plegado (*Folding funnel*) - cooperativo.

*Distintas estructuras iniciales desplegadas total o parcialmente pueden tomar diversos caminos (marcados con los puntos y líneas negras) para llegar al estado nativo **N**.*

La representación del embudo de plegado mostrado en la Figura 1.10, representa a un plegado totalmente cooperativo, casi instantáneo donde la proteína establece todo sus contactos nativos (que dan origen a la estructura tridimensional en un solo paso). La realidad es que el proceso de plegado implica la formación de contactos transitorios que luego se rompen para formar contactos nativos, por lo que una representación más fidedigna del proceso es la que se muestra en la Figura 1.11 y Figura 1.12.

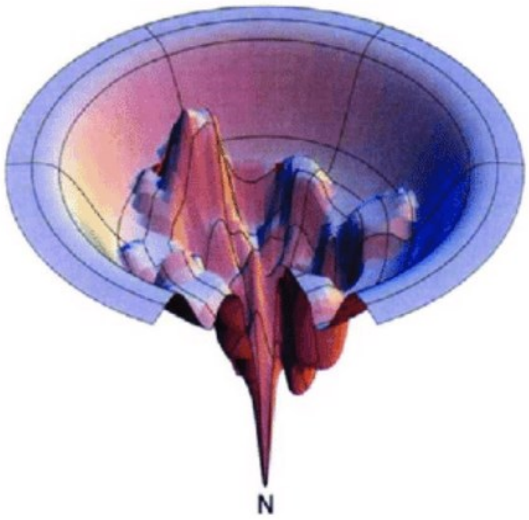


Figura 1.11 Embudo de plegado (*Folding Funnel*).

Los máximos representan estados de transición entre distintos intermediarios del proceso de plegado. También existen depresiones que representan intermediarios que son estables pero que no son el estado nativo.

La Figura 1.12. muestra un corte transversal de este embudo, donde la rugosidad del fondo del 'embudo de plegado' implica la existencia de distintas conformaciones nativas que se encuentran en equilibrio dinámico en cada mínimo local.

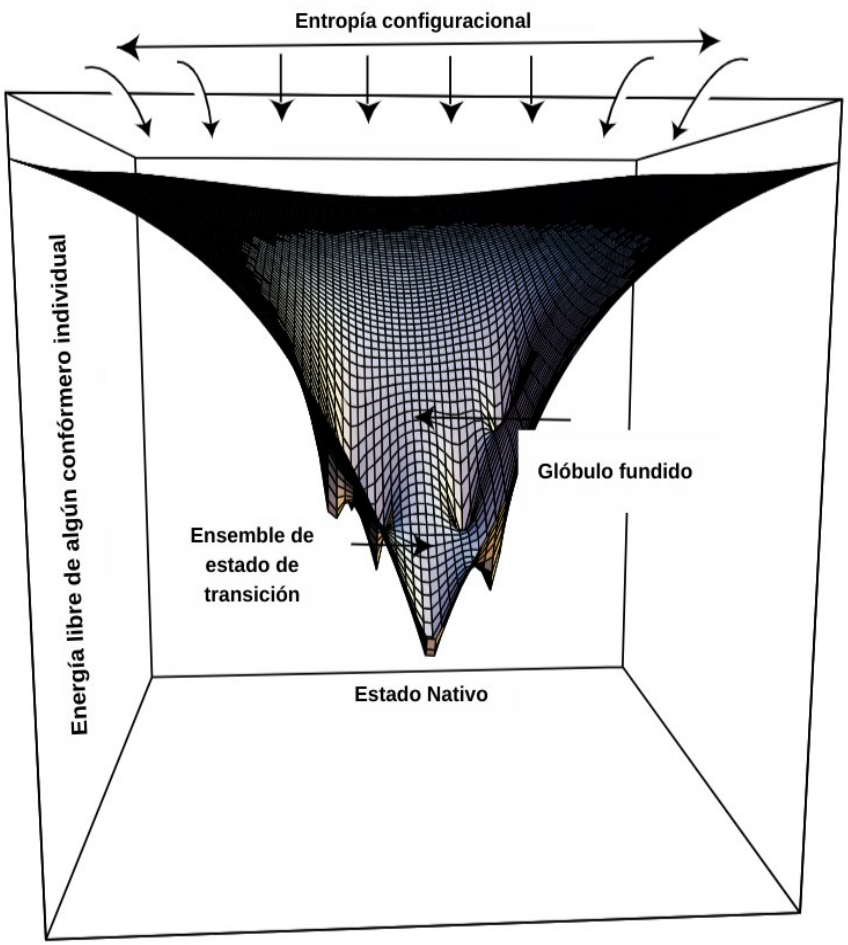


Figura 1.12 Plegado proteico y el paisaje energético.

Bajo esta interpretación no existe un único camino que conduzca al correcto plegado de la cadena polipeptídica. En una superficie con forma de embudo, existen distintas rutas en las cuales es probable que la energía de la proteína disminuya hasta llegar al fondo rugoso de dicho embudo. Figura extraída y adaptada de (Onuchic et al., 1996).

Las diferencias energéticas entre los distintos confórmeros nativos pueden ser grandes o pequeñas (hecho reflejado en la altura de las barreras energéticas que separan cada uno de los estados conformacionales en el paisaje de energía libre), pero todas estos confórmeros representan estructuras a las cuales puede acceder una secuencia proteica. Los confórmeros no se encuentran igualmente representados en su entorno fisiológico y esto se debe principalmente a la energía relativa de cada una de dichas estructuras (Bryngelson and Wolynes, 1989) (Frauenfelder et al., 1991).

La idea de paisajes energéticos fue posteriormente extendida a una concepción dinámica donde se considera el entorno en el cual se encuentra una proteína. El entorno puede modificar el paisaje energético y entonces alterar la rugosidad del fondo del embudo y cambiar la distribución de las distintas poblaciones de confórmeros (Kumar et al., 2000). El efecto del entorno puede ser sobre las barreras energéticas que separan a los distintos confórmeros haciéndolas más bajas o más altas, es decir siendo más fáciles o más difícil de interconvertir (Figura 1.13, panel A) o bien cambiando la estabilidad relativa de los distintos confórmeros ver (Figura 1.13, panel B). Ambos esquemas describen lo que se denominó como *population-shift*.

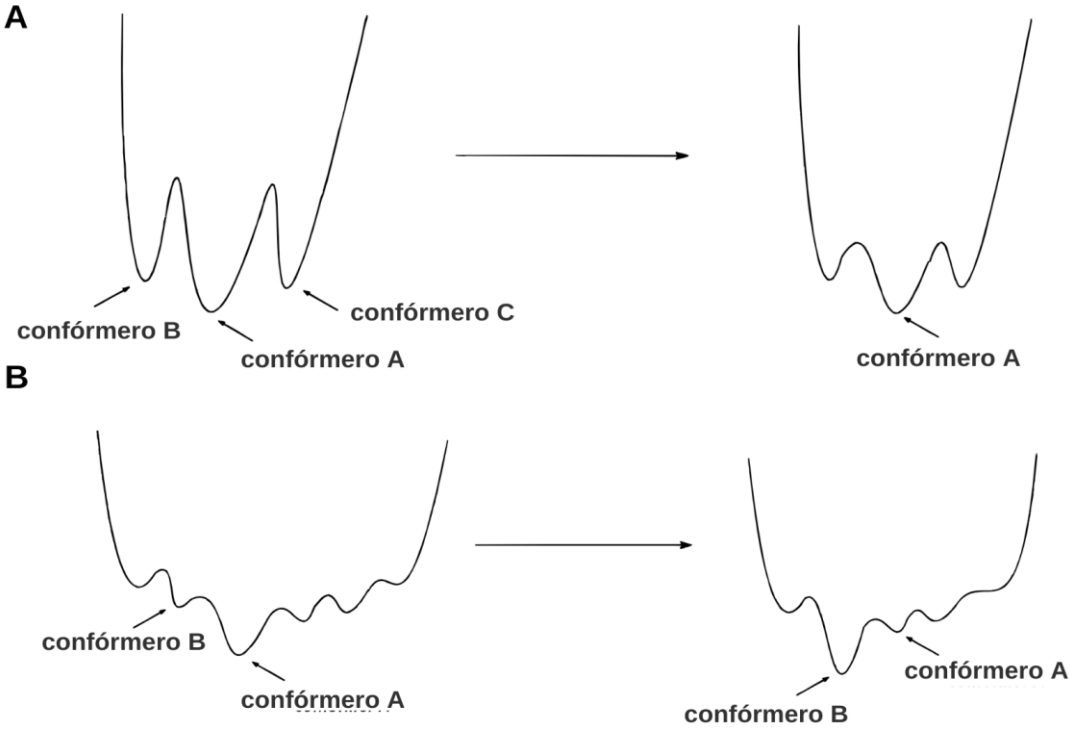


Figura 1.13 Esquema representativo del proceso de population shift.

En el panel A) de la figura se muestra cómo el entorno altera el paisaje de energía haciendo que las barreras energéticas que separan los estados conformaciones sean más bajas. En el panel B) de la figura se muestra un esquema de cómo el entorno podría alterar la proporción de las distintas poblaciones de confórmeros. Ambos esquemas describen lo que en conjunto se denominó population shift. Imagen extraída y adaptada de (Kumar et al., 2000).

1.1.3 El estado nativo hoy: el ensemble conformacional

La visión actual del estado nativo, integra lo abarcado en la sección 1.1.1 y en la sección 1.1.2 y se concibe como una colección de estructuras o confórmeros que se encuentran en equilibrio dinámico. Cada uno de los diferentes confórmeros representan arreglos estructurales que puede adoptar una misma secuencia proteica y que se interconvierten sin ruptura de enlaces covalentes. Estas conformaciones pueden corresponder con un mínimo de energía libre local. A esta colección de confórmeros se la denomina como ensemble nativo (del inglés *native ensemble*) de una proteína (Tsai et al., 1999) (Wei et al., 2016). La primera evidencia experimental de la existencia de ensembles conformacionales surgió por la cristalización de la hemoglobina con sus dos estados T (desoxigenada) y R (oxigenada) (Perutz, 1976) donde cada una de las dos conformaciones cumple un rol específico en relación al proceso de unión del oxígeno y tiene, una vinculación directa con el desplazamiento del equilibrio químico de dicho proceso (Figura 1.14).

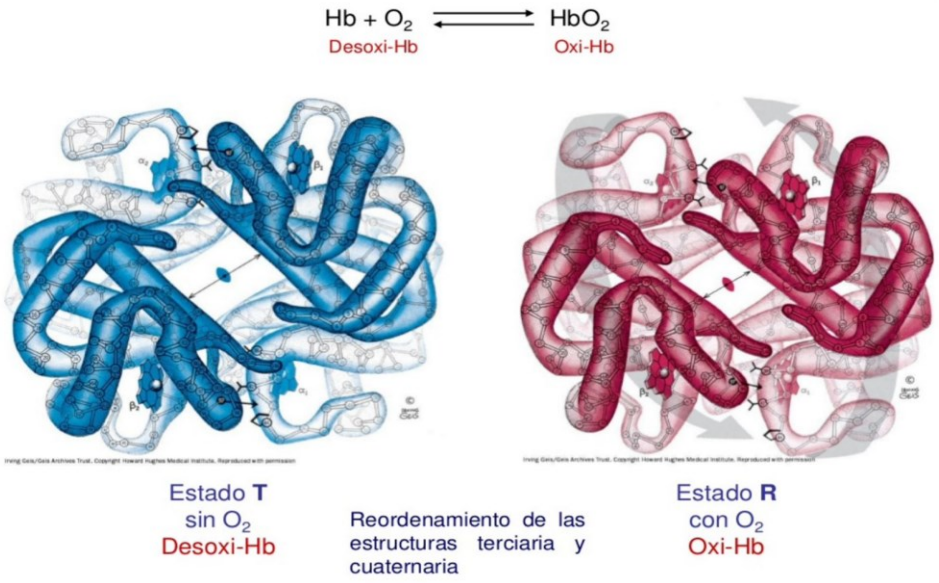


Figura 1.14 Esquemas de las formas T y R de Hemoglobina.

Modificado de la figura original de Irving Geis, The Geis Archives, Howard Hughes Medical Institute (http://www.hhmi.org/).

Como se mencionara en la sección anterior, las distribuciones relativas de los confórmeros son entorno dependientes y de allí que la función biológica de una proteína va a estar también determinada por la distribución de los distintos sub-estados conformacionales y de su redistribución frente a cambios en las condiciones del entorno (Kumar et al., 2000). Esta noción del estado nativo permite explicar muchos fenómenos tal como los observados por Milstein y colaboradores (Foote and Milstein, 1994), donde encontraron que existían distintas conformaciones (llamados isómeros en su trabajo) en los anticuerpos, la comprensión de procesos de alosterismo y cooperativismo (Tsai and Nussinov, 2014) (Cui and Karplus, 2008), la función biológica de las proteínas (Eisenmesser et al., 2005), fenómenos de reconocimiento molecular (Boehr et al., 2009; Nussinov et al., 2013), la catálisis enzimática y el origen de nuevas funciones (Tokuriki and Tawfik, 2009).

1.1.4 Cuantificación de la diversidad conformacional y distribución de movilidad

Los distintos confórmeros que definen un ensemble nativo pueden diferir poco o mucho en su estructura terciaria y estas diferencias estructurales se describen como la diversidad conformacional del ensemble o entre confórmeros, y va a depender de la cantidad de confórmeros diferentes en un ensemble nativo, a su vez el número de mínimos en el fondo del embudo rugoso reflejando las estabilidades de estas estructuras alternativas que se interconvierten entre sí. Una forma de cuantificar la diversidad conformacional es a través de la superposición estructural de confórmeros. En su forma más sencilla que es la superposición de cuerpo rígido, se comienza con un alineamiento de las secuencias para asignar los residuos equivalentes, seguido de una traslación al mismo centro de masa y una posterior rotación de una de las estructuras. Estas rotaciones se realizan hasta minimizar la distancia cuadrática media (**R**oot **M**ean **S**quare **D**eviation, RMSD por sus siglas en inglés) entre los átomos equivalentes de las estructuras, con el algoritmo de Kabsch (Kabsch, 1978). Normalmente esta

superposición es efectuada solamente con las coordenadas espaciales de los Cα de la proteína (Figura 1.15).

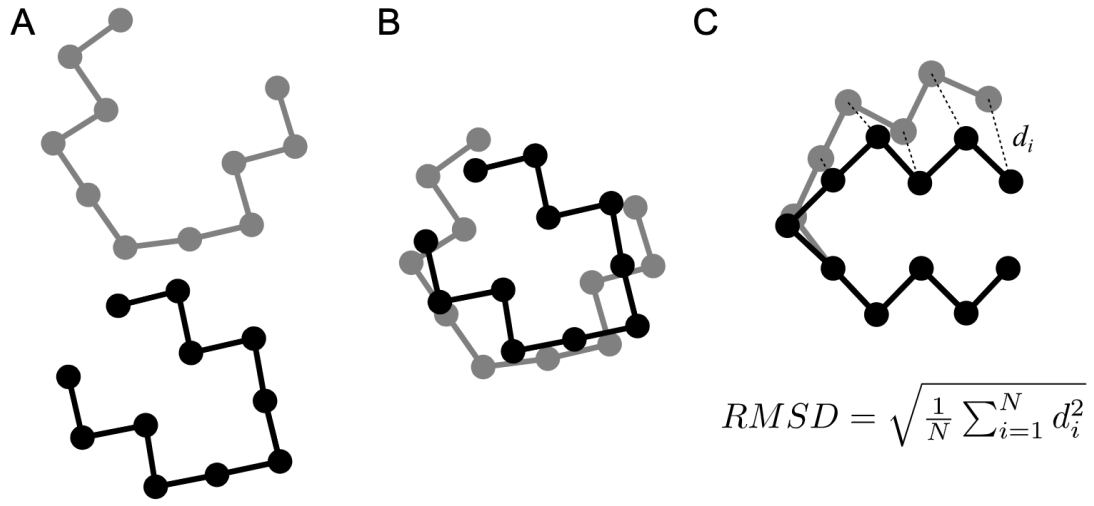


Figura 1.15 Esquema de una superposición estructural de cuerpo rígido.

*En el panel A se muestran dos conformaciones de una misma proteína (uno en color negro y otro en color gris). Estas dos conformaciones se trasladan al mismo centro de masa (panel B de la figura) para luego rotar cada una de las estructuras para de manera tal de obtener el mínimo valor de RMSD de los **Cα** de los residuos equivalentes, panel C. Esta imagen fue extraída de (Zea, 2015).*

El RMSD puede calcularse para cualquier par de estructuras superpuestas, según la Ecuación 1.

Ecuación 1: Cálculo de RMSD.

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N} d_i^2}{N}}$$

Donde **di** es la distancia entre un par átomos equivalentes, normalmente expresada en Ångströms (Å). Si dos estructuras son idénticas el RMSD correspondiente toma un valor de 0 Å, y su valor aumentará a medida que las estructuras difieran entre sí. Es por eso

que se considera a la medida del RMSD como un indicador confiable de variabilidad estructural cuando se calcula entre confórmeros de una misma proteína. Como se puede deducir de la fórmula, dado que el RMSD depende del número de átomos incluidos, N, los valores calculados para pares de estructuras de diferentes longitudes no pueden ser directamente comparados (Irving et al., 2001). Para evitar esta dependencia es posible usar un RMSD normalizado a 100 residuos, según Ecuación 2.

Ecuación 2: Cálculo de $RMSD_{100}$.

$$RMSD_{100} = \frac{RMSD}{1 + ln\ ln \sqrt{\frac{N}{100}}}$$

Donde RMSD está calculado según la Ecuación 1 para una superposición estructural de Cα, y N es el número de aminoácidos superpuestos (Carugo and Pongor, 2001).
Otra medida de variabilidad estructural interesante y que es utilizada en esta tesis es el RMSF (**R**oot **M**ean **S**quare **F**luctuation, RMSF por sus siglas en inglés), que está dado por el valor de RMSD promediado en el tiempo. El RMSF es útil cuando el sistema es dinámico y fluctúa con el tiempo .

Como ya se mencionó anteriormente dentro de un ensemble conformacional existen diferencias estructurales entre los distintos confórmeros y estas puede deberse a diferentes aspectos, como por ejemplo, el movimiento relativo de dominios (Gerstein and Echols, 2004), re-arreglos de elementos de estructura secundaria y terciaria (Gerstein and Krebs, 1998) o movimientos de loops (Gu et al., 2015) (Figura 1.16). La magnitud de las diferencias estructurales, causadas por los movimientos antes mencionados están directamente relacionados con la diversidad conformacional. Para poder medir cuál era la magnitud de estas diferencias, Burra y colaboradores en su trabajo (Burra et al., 2009) usaron el RMSD de los Cα, para medir diferencias estructurales entre confórmeros. Se trató de un análisis a gran escala para el cual se utilizaron colecciones redundantes de

estructuras cristalizadas para una misma proteína, es decir que se se utilizaron colecciones de ensambles nativos. Se encontró que la diversidad conformacional de las proteínas presenta una distribución con un pico en ~ 0.5 Å (RMSD máximo observado: 23.7 Å). La distribución no era simétrica y tenía un sesgo hacia RMSD (Figura 1.17).

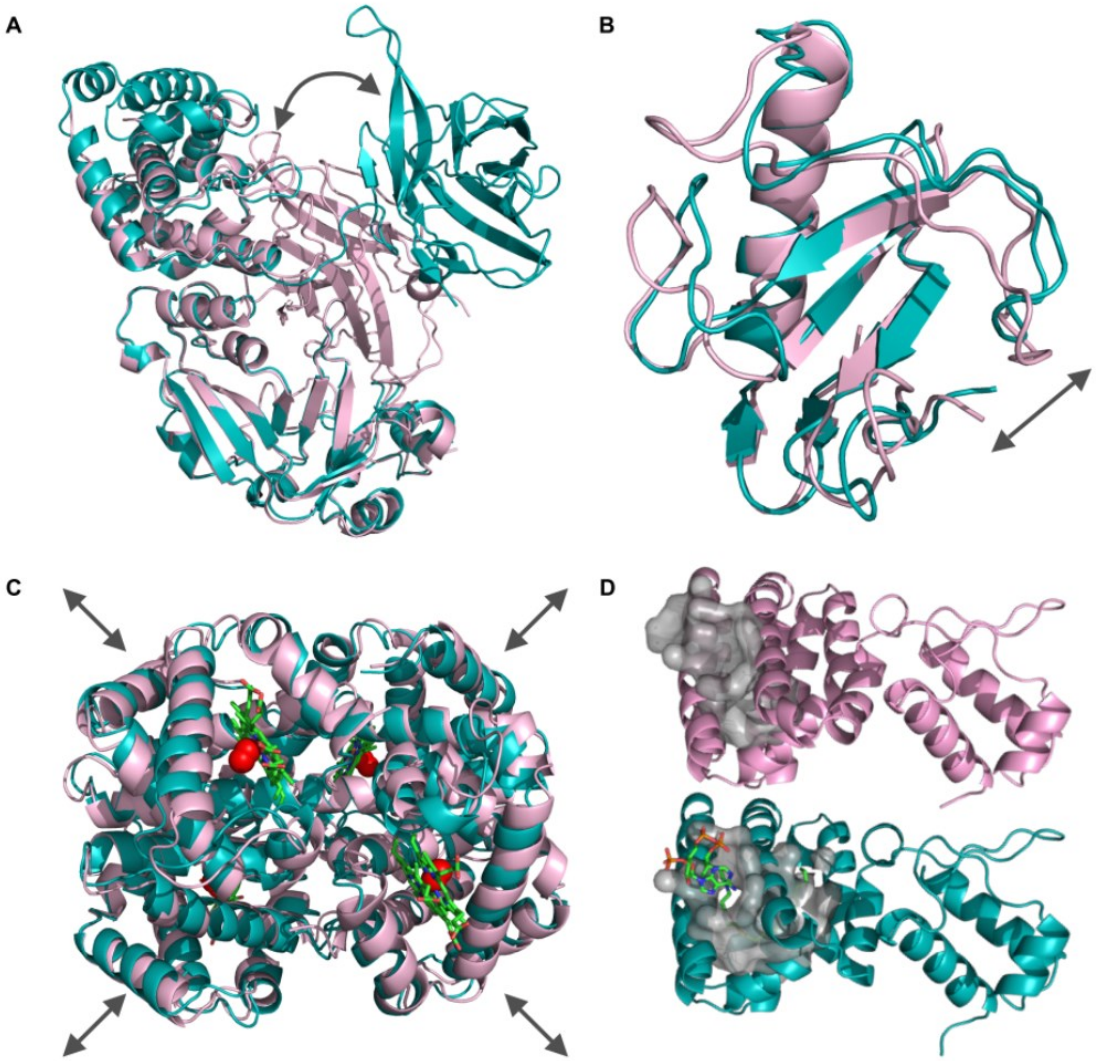


Figura 1.16 Ejemplos de diversidad conformacional en proteínas.

Panel A) Muestra la toxina de difteria en complejo con NAD (PDB: 1TOX_A) en celeste y cristalizada en condiciones ácidas en color rosa (PDB: 4OW6_A). El RMSD de ***Cα*** *entre los confórmeros en la región superpuesta es de 1.78 Å. Esta proteína muestra un movimiento de bisagra de hasta 65Å que permite la rotación del dominio indicada con la flecha gris. Panel B) Se observa a la ribonucleasa F1 en complejo con ácido piroglutámico en celeste (PDB 1RCK_A, modelo 21 obtenido por NMR) y la forma libre de ligando en rosa (PDB: 1FUS_A).*

Estas estructuras muestran movimientos de deslizamiento indicados por la flecha gris, debidos principalmente a la acción de las cadenas laterales de los residuos en la región, mostrando un RMSD de 1.99 Å. C) Se observa a la hemoglobina en su estado R unido a oxígeno en celeste (PDB: 1HHO) y en el estado T en rosa (PDB: 2HHB). Las dos conformaciones muestran una RMSD de 2.34 Å debido al movimiento de cuerpo rígido indicado por la flecha gris. D) Se observa al factor de transcripción FadR en la forma apo en rosa (PDB: 1E2X_A) y en el estado holo en complejo con miristoil-CoA en celeste (PDB: 1H9G_A) que presentan un RMSD de 1.28 A. La diferencia principal entre los confórmeros es un cambio en el volumen de la cavidad (1586A3) indicada por la superficie gris y el número de túneles que permiten al ligando acceder al sitio activo. Imagen extraída de (Parisi et al., 2015).

De la distribución presentada en la Figura 1.17 es posible inferir que la mayoría de las proteínas presentan una diversidad conformacional promedio que es comparable con el error experimental de la cristalografía de rayos X (aproximadamente 0.4 Å (Kuriyan et al., 1987)). Si se entiende el RMSD entre confórmeros de una misma proteína como una 'medida' de la importancia del movimiento y la dinámica de la proteína para cumplir con su función biológica, con los resultados previamente comentados se puede concluir que una buena parte del universo proteico no necesita grandes movimientos que implican cambios a nivel estructural o incluso ninguno para poder llevar a cabo correctamente su función biológica. Se volverá a estas observaciones en la sección siguiente.

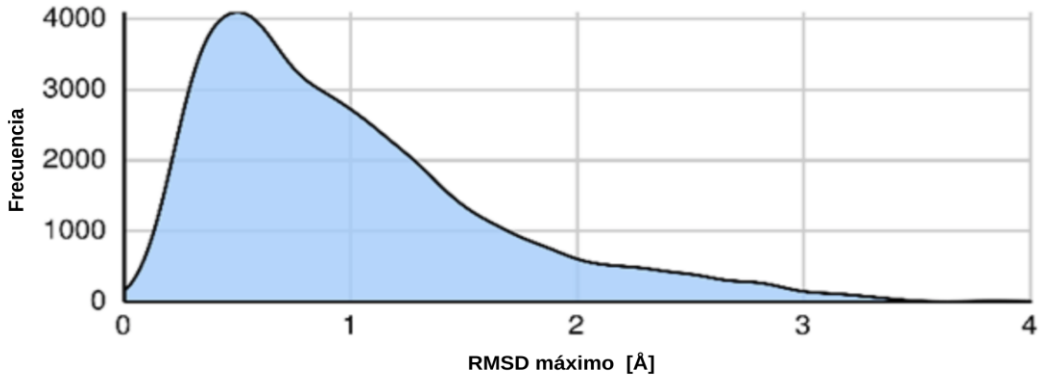


Figura 1.17 Distribución de la diversidad conformacional máxima para una proteína.

Distribución de diversidad conformacional de pares de confórmeros que tuvieran el máximo valor de RMSD. Figura tomada y adaptada de (Monzon et al., 2017a).

1.1.5 Proteínas Intrínsecamente Desordenadas (IDPs)

Como se puede observar de las secciones anteriores, los trabajos referidos a tratar de elucidar la relación entre la estructura y la función biológica de las proteínas fueron planteados con base en proteínas globulares, centrales y paradigmáticas en los libros de texto de bioquímica. Los experimentos de desnaturalización y renaturalización llevados a cabo por Anfinsen (Anfinsen, 1973), demostraron que la pérdida del estado '*plegado'* implicaba una pérdida de la función biológica. El paradigma establecido de la biología estructural sostenía entonces que la existencia de una estructura plegada es una condición necesaria para la función biológica. Estas ideas estaban condicionadas por las herramientas experimentales de la época, como la cristalografía de rayos X. Sin embargo, con esta misma metodología, se observó que átomos de ciertas proteínas no producían una señal de difracción nítida siendo imposible la determinación de sus coordenadas. Esto podía suceder debido a errores en la técnica experimental tales como defectos en el cristal o alteraciones en el proceso de purificación de la proteína. Sin embargo, otras de las razones por las cuales algunos átomos, cadenas laterales o residuos en una proteína podrían no dar una señal precisa en su difracción es la variación en su posición en las distintas moléculas que componen el cristal. El concepto de desorden estructural en proteínas fue introducido hace alrededor 20 años (Romero et al., 1998) y aquellos residuos para los cuales no se puede determinar sus coordenadas se denominan *'missing'*[2] y se los asume, *'desordenados'* o desestructurados por la falta de constancia en sus coordenadas en las diferentes moléculas que componen el cristal *(DeForte and Uversky, 2016)*. Siguiendo la visión clásica de la relación estructura-función que sostiene que es necesaria una estructura para la función biológica de la proteína, se esperaba que estos residuos, regiones o segmentos sin coordenadas asignadas no tuvieran importancia sobre la actividad biológica. Sin embargo, fueron numerosos los trabajos que demostraron que estos residuos '*desestructurados'* eran esenciales para el correcto funcionamiento de la proteína (Bloomer et al., 1978) (Bode et al., 1978) (Alber

[2] En este trabajo optamos por seguir utilizando el término en inglés missing y no su traducción perdidos, ya que missing es el término que se usa en literatura y porque su traducción perdido no ilustra bien su propiedad.

et al., 1983) (Spolar and Record, 1994). Estas evidencias pusieron en jaque el paradigma de la necesidad de estructura para llevar adelante la función biológica (Figura 1.18).

Figura 1.18 Impacto en la Biología Estructural por el descubrimiento de las IDPs.

Figura extraída de (Chouard, 2011).

El reconocimiento de la existencia de regiones o incluso, proteínas completamente *desestructuradas* dio lugar al origen a una nueva concepción del paradigma relación estructura-función (Figura 1.19).

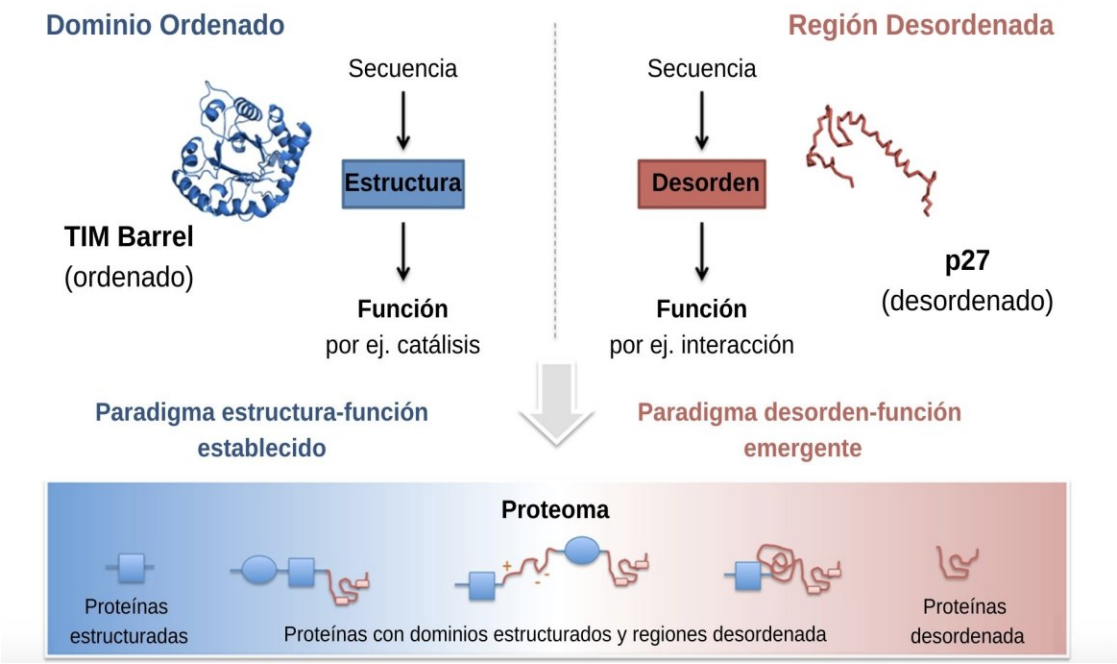


Figura 1.19 Esquema de la evolución del paradigma de relación estructura-función.

Figura extraída y adaptada de (van der Lee et al., 2014).

Desde la primera proteína cristalizada (Kendrew et al., 1958) en 1958 hasta la actualidad, el campo de la biología estructural creció enormemente, en parte debido al refinamiento y al desarrollo de nuevas técnicas para la obtención de información estructural, desarrollo de nuevas técnicas como como NMR, SAXS y *Single Molecule Spectroscopy*. El crecimiento de la base de datos PDB (Protein Data Bank, https://www.rcsb.org/) gracias a estas nuevas metodologías antes mencionadas, redundó en un aumento en la disponibilidad de información estructural que permitió la caracterización de ensembles conformacionales complejos. Con toda esta nueva información disponible, el universo de tipos estructurales proteicos se extendió más allá de las fronteras establecidas históricamente por las proteínas globulares: existen proteínas desordenadas (Tompa, 2002), repetitivas (Kajava, 2001), circulares (Trabi and Craik, 2002), anudadas (Mansfield, 1994), y la lista sigue creciendo.

Además, estas nuevas técnicas permitieron la determinación de estructuras imposibles de obtener por cristalografía de rayos X y por lo tanto fue posible mejorar y precisar la caracterización de los residuos desordenados de las proteínas, tanto que se pudieron

encontrar proteínas totalmente funcionales sin una estructura aparente: desordenadas de extremo a extremo (Daughdrill et al., 1997; Kriwacki et al., 1996) (Fletcher and Wagner, 1998) (Wright and Dyson, 1999). Las Proteínas Intrínsecamente Desordenadas (IDPs por sus siglas en inglés) son proteínas que se caracterizan por la carencia de una estructura terciaria estable bajo condiciones fisiológicas (Siltberg-Liberles et al., 2011) (Tompa, 2002). Lejos de tratarse de polímeros con una estructura azarosa o ensembles que podrían corresponder a estructuras simples como las *random coil*, es cada vez más evidente que los ensembles de las IDPs no son completamente desordenados y muestran una organización estructural transitoria de corto y largo alcance (Tompa, 2011) (Figura 1.20).

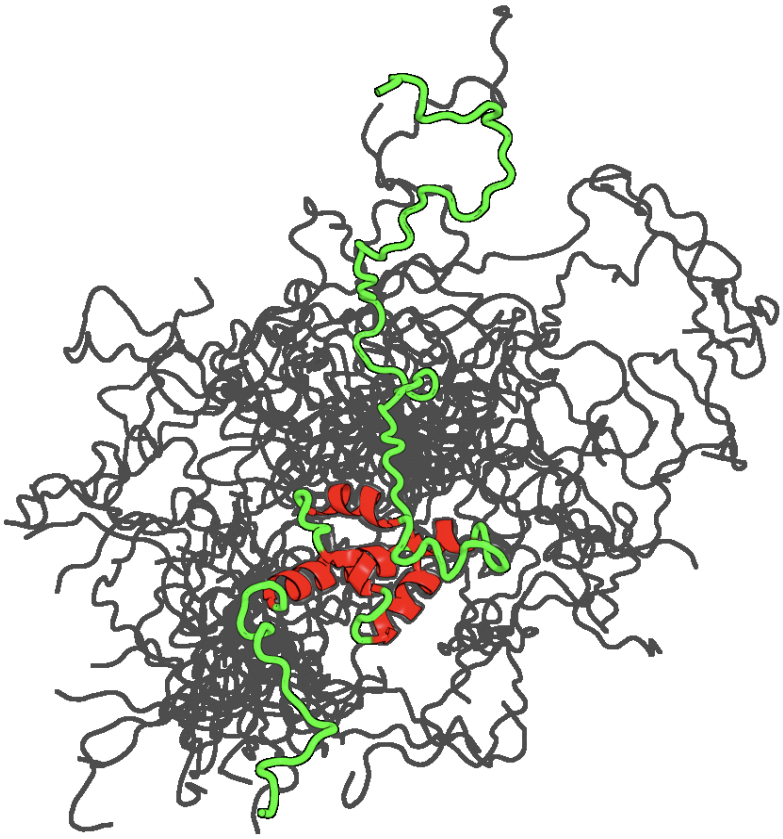


Figura 1.20. Representación de IDPs.

Representación de la estructura del dominio PABC de unión a la proteína poly(A) humana. Código PDB: 1G9L_A. Ensemble conformacional compuesto por 30 estructuras. En rojo y verde se muestra una de esas conformaciones y en gris el ensemble completo.

Los ensembles de las IDPs están compuestos por una gran cantidad de confórmeros y con pequeñas barreras energéticas existentes entre ellos (Varadi et al., 2014) y por lo tanto estas proteínas presentan un alto grado de diversidad conformacional que puede en algunos casos llegar a estar limitada por las interacciones entre residuos, originando de esta manera una mezcla compleja de confórmeros en el ensemble (Sormanni et al., 2017; Varadi et al., 2014) (Figura 1.21). Si se observa el panel superior derecho, se puede concluir que las proteínas estructuradas y ordenadas tienen paisajes energéticos con una conformación de preferencia. Por otro lado, si se observa el panel superior izquierdo, se puede ver que las proteínas desordenadas muestran en su estado nativo un paisaje energético rugoso (lo que implica la existencia de múltiples conformaciones) y plano (lo que indica la facilidad de interconversión entre los confórmeros).

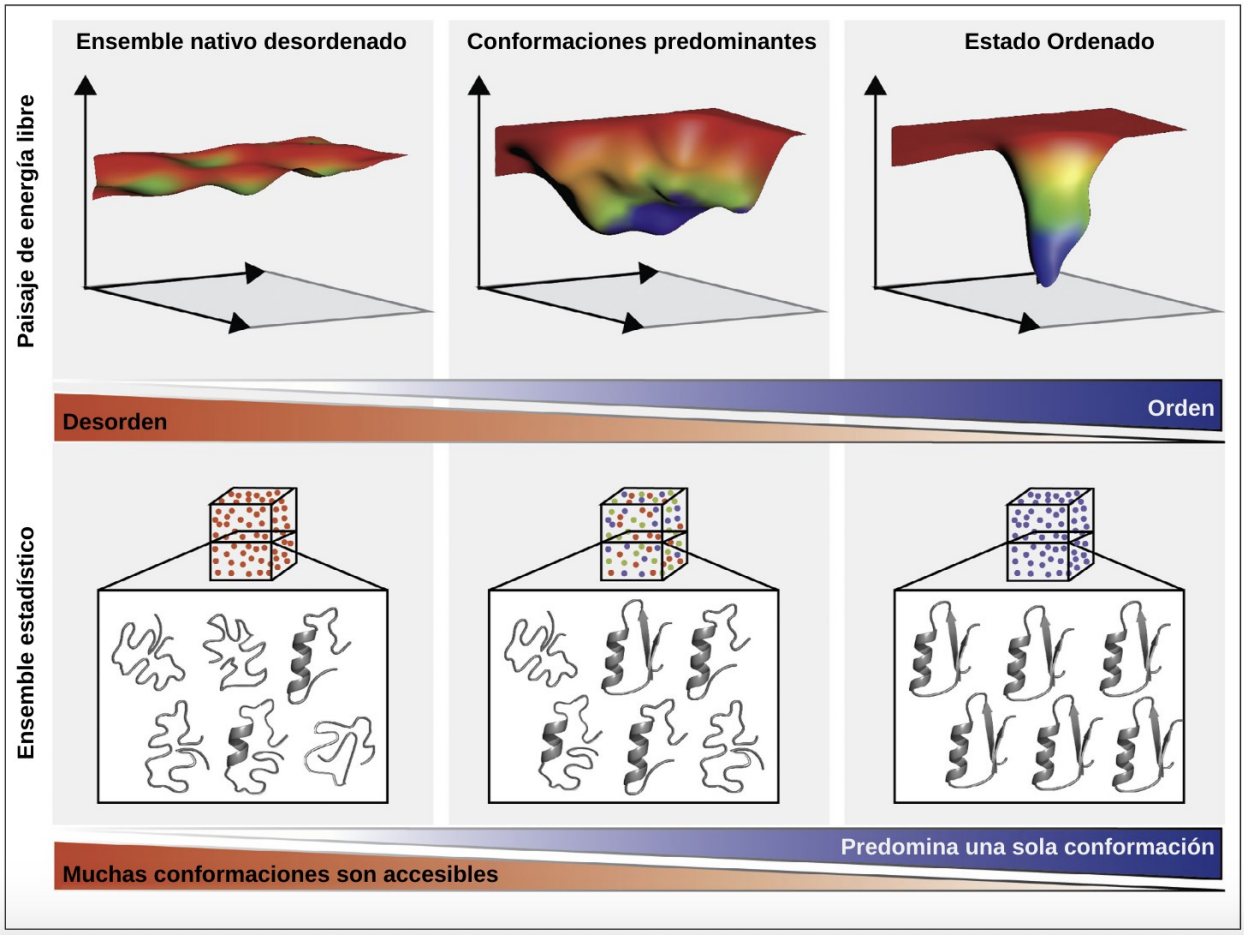


Figura 1.21 El paisaje de energía libre y los ensembles conformacionales de IDRs o IDPs.

El paisaje de energía libre representa esquemáticamente cualquier conformación posible (ejes x-y) que una proteína puede adoptar, así como su respectiva energía libre (eje z). En la fila superior se representa el paisaje de energía libre y describe la probabilidad de encontrar algún confórmero en un ensemble de estados conformacionales (ensemble estadístico, mostrado en la fila inferior). Imagen extraída y adaptada de (Flock et al., 2014).

Los ensambles nativos de las IDPs o regiones desordenadas IDRs son extremadamente complejos y su estudio supone un desafío para la biología estructural.

El desorden estructural, ya sea que esté presente en pequeñas regiones (IDRs) o en proteínas enteras (IDPs) se encuentra ampliamente distribuido. Se estima que entre el 10 y el 35% de proteínas procariotas y que el 15 al 45% de proteínas eucariotas poseen regiones desordenadas de al menos 30 residuos consecutivos de longitud (Peng et al., 2015) (Xue et al., 2012). Muchas de ellas cumplen roles claves para la célula principalmente relacionados con señalización celular y reconocimiento molecular (Habchi et al., 2014) (van der Lee et al., 2014) por lo que entender y conocer su biología resulta de vital importancia. En el trabajo de Dunker y colaboradores (Dunker et al., 2002) se tomó información curada manualmente de 150 proteínas que contenían al menos 30 residuos consecutivos desordenados y pudieron establecer 28 roles relacionados a la presencia de desorden entre los que se incluyen: interacción de proteína con proteínas, con ácidos nucleicos (ADN y ARNs). En el caso de interacción con ARNs se encuentran ejemplos que involucran ARN de transferencia (ARNt), ARN ribosomal (ARNr), ARN mensajero (ARNm) y ARN genómico. También se reconocen interacciones de estas proteínas con lípidos, unión a sustratos, cofactores y metales, y de sitios de modificación post-traduccional (acetilación, glicosilación, acetilación y fosforilación). En un estudio posterior (Xie et al., 2007) se confirmó este amplio repertorio funcional en la que están involucradas las IDPs o IDRs, que esquemáticamente, se pueden clasificar en tres grandes grupos (Figura 1.22).

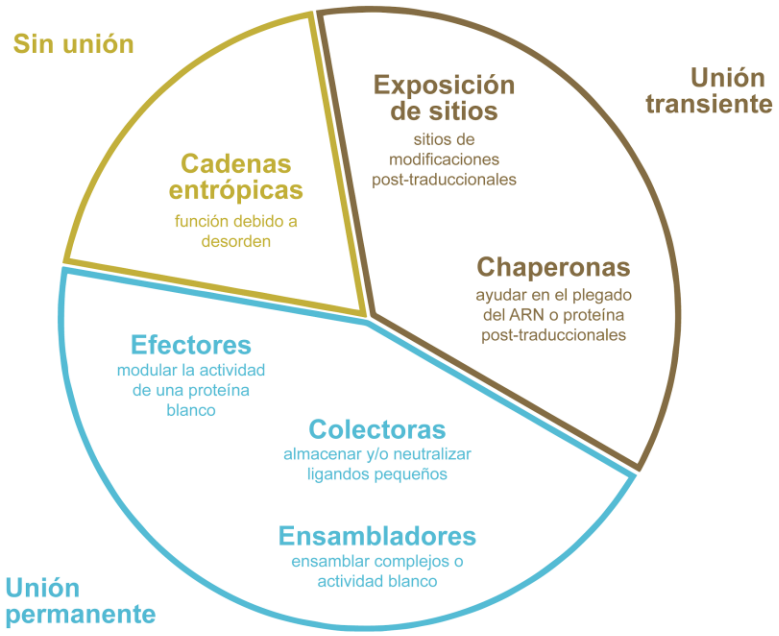


Figura 1.22 Funcionalidad de las IDPs.

Las funcionalidades de las IDPs pueden entenderse en 3 grandes grupos. Pueden no unirse a ligando y llevar a cabo funciones de cadena entrópicas (mostradas en lila) tales como linkers o spacers, o pueden unirse de manera permanente a efectores (mostradas en azul) tales como proteínas, otras macromoléculas o moléculas pequeñas. También pueden unirse de manera transitoria a un blanco (mostradas en verde) por ejemplo en el display de sitios de modificación post traduccional (imagen tomada y adaptada de (van der Lee et al., 2014).

Desde el momento del reconocimiento del desorden estructural hace más de 20 años, además de su caracterización experimental también se ha avanzado en el desarrollo de estrategias teórico-computacionales (Liu et al., 2019), (Atkins et al., 2015) (He et al., 2009). En la predicción del desorden existen métodos basados en distintos principios y técnicas computacionales y por el momento no existe un estándar completamente definido para la asignación de desorden. En líneas generales los algoritmos de predicción de desorden pueden clasificarse en cuatro grandes grupos:

(1) Predicción de desorden a partir de la secuencia y la propensión de determinados aminoácidos vinculados al desorden. Aplican estadística básica sobre las características fisicoquímicas de los aminoácidos FoldIndex (Prilusky et al., 2005),

sobre la tendencia de adoptar cierta estructura secundaria como GlobPlot (Linding et al., 2003b) o en regiones con factores de alta temperatura estructural DisEMBL (Linding et al., 2003a)

(2) Predictores basados en funciones de pseudo-energía (por ejemplo a través de potenciales de contacto) que se basan en estadísticas derivadas de contactos entre residuos en proteínas plegadas para poder determinar residuos con alta energía (y por lo tanto desordenados) IUPred (Mészáros et al., 2018) (Dosztányi et al., 2005).

(3) Predicción de desorden por métodos de aprendizaje automatizado (*machine learning* en inglés) sobre estructuras determinadas por cristalografía de rayos X: PrDOS (Ishida and Kinoshita, 2007), DISORPRED3 (Jones and Cozzetto, 2015), Spritz (Vullo et al., 2006), AUCpreD (Wang et al., 2016), ESpritz (Walsh et al., 2012), SPINE-D (Zhang et al., 2017).

(4) Meta predictores que integran la información de predicción de distintos softwares citas MFDp (Mizianty et al., 2010), MobiDBLite (Necci et al., 2017), CSpritz (Walsh et al., 2011), PONDR-FIT (Xue et al., 2010)

En esta tesis en particular se utilizó el predictor ESpritz (Walsh et al., 2012) que se trata de un predictor que utiliza redes neuronales para predecir distintos tipos de desorden utilizando información secuencial.

Tal fue el crecimiento de la disponibilidad de información sobre este tipo de proteínas que fue necesario desarrollar recursos informáticos que centralizaran y organizaran dicha información en forma de bases de datos MobiDB (Piovesan et al., 2018), IDEAL (Ota et al., 2013), DIBS (Schad et al., 2018), MFIB (Fichó et al., 2017) y las recientemente actualizadas DisProt (Hatos et al., 2020) y PED (Lazar et al., 2021). Más aún, las IDPs fueron incorporadas en la edición del año 2002 de las CASP (*Critical Assessment of Protein Structure Prediction*, por sus siglas en inglés, http://predictioncenter.org),

competencias bianuales en las que distintos grupos de investigación y desarrollo buscan predecir la estructura de una proteína a partir de su secuencia. Sin embargo, a los pocos años, las predicciones de desorden fueron retiradas del CASP fundamentalmente por falta de estandarización. Recientemente, en 2020 se reanudó el denominado *Critical Assessment of Protein Intrinsic Disorder Prediction* (CAID) cuyos primeros resultados fueron recientemente aceptados en la revista *Nature Methods* (Necci et al., 2020).

Sin lugar a dudas la existencia de las IDPs refleja que el término "proteínas" dista mucho de haber alcanzado un concepto definitivo de estas macromoléculas y la importancia de la consideración del desorden no es menor para mejorar la comprensión de la relación estructura-función. En este sentido y retomando el trabajo de Burra (Burra et al., 2009), previamente comentado, se rescataba que la mayoría de las proteínas no cambian de conformación para poder llevar a cabo correctamente su función biológica. Sin embargo, cuando se analizó siguiendo el protocolo de este mismo trabajo sobre un conjunto de datos curado de confórmeros extraídos de la base de datos de diversidad conformacional, CoDNaS (***C****onformational* ***D****iversity of* ***N****ative* ***S****tate*) (Monzon et al., 2016), se reconocieron 3 grandes grupos de mecanismos conformacionales de relación estructura-función, establecidos características estructurales y dinámicas del ensemble conformacional (Monzon et al., 2017a) :

- Proteínas rígidas: todos sus confórmeros son ordenados y tienen un RMSD máximo promedio entre confórmeros de 0.83 Å. Estas proteínas no requieren de grandes movimientos para cumplir con su función biológica. La manera propuesta de funcionamiento de este tipo de proteínas es la apertura y cierre de túneles.

- Proteínas parcialmente desordenadas: que tienen en promedio el 67% de sus confórmeros desordenados y presentan RMSD máximo promedio entre confórmeros de 1.11 Å.

- Proteínas maleables: solamente el 25% de sus confórmeros son desordenados y presentan RMSD máximo promedio entre confórmeros de 1. Å.

Si bien estos resultados son novedosos y nos dan valiosa información biológica sobre las proteínas, se debe tener en cuenta que se basan en el estudio de proteínas globulares, mayormente ordenadas (estructuradas) cuyas estructuras fueron obtenidas por cristalografía de rayos X. La inclusión de resultados obtenidos por otras técnicas ha permitido ampliar estas categorías y así entender mejor el rol del desorden y las restricciones evolutivas que soportan este tipo de proteínas.

1.2 Evolución de Proteínas

Hasta este momento de la tesis se habló sobre el estado nativo de las proteínas y la relevancia de los distintos modelos para explicar la función biológica de una proteína. El enfoque primario para comprender la función biológica implica estudiar cómo las proteínas interactúan con sus ligandos, efectores, moduladores y otras macromoléculas. Sin embargo, un estudio más integral tendría que incorporar la variación[3] y modulación de dichas funciones desde su origen, esto es contemplar un enfoque evolutivo. Quedarnos únicamente con el aspecto estructural y fenomenológico, de las proteínas sería negar la enorme importancia que tienen otros factores sobre la función biológica de las mismas como lo son los procesos evolutivos. En esta sección se tratarán generalidades sobre evolución molecular, modelos de evolución molecular y su evaluación.

1.2.1 Evolución molecular

Las proteínas son entidades dinámicas que cambian su secuencia a lo largo del tiempo por procesos aleatorios de base molecular como mutaciones o recombinaciones ocurridas a nivel de ADN (o ARN). Estas modificaciones están sujetas a procesos de

[3] En buena parte de la bibliografía de Evolución Molecular, los términos variación y mutación suelen usarse como sinónimos

selección natural y a la deriva génica y de acuerdo al impacto que generen en el organismo y pueden o no ser mantenidas a lo largo del tiempo en el acervo genético de una población. La vida se originó hace más de 4000 millones de años y toda la historia de variación en las secuencias viables para los organismos y que se fijaron en la población quedó registrada en los genomas de los seres vivos que vivieron y viven actualmente en el planeta Tierra. Estudiar y comprender cuáles fueron los eventos moleculares que tuvieron lugar en los genomas de los seres vivos, es otra forma de estudiar la historia de la vida en la Tierra (Zuckerkandl and Pauling, 1965).

La homología es la propiedad que indica la presencia de similitud debido a un origen común. De aquí que, los genomas, los genes, las regiones no codificantes o las proteínas homólogos son aquellos que comparten un origen común, o dicho de otro modo, que comparten un ancestro común. Las proteínas denominadas ortólogas, son proteínas que tienen un ancestro común y cuya divergencia se dio por un proceso de especiación. En cambio, las proteínas parálogas se originan por procesos de duplicación génica (Figura 1.23).

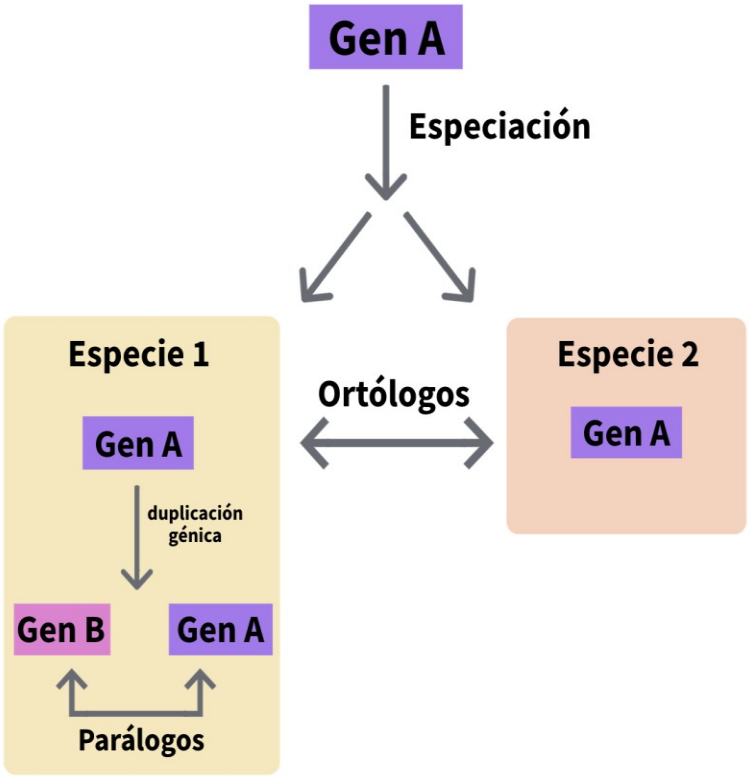


Figura 1.23 Esquema de relación entre proteínas ortólogas y parálogas.

Las proteínas ortólogas son producto de un proceso de especiación y las parálogas son producto de procesos de duplicación génica dentro de una misma especie. En la figura se esquematiza con genes porque son estos los que llevan la información codificante de las proteínas.

Las proteínas ortólogas tienden a conservar una función similar en los distintos organismos. Esta función, según el paradigma clásico, debe estar sostenida por una estructura tridimensional determinada. Por lo tanto, las proteínas homólogas presentan variantes secuenciales que preservan la estructura. Sin embargo, y de acuerdo a lo mencionado previamente, las proteínas desordenadas también sufren cambios en sus secuencias y cumplen funciones, por lo que la preservación de la función también está presente sin que implique la restricción impuesta por la estructura. Más adelante y dentro del trabajo realizado durante esta tesis, profundizamos sobre este tema.

Independientemente del tipo de proteína, si se reclutan secuencias de proteínas homólogas y se realiza un alineamiento múltiple de secuencias (***M**ultiple **S**equence **A**lignment* MSA por sus siglas en inglés, acrónimo que se utilizará de aquí en adelante), es posible observar que mientras algunas posiciones se conservan, otras varían casi al azar. Independientemente del tiempo de divergencia entre las secuencias de proteínas utilizadas para la construcción del MSA, una posición se considera conservada cuando para ese sitio, todas o una gran parte de las secuencias tienen el mismo aminoácido o bien aminoácidos que presentan características fisicoquímicas similares (Figura 1.24). Si bien esta tesis se ha centrado en la descripción de un alineamiento múltiple de secuencias de aminoácidos, también podríamos extender a alineamientos de secuencias de nucleótidos o ribonucleótidos, independientemente que sean regiones codificantes o no. La conservación diferencial en cualquiera de ellos es informativa.

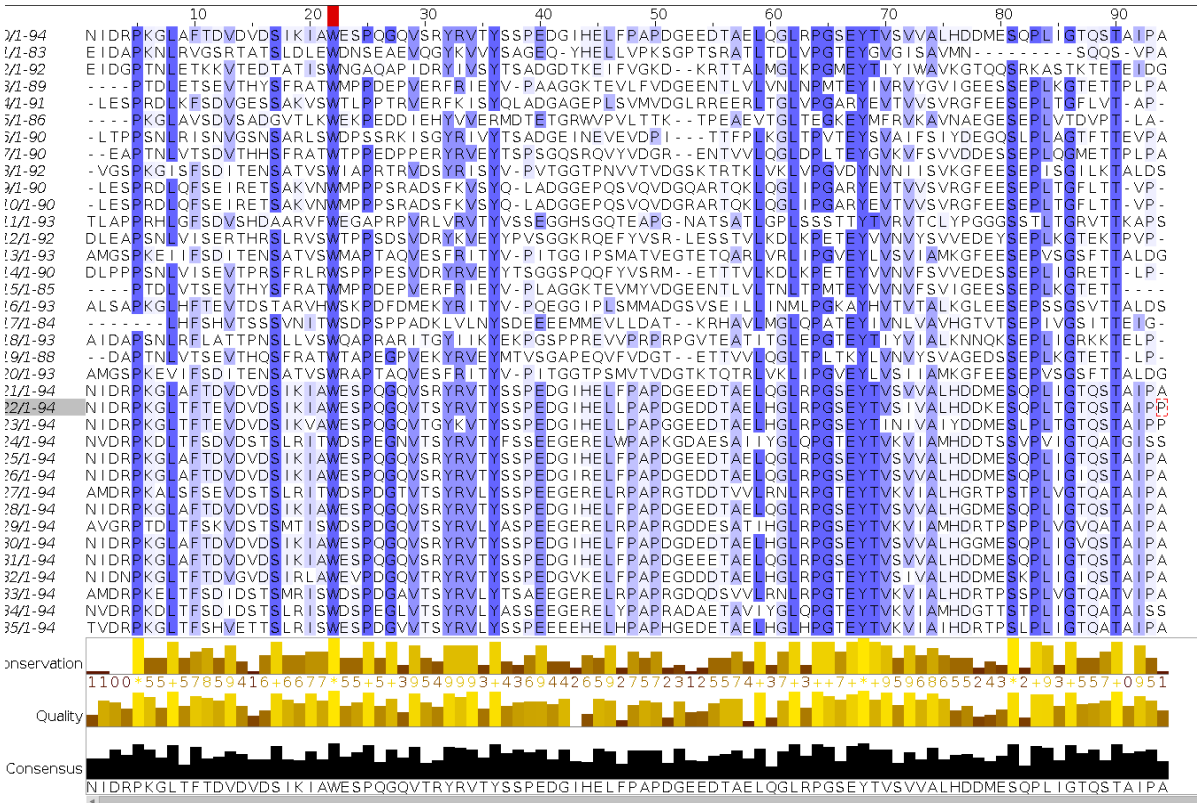


Figura 1.24 Ejemplo de un alineamiento múltiple de secuencias de aminoácidos (MSA).

A partir de un MSA es posible observar que existen posiciones más conservadas (columnas en azul) que otras.

Los mecanismos por los que una secuencia de ácidos nucleicos puede cambiar en el tiempo son 3: sustituciones, deleciones e inserciones. Las sustituciones son el tipo de cambio más frecuente, siendo a su vez las transiciones más frecuentes que las transversiones. Estos cambios, si ocurren en regiones codificantes del genoma pueden dar origen a cambios en las secuencias de las proteínas correspondientes. Aún, cuando no cambien la cadena polipeptídica final pueden causar alteraciones en la síntesis de la proteína correspondiente por falta de la maquinaria necesaria para procesar el cambio. Asimismo la velocidad con que una secuencia varía en el tiempo no es uniforme y se han planteado cuatro puntos de control a través de los cuales la selección natural puede modular lo que en el campo se llama velocidad de mutación: fidelidad de la replicación, exposición a mutágenos y mecanismos de reparación y buffering (Lynch et al., 2016) (Baer et al., 2007).

Desde la perspectiva de la evolución molecular, el origen de variaciones secuenciales en la población y su posible conservación puede ser explicada según cuál fue el proceso que operó mayormente en el curso de la evolución. Muy sucintamente, estos procesos se pueden explicar por las teorías denominadas seleccionista o neutralista (respectivamente moduladas por la selección natural o por la deriva génica (Futuyma, 2009)).

- Para la rama *seleccionista*, aquellas variantes que producen una tasa diferencial de reproducción a los organismos que las portan, es decir que otorgan una ventaja adaptativa frente a otras variables, serán seleccionadas positivamente y serán *fijadas* en el acervo génico de la población.

- Para la rama de evolución *neutralista (Kimura, 1968)* (King and Jukes, 1969), la mayor parte de las variaciones en secuencia que fueron fijadas en la población no otorgaron necesariamente una ventaja adaptativa y son en gran medida neutras y por lo tanto el curso de la evolución de secuencias va a estar determinado primariamente por procesos de deriva génica y no por un proceso de selección. Esta rama sostiene que la mayor parte de las sustituciones observadas en secuencias de proteínas homólogas se consideran neutrales en el sentido que no afectan de manera notable a la función biológica de la proteína en cuestión ni su síntesis y que por lo tanto no tienen mayor impacto en la funcionalidad de la célula o del organismo. Sin embargo, es posible observar que existen variaciones en las velocidades de sustitución entre distintas clases de proteínas y Kimura (Kimura, 1968) explica que la razón de dicha variación se debe al efecto que ejerce la selección natural sobre la conservación de la estructura. Aquellos residuos que comprometen la función biológica de la proteína, debido a que están estrechamente relacionados a la misma (una relación sencilla es pensar en aminoácidos que forman parte del sitio activo) tienden a variar menos durante el curso de la evolución (Kimura, 1968) y que por lo tanto van a estar más conservados que (Conrad, 1977) (Kimura and Ohta, 1974). Podríamos limitar

estas afirmaciones a la preservación de la función, dejando de lado la estructura en vista de los diferentes tipos de proteínas.

1.2.2 El rol de la estructura proteica durante el curso de la evolución

En este punto es evidente que existe una estrecha relación entre la secuencia, la estructura y la función biológica de una proteína. Perutz en el año 1965 (Perutz et al., 1965), tomó la información estructural disponible de proteínas pertenecientes a la familia de las globinas y recolectó 18 secuencias homólogas. A pesar de la escasa información disponible en aquella época, Perutz realizó una observación que daría el puntapié inicial al estudio del rol de la estructura proteica en el curso de la evolución. En este trabajo pudo determinar que, pese a la divergencia secuencial observada, las proteínas presentaban un plegamiento muy parecido y además observó que aquellos residuos que se encontraban formando parte del núcleo estructural o *'core'* eran principalmente no polares y se estaban altamente conservados en las secuencias que habían sido utilizadas para el estudio. De este modo los residuos formadores del '*core*' presentaban un patrón de conservación diferente al resto de los residuos. Las proteínas de la familia de las globinas fueron utilizadas en numerosos trabajos posteriores al de Perutz y se intentó explicar la conservación diferencial de los residuos que formaban parte del *'core'* proteico. En estos trabajos se determinó que era esencial el mantenimiento de un núcleo hidrofóbico (Ptitsyn, 1974) y el volumen de permanencia del core (Lesk and Chothia, 1980).

Lesk y Chothia extienden el estudio de la adaptación de la estructura a las distintas sustituciones descrito inicialmente para las globinas, a distintas familias estructurales. En este trabajo estudiaron la relación entre la divergencia secuencial (evaluada como el porcentaje de identidad entre las secuencias utilizadas en el MSA) y la divergencia estructural (medida con el RMSD) y encontraron que dicha relación respondía a un comportamiento exponencial decreciente (Chothia and Lesk, 1986) (Chothia and Lesk, 1987), Figura 1.25.

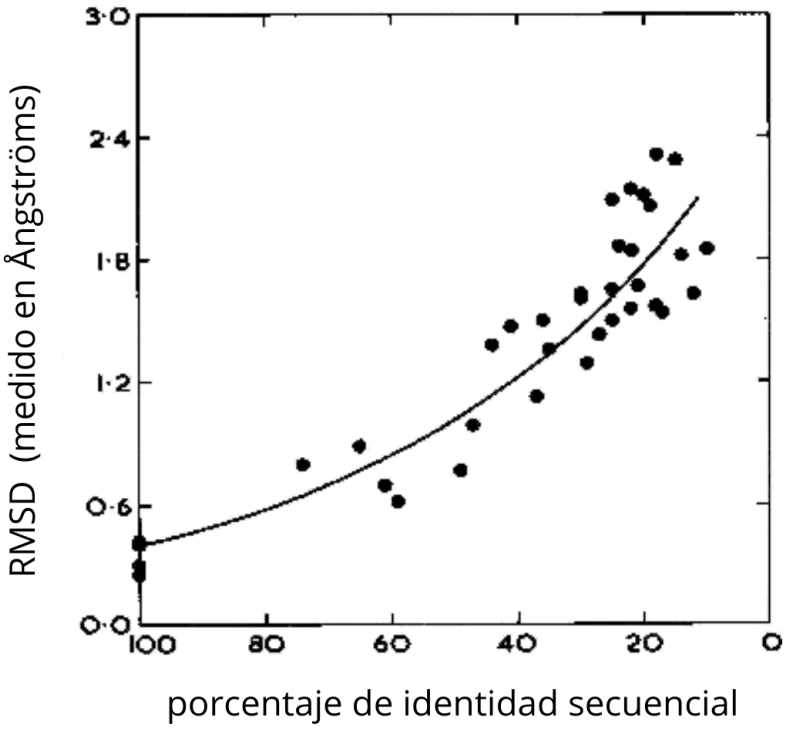


Figura 1.25 Relación entre la divergencia estructural y la divergencia secuencial.

Relación entre el grado de divergencia secuencial entre dos proteínas (medido con base en el porcentaje de identidad secuencial) y el grado de divergencia estructural (medido con base en RMSD) (Chothia and Lesk, 1986).

En trabajos posteriores, se estudió la relación entre la divergencia secuencial y otros parámetros de divergencia estructural (conservación de accesibilidad al solvente, estructura secundaria, número de inserciones y deleciones, número de contactos establecidos entre residuos) y en todos se encontró que existe una relación decreciente (no estrictamente exponencial) entre el grado de divergencia secuencial y el grado de divergencia estructural entre dos proteínas homólogas (Wood and Pearson, 1999),(Flores et al., 1993; Russell and Barton, 1992). Las proteínas que tenían un grado de divergencia no mayor al 25% de identidad secuencial, tenían valores bajos de RMSD lo que significa que secuencias muy diferentes pueden adoptar el mismo plegamiento. El corolario clave de todos estos resultados es que la estructura diverge menos que la

secuencia (Illergård et al., 2009) ya que la conservación de la estructura asegura la conservación de la función biológica. De esta forma, necesariamente, la conservación de la estructura de una proteína debe imponer restricciones sobre la divergencia de las secuencias, esto es sobre el patrón de sustitución de los aminoácidos. Nuevamente, se puede agregar que se podrá pensar en las restricciones que operan sobre las proteínas desordenadas.

1.2.3 El Patrón de sustitución y los modelos de evolución

Como se concluye de la sección anterior, la conservación del plegamiento de una proteína puede imponer una fuerte presión de selección sobre la divergencia de secuencia de la proteína. Como consecuencia algunas posiciones tenderán a estar más conservadas que otras (Figura 1.24), dando lugar a patrones de variabilidad característicos denominados patrones de sustitución y que son insumo de información clave para el estudio de la evolución molecular. Una forma de estudiar y caracterizar los eventos moleculares que tuvieron lugar para dar origen a la variabilidad observada en las secuencias partiendo de un MSA es a través de los modelos de evolución molecular. En líneas generales muchos de estos modelos son modelos matemáticos probabilísticos que se encuentran enmarcados dentro del formalismo de los procesos de Markov[4] (de tiempo homogéneo) cuya característica principal es que la probabilidad de que un evento tenga lugar solamente depende del estado actual y no de la historia de dicho estado. A continuación se dan las generalidades matemáticas en relación a los modelos de evolución, las ecuaciones que se muestran y su desarrollo pueden encontrarse en (Kosiol and Goldman, 2011). En el marco de la evolución molecular de proteínas los estados son los aminoácidos y el evento a modelar es la variación en dicho sitio. Se puede pensar que en una secuencia aminoacídica, donde cada sitio o posición evoluciona independientemente uno de otro, la sustitución del aminoácido ***i*** por el ***j*** durante un

[4] Los procesos de Markov son ampliamente utilizados en otras áreas disciplinares como la meteorología y economía entre otras.

determinado tiempo ***t*** tiene una probabilidad $\mathbf{P_{ij}(t)}$ de ocurrir. Dicha probabilidad solamente depende de su estado actual y puede ser expresada como:

Ecuación 3: Expresión matemática de la matriz de intercambiabilidad P(t).

$$P(t) = e^{tQ}$$

Donde la función exponencial puede ser escrita como:

Ecuación 4: Expresión matemática de la matriz de intercambiabilidad P(t) mediante expansión del exponencial como serie.

$$e^{tQ} = I + tQ + \frac{(tQ)^2}{2!} + \frac{(tQ)^3}{3!} + \cdots$$

Donde **I** es la matriz de identidad y **Q** es la matriz de velocidad que contiene las velocidades relativas de sustitución de ***i*** → ***j***. Bajo la asunción de un modelo de Markov homogéneo, la matriz **Q** es independiente del tiempo, lo que implica que el proceso de evolución es homogéneo (es decir que los patrones de sustitución observados son los mismos ya sea en alta divergencia o baja divergencia secuencial), lo que le confiere a **Q** la propiedad de reversibilidad y por ende la velocidad de sustitución del aminoácido i por el j es igual a la velocidad de sustitución del aminoácido j por el i, entonces:

Ecuación 5: Reversibilidad de la matriz Q.

$$Q_{ij} = Q_{ji}$$

Esta matriz es simétrica y tiene una dimensión de N x N donde N es el número de posibles estados: los 4 nucleótidos para secuencias de ADN para modelos de ácidos nucleicos,

los 20 aminoácidos para modelos basados en proteínas y los 64 codones para modelos evolutivos basados en codones. El elemento $\mathbf{Q_{ii}}$ de la matriz es **1** menos la suma de todos los elementos de la fila distintos a $\mathbf{Q_{ii}}$:

Ecuación 6: Expresión para la matriz de velocidad Q_{ii}.

$$Q_{ii} = -\sum_{j \neq i}^{N} Q_{ij}$$

Si el proceso de Markov se deja 'evolucionar' por un tiempo largo, la probabilidad del estado converge a un valor que es independiente del estado inicial y estará determinada por la frecuencia de equilibrio que puede ser determinada resolviendo

Ecuación 7: Expresión matemática de la matriz (Q).

$$\pi Q = 0$$

Como el tiempo **t** y **Q** no pueden ser determinadas en forma independiente por lo que sí se puede estimar es su producto **tQ** y de esa manera **Q** puede ser normalizada por cualquier factor. En evolución molecular **Q** es expresada como el promedio de las velocidades de sustitución y de esa manera la distancia evolutiva, es decir el tiempo, puede ser medido en unidades de sustitución por sitio.

Recapitulando, los modelos de evolución deben contener:

- Una matriz de velocidad de sustitución **(Q)**
- Un vector con las frecuencias de equilibrio.

Como consecuencia, una vez obtenida **Q** (la matriz de velocidad de intercambio) es posible derivar una matriz de intercambiabilidad **P** (matriz de intercambiabilidad) con la Ecuación 3, que da cuenta de la probabilidad de intercambio de un nucleótido por otro,

o de un codón por otro, o de un aminoácido por otro según el modelo esté basado en ácidos nucleicos, codones o proteínas, respectivamente. Con el crecimiento de información secuencial y la disponibilidad de mejores herramientas computacionales, los modelos de evolución son más que numerosos donde cada uno de ellos toma diferentes aproximaciones para estimar la matriz **Q**.

Una forma posible de clasificar a los modelos es en:

- *Paramétricos (o mecanicistas)*: estos modelos tienen la particularidad de tener en cuenta para la estimación de la matriz **Q**, parámetros relacionados con las propiedades fisicoquímicas o mecanismos biológicos (ej. Sesgos de sustituciones en el caso de modelos de ADN, degeneración del código genético, etc.) para estimar las tendencias relativas de sustitución de nucleótidos, codones o aminoácidos. A modo de ejemplo una observación frecuente en este tipo de modelos es que establecen que las transiciones (cambio de una purina por una purina o de una pirimidina por otra pirimidina) son más frecuentes que las transversiones (cambio de una purina por una pirimidina o viceversa) o bien que los intercambios entre aminoácidos que son similares fisicoquímicamente son más frecuentes que cambios entre aminoácidos disímiles. La ventaja de este tipo de modelos es que la estimación **Q** es específica para cada conjunto de datos utilizados para su construcción.

- *Empíricos*: estos modelos estiman la matriz **Q** de manera empírica, utilizando la información contenida en un gran número de secuencias. como por ejemplo contando los intercambios de aminoácidos que tuvieron lugar entre secuencias que estén relacionadas de manera cercana. A diferencia de los modelos paramétricos, la matriz **Q** es fija y es potencialmente aplicable a cualquier conjunto de secuencias. Si bien estos modelos son menos interpretativos del proceso de evolución que los modelos mecanicistas, debido a la facilidad de su implementación fueron ampliamente estudiados y utilizados.

Otra forma de clasificar a los modelos de evolución, que no es excluyente a la recientemente mencionada, es en función a si tienen en cuenta (restrictos) o no (irrestrictos) a la estructura tridimensional de las proteínas a la hora de establecer las matrices de sustitución aminoacídica. En la próxima sección, se mencionarán generalidades de estos modelos.

1.2.4 Modelos de evolución irrestrictos

Los modelos que se detallan a continuación son modelos irrestrictos, es decir que no tienen en cuenta de manera explícita la restricción impuesta por la estructura sobre la divergencia de las secuencias para la construcción de la matriz **Q**. Margaret Dayhoff y colaboradores, en el año 1978 recopilaron las secuencias disponibles de 78 proteínas distintas, analizaron todas las secuencias disponibles de proteínas homólogas que tuvieran al menos un 85% de identidad secuencial. Luego de alinearlas contaron los cambios observados construyendo una matriz de intercambiabilidad llamada 1 PAM (por *point-accepted mutation*) con la cual se puede construir una matriz de velocidad (**Q**) conocida como matriz Dayhoff (Dayhoff et al., 1978). Jones y colaboradores, mejoraron la estimación de la matrices de sustitución al incorporar una cantidad mucho más grande de secuencias, dando origen a la matriz conocida como JTT (David T. Jones et al., 1992). Tiempo más tarde se propusieron otras matrices como la WAG y (Whelan and Goldman, 2001). En la actualidad se cuenta con varios modelos de evolución parametrizados de distintas formas (para un review ver (Thorne, 2000)).

Tanto la matriz de Dayhoff como la JTT caen en la clasificación de modelos empíricos, y fueron y son aplicadas a cualquier conjunto de secuencias cuya evolución se desee estudiar. Para evitar la excesiva generalidad de dichas matrices otros autores utilizaron grupos específicos de secuencias, por ejemplo conjunto de secuencias de proteínas mitocondriales de vertebrados(mtREV) (Adachi and Hasegawa, 1996) , secuencias de proteínas de mamíferos (MTMAM) (Yang, 1998) o secuencias de proteínas del cloroplasto (cpREV) (Adachi et al., 2000). En 1994, Yang desarrolla un modelo de evolución basado en ácidos nucleicos que además de estimar la matriz **Q** y contener la

frecuencia de equilibrio de los nucleótidos, estima la heterogeneidad de las velocidades de evolución entre las distintas posiciones de una secuencia introduciendo una distribución gamma (Yang, 1994). Esta consideración permitió la mejora sustancial de los modelos de evolución. El desarrollo de modelos continuó incorporando nuevos parámetros y aumentando su complejidad y mejorando su capacidad predictiva, pero quitando interpretabilidad biológica a cada uno de los parámetros derivados del modelo. Es decir, un aumento en el número de parámetros no necesariamente redunda en una mejora del conocimiento del proceso que se quiere modelar, aunque las estimaciones den mejores resultados numéricos. Algunas de las matrices más usadas en este tipo de modelos se muestran en la Figura 1.26.

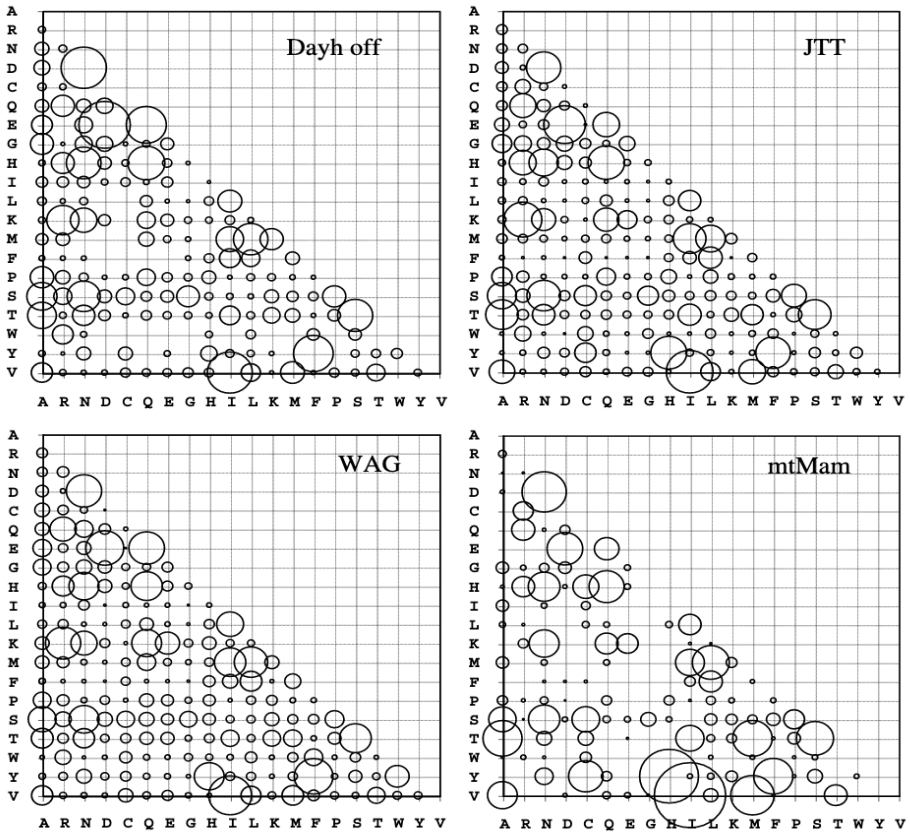


Figura 1.26 Ejemplos de matrices de sustitución basadas en proteínas.

Matrices de sustitución estimadas bajo distintos modelos empíricos: DAYHOFF (Dayhoff et al., 1978), JTT (David T. Jones et al., 1992), WAG, MTMAM (Yang, 1998). El volumen de las burbujas representa el grado de intercambiabilidad entre los aminoácidos, cuanto más grande es la burbuja mayor probabilidad de intercambio entre esos dos aminoácidos. Imagen extraída de (Yang, 2006).

1.2.5 Modelos de evolución restrictos

Como se mencionó en la sección 1.2.2 la estructura proteica diverge menos que la secuencia, por lo tanto, es posible pensar que la evolución de las secuencias debe estar modulada por la estructura. Los modelos de evolución restrictos surgen en respuesta a la necesidad de contar con un modelo de evolución donde el patrón de sustitución secuencial surja como consecuencia directa de la conservación de la estructura proteica. Una primera aproximación para determinar el grado de restricción que impone la estructura podría ser por ejemplo la accesibilidad al solvente de los distintos residuos. Así se ha observado que aquellos residuos que se encuentran 'enterrados' y que tienen una menor accesibilidad al solvente tienen tasas de evolución más bajas que los que se encuentran expuestos al solvente (Franzosa and Xia, 2009). Otros estudios mostraron que la cantidad de contactos nativos provee una mejor aproximación para las tasas de evolución que la accesibilidad al solvente (Yeh et al., 2014b). Se podría pensar entonces que los contactos nativos de la estructura tridimensional de la proteína imponen restricciones evolutivas considerables sobre la divergencia secuencial. Lo clave a entender de estos trabajos es que según el rol que los residuos tengan en el correcto plegamiento y en la estabilidad de la estructura terciaria, van a tener tasas de evolución más lentas o más rápidas. De allí la importancia de su efecto sobre las velocidades de sustitución (**Q**) y por ende en las probabilidades de sustitución (**P(t)**). Los modelos de evolución molecular restrictos tienen en cuenta todos estos aspectos a la hora de estimar las probabilidades de sustitución.

Los modelos de restricción de *estabilidad*, predicen el posible efecto en la energía libre del estado nativo de la proteína si tiene lugar una mutación y allí derivan la probabilidad de observar una sustitución (Bastolla et al., 2017; Goldstein, 2011; Serohijos and Shakhnovich, 2014). Con este objetivo interpretan a la proteína como un mapa de contactos nativos, simulan una mutación y calculan la variación de la energía libre de la matriz de contacto, asumiendo que la estructura se mantiene a pesar de la ocurrencia de la mutación. A diferencia de los modelos de restricción de *estabilidad*, los modelos de restricción *estructural* estiman el grado de perturbación estructural causada por una

mutación, es decir que asumen que la estructura puede cambiar por efecto de una mutación. Dentro de este grupo se encuentra el SCPE (Parisi and Echave, 2001) (por sus siglas en inglés *Structurally Constrained Protein Evolution*, acrónimo utilizado de aquí en adelante) que debido a su rol central en esta tesis será detallado en la sección siguiente.

1.2.5.1 Modelo SCPE

El SCPE es un modelo de evolución que busca entender cómo el plegamiento proteico impone restricciones estructurales que modulan la variación secuencial durante la evolución molecular. El objetivo de este modelo es simular computacionalmente la evolución de secuencias de proteínas tomando el cumplimiento de restricciones estructurales, dadas por el plegamiento, como único criterio de selección. En este proceso evolutivo simulado, las mutaciones en dicha secuencia tendrán o no lugar en función del grado de perturbación estructural que hayan introducido. Si una mutación desestabiliza completamente la estructura de la proteína, dicha mutación es considerada deletérea y por lo tanto rechazada. De otra manera, es aceptada.

Una posible manera de estimar el grado de perturbación estructural causado por una mutación podría ser a través de la observación de un incremento en la energía libre de la proteína o alguna función derivada de ella. Así, es posible seleccionar secuencias que se adecuen en grado variable a la estructura de la proteína que se está estudiando y por lo tanto la simulación de la evolución consistiría en establecer un valor de corte de adecuación estructural y en función de ello aceptar o rechazar sustituciones. El modelo SCPE simula la divergencia secuencial bajo una determinada presión de selección dado por un valor de corte de adecuación estructural. El SCPE realiza **n** simulaciones independientes para poder alcanzar un grado de significancia estadística adecuado.
En líneas generales el modelo del SCPE consta de la repetición de una serie de pasos representados en la Figura 1.27.

Figura 1.27 Esquema general del algoritmo empleado para el SCPE.

Cada ronda de mutación de cada simulación del SCPE consiste en:

1) Se toma la secuencia de la proteína de interés que además con estructura tridimensional conocida (estructura de referencia). Esta secuencia es la considerada como 'ancestral'. Esta secuencia representa el tiempo t = 0 de la simulación.

2) La secuencia S_0 es mutada a una secuencia S_t (la letra t proviene de *trial*). Cada secuencia mutada son secuencias *trials* o pruebas y fueron referenciadas así dentro del modelo debido a que prueban si dichas secuencias pueden adoptar una determinada estructura.

3) La secuencia proteica S_t es forzada a adoptar la estructura de referencia. Si en cambio la mutación al azar produce la ocurrencia del mismo aminoácido, se vuelve al paso 2 es decir no se hace ninguna estimación energética debido a que $S_t = S_0$.

4) Se calcula un puntaje o *score* de adecuación estructural. Este score está dado por la siguiente ecuación:

Ecuación 8: Expresión matemático del score utilizado por el SCPE.

$$S = \left\{ \sum_{p=1}^{N} \left(E_{mut}\,(p) - E_{ref}\,(p) \right)^2 \right\}^{\frac{1}{2}}$$

Donde N es el número de sitios de la secuencia proteica, E_{mut} y E_{ref}, son las energías de la posición p en la secuencia mutada y en la secuencia de referencia respectivamente. Este *score* puede ser global si se usa como energía de referencia la energía de la secuencia de referencia, o puede ser local si la energía de referencia está dada por la secuencia aceptada en el paso inmediatamente anterior de mutación. Sin entrar en detalles metodológicos sobre el cálculo de energía, este *score* mide la diferencia entre la energía del estado plegado de la secuencia ancestral (S_0, si es global, $S_{(t-1)}$ si es local) y la energía del estado plegado de la secuencia *trial* S_t. Tanto E_{mut} y E_{ref} son obtenidas siguiendo:

Ecuación 9: Ecuación de Boltzman para el cálculo de energía de interacción entre dos residuos.

$$E_{ij} = -kTln(\frac{p_{ij}}{p^*})$$

Donde $\mathbf{p_{ij}}$ es la distribución para el par de residuos ***i j*** calculadas a partir de la estructura proteica y **p*** es un estado de referencia.

Para ilustrar esto último, en la Figura 1.28 se muestra la energía total (utilizando la ecuación-x-tesis (la de la energía) para la familia de proteínas Lpxa y tal como se puede apreciar la mutación Ile38Glu (catalogada como una sustitución no conservativa que cambia un aminoácido alifático por uno cargado a pH=7) produce un aumento en la energía total del sistema para esa posición, mientras que la sustitución Ile38Val (sustitución conservativa) no produce cambios significativos en la energía total del sistema para esa posición.

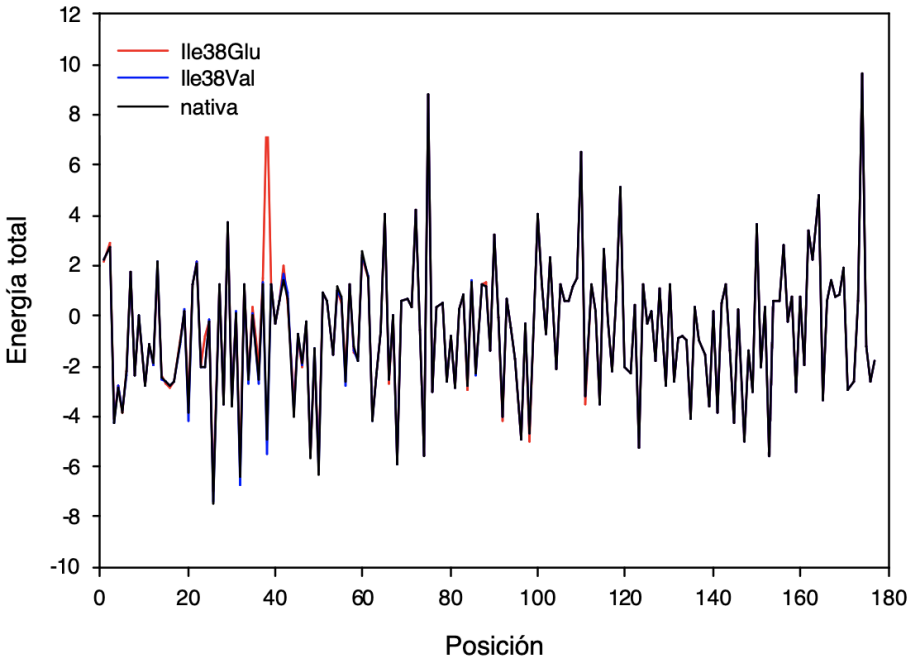


Figura 1.28. Perfil energético de la proteína Lpxa.

Perfil energético de la proteína Lpxa mutada y nativa, demostrando los efectos de sustituciones no conservativas (rojo) y conservativas (azul) en la energía del sistema (expresada de E/kT, donde k es la constante de Boltzman y T la temperatura en grados Kelvin, figura extraída de (Parisi, 2001).

5) Si el *score*
    a) Es mayor al valor de corte establecido, esa mutación es rechazada.
    b) Es menor al valor de corte establecido, esa mutación es aceptada.

Como resultado de las corridas independientes y las rondas de mutación en cada corrida, el SCPE deriva matrices de sustitución que son específicas por sitio, es decir que para cada posición de la proteína fue posible obtener una matriz de sustitución P_{ij} (Fornasari et al., 2002).

En resumen, el SCPE es una herramienta que ha demostrado ser muy útil para estudiar la relación entre la divergencia secuencial y estructural (Fornasari et al., 2007, 2002; Parisi and Echave, 2001), validar modelos estructurales de proteínas con base evolutiva (Palopoli et al., 2013), evaluar el impacto de la diversidad conformacional sobre el patrón de sustitución (Juritz et al., 2013) por nombrar algunos ejemplos. Debe quedar claro que el SCPE permite derivar el patrón de sustitución sitio-específico para una determinada posición de una determinada proteína/estructura. Cada estructura o conformación implica obtener matrices de sustitución específicas para esa estructura permitiendo de esta forma testear hipótesis sobre la conservación de la estructura o sobre una conformación específica durante la evolución.

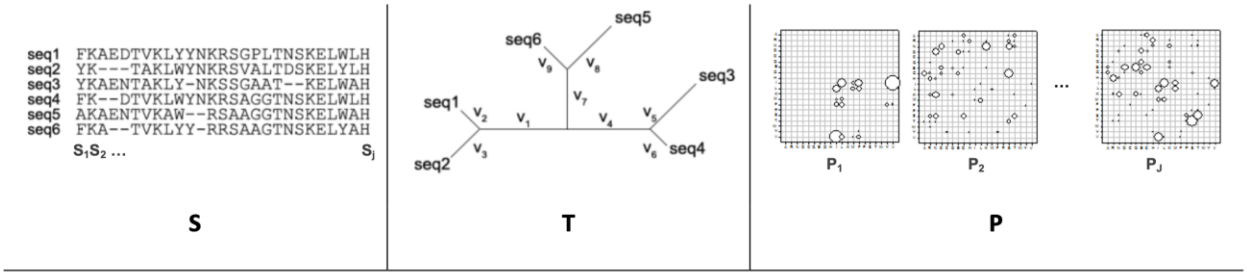
1.2.6 Evaluación de modelos de evolución

En este punto está claro que el universo de modelos de evolución es extenso y se sigue expandiendo. ¿Cómo es posible determinar la eficiencia de un modelo determinado para explicar el patrón de sustitución observado en un MSA? ¿Es posible que haya modelos que son ‘mejores’ o ‘peores’ según el tipo de estudio que se quiera realizar? Está claro que a medida que se desarrollaban modelos de evolución era necesario estimar si estos modelos de evolución mejoraban o empeoraban las estimaciones realizadas por modelos ya existentes.

1.2.6.1 *Maximum Likelihood* (generalidades)

La efectividad de cualquier modelo de evolución en explicar el patrón de sustitución observado de un conjunto de secuencias de un MSA puede evaluarse por medio del

criterio de *Maximum Likelihood*[5] (ML, por sus siglas en inglés) o de Máxima Verosimilitud. Estadísticamente hablando, el criterio de ML busca maximizar el *likelihood* (L) de observar el conjunto de datos obtenido a partir del modelo que se está evaluando. En el cálculo de ML aplicado a evaluación de modelos de evolución molecular, se emplean la topología de un árbol filogenético (**T**), un conjunto de distancias (cantidad de sustituciones no sinónimas) derivadas de secuencias presentes en un MSA (**S**, en el esquema) y conectadas entre sí según la topología del árbol **T**, y las probabilidades de intercambio entre aminoácidos representadas por una matriz de sustitución (modelo a evaluar). Para cada posición **p** en el MSA, es posible estimar el *likelihood* $\mathbf{L_p}$ de que dicha posición haya evolucionado de acuerdo a los patrones de sustitución que son inferidos según el modelo de sustitución representado por la topología del árbol y sus distancias. En esta tesis se utilizó el programa HyPhy (Pond et al., 2005) ya que es capaz de utilizar matrices de sustitución sitio-específicas como las derivadas en las simulaciones del SCPE. El HyPhy optimiza las distancias v1, v2, ..., vn, para que, dados los parámetros del modelo, el *likelihood* para dicha posición **Lp** sea máximo. En la Figura 1.29 se muestra un esquema del cálculo de *likelihood*.



$$L_p = (S_p \mid T; v_1, v_2 ... v_n; P_p)$$

Figura 1.29 Representación del significado del likelihood y la manera de computarlo.

S_p representa a una columna (en este caso la columna p) de un MSA (en este esquema compuesto por 6 secuencias). T es el árbol inferido para dicho conjunto de secuencias y v_1 a v_n indican la longitud de cada rama del árbol. L_p es el likelihood para la posición p. El maximun likelihood (o máxima verosimilitud) para toda la proteína se obtiene como la suma de los valores L_p a lo largo de toda la proteína, como $L = \prod_j^n L_j$ o de manera equivalente $lnL = \sum_{j=1}^{n} lnL_j$. Figura extraída de (Palopoli, 2011).

[5] Si bien existe la expresión en español que es máxima verosimilitud, en este trabajo se utilizará el término en inglés *Maximum Likelihood*.

Objetivos

Objetivos generales

El tema central del presente trabajo de tesis es la caracterización del estado nativo de las proteínas desordenadas (IDPs), utilizando métodos computacionales. Estas proteínas se caracterizan por poseer una alta flexibilidad y por ende carecen de una estructura tridimensional estable en condiciones nativas. Si bien se las conoce como "desordenadas" quizás el término es abusivo e induce a confusión. Las IDPs no son polímeros con estructuras azarosas, sino que dentro de su enorme flexibilidad y diversidad conformacional poseen un cierto "orden" que está íntimamente relacionado con la función biológica. En Biología Estructural y Evolución Molecular existe una vieja tradición para estudiar la relación estructura-función en proteínas, que históricamente, han tomado como modelos de estudio casi exclusivamente a las proteínas globulares, ordenadas y paradigmáticas de lo que "representa" una proteína. La hipótesis central del presente trabajo es que dichos métodos también pueden ser utilizados en las IDPs, para caracterizar su, aparentemente contradictoria, relación "estructura-función".

El objetivo general de este trabajo es el estudio de proteínas mediante herramientas bioinformáticas basadas en métodos evolutivos. En particular este trabajo se centra en dilucidar la relación entre la ausencia de estructura y el proceso de evolución molecular de las IDPs. El estudio de los ensembles conformacionales de IDPs desde un enfoque evolutivo no solo permitirá la caracterización de su estado nativo y su íntima relación con la función biológica, sino que además promoverá el desarrollo de nuevos métodos y herramientas bioinformáticas.

Objetivos específicos

- Estudiar la presencia de condicionamientos estructurales en la divergencia secuencial en IDPs utilizando modelos evolutivos con base estructural.

- Estudiar el comportamiento diferencial de velocidades de evolución en regiones ordenadas y desordenadas de las IDPs y su relación con las características estructurales de las mismas.

- Analizar la información evolutiva presente en todo el ensemble para poder detectar conformaciones biológicamente relevantes en ensembles conformacionales.

2 Estudio de las restricciones estructurales en proteínas intrínsecamente desordenadas

2.1 Resumen

Las conformaciones a las que puede acceder una proteína están estabilizadas en gran medida por el establecimiento de contactos entre residuos[6]. Estos contactos actúan como restricciones estructurales que modulan la divergencia de las secuencias, probablemente porque las mismas participan en la estabilización de conformaciones, estructuras o complejos que son de vital importancia para el mantenimiento de la función biológica. En esta parte del trabajo se estudió la proporción de sitios que poseen restricciones estructurales presentes en dos grandes tipos de proteínas: las extensamente conocidas proteínas globulares (mayormente ordenadas y poco flexibles) y las 'novedosas' proteínas intrínsecamente desordenadas IDPs, (mayormente sin estructura terciaria y altamente flexibles). Estas restricciones estructurales que se evidencian en el proceso evolutivo fueron determinadas utilizando un modelo de evolución molecular con base estructural llamado SCPE (descrito en la 1.2.5.1).

Se encontró que en los ensembles conformacionales de ambos tipos de proteínas, al menos un 40% de sus sitios, evolucionan bajo restricciones estructurales y en promedio el 27% de las posiciones desordenadas de las IDPs son sitios restringidos evolutivamente. A pesar de las diferencias en la complejidad y en la estructura entre proteínas globulares e IDPs, ambos tipos de proteínas presentan restricciones estructurales similares por lo que se puede deducir que la presión de selección impuesta

[6] En esta tesis los términos *contactos entre residuos* y *contactos terciarios* hacen referencia al mismo concepto.

sobre la estructura para el mantenimiento de la función biológica es similar en ambos tipos de proteínas.

Con estos resultados se hace evidente la importancia de la información evolutiva a la hora de obtener resultados que permitan caracterizar biológicamente a los ensembles conformacionales complejos como los de las IDPs.

2.2 Introducción

Como ya se mencionó anteriormente, el estado nativo de una proteína puede ser descrito por el conjunto de distintas conformaciones a las que la secuencia de dicha proteína puede acceder. Este estado nativo, se denomina ensemble conformacional y es esencial para entender la biología las proteínas (Tsai et al., 1999; Wei et al., 2016). Los ensembles de IDPs pueden estar formados por numerosas conformaciones: hay ensembles de 10 confórmeros, de 20 confórmeros e incluso existen más de 60. En cambio, las proteínas globulares cuentan con una menor cantidad de confórmeros en su estado nativo. Entre estas, se pueden encontrar proteínas con poca flexibilidad en su esqueleto carbonado, llamadas rígidas y que poseen un solo confórmero en su ensemble (Monzon et al., 2017a) como es el caso de la celulasa de *C. cellulolyticum* (Parsiegla et al., 2000). Un ensemble conformacional paradigmático de proteínas globulares con dos confórmeros es el de hemoglobina (Perutz et al., 1964) Figura 1.14, mientras que la proteína activadora por catabolitos (Popovych et al., 2006) y la glucoquinasa humana son ejemplos de ensembles proteicos con tres conformaciones (Kamata et al., 2004). También existen ensembles de proteínas globulares más complejos, en general se tratan de proteínas que están formadas por distintas cadenas, como el caso de la ATP sintasa mitocondrial (Zhou et al., 2015) que puede llegar a tener hasta siete confórmeros. Si bien estos datos provienen de casos particulares de estudio, no existe al momento un estudio a gran escala sobre la distribución del número de confórmeros en el estado nativo de proteínas globulares. A medida que la flexibilidad estructural aumenta, el número de confórmeros que describen al estado nativo, también aumenta, dando lugar a la aparición de ensembles conformacionales muy complejos, como lo son los de las proteínas intrínsecamente desordenadas (IDPs) o regiones desordenadas (IDRs). En las IDPs e IDRs las transiciones orden desorden se observan frecuentemente, fenómeno que puede

estar asociado al proceso de unión a ligando (Zea et al., 2016) o simplemente reflejando la heterogeneidad en la composición de este tipo particular de ensembles proteicos (Monzon et al., 2017a) (DeForte and Uversky, 2016).

En este capítulo de la tesis, se estudió el nivel de restricciones estructurales presentes en los ensembles de IDPs y se los comparó con el nivel de restricciones estructurales observados en ensembles de proteínas globulares. Las restricciones estructurales pueden ser estudiadas a través de métodos directos tal como puede ser la medición de contactos entre residuos en algún confórmero dado o algún otro parámetro derivado de la información estructural como puede ser la densidad de contactos (promedio de contactos entre residuos por residuo) o redes de interacción (Piovesan et al., 2016). Sin embargo, en el caso de ensembles complejos como los de IDPs, es frecuente encontrar contactos que son artefactos o arreglos estructurales que son irrelevantes en relación a la biología, haciendo difícil establecer entonces la importancia biológica de los mismos (Sormanni et al., 2017). Esto se debe principalmente al grado de movilidad y de flexibilidad de los ensembles de IDPs. Para estudiar el impacto de las múltiples conformaciones y multitud de contactos con distinto grado de permanencia (porcentaje de confórmeros que poseen el mismo contacto), en este trabajo se propone un enfoque evolutivo para estudiar las restricciones estructurales en ensembles conformacionales y derivar su importancia biológica.

En la sección 1.2.2 se expuso que la conservación de la estructura proteica restringe la divergencia secuencial, modulando de esta manera el patrón de sustitución aminoacídica para determinadas posiciones que podrían, en principio, ser importantes para mantener la función de la proteína (Overington et al., 1990) (Worth et al., 2009). Las posiciones que están restringidas estructuralmente se ponen en evidencia a partir de alineamientos múltiples de secuencia (MSA), como aquellas posiciones que se encuentran diferencialmente conservadas o que muestran un determinado patrón de sustitución de aminoácidos. Esto puede deberse a que esas posiciones se encuentran sujetas a procesos de co-evolución o porque presentan cierto sesgo fisicoquímico debido a su importancia en el mantenimiento de la estructura, la cinética del plegado, la interacción

con otras proteínas o ligandos y la dinámica de la proteína. Un ejemplo de este último fenómeno puede ser la conservación de determinados residuos debido a que la interacción de los mismos incrementa la estabilidad estructural u otorga movilidad a la proteína. Los patrones de sustitución que poseen conservación estructural, han sido ampliamente explotados para: mejorar los modelos existentes de evolución molecular, (Bastolla et al., 1999) (Kleinman et al., 2010; Parisi and Echave, 2001), explicar la heterogeneidad en la velocidad de evolución (Echave et al., 2016), realizar predicciones funcionales (Simon et al., 2002), comparar los procesos de sustitución aminoacídica tanto en proteínas ordenadas como en proteínas desordenadas (Karpati et al., 1988) y en la deducción (inferencia) de ciertos plegamientos de estructura terciaria (Surkont and Pereira-Leal, 2015), entre otros. La información evolutiva puede ser utilizada también para predecir contactos nativos y establecer modelos estructurales para dominios globulares (Hopf et al., 2014), (Booth and Bradford, 1963), (Ovchinnikov et al., 2014). Recientemente, estos métodos fueron adaptados para predecir de manera satisfactoria estados globulares en IDPs y para mostrar las restricciones evolutivas presentes en interfases proteicas entre proteínas ordenadas y desordenadas, lo cual demuestra nuevamente la importancia de estudiar las restricciones estructurales durante el curso de la evolución (Ovchinnikov et al., 2014) (Pancsa et al., 2018). En los últimos meses el caso paradigmático de la importancia del uso de información evolutiva para la predicción de contactos lo constituye el uso de inteligencia artificial representado por el algoritmo desarrollado por Google Alphafold (Senior et al., 2020) mostrando un éxito arrollador sobre otros métodos para estimar la estructura terciaria de proteínas usando información evolutiva.

En la sección 1.2.1 se mencionó que los modelos de evolución molecular pueden describir de manera satisfactoria los patrones de sustitución observados en MSA (Thorne, 2000). Existen numerosos modelos de evolución, donde cada uno es generado asumiendo distintas hipótesis evolutivas respecto al patrón de sustitución de la proteína en estudio. En este trabajo el foco estuvo puesto en evaluar el desempeño de dos grandes grupos de modelos de evolución proteica a la hora de explicar el patrón de sustitución secuencial observado en un MSA. El primer grupo es el de los modelos

irrestrictos, que son aquellos modelos de base netamente secuencial y que no tienen en cuenta la información estructural. Algunos ejemplos: JTT (David T. Jones et al., 1992), Dayhoff (Dayhoff et al., 1978) y WAG (Whelan and Goldman, 2001). El segundo grupo es el de los modelos restrictos, que a diferencia de los irrestrictos tienen en cuenta de manera explícita el rol de la estructura en la evolución y divergencia de secuencias (Bastolla et al., 2017; Goldstein, 2011; Serohijos and Shakhnovich, 2014). En particular, se evaluó cuál era el desempeño del modelo de evolución SCPE (Parisi and Echave, 2001).

Para poder comparar el desempeño de los distintos modelos en la explicación de los patrones de sustitución observados y decidir cuál es la hipótesis que mejor explica la evolución de la proteínas en este estudio, se utilizaron métodos de *Maximum Likelihood* (ML).A pesar que las proteínas ordenadas y las IDPs difieren enormemente en sus características estructurales (Figura 1.14 y Figura 1.20, respectivamente), se encontró que cuando se consideran todos los confórmeros presentes en ensembles de proteínas globulares e IDPs, las restricciones estructurales en este tipo de proteínas son similares a las restricciones estructurales observadas para ensembles de proteínas globulares.

2.3 Resultados

2.3.1 Descripción del conjunto de datos

Desde el punto de vista del universo conocido de proteínas, sus distintos ensembles contienen conformaciones conectadas entre sí por medio de dinámicas muy diversas. Si bien no está probado fehacientemente, la idea generalizada es que existe un continuo en el espacio dinámico de las proteínas (Motlagh et al., 2014). En uno de los extremos de este continuo se encuentran las proteínas rígidas, cuyos confórmeros presentan escasa diferencia entre sí en su esqueleto carbonado (evidenciado por sus bajos valores de RMSD de los carbonos alfa) y solo exhiben diversidad conformacional a nivel de residuo (Monzon et al., 2017a). La presencia de desorden podría ser evidenciada debido

al aumento en la diversidad conformacional a nivel del esqueleto carbonado, donde el surgimiento de una dinámica de corto alcance permitiría la ampliación del muestreo en el espacio conformacional (Berlow et al., 2015). En la Figura 2.1 se muestran distintos tipos de ensembles conformacionales en función de la diversidad conformacional que presentan.

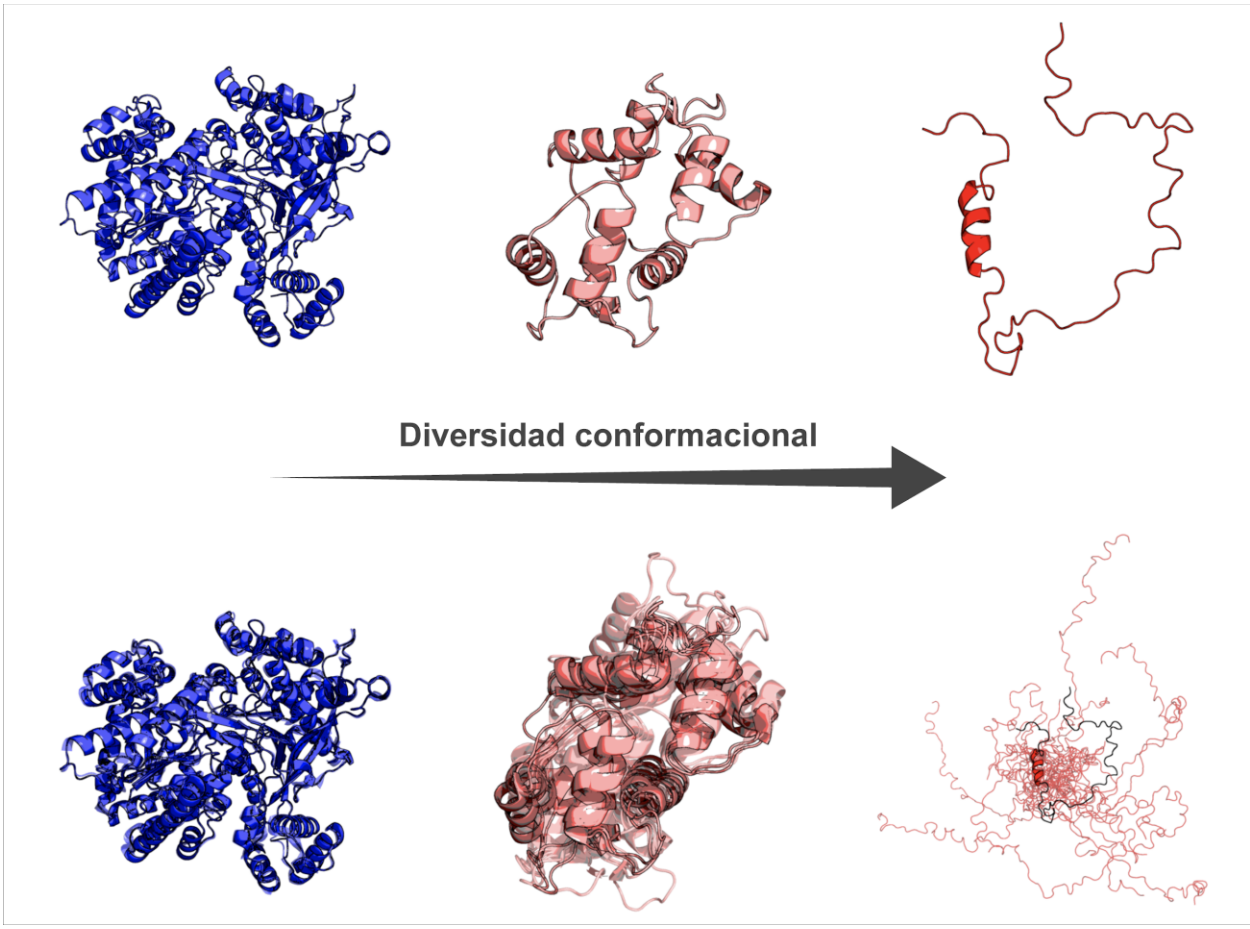


Figura 2.1 Diversidad conformacional en función de la flexibilidad.

Distintos ensembles proteicos en función del incremento en la flexibilidad (dado por la dirección de la flecha). El panel superior de la figura muestra un solo confórmero, mientras que el panel inferior muestra todos los confórmeros disponibles. En la izquierda se muestra a la maltodextrina fosforilasa una proteína rígida (solamente el 6.53 % de sus residuos se encuentra desordenado) tomada como representante de proteínas ordenadas (códigos PDB: 1AHP_A, 1AHP_B, 1L5V_B). En la derecha se muestra a la fosfoproteína soluble del tilacoide como representante de una típica proteína intrínsecamente desordenada con el 100% de sus residuos desordenados (códigos PDB: 2FFT_A). En el centro del panel se muestra a la calmodulina como una estructura intermedia en el continuo estructural con 10.64 % de sus residuos desordenados (códigos PDB: 2FOT_A, 1LIN_A, 1NIW_E, 3G43_A, 2BE6_A, 1CDL_A, 3GP2_A, 4L79_B, 1CCL_A).

En el extremo del continuo estructural se encuentran las proteínas globulares, en la parte izquierda de la Figura 2.1 donde se muestra una proteína globular típica (panel superior) y su ensemble conformacional (panel inferior). Estas proteínas por lo general tienen una gran proporción de estructura secundaria cuyo arreglo espacial define una “única” estructura terciaria y un core o núcleo central hidrofóbico. La gran cantidad de interacciones entre residuos en el núcleo hidrofóbico restringe la velocidad de evolución de estos sitios en comparación con residuos que se encuentran expuestos al solvente. Además el núcleo hidrofóbico contiene suficiente información estructural para definir el arreglo en el espacio de la estructura terciaria (Pereira de Araujo and Onuchic, 2009).
Las proteínas globulares también pueden tener un estado nativo que esté descrito por distintos confórmeros y no una única estructura en pos de poder llevar a cabo su función biológica, Figura 2.1 (panel central), dando lugar a restricciones estructurales adicionales al patrón de sustitución de dicha proteína (Zea et al., 2013). En el panel central de la Figura 2.1 se muestra una proteína globular y su correspondiente ensemble, que presenta regiones ordenadas y también regiones muy flexibles que poseen un comportamiento dinámico muy distinto y que puedan pueden dar lugar al origen de regiones desordenadas de distinta longitud. Se aprecia además el incremento en la diversidad conformacional de la proteína, esto es las diferencias estructurales entre los confórmeros son más marcadas. En el panel inferior derecho de la Figura 2.1 se muestra un típico ensemble conformacional de IDP, se puede observar que el ensemble presenta una gran diversidad de conformaciones distintas determinadas experimentalmente por NMR. Estos ensembles muestran una alta flexibilidad en sus cadenas y pueden eventualmente presentar algunas pequeñas transiciones a segmentos ordenados con estructura secundaria o terciaria (Forman-Kay and Mittag, 2013).

Como se describe en la sección de métodos de este capítulo, se utilizaron dos conjuntos de datos curados a mano: uno de proteínas ordenadas y el otro de IDPs. El conjunto de datos de proteínas ordenadas estaba compuesto por 183 proteínas con estructura conocida a través de métodos cristalográficos y que no tuvieran residuos catalogados como *'missing'* ya que los mismos pueden estar asociados a segmentos altamente

flexibles. El otro conjunto de datos contenía proteínas desordenadas representadas por 93 ensembles derivados de NMR para distintas proteínas. El grado de desorden fue predicho utilizando el predictor ESpritz (información más detallada se puede encontrar en la sección de métodos de este capítulo). Como se muestra en la Figura 2.2, las proteínas ordenadas muestran un bajo nivel de desorden predicho a diferencia de las proteínas del conjunto de datos desordenado. En promedio el 58% de los residuos de las proteínas desordenadas, son desordenados (con un mínimo de 40% y un máximo de 98%). Es de esperar que estas proteínas cuenten con más de la mitad de su estructura en estado desordenado y presenten solamente pequeñas regiones globulares.

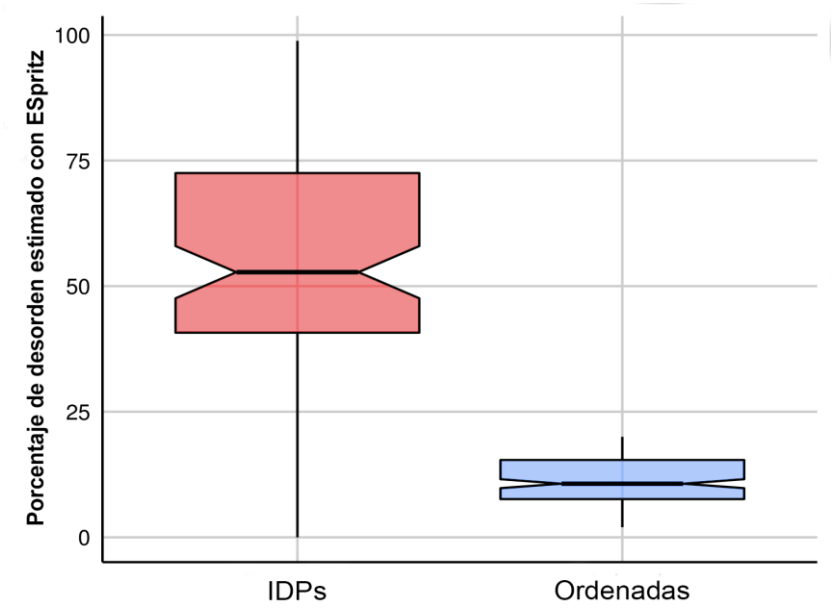


Figura 2.2 Distribución del porcentaje de desorden en IDPs y proteínas globulares.

Estimación del contenido de desorden utilizando NMR-ESpritz (Walsh et al., 2012). y Mobi 2.0 (Piovesan and Tosatto, 2018) para las proteínas desordenadas y ESpritz X-Ray (Walsh et al., 2012). para las proteínas ordenadas. Se puede observar que las proteínas ordenadas presentan una menor proporción de desorden incluso menor que el valor reportado en la estimación del método (Walsh et al., 2012).

Los MSA utilizados para estudiar la proporción de sitios estructuralmente restringidos desde un punto de vista evolutivo para los dos conjuntos de datos fueron extraídos de la base de datos HSSP (ver la sección de métodos de este capítulo). Estos MSA fueron

estudiados y se omitieron aquellas proteínas que tuvieran regiones con alto porcentaje de inserciones y deleciones y se tomaron en cuenta únicamente secuencias que tuvieran al menos un 30% de identidad respecto a la secuencia de la proteína de interés para la cual es necesario que se conociera su estructura tridimensional. Se tomó este porcentaje de 30% de identidad como cota inferior para evitar en la medida de lo posible la incorporación de proteínas con gran divergencia secuencial y posiblemente estructural y/o funcional. Información adicional puede encontrarse en la sección de métodos de este capítulo.

2.3.2 Determinación de restricciones estructurales a partir de los contactos físicos

Una manera de evaluar las restricciones estructurales en ensembles conformacionales es cuantificando la cantidad de contactos entre residuos. En general, los contactos entre átomos se dan por el establecimiento de interacciones débiles como puentes de hidrógeno, van der Waals e interacciones de tipo coulómbicas que en general estabilizan los distintos niveles estructurales de la proteína. Este efecto estabilizador derivado de la formación de interacciones explica la conservación de las propiedades fisicoquímicas de los aminoácidos observado en numerosas posiciones de las proteínas globulares u ordenadas: el proceso de sustitución tiende a conservar el número y tipo de interacciones débiles que establecen los aminoácidos en determinadas posiciones u entornos estructurales (Overington et al., 1990) (Overington et al., 1992). Para cuantificar el número de restricciones estructurales, se decidió estudiar las distribuciones de los contactos para cada confórmero de cada ensemble del conjunto de datos, esto es se calculó para una proteína el porcentaje de posiciones que tenían al menos un contacto (la definición de contactos terciarios puede encontrarse en la sección de métodos de este capítulo). Se encontró que las proteínas ordenadas presentan en promedio un porcentaje de contactos del ~94, mientras que para las IDPs este promedio es del ~74, siendo esta diferencia estadísticamente significativa (Figura 2.3).

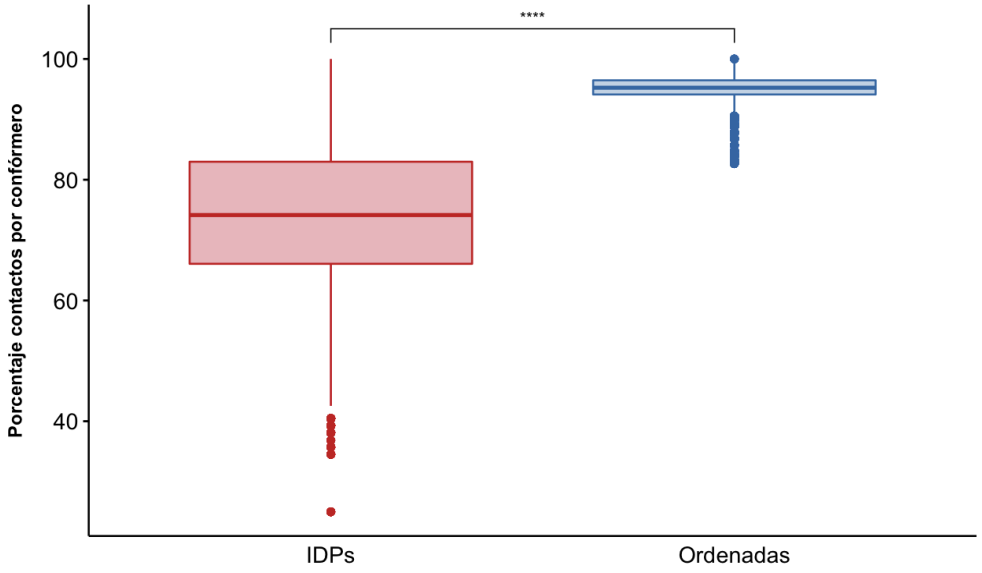


Figura 2.3 Porcentaje de posiciones con contactos entre residuos por confórmero.

*Tal como se puede apreciar las posiciones ordenadas tienen un porcentaje muy alto de contactos por confórmero. Para las comparaciones estadísticas se realizó un test de Wilcoxon, los asteriscos representan valores de distinta significancia: **** representan valores-p <= 0.0001; *** representan valores-p <= 0.001; ** representan valores-p <= 0.01; * representa un valor-p <= 0.05; y la expresión ns (no significativo) implica que se obtuvo un valor-p > 0.05.*

En principio este resultado es razonable ya que se asumió que las conformaciones individuales de IDPs son más flexibles que las proteínas ordenadas y de ahí que posean menor cantidad de restricciones estructurales. Ahora bien, no es posible comprender a las proteínas si no se tienen en cuenta todo el ensemble conformacional. Como se mencionó anteriormente la enorme diversidad conformacional de las IDPs puede hacer que algunos confórmeros en una determinada posición establezcan contactos con ciertos residuos y que esa misma posición en otro confórmero no establezca ningún contacto (Figura 2.4). Para cuantificar la información de contactos se acumuló la información derivada de los contactos para cada confórmero en cada ensemble proteico de cada conjunto de datos. Esto significa que a lo largo del ensemble una posición era catalogada como un sitio de contacto si al menos un confórmero de todo el ensemble registraba al menos un contacto en dicha posición (Figura 2.5) .Si una posición está haciendo contacto

con 2 o más posiciones, esta posición de todas maneras se contabiliza como una posición de contacto. La información acumulada, es una aproximación razonable, ya que retiene la información sobre las contribuciones que cada confórmero en particular hace a la totalidad del ensemble para poder mantener la función biológica (Wei et al., 2016).

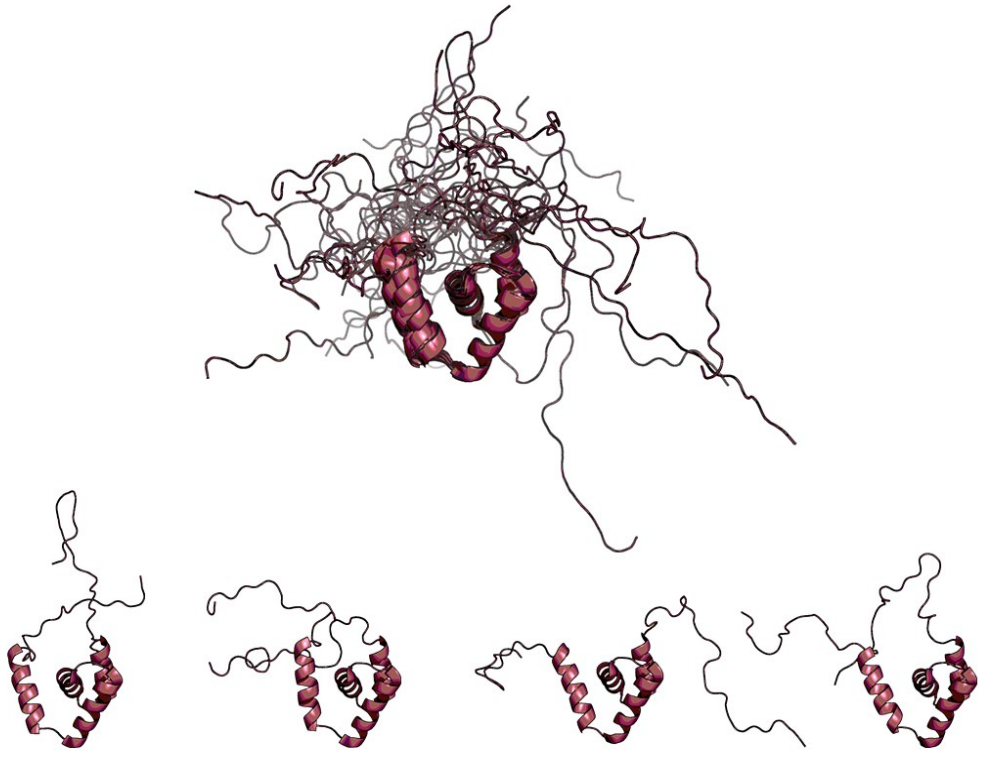


Figura 2.4 Representación de distintos confórmeros de un ensemble proteico.

Representación en formato del dominio UBA de ratón (código PDB 1WJ7_A, obtenida por NMR). En el panel superior se muestra el ensemble completo. En el panel inferior se muestran distintas conformaciones, de izquierda a derecha: modelo 19, modelo 1, modelo 7, modelo 1. La enorme flexibilidad de las regiones desordenadas hace que las conformaciones tengan diferencias estructurales notables y que por lo tanto la distribución de contactos por posición sea distinta.

Posición	1	2	3	4	5	6	7	8	9	10
Confórmero 1	1	0	1	1	1	0	0	0	0	0
Confórmero 2	0	0	0	0	1	1	0	0	0	1
Confórmero 3	1	0	1	0	1	0	0	1	0	1
Sitio de Contacto	1	0	1	1	1	1	0	1	0	1

Figura 2.5 Esquema del cálculo de sitios de contactos.

Dado un ensemble que contiene 3 confórmeros, para cada confórmero se calcula si una determinada posición está haciendo contactos (1) o no (0) con otro residuo. Si hay al menos un confórmero que presenta contactos en una determinada posición (como puede ser el caso de la posición 8 del esquema donde solo el confórmero 3 presenta contacto en dicha posición),este sitio es considerado un sitio de contacto. Si ninguno de los confórmeros presenta contactos en una determinada posición como la posición 9 del esquema, esa posición no es sitio de contacto). En el ejemplo ilustrado en este esquema 7 de las 10 posiciones de la proteína son sitios de contactos a lo largo de todo el ensemble.

Se pudo determinar, tal como se muestra en la Figura 2.6, que para los dos tipos de proteínas, la mayoría de las posiciones de las proteínas presenta al menos un contacto (a lo largo de todo el ensemble). A pesar de que las IDPs presentan una enorme flexibilidad estructural, tienen la misma cantidad de sitios de contactos acumulados a lo largo de todo el ensemble que las proteínas globulares (3er cuartil 100% y 97% para IDPs y proteínas globulares respectivamente)(Figura 2.6). Esto demuestra que cuando se tienen en cuenta la información disponible en el ensemble completo, la gran mayoría de las posiciones de las IDPs tanto como la de las proteínas ordenadas, se encuentran bajo restricciones estructurales que son comparables y que están debidas al menos a la presencia de contactos entre residuos.

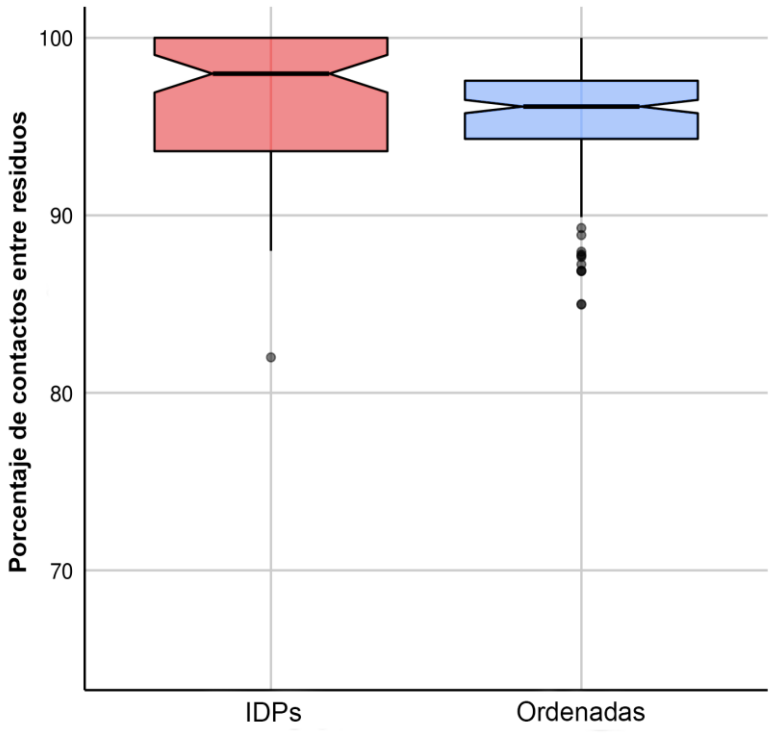


Figura 2.6 Porcentaje de posiciones con contactos entre residuos acumulado a lo largo del ensemble.

El porcentaje de posiciones con contactos entre residuos para el conjunto de datos de proteínas desordenadas y ordenadas tiene un valor promedio medio de 96.1%.

2.3.3 Determinación de restricciones estructurales por medio de métodos evolutivos.

En la sección anterior se estudiaron las restricciones estructurales debidas mayormente al establecimiento de contactos terciarios entre residuos, práctica muy bien establecida y casi fundacional de la Biología estructural de proteínas (Xia and Li, 1998), (Overington et al., 1992), (Overington et al., 1990), (Wood and Pearson, 1999), (Tourasse and Li, 2000). En esta sección se propone estudiar esas restricciones desde un enfoque evolutivo donde no solamente se tenga en cuenta el mantenimiento de una estructura dada sino las restricciones impuestas por la misma en la divergencia de las secuencias. El patrón de sustitución de ambos tipos de proteínas es sustancialmente distinto: las

IDPs, por un lado, presentan una alta conservación en la composición de aminoácidos (Moesa et al., 2012) mientras que las proteínas ordenadas presentan un patrón de sustitución sitio-específico conservado (Fornasari et al., 2002). Además, tanto las IDPs como las IDRs presentan velocidades de evolución y de inserciones y deleciones mucho más altas en comparación con las proteínas ordenadas (Forman-Kay and Mittag, 2013; Siltberg-Liberles et al., 2011) (Brown et al., 2011). Si se tiene en cuenta todo esto, el grado de restricción estructural encontrado derivado del conjunto de confórmeros del ensemble para las IDPs es más elevado de lo esperado. En este punto la hipótesis de trabajo fue la siguiente y básicamente deriva de la teoría neutral de evolución molecular. Si una posición es importante para mantener la estructura y/o la función de la proteína, esta posición tendrá un patrón de sustitución diferencial a otras posiciones que no son relevantes para la función y/o estructura de la proteína. Dicho de otro modo, si los contactos observados entre residuos son relevantes para la conformación, estructura, función de la proteína, estas posiciones deberían mostrar un patrón de sustitución diferencial. La hipótesis de trabajo entonces viró a considerar que los contactos observados entre residuos no son suficiente como para evaluar la importancia biológica de los mismos. En este punto entonces se considera complementar la información de los contactos físicos con estudios evolutivos del patrón de sustitución.

Mediante comparaciones de *Maximum Likelihood* (ML) (Figura 2.14) se estudió si el patrón de sustitución observado en el conjunto de proteínas homólogas a la proteína bajo estudio, era mejor explicado por un modelo evolutivo que utiliza la información estructural (SCPE) o por otros modelos que no tienen en cuenta esta información como (JTT, Dayhoff, WAG). Este análisis se realizó en forma sitio-específica y para todas las proteínas y confórmeros de los dos conjuntos de datos. Cuando se utiliza el criterio de ML, el árbol utilizado en la estimación del *likelihood* no afecta la selección de uno u otro modelo, en tanto la topología del árbol no esté definida al azar (Posada and Crandall, 2001) y pueden compararse sobre el mismo MSA los valores de ML obtenidos para distintos modelos en este caso SCPE, JTT, Dayhoff y WAG. Aquellas posiciones que se encuentren sometidas a restricciones estructurales definidas, por ejemplo en caso de posiciones que son claves para la estructura proteica, es de esperar que el valor de *maximum likelihood* alcanzado con las matrices SCPE sea superior al obtenido con JTT,

Dayhoff o WAG. Estos sitios fueron como sitios restringidos SC (por la expresión en inglés, vinculada al modelo utilizado, Structurally Constrained para referir a estructuralmente restringidos). Más detalles sobre la manera en que estos sitios fueron definidos puede encontrarse en la sección de métodos de este capítulo.

En una primera instancia, se calculó el porcentaje de sitios SC para confórmeros individuales, es decir sin acumular a lo largo del ensemble y se encontró que la proporción de sitios restringidos estructuralmente era de 32.1% para IDPs y 36.1% para proteínas ordenadas (Figura 2.7) mostrando un valor menor que el obtenido a partir de la información de los contactos terciarios (Figura 2.3). Esto podría evidenciar que la información proveniente únicamente de contactos físicos no es del todo sensible para evidenciar los condicionamientos estructurales a los que está sometida una proteína.

Una vez definidos los sitios SC para cada confórmero de cada ensemble, se calculó el porcentaje de sitios restringidos evolutivamente a lo largo de todo el ensemble de manera análoga al cálculo de porcentaje de sitios de contacto (Figura 2.8). Si una posición en al menos un confórmero es catalogada como SC, entonces esa posición es registrada como SC en el ensemble.

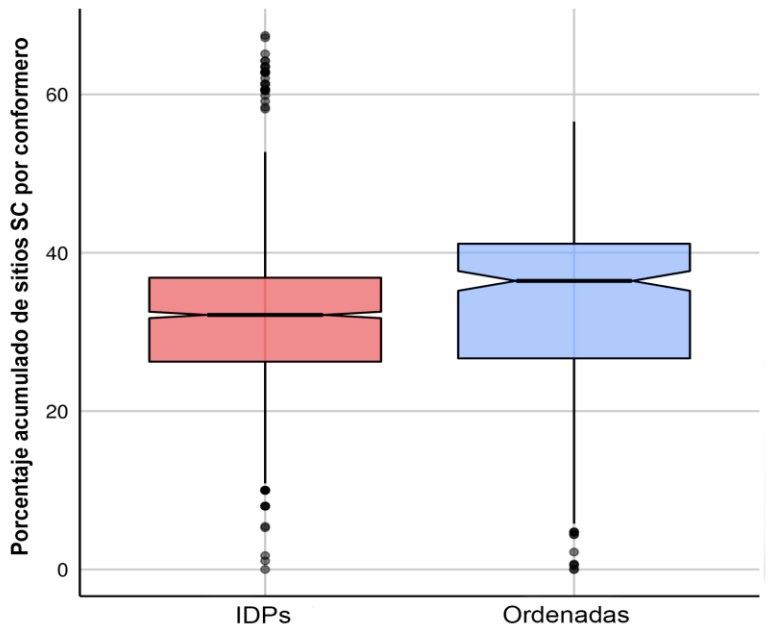


Figura 2.7 Distribución de sitios estructuralmente restringidos (SC) por confórmero.

Mediana de 32.1% de sitios para las proteínas desordenadas y un 36.1% para las proteínas globulares.

Posición	1	2	3	4	5	6	7	8	9	10
Confórmero 1	0	0	0	1	0	1	0	0	0	0
Confórmero 2	0	0	0	1	1	1	0	1	0	0
Confórmero 3	0	0	1	0	0	0	0	1	0	1
sitio SC	0	0	1	1	1	1	0	1	0	1

Figura 2.8 Esquema de determinación de sitios estructuralmente restringidos (SC) evolutivamente a lo largo del ensemble.

Para cada confórmero, una determinada posición es catalogada como SC (1) o no SC (0). Si para una determinada posición al menos un confórmero presenta restricciones evolutivas (SC) entonces esa posición es SC en todo el ensemble (ejemplo posición 3,5,10 del esquema). Si una posición es catalogada como SC en más de un confórmero, de todas manera se cataloga simplemente como SCa lo largo del ensemble (posiciones 4,6 y 8 del esquema). Si una posición no es catalogada como SC en ningún confórmero entonces ese sitio es no SC para todo el ensemble (posiciones 1,2,7 y 9 del esquema)

Bajo esta nueva forma de determinar las restricciones estructurales, sorprendentemente se encontró que el porcentaje de sitios estructuralmente restringidos (SC) a lo largo de cada ensemble es casi el mismo en ambos tipos de proteínas. Las proteínas desordenadas presentan un promedio de 41.6 % de sitios SC y las ordenadas un promedio de 40.5% de sitios SC (Figura 2.9) y que este porcentaje es mucho menor que el calculado cuando se considera únicamente la cantidad acumulada de contactos entre residuos (Figura 2.6).

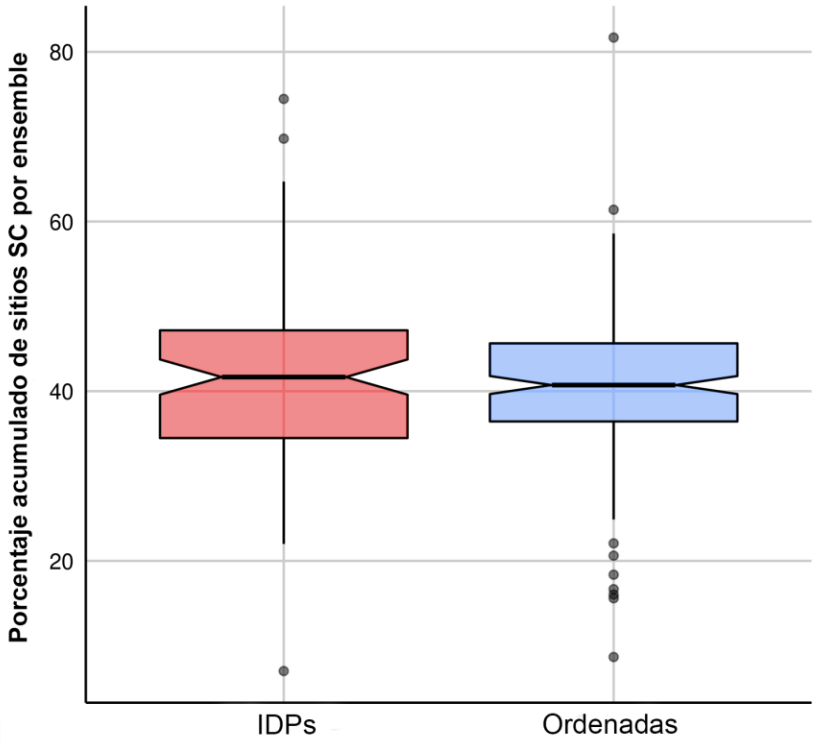


Figura 2.9 Distribución del número de sitios restringidos estructuralmente (SC) acumulados a lo largo del ensemble.

Las proteínas desordenadas presentan un 41.6% de SC sites, mientras que las proteínas globulares presentan un 40.5%. Las distribuciones son estadísticamente similares. La significancia estadística se calculó usando un test Kolmogorov-Smirnov que arrojó un p-valor = 0.39 y Mann-Whitney-Wilcoxon test con un p-valor = 0.45.

A partir de estos resultados se puede concluir que, por un lado, los métodos evolutivos son más sensibles que los contactos a la hora de evidenciar los condicionamientos

estructurales. Por otro lado, se esperaría encontrar que los porcentajes de sitios SC en confórmeros individuales fuera mucho menor al porcentaje de sitios SC en el ensemble completo, por lo que resulta llamativo encontrar que en el porcentaje sitios SC para confórmeros individuales en IDPs (36.1%, Figura 2.7) no es un valor estadísticamente diferente del porcentaje de sitios SC en el ensemble entero (41.6%, Figura 2.9). Podría ser entonces un indicio de información estructural en el ensemble es redundante. Un desarrollo más profundo sobre esto último se puede encontrar en la sección 4.1.

2.3.4 Sitios restringidos en proteínas desordenadas

Como se mencionó anteriormente las IDPs, a pesar de ser mayormente desordenadas, pueden poseer regiones globulares ya sean transitorias o permanentes (
Figura 2.10), que al presentar mayor cantidad de contactos entre residuos podrían causar un incremento en las restricciones estructurales. Para poder ahondar sobre esta hipótesis, se investigó cómo era la distribución de sitios SC en proteínas desordenadas. Si se tiene en mente que los sitios SC son aquellos que modulan la divergencia de la secuencia en esa posición específica, es esperable que estos sitios estuvieran localizados en regiones ordenadas o globulares de IDPs debido a que en estas regiones es donde se encuentra la mayor restricción estructural debida al establecimiento de contactos y que no se encuentren en posiciones desordenadas. Sin embargo, cuando se estiman los sitios SC únicamente en proteínas desordenadas, encontramos que estos sitios se encuentran en promedio en el 40% de las posiciones ordenadas y en promedio en el 27% de las posiciones desordenadas. Es decir que las restricciones estructurales encontradas en proteínas desordenadas no solo están impuestas sobre posiciones ordenadas, las posiciones desordenadas también se encuentran restringidas estructuralmente (Figura 2.11).

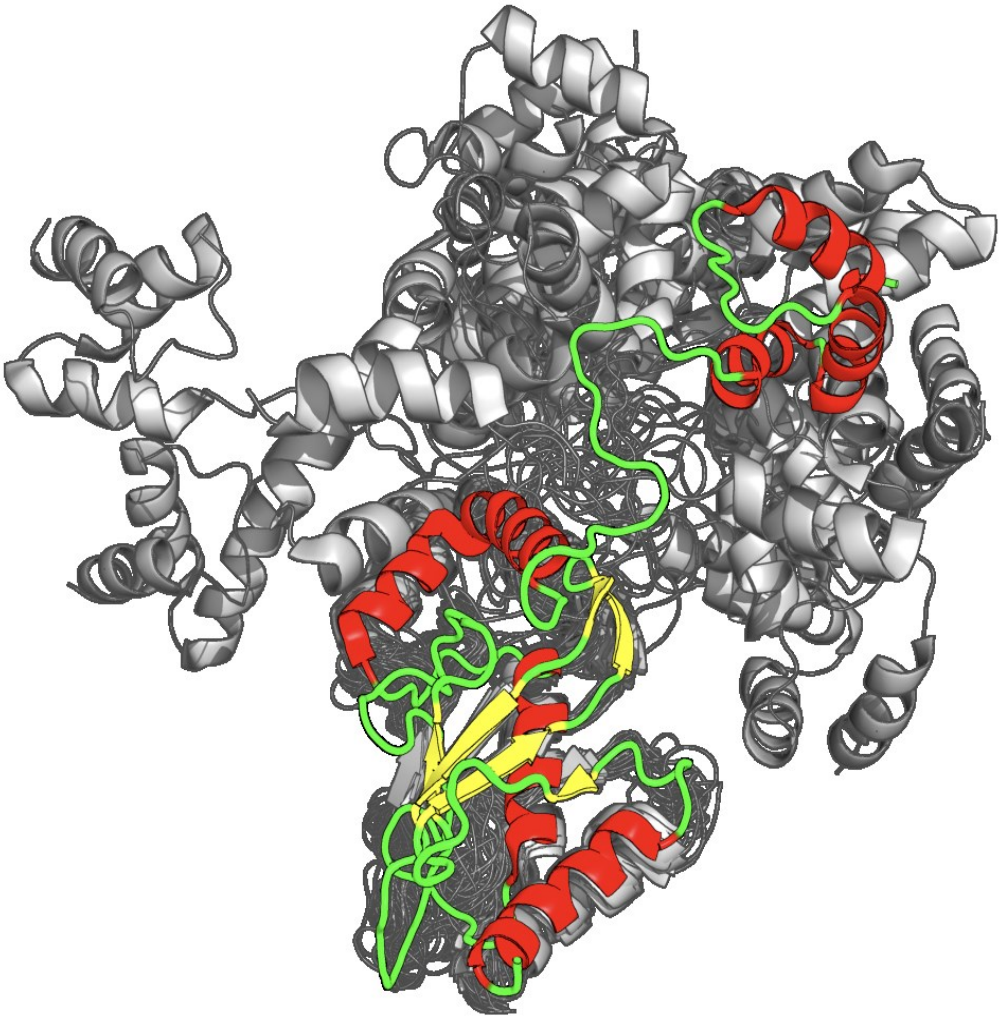


Figura 2.10 Representación del ensemble conformacional conformacional de IDP.

Representación esquemática de la enzima conjugadora de ubiquitina de clase II, Ubc1 (código PDB: 1TTE_A). Se puede observar que esta proteína posee elementos de estructura secundaria mostrados en rojo y también regiones flexibles (mostradas en verde). Para simplificación en la interpretación de la imagen, se coloreó solo un confórmero mientras que el resto del ensemble se muestra en gris.

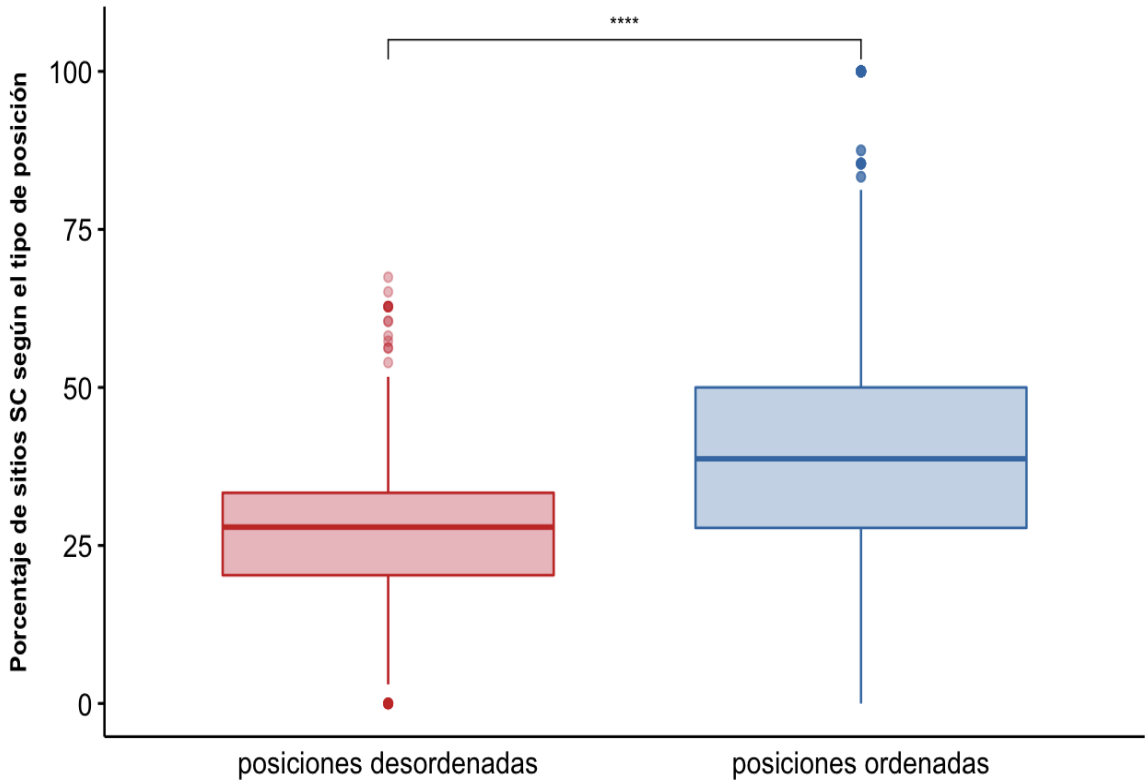


Figura 2.11 Distribución de sitios estructuralmente restringidos (SC) en IDPs

Boxplots mostrando la sitios estructuralmente restringidos (SC) en IDPs que caen en posiciones desordenadas (media 27%) y en posiciones ordenadas (media 40%).

En la Figura 2.12 se muestran en color rojo los sitios restringidos SC en color rojo y tal como se puede deducir las proteínas desordenadas pueden tener un alto grado de diversidad conformacional, sin embargo entre todas esas distintas conformaciones se puede observar que algunas de ellas adoptan una estructura más globular que otras, posiblemente incrementando de esta manera el número de contactos por sitio.

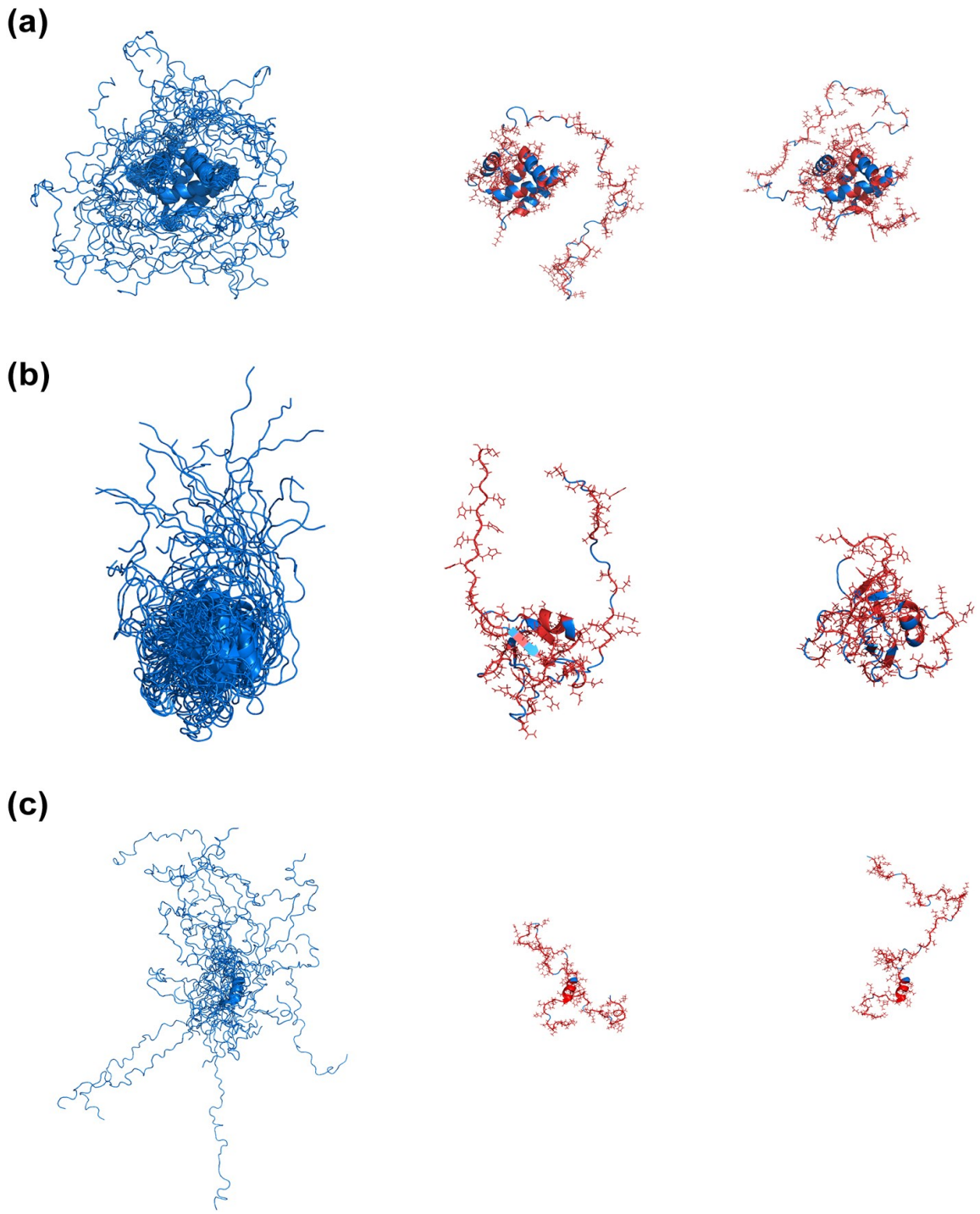


Figura 2.12 Se muestra la distribución de sitios estructuralmente restringidos (SC) en ensembles de IDPs.

Los sitios SC se muestran en color rojo. A la izquierda de los tres paneles A (código PDB: 2JRF_A), B (código PDB: 2ADZ_A) y C (código PDB: 5MRG_A) se muestra el ensemble completo de las citadas proteínas, mientras que en el centro y a la derecha se muestran confórmeros individuales de dichos ensembles.

En este trabajo también se estudió qué porcentaje de los sitios SC tenían contactos en el 100% de sus confórmeros. Como se puede apreciar en la Figura 2.13 el 51% de los sitios SC presentan contactos en la totalidad de los confórmeros que forman parte del ensemble. Sin embargo, tal como se puede ver en la Figura 2.13 existe una cola en la distribución hacia valores más bajos implicando que a lo mejor solamente algunas pocas conformaciones entre el conjunto total del ensemble pueden presentar contactos entre residuos que modulen el patrón de sustitución para esas posiciones determinadas.

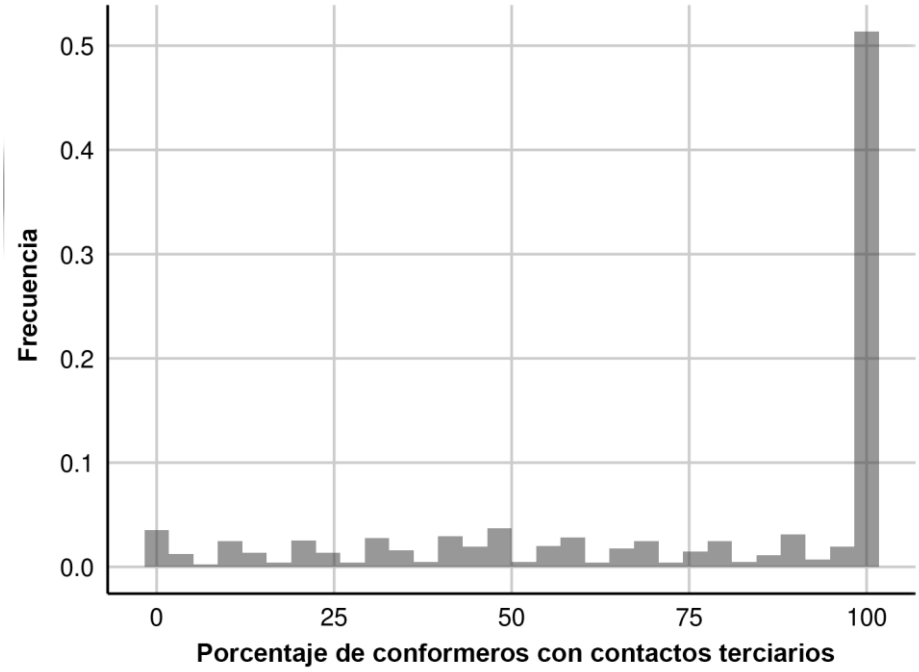


Figura 2.13 Porcentaje de conformaciones con contactos terciarios en sitios sitios estructuralmente restringidos (SC).

Aproximadamente el 51% de los sitios SC tienen contactos en el 100% del ensemble.

2.4 Discusión

En ambos tipos de proteínas, el número de posiciones que tienen contactos terciarios (determinados por los residuos de los aminoácidos y no por el esqueleto carbonado de la proteína (ver sección métodos de este capítulo)) entre residuos acumulado a lo largo de todas las conformaciones del ensemble, es alrededor del 100% (Figura 2.6) sin embargo, el porcentaje de sitios estructuralmente restringidos, determinados no solo desde el punto de vista estructural sino desde una aproximación evolutiva, es mucho menor (alrededor del 40% Figura 2.9). Esta reducción es esperable, si se tiene presente que no todos los contactos físicos establecidos entre residuos son igualmente relevantes para la estabilización de la estructura terciaria (Sathyapriya et al., 2009). También es posible que los modelos estructurales y los potenciales energéticos para su derivación sean inadecuados y la incertidumbre en las coordenadas atómicas impacten en la diferencia observada entre la cantidad de contactos físicos y la restricción estructural impuesta por la evolución (Garbuzynskiy et al., 2005),(Spronk et al., 2002), (Spronk et al., 2003).

Esta disminución entre el número de contactos físicos y los derivados del análisis evolutivo, también puede ser atribuida a la falta de información estructural confórmero específica en los alineamientos múltiples (Spronk et al., 2003). Esta última razón podría tener efecto sobre las matrices de sustitución generadas por el SCPE, debido a que las mismas son sitio-específicas y plegamiento y/o confórmero específicas y fueron evaluadas utilizando alineamientos de secuencias de proteínas homólogas. Estos MSA contienen información evolutiva de distintos tipos como pueden ser divergencia estructural (Illergård et al., 2009) o adaptaciones en la dinámica molecular (Monzon et al., 2017b; Zea et al., 2018) que podrían en cierta medida modificar el patrón de contactos en proteínas homólogas. Por todo lo anteriormente dicho, es esperable que ese 40% de sitios estructuralmente restringidos promediados a lo largo de todo el ensemble, en ambos tipos de proteínas, no capturen contactos o interacciones entre residuos que son sutiles y probablemente originados como consecuencia de adaptaciones funciones para confórmeros particulares. En concordancia con esta última observación, recientemente

se ha demostrado que usando MSAs (Zea et al., 2018) es posible recuperar un patrón más conservado de residuos entre contactos que cuando se utilizan métodos de co-evolución y de evolución acoplada.

Otro resultado importante es que tanto las proteínas ordenadas como las desordenadas muestran restricciones estructurales impuestas sobre la divergencia secuencial que son comparables tal como se muestra en la Figura 2.9. Este resultado sugiere que las contribuciones individuales de cada confórmero en los ensembles de IDPs son importantes para mantener la función biológica, como ya hace tiempo está establecido para las proteínas ordenadas y que más recientemente se ha empezado a observar para proteínas desordenadas. (Siltberg-Liberles et al., 2011; Wei et al., 2016) (Brown et al., 2011). Estas pequeñas contribuciones de cada confórmero en ensembles de IDPs dan lugar a la misma proporción de restricciones estructurales encontradas en proteínas ordenadas, posiblemente con el distinto peso o importancia que tengan en relación al rol biológico.

Los resultados aquí presentados, demuestran la importancia de análisis evolutivos en la discriminación de contactos físicos entre residuos que guarden información biológicamente relevante de aquellos que no lo son.

2.5 Métodos específicos del capítulo

2.5.1 Construcción de los conjuntos de datos

Las proteínas ordenadas o globulares fueron obtenidas de la base de datos CoDNaS (Monzon et al., 2016). Teniendo en cuenta la presencia de residuos faltantes catalogados como '*missing*' como principal indicador de regiones intrínsecamente desordenadas (IDRs) (Dunker et al., 2001) y se seleccionaron 183 proteínas no tuvieran residuos '*missing*' en cada uno de los confórmeros del ensamble. Estas proteínas tienen al menos 5 confórmeros depositados en la base de datos, lo que otorgó seguridad en la estimación

de la variabilidad conformacional (Best et al., 2006). Para el conjunto de datos de proteínas ordenadas, sólo se tomaron los confórmeros que tuvieran el máximo RMSD entre todos los confórmeros del ensemble. Estas proteínas mostraban niveles insignificantes de desorden predicho con ESpritz X-ray (Walsh et al., 2012)

Para el caso de las proteínas desordenadas, se predijo y se estimó el desorden en todas los NMR disponibles en la PDB (a mayo 2018), utilizando NMR-ESpritz y Mobi 2.0 (Piovesan and Tosatto, 2018). Luego de un proceso de curado a mano, en donde se tuvo en cuenta la biología de la proteína y su longitud, el set de datos quedó finalmente conformado con 93 ensembles de NMR que tenían más del 40% de sus residuos predichos como desordenados Figura 2.2.

2.5.2 Estimación de contactos terciarios

Los contactos terciarios entre residuos fueron estimados como aquellos en donde los átomos pesados de dos residuos que no estuvieran posiciones contiguas y que tuvieran sus respectivas esferas de van der Waals a una distancia menor a 1.0 Å. Los residuos de largo alcance fueron determinados usando la misma definición que para los contactos terciarios, pero esta vez considerando dos posiciones que estuvieran al menos a 4 residuos contiguos de distancia.

2.5.3 Estimación de sitios estructuralmente restringidos SC

Para las proteínas desordenadas se tuvieron en consideración todos los confórmeros presentes en el ensemble de NMR. En el caso de los ensembles de proteínas globulares solo se tuvieron en cuenta para el análisis, el par de confórmeros de máximo RMSD de acuerdo con la información provista en CoDNaS (Monzon et al., 2016).

Para cada confórmero de cada ensemble de cada conjunto de datos se corrió el modelo SCPE (Parisi and Echave, 2001). El SCPE deriva matrices de sustitución sitio-específicas, mediante simulaciones evolutivas bajo condiciones neutrales para la

conservación de plegamiento (Fornasari et al., 2002; Parisi and Echave, 2005). Como fue explicado en capítulos anteriores, el SCPE utiliza cálculos energéticos para evaluar el grado de perturbación estructural como consecuencia de la introducción de una mutación no sinónima, durante el proceso de simulación. Utilizando estimaciones de *Maximum Likelihood* (ML) es posible comparar las matrices del SCPE con modelos que carecen de información estructural como JTT (D T Jones et al., 1992), Dayhoff (Dayhoff et al., 1978) and WAG (Whelan and Goldman, 2001). En la Figura 2.14 se puede encontrar un esquema del flujo de trabajo utilizado para el cálculo de ML sitio-específica.

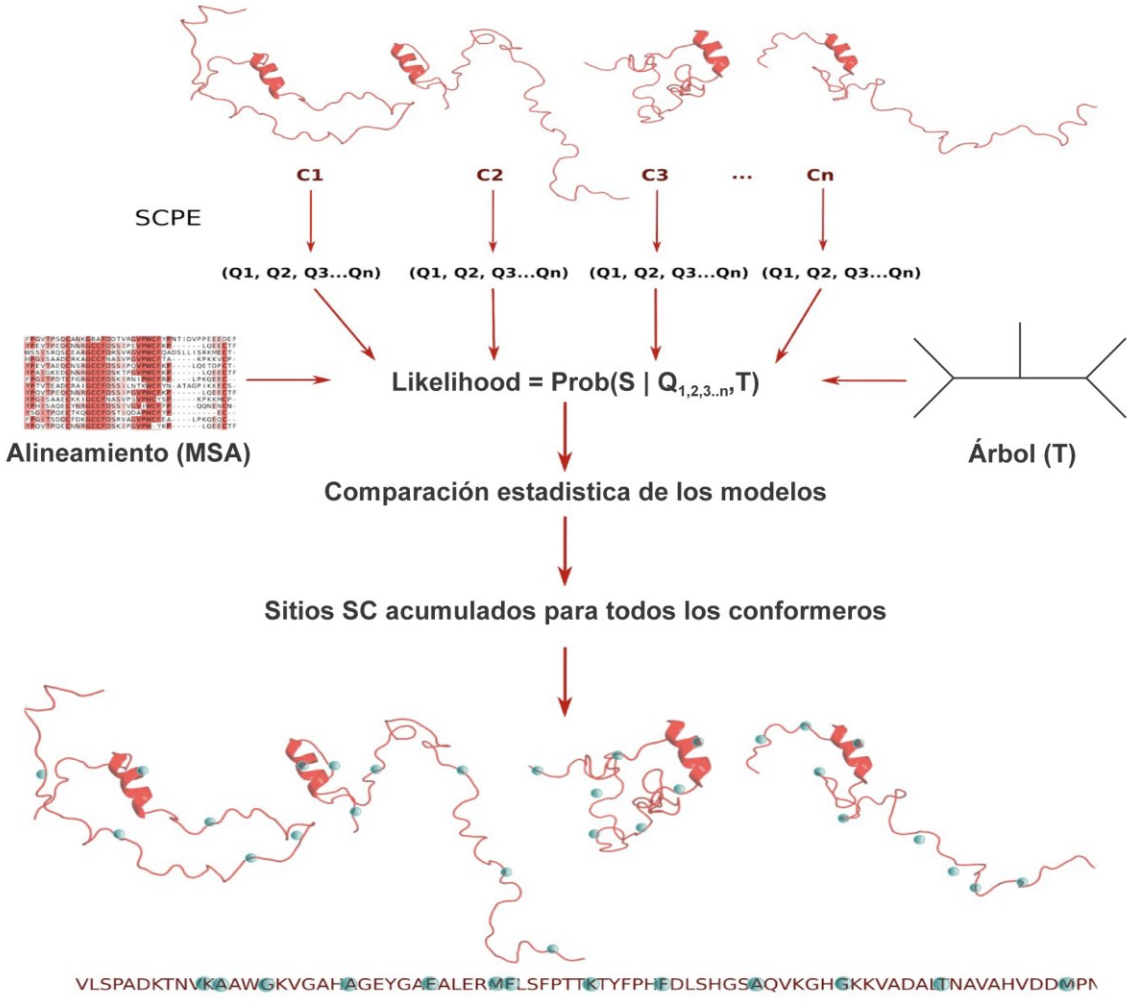


Figura 2.14 Esquema de flujo de trabajo.

Descripción de la estimación de sitios restringidos estructuralmente (sitios SC). Con el modelo SCPE, es posible obtener para cada confórmero un conjunto de matrices de sustitución sitio-específicas (Q1,Q2..Qn) y que a su vez son estructura-dependientes y por lo tanto son confórmero específica. Para cada proteína es posible determinar cuán bien cada una de estas matrices explica el patrón de sustitución observado usando métodos de máxima verosimilitud (ML). El valor del Log(L) da cuál es la mejor probabilidad de observar un determinado patrón de sustitución en un MSA (representado como S) en concordancia con un determinado árbol filogenético (representado como T) bajo un determinado modelo de evolución descrito por Q1,Q2...Qn. En este trabajo cuando el SCPE explica mejor el patrón de sustitución observado frente a modelos que no consideran información estructural (irrestrictos) son considerados sitios restringidos y además esa posición es una posición de contactos, entonces ese sitio es inferido como restringido estructuralmente (sitios SC). Esa información es luego acumulada a lo largo de todo en el ensemble.

Los cálculos sitio-específicos de ML, fueron realizados utilizando el paquete HYPHY (Pond et al., 2005), para tal fin eran necesarios un MSA y los árboles filogenéticos correspondientes. Los MSA que se utilizaron para el análisis de ML fueron obtenidos de la base de datos HSSP (Touw et al., 2015) y los árboles filogenéticos fueron calculados con el paquete Phylip (Baum, 1989) utilizando un algoritmo de Neighbour-joining.

Para cada posición de cada confórmero de cada ensemble es posible calcular un valor de log(L) asociado a cada uno de los modelos utilizados para la comparación. La evaluación estadística del ajuste de cada modelo se realizó a través del coeficiente AIC (Akaike Information Criterion o criterio de información de Akaike) (Akaike, 1974), definido para un modelo dado según:

Ecuación 10: Cálculo del AIC.

$$AIC_{model} = -2lnL_{p,model} + 2K$$

Donde $\mathbf{L_p}$ es el valor de *likelihood* obtenido para una posición para un determinado modelo y **K** es el número de parámetros del modelo.

El AIC se puede a la comparación de la evolución global o sitio-específica de la proteína. Una ventaja de este criterio es que permite la comparación directa de múltiples modelos de evolución a través de la diferencia de sus valores de AIC (Guthery et al., 2003), calculada según:

Ecuación 11: Cálculo de ΔAIC.

$$\Delta AIC = AIC_i + AIC_{min}$$

Donde **AIC_i**, es el **AIC** obtenido para cada modelo ***i*** a evaluar y **AIC_{min}** el menor de todos los AIC. Si el modelo presentaba un ΔAIC <= 2 (valor aceptado como límite superior) indicaba que la evidencia apoya sustancialmente al modelo y entonces era considerado como un modelo muy bueno, si el valor de ΔAIC estaba entre 2 y 4, entonces ese modelo era considerado moderado, mientras que si ese valor era > 10, ese modelo era malo. Es importante tener presente que en este trabajo se comparan 4 modelos de evolución de manera simultánea. En la Figura 2.15 se muestra un esquema de comparación para una determinada posición:

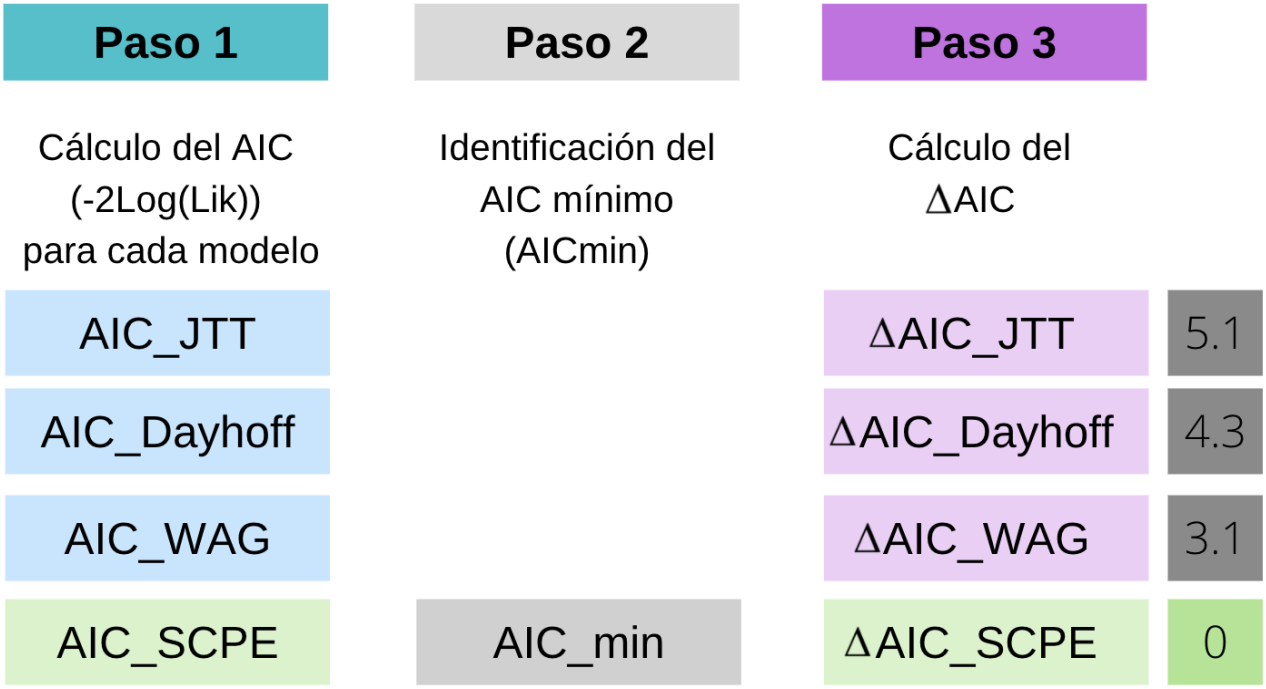


Figura 2.15 Esquema de flujo de comparación de modelos para una determinada posición.

En el primer paso se calcula el AIC para cada modelo. En el segundo paso se identifica cuál de los 4 valores de AIC es el mínimo (en el esquema lo obtiene el SCPE) y en el paso 3 se calculó el valor ΔAIC que cuando da valores <=2 se puede considerar que ese modelo es adecuado.

Basados en la información obtenida por los análisis de ML una determinada posición era inferida como un sitio que evoluciona bajo restricciones estructurales (llamado SC), si simultáneamente para una posición determinada se daban dos condiciones:

(1) La matriz de sustitución sitio-específica generada por el SCPE explicaba mejor el patrón de sustitución observado para esa posición que alguno de los modelos irrestrictos utilizados para la comparación, es decir que obtenía un valor de ΔAIC_SCPE <= 2. Esto indicaba la restricción impuesta por la estructura a la divergencia secuencial es decir teniendo en cuenta la información derivada de la evolución de las proteínas y de su estructura.

(2) Ninguno de los modelos irrestrictos provee una buena explicación para el patrón de sustitución observado. Es decir que los ΔAIC_JTT, ΔAIC_Dayhoff y el ΔAIC_WAG tenían valores > 2.

Esta aclaración es importante, debido a que se compara el desempeño del SCPE contra 3 modelos irrestrictos. Se puede dar la situación que una determinada posición sea explicada por el SCPE (restricto) y por JTT (irrestricto) y en ese caso el sitio no se considera estructuralmente restringido, porque hay al menos un modelo restricto que explica bien su patrón de sustitución (Figura 2.16). Estas posiciones donde el SCPE y algún modelo irrestricto tienen un nivel de explicación similar (sus ML son indistinguibles) podría ocurrir con posiciones expuestas al solvente o con bajo número de contactos. El SCPE en estos casos da matrices de sustitución que son similares a la de los modelos irrestrictos ya que varios aminoácidos para esa posición podrían dar un buen score de aceptación y por lo cual la matriz de sustitución obtenida para ese sitio podría parecerse a la de un modelo de sustitución irrestricto.

Posición	1	2	3	4	5	6	7	8	9	10
Modelo JTT	1	0	1	1	1	0	0	0	0	0
Modelo Dayhoff	0	0	0	0	1	1	0	0	0	1
Modelo WAG	1	0	1	0	1	0	0	1	0	1
Modelo SCPE	0	1	0	1	0	0	1	0	1	0
Sitio SC	0	1	0	0	0	0	1	0	1	0

Figura 2.16 Esquema de definición de sitio sitios estructuralmente restringidos (SC).

Si para una posición una determinada posición es explicada por uno de los modelos citados a la izquierda, se le asigna 1 y si no 0. Un sitio SC es aquel que únicamente es explicado por el modelo de evolución con base estructural SCPE (posiciones 2 ,7 y 8 en el esquema). Los sitios irrestrictos son aquellos que son explicados por al menos un modelo de evolución irrestricto (posiciones 1,3,5,6,10 en el esquema), sin importar si para ese sitio el SCPE tuvo un buen desempeño (posición 4 en el esquema). La metodología de evaluación de desempeño se explica más arriba en esta misma sección.

3 Las velocidades de evolución de IDPs son modeladas por los ensembles conformacionales.

3.1 Resumen

Como ya fue mencionado en el curso de esta tesis, las Proteínas Intrínsecamente Desordenadas (IDPs) carecen de una estructura terciaria estable bajo condiciones fisiológicas. Su estado nativo está descrito por un ensemble conformacional formado por confórmeros parcial o completamente desplegados. Estas proteínas presentan un desafío a la Biología debida a su dinámica compleja, su composición tan característica y su gran diversidad conformacional. Utilizando información estructural obtenida por NMR, en esta sección de la tesis se encontró que las IDPs presentan tasas de evolución sitio-específicas muy heterogéneas dentro de una misma proteína, debidas principalmente a diversas restricciones estructurales originadas por el establecimiento de contactos físicos entre residuos. Se pudo encontrar además que los perfiles de velocidad correlacionan con la diversidad conformacional observada experimentalmente en este tipo de proteínas, permitiendo la descripción de diferentes patrones conformacionales, posiblemente relacionados con la relación estructura-función de la proteína. Estos resultados por un lado sugieren que los contactos entre residuos en las IDPs restringen la velocidad de evolución de manera tal de conservar el comportamiento dinámico del ensemble conformacional y por otro lado que las velocidades de evolución pueden ser utilizadas como una aproximación sobre la diversidad conformacional de las IDPs.

3.2 Introducción

El estado nativo de las proteínas está compuesto por múltiples confórmeros que se encuentran en un equilibrio dinámico, en conjunto todas estas estructuras forman lo que se conoce como ensemble nativo (Wei et al., 2016). Para el caso de proteínas globulares

(conocidas como “ordenadas”) o dominios globulares, las barreras energéticas presentes para las transiciones entre los distintos confórmeros son relativamente altas, permitiendo de esta manera explorar, mediante distintos métodos y técnicas experimentales, conformaciones discretas y bien establecidas (Monzon et al., 2016). Dentro del continuo estructural del espacio conformacional de las proteínas (mencionado en la introducción de la tesis) se puede encontrar en un extremo de dicho continuo, confórmeros que presentan pequeñas diferencias estructurales entre ellos (comportamiento típico de proteínas llamadas rígidas) donde movimientos pequeños o rotaciones en residuos que no involucran traslaciones a nivel del esqueleto carbonado, permiten la llegada de ligandos al sitio activo de la proteína, la apertura de túneles o el alargamiento de cavidades (Figura 1.16) (Monzon et al., 2017a). Los movimientos de loops, el desplazamiento de elementos de estructura secundaria y movimientos de rotación de dominios contribuyen al incremento de las diferencias estructurales entre confórmeros (Gerstein and Krebs, 1998).

En el otro extremo del continuo estructural se encuentran las IDPs que presentan una alta plasticidad y barreras energéticas bajas entre los distintos confórmeros (Berlow et al., 2015). Las características distintivas de este conjunto de proteínas no suponen una restricción para su distribución en los distintos reinos y taxas. Por ejemplo, aproximadamente el 40% de los proteomas de eucariotas (Lobanov and Galzitskaya, 2015) y aproximadamente el 46% de entradas de Uniprot (The UniProt Consortium, 2017) contienen regiones desordenadas cortas (Necci et al., 2016). Más aún es bien conocido el hecho que estas proteínas juegan un rol crucial en muchas enfermedades neurodegenerativas y en desórdenes sistémicos (Uversky et al., 2009).

Tanto para las proteínas globulares (u ordenadas) como para las IDPs, el concepto del ensemble estructural nativo es clave para dos explicaciones alternativas sobre la descripción de una serie de eventos complejos relacionados con la unión a ligando (Nussinov et al., 2014). Estas dos explicaciones son: el modelo de selección conformacional (de pre-equilibrio) y el modelo de encaje o ajuste inducido mencionados anteriormente al tratar los modelos de Daniel Koshland y de Jacques Monod.

Brevemente, en el modelo de selección conformacional, un conjunto de conformaciones que se encuentran pobremente representadas puede presentar una alta afinidad para un determinado ligando biológico. Los eventos de unión producen un cambio en el equilibrio conformacional que produce un incremento en la concentración relativa de estas conformaciones competentes fenómeno denominado *population shift*. Por otro lado, el modelo de ajuste inducido propone que la adopción de una conformación preferida ocurre luego del evento de unión al ligando, en un proceso que puede involucrar cambios conformacionales y transiciones orden-desorden en el caso de IDPs. Cualquiera sea el modelo apropiado para explicar la biología de una proteína determinada (Hammes et al., 2009), la pre-existencia de conformaciones escasamente pobladas competentes o la reestructuración conformacional posterior a la unión de ligando, trae aparejado la noción que las IDPs no son proteínas completamente desestructuradas (Berlow et al., 2015) y que existen determinadas conformaciones que contienen información estructural residual que juega un papel clave en la función biológica de dicha proteína (Berlow et al., 2015; Davey, 2019; Mészáros et al., 2019; Reisen et al., 2010; Zea et al., 2016).

Un parámetro que es extensamente utilizado para predecir las velocidades de evolución es el nivel de expresión génica (Drummond et al., 2005) encontrándose una relación inversa. Esto último es entendiendo a las velocidades de evolución de las proteínas como un parámetro global, sin embargo, cada posición de cada proteína puede evolucionar bajo una presión de selección distinta del resto de las posiciones y dar origen a velocidades de evolución heterogéneas en una misma proteína. Por ejemplo para las proteínas ordenadas, los sitios que presentan un gran número de contactos entre residuos o con una alta densidad de empaquetamiento (como aquellos residuos que forman parte del núcleo proteico) o con baja exposición al solvente, tienden a evolucionar más lento (Franzosa and Xia, 2009; Tóth-Petróczy and Tawfik, 2011; Yeh et al., 2014a). Muchos estudios encontraron correlaciones entre velocidades de evolución sitio-específicas y distintas características estructurales (para un review ver: (Echave et al., 2016)).

¿Qué sucede con las IDPs e IDRs? Es preciso recordar que el universo proteico es mucho más amplio y no solamente comprende proteínas globulares. En una primera aproximación las regiones desordenadas tienden a evolucionar más rápido que las proteínas ordenadas (Brown et al., 2002). Esto es un resultado esperable debido a que sus residuos presentan un bajo número de contactos entre residuos (Dosztányi et al., 2005) y una alta exposición al solvente (Davey, 2019). A pesar que trabajos previos detectaron sitios restringidos evolutivamente en IDPs (Marchetti et al., 2019; Pancsa et al., 2018; Toth-Petroczy et al., 2016), el rol que estos sitios tienen en las IDPs con una alta flexibilidad estructural, no está claro. En proteínas que poseen una alta diversidad conformacional (como es el caso de las IDPs) podría esperarse que algunos confórmeros que se encuentran parcialmente plegados y que presentan elementos de estructura secundaria transitorios, impongan condicionamientos estructurales específicos y aditivos sobre las velocidades de evolución sitio-específicas. En este capítulo de la tesis, se describen las metodologías y resultados, para respaldar esta última hipótesis mediante la evaluación del impacto que tiene la información estructural de IDPs derivada de datos experimentales, en la velocidad de evolución. Incluso, podría suceder que estos condicionamientos estructurales derivados del ensemble nativo revelen la existencia de distintos tipos de funcionalidad de IDPs, en especial aquellas relacionadas con transiciones orden-desorden posterior a la unión a ligando (Davey, 2019).

3.3 Resultados

3.3.1 Los perfiles de velocidad permiten la caracterización de la organización estructural de las IDPs

El conjunto de datos utilizado en este trabajo consistió en 310 estructuras de IDPs obtenidas por NMR. Estas proteínas contienen al menos el 40% de sus posiciones desordenadas (para una explicación más detallada sobre la estimación de desorden ver los métodos específicos de este capítulo). La Figura 3.1 ilustra una estructura típica de una IDP distinguiendo entre regiones ordenadas (azul) y desordenadas (rojas). De la

comparación entre las velocidades de evolución normalizadas de ambos tipos de regiones, es decir de regiones ordenadas y regiones desordenadas de IDPs, presentes en todo el conjunto de datos, se puede deducir que las regiones desordenadas presentes en las IDPs tienen tasas de evolución significativamente más altas que las regiones ordenadas de las mismas proteínas (Figura 3.2).

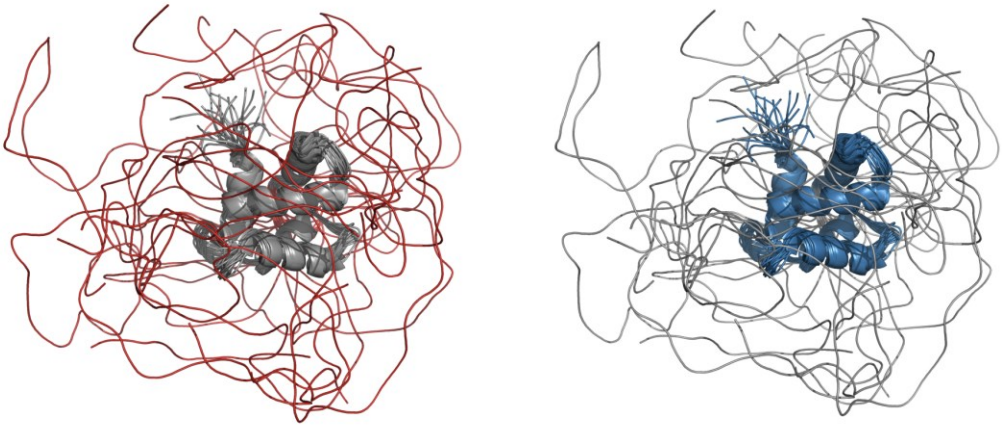


Figura 3.1 Representación de un ensemble IDPs.

Ambas figuras son una representación hecha con el software PyMol de la misma proteína promotora de polimerización de tubulina de humanos (TPPP3 código PDB: 2JRF). En la figura de la izquierda se resalta en color rojo las posiciones desordenadas de esta proteína mientras que en el lado derecho de la figura se muestra en color azul las posiciones ordenadas.

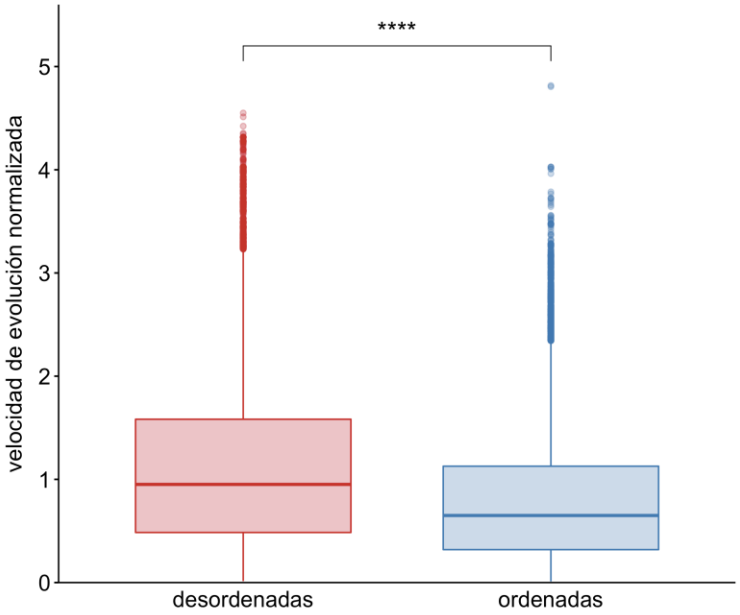


Figura 3.2 Boxplots representando la distribución de las velocidades de evolución normalizadas por posición.

*Las velocidades normalizadas para las posiciones desordenadas son mostradas en rojo (mediana de 0.95) y las posiciones ordenadas son mostradas en azul (mediana de 0.65) del conjunto de datos entero de IDPs. Para las comparaciones estadísticas se realizó un test de Wilcoxon, los asteriscos representan valores de distinta significancia: **** representan valores-p <= 0.0001; *** representan valores-p <= 0.001; ** representan valores-p <= 0.01; * representa un valor-p <= 0.05; y la expresión ns (no significativo) implica que se obtuvo un valor-p > 0.05.*

Esta heterogeneidad observada en las velocidades de evolución de IDPs, se comprende mejor si se estudia cuál es la relación de los perfiles de velocidades posición-específica con la variación estructural observada en el ensemble proteico. Se encontró que las velocidades de evolución por sitio muestran una alta correlación con su variación estructural en el ensemble, medida con el parámetro RMSF (Root Mean Square Fluctuation) de sus carbonos alfa tal como se muestra en la Figura 3.3. Cada panel de la figura colecta proteínas que presentan un arreglo estructural de desorden determinado (ver en la Figura 3.3 paneles de la izquierda y la descripción de la figura para más información sobre las proteínas utilizadas) y que por lo tanto presentan perfiles de velocidad de evolución similares (Figura 3.3, panel izquierdo) así como también perfiles de RMSF similares (Figura 3.3, panel central). Cabe destacar que las proteínas representadas en cada panel no son homólogas entre sí, son proteínas funcional y secuencialmente alejadas que sin embargo poseen la misma organización estructural en cuando al patrón de orden-desorden.

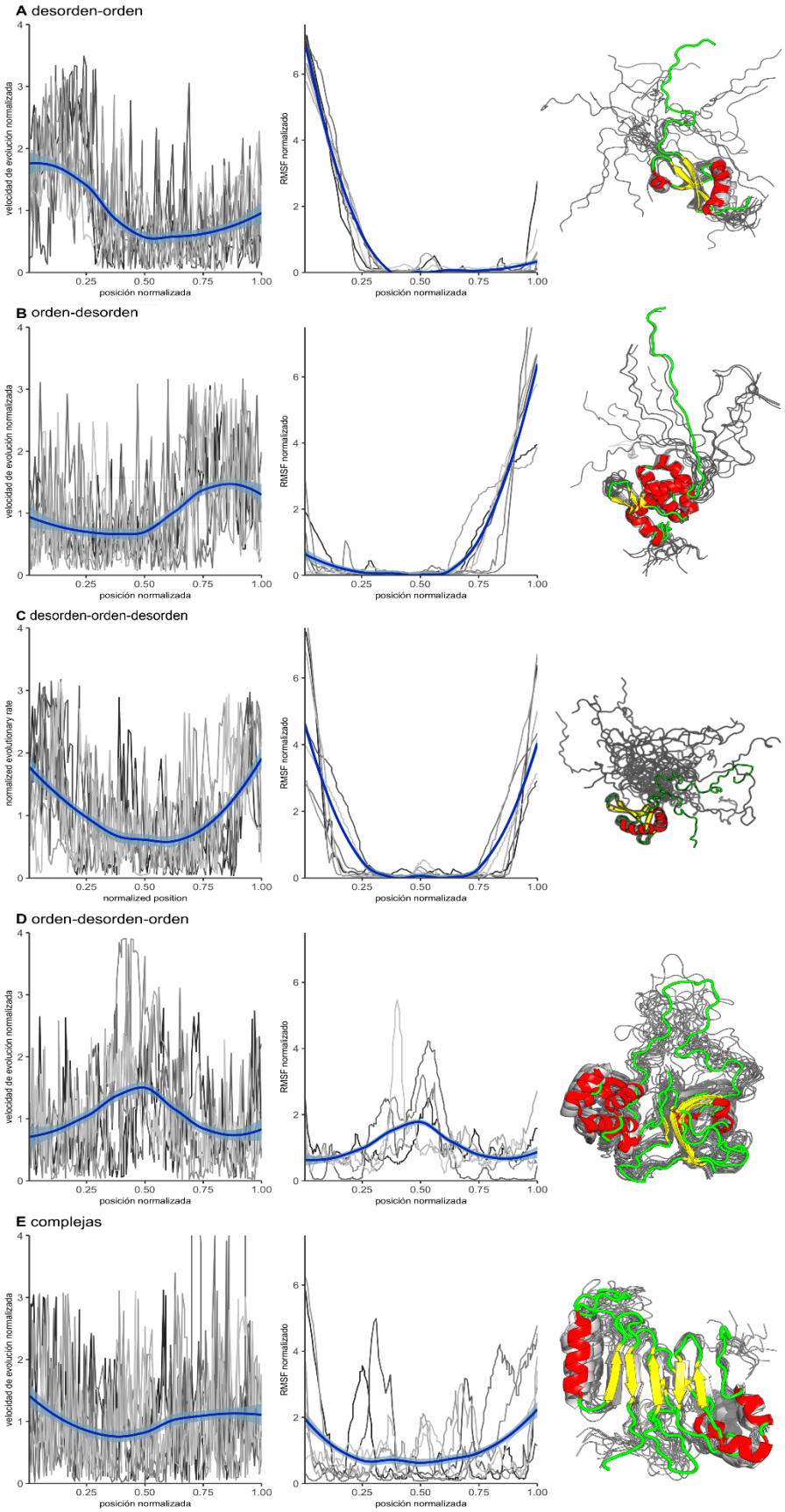

A desorden-orden
velocidad de evolución normalizada
RMSF normalizado
posición normalizada
B orden-desorden
velocidad de evolución normalizada
RMSF normalizado
posición normalizada
C desorden-orden-desorden
normalized evolutionary rate
RMSF normalizado
normalized position
posición normalizada
D orden-desorden-orden
velocidad de evolución normalizada
RMSF normalizado
posición normalizada
E complejas
velocidad de evolución normalizada
RMSF normalizado
posición normalizada

Figura 3.3 Perfiles de velocidad normalizada y de RMSF.

A la izquierda: perfiles de velocidad de evolución normalizada; en el centro perfiles del RMSF normalizados; a la derecha proteína representativa de cada uno de los arreglos estructurales encontrados. En azul las curvas obtenidas por regresión LOESS. Los perfiles en gris corresponden a las cadenas proteicas para cada grupo y debajo los códigos PDB. , Marcadas con asteriscos están las proteínas utilizadas para la representación esquemática mostrada a la derecha de cada panel. ***Panel A****: Proteínas que presentan un arreglo estructural del tipo desorden-orden (PDBids: 2cqrA, 2k3aA, 2khmA, 2myxA, 2rsmA, 2rsoA, *2rvlA).* ***Panel B****: Proteínas que presentan un arreglo estructural del tipo orden-desorden (PDBids: 1uffA, 1ug1A, 1jvrA, *2hmxA, 2k4kA, 2k5fA, 2ru8A).* ***Panel C****: Proteínas que presentan un arreglo estructural del tipo desorden-orden-desorden (PDBids: 1ixdA, *1u6fA, 1wj7A, 2dh7A, 2ecbA, 2ejmA, 2k8pA).* ***Panel D:*** *Proteínas que presentan un arreglo estructural del tipo orden-desorden-orden (PDBids: 1qu6A, 2adzA, 2lbcA, 2l3sA, *2mphA).* ***Panel E:*** *Proteínas que presentan un arreglo estructural complejo (PDBids: 1mm4A, 1sm7A, 1tteA, 1vzsA, *2aivA, 2afjA, 2n1rA).*

Si se observa el panel de la derecha de la Figura 3.3, se pueden ver representaciones esquemáticas de proteínas con determinados arreglos estructurales y vale la pena observar que las regiones desordenadas pueden encontrarse en distintas partes de las proteínas, pudiendo estar presentes en la parte central de la proteína o en los extremos (entiéndase por extremos amino o carboxilo terminal los correspondientes a la proteína o bien a los límites entre dominios, ya que un tercio de los polipéptidos presentes en el conjunto de datos, son dominios se encuentran presente en la estructura al menos a 30 residuos de los extremos de la proteína). Estos perfiles de velocidad de evolución pueden ser utilizados para distinguir como mínimo 5 tipos distintos de organizaciones estructurales en proteínas:

1) *Proteínas que contienen una región N-terminal desordenada seguida de una región globular (panel* ***A*** *de la* Figura 3.3*)*. Como representante de este grupo se tomó al dominio N-terminal de la heterocromatina 1 de murino (Shimojo et al., 2016) (código PDB: 2RVL). El extremo N-terminal de esta proteína es polianfolítico y su flexibilidad es regulada por la fosforilación de sus serinas, lo que altera la distribución del ensemble conformacional, desplazándolo hacia aquellos confórmeros donde hay unión con la histona H3. Si se observa el lado izquierdo del panel A, se puede ver que las velocidades de evolución son más altas en la

región de alta flexibilidad que en este caso se corresponde con el extremo amino terminal.

2) *Proteínas que contienen una región N-terminal ordenada seguida de una región desordenada (panel **B** de la* Figura 3.3*).* En esta caso se tomó como representante para la esquematización dominio alfa-hélice de la precursor de la poliproteína Gag del HIV-1 (código PDB: 2HMX). En esta poliproteína, la región desordenada une MA al dominio N- terminal de la proteína CA. Recientemente se sugirió que la contracción de ese linker desordenado, puede permitir la interacción entre la MA (inmadura) y el dominio CA, regulando la accesibilidad de la proteasa viral para la liberación de proteínas maduras (Samsudin et al., 2020). Tal como es el caso del grupo anteriormente comentado, estas proteínas presentan velocidades de evolución más altas en su extremo flexible que en este caso es el extremo carboxilo terminal.

3) *Proteínas que presentan una organización estructural con una alternancia desorden-orden-desorden (panel **C** de la* Figura 3.3*).* En este caso se tomó como representante a la proteína UBI-1 del parásito *Trypanosoma cruzi*. La proteína contiene un motivo de reconocimiento de RNA flanqueado por regiones N-terminal y C-terminal desordenadas (código PDB: 1U6F) (Volpon et al., 2005). La región ordenada es evidenciada por una disminución concomitante tanto en el perfil de velocidad como en el perfil de flexibilidad.

4) *Proteínas que presentan una organización estructural con una alternancia orden-desorden-orden (panel **D** de la* Figura 3.3*).* Como representante de este grupo se tomó a la inmunofilina nuclear FKBP25 (código PDB: 2MPH). Esta proteína presenta un dominio N-terminal HLH y dominio C-terminal de unión a FK506 que se encuentran unidos por un conector (*linker*) largo y flexible que permite el correcto reconocimiento del ADN por cualquiera de los dos dominios globulares (Prakash et al., 2016).

5) *Proteínas que presentan una organización estructural de desorden complejo (panel **E** de la* Figura 3.3*).* Estas proteínas se caracterizan por tener regiones desordenadas cortas, embebidas en dominios globulares o formando parte de regiones de transición orden-desorden. Como ejemplo de este tipo de proteínas, se muestra una representación de la proteína del poro nuclear de levaduras NUP116P (código PDB: 2AIV) (Robinson et al., 2005), y tal como se puede observar, esta proteína presenta regiones con estructura secundaria bien definida intercalada por loops flexibles.

3.3.2 Efectos de los contactos entre residuos en las velocidades de evolución.

Está bien establecido que la estructura de la proteína tiene un gran impacto en las velocidades de evolución (Echave et al., 2016). Las velocidades de evolución en posiciones ordenadas y desordenadas podrían estar condicionadas o podrían ser explicadas si se consideran el número de contactos entre residuos. El promedio de contactos por posición en posiciones ordenadas es de 3.44, mientras que para las posiciones desordenadas es de 1.75, es decir que en promedio las posiciones ordenadas tienen el doble de contactos físicos que las posiciones desordenadas. Como se puede observar en la Figura 3.2 las posiciones desordenadas de las IDPs presentan tasas de evolución más altas pero también algunas posiciones desordenadas pueden evolucionar tan lento como posiciones ordenadas. Entonces, ¿cómo es posible que a pesar de tener en promedio menor cantidad de contactos terciarios, las posiciones desordenadas puedan presentar tasas evolutivas comparables con las tasas evolutivas de posiciones ordenadas? Los ensembles de IDPs están formados por conformaciones que presentan una alta tasa de interconversión y que además son estructuralmente muy diferentes entre sí por lo que la cantidad de contactos entre residuos para una determinada posición puede cambiar sustancialmente entre distintas conformaciones (Figura 2.4). Entonces, fue necesario derivar alguna métrica estructural que diera cuenta de la información presente en todo el ensemble. Una de esas métricas es el promedio de contactos por posición a lo largo del ensemble que se calculó de la siguiente manera (Figura 3.4).

Posición	1	2	3	4	5
Confórmero 1	2	0	0	1	8
Confórmero 2	1	2	0	1	1
Confórmero 3	2	0	0	1	0
Confórmero 4	3	2	0	1	7
promedio de contactos	2	1	0	1	4

Figura 3.4 Esquema de cálculo del promedio de contactos a lo largo del ensemble.

El promedio de contactos se calcula simplemente promediando la cantidad de contactos observados para una determinada posición a lo largo del ensemble.

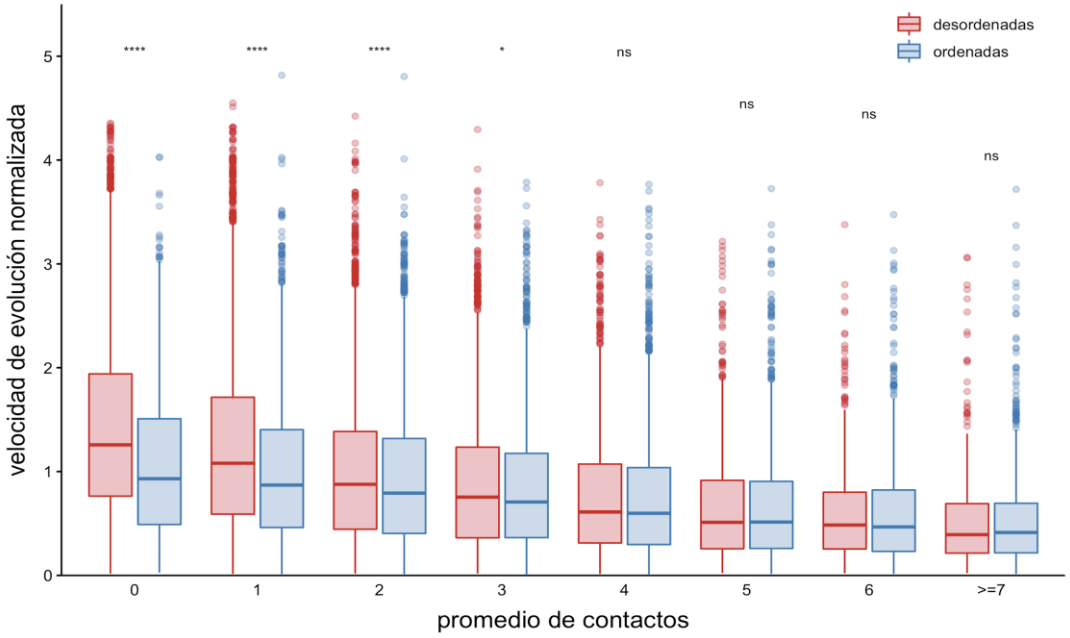


Figura 3.5 Velocidad de evolución normalizada en función del promedio de contactos.

*Los boxplots muestran la distribución de la velocidad normalizada en función del promedio de contactos por posición a lo largo de todo el ensemble. Para las comparaciones estadísticas se realizó un test de Wilcoxon, los asteriscos representan valores de distinta significancia: **** representan valores-p <= 0.0001; *** representan valores-p <= 0.001; ** representan valores-p <= 0.01; * representa un valor-p <= 0.05; y la expresión ns (no significativo) implica que se obtuvo un valor-p > 0.05.*

Si se considera el promedio de número de contactos por posición a lo largo de todo el ensemble, se puede observar una tendencia negativa entre la velocidad de evolución normalizada y el promedio de contactos, independientemente si esa posición es ordenada o desordenada (Figura 3.5). Es interesante notar asimismo que las posiciones ordenadas y las posiciones desordenadas tienden a presentar tasas de evolución similares cuando el número de contactos aumenta. Este mismo resultado se obtiene cuando se compara con el máximo, el mínimo y la moda en el número de contactos derivados de todo el ensemble (Figura 3.6).

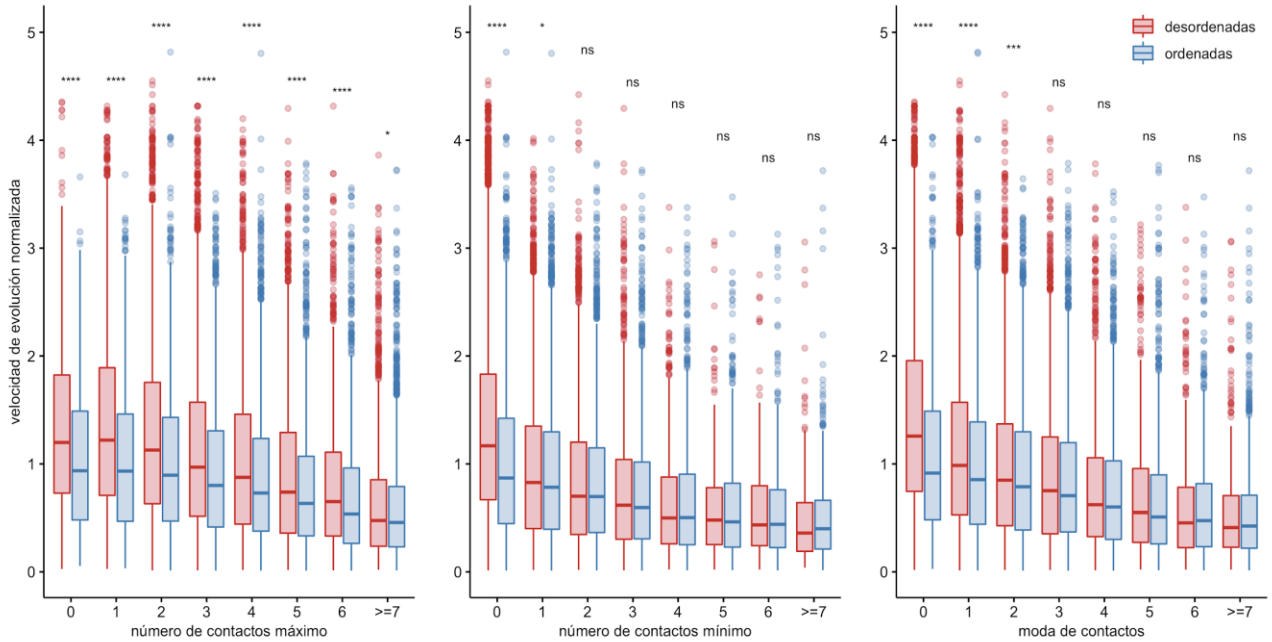


Figura 3.6 Velocidad de evolución normalizada en función de información de contactos.

*Los Boxplots muestran las distribuciones de las velocidades de evolución normalizadas en función de distintas métricas estructurales. Cada una de estas métricas fueron derivadas del ensemble como el máximo, el mínimo y la mode de contactos para una determinada posición a lo largo del ensemble. Para las comparaciones estadísticas se realizó un test de Wilcoxon, los asteriscos representan valores de distinta significancia: **** representan valores-p <= 0.0001; *** representan valores-p <= 0.001; ** representan valores-p <= 0.01; * representa un valor-p <= 0.05; y la expresión ns (no significativo) implica que se obtuvo un valor-p > 0.05.*

Otra métrica estructural utilizada para caracterizar el comportamiento dinámico de las IDPs, fue el parámetro que se denominó “robustez”[7] de contactos que contabilizaba la fracción de confórmeros con al menos un contacto por sitio. Entonces en el caso de una robustez o fracción de contactos del 50% significa que para esa posición la mitad de los confórmeros que formaban parte del ensemble estaban haciendo contacto. Para esta determinación no importaba si esa posición realizaba 4 contactos con otro residuo o 1 era contabilizada como una posición con al menos un contacto. El esquema del cálculo de dicha fracción se ejemplifica en la Figura 3.7.

Posición	1	2	3	4	5
Confórmero 1	0	0	0	1	0
Confórmero 2	0	1	0	1	1
Confórmero 3	1	0	0	1	1
Confórmero 4	0	1	0	1	1
robustez de contactos (%)	25	50	0	100	75

Figura 3.7 Esquema de cálculo de robustez de contacto para un ensemble.

Para el esquema se ejemplifica con un ensemble de 4 confórmeros. La robustez de contactos se muestra expresada en porcentaje, pero su expresión en fracción es análoga. Es decir, una robustez de 0.5 equivale a una robustez del 50% y viceversa. Si una posición no establece ningún contacto con otros residuos a esa posición se le asigna un 0. Si una posición establece un contacto o más con otra posición se le asigna un 1. La robustez de contactos se calcula contando para qué fracción del ensemble entero esa posición establece contacto. Por ejemplo, en el esquema la posición 1 presenta solamente contactos en el confórmero 3, entonces la fracción de contactos se calcula como 1 dividido el total de confórmeros en el ensemble (4), lo que arroja una robustez del 25%. Mientras que la posición 5 presenta contacto en el confórmero 2, 3 y el 4, dando lugar a una robustez del 75%.

[7] El concepto de robustez también puede encontrarse referido con la expresión fracción de contactos. En esta tesis robustez y fracción representan lo mismo.

.Además, La fracción de contactos puede ser interpretada como la robustez que tiene esa posición para mantener al menos un contacto a lo largo del ensemble (de allí el uso de la expresión robustez). Una posición que tiene contactos en el 70% de los confórmeros es más robusta que una posición que solamente la tiene en el 10% de las conformaciones pertenecientes al ensemble y por eso se definió a esa fracción como robustez de contacto. Como era de esperar las posiciones desordenadas tienen en promedio valores más bajos de robustez de contactos (~0.67) que las posiciones ordenadas (~0.92). Esto significa que las posiciones ordenadas tienden a mantener los contactos a lo largo de todo el ensemble (Figura 3.8).

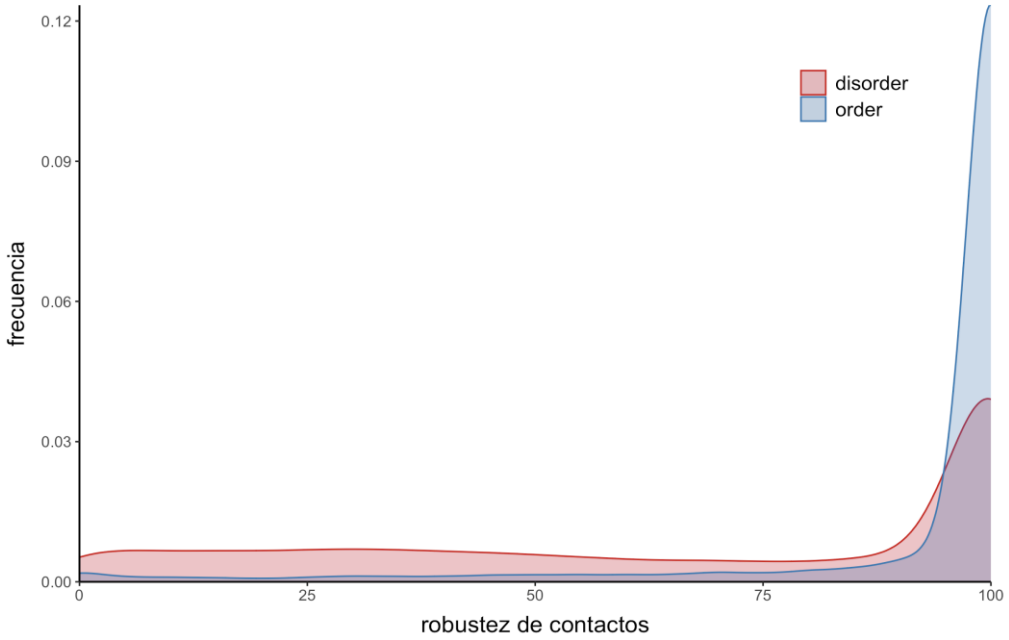


Figura 3.8 Distribución de la robustez de contactos.

Las posiciones desordenadas son mostradas en rojo y posiciones ordenadas en azul.

Que las posiciones ordenadas tengan en promedio mayor robustez de contacto que las posiciones desordenadas es de esperarse debido a su naturaleza y rol en el mantenimiento de una estructura tridimensional determinada, establecida principalmente por contactos terciarios entre residuos. Como se puede ver en la Figura 3.8 existen posiciones desordenadas que también presentan valores altos de robustez, alrededor

del 40% de las posiciones desordenadas presenta un valor de robustez de contactos del 100%. Entonces evidentemente el mantenimiento de contactos en determinadas posiciones resulta importante a pesar de que una posición esté predicha como desordenada, ya que esa predicción no implica necesariamente una pérdida de características estructurales. El desorden es un elemento estructural y tiene que ser tenido en cuenta para entender el rol biológico de las IDPs.

Entonces, ¿cuál es la relación entre la velocidad de evolución y la fracción o robustez de contactos para las posiciones desordenadas? En una primera aproximación la robustez de contactos puede entenderse como determinante si toma el valor 0 (ningún confórmero tiene contactos en esa posición) o si toma el valor 100 (todos los confórmeros mantienen al menos un contacto en esa posición). Si se observa la Figura 3.9, se puede concluir que la necesidad de mantener al menos un contacto en el 100% de las conformaciones restringe la velocidad de evolución para esa posición. El desorden es también un elemento estructural que modula la divergencia secuencial.

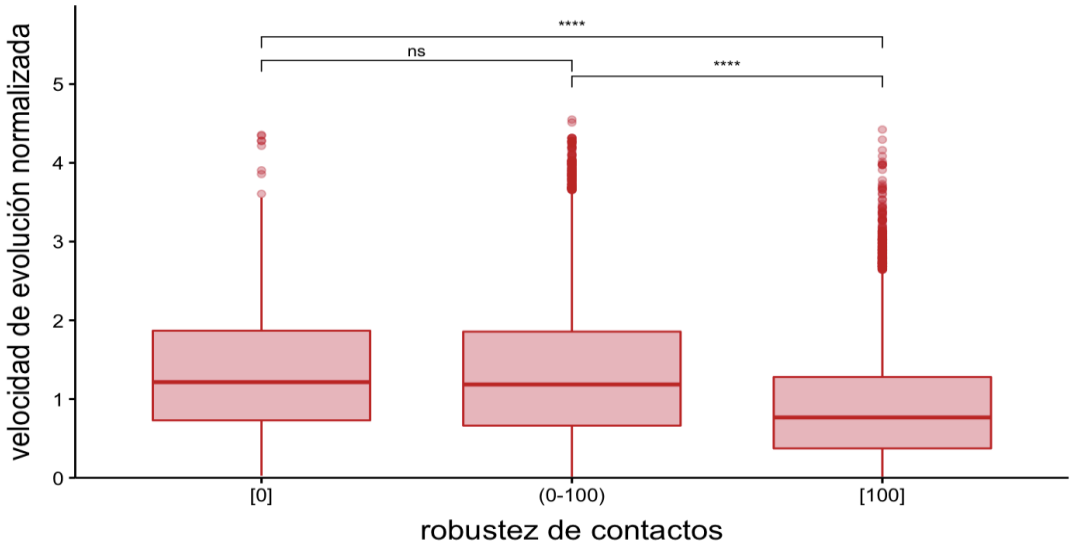


Figura 3.9 Distribución de las velocidades de evolución normalizadas en función de la robustez de contactos.

*En esta figura se resalta la relación entre las velocidades normalizadas de evolución para posiciones desordenadas con robustez nula (0 %) o completa (100 %) o algún valor dentro del intervalo mencionado (0-100). Para las comparaciones estadísticas se realizó un test de Wilcoxon, los asteriscos representan valores de distinta significancia: **** representan valores-p <= 0.0001; *** representan valores-p <= 0.001; ** representan valores-p <= 0.01; * representa un valor-p <= 0.05; y la expresión ns (no significativo) implica que se obtuvo un valor-p > 0.05.*

Si ahora se analiza el continuo de robustez en las posiciones desordenadas en función de las velocidades normalizadas de evolución puede observarse una tendencia similar que con el promedio de contactos: la velocidad de evolución se vuelve más lenta a medida que se incrementa la robustez de contactos ver Figura 3.10. Esto sugiere que el la presencia del desorden podría ayudar a aumentar las velocidades de evolución en esas posiciones.

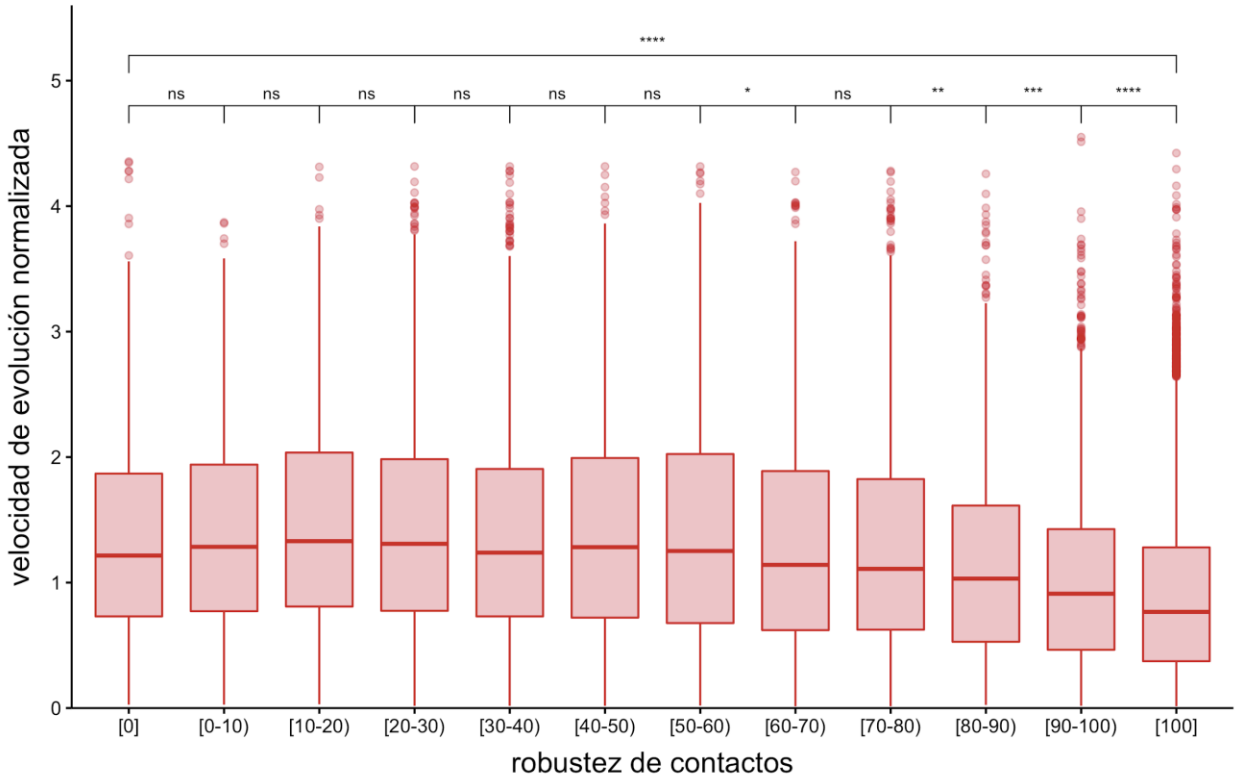


Figura 3.10 Velocidad de evolución normalizada en función de la robustez de contacto.

*En esta figura se resalta la relación entre las velocidades normalizadas de evolución para posiciones desordenadas y la robustez de contactos (expresada en porcentaje) clasificadas en intervalos más pequeños. Para las comparaciones estadísticas se realizó un test de Wilcoxon, los asteriscos representan valores de distinta significancia: **** representan valores-p <= 0.0001; *** representan valores-p <= 0.001; ** representan valores-p <= 0.01; * representa un valor-p <= 0.05; y la expresión ns (no significativo) implica que se obtuvo un valor-p > 0.05.*

Como ya se mencionó antes, las IDPs son altamente dinámicas entonces es posible pensar que otro de los factores que pueden llegar a modular la velocidad de evolución en residuos desordenados es su interacción con regiones desordenadas y regiones ordenadas. Para poder responder esto, se estudió cómo era la velocidad normalizada de evolución, el promedio de los contactos por sitio y los valores de RMSF normalizados en posiciones desordenadas en función a la distancia al residuo ordenado más cercano. La velocidad de evolución normalizada aumenta a medida que la distancia al residuo ordenado más cercano aumenta (Figura 3.11). Una tendencia similar es observada entre el RMSF normalizado, que es un indicador del grado de flexibilidad en esa posición (Figura 3.12).

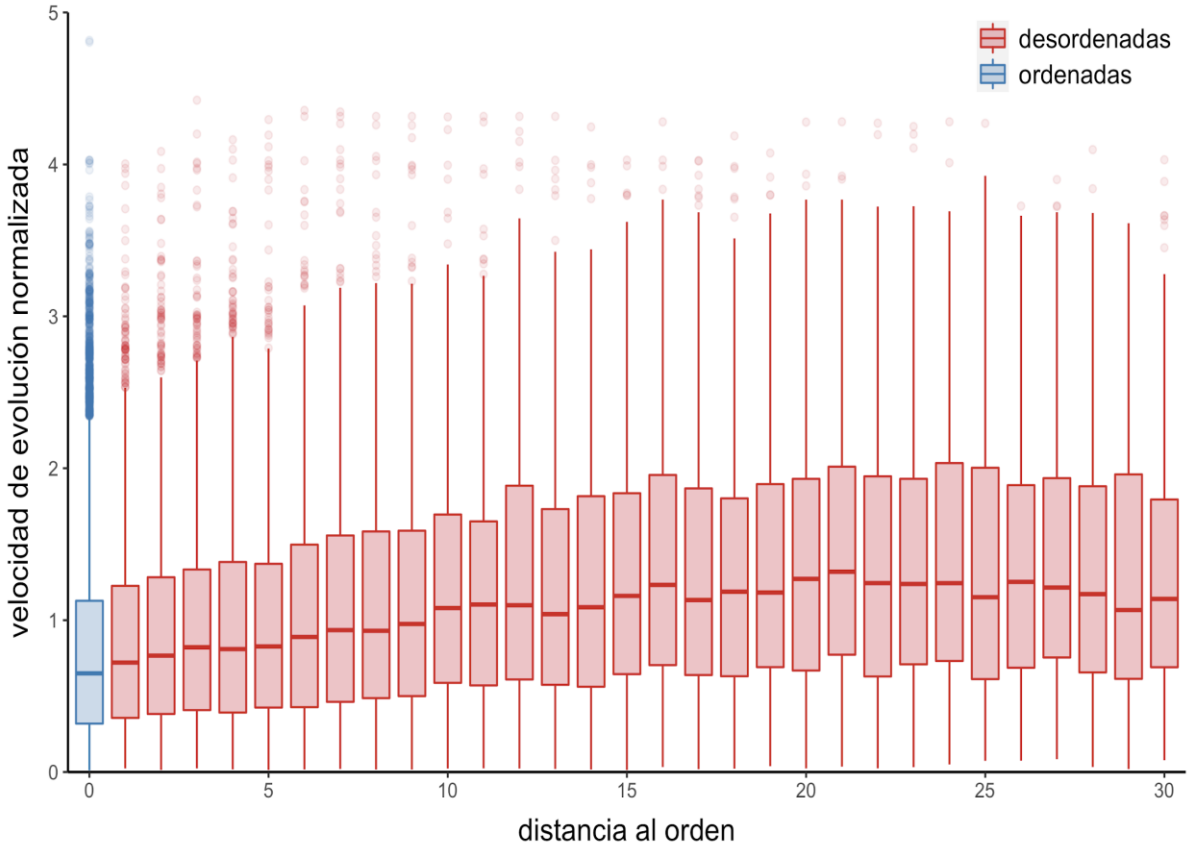


Figura 3.11 Distribución de las velocidades de evolución normalizadas en función de la distancia al residuo ordenado más cercano.

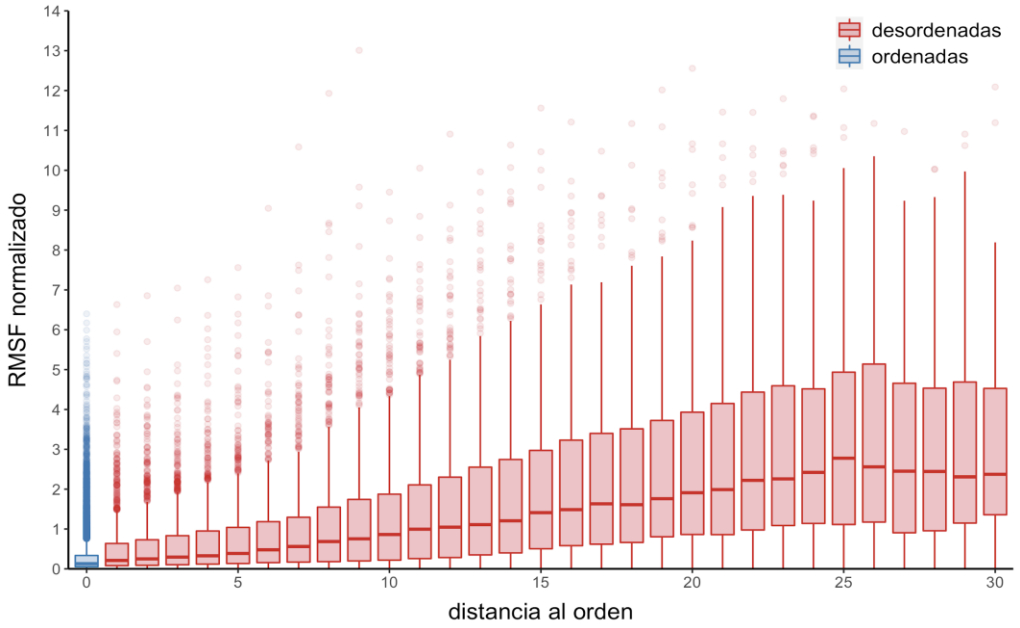


Figura 3.12 Distribución de RMSF normalizados en función de la distancia al residuo ordenado más cercano.

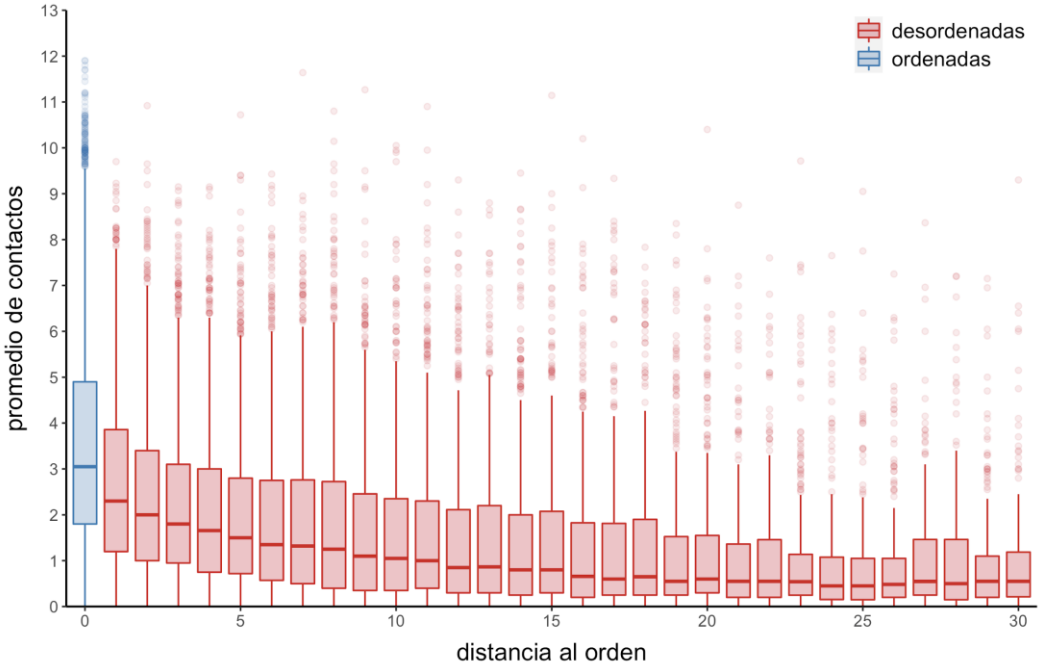


Figura 3.13 Distribución del promedio de contactos en función de la distancia al residuo ordenado más cercano.

Mientras que el promedio de contactos disminuye a medida que aumenta la distancia al orden (Figura 3.13), este último resultado a su vez está acompañado con un incremento en la proporción de contactos establecidos entre regiones desordenadas (Figura 3.14).

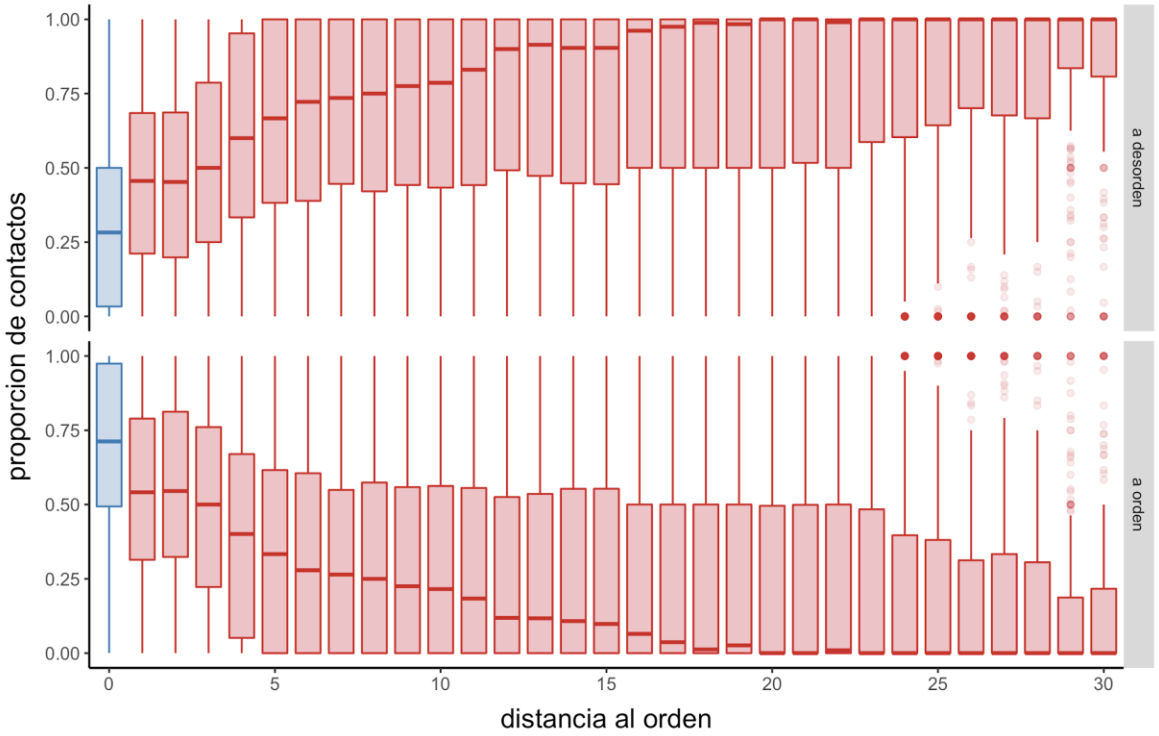


Figura 3.14 Distribución de la proporción de contactos.

Boxplots que muestran la distribución de la proporción de contactos establecidas con posiciones desordenadas (parte superior) o posiciones ordenadas (parte inferior) en función a la distancia secuencial con el residuo ordenado más cercano.

Esos últimos resultados muestran que el entorno estructural donde se encuentra una posición desordenada (establecida por su distancia secuencial a regiones ordenadas) tienen un gran impacto sobre las velocidades de evolución normaliza, el RMSF y el promedio de contactos para esa posición, influyendo incluso sobre la proporción de contactos hacia posiciones ordenadas. Con todo lo dicho anteriormente se puede concluir que las restricciones que imponen las regiones ordenadas en las velocidades de evolución para mantener una estructura tridimensional dada, también son impuestas por

el desorden en pos de mantener una estructura dinámica que es clave en el rol biológico de las IDPs.

Todos los resultados comentados en esta sección que involucran información derivada de contactos como métricas estructurales (promedio, robustez, moda, máximo y mínimo) también fueron calculados excluyendo a aquellos contactos que estuvieran establecidos entre residuos que se encontraran distanciados a menos de 4 residuos secuencialmente consecutivos. De esta manera se elimina la posibilidad que las restricciones impuestas estuvieran debidas no por el desorden, sino por la aparición de elementos de estructura secundaria transitoria debida al establecimiento de interacciones locales. Las tendencias observadas son las mismas a las informadas en esta sección, con el fin de no entorpecer la lectura del resto del capítulo, las figuras se pueden encontrar en el Apéndice A.

3.3.3 Relación entre el desorden interno y el desorden terminal y las velocidades de evolución

De las figuras (Figura 3.11, Figura 3.13) se puede deducir que las velocidades de evolución en posiciones desordenadas se encuentran condicionadas al entorno en el que se encuentran, esto es por ejemplo si están más cerca o alejadas de una región ordenada. De las curaciones visuales realizadas en el conjunto de datos fue posible entender que el desorden no se comporta de la misma manera si se encuentra en un extremo de la proteína (terminal) o si se encuentra entre regiones ordenadas (interno). En la Figura 3.15 se muestra un esquema de definición de desorden terminal y desorden interno.

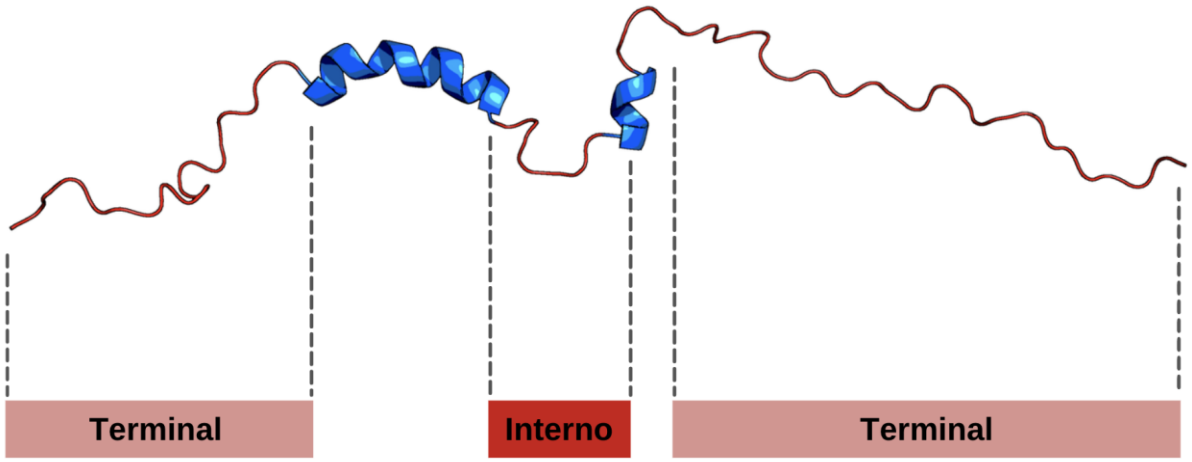


Figura 3.15 Esquema de definición de desorden terminal y de desorden interno.

Representación esquemática de la proteína (completar nombre) código PDB: 1KN7_A. Las regiones desordenadas que se encuentran entre regiones ordenadas son denominadas interna y las que se encuentran ya sea en el extremo N-terminal o en el extremo C-terminal.

Luego de definir estas regiones se graficaron las velocidades de evolución normalizada (Figura 3.16, panel superior) y el promedio de los contactos (Figura 3.16, panel inferior). en función de la distancia al residuo ordenado más cercano. Se puede observar que para posiciones que se encuentran en las regiones de desorden interno, a medida que aumenta su distancia al orden aumentan sus velocidades de evolución normalizada (Figura 3.16, panel superior) esto puede estar vinculado a la menor cantidad de contactos establecidos en estas regiones (Figura 3.16, panel inferior). Una posición de desorden interno que se encuentra a al menos 30 residuos de distancia de un residuo ordenado, se comporta como un loop terminal, esto es tiene un grado de flexibilidad muy alta casi sin interacción con el resto de la proteína. En cambio, para posiciones se encuentran en las regiones de desorden terminal si bien la cantidad de contactos promedio disminuye a medida que aumenta la distancia al orden, no se observa la misma tendencia para la velocidad normalizada de evolución. Una región desordenada terminal puede tener más libertad de movimiento para poder interactuar con regiones ordenadas que se encuentren muy distantes secuencialmente o bien con otras proteínas o ligandos.

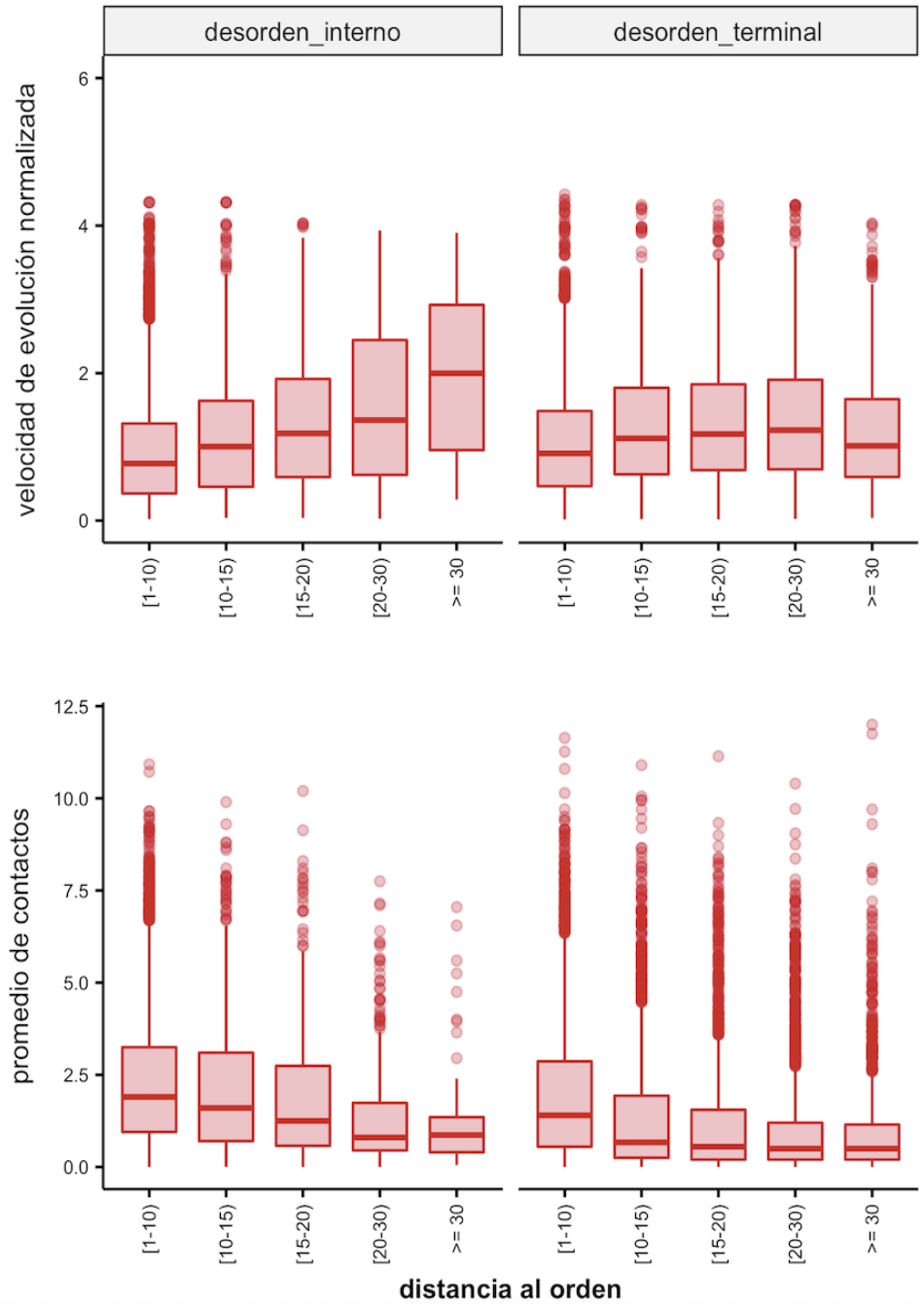


Figura 3.16 Velocidad de evolución normalizada y promedio de contactos para desorden para distintos tipos de desorden.

Boxplots que muestran la distribución de las velocidades normalizadas (superior) y el promedio de contactos (inferior) para posiciones que se encuentran en regiones de desorden interno (izquierda) y para posiciones que se encuentran en regiones de desorden terminal.

3.4 Discusión

En una primera aproximación, la heterogeneidad de las velocidades de evolución en las IDPs (ver Figura 3.2, Figura 3.3, Figura 3.11), puede ser explicada como función de los contactos terciarios por posición en los distintos confórmeros del ensemble (ver Figura 3.5, Figura 3.9, Figura 3.10, Figura 3.13). Estos resultados sugieren que el estado del arte y todo lo que se conoce hasta el momento en relación con las tasas evolutivas de proteínas globulares, puede extenderse a proteínas que carecen de una estructura estable y definida como las IDPs. Sin embargo, este tipo de proteínas como ya se mencionó poseen ciertas particularidades. La mayor diferencia entre las proteínas globulares y las IDPs es la enorme diversidad estructural que estas proteínas presentan entre los distintos confórmeros que forman parte del ensemble nativo. Este comportamiento particular implica una variabilidad muy marcada en el número de contactos por sitio en cada confórmero.

Las variaciones en la cantidad de contactos por posición debido a cambios conformacionales fueron extensamente ignorados en los estudios evolutivos (excepto para algunos ejemplos como citas (Sharir-Ivry and Xia, 2017; Zea et al., 2013)), y como se deduce de este capítulo son imposibles de negar en los ensembles de IDP. Estas variaciones imponen restricciones acumulativas en las velocidades de evolución (Saldaño et al., 2016) que podrían reflejar la importancia de las conformaciones transitorias o preferentes en el ensemble (Davey, 2019). Es posible que la heterogeneidad de la velocidad de evolución en regiones desordenadas, se deban parcialmente a la presencia de módulos funcionales involucrados en interacciones proteína-proteína, unión a ligando o modificaciones post-traduccionales. Entre estos, los motivos lineales cortos (**S**hort **Li**near **M**otivs, SLiMs por sus siglas en inglés) son típicos sitios de interacción encontrados en IDPs, usualmente estos motivos se encuentran más conservados que los residuos que se encuentran a su alrededor (van der Lee et al., 2014). Si bien no se detectaron sitios de unión cuando se buscaron SLiMs conocidos en

la base de datos ELM (Gouw et al., 2018) aun así, podrían existir conformaciones ‘predilectas’ que podrían llegar a contener estructuras secundarias residuales relacionadas con las interacciones entre proteínas.

Las IDPs presentes en este conjunto de datos tienen diferentes patrones estructurales con alternancia entre regiones ordenadas y desordenadas que son evidenciadas en los perfiles de velocidad. Dado que estos perfiles reflejan la preferencia por determinadas conformaciones, podrían ser usados para inferir y clasificar comportamientos globales de IDPs. A su vez se observó que existe una relación entre las velocidades de evolución y la cercanía con regiones ordenadas, dando cuenta de los efectos moduladores que tienen los dominios plegados o estructurados en las velocidades de evolución de regiones desordenadas adyacentes, dando lugar a la pregunta sobre la ocurrencia de procesos de coevolución (Mittal et al., 2018) en IDPs.

Estos resultados sugieren que los estudios estructurales con base evolutiva sobre IDPs, no sólo son posibles sino que también son una aproximación que promete revelar importante información sobre la evolución y la biología de proteínas intrínsecamente desordenadas.

3.5 Métodos específicos del capítulo

3.5.1 Construcción del conjunto de datos

El conjunto de datos utilizados se construyó a partir de IDPs que estuvieran depositadas en la PDB (Rose et al., 2015) y que por lo tanto su estructura fuera conocida y obtenida mediante técnicas de NMR. Para tal fin, se descargaron todos los archivos de coordenadas estructurales obtenidos por NMR y sobre todo ese conjunto se predijo el porcentaje de desorden con el método ESpritz (Walsh et al., 2012) y se clasificaron como IDPs aquellas proteínas que tuvieran al menos un 40% de sus residuos predichos ser desordenados. Luego se obtuvieron alineamientos múltiples de proteínas homólogas

presentes en la base de datos HSSP (Dodge et al., 1998), filtrando y seleccionando aquellos MSA que tuvieran sólo un %5 de gaps en promedio por posición y con porcentajes de identidad entre el 33 y el 100% contra la secuencia de referencia de PDB y que tuvieran un cobertura de al menos el 60% con esta última (median identity percentage of 64%). Los alineamientos utilizados tenían en promedio 96 secuencias y un promedio de porcentaje de gaps en las posiciones de 7.38%. Para asegurarse que en el conjunto de datos consistía en proteínas que fueron menos del 50% idénticas se utilizó el programa CD-HIT (Fu et al., 2012).Todas las proteínas que cumplieran con las condiciones antes mencionadas fueron posteriormente inspeccionadas visualmente y se seleccionaron aquellas que tuvieran un contenido de desorden evidente en la estructura 3D. Este paso es importante porque los predictores pueden (como todo método) tener errores en la asignación de desorden. Además, se eliminaron proteínas quiméricas.

El conjunto de datos final y curado consistió en 310 IDPs, con una longitud promedio de aproximadamente 130 residuos (mínimo = 58, máximo = 37) y número promedio de confórmeros por ensemble de aproximadamente 20 (mínimo = 6, máximo = 60). En este conjunto de datos las regiones desordenadas que consisten de más de 5 residuos tienen una longitud promedio de 23.6 aminoácidos, con un máximo de 263 aminoácidos. Algunas de las IDPs utilizadas para el estudio se pueden observar en la Figura 3.17.

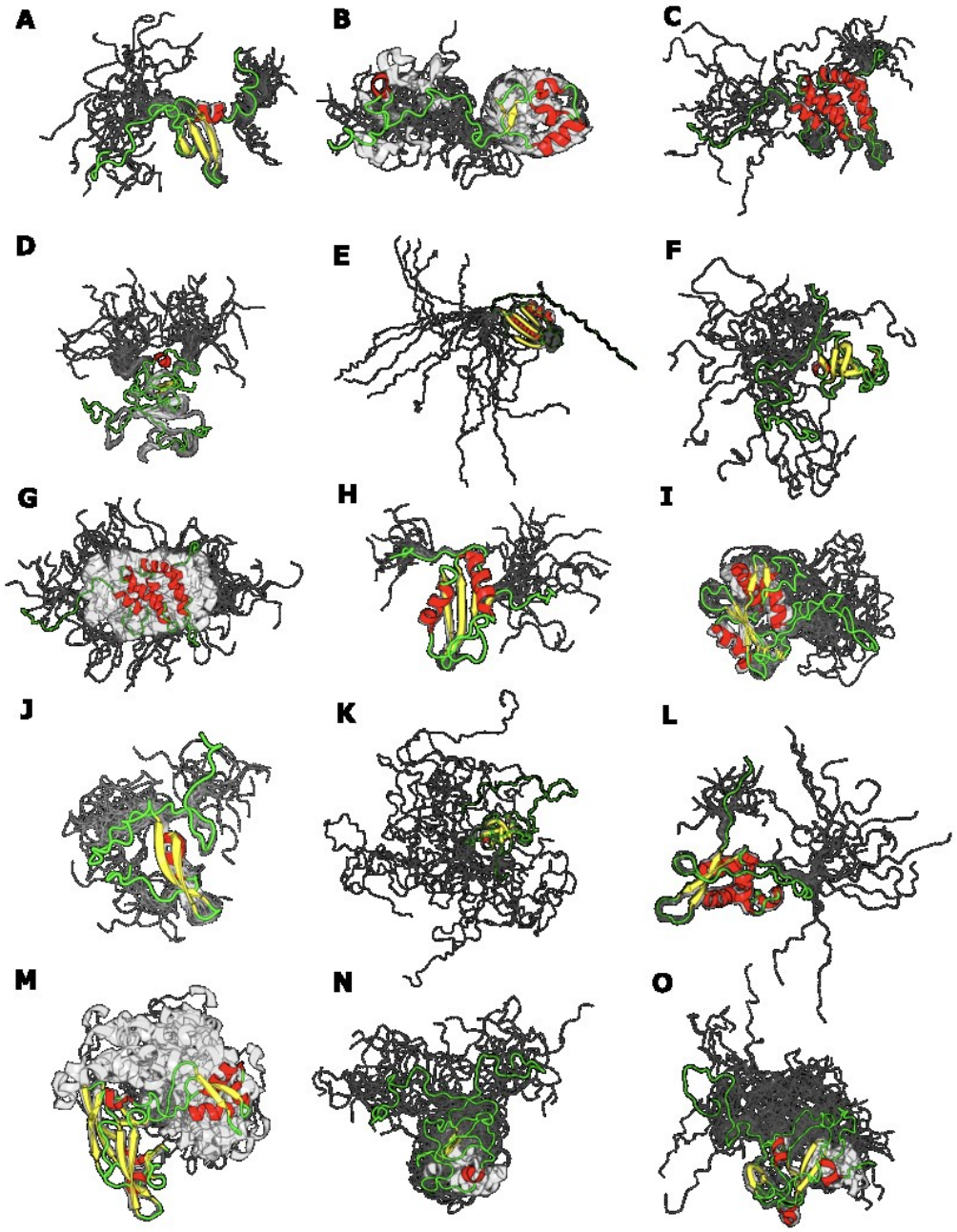


Figura 3.17 Representación de IDPs presentes en el conjunto de datos.

Códigos PDB: a)1AP0_A, b)1WWK_A, c)1INZ_A, d)1IXD_A, e)2LXF_A, f)2COY_A, g)1EJ5_A, h)2DGQ_A, i)2BBU_A, j)2JZ6_A, k)2CP5_A, l)2KPM_A, m)2L9Y_A, N)1JFN_A, o)2N1I_A.

3.5.2 Estimación de las velocidades de evolución

Para el cálculo de las velocidades de evolución se utilizó el programa Rate4Site (Pupko et al., 2002) con los alineamientos descritos anteriormente. Este programa estima las velocidades de evolución sitio-específicas mediante métodos Bayesianos bajo la hipótesis de independencia entre los sitios, reconstruye un árbol filogénetico para las secuencias del MSA mediante un algoritmo de neighbor-joining y distancias estimadas por máxima verosimilitud. En este trabajo se utilizó el modelo de JTT (D T Jones et al., 1992). Para cada proteína, los valores de velocidad de evolución sitio-específicos sin normalizar (datos crudos) del Rate4Site, se normalizaron dividiéndolos por el valor promedio de velocidad de la proteína entera tal como se indica en (Sydykova and Wilke, 2017). El test de Wilcoxon fue utilizado para evaluar las diferencias significativas en las medias de las velocidades de evolución normalizadas entre los subconjuntos de posiciones ordenadas y desordenadas.

3.5.3 Información estructural en confórmeros y ensembles

Todos los resultados comentados en esta sección que involucran información derivada de contactos como métricas estructurales (promedio, robustez, moda, máximo y mínimo) fueron derivados para cada confórmero de cada ensemble nativo de IDPs presente en el conjunto de datos y la información de contactos fue obtenida como el valor absoluto de los contactos terciarios por posición (definido como la distancia mínima entre las esferas de Van der Waals, para cualesquiera dos átomos pesados de las cadenas laterales de los aminoácidos, utilizando un valor de corte de 1.0 Å (Berrera et al., 2003). Estas métricas fueron estimadas de distintas maneras: el promedio, la moda, el máximo y el mínimo número de contactos por posición fueron calculados sobre una misma posición a lo largo de todo el ensamble. La fracción o robustez de contactos por posición fue definida y calculada como la fracción de confórmeros que tienen al menos un contacto

para esa posición. Por ejemplo, un valor de robustez 0.5 (o del 50%) significa que la mitad de los confórmeros observados tiene al menos un contacto en esa posición (Figura 3.7). Para cada posición desordenada, la distancia a un residuo ordenado fue calculada como la mínima distancia consecutiva al residuo ordenado más cercano. La flexibilidad fue estimada usando el RMSF de los **Cα** , calculados según MIToS (Zea et al., 2017). Los valores de RMSF fueron normalizados, dividiendo los valores RMSF de cada posición, por el promedio de los RMSF del ensamble. El mismo paquete fue utilizado para calcular los valores de los valores de RMSD de toda la proteína.

3.5.4 Obtención de los perfiles de velocidad y patrones conformacionales

Para cada proteína en el conjunto de datos se estimó su velocidad de evolución y los perfiles de RMSF por posición y se derivaron las respectivas medidas normalizadas, tal como explicadas en la sección 3.3.1. Como las proteínas tienen longitudes distintas, para poder hacer comparaciones válidas entre ellas, las posiciones fueron renumeradas al rango [0,1], dividiendo cada posición por la longitud total de la proteína. De esta manera, para cada proteína fue posible obtener patrones individuales como los mostrados en la Figura 3.18. Se corroboraron visualmente los arreglos de orden/desorden en los elementos estructurales de cada proteína para poder definir sus patrones estructurales y los perfiles individuales fueron luego agrupados y acumulados en función de las organizaciones estructurales de desorden antes mencionadas dan lugar a los perfiles mostrados en la Figura 3.3. Se realizó una regresión LOESS suavizada para poder modelar el comportamiento y la variabilidad de la información (línea azul de Figura 3.3).

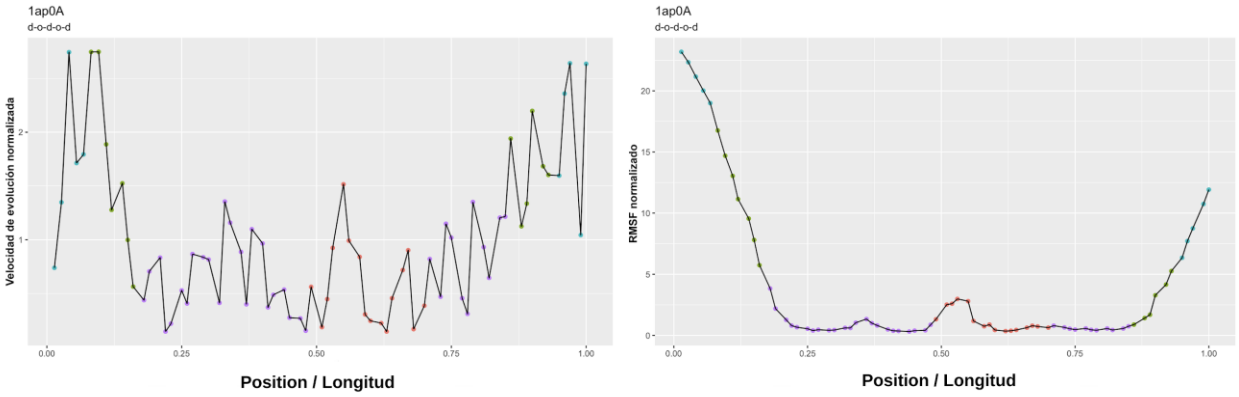


Figura 3.18 Perfiles individuales de velocidad de evolución normalizada y RMSF normalizado

Se muestra un perfil individual para la velocidad normalizada (izquierda) y un perfil de RMSF normalizado (derecha). En el eje x se grafica la posición normalizada de la longitud para poder acumular los perfiles de proteínas distintas. Este perfil corresponde al dominio de enlace de cromatina (Cromo) de la proteína modificadora 1 de ratón (código PDB: 1AP0_A). Esta proteína tiene 2 regiones ordenadas flanqueadas por regiones desordenadas, tal como queda evidenciado en el perfil de velocidad (izquierda) y el de RMSF (derecha)

3.6 Disponibilidad de datos y código.

El código y todas las tablas utilizadas para este capítulo puede encontrarse en:

https://gitlab.com/sbgunq/publications/palopoli-marchetti-2020-rates

4 Estudio y evaluación de la información estructural en ensembles conformacionales

Los resultados comentados en los capítulos 2 y 3 dejan en claro que no es posible entender a las IDPs sin tener en cuenta sus características dinámicas dada por la complejidad de sus ensambles nativos. Como ya se mencionó, los ensambles nativos de IDPs son un conjunto de numerosas conformaciones y algunos ensembles pueden estar descritos por 10 conformaciones, otros por 20 y otros por más de 60. La pregunta que es necesario realizar es si todas estas conformaciones tienen el mismo impacto en la función biológica de las proteínas, es decir: ¿los ensembles de IDPs son redundantes en cuanto a la información estructural que poseen?

En esta sección se muestran resultados que intentan responder esta pregunta utilizando las metodologías y conjuntos de datos descritos en el capítulo 2 y 3. Para mayor facilidad de interpretación de estos resultados cada uno tendrá su respectiva sección en este capítulo.

4.1 Evaluación de la información estructural en ensembles mediante el modelo de evolución SCPE.

A partir del conjunto de datos y la metodología descrita en el capítulo 2, se estudió si los sitios evolucionando bajo un régimen de condicionamiento estructural (sitios SC) provenía de todos los confórmeros analizados en el ensemble o si, dicha información evolutiva provenía de unos pocos confórmeros que por su importancia biológica y sus propiedades estructurales impactaban más en la divergencia de las secuencias. ¿Cuántos confórmeros son requeridos para describir de manera completa las restricciones estructurales observados en los MSA de todo un ensemble?

4.1.1 Resultados

Para responder las inquietudes antes planteadas, se calculó cuál es la cantidad mínima de confórmeros que son necesarios para poder alcanzar el porcentaje de sitios SC acumulados a largo de todo el ensemble y se encontró que en promedio esa cantidad corresponde a aproximadamente 11 confórmeros (Figura 4.1) para proteínas desordenadas.

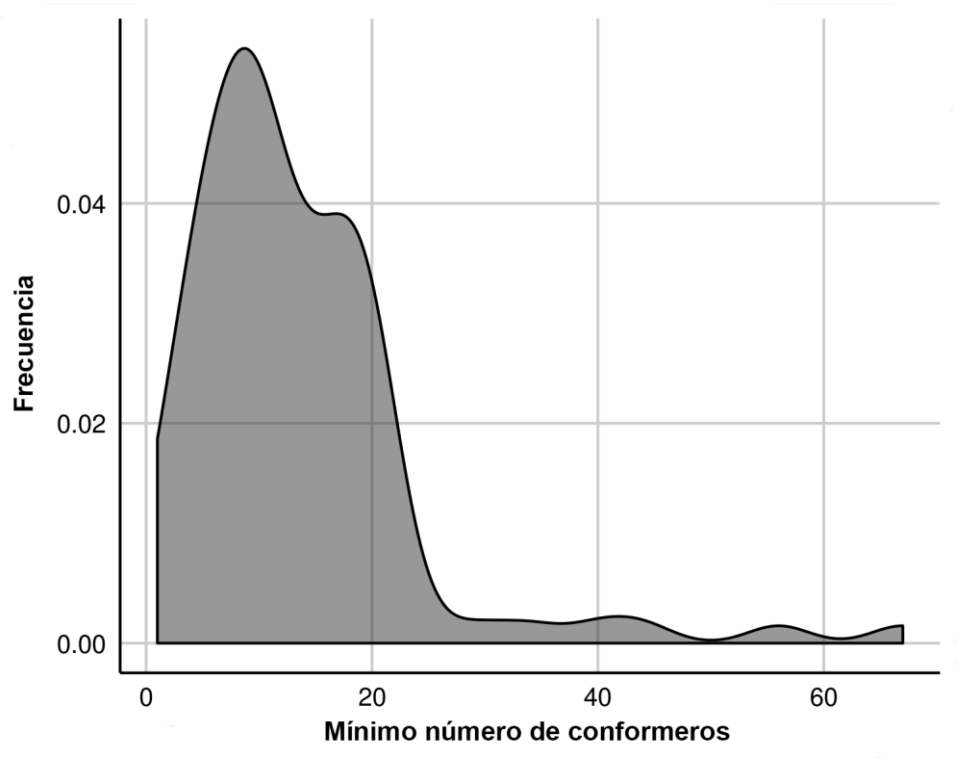


Figura 4.1 Distribución de mínimo número de confórmeros.

Se muestra la distribución del mínimo número de confórmeros que son necesarios para alcanzar el porcentaje de sitios SC acumulado a lo largo de todo el ensemble para cada uno de los 93 ensembles utilizados para el conjunto de datos de proteínas desordenadas. Esta distribución presenta un mínimo de un confórmero, un promedio de 11 y un máximo de alrededor 64.

Para las proteínas ordenadas ese valor es de aproximadamente 1.5 (no mostrado en la figura) y es consistente con la evidencia experimental disponible ya que la mayor parte de las proteínas ordenadas tiene baja diversidad conformacional y por ello son llamadas

rígidas (Monzon et al., 2017a) o poseen una cantidad pequeña de confórmeros (2 mayormente) relacionados a los estados libre o unido a ligando (Gutteridge and Thornton, 2005) (Amemiya et al., 2012). Sin embargo, estimar un número mínimo de conformaciones para ensembles de proteínas desordenadas no es una tarea fácil o sencilla debido a la alta complejidad que presentan estos ensembles.

El método propuesto en este trabajo, basado no solo en la cuantificación de los contactos físicos entre residuos sino en la implicancia que estos tienen en la restricción estructural impuesta durante el curso de la evolución, ofrece una buena aproximación en la determinación de la cantidad de conformaciones distintas que son necesarias para explicar las restricciones estructurales originadas debido a la presión de selección en dicho ensemble. Dado que en promedio los ensembles de IDPs están formados por 20 conformaciones, los resultados de este trabajo muestran que estos ensembles son altamente redundantes.

4.1.2 Metodología

Para poder determinar cuál era la mínima cantidad de confórmeros necesaria para explicar la restricción evolutiva observada en todo el ensemble, para cada proteína y su correspondiente ensemble del conjunto de datos se tomaron combinaciones usando la siguiente fórmula:

Ecuación 12: Fórmula de obtención de combinaciones.

$$\frac{n!}{k!\,(n-k)!}$$

Donde k es el número de conformaciones combinados y n es el número total de confórmeros presentes en dicho ensemble y en cada una de esas combinaciones se

acumularon los sitios SC de cada confórmero. Luego se registró para cada ensemble cuál era la mínima cantidad de confórmeros que combinados reportaban la restricción presente en todo el ensemble.

4.1.3 Discusión

Se encontró que el número de confórmeros en ensembles de IDPs para alcanzar el nivel total de restricciones estructurales observado es alrededor de 11 (Figura 4.1) de los confórmeros presentes en los ensembles (~20). Esto significa que los ensembles de IDPs son redundantes y que las conformaciones que son biológicamente importantes son pocas en relación a lo esperado dada su alta flexibilidad. Estos resultados concuerdan con la idea que distintos confórmeros del ensemble pueden estar involucrados de manera directa con la función de la proteína, pero también pueden ser importantes como mínimos locales en la inter conversión entre distintas conformaciones que son biológicamente relevantes (Turoverov et al., 2010).
Los resultados comentados en esta sección remarcan la importancia de los análisis evolutivos a la hora de determinar cuáles son las conformaciones relevantes en ensembles de alta complejidad como los de IDPs.

4.2 Evaluación de la información estructural en ensembles mediante velocidades de evolución.

Es necesario recordar que en el capítulo 3 se estudió cómo eran las velocidades de evolución en proteínas desordenadas en función de alguna métrica estructural derivada de la información de contactos. Utilizando el conjunto de datos y la metodología descritos en el capítulo 3 se buscó evaluar el grado de información estructural presente en ensembles de IDPs.

4.2.1 Resultados

Para poder explorar con mayor detalle cómo es la relación entre los ensembles conformacionales de IDPs y las velocidades de evolución, se determinó si las velocidades de evolución sitio-específicas correlacionaban mejor con los contactos para cualquier confórmero individual o con la información de contactos promediada a lo largo de todo el ensemble. Para cada proteína se calculó la correlación de Pearson (derivando el coeficiente de correlación rho (**ρ**)) entre la velocidad de evolución sitio-específica para una posición y una métrica estructural, en particular el promedio de contactos y la robustez. Luego, se comparó con cuál de estos dos descriptores las velocidades de evolución obtenían una mejor correlación.

Para el 75% de las proteínas del conjunto de datos utilizado, la correlación más fuerte se obtuvo cuando se tuvo en cuenta la información estructural de una sola conformación (en promedio **ρ** = -0.3692). Podría esperarse que las restricciones funcionales relevantes a la evolución estuvieran impuestas por aquellas conformaciones que tienen el mayor número de contactos (que se asume que es la conformación más cercana a una conformación cerrada de la proteína) o aquellas conformaciones que poseen el mínimo número de contactos (posiblemente la forma abierta de la proteína). Sin embargo, nuestros resultados indican que las conformaciones cerradas y abiertas no muestran los mejores coeficientes de correlación. En el 25% restante de las proteínas, la mejor correlación se obtuvo con parámetros que dan cuenta de la distribución de contactos a lo largo del ensemble completo: el 21% (promedio **ρ** = -0.5448) de las proteínas la mejor correlación se obtuvo con la robustez de contactos mientras que para el ~3% la mejor correlación se obtuvo con el promedio de contactos (promedio **ρ** = -0.4683). Estos resultados indican que para un cuarto de las proteínas del conjunto de datos la consideración de la información estructural presente en todo el ensemble permite una mejor explicación sobre los cambios evolutivos observados en el conjunto de proteínas homólogas.

Ahora bien, las correlaciones hechas hasta acá fueron hechas considerando una sola conformación o bien considerando todo el ensemble, la siguiente pregunta para realizar es: ¿existen combinaciones entre 1 y N (donde N es el número de conformaciones totales presentes en el ensemble) que tenga información estructural derivada de contactos que expliquen mejor los perfiles de velocidad observados? En la sección 4.1 se mencionó que los ensembles de IDPs son redundantes en término de la restricción estructural impuesta en la evolución, debido a que solamente se necesitan alrededor de 10 confórmeros en promedio para explicar la restricción estructural presente en alineamiento de proteínas homólogas de IDPs. Tomando esta idea como directriz, se exploraron todas las combinaciones posibles entre los distintos confórmeros para cada ensemble del conjunto de datos, para estudiar si existe un sub-conjunto de conformaciones en particular que den lugar a mejores valores de correlación entre alguna métrica y las velocidades de evolución normalizadas. Para poder contestar esa pregunta, se obtuvieron todas las combinaciones posibles entre cualquier confórmero en cada ensemble del conjunto de datos. Lo que se combinaba era la información estructural para cada residuo (ya sea el promedio de contactos o la robustez) de los confórmeros que fueron seleccionados. Luego se calculó el coeficiente de correlación de Pearson (rho $\boldsymbol{\rho}$) entre las velocidades de evolución y el promedio de los contactos (referido de aquí en adelante como $\boldsymbol{\rho_{ave}}$) y el coeficiente para la correlación entre las velocidades de evolución y la robustez (referido de aquí en adelante como $\boldsymbol{\rho_{crob}}$). Se calculó la significancia estadística para cada combinación utilizando un test de student (t-test) que determina si la distribución de los coeficientes $\boldsymbol{\rho}$ es diferente a la distribución de coeficientes generados por distribuciones al azar (que se realizaron mediante un bootstrapping (para más detalles, ver los métodos específicos de esta sección).

Se encontró que en ~52% de las proteínas, las velocidades de evolución correlación mejor con la fracción de contactos de los confórmeros combinados (promedio de los coeficientes $\boldsymbol{\rho_{crob}}$ = -0.5254), que con cualquier otra información disponible para confórmeros individuales o parámetros de información estructural derivados de todo el ensemble. En el mismo sentido, para el ~30% de los casos, las correlaciones entre el número promedio de los contactos para una determinada combinación (promedio de los coeficientes $\boldsymbol{\rho_{ave}}$ = -0.4154), dio mejores valores de correlación que el resto de los

parámetros, ya sea para un confórmero individual o información promediada a lo largo de todo el ensemble mencionados anteriormente. Solo para el ~17%, de los ensambles conformacionales estudiados la mejor correlación entre las velocidades de evolución y la información de contactos estaba dada para un confórmero individual, arrojando los valores más bajos de correlación obtenidos promedio ρ_{indiv} *= -0.3695. Un esquema que resume los valores de correlación obtenidos puede verse en la*

Tabla 1.

	Promedio de correlación de todo el ensemble o de confórmeros individuales		Promedio de correlaciones obtenidas de la combinación de un subconjunto de confórmeros	
	Correlación Individual	Todo el ensemble	Correlación Individual	Combinación de un subconjunto dado
Promedio	-0.3692 (**75%**)	-0.4683 (**21%**)	-0.3695 (**17%**)	-0.4154 (**30%**)
Robustez		-0.5448 (**23%**)		-0.5254 (**52%**)

Tabla 1: Valores de los promedios de correlación obtenidos para confórmeros individuales o para un subconjunto de confórmeros.

Entre paréntesis se muestran los porcentajes de proteínas del conjunto de datos para las cuales esa condición daba la mejor correlación

Tanto para la fracción o robustez de contactos, como para el promedio, las correlación pueden ser maximizadas si se tiene en cuenta la información contenida en un subconjunto de conformaciones presentes en el ensemble. Estadísticamente se esperaría que a medida que el espacio combinatorio aumenta, es más probable que la mejor correlación aparezca por azar (por ejemplo, la mejor correlación debería encontrarse cuando se combinan 10 confórmeros en ensembles con 20 conformaciones distintas). Sin embargo, la mejor correlación entre la fracción o robustez de contactos y las velocidades de evolución se obtuvo cuando se combinaban en promedio 4 confórmeros, mientras que la mejor correlación entre el promedio del número de contactos y las velocidades de evolución se obtuvo cuando se combinaban en promedio

2.8 confórmeros y nunca se obtuvieron mejores valores de correlación cuando se combinaban más de 8 conformaciones (Figura 4.2).

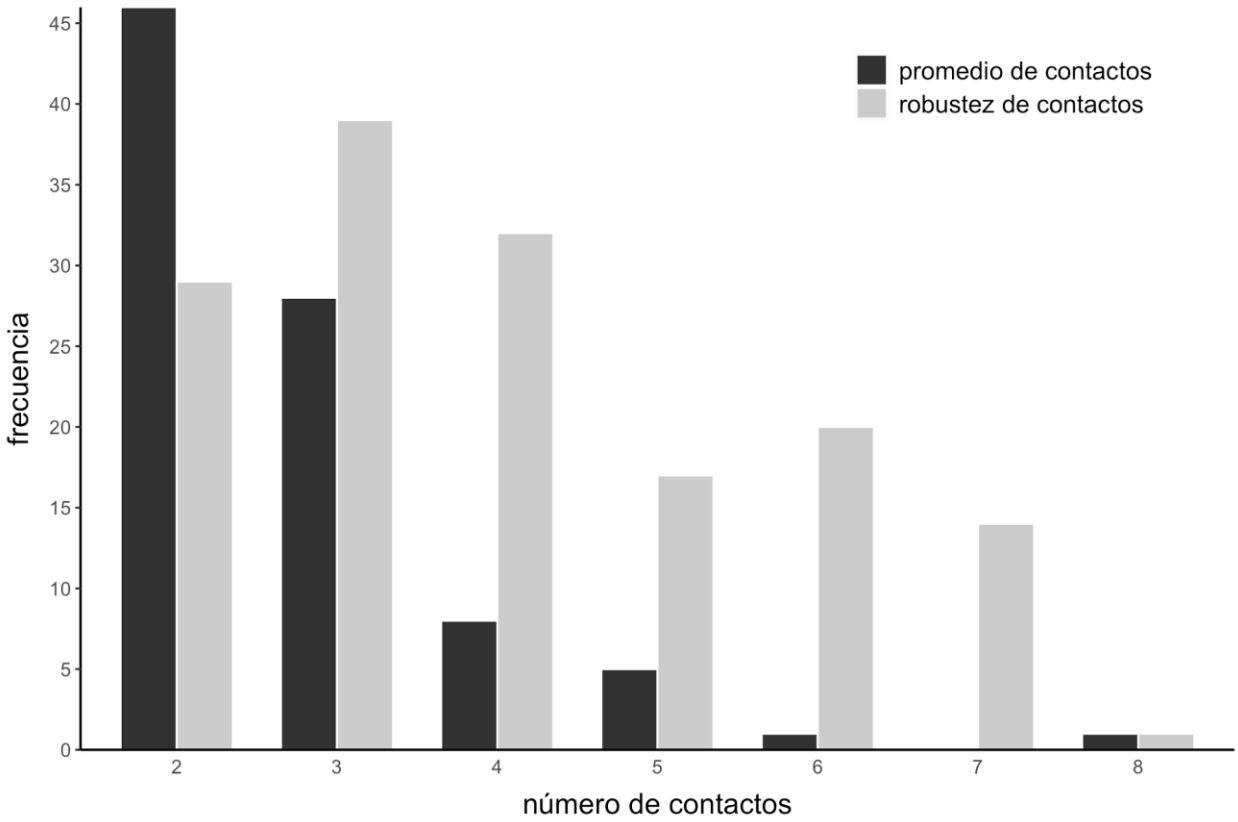


Figura 4.2 Conformaciones necesarias para la obtención de la mejor correlación.

Histograma que muestra el número de diferentes confórmeros necesarios para obtener la mejor combinación entre las velocidades de evolución normalizadas y la robustez (gris) y el promedio de número de contactos (negro).

Todos estos resultados sugieren que la información estructural proveniente de la combinación de determinados confórmeros representando buenos candidatos para explorar la diversidad conformacional de proteínas, así como su importancia biológica en concordancia con lo expresado en la sección 4.1. Estos resultados remarcan la importancia y el peso que tiene el ensemble conformacional sobre las velocidades de evolución ya que la correlación mejoró notablemente cuando se tuvo en cuenta diferentes descriptores de los contactos en el ensemble conformacional y cuando se obtuvieron

combinaciones de confórmeros. Dichas mejoras alcanzaron un 83% de las proteínas estudiadas.

La siguiente pregunta a contestar es ¿por qué para algunas proteínas la mejor correlación se obtiene con la robustez de contactos, mientras que para otras se obtiene con el promedio de los contactos? Es importante tener en cuenta que la robustez de contactos puede ser una buena métrica estructural para representar a IDPs que tienen una variación estructural considerable entre los confórmeros, de ahí que el número de contactos varíe notablemente entre confórmeros. Por otro lado, el promedio de los contactos puede capturar el comportamiento del ensemble cuando los contactos muestran una variación menor entre confórmeros. En la Figura 4.3 se muestran representantes de ambos tipos de proteínas.

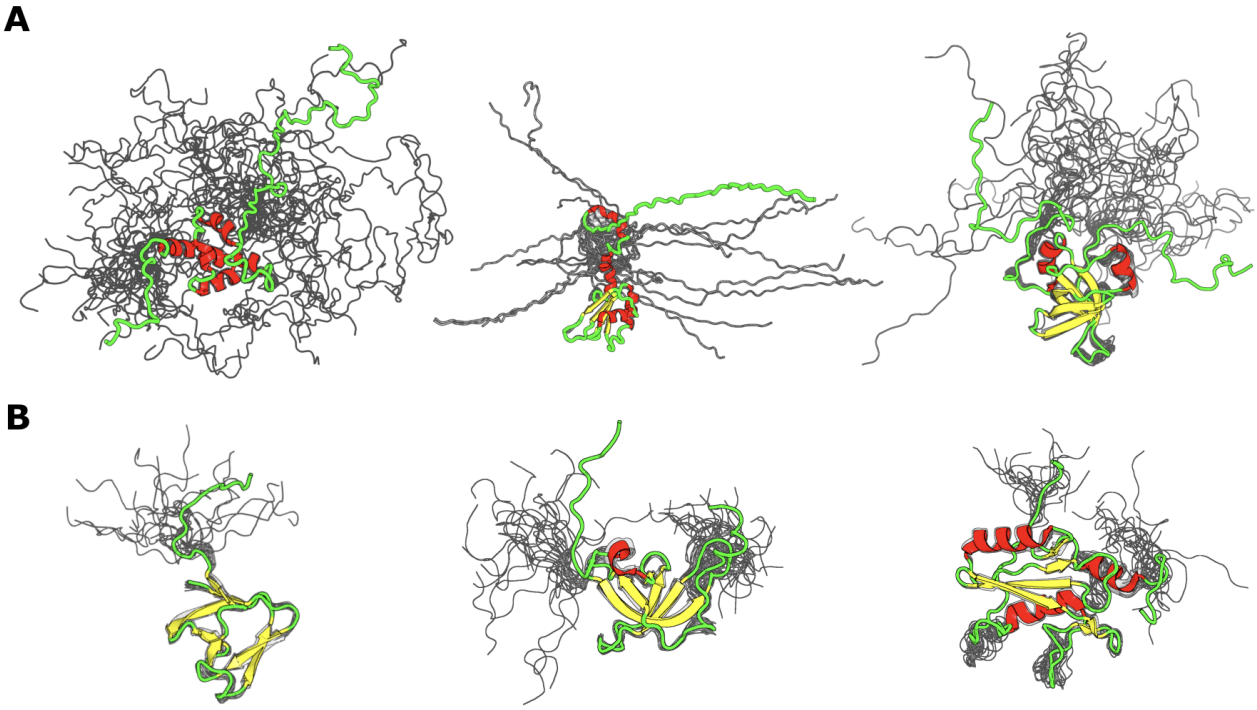


Figura 4.3 Representación de ensembles proteicos para cada grupo de correlación.

Representación de ensembles proteicos para las cuales la mejor correlación fue obtenida si se combinaba la información estructural de 2 o más confórmeros. ***Panel A:*** *ejemplos de ensembles de IDPs para las cuales la fracción de contactos arroja el mejor valor de correlación. De izquierda a derecha códigos PDB: 1G9L_A (dominio PABC de la proteína de unión poly(a)*

humana), 2M70_A (Citrus sinensis proteína de unión a poli (A) 1), 1D7Q_A (factor de iniciación de traducción humano EIF1A). ***Panel B:*** *Ejemplos de ensembles de IDPs para los cuales el promedio de los contactos correlaciona mejor con la velocidad normalizada de evolución (de izquierda a derecha códigos PDB: 2KRS_A (Dominio SH3 de Clostridium perfringens putativo de la enterotoxina CPF_0587), 2BUD_A (Dominio cromo putativo de Drosophila melanogaster de machos, ausente en la primera proteína (MOF), 2EO6_A dominio SH2 de ratón de la proteína enlazadora de células B BLNK).*

Se hipotetizó que las proteínas que presentan la mejor correlación con la fracción de contactos (como las que se muestran en la Figura 4.3, panel A) pueden llegar a ser más flexibles, y sus posiciones que están evolutivamente restringidas, pueden establecer contactos en determinado confórmeros y no en otros. En cambio, las proteínas que presentan una mejor correlación con el promedio de los contactos (tal como se muestran en la Figura 4.3, el panel B), son estructuras que son más cerradas donde el número de contactos en cada posición se mantiene en las distintas conformaciones del ensemble. Con el fin de probar esta hipótesis se estimó el valor de RMSD y el radio de giro entre todos los confórmeros en el ensemble y calculamos su valor máximo, mínimo y promedio. El radio de giro es una medida que aporta información sobre el grado de compactación de una proteína.

Se encontró que para el conjunto de proteínas en donde la robustez de contactos provee la mejor correlación con las velocidades de evolución el promedio del promedio de RMSD es de 19.77 Å y el promedio de los RMSD máximos es 35.2 Å mientras que el promedio del valor de radio de giro medio en el ensemble es 23.13Å. Para el conjunto de proteínas que muestran la mejor correlación con el promedio de los contactos y la velocidad de evolución, tienen valores más bajos de RMSD el promedio del promedio de RMSD es 11.12 Å y el promedio de los RMSD máximos es 20.28 Å y valores más bajos de radio de giro (promedio 18.27Å) (
Tabla 2). Tanto el RMSD como el radio de giro en el ensemble dan evidencia extra para respaldar la hipótesis planteada en la que cuando las velocidades de evolución son mejor explicadas con el promedio de los contactos deben presentar conformaciones menos extendidas y menos flexibles.

Mejor correlación	Robustez	Promedio
RSMD promedio	19.77 Å	11.12 Å
RMSD máximo	35.2 Å	20.28 Å
RG promedio	23.13 Å	18.27 Å

Tabla 2: Valores de RMSD promedio, RMSD máximo y RG promedio para las mejores correlaciones obtenidas con la robustez de contactos o el promedio.

4.2.2 Metodología

Para los análisis de correlación para confórmeros individuales se calcularon los coeficientes de correlación **ρ** (rho) de Pearson para cada confórmero, entre los valores de velocidad normalizados y el número de contactos por posición. Se repitió este análisis a nivel de ensemble para los análisis de correlación del ensemble entero, obteniendo las correlaciones de Pearson, entre las velocidades de evolución y el promedio de los contactos o la fracción de contactos por posición entre todos los confórmeros del ensemble. Luego se determinó la descripción de contactos que mejor correlacionó con las velocidades de evolución para cada proteína, comprando los valores de correlación. Para los análisis de correlación entre confórmeros combinados se calcularon todas las combinaciones posibles, usando la Ecuación 12. Para cada una de estas combinaciones, se calcularon los coeficientes (**ρ**) de correlación de Pearson, contra las velocidades de evolución de dos maneras: usando el promedio de los residuos de contactos ($\boldsymbol{\rho_{ave}}$) y utilizando la fracción de contactos ($\boldsymbol{\rho_{crob}}$). Para cada combinación y proteína, se compararon los valores de ρ con la distribución de bootstrapping (con 10.000 re-sampleos) para evaluar si cada ρ obtenido, era estadísticamente distinta de una distribución random, utilizando un t-test (con un nivel de significancia de 0.01).

El cálculo del RMSD fue descrito en tal sección. Y para cada confórmero se calculó el radio de giro con MMTSB Tool Set (Feig et al., 2004).

4.2.3 Discusión

De acuerdo con los resultados comentados, alrededor del 80% de las IDPs estudiadas presentan una mejor correlación entre las velocidades de evolución y los perfiles de contactos, cuando la información de contactos es tomada para un subconjunto particular de confórmeros y no uno en particular. Si se tiene en cuenta que en promedio estos ensembles están formados por 20 conformaciones las mejores correlaciones se obtienen cuando se combina la información de pocos confórmeros (< 8). Esto último puede deberse a que posiblemente existan conformaciones 'preferidas' que posiblemente estén relacionadas con la estructura y la función de la proteína (Fuxreiter, 2012; Kim and Chung, 2020). Estos resultados están en concordancia con los resultados mostrados en 4.1 y además abren la posibilidad de nuevos trabajos bioinformáticos como herramientas para el curado de bases de datos de IDPs (Varadi et al., 2014). Debido a que la mayoría de las correlaciones más altas entre la velocidad de evolución, fueron obtenidas con la fracción de contactos, esto soporta la visión actual que las IDPs son ensembles altamente dinámicos con múltiples interacciones débiles, establecidas a lo largo de toda la proteína, posiblemente para permitir la estabilización del ensemble y la interacción con otras proteínas (Borgia et al., 2018; Brodsky et al., 2020; Das and Pappu, 2013; Protter et al., 2018). En general, la fracción de contactos refleja mejor las velocidades de evolución que el promedio de los contactos. Esto es así posiblemente debido a la enorme cantidad de interacciones dinámicas que son necesarias para mantener el ensemble (Mittal et al., 2018).

5 Conclusiones generales y perspectivas a futuro

En este trabajo queda de manifiesto que un enfoque evolutivo es necesario e imprescindible a la hora de estudiar y caracterizar ensembles complejos de proteínas como las IDPs. Estas proteínas a pesar de su alta flexibilidad poseen condicionamientos estructurales, cuya presencia queda en evidencia claramente, y que no solo ocurren en posiciones ordenadas sino también en posiciones desordenadas, lo que demuestra que el desorden es un elemento de la organización espacial necesario para la función biológica de IDPs. Estos condicionamientos estructurales son debidos mayormente al establecimiento de contactos terciarios en conformaciones particulares. Así como se sabe que el mantenimiento de determinados residuos es importante para la función biológica, esta idea se puede extender a conformaciones. El ensemble conformacional es clave para la función biológica de estas proteínas y las distintas conformaciones presentes en el ensemble soportan presiones selectivas diferenciales, modulando la evolución de dichas proteínas.

Si bien en este trabajo no se pudo determinar las características particulares para que determinadas conformaciones fueran biológicamente relevantes, se demostró desde un enfoque evolutivo que los ensembles conformacionales de IDPs son redundantes en cuanto a información estructural. La pregunta a hacerse entonces es, ¿por qué son descritos por tantos confórmeros? ¿En qué aspectos de las distintas conformaciones actúa la presión de selección? Lamentablemente las estructuras estimadas por NMR no poseen información sobre la abundancia relativa de cada confórmero en el ensemble. De la misma forma, aún es difícil establecer experimentalmente cuál de las conformaciones posee un rol distintivo dificultando el avance en el área

Se demostró que las velocidades de evolución son un buen parámetro para capturar organizaciones estructurales determinadas. Si bien con los resultados obtenidos y

mostrados en este trabajo no es posible ser del todo concluyente, nuestros resultados preliminares obtenidos recientemente indicarían que las velocidades de evolución pueden también servir para caracterizar regiones desordenadas de manera más sensible pudiendo diferenciar distintos tipos estructurales de desorden, esto es terminal o interno dependiendo de su ubicación. Es interesante destacar que nuestra metodología conjuga la información experimental (por ej. la determinación de un contacto físico entre dos aminoácidos en una determinada estructura o conformación) con la importancia de esa interacción estimada a partir de cómo la misma impacta en el patrón evolutivo de la familia de proteínas. Esta combinación de dos tipos de información bien distintas nos permitieron refinar y mejorar la información derivada de los contactos físicos entre residuos.

A pesar de que han pasado aproximadamente 30 años desde el descubrimiento de las IDPs en la actualidad siguen siendo desconocidas por buena parte de la comunidad científica. Es necesario comprender que las IDPs no son excepciones a una regla, sino que son la punta de un iceberg de un mundo estructural inmenso y que posiblemente aún no hemos convergido en su exploración ni entendimiento. Recientemente, se han caracterizado proteínas cuyas estructuras no pueden asociarse o derivarse de ningún plegamiento conocido en la actualidad. Estas proteínas se denominan “dark” (o “dark proteome”) haciendo referencia a la materia oscura detectada por físicos y astrónomos. Como en el caso de la materia oscura que los físicos saben que existe pero que desconocen su naturaleza, lo mismo ocurre a nivel de los proteomas de casi todos los tipos de organismos: sabemos que hay regiones codificantes que expresan proteínas cuyas propiedades estructurales, su relación estructura-función son completamente desconocidas e imposibles de predecir a partir del universo de proteínas conocido. Es por eso que el desarrollo de métodos evolutivos y/o evolutivos/estructurales aplicables a proteínas como las IDPs que desafían el concepto de la relación estructura-función, sirva como motor para explorar nuevas clases de proteínas aún inexploradas o desconocidas. A pesar de la importancia y necesidad del enfoque evolutivo, la disponibilidad de herramientas basadas en métodos evolutivos que sean confórmero-específicas es baja, por eso creemos que es necesario profundizar más sobre la relación entre desorden y

evolución. Creemos que estos avances darán las bases propicias para el desarrollo de herramientas que permitan una mayor compresión sobre IDPs. Es por ello que aún resta mucho hacer, por mencionar algunas:

- Predicción de desorden a partir de parámetros que no necesiten del dato de la estructura como las velocidades de evolución o co-evolución para inferir presencia de desorden y orden en proteínas cuya estructura es desconocida.

- Desarrollo de métodos con un enfoque evolutivo (como el que se presenta en este trabajo) para la curación de ensembles de IDPs obtenidos métodos computacionales como dinámica molecular. Estos métodos pueden basarse en análisis de secuencias homólogas para extraer información de posiciones relevantes biológicamente y descartar aquellas que derivan meramente de la divergencia. Los métodos de coevolución y de estimación de velocidades de evolución pueden ser útiles también en este objetivo.

6 Listado de abreviaturas utilizados en esta tesis

- Å: ***Å**ngströms*
- ADN: ***Á**cido **D**esoxirribo**n**ucleico*
- ARN: ***Á**cido **R**ibo**n**ucleico*
- Cα: ***C**arbonos alfa*
- CASP: ***C**ritical **A**ssessment of **P**rotein **S**tructure **P**rediction*. (Evaluación de predicción de estructura proteica)
- CAID: **C**ritical **A**ssessment of protein **I**ntrinsic **D**isorder prediction. (Evaluación de predicción de desorden intrínseco).
- CoDNaS: ***C**onformational **D**iversity of **Na**tive **S**tate in proteins* (http://ufq.unq.edu.ar/codnas/)
- IDP: ***I**ntrinsically **D**isordered **P**rotein* (Proteínas intrínsecamente desordenadas)
- IDR: ***I**ntrinsically **D**isordered **R**egion* (Regiones intrínsecamente desordenadas)
- LOESS: ***L**ocally **E**stimated **S**catterplot **S**moothing* (Regresión local)
- MSA: ***M**ultiple **S**equence **A**lignment* (Alineamiento múltiple de secuencias)
- ML: ***M**aximum **L**ikelihood* (Máxima Verosimilitud)
- NMR: ***N**uclear **M**agnetic **R**esonance* (Resonancia Magnética Nuclear)
- PDB: ***P**rotein **D**ata **B**ank* (http://www.wwpdb.org/)
- RMSF ***R**oot **M**ean **S**quare **F**luctuation* (Fluctuación cuadrática media)
- RMSD: ***R**oot **M**ean **S**quare **D**eviation* (Desviación cuadrática media) raíz cuadrada del valor medio de los cuadrados
- SAXS: ***S**mall **A**ngle **X**-ray **S**cattering* (Dispersión de rayos X de ángulo reducido).
- SCPE: ***S**tructurally **C**onstrained model **P**rotein **E**volution* (Modelo de evolución de proteínas con restricciones estructurales)
- SLiMs: ***S**hort **Li**near **M**otifs* (Motivos lineales cortos)

7 Apéndice A

A continuación, se muestran las figuras correspondientes a los resultados mostrados en la sección 3.3.2, considerando solo aquellas posiciones sin interacciones locales, es decir que poseen contactos con residuos que se encuentra a 4 residuos o más de distancia. Otra aclaración que es importante tener en cuenta es que solamente se consideran posiciones que tienen al menos un contacto (por eso no existe la robustez 0 en estos gráficos).

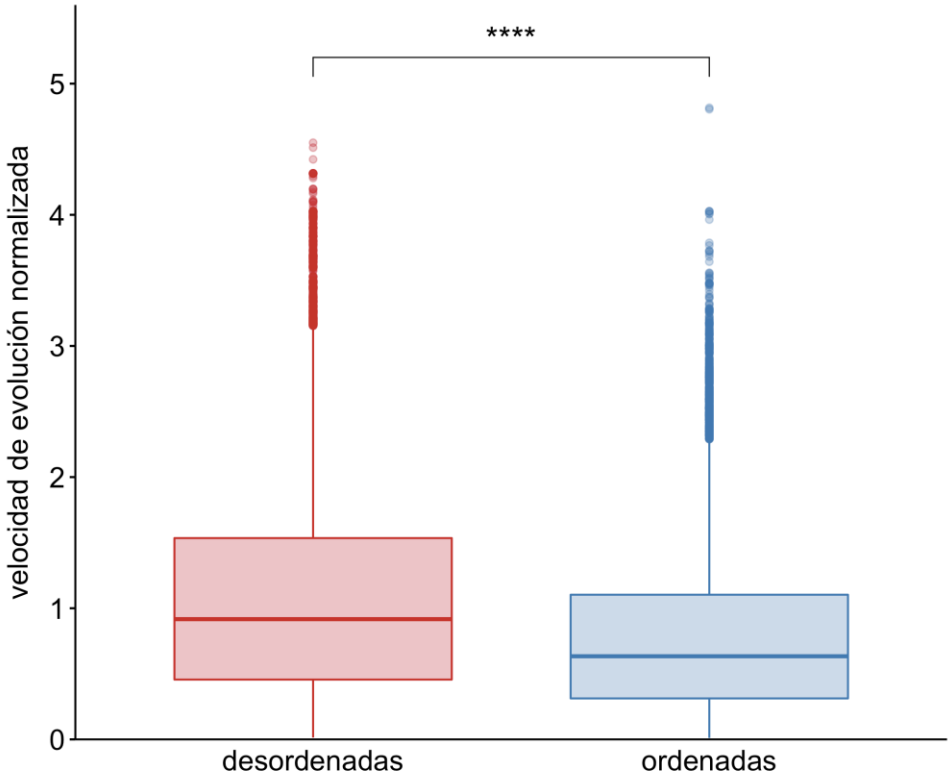


Figura 7.1 Boxplots representando la distribución de las velocidades de evolución normalizadas por posición.

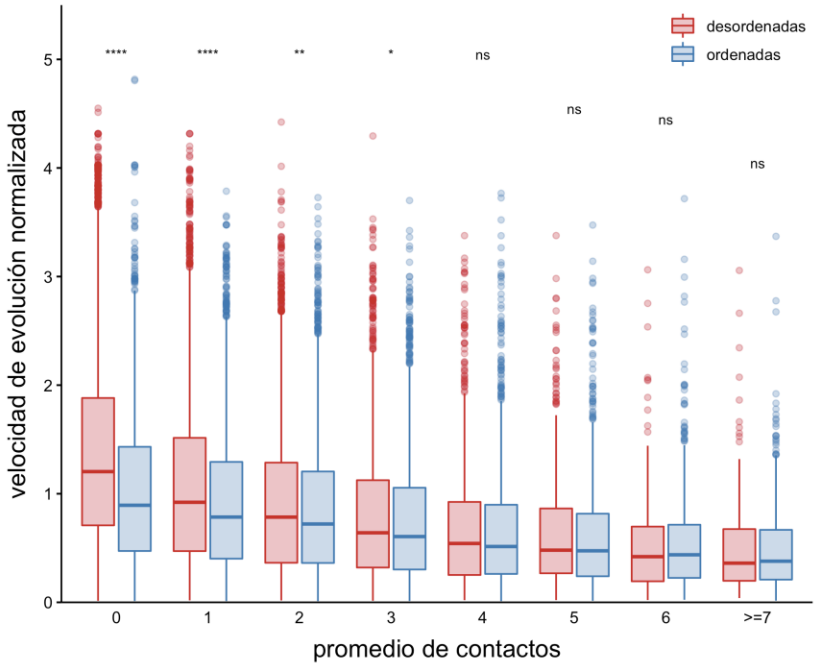


Figura 7.2. Velocidad de evolución normalizada en función del promedio de contactos.

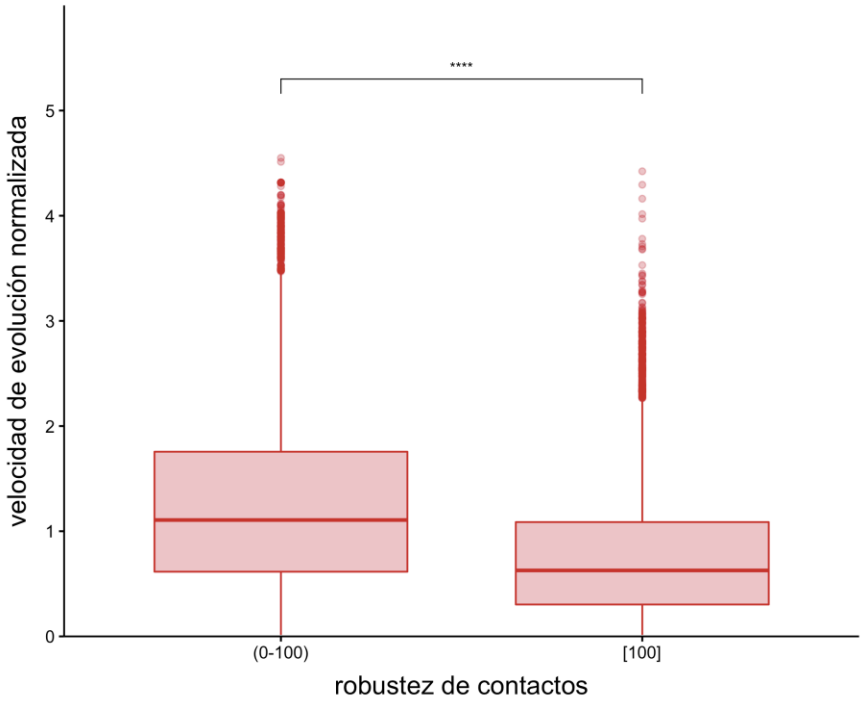


Figura 7.3 Distribución de las velocidades de evolución normalizadas en función de la robustez de contactos.

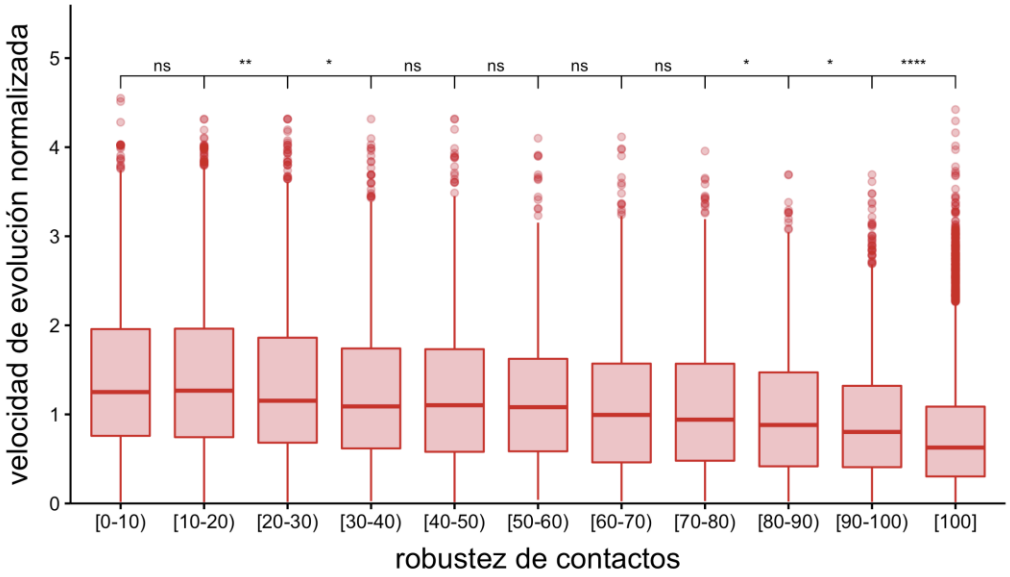


Figura 7.4 Velocidad de evolución normalizada en función de la robustez de contacto.

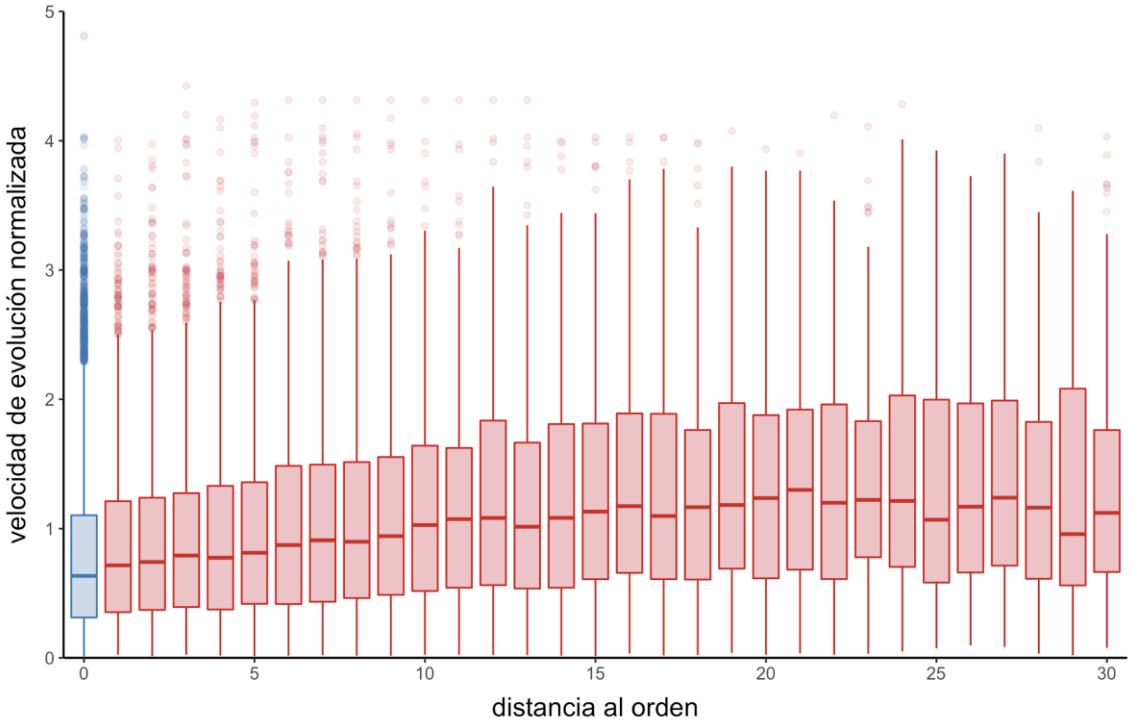


Figura 7.5 Distribución de las velocidades de evolución normalizadas en función de la distancia al residuo ordenado más cercano.

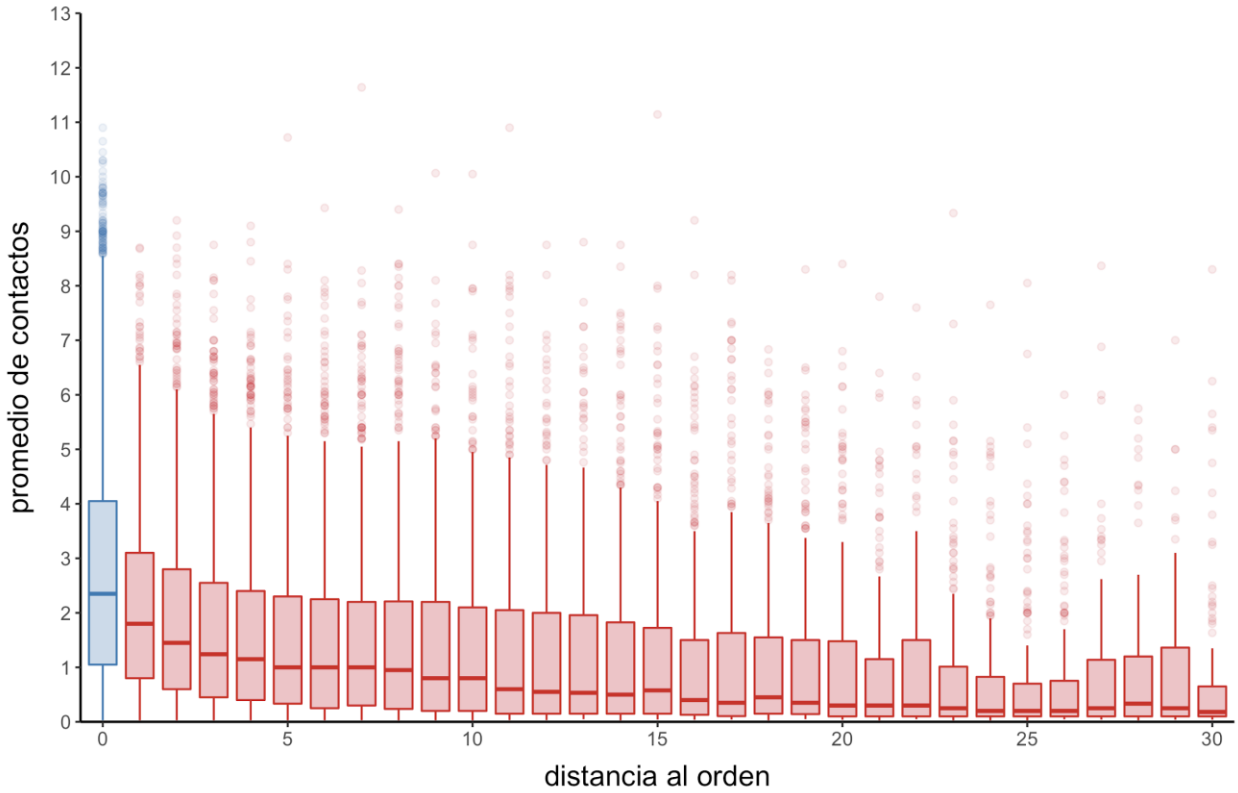


Figura 7.6 Distribución del promedio de contactos en función de la distancia al residuo ordenado más cercano.

8 Bibliografía

Adachi, J., Hasegawa, M., 1996. Model of amino acid substitution in proteins encoded by mitochondrial DNA. J. Mol. Evol. 42, 459–468. doi:10.1007/BF02498640

Adachi, J., Waddell, P.J., Martin, W., Hasegawa, M., 2000. Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. J. Mol. Evol. 50, 348–358. doi:10.1007/s002399910038

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Automat. Contr. 19, 716–723. doi:10.1109/TAC.1974.1100705

Alber, T., Gilbert, W.A., Ponzi, D.R., Petsko, G.A., 1983. The role of mobility in the substrate binding and catalytic machinery of enzymes. Ciba Found. Symp. 93, 4–24.

Amemiya, T., Koike, R., Kidera, A., Ota, M., 2012. PSCDB: a database for protein structural change upon ligand binding. Nucleic Acids Res. 40, D554-8. doi:10.1093/nar/gkr966

Anfinsen, C.B., 1973. Principles that govern the folding of protein chains. Science 181, 223–230. doi:10.1126/science.181.4096.223

Atkins, J.D., Boateng, S.Y., Sorensen, T., McGuffin, L.J., 2015. Disorder prediction methods, their applicability to different protein targets and their usefulness for guiding experimental studies. Int. J. Mol. Sci. 16, 19040–19054. doi:10.3390/ijms160819040

Baer, C.F., Miyamoto, M.M., Denver, D.R., 2007. Mutation rate variation in multicellular eukaryotes: causes and consequences. Nat. Rev. Genet. 8, 619–631. doi:10.1038/nrg2158

Bastolla, U., Dehouck, Y., Echave, J., 2017. What evolution tells us about protein physics, and protein physics tells us about evolution. Curr. Opin. Struct. Biol. 42, 59–66. doi:10.1016/j.sbi.2016.10.020

Bastolla, U., Roman, H.E., Vendruscolo, M., 1999. Neutral evolution of model proteins: diffusion in sequence space and overdispersion. J. Theor. Biol. 200, 49–64. doi:10.1006/jtbi.1999.0975

Baum, B.R., 1989. *phylip: phylogeny inference package. version 3.2* . joel felsenstein. Q.

Rev. Biol. 64, 539–541. doi:10.1086/416571

Berlow, R.B., Dyson, H.J., Wright, P.E., 2015. Functional advantages of dynamic protein disorder. FEBS Lett. 589, 2433–2440. doi:10.1016/j.febslet.2015.06.003

Berrera, M., Molinari, H., Fogolari, F., 2003. Amino acid empirical contact energy definitions for fold recognition in the space of contact maps. BMC Bioinformatics 4, 8. doi:10.1186/1471-2105-4-8

Best, R.B., Lindorff-Larsen, K., DePristo, M.A., Vendruscolo, M., 2006. Relation between native ensembles and experimental structures of proteins. Proc Natl Acad Sci USA 103, 10901–10906. doi:10.1073/pnas.0511156103

BLAKE, C.C.F., KOENIG, D.F., MAIR, G.A., NORTH, A.C.T., PHILLIPS, D.C., SARMA, V.R., 1965. Structure of Hen Egg-White Lysozyme: A Three-dimensional Fourier Synthesis at 2 Å Resolution. Nature 206, 757–761. doi:10.1038/206757a0

Bloomer, A.C., Champness, J.N., Bricogne, G., Staden, R., Klug, A., 1978. Protein disk of tobacco mosaic virus at 2.8 A resolution showing the interactions within and between subunits. Nature 276, 362–368. doi:10.1038/276362a0

Bode, W., Schwager, P., Huber, R., 1978. The transition of bovine trypsinogen to a trypsin-like state upon strong ligand binding. The refined crystal structures of the bovine trypsinogen-pancreatic trypsin inhibitor complex and of its ternary complex with Ile-Val at 1.9 A resolution. J. Mol. Biol. 118, 99–112.

Boehr, D.D., Nussinov, R., Wright, P.E., 2009. The role of dynamic conformational ensembles in biomolecular recognition. Nat. Chem. Biol. 5, 789–796. doi:10.1038/nchembio.232

Booth, V.H., Bradford, M.P., 1963. Tocopherol contents of vegetables and fruits. Br. J. Nutr. 17, 575–581. doi:10.1079/bjn19630060

Borgia, A., Borgia, M.B., Bugge, K., Kissling, V.M., Heidarsson, P.O., Fernandes, C.B., Sottini, A., Soranno, A., Buholzer, K.J., Nettels, D., Kragelund, B.B., Best, R.B., Schuler, B., 2018. Extreme disorder in an ultrahigh-affinity protein complex. Nature 555, 61–66. doi:10.1038/nature25762

Brodsky, S., Jana, T., Mittelman, K., Chapal, M., Kumar, D.K., Carmi, M., Barkai, N., 2020. Intrinsically disordered regions direct transcription factor in vivo binding specificity. Mol. Cell 79, 459-471.e4. doi:10.1016/j.molcel.2020.05.032

Brown, C.J., Johnson, A.K., Dunker, A.K., Daughdrill, G.W., 2011. Evolution and disorder. Curr. Opin. Struct. Biol. 21, 441–446. doi:10.1016/j.sbi.2011.02.005

Brown, C.J., Takayama, S., Campen, A.M., Vise, P., Marshall, T.W., Oldfield, C.J., Williams, C.J., Dunker, A.K., 2002. Evolutionary rate heterogeneity in proteins with long disordered regions. J. Mol. Evol. 55, 104–110. doi:10.1007/s00239-001-2309-6

Bryngelson, J.D., Wolynes, P.G., 1989. Intermediates and barrier crossing in a random energy model (with applications to protein folding). J. Phys. Chem. 93, 6902–6915. doi:10.1021/j100356a007

Burra, P.V., Zhang, Y., Godzik, A., Stec, B., 2009. Global distribution of conformational states derived from redundant models in the PDB points to non-uniqueness of the protein structure. Proc Natl Acad Sci USA 106, 10505–10510. doi:10.1073/pnas.0812152106

Carugo, O., Pongor, S., 2001. A normalized root-mean-square distance for comparing protein three-dimensional structures. Protein Sci. 10, 1470–1473. doi:10.1110/ps.690101

Chothia, C., Lesk, A.M., 1986. The relation between the divergence of sequence and structure in proteins. EMBO J. 5, 823–826. doi:10.1002/j.1460-2075.1986.tb04288.x

Chothia, C., Lesk, A.M., 1987. The evolution of protein structures. Cold Spring Harb. Symp. Quant. Biol. 52, 399–405. doi:10.1101/sqb.1987.052.01.046

Chouard, T., 2011. Structural biology: Breaking the protein rules. Nature 471, 151–153. doi:10.1038/471151a

Conrad, M., 1977. Evolutionary adaptability of biological macromolecules. J. Mol. Evol. 10, 87–91. doi:10.1007/BF01796137

Cui, Q., Karplus, M., 2008. Allostery and cooperativity revisited. Protein Sci. 17, 1295–1307. doi:10.1110/ps.03259908

Das, R.K., Pappu, R.V., 2013. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. Proc Natl Acad Sci USA 110, 13392–13397. doi:10.1073/pnas.1304749110

Daughdrill, G.W., Chadsey, M.S., Karlinsey, J.E., Hughes, K.T., Dahlquist, F.W., 1997. The C-terminal half of the anti-sigma factor, FlgM, becomes structured when bound to its target, sigma 28. Nat. Struct. Biol. 4, 285–291. doi:10.1038/nsb0497-285

Davey, N.E., 2019. The functional importance of structure in unstructured protein regions. Curr. Opin. Struct. Biol. 56, 155–163. doi:10.1016/j.sbi.2019.03.009

Dayhoff, Schwartz, Orcutt, 1978. A Model of Evolutionary Change in Proteins. National Biomedical Research Foundation.

DeForte, S., Uversky, V.N., 2016. Resolving the ambiguity: Making sense of intrinsic disorder when PDB structures disagree. Protein Sci. 25, 676–688. doi:10.1002/pro.2864

Di Cera, E., 2020. Mechanisms of ligand binding. Biophys. Rev. 1, 011303. doi:10.1063/5.0020997

Dobzhansky, T., 1973. Nothing in Biology Makes Sense except in the Light of Evolution. Am. Biol. Teach. 35, 125–129. doi:10.2307/4444260

Dodge, C., Schneider, R., Sander, C., 1998. The HSSP database of protein structure-sequence alignments and family profiles. Nucleic Acids Res. 26, 313–315. doi:10.1093/nar/26.1.313

Dosztányi, Z., Csizmók, V., Tompa, P., Simon, I., 2005. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J. Mol. Biol. 347, 827–839. doi:10.1016/j.jmb.2005.01.071

Drummond, D.A., Bloom, J.D., Adami, C., Wilke, C.O., Arnold, F.H., 2005. Why highly expressed proteins evolve slowly. Proc Natl Acad Sci USA 102, 14338–14343. doi:10.1073/pnas.0504070102

Dunker, A.K., Brown, C.J., Lawson, J.D., Iakoucheva, L.M., Obradović, Z., 2002. Intrinsic disorder and protein function. Biochemistry 41, 6573–6582. doi:10.1021/bi012159+

Dunker, A.K., Lawson, J.D., Brown, C.J., Williams, R.M., Romero, P., Oh, J.S., Oldfield, C.J., Campen, A.M., Ratliff, C.M., Hipps, K.W., Ausio, J., Nissen, M.S., Reeves, R., Kang, C., Kissinger, C.R., Bailey, R.W., Griswold, M.D., Chiu, W., Garner, E.C., Obradovic, Z., 2001. Intrinsically disordered protein. J. Mol. Graph. Model. 19, 26–59. doi:10.1016/S1093-3263(00)00138-8

Echave, J., Spielman, S.J., Wilke, C.O., 2016. Causes of evolutionary rate variation among protein sites. Nat. Rev. Genet. 17, 109–121. doi:10.1038/nrg.2015.18

Eisenmesser, E.Z., Millet, O., Labeikovsky, W., Korzhnev, D.M., Wolf-Watz, M., Bosco,

D.A., Skalicky, J.J., Kay, L.E., Kern, D., 2005. Intrinsic dynamics of an enzyme underlies catalysis. Nature 438, 117–121. doi:10.1038/nature04105

Feig, M., Karanicolas, J., Brooks, C.L., 2004. MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology. J. Mol. Graph. Model. 22, 377–395. doi:10.1016/j.jmgm.2003.12.005

Fichó, E., Reményi, I., Simon, I., Mészáros, B., 2017. MFIB: a repository of protein complexes with mutual folding induced by binding. Bioinformatics 33, 3682–3684. doi:10.1093/bioinformatics/btx486

Fischer, E., 1894. Einfluss der configuration auf die wirkung der enzyme. II. Ber. Dtsch. Chem. Ges. 27, 3479–3483. doi:10.1002/cber.189402703169

Fletcher, C.M., Wagner, G., 1998. The interaction of eIF4E with 4E-BP1 is an induced fit to a completely disordered protein. Protein Sci. 7, 1639–1642. doi:10.1002/pro.5560070720

Flock, T., Weatheritt, R.J., Latysheva, N.S., Babu, M.M., 2014. Controlling entropy to tune the functions of intrinsically disordered regions. Curr. Opin. Struct. Biol. 26, 62–72. doi:10.1016/j.sbi.2014.05.007

Flores, T.P., Orengo, C.A., Moss, D.S., Thornton, J.M., 1993. Comparison of conformational characteristics in structurally similar protein pairs. Protein Sci. 2, 1811–1826. doi:10.1002/pro.5560021104

Foote, J., Milstein, C., 1994. Conformational isomerism and the diversity of antibodies. Proc Natl Acad Sci USA 91, 10370–10374. doi:10.1073/pnas.91.22.10370

Forman-Kay, J.D., Mittag, T., 2013. From sequence and forces to structure, function, and evolution of intrinsically disordered proteins. Structure 21, 1492–1499. doi:10.1016/j.str.2013.08.001

Fornasari, M.S., Parisi, G., Echave, J., 2002. Site-specific amino acid replacement matrices from structurally constrained protein evolution simulations. Mol. Biol. Evol. 19, 352–356. doi:10.1093/oxfordjournals.molbev.a004089

Fornasari, M.S., Parisi, G., Echave, J., 2007. Quaternary structure constraints on evolutionary sequence divergence. Mol. Biol. Evol. 24, 349–351. doi:10.1093/molbev/msl181

Franzosa, E.A., Xia, Y., 2009. Structural determinants of protein evolution are context-

sensitive at the residue level. Mol. Biol. Evol. 26, 2387–2395. doi:10.1093/molbev/msp146

Frauenfelder, H., McMahon, B.H., Fenimore, P.W., 2003. Myoglobin: the hydrogen atom of biology and a paradigm of complexity. Proc Natl Acad Sci USA 100, 8615–8617. doi:10.1073/pnas.1633688100

Frauenfelder, H., Sligar, S.G., Wolynes, P.G., 1991. The energy landscapes and motions of proteins. Science 254, 1598–1603. doi:10.1126/science.1749933

Futuyma, D., 2009. Evolution, Second Edition, 2nd Edition. ed. Sinauer Associates, Inc., Sunderland, Mass.

Fuxreiter, M., 2012. Fuzziness: linking regulation to protein dynamics. Mol. Biosyst. 8, 168–177. doi:10.1039/c1mb05234a

Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W., 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28, 3150–3152. doi:10.1093/bioinformatics/bts565

Garbuzynskiy, S.O., Melnik, B.S., Lobanov, M.Y., Finkelstein, A.V., Galzitskaya, O.V., 2005. Comparison of X-ray and NMR structures: is there a systematic difference in residue contacts between X-ray- and NMR-resolved protein structures? Proteins 60, 139–147. doi:10.1002/prot.20491

Gerstein, M., Echols, N., 2004. Exploring the range of protein flexibility, from a structural proteomics perspective. Curr. Opin. Chem. Biol. 8, 14–19. doi:10.1016/j.cbpa.2003.12.006

Gerstein, M., Krebs, W., 1998. A database of macromolecular motions. Nucleic Acids Res. 26, 4280–4290. doi:10.1093/nar/26.18.4280

Goldstein, R.A., 2011. The evolution and evolutionary consequences of marginal thermostability in proteins. Proteins 79, 1396–1407. doi:10.1002/prot.22964

Gouw, M., Michael, S., Sámano-Sánchez, H., Kumar, M., Zeke, A., Lang, B., Bely, B., Chemes, L.B., Davey, N.E., Deng, Z., Diella, F., Gürth, C.-M., Huber, A.-K., Kleinsorg, S., Schlegel, L.S., Palopoli, N., Roey, K.V., Altenberg, B., Reményi, A., Dinkel, H., Gibson, T.J., 2018. The eukaryotic linear motif resource - 2018 update. Nucleic Acids Res. 46, D428–D434. doi:10.1093/nar/gkx1077

Gu, Y., Li, D.-W., Brüschweiler, R., 2015. Decoding the mobility and time scales of protein

loops. J. Chem. Theory Comput. 11, 1308–1314. doi:10.1021/ct501085y

Guthery, F.S., Burnham, K.P., Anderson, D.R., 2003. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. The Journal of Wildlife Management 67, 655. doi:10.2307/3802723

Gutteridge, A., Thornton, J., 2005. Conformational changes observed in enzyme crystal structures upon substrate binding. J. Mol. Biol. 346, 21–28. doi:10.1016/j.jmb.2004.11.013

Habchi, J., Tompa, P., Longhi, S., Uversky, V.N., 2014. Introducing protein intrinsic disorder. Chem. Rev. 114, 6561–6588. doi:10.1021/cr400514h

Hammes, G.G., Chang, Y.-C., Oas, T.G., 2009. Conformational selection or induced fit: a flux description of reaction mechanism. Proc Natl Acad Sci USA 106, 13737–13741. doi:10.1073/pnas.0907195106

Harrison, S.C., Durbin, R., 1985. Is there a single pathway for the folding of a polypeptide chain? Proc Natl Acad Sci USA 82, 4028–4030. doi:10.1073/pnas.82.12.4028

Hatos, A., Hajdu-Soltész, B., Monzon, A.M., Palopoli, N., Álvarez, L., Aykac-Fas, B., Bassot, C., Benítez, G.I., Bevilacqua, M., Chasapi, A., Chemes, L., Davey, N.E., Davidović, R., Dunker, A.K., Elofsson, A., Gobeill, J., Foutel, N.S.G., Sudha, G., Guharoy, M., Horvath, T., Piovesan, D., 2020. DisProt: intrinsic protein disorder annotation in 2020. Nucleic Acids Res. 48, D269–D276. doi:10.1093/nar/gkz975

He, B., Wang, K., Liu, Y., Xue, B., Uversky, V.N., Dunker, A.K., 2009. Predicting intrinsic disorder in proteins: an overview. Cell Res. 19, 929–949. doi:10.1038/cr.2009.87

Hopf, T.A., Schärfe, C.P.I., Rodrigues, J.P.G.L.M., Green, A.G., Kohlbacher, O., Sander, C., Bonvin, A.M.J.J., Marks, D.S., 2014. Sequence co-evolution gives 3D contacts and structures of protein complexes. elife 3. doi:10.7554/eLife.03430

Illergård, K., Ardell, D.H., Elofsson, A., 2009. Structure is three to ten times more conserved than sequence--a study of structural response in protein cores. Proteins 77, 499–508. doi:10.1002/prot.22458

Irving, J.A., Whisstock, J.C., Lesk, A.M., 2001. Protein structural alignments and functional genomics. Proteins 42, 378–382. doi:10.1002/1097-0134(20010215)42:3<378::aid-prot70>3.0.co;2-3

Ishida, T., Kinoshita, K., 2007. PrDOS: prediction of disordered protein regions from

amino acid sequence. Nucleic Acids Res. 35, W460-4. doi:10.1093/nar/gkm363

James, L.C., Tawfik, D.S., 2003. Conformational diversity and protein evolution--a 60-year-old hypothesis revisited. Trends Biochem. Sci. 28, 361–368. doi:10.1016/S0968-0004(03)00135-X

Jones, D.T., Cozzetto, D., 2015. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. Bioinformatics 31, 857–863. doi:10.1093/bioinformatics/btu744

Jones, David T., Taylor, W.R., Thornton, J.M., 1992. The rapid generation of mutation data matrices from protein sequences. Bioinformatics 8, 275–282. doi:10.1093/bioinformatics/8.3.275

Jones, D T, Taylor, W.R., Thornton, J.M., 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. 8, 275–282.

Juritz, E., Palopoli, N., Fornasari, M.S., Fernandez-Alberti, S., Parisi, G., 2013. Protein conformational diversity modulates sequence divergence. Mol. Biol. Evol. 30, 79–87. doi:10.1093/molbev/mss080

Kabsch, W., 1978. A discussion of the solution for the best rotation to relate two sets of vectors. Acta Cryst. A 34, 827–828. doi:10.1107/S0567739478001680

Kajava, A.V., 2001. Review: proteins with repeated sequence--structural prediction and modeling. J. Struct. Biol. 134, 132–144. doi:10.1006/jsbi.2000.4328

Kamata, K., Mitsuya, M., Nishimura, T., Eiki, J.-I., Nagata, Y., 2004. Structural basis for allosteric regulation of the monomeric allosteric enzyme human glucokinase. Structure 12, 429–438. doi:10.1016/j.str.2004.02.005

Karpati, G., Carpenter, S., Durham, H., 1988. A hypothesis for the pathogenesis of amyotrophic lateral sclerosis. Rev Neurol (Paris) 144, 672–675.

Karush, F., 1950. Heterogeneity of the Binding Sites of Bovine Serum Albumin. J. Am. Chem. Soc. 72, 2705–2713. doi:10.1021/ja01162a099

Kendrew, J., Bodo, G., Dintzis, H., Parrish, R., Wyckoff, H., Phillips, D., 1958. A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. Nature 181, 662–666. doi:10.1038/181662a0

Kimura, M., 1968. Evolutionary rate at the molecular level. Nature 217, 624–626. doi:10.1038/217624a0

Kimura, M., Ohta, T., 1974. On some principles governing molecular evolution. Proc Natl Acad Sci USA 71, 2848–2852. doi:10.1073/pnas.71.7.2848

Kim, J.-Y., Chung, H.S., 2020. Disordered proteins follow diverse transition paths as they fold and bind to a partner. Science 368, 1253–1257. doi:10.1126/science.aba3854

King, J.L., Jukes, T.H., 1969. Non-Darwinian evolution. Science 164, 788–798. doi:10.1126/science.164.3881.788

Kleinman, C.L., Rodrigue, N., Lartillot, N., Philippe, H., 2010. Statistical potentials for improved structurally constrained evolutionary models. Mol. Biol. Evol. 27, 1546–1560. doi:10.1093/molbev/msq047

Koshland, D.E., 1958. Application of a Theory of Enzyme Specificity to Protein Synthesis. Proc Natl Acad Sci USA 44, 98–104. doi:10.1073/pnas.44.2.98

Kosiol, C., Goldman, N., 2011. Markovian and non-Markovian protein sequence evolution: aggregated Markov process models. J. Mol. Biol. 411, 910–923. doi:10.1016/j.jmb.2011.06.005

Kriwacki, R.W., Hengst, L., Tennant, L., Reed, S.I., Wright, P.E., 1996. Structural studies of p21Waf1/Cip1/Sdi1 in the free and Cdk2-bound state: conformational disorder mediates binding diversity. Proc Natl Acad Sci USA 93, 11504–11509. doi:10.1073/pnas.93.21.11504

Kumar, S., Ma, B., Tsai, C.J., Sinha, N., Nussinov, R., 2000. Folding and binding cascades: dynamic landscapes and population shifts. Protein Sci. 9, 10–19. doi:10.1110/ps.9.1.10

Kuriyan, J., Karplus, M., Petsko, G.A., 1987. Estimation of uncertainties in X-ray refinement results by use of perturbed structures. Proteins 2, 1–12. doi:10.1002/prot.340020102

Lazar, T., Martínez-Pérez, E., Quaglia, F., Hatos, A., Chemes, L.B., Iserte, J.A., Méndez, N.A., Garrone, N.A., Saldaño, T.E., Marchetti, J., Rueda, A.J.V., Bernadó, P., Blackledge, M., Cordeiro, T.N., Fagerberg, E., Forman-Kay, J.D., Fornasari, M.S., Gibson, T.J., Gomes, G.-N.W., Gradinaru, C.C., Piovesan, D., 2021. PED in 2021: a major update of the protein ensemble database for intrinsically disordered proteins. Nucleic Acids Res. 49, D404–D411. doi:10.1093/nar/gkaa1021

Lesk, A.M., Chothia, C., 1980. How different amino acid sequences determine similar

protein structures: the structure and evolutionary dynamics of the globins. J. Mol. Biol. 136, 225–270. doi:10.1016/0022-2836(80)90373-3

Levinthal, C., 1968. Are there pathways for protein folding? J. Chim. Phys. 65, 44–45. doi:10.1051/jcp/1968650044

Linding, R., Jensen, L.J., Diella, F., Bork, P., Gibson, T.J., Russell, R.B., 2003a. Protein disorder prediction: implications for structural proteomics. Structure 11, 1453–1459. doi:10.1016/j.str.2003.10.002

Linding, R., Russell, R.B., Neduva, V., Gibson, T.J., 2003b. GlobPlot: Exploring protein sequences for globularity and disorder. Nucleic Acids Res. 31, 3701–3708. doi:10.1093/nar/gkg519

Liu, Y., Wang, X., Liu, B., 2019. A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction. Brief. Bioinformatics 20, 330–346. doi:10.1093/bib/bbx126

Lobanov, M.Y., Galzitskaya, O.V., 2015. How common is disorder? occurrence of disordered residues in four domains of life. Int. J. Mol. Sci. 16, 19490–19507. doi:10.3390/ijms160819490

Lynch, M., Ackerman, M.S., Gout, J.-F., Long, H., Sung, W., Thomas, W.K., Foster, P.L., 2016. Genetic drift, selection and the evolution of the mutation rate. Nat. Rev. Genet. 17, 704–714. doi:10.1038/nrg.2016.104

Mansfield, M.L., 1994. Are there knots in proteins? Nat. Struct. Mol. Biol. 1, 213–214. doi:10.1038/nsb0494-213

Marchetti, J., Monzon, A.M., Tosatto, S.C.E., Parisi, G., Fornasari, M.S., 2019. Ensembles from Ordered and Disordered Proteins Reveal Similar Structural Constraints during Evolution. J. Mol. Biol. 431, 1298–1307. doi:10.1016/j.jmb.2019.01.031

Mészáros, B., Dobson, L., Fichó, E., Tusnády, G.E., Dosztányi, Z., Simon, I., 2019. Sequential, structural and functional properties of protein complexes are defined by how folding and binding intertwine. J. Mol. Biol. 431, 4408–4428. doi:10.1016/j.jmb.2019.07.034

Mészáros, B., Erdos, G., Dosztányi, Z., 2018. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. Nucleic Acids Res.

46, W329–W337. doi:10.1093/nar/gky384

Mirsky, A.E., Pauling, L., 1936. On the structure of native, denatured, and coagulated proteins. Proc Natl Acad Sci USA 22, 439–447. doi:10.1073/pnas.22.7.439

Mittal, A., Holehouse, A.S., Cohan, M.C., Pappu, R.V., 2018. Sequence-to-Conformation Relationships of Disordered Regions Tethered to Folded Domains of Proteins. J. Mol. Biol. 430, 2403–2421. doi:10.1016/j.jmb.2018.05.012

Mizianty, M.J., Stach, W., Chen, K., Kedarisetti, K.D., Disfani, F.M., Kurgan, L., 2010. Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources. Bioinformatics 26, i489-96. doi:10.1093/bioinformatics/btq373

Monod, J., Wyman, J., Changeux, J.P., 1965. On the nature of allosteric transitions: a plausible model. J. Mol. Biol. 12, 88–118. doi:10.1016/S0022-2836(65)80285-6

Monzon, A.M., Rohr, C.O., Fornasari, M.S., Parisi, G., 2016. CoDNaS 2.0: a comprehensive database of protein conformational diversity in the native state. Database (Oxford) 2016. doi:10.1093/database/baw038

Monzon, A.M., Zea, D.J., Fornasari, M.S., Saldaño, T.E., Fernandez-Alberti, S., Tosatto, S.C.E., Parisi, G., 2017a. Conformational diversity analysis reveals three functional mechanisms in proteins. PLoS Comput. Biol. 13, e1005398. doi:10.1371/journal.pcbi.1005398

Monzon, A.M., Zea, D.J., Marino-Buslje, C., Parisi, G., 2017b. Homology modeling in a dynamical world. Protein Sci. 26, 2195–2206. doi:10.1002/pro.3274

Motlagh, H.N., Wrabl, J.O., Li, J., Hilser, V.J., 2014. The ensemble nature of allostery. Nature 508, 331–339. doi:10.1038/nature13001

Necci, M., Piovesan, D., CAID Predictors, DisProt Curators, Tosatto, S.C.E., 2020. Critical assessment of protein intrinsic disorder prediction. BioRxiv. doi:10.1101/2020.08.11.245852

Necci, M., Piovesan, D., Dosztányi, Z., Tosatto, S.C.E., 2017. MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins. Bioinformatics 33, 1402–1404. doi:10.1093/bioinformatics/btx015

Necci, M., Piovesan, D., Tosatto, S.C.E., 2016. Large-scale analysis of intrinsic disorder flavors and associated functions in the protein sequence universe. Protein Sci. 25,

2164–2174. doi:10.1002/pro.3041

Nussinov, R., Ma, B., Tsai, C.-J., 2014. Multiple conformational selection and induced fit events take place in allosteric propagation. Biophys. Chem. 186, 22–30. doi:10.1016/j.bpc.2013.10.002

Nussinov, R., Ma, B., Tsai, C.-J., Csermely, P., 2013. Allosteric conformational barcodes direct signaling in the cell. Structure 21, 1509–1521. doi:10.1016/j.str.2013.06.002

Onuchic, J.N., Socci, N.D., Luthey-Schulten, Z., Wolynes, P.G., 1996. Protein folding funnels: the nature of the transition state ensemble. Fold. Des. 1, 441–450. doi:10.1016/S1359-0278(96)00060-0

Ota, M., Koike, R., Amemiya, T., Tenno, T., Romero, P.R., Hiroaki, H., Dunker, A.K., Fukuchi, S., 2013. An assignment of intrinsically disordered regions of proteins based on NMR structures. J. Struct. Biol. 181, 29–36. doi:10.1016/j.jsb.2012.10.017

Ovchinnikov, S., Kamisetty, H., Baker, D., 2014. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. elife 3, e02030. doi:10.7554/eLife.02030

Overington, J., Donnelly, D., Johnson, M.S., Sali, A., Blundell, T.L., 1992. Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. Protein Sci. 1, 216–226. doi:10.1002/pro.5560010203

Overington, J., Johnson, M.S., Sali, A., Blundell, T.L., 1990. Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. Proc. Biol. Sci. 241, 132–145. doi:10.1098/rspb.1990.0077

Palopoli, N., 2011. Desarrollo de un métododе validación de estructura terciaria de proteínascon base evolutivay su aplicación en la caracterización de almidón sintasas (Doctoral dissertation). Structural Bioinformatics Group. Universidad Nacional de Quilmes.

Palopoli, N., Lanzarotti, E., Parisi, G., 2013. BeEP Server: Using evolutionary information for quality assessment of protein structure models. Nucleic Acids Res. 41, W398-405. doi:10.1093/nar/gkt453

Pancsa, R., Schad, E., Tantos, A., Tompa, P., 2019. Emergent functions of proteins in non-stoichiometric supramolecular assemblies. Biochim. Biophys. Acta Proteins Proteom. 1867, 970–979. doi:10.1016/j.bbapap.2019.02.007

Pancsa, R., Zsolyomi, F., Tompa, P., 2018. Co-Evolution of Intrinsically Disordered Proteins with Folded Partners Witnessed by Evolutionary Couplings. Int. J. Mol. Sci. 19. doi:10.3390/ijms19113315

Parisi, G., 2001. Desarrollo de un modelo de evolución molecularcon restricciones estructurales (Doctoral dissertation). Universidad Nacional de Quilmes.

Parisi, G., Echave, J., 2001. Structural constraints and emergence of sequence patterns in protein evolution. Mol. Biol. Evol. 18, 750–756. doi:10.1093/oxfordjournals.molbev.a003857

Parisi, G., Echave, J., 2005. Generality of the structurally constrained protein evolution model: assessment on representatives of the four main fold classes. Gene 345, 45–53. doi:10.1016/j.gene.2004.11.025

Parisi, G., Zea, D.J., Monzon, A.M., Marino-Buslje, C., 2015. Conformational diversity and the emergence of sequence signatures during evolution. Curr. Opin. Struct. Biol. 32, 58–65. doi:10.1016/j.sbi.2015.02.005

Parsiegla, G., Reverbel-Leroy, C., Tardif, C., Belaich, J.P., Driguez, H., Haser, R., 2000. Crystal structures of the cellulase Cel48F in complex with inhibitors and substrates give insights into its processive action. Biochemistry 39, 11238–11246. doi:10.1021/bi001139p

Pauling, L., 1940. A theory of the structure and process of formation of antibodies*. J. Am. Chem. Soc. 62, 2643–2657. doi:10.1021/ja01867a018

Pauling, L., Corey, R.B., Branson, H.R., 1951. The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. Proc. Natl. Acad. Sci. USA 37, 205–211. doi:10.1073/pnas.37.4.205

Peng, Z., Yan, J., Fan, X., Mizianty, M.J., Xue, B., Wang, K., Hu, G., Uversky, V.N., Kurgan, L., 2015. Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. Cell. Mol. Life Sci. 72, 137–151. doi:10.1007/s00018-014-1661-9

Pereira de Araujo, A.F., Onuchic, J.N., 2009. A sequence-compatible amount of native burial information is sufficient for determining the structure of small globular proteins. Proc Natl Acad Sci USA 106, 19001–19004. doi:10.1073/pnas.0910851106

Perutz, M.F., 1976. Structure and mechanism of haemoglobin. Br. Med. Bull. 32, 195–

208. doi:10.1093/oxfordjournals.bmb.a071363

Perutz, M.F., Bolton, W., Diamond, R., Muirhead, H., Watson, H.C., 1964. Structure of haemoglobin. an x-ray examination of reduced horse haemoglobin. Nature 203, 687–690. doi:10.1038/203687a0

Perutz, M.F., Kendrew, J.C., Watson, H.C., 1965. Structure and function of haemoglobin: II. Some relations between polypeptide chain configuration and amino acid sequence. J. Mol. Biol. 13, 669–678. doi:10.1016/S0022-2836(65)80134-6

Piovesan, D., Minervini, G., Tosatto, S.C.E., 2016. The RING 2.0 web server for high quality residue interaction networks. Nucleic Acids Res. 44, W367-74. doi:10.1093/nar/gkw315

Piovesan, D., Tabaro, F., Paladin, L., Necci, M., Micetic, I., Camilloni, C., Davey, N., Dosztányi, Z., Mészáros, B., Monzon, A.M., Parisi, G., Schad, E., Sormanni, P., Tompa, P., Vendruscolo, M., Vranken, W.F., Tosatto, S.C.E., 2018. MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins. Nucleic Acids Res. 46, D471–D476. doi:10.1093/nar/gkx1071

Piovesan, D., Tosatto, S.C.E., 2018. Mobi 2.0: an improved method to define intrinsic disorder, mobility and linear binding regions in protein structures. Bioinformatics 34, 122–123. doi:10.1093/bioinformatics/btx592

Pond, S.L.K., Frost, S.D.W., Muse, S.V., 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21, 676–679. doi:10.1093/bioinformatics/bti079

Popovych, N., Sun, S., Ebright, R.H., Kalodimos, C.G., 2006. Dynamically driven protein allostery. Nat. Struct. Mol. Biol. 13, 831–838. doi:10.1038/nsmb1132

Prakash, A., Shin, J., Rajan, S., Yoon, H.S., 2016. Structural basis of nucleic acid recognition by FK506-binding protein 25 (FKBP25), a nuclear immunophilin. Nucleic Acids Res. 44, 2909–2925. doi:10.1093/nar/gkw001

Prilusky, J., Felder, C.E., Zeev-Ben-Mordehai, T., Rydberg, E.H., Man, O., Beckmann, J.S., Silman, I., Sussman, J.L., 2005. FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. Bioinformatics 21, 3435–3438. doi:10.1093/bioinformatics/bti537

Protter, D.S.W., Rao, B.S., Van Treeck, B., Lin, Y., Mizoue, L., Rosen, M.K., Parker, R., 2018. Intrinsically disordered regions can contribute promiscuous interactions to

RNP granule assembly. Cell Rep. 22, 1401–1412. doi:10.1016/j.celrep.2018.01.036

Ptitsyn, O.B., 1974. Invariant features of globin primary structure and coding of their secondary structure. J. Mol. Biol. 88, 287–300. doi:10.1016/0022-2836(74)90482-3

Pupko, T., Bell, R.E., Mayrose, I., Glaser, F., Ben-Tal, N., 2002. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics 18 Suppl 1, S71-7. doi:10.1093/bioinformatics/18.suppl_1.s71

Reisen, F., Weisel, M., Kriegl, J.M., Schneider, G., 2010. Self-organizing fuzzy graphs for structure-based comparison of protein pockets. J. Proteome Res. 9, 6498–6510. doi:10.1021/pr100719n

Robinson, M.A., Park, S., Sun, Z.-Y.J., Silver, P.A., Wagner, G., Hogle, J.M., 2005. Multiple conformations in the ligand-binding site of the yeast nuclear pore-targeting domain of Nup116p. J. Biol. Chem. 280, 35723–35732. doi:10.1074/jbc.M505068200

Romero, P., Obradovic, Z., Kissinger, C.R., Villafranca, J.E., Garner, E., Guilliot, S., Dunker, A.K., 1998. Thousands of proteins likely to have long disordered regions. Pac. Symp. Biocomput. 437–448.

Rose, P.W., Prlić, A., Bi, C., Bluhm, W.F., Christie, C.H., Dutta, S., Green, R.K., Goodsell, D.S., Westbrook, J.D., Woo, J., Young, J., Zardecki, C., Berman, H.M., Bourne, P.E., Burley, S.K., 2015. The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. Nucleic Acids Res. 43, D345-56. doi:10.1093/nar/gku1214

Russell, R.B., Barton, G.J., 1992. Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. Proteins 14, 309–323. doi:10.1002/prot.340140216

Saldaño, T.E., Monzon, A.M., Parisi, G., Fernandez-Alberti, S., 2016. Evolutionary conserved positions define protein conformational diversity. PLoS Comput. Biol. 12, e1004775. doi:10.1371/journal.pcbi.1004775

Samsudin, F., Gan, S.K.-E., Bond, P.J., 2020. The Structural Basis for Gag Non-Cleavage Site Mutations in Determining HIV-1 Viral Fitness. BioRxiv. doi:10.1101/2020.07.05.188326

Sathyapriya, R., Duarte, J.M., Stehr, H., Filippis, I., Lappe, M., 2009. Defining an essence of structure determining residue contacts in proteins. PLoS Comput. Biol. 5, e1000584. doi:10.1371/journal.pcbi.1000584

Schad, E., Fichó, E., Pancsa, R., Simon, I., Dosztányi, Z., Mészáros, B., 2018. DIBS: a repository of disordered binding sites mediating interactions with ordered proteins. Bioinformatics 34, 535–537. doi:10.1093/bioinformatics/btx640

Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A.W.R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D.T., Silver, D., Kavukcuoglu, K., Hassabis, D., 2020. Improved protein structure prediction using potentials from deep learning. Nature 577, 706–710. doi:10.1038/s41586-019-1923-7

Serohijos, A.W.R., Shakhnovich, E.I., 2014. Merging molecular mechanism and evolution: theory and computation at the interface of biophysics and evolutionary population genetics. Curr. Opin. Struct. Biol. 26, 84–91. doi:10.1016/j.sbi.2014.05.005

Sharir-Ivry, A., Xia, Y., 2017. The impact of native state switching on protein sequence evolution. Mol. Biol. Evol. 34, 1378–1390. doi:10.1093/molbev/msx071

Shimojo, H., Kawaguchi, A., Oda, T., Hashiguchi, N., Omori, S., Moritsugu, K., Kidera, A., Hiragami-Hamada, K., Nakayama, J.-I., Sato, M., Nishimura, Y., 2016. Extended string-like binding of the phosphorylated HP1α N-terminal tail to the lysine 9-methylated histone H3 tail. Sci. Rep. 6, 22527. doi:10.1038/srep22527

Shin, Y., Brangwynne, C.P., 2017. Liquid phase condensation in cell physiology and disease. Science 357. doi:10.1126/science.aaf4382

Siltberg-Liberles, J., Grahnen, J.A., Liberles, D.A., 2011. The evolution of protein structures and structural ensembles under functional constraint. Genes (Basel) 2, 748–762. doi:10.3390/genes2040748

Simon, A.L., Stone, E.A., Sidow, A., 2002. Inference of functional regions in proteins by quantification of evolutionary constraints. Proc Natl Acad Sci USA 99, 2912–2917. doi:10.1073/pnas.042692299

Sormanni, P., Piovesan, D., Heller, G.T., Bonomi, M., Kukic, P., Camilloni, C., Fuxreiter, M., Dosztanyi, Z., Pappu, R.V., Babu, M.M., Longhi, S., Tompa, P., Dunker, A.K.,

Uversky, V.N., Tosatto, S.C.E., Vendruscolo, M., 2017. Simultaneous quantification of protein order and disorder. Nat. Chem. Biol. 13, 339–342. doi:10.1038/nchembio.2331

Spolar, R.S., Record, M.T., 1994. Coupling of local folding to site-specific binding of proteins to DNA. Science 263, 777–784. doi:10.1126/science.8303294

Spronk, C.A.E.M., Linge, J.P., Hilbers, C.W., Vuister, G.W., 2002. Improving the quality of protein structures derived by NMR spectroscopy. J. Biomol. NMR 22, 281–289. doi:10.1023/a:1014971029663

Spronk, C.A.E.M., Nabuurs, S.B., Bonvin, A.M.J.J., Krieger, E., Vuister, G.W., Vriend, G., 2003. The precision of NMR structure ensembles revisited. J. Biomol. NMR 25, 225–234. doi:10.1023/a:1022819716110

Surkont, J., Pereira-Leal, J.B., 2015. Evolutionary patterns in coiled-coils. Genome Biol. Evol. 7, 545–556. doi:10.1093/gbe/evv007

Sydykova, D.K., Wilke, C.O., 2017. Calculating site-specific evolutionary rates at the amino-acid or codon level yields similar rate estimates. PeerJ 5, e3391. doi:10.7717/peerj.3391

The UniProt Consortium, 2017. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 45, D158–D169. doi:10.1093/nar/gkw1099

Thorne, J.L., 2000. Models of protein sequence evolution and their applications. Curr. Opin. Genet. Dev. 10, 602–605. doi:10.1016/S0959-437X(00)00142-8

Tokuriki, N., Tawfik, D.S., 2009. Protein dynamism and evolvability. Science 324, 203–207. doi:10.1126/science.1169375

Tompa, P., 2002. Intrinsically unstructured proteins. Trends Biochem. Sci. 27, 527–533. doi:10.1016/s0968-0004(02)02169-2

Tompa, P., 2011. Unstructural biology coming of age. Curr. Opin. Struct. Biol. 21, 419–425. doi:10.1016/j.sbi.2011.03.012

Toth-Petroczy, A., Palmedo, P., Ingraham, J., Hopf, T.A., Berger, B., Sander, C., Marks, D.S., 2016. Structured States of Disordered Proteins from Genomic Sequences. Cell 167, 158-170.e12. doi:10.1016/j.cell.2016.09.010

Tóth-Petróczy, A., Tawfik, D.S., 2011. Slow protein evolutionary rates are dictated by surface-core association. Proc Natl Acad Sci USA 108, 11151–11156.

doi:10.1073/pnas.1015994108

Tourasse, N.J., Li, W.H., 2000. Selective constraints, amino acid composition, and the rate of protein evolution. Mol. Biol. Evol. 17, 656–664. doi:10.1093/oxfordjournals.molbev.a026344

Touw, W.G., Baakman, C., Black, J., te Beek, T.A.H., Krieger, E., Joosten, R.P., Vriend, G., 2015. A series of PDB-related databanks for everyday needs. Nucleic Acids Res. 43, D364-8. doi:10.1093/nar/gku1028

Trabi, M., Craik, D.J., 2002. Circular proteins — no end in sight. Trends Biochem. Sci. 27, 132–138. doi:10.1016/S0968-0004(02)02057-1

Tsai, C.-J., Nussinov, R., 2014. A unified view of “how allostery works”. PLoS Comput. Biol. 10, e1003394. doi:10.1371/journal.pcbi.1003394

Tsai, C.J., Kumar, S., Ma, B., Nussinov, R., 1999. Folding funnels, binding funnels, and protein function. Protein Sci. 8, 1181–1190. doi:10.1110/ps.8.6.1181

Turoverov, K.K., Kuznetsova, I.M., Uversky, V.N., 2010. The protein kingdom extended: ordered and intrinsically disordered proteins, their folding, supramolecular complex formation, and aggregation. Prog. Biophys. Mol. Biol. 102, 73–84. doi:10.1016/j.pbiomolbio.2010.01.003

Uversky, V.N., Oldfield, C.J., Midic, U., Xie, H., Xue, B., Vucetic, S., Iakoucheva, L.M., Obradovic, Z., Dunker, A.K., 2009. Unfoldomics of human diseases: linking protein intrinsic disorder with diseases. BMC Genomics 10 Suppl 1, S7. doi:10.1186/1471-2164-10-S1-S7

van der Lee, R., Buljan, M., Lang, B., Weatheritt, R.J., Daughdrill, G.W., Dunker, A.K., Fuxreiter, M., Gough, J., Gsponer, J., Jones, D.T., Kim, P.M., Kriwacki, R.W., Oldfield, C.J., Pappu, R.V., Tompa, P., Uversky, V.N., Wright, P.E., Babu, M.M., 2014. Classification of intrinsically disordered regions and proteins. Chem. Rev. 114, 6589–6631. doi:10.1021/cr400525m

Varadi, M., Kosol, S., Lebrun, P., Valentini, E., Blackledge, M., Dunker, A.K., Felli, I.C., Forman-Kay, J.D., Kriwacki, R.W., Pierattelli, R., Sussman, J., Svergun, D.I., Uversky, V.N., Vendruscolo, M., Wishart, D., Wright, P.E., Tompa, P., 2014. pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins. Nucleic Acids Res. 42, D326-35. doi:10.1093/nar/gkt960

Volpon, L., D'Orso, I., Young, C.R., Frasch, A.C., Gehring, K., 2005. NMR structural study of TcUBP1, a single RRM domain protein from Trypanosoma cruzi: contribution of a beta hairpin to RNA binding. Biochemistry 44, 3708–3717. doi:10.1021/bi047450e

Vullo, A., Bortolami, O., Pollastri, G., Tosatto, S.C.E., 2006. Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines. Nucleic Acids Res. 34, W164-8. doi:10.1093/nar/gkl166

Walsh, I., Martin, A.J.M., Di Domenico, T., Tosatto, S.C.E., 2012. ESpritz: accurate and fast prediction of protein disorder. Bioinformatics 28, 503–509. doi:10.1093/bioinformatics/btr682

Walsh, I., Martin, A.J.M., Di Domenico, T., Vullo, A., Pollastri, G., Tosatto, S.C.E., 2011. CSpritz: accurate prediction of protein disorder segments with annotation for homology, secondary structure and linear motifs. Nucleic Acids Res. 39, W190-6. doi:10.1093/nar/gkr411

Wang, S., Ma, J., Xu, J., 2016. AUCpreD: proteome-level protein disorder prediction by AUC-maximized deep convolutional neural fields. Bioinformatics 32, i672–i679. doi:10.1093/bioinformatics/btw446

Wei, G., Xi, W., Nussinov, R., Ma, B., 2016. Protein ensembles: how does nature harness thermodynamic fluctuations for life? the diverse functional roles of conformational ensembles in the cell. Chem. Rev. 116, 6516–6551. doi:10.1021/acs.chemrev.5b00562

Whelan, S., Goldman, N., 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. 18, 691–699. doi:10.1093/oxfordjournals.molbev.a003851

Wood, T.C., Pearson, W.R., 1999. Evolution of protein sequences and structures. J. Mol. Biol. 291, 977–995. doi:10.1006/jmbi.1999.2972

Worth, C.L., Gong, S., Blundell, T.L., 2009. Structural and functional constraints in the evolution of protein families. Nat. Rev. Mol. Cell Biol. 10, 709–720. doi:10.1038/nrm2762

Wright, P.E., Dyson, H.J., 1999. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J. Mol. Biol. 293, 321–331. doi:10.1006/jmbi.1999.3110

Xia, X., Li, W.H., 1998. What amino acid properties affect protein evolution? J. Mol. Evol. 47, 557–564. doi:10.1007/pl00006412

Xie, H., Vucetic, S., Iakoucheva, L.M., Oldfield, C.J., Dunker, A.K., Uversky, V.N., Obradovic, Z., 2007. Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. J. Proteome Res. 6, 1882–1898. doi:10.1021/pr060392u

Xue, B., Dunbrack, R.L., Williams, R.W., Dunker, A.K., Uversky, V.N., 2010. PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. Biochim. Biophys. Acta 1804, 996–1010. doi:10.1016/j.bbapap.2010.01.011

Xue, B., Dunker, A.K., Uversky, V.N., 2012. Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. J. Biomol. Struct. Dyn. 30, 137–149. doi:10.1080/07391102.2012.675145

Yang, 2006. Computational Molecular Evolution.

Yang, Z., 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39, 306–314. doi:10.1007/BF00160154

Yang, Z., 1998. On the best evolutionary rate for phylogenetic analysis. Syst. Biol. 47, 125–133. doi:10.1080/106351598261067

Yeh, S.-W., Huang, T.-T., Liu, J.-W., Yu, S.-H., Shih, C.-H., Hwang, J.-K., Echave, J., 2014a. Local packing density is the main structural determinant of the rate of protein sequence evolution at site level. Biomed Res. Int. 2014, 572409. doi:10.1155/2014/572409

Yeh, S.-W., Liu, J.-W., Yu, S.-H., Shih, C.-H., Hwang, J.-K., Echave, J., 2014b. Site-specific structural constraints on protein sequence evolutionary divergence: local packing density versus solvent exposure. Mol. Biol. Evol. 31, 135–139. doi:10.1093/molbev/mst178

Zea, D.J., 2015. Evolución de proteínas y diversidad conformacional (Doctoral dissertation). Grupo de Bioinformatica Estructural. Universidad Nacional de Quilmes.

Zea, D.J., Anfossi, D., Nielsen, M., Marino-Buslje, C., 2017. MIToS.jl: mutual information tools for protein sequence analysis in the Julia language. Bioinformatics 33, 564–565. doi:10.1093/bioinformatics/btw646

Zea, D.J., Miguel Monzon, A., Fornasari, M.S., Marino-Buslje, C., Parisi, G., 2013. Protein conformational diversity correlates with evolutionary rate. Mol. Biol. Evol. 30, 1500–1503. doi:10.1093/molbev/mst065

Zea, D.J., Monzon, A.M., Gonzalez, C., Fornasari, M.S., Tosatto, S.C.E., Parisi, G., 2016. Disorder transitions and conformational diversity cooperatively modulate biological function in proteins. Protein Sci. 25, 1138–1146. doi:10.1002/pro.2931

Zea, D.J., Monzon, A.M., Parisi, G., Marino-Buslje, C., 2018. How is structural divergence related to evolutionary information? Mol. Phylogenet. Evol. 127, 859–866. doi:10.1016/j.ympev.2018.06.033

Zhang, T., Faraggi, E., Li, Z., Zhou, Y., 2017. Intrinsic Disorder and Semi-disorder Prediction by SPINE-D. Methods Mol. Biol. 1484, 159–174. doi:10.1007/978-1-4939-6406-2_12

Zhou, A., Rohou, A., Schep, D.G., Bason, J.V., Montgomery, M.G., Walker, J.E., Grigorieff, N., Rubinstein, J.L., 2015. Structure and conformational states of the bovine mitochondrial ATP synthase by cryo-EM. elife 4, e10180. doi:10.7554/eLife.10180

Zuckerkandl, E., Pauling, L., 1965. Molecules as documents of evolutionary history. J. Theor. Biol. 8, 357–366. doi:10.1016/0022-5193(65)90083-4