Monzón, Alexander Miguel

Estudio de los factores biológicos y fisicoquímicos que modulan la diversidad conformacional del estado nativo proteico

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

Estudio de los factores biológicos y fisicoquímicos que modulan la diversidad conformacional del estado nativo proteico

TESIS DOCTORAL

Alexander Miguel Monzon

monzon.alexander@gmail.com

Resumen

El concepto de estado nativo ha evolucionado en el tiempo junto a las técnicas de caracterización de proteínas. Hoy en día el estado nativo se describe por lo que denominamos ensamble, el cual está compuesto por confórmeros que se encuentran en equilibrio dinámico. Los cambios conformacionales que experimenta una proteína en su estado nativo es lo que se conoce como diversidad conformacional, y es un concepto clave para la comprensión de procesos como el cooperativismo y el alosterismo, la función biológica, el reconocimiento molecular, la catálisis enzimática y el origen de nuevas funciones entre otros tantos ejemplos. De esta forma, creemos firmemente que la relación estructura-función de las proteínas debiera reformularse en términos dinámicos, esto es, que se considere la diversidad conformacional para llegar a conclusiones biológicas más relevantes.

El presente trabajo se centra en comprender y explorar la diversidad conformacional de proteínas y su estrecha relación con la función biológica. Para este propósito hemos desarrollado una base de datos de diversidad conformacional denominada CoDNaS (\Conformational Diversity of the Native State") la cual permite estudiar la diversidad conformacional a partir de las estructuras de proteínas redundantes depositadas en PDB y que han sido obtenidas en distintas condiciones experimentales. Utilizando los varios miles de proteínas depositadas y anotadas en CoDNaS, derivamos en distintos estudios conceptos generales sobre el comportamiento de la diversidad conformacional en función de distintos aspectos biológicos, sicoquímicos y biofísicos. Así, hemos podido explicar la distribución global de diversidad conformacional, identificando al menos cuatro relaciones estructura-dinámica-función que emergen del análisis del ensamble conformacional que posee cada proteína. Estas características compartidas entre los diferentes grupos podrán representar mecanismos conformacionales relacionados con sus funciones biológicas. Además, re-examinamos la muy bien establecida relación entre divergencia estructural y divergencia secuencial a la luz de la diversidad conformacional.

Encontramos que la consideración de la diversidad conformacional impacta fuertemente en la correlación de la relación secuencia-estructura lo que repercute directamente en la contabilidad de los métodos de modelado por homología.

Nuestros resultados ofrecen nuevas perspectivas dentro de la bioinformática y biología estructural de proteínas. Esperamos que los mismos redunden en una mejor comprensión de los mecanismos subyacentes a la función proteica como así también en la generación de herramientas bioinformáticas más precisas.

Palabras claves: diversidad conformacional, estado nativo, ensamble conformacional, función proteica, dinámica de proteínas, bioinformática.

Universidad Nacional de Quilmes

Departamento de Ciencia y Tecnología

Estudio de los factores biológicos y fisicoquímicos que modulan la diversidad conformacional del estado nativo proteico

Lic. Alexander Miguel Monzon

Tesis a presentar para optar por el título de Doctor de la Universidad Nacional de Quilmes

Director de tesis: Dr. Gustavo Parisi

Co-Director de tesis: Dra. Silvina Fornasari

Consejero de estudios: Dr. Adolfo Iribarren

Lugar de trabajo:

Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes.

Bernal, 2017

Índice general

Índice de figuras

Abreviaciones

PDB . . . . . *Protein Data Bank*

RMN . . . . . *Resonancia Magnética Nuclear*

DRX . . . . . *Difracción por Rayos-X*

RMSD . . . . *Cα "Root Mean Square Deviation" o Desviación cuadrática media*

IDPs . . . . . *"Intrinsically Disordered Proteins" o Proteínas intrínsecamente desordenadas*

IDRs . . . . . *"Intrinsically Disordered Regions" o Regiones intrínsecamente desordenadas*

CoDNaS . . . *Base de datos "Conformational Diversity of the Native State"*

TBM . . . . . *"Template-based modeling"*

NCBI . . . . . *"National Center for Biotechnology Information"*

Å . . . . . *Angstroms*

Agradecimientos

A Gustavo, mi director, papá, amigo, que me ha transmitido y contagiado la pasión por investigar. Gracias por todo lo que me has enseñado estos años, no solo a nivel científico sino también como persona, por haber sido mi papá científico y mi segundo padre más cercano en Buenos Aires. Gracias por abrirme las puertas del grupo desde que ingresé hace ya 6 años. Gracias por siempre alentarme a seguir adelante y darme la libertad de poder decidir y elegir durante este proceso. Gracias por tu calidez humana ante todo, y por brindarme todas las herramientas necesarias para poder desarrollar mi carrera científica. Siempre te voy a estar agradecido.

A Silvina, nuestra mamá dentro del grupo, gracias por preocuparte por nosotros, por estar cuando te necesité y por ser la gran persona que sos.

A Dieguito Zea, mi hermanito científico, gracias por estar ahí cuando recién comencé este camino. Gracias por todo lo que me enseñaste, las horas de charlas y discusiones. Todo este proceso no hubiese sido igual sin nuestra amistad, colaboraciones y trabajos en conjunto.

A Cristina Marino-Buslje, gracias por todos estos años de intensa colaboración y aprendizaje juntos. Gracias por estar siempre más allá de lo científico y poder contar con tu apoyo.

A mis compañeros del grupo SBG de la UNQ, a los más viejitos, Marcia y Nico. Gracias por su compañía incondicional, charlas, cervezas, comidas, por ser mis amigos más allá de colegas. A los más nuevos, Julia, Guille, Ana y Cristian. Gracias por siempre confiar en mí. Estoy seguro que todo este camino no hubiese sido igual sin su compañía.

A nuestro grupo hermano comandado por Sebastián Fernández-Alberti. Gracias Seba por ser siempre un buen consejero, por tu sentido del humor y por tu mirada científica que fue de gran importancia en nuestros trabajos. Gracias Tade y Pato, por su compañía, amistad y divertidas charlas, y a "los cubanos" por traer el caribe a la oficina.

A mis colegas amigos, con los que vengo compartiendo esta profesión desde que comenzamos allá en Oro Verde y hoy seguimos transitando junto este camino en Buenos Aires y

algunos otros por el mundo. Gracias por todos los buenos momentos compartidos y por estar siempre que los necesité. A Gonza, por ser siempre un buen consejero y amigo, por los viajes y charlas compartidas y por motivarme siempre a sacar un poco más de mí.

Al Consejo de estudiantes de la ISCB y al RSG-Argentina. Gracias por todos estos años de compartir proyectos juntos, organizar simposios, dictar workshops y por enseñarme lo que es trabajar en equipo.

A la A2B2C, por abrirme las puertas y darme un lugar para poder llevar acabo distintas actividades, organización de eventos y estar siempre predispuestos a colaborar con los estudiantes. Gracias por darme confianza.

Al grupo *Biocomputing UP* dirigido por Silvio Tosatto. Gracias por haberme recibido siempre tan cordialmente y por haberme dejado colaborar con Uds. durante estos años.

A mis amigos de toda la vida de Paraná, "los pibes" y Juanma. Gracias por estar siempre a pesar de la distancia, por los asados, cervezas, salidas y charlas compartidas, que nunca faltan cuando estoy en la ciudad.

A Iara, el amor de mi vida. Gracias por ser mi pie de apoyo durante estos 5 años, por entenderme y apoyarme siempre, por darme buenos consejos y confianza cuando la necesité. Gracias por sobre todas las cosas por brindarme tu amor incondicional todo este tiempo y por ser la gran persona que sos. Te amo muchísimo, sin tu presencia nada hubiese sido igual.

A mi familia, mis padres y hermana. Gracias por enseñarme lo que es el esfuerzo y la dedicación, por siempre depositar su confianza en mí, por apoyarme en éste camino que emprendí en Buenos Aires hace 5 años, como así también en todos mis proyectos. Gracias por estar ahí siempre que los necesité y ser mis mayores maestros en esta vida. Sin Uds. hoy no estaría acá y no sería la persona que soy. Gracias por todo viejos! Los amo!

A la memoria de mi abuela Angélica, que partió cuando comencé este nuevo ciclo en mi vida, y a la de mi tío Rubén. Gracias por estar siempre presentes y por todas las enseñanzas que me dejaron, llevo conmigo los mejores recuerdos juntos.

Resumen

El concepto de estado nativo ha evolucionado en el tiempo junto a las técnicas de caracterización de proteínas. Hoy en día el estado nativo se describe por lo que denominamos ensamble, el cual está compuesto por confórmeros que se encuentran en equilibrio dinámico. Los cambios conformacionales que experimenta una proteína en su estado nativo es lo que se conoce como diversidad conformacional, y es un concepto clave para la comprensión de procesos como el cooperativismo y el alosterismo, la función biológica, el reconocimiento molecular, la catálisis enzimática y el origen de nuevas funciones entre otros tantos ejemplos. De esta forma, creemos firmemente que la relación estructura-función de las proteínas debiera reformularse en términos dinámicos, esto es, que se considere la diversidad conformacional para llegar a conclusiones biológicas más relevantes.

El presente trabajo se centra en comprender y explorar la diversidad conformacional de proteínas y su estrecha relación con la función biológica. Para este propósito hemos desarrollado una base de datos de diversidad conformacional denominada CoDNaS (*"Conformational Diversity of the Native State"*) la cual permite estudiar la diversidad conformacional a partir de las estructuras de proteínas redundantes depositadas en PDB y que han sido obtenidas en distintas condiciones experimentales. Utilizando los varios miles de proteínas depositadas y anotadas en CoDNaS, derivamos en distintos estudios conceptos generales sobre el comportamiento de la diversidad conformacional en función de distintos aspectos biológicos, fisicoquímicos y biofísicos. Así, hemos podido explicar la distribución global de diversidad conformacional, identificando al menos cuatro relaciones estructura-dinámica-función que emergen del análisis del ensamble conformacional que posee cada proteína. Estas características compartidas entre los diferentes grupos podrían representar mecanismos conformacionales relacionados con sus funciones biológicas. Además, re-examinamos la muy bien establecida relación entre entre divergencia estructural y divergencia secuencial a la luz de la diversidad conformacional. Encontramos que la consideración de la diversidad conformacional impacta fuertemente en la correlación de la relación secuencia-estructura lo que repercute directamente en la confiabilidad de los métodos de modelado por homología.

Nuestro resultados ofrecen nuevas perspectivas dentro de la bioinformática y biología es-

tructural de proteínas. Esperamos que los mismos redunden en una mejor comprensión de los mecanismos subyacentes a la función proteica como así también en la generación de herramientas bioinformáticas más precisas.

Palabras claves: diversidad conformacional, estado nativo, ensamble conformacional, función proteica, dinámica de proteínas, bioinformática.

Publicaciones

Como resultado de esta tesis doctoral se han publicado los siguiente artículos académicos.

En el marco de esta tesis:

1. *Protein conformational diversity correlates with evolutionary rate.* Zea, D., **Monzon, A. M.**, Fornasari, M. S., Marino-Buslje, C., Parisi, G. Molecular Biology and Evolution, 2013, 30(7), 1500–1503. https://doi.org/10.1093/molbev/mst065

2. *CoDNaS: a database of Conformational Diversity in the native state of proteins.* **Monzon, A. M.**, Juritz, E., Fornasari, M. S., Parisi, G. Bioinformatics, 2013, doi: 10.1093/bioinformatics/btt405.

3. *Conformational diversity and the emergence of sequence signatures during evolution* Parisi G., Zea, D., **Monzon, A. M.**, Marino-Buslje, C. Current Opinion in Structural Biology, vol. 32, ISSN: 0959-440X, doi:10.1016/j.sbi.2015.02.005.

4. *Evolutionary Conserved Positions Define Protein Conformational Diversity.* Saldaño, T. E., **Monzon, A. M.**, Parisi, G., Fernandez-Alberti, S. PLoS Computational Biology, 2016, 12(3), e1004775. https://doi.org/10.1371/journal.pcbi.1004775

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

6. *Disorder transitions and conformational diversity cooperatively modulate biological function in proteins.* Zea, D., **Monzon, A. M.**, Gonzalez, C., Fornasari, M. S., Tosatto, S., Parisi, G. Protein Science, 2016, 25(6):1138-46. doi: 10.1002/pro.2931.

7. *Addressing the Role of Conformational Diversity in Protein Structure Prediction.* **Monzon, A. M.**, Palopoli, N.,Fornasari, M. S., Parisi, G. PloS One, 2016, 11(5):e0154923, doi: 10.1371/journal.pone.0154923

8. *Conformational diversity analysis reveals three functional mechanisms in proteins.* **Monzon, A. M**, Zea D.J., Fornasari M.S., Saldaño T.E., Fernandez-Alberti S., Tosatto S.C.E., Parisi G. PLOS Computational Biology, 2017, 13(2): e1005398. doi: 10.1371/journal.pcbi.1005398

9. *On the dynamical incompleteness of the Protein Data Bank.* **Monzon, A. M.**, Marino-Buslje, C., Zea D.J., Fornasari M.S., Parisi G. Briefings in Bioinformatics. doi:10.1093/bib/bbx084, Agosto 2017.

10. *Homology Modeling in a dynamical World* **Monzon, A. M.**, Zea D.J., Marino-Buslje, C., Parisi G. Protein Science. doi:10.1002/pro.32742017. Agosto 2017.

11. *MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins.* Piovesan, D., Tabaro, F., Paladin, L., Necci, M., Mičetić, 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. Nucleic Acids Research. doi:10.1093/nar/gkx1071. Noviembre 2017.

12. *Large scale analysis of protein conformational transitions from aqueous to non-aqueous media.* Velez Rueda, A. J., **Monzon, A. M.**, Ardanaz A. M., Iglesias, L. E., Parisi, G. BMC Bioinformatics. 2da revisión. Noviembre 2017.

13. *How is structural divergence related to evolutionary information?.* Zea D.J., **Monzon, A. M.**, Marino-Buslje, C., Parisi G. Molecular Phylogenetics and Evolution. 2da revisión. Diciembre 2017.

Capítulos de libros:

1. *Exploring protein conformational diversity.* **Monzon, A. M.**, Fornasari, M. S., Zea, D. J., Parisi, G. Methods in Molecular Biology. Capítulo de libro. En prensa. Octubre 2017.

Artículos educacionales y de divulgación:

1. *Highlights of the 1st Argentine Symposium of Young Bioinformatics Researchers (1SAJIB) organized by the ISCB RSG-Argentina.* Parra R.G., Defelipe L.A., Guzovsky AB,

Monzon, A. M., Cravero F., Mancini E., Moreyra N., Padilla Franzotti C.L., Revuelta M.V., Freiberger M.I., Gonzalez N.N., Gonzalez G.A., Orts F., Stocchi N., Hasenahuer M.A., Teppa E., Zea D.J., Palopoli N. PeerJ, 2017.
Preprints:https://doi.org/10.7287/peerj.preprints.2494v1

2. *ISCB-Student Council Narratives: Strategical development of the ISCB-Regional Student Groups in 2016*. Shome S., Meysman P., Parra R.G., **Monzon, A. M.**, et al. F1000Research, 5(ISCB Comm J):2882. doi: 10.12688/f1000research.10420.1. Diembre 2016.

3. *A report on the "International Society for Computational Biology - Latin America (ISCB-LA)" Bioinformatics Conference 2016*. Palopoli, N., **Monzon, A. M.**, Parisi, G., Chernomoretz, A., Agüero, F. EMBnet.Journal, 2017, 23, e883.
doi:http://dx.doi.org/10.14806/ej.23.0.883

4. *Second ISCB Latin American Student Council Symposium (LA-SCS) 2016*. **Monzon, A. M.**, Hasenahuer, M.A., Mancini, E et al. F1000Research, 6(ISCB Comm J):1491. doi: 10.12688/f1000research.12321.1. Agosto 2017.

Artículos fuera del marco de la tesis:

1. *Seasonal variation in genetic population structure of sábalo (Prochilodus lineatus) in the Lower Uruguay River*. Rueda, E.C., Carriquiriborde, P., **Monzon, A. M.**, Somoza, G. M., Ortí G. Genetica, 2013, 141(7–9), 401–7. https://doi.org/10.1007/s10709-013-9739-0

Capítulo 1

Introducción general

1.1. Concepto del estado nativo proteico y la relación estructura-función

A principios de siglo XX, comenzó a acumularse conocimiento sobre cómo funcionaban las proteínas a pesar de que en ese momento no se tenía información de la existencia de la estructura primaria, no se conocían estructuras experimentales y aún no se habían descubierto todos los aminoácidos. Así, fueron Emil Fischer y Franz Hofmeister quienes co-descubrieron el enlace peptídico en 1902 explicando como se unían los aminoácidos para formar las proteínas, o como Albrecht Kossel predecía que la función de las proteínas estaría relacionada con el tipo y disposición espacial de los aminoácidos que la componen [Kossel, 1898]. Años más tarde, en 1936, Mirsky y Pauling definen al estado nativo de una proteína (con propiedades específicas) como una cadena polipeptídica contínua y no ramificada, que no se interrumpe a lo largo de la molécula (o, en algunos casos de dos o más cadenas de este tipo), que adopta un plegado con una conformación definida de forma única y estabilizada mediante enlaces de hidrógeno. Consideraban que las propiedades que poseían las proteínas en su estado nativo eran atribuidas a la conformación única que adquirían. Además, en este trabajo también explican el concepto de proteína desnaturalizada, el cual caracterizan como la pérdida de la conformación nativa [Mirsky and Paulin, 1936]. Pauling propone cuatro años más tarde, que los anticuerpos podrían adoptar diferentes conformaciones de energía similar para poder asu-

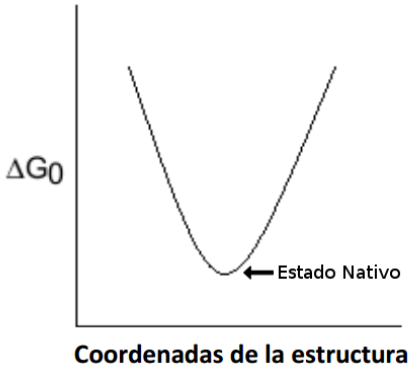

Figura 1.1: Paisaje de energía libre de las coordenadas de una estructura proteica en el espacio. El estado nativo se encuentra en el mínimo de energía y corresponde a una única estructura nativa.

mir la conformación complementaria al antígeno [Pauling, 1940]. Fred Karush extendió en 1950 este concepto al proponer que el estado nativo de las proteínas podría contener diferentes confórmeros con energías similares en equilibrio dinámico para explicar la heterogeneidad de unión de la seroalbúmina [Karush, 1950]. Él define como adaptabilidad configuracional al hecho que diferentes confórmeros pueden tener diferentes afinidades para los ligandos. Lamentablemente su trabajo fue mayormente ignorado a pesar de describir muy tempranamente las ideas que hoy día explican la relación estructura-función de las proteínas. Estos primeros estudios dentro de la bioquímica estructural de proteínas, proponían la idea de un estado nativo situado en un mínimo de energía con una conformación estructural única (o más de una con energía similar) (Figura 1.1). Se consideraba que la proteína se encontraba en un estado funcional cuando estaba plegada y no funcional cuando se perdía dicha estructura, es decir se desnaturalizaba.

En estos años empezaron a surgir diferentes modelos para explicar la relación entre la estructura-función de las proteínas. Uno de los primeros modelos propuesto por Emil Fischer [Fischer, 1894] fue el conocido como "llave y cerradura". Con la evidencia experimental que sostiene que muchas enzimas son sumamente específicas, Fischer propone en este modelo que la proteína adopta una conformación estructural única, que le confiere una alta especificidad para la unión con el ligando. Considera a la estructura de la proteína como una cerradura y al sustrato como a una llave que encaja de forma perfecta en dicha cerradura (Figura 1.2).

Años más tarde en 1958, Daniel Koshland [Koshland et al., 1958] observó que el modelo de "llave y cerradura" no explicaba ciertas discrepancias que se daban entre algunos tipos de

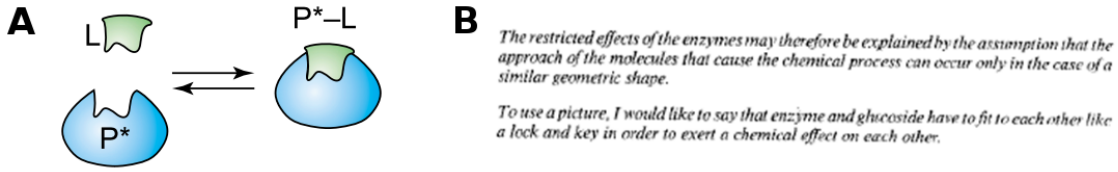

Figura 1.2: A) Representación esquemática del modelo "llave y cerradura". P* representa la conformación activa de la proteína, mientras que L es el ligando. Imagen extraída de [James and Tawfik, 2003]. B) Extracto del manuscrito original en donde Fischer define por primera vez el concepto de llave y cerradura [Fischer, 1894].

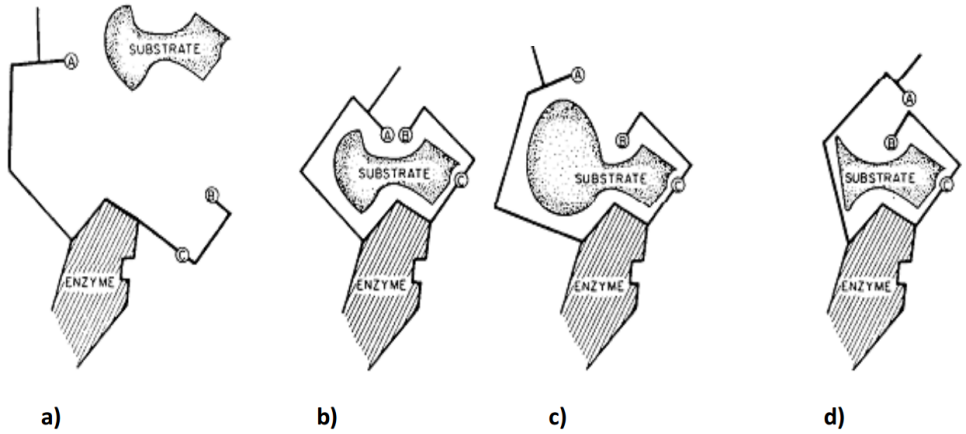

Figura 1.3: Modelo esquemático del "ajuste inducido". Las líneas negras indican las cadenas de la proteína que contienen grupos catalíticos A y B y un grupo de unión C. a) el substrato y la enzima se encuentran disociados. b) el sustrato induce un cambio conformacional en las cadenas de las proteínas para llevar a A y B a la alineación apropiada para la unión con el ligando. c) un grupo voluminoso adherido al substrato, evita la alineación correcta de A y B. d) la deleción del grupo elimina la acción yuxtapuesta sobre la cadena que contiene A, por lo que termodinámicamente el complejo tiene una alineación incorrecta de A y B. Imagen extraída de [Koshland et al., 1958]

reacciones químicas, tales como la inhibición no competitiva, la falta de actividad hidrolítica de las quinasas y principalmente la unión de ligandos de diferentes tamaños y formas. De esta manera surge el modelo conocido como "ajuste inducido" bajo las siguientes premisas: a) para la acción enzimática es requerida la orientación precisa de los grupos catalíticos, b) el substrato causa un cambio apreciable en la relación tridimensional de los aminoácidos del sitio activo, y c) los cambios en la estructura de la proteína causados por la unión con el substrato, posicionarán los grupos catalíticos en la alineación adecuada, mientras que si no es el sustrato indicado, este proceso no tendrá lugar [Koshland et al., 1958]. En la Figura 1.3 se ilustra este concepto y cómo este podrían explicar las anomalías observadas en el caso de presentarse un sustrato inadecuado.

Para demostrar esta nueva característica de flexibilidad, no tomada en cuenta en el modelo "llave cerradura", Koshland eligió la enzima fosfoglucomutasa (cuya reacción presenta

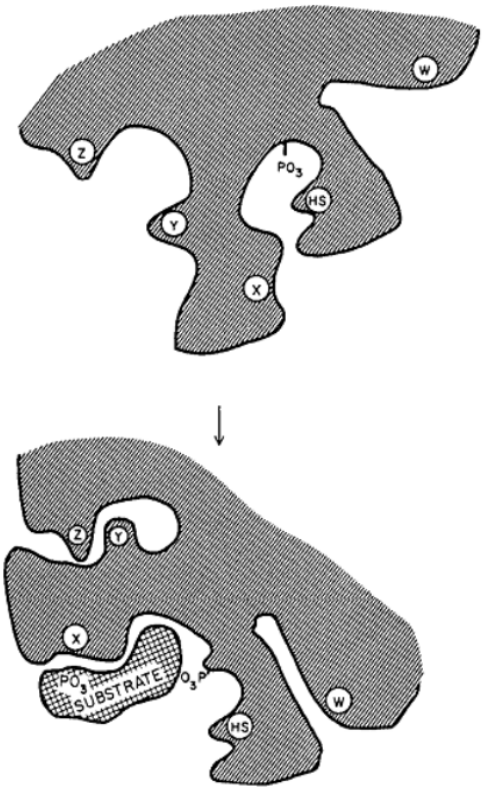

Figura 1.4: Ajuste inducido de la fosfoglucomutasa. En la parte superior podemos observar la enzima en ausencia del substrato. En la imagen inferior se muestra el cambio conformacional que lleva a la exposición del grupo –SH y esconde los grupos Z, Y, Z y W. Imagen extraída de [Koshland et al., 1958].

una alta similitud con la hexoquinasa y se esperaba que sea una enzima que realice el "ajuste inducido"). Sin conocer la estructura terciaria de la proteína, Koshland detectó que en presencia y ausencia del ligando existía una diferencia en el número de tioles libres titulables. El concepto de este experimento lo podemos observar en la Figura 1.4, donde la presencia de tioles libres titulables con posterioridad a la unión de sustrato manifiesta un cambio conformacional "inducido" por el mismo. Este resultado puso en evidencia ciertas inconsistencias en el modelo de "llave y cerradura" para explicar la relación estructura-función en proteínas. De esta manera, el modelo de "ajuste inducido", sin dejar de incluir la idea de la especificidad del sitio activo propuesta por Fischer, toma en cuenta adicionalmente el concepto de flexibilidad o cambio conformacional inducido por el ligando [Koshland, 1994].

De esta manera, la idea que las proteínas poseen una única estructura estática en su estado nativo fue siendo dejada de lado para ser considerada como una estructura con cierta flexibilidad inherente a su función biológica.

En 1965 Monod tomó en consideración la dinámica de la estructura proteica para describir

su función biológica [Monod et al., 1965] pero desde un punto de vista completamente distinto al de Koshland. Monod se basó en los resultados experimentales de Umbarger utilizando la L-treonina deaminasa [Umbarger and Brown, 1957] y en los resultados sobre la aspartato transcarbamoilasa obtenidos por Gerhart en 1962 [Gerhart and Pardee, 1962]. Distintos estudios llevados a cabo por Jean-Pierre Changeux [Changeux, 2011] sobre la L-threonina deaminasa y su actividad en presencia de urea y desnaturalización por temperatura, además de contar con el conocimiento de los confórmeros de la hemoglobina obtenidos por Max Perutz; condujeron a pensar que las proteínas existían en un equilibrio conformacional independiente de la existencia del sustrato.

A partir de estos trabajos surge el modelo de Monod – Wyman – Changeux (MWC), basado en un equilibrio entre al menos dos confórmeros pre-existentes al agregado de ligando. Este modelo puede ser considerado como uno de los primeros en tomar en cuenta explícitamente la pre-existencia de un conjunto de confórmeros en equilibrio. A diferencia del modelo de Koshland, en este modelo los confórmeros existen con anterioridad al agregado de ligando, reconociéndose así la existencia de un estado nativo más complejo (Figura 1.5). Entendemos por confórmero de una proteína a las distintas estructuras que puede adoptar una misma secuencia por rotación de enlaces simples y que corresponden a un mínimo de energía potencial específico. El cambio entre conformaciones se produce mediante rotaciones y cambios entre las interacciones de los residuos, dando lugar distintas conformaciones. Posteriormente a los trabajos de Koshland y Monod, la dinámica de la estructura proteica comenzó a tenerse en cuenta como un factor importante para describir su función biológica. Un ejemplo muy bien caracterizado es el de la mioglobina, que posee un grupo hemo con un átomo de hierro, y cuya función es almacenar y transportar oxígeno a los músculos. Sin embargo, la mioglobina luego de unir el oxígeno para su transporte a las células somáticas, debe ser capaz de desprenderse de él. Para llevar a cabo este proceso, debe experimentar cambios conformacionales en su estructura de tal manera que se alteren las constantes de disociación.

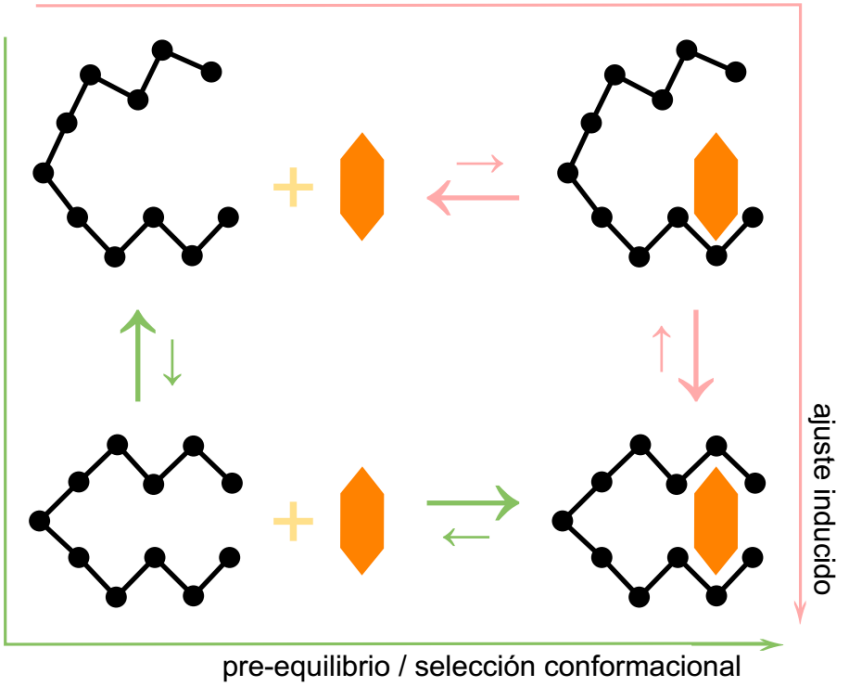

Figura 1.5: Representación esquemática de una proteína en sus formas abiertas (arriba) y cerradas (abajo) junto con su ligando (naranja). En el modelo de pre-equilibro o de selección conformacional, ambos confórmeros se encuentran presentes en la solución antes de la interacción con el ligando. El ligando se une al confórmero cerrado, desplazando el equilibrio hacia éste. Por el contrario, en el modelo del ajuste inducido, el ligando se une a la forma abierta disparando un cambio conformacional que termina en el confórmero cerrado

1.1.1. Definición actual del estado nativo

Es así como hoy en día, la "visión" actual [James et al., 2003] de la naturaleza del estado nativo proteico se encuentra incluida en la teoría del plegado proteico [Bryngelson and Wolynes, 1989, Bryngelson et al., 1995, Onuchic and Wolynes, 2004] definida por el modelo "embudo" (Figura 1.6), en donde la rugosidad de su fondo indica la existencia de un conjunto de confórmeros en equilibrio dinámico (Figura 1.7) [Tsai et al., 1999, Boehr et al., 2006, Volkman et al., 2001]. Acorde a esta visión, las poblaciones de estos confórmeros estructurales siguen distribuciones estadísticas termodinámicas y las barreras energéticas que hay entre ellos definen los cambios conformacionales que experimenta la estructura proteica. A esto es lo que hoy en día se lo conoce como "ensamble" nativo (del inglés *native ensemble*) [Wei et al., 2016]. El grado de diversidad conformacional está relacionada con la extensión del extremo rugoso del paisaje energético en forma de "embudo", incluyendo la distribución y la altura de las barreras energéticas entre confórmeros [Wei et al., 2016]. La población de confórmeros que constituyen el ensamble nativo está relacionada con el plegado proteico [Keskin et al., 2000],

con la presencia de ciertas mutaciones [Sinha and Nussinov, 2001], y con la historia evolutiva de la proteína [Maguid et al., 2006].

La idea inicial de paisajes energéticos "estáticos", como el ejemplificado en la Figura 1.7, fue extendida posteriormente a la noción de paisajes energéticos dinámicos para incluir el efecto del medio (solvente, condiciones, etc) en su forma general [Kumar et al., 2000]. Esta idea de paisajes dinámicos sustenta la hipótesis por la cual el ligando selecciona la conformación con mayor afinidad, así como los antígenos seleccionan el anticuerpo de mayor afinidad en la respuesta inmunológica [Foote and Milstein, 1994]. Adicionalmente, a pesar que el confórmero que une con mayor afinidad al ligando puede presentar una mayor energía relativa, los confórmeros que pertenecen a estados poco poblados en el "ensamble" podrían unirse al ligando y disparar alteraciones en ese potencial, y de esta forma desplazar el equilibrio hacia la conformación unida al mismo (Figura 1.8) [Kumar et al., 2000]. De esta forma, la descripción de paisajes dinámicos ofrece una visión central para explicar la relación estructura-dinámica-función de las proteínas [James et al., 2003].

Estos conceptos resultan fundamentales para la comprensión de procesos como el cooperativismo y el alosterismo [Papaleo et al., 2016, Tsai and Nussinov, 2014], la función biológica [Wolf-Watz et al., 2004, Kern et al., 2005, Eisenmesser et al., 2005], el reconocimiento molecular [Boehr et al., 2009a, Nussinov et al., 2013], la catálisis enzimática y el origen de nuevas funciones entre otros ejemplos [Tokuriki and Tawfik, 2009]. Básicamente los paisajes dinámicos y la existencia de la diversidad conformacional son la expresión estructural de las constantes de afinidad de un ligando por su receptor, definidas por las respectivas *kon* y *koff* (constantes cinéticas de unión y liberación respectivamente) que necesariamente cambiarán en función de los confórmeros considerados.

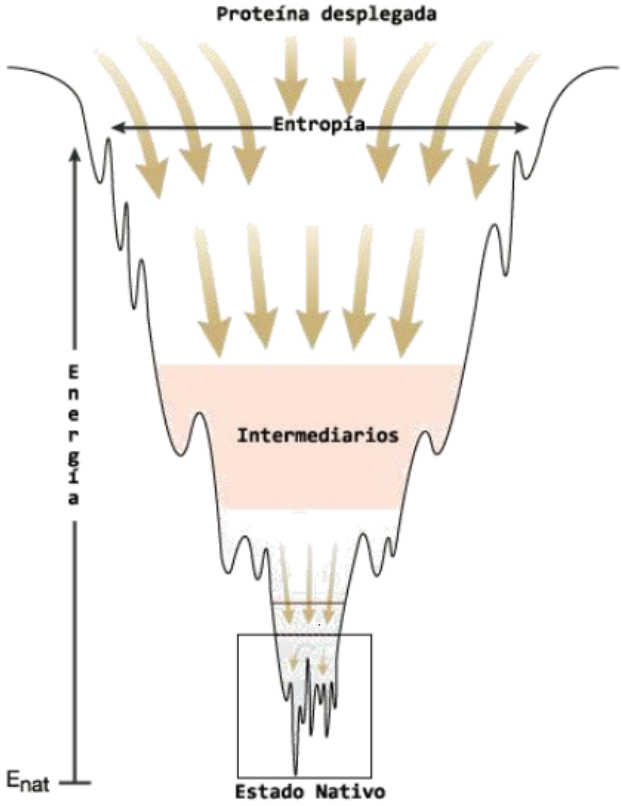

Figura 1.6: Paisaje energético en forma de "embudo" que explica el plegado proteico. El proceso de plegado de una proteína no sigue un camino único y específico, sino que sigue una especie de embudo energético en el cual la proteína puede adoptar diferentes rutas para minimizar su energía y llegar al estado nativo. El fondo del embudo rugoso describe el estado nativo de la proteína formado por un conjunto de confórmeros en equilibrio, con relativamente menor energía que los estados desnaturalizados. Por el contrario, la apertura del embudo está dominada por la enorme cantidad de estados desnaturalizados y de ahí con gran entropía. Imagen adaptada de `http://www.nature.com/horizon/proteinfolding/figures/summ_f1.html`.

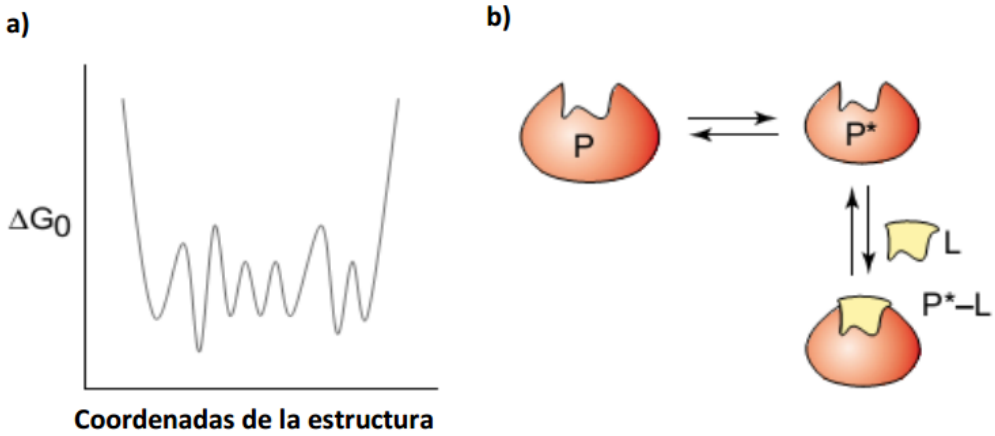

Figura 1.7: La "visión" actual asume un conjunto de confórmeros en equilibrio con una energía libre similar (a), y un modo de función basado en un equilibrio entre dos o más confórmeros pre-existentes (b). Imágenes extraídas y adaptadas de [James and Tawfik, 2003].

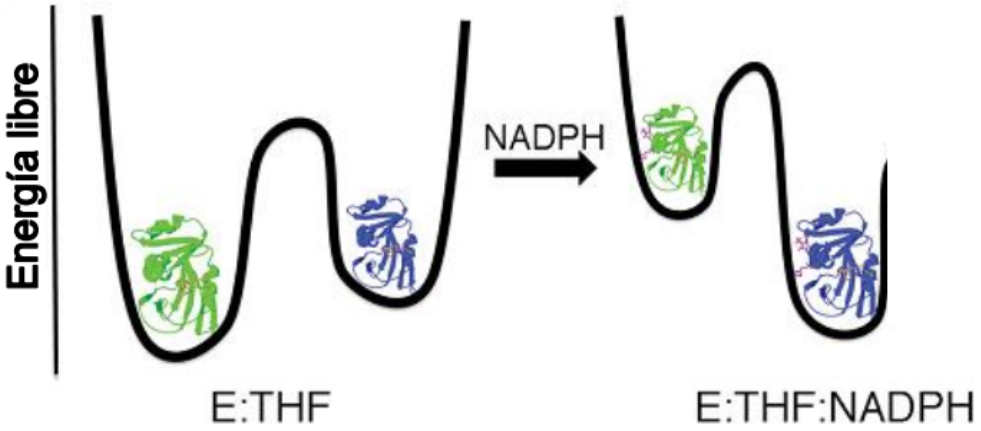

Figura 1.8: La unión del cofactor NADPH a la enzima dihidrofolato reductasa cambia la energía libre del paisaje energético de la misma lo que se denomina "paisajes dinámicos". En la figura se indica cómo la unión de un ligando (NADPH en este caso) afecta la energía libre del sistema favoreciendo en este caso la ocurrencia del confórmero azul. Imagen extraída y adaptada de [Boehr et al., 2006]

1.2. Diversidad conformacional en el estado nativo proteico

Una de las primeras evidencias experimentales de diversidad conformacional fue encontrada en 1959, cuando Max Perutz determinó la estructura tridimensional de la hemoglobina en dos estados o conformaciones utilizando la técnica de difracción por Rayos X. La hemoglobina cuenta con cuatro cadenas polipeptídicas que forman un tetrámero, el cual se encuentra en equilibrio entre dos estados (Figura 1.9): el T o "desoxi" y R u "oxi" [Perutz et al., 1960]. En el estado desoxi, la hemoglobina se encuentra estructuralmente "tensa" (debido a la presencia de un mayor número de interacciones débiles, es una estructura más rígida que el otro confórmero y de aquí el nombre de T por *"tense"*). Estas mismas uniones estabilizantes adicionales entre sus subunidades se oponen a la unión del oxígeno. En su estado oxi, las interacciones que estabilizan el tetrámero se encuentran más "relajadas" (de aquí el nombre de R por *"relaxed"*) favoreciendo la unión del oxígeno al grupo hemo. Una vez que el O_2 se une a un sitio activo de la hemoglobina, el átomo de hierro (Fe_2) se oxida a (Fe_3). La unión del O_2 al átomo de hierro resulta en un cambio conformacional en el residuo de histidina hacia porfirina en la estructura de la hemoglobina, que finalmente resulta en un aumento de afinidad por el O_2

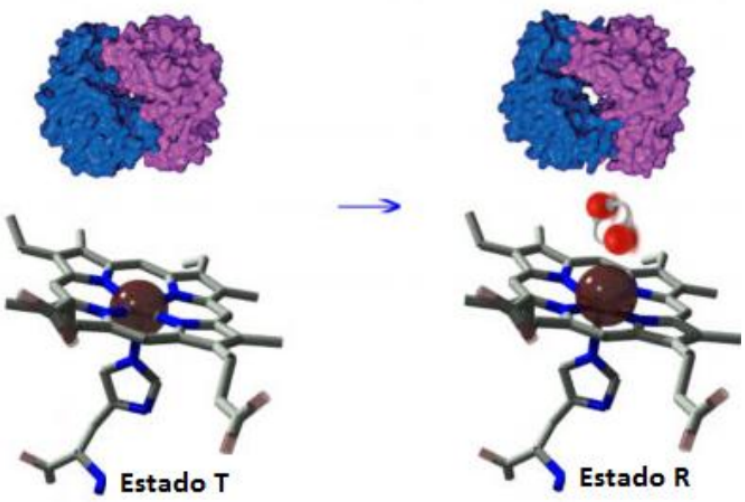

Figura 1.9: Transición entre el estado T y R de la hemoglobina.

simplemente por un aumento relativo del confórmero R. El movimiento asociado al grupo que contiene la histidina dará lugar a un cambio conformacional en el resto de la estructura de la hemoglobina. El grupo COO^- está ahora interactuando con la interfaz alfa-beta causando cambios conformacionales de los sitios activos vecinos. Estos cambios conformacionales, que producen el desplazamiento del equilibrio entre las dos conformaciones, se traducirán en un aumento de la afinidad del O_2 a la hemoglobina. Adicionalmente, permiten explicar importantes propiedades funcionales, tales como el cooperativismo de unión del oxígeno y las propiedades alostéricas debidas a la unión de distintos ligandos como el protón y moléculas orgánicas como el difosfoglicerato, entre otras.

Siguiendo el modelo de Monod de pre-equilibrio, en todo nuestro trabajo entendemos por diversidad conformacional a la existencia de más de una conformación de la misma proteína en el equilibrio, independientemente de la presencia o no de un ligando que induzca el cambio conformacional. Recientes evidencias experimentales apoyan el modelo de pre-equilibrio en contraposición al del modelo de ajuste inducido [Zoete et al., 2002, Nussinov and Ma, 2012, Gardino et al., 2009, Boehr et al., 2009b]. A modo de ejemplo, podemos mencionar el trabajo realizado por *Macol et al.*, donde utilizan una forma mutada de la aspartato transcarbomilasa de *E. coli* para comprobar si la enzima seguía el modelo de ajuste inducido o pre-equilibrio de Monod. Recordemos que ambos modelos son indistinguibles desde el punto de vista cinético y que por esta razón deben explorarse en forma estructural. Brevemente, la aspartato transcarbomilasa es una enzima hetero-12-mérica compuesta por seis cadenas ca-

talíticas con un sitio activo cada una, y otras seis que carecen de actividad catalítica. Para su función esta enzima posee dos estados T y R como la hemoglobina. En el trabajo mencionado, muestran que usando una forma mutada de esta enzima, obtienen una forma quimérica la cual sólo permite unir el sustrato a uno de los seis sitios activos. La unión del sustrato a sólo uno de sus sitios permite el paso desde la conformación T a R [Macol et al., 2001]. Lo que se demuestra en este trabajo condice con el modelo de Monod, ya que en el pre-equilibrio existe la conformación con sólo un sitio activo que al unir el ligando desplaza el equilibrio conformacional a la conformación R, provocando cambios conformacionales en los cinco sitios restantes. Esto se ha reafirmado en otros trabajos utilizando diferentes técnicas experimentales, donde efectivamente demuestran que los estados T y R de la aspartato transcarbomilasa pre-existen en equilibrio dinámico [Velyvis et al., 2007].

La extensión de la diversidad conformacional observada en proteínas va desde fluctuaciones en las cadenas laterales (apertura y cierre de túneles y cavidades) a movimientos de loops, estructura secundaria y reordenamientos de la estructura terciaria (Figura 1.10 y Figura 1.11). Las técnicas de Resonancia Magnética Nuclear (RMN) han sido particularmente efectivas en revelar la diversidad conformacional de las proteínas, ya que ha demostrado que éstas pueden adoptar en solución conformaciones alternativas incluso en la ausencia de un ligando [Csermely et al., 2010]. Un ejemplo es el dominio N – terminal SH_3 de la proteína drk de *Drosophila*, el cual puede adoptar aproximadamente 60 conformaciones estables y diferentes, presentando significativas diferencias estructurales entre algunas de ellas [Choy and Forman-Kay, 2001]. Otro ejemplo de diversidad conformacional muy bien caracterizado a nivel de estructura secundaria, se da en el dominio de unión a ADN del represor Arc. El Arc N11L mutante se convierte espontáneamente desde una hoja β a una estructura de hélice 3_{10} [Cordes et al., 2000]. En ausencia de ligando (ADN), estas dos conformaciones parecerían estar en pre-equilibrio en proporciones casi idénticas.

Quizás una de las manifestaciones más llamativas de la diversidad conformacional proteica es en aquellas proteínas que se encuentran parcial o completamente desordenadas en su estado nativo. En los últimos 20 años, la biología estructural en conjunto con la bioinformática han puesto gran esfuerzo en caracterizar proteínas y/o regiones que carecen de

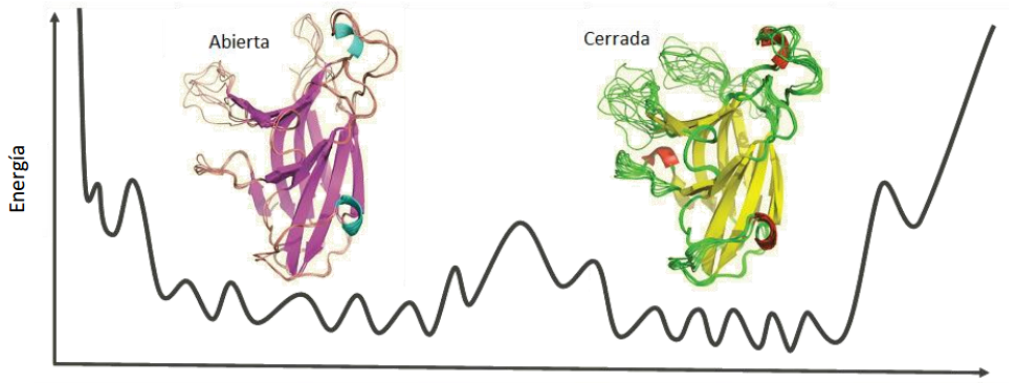

Figura 1.10: Diversidad conformacional del receptor Eph. Consta de dos conformaciones, una abierta y otra cerrada, y además presentan movimientos en su estructura secundaria y regiones de loops. Imagen extraída y adaptada de [Nussinov and Ma, 2012].

una estructura terciaria en condiciones fisiológicas [Dunker et al., 2008, Dyson and Wright, 2005, Tompa, 2002]. Estas proteínas desafían el conocido paradigma de estructura-función ya que no necesitan de una estructura tridimensional definida para cumplir su función y participan en muchos procesos biológicos como unión de ARN y ADN, transcripción, traducción, regulación del ciclo celular, como así también en patologías asociadas al plegado incorrecto de la estructura y agregación [van der Lee et al., 2014, Chiti and Dobson, 2006, Wright and Dyson, 1999]. Estas proteínas poseen múltiples conformaciones en su estado nativo que están separadas por barreras de baja energía libre en el paisaje energético, y consecuentemente pueden fluctuar entre diferentes conformaciones constantemente, dando lugar a ensambles muy dinámicos (ver Figura 1.12). Sin embargo, en algunos casos adquieren una estructura ordenada o parcialmente ordenada cuando se encuentran en presencia de algún ligando o proteína de interacción [Abagyan and Batalov, 1997, Zea et al., 2016, DeForte and Uversky, 2016].

Este tipo de proteínas muestran las diferencias estructurales entre sus confórmeros que pueden variar drásticamente y ser tan globales como para afectar toda la estructura de la proteína (por ejemplo, una conformación podría estar completamente ordenada y la otra no). Podemos decir que las proteínas se encuentran en un contínuo conformacional que va desde una proteína completamente estructurada a una completamente desordenada. Este espectro comprende dominios bien plegados que no muestran desorden global o sólo desorden local en loops o *tails*, proteínas multidominio conectados por regiones desordenadas, proteínas en estado de glóbulo fundido o *"molten globule"* compactos que contienen una extensa estructura secundaria, glóbulos colapsados formados por segmento de estructura secundaria, estados

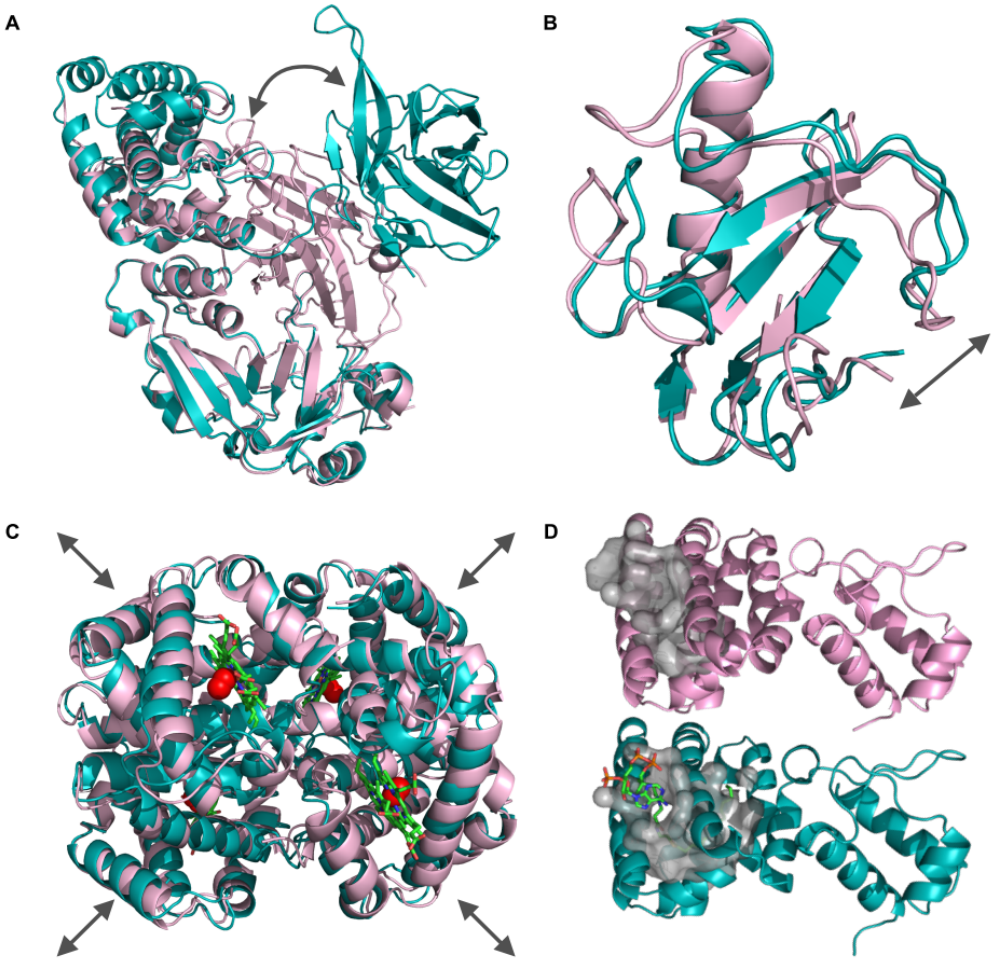

Figura 1.11: A) Muestra la toxina de difteria en complejo con NAD (PDB: 1TOX, cadena A) en celeste y cristalizada en condiciones ácidas en color rosa (PDB: 4OW6, cadena A). El RMSD de Cα entre los confórmeros en la región superpuesta es de 1.78 $\mathring{A}$. Esta proteína muestra un movimiento de bisagra de hasta 65$\mathring{A}$ que permite la rotación del dominio indicada con la flecha gris. B) Se observa a la ribonucleasa F1 en complejo con ácido piroglutámico en celeste (PDB 1RCK, cadena A del modelo 21 obtenido por RMN) y la forma libre de ligando en rosa (PDB: 1FUS, cadena 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 $\mathring{A}$. 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 $\mathring{A}$ 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, cadena A) y en el estado *holo* en complejo con miristoil-CoA en celeste (PDB: 1H9G, cadena A) que presentan un RMSD de 1.28 $\mathring{A}$. La diferencia principal entre los confórmeros es un cambio en el volumen de la cavidad (1586$\mathring{A}^3$) indicada por la superficie gris y el número de túneles que permiten al ligando acceder al sitio activo. Imágen extraída de [Parisi et al., 2015]

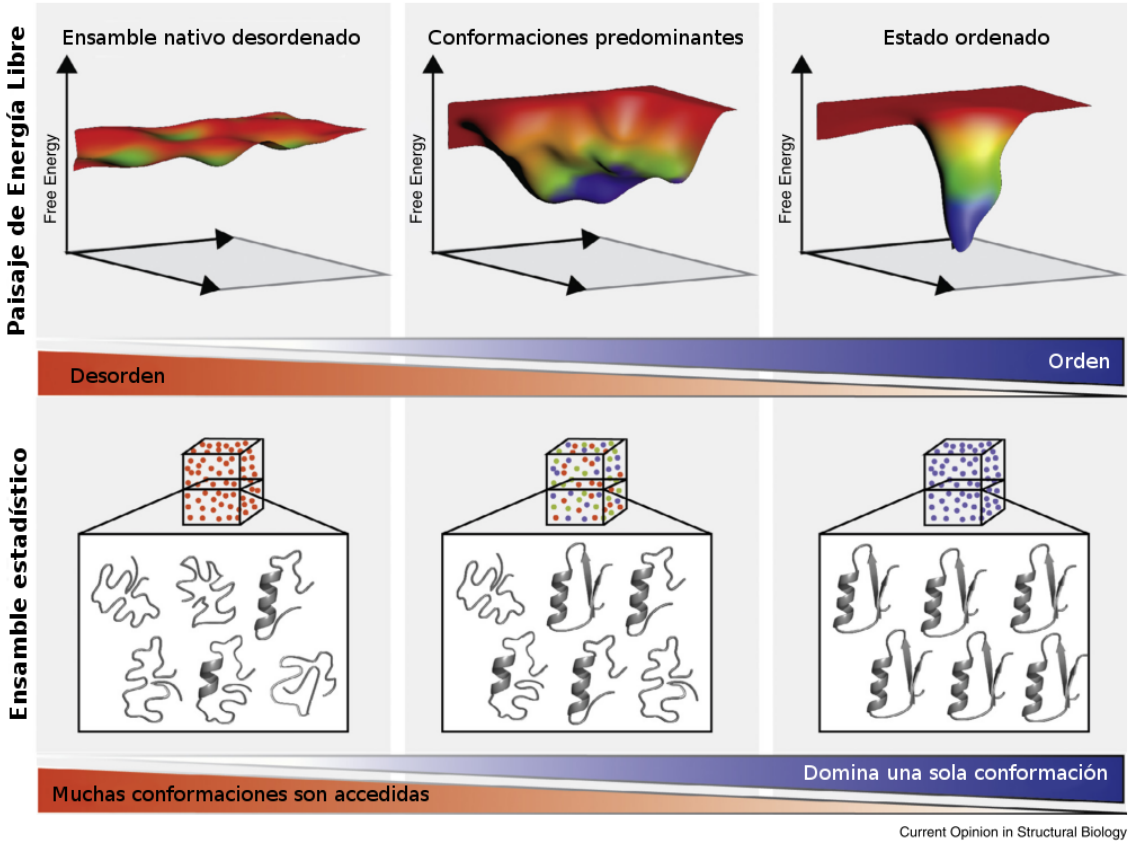

Figura 1.12: El paisaje de energía libre y los ensambles conformacionales de IDRs o IDPs se pueden ajustar dinámicamente. El paisaje de energía libre representa esquemáticamente cualquier conformación posible (ejes x-y) que una proteína puede adoptar, así como su respectiva energía libre (eje z) (paisaje de energía libre, fila superior), y por lo tanto describe la probabilidad de encontrar algún confórmero en un ensamble de estados conformacionales (ensamble estadístico, fila inferior). Mientras que las proteínas estructuradas u ordenadas tienen paisajes energéticos con una conformación claramente preferida (mínimo global, panel derecho), las proteínas desordenadas en su forma nativa tienen un paisaje energético rugoso y plano que permite que múltiples conformaciones sean termodinámicamente favorables y se encuentren en el ensamble estadístico (panel izquierdo). Varios factores pueden modular el paisaje de energía libre de IDRs o IDPs y restringir el ensamble estadístico a conformaciones menos definidas (panel central). A medida que disminuye la entropía conformacional, también lo hace el número de estados conformacionales observados. En el estado más ordenado, el paisaje de proteínas desordenadas se convierte en un paisaje en forma de embudo de proteínas estructuradas, desplazando el equilibrio del ensamble conformacional hacia una conformación predominante (panel derecho). Imágen extraída y adaptada de [Flock et al., 2014]

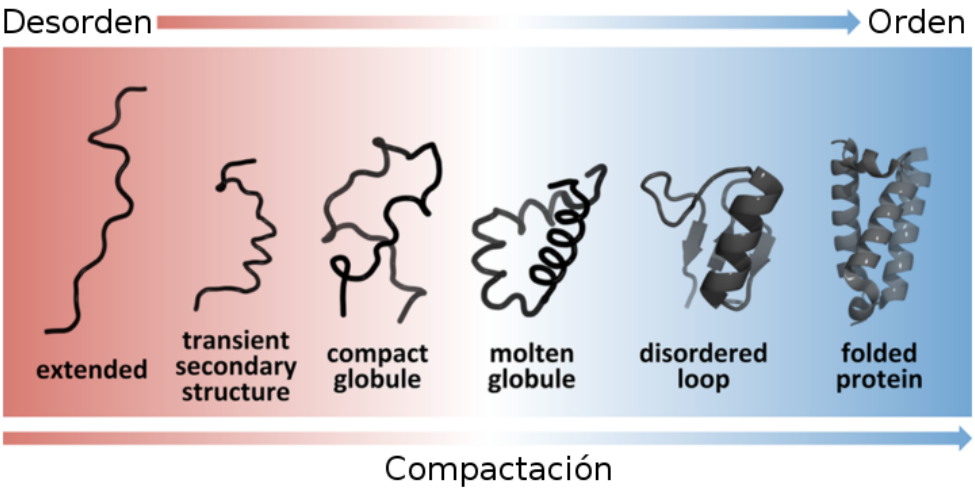

Figura 1.13: El gradiente de color representa el continuo de los estados conformacionales que se extienden desde altamente dinámicos, ensambles conformacionales extendidos (rojos) hacia compactos, restringidos dinámicamente, estados completamente ordenados (azul). Los estados dinámicamente desordenados están representados por líneas negras gruesas y estructuras plegadas estables como cartoon. Imágen extraída y adaptada de [van der Lee et al., 2014]

desplegados que pueblan transitoriamente elementos locales de la estructura secundaria, y estados muy extendidos que se parecen a un ovillo de hilo (Figura 1.13). Las IDPs o IDRs son altamente dinámicas y puede fluctuar rápidamente en este contínuo de estructuras [van der Lee et al., 2014]. Desde un punto de vista evolutivo, las proteínas intrínsecamente desordenadas proporcionan una poderosa demostración del hecho que una estructura tridimensional definida no es un requisito previo para la función proteica [Wright and Dyson, 1999, Abagyan and Batalov, 1997, van der Lee et al., 2014].

Es este trabajo de tesis estudiaremos la diversidad conformacional a nivel de los movimientos *backbone*. Sin embargo, como podemos ver en la Figura 1.12 existen proteínas donde predomina una sola conformación; en el capítulo 4 veremos que las poblaciones de confórmeros de estas proteínas no difieren en su *backbone* sino en la disposición de cadenas laterales R de los aminoácidos. Veremos que éstos movimientos locales están asociados con la apertura y cierre de túneles o canales o en el cambio de volumen de cavidades en la proteína.

1.3. Diversidad conformacional y función proteica

Como mencionamos anteriormente, el concepto de la diversidad conformacional se encuentra en estrecha relación con la función de las proteínas. Una consecuencia de la diversidad con-

formacional es que provee un mecanismo para diversificar la capacidad funcional. Una proteína que adopta diferentes conformaciones, en principio, podría presentar diferentes funciones [Ma et al., 2002]. A partir de esto, surgen los conceptos de promiscuidad y multi-especificidad funcional en proteínas. El primero hace referencia a aquéllas que presentan diferentes funciones utilizando el mismo sitio activo y el segundo se refiere a una proteína que ejerce una función similar (por ejemplo, unión) en ligandos claramente diferentes, probablemente utilizando diferentes residuos de sitio activo [James et al., 2003]. A pesar de la dificultad de encontrar ejemplos precisos de diversidad conformacional mediada por multi-especificidad o promiscuidad, se han caracterizado algunos sistemas. Por ejemplo, un anticuerpo y la albúmina de suero humano, que son proteínas sin relación alguna, comparten características similares en su sitio activo, el cual es hidrofóbico y posee una lisina en su cadena lateral que puede actuar como base (Figura 1.14) [James and Tawfik, 2001].

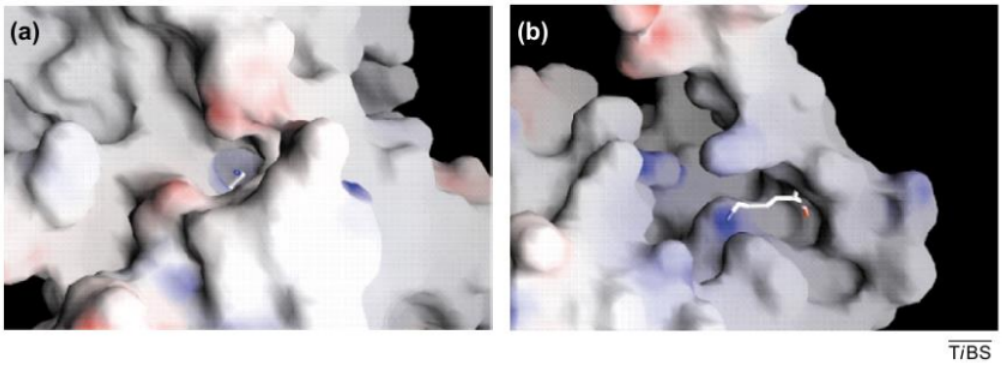

Figura 1.14: Un anticuerpo (a) y la albúmina de suero humano (b) comparten un patrón de unión promiscuo y similares actividades catalíticas. Ambas proteínas tienen un sitio activo hidrofóbico con una lisina en la cadena lateral que puede actuar como una base o interactuar electrostáticamente con ligandos cargados negativamente. Estas dos proteínas no poseen relación alguna en secuencia y estructura y tampoco han evolucionado hacia las actividades de unión promiscua y acción catalítica que poseen [James and Tawfik, 2001, Hollfelder et al., 1996]. Esto indica que las características del sitio activo podrían ser promiscuas reclutadas en un contexto mecanicista. Imagen extraída de [James and Tawfik, 2003].

Otro ejemplo en donde se encuentran cambios conformacionales más significativos respecto a la función es en el anticuerpo monoclonal denominado SPE7. Éste puede adoptar diferentes conformaciones de su sitio de unión, permitiéndole la unión de múltiples antígenos no relacionados (Figura 1.15) [James et al., 2003].

A partir de los ejemplos anteriores podemos ver que la diversidad conformacional juega un papel muy importante, proporcionando un mecanismo de diversificación funcional. Muchas

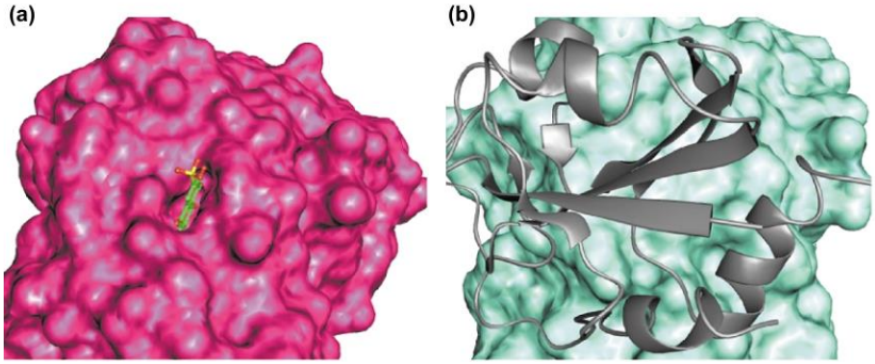

Figura 1.15: Diversidad conformacional del anticuerpo SPE7. Dos conformaciones diferentes de SPE7 le permiten unirse a antígenos que no se encuentran completamente relacionados en todo el espectro estructural. Una conformación (a) tiene un sitio de unión profundo y estrecho que une el hapteno para el cual se utilizó SPE7 [dinitrofenol (DNP)]. La otra conformación (b) tiene una superficie de unión plana, que promiscuamente une una antígeno proteico (gris). La diferencia en las superficies de unión se pueden observar en el centro de cada imagen; hay una profunda cavidad en la estructura unida a DNP la cual está ausente en la conformación a la cual se une la proteína (gris). Este análisis estructural se apoya en estudios experimentales que indican la existencia de diferentes isómeros pre-existentes. Pareciera ser que cada ligando selecciona un isómero a su favor, y por lo tanto, desplaza el equilibrio en su dirección. Imagen extraída de [James and Tawfik, 2003]

proteínas exhiben múltiples funciones celulares y algunas enzimas catalizan promiscuamente otras reacciones que no son las habituales [Khersonsky et al., 2006]. Esta promiscuidad y multi-especificidad podría ser atribuida a varios ligandos o substratos, que desplazan el equilibrio hacia las conformaciones menos pobladas del ensamble (Figura 1.16) [Dean and Thornton, 2007, Tokuriki and Tawfik, 2009].

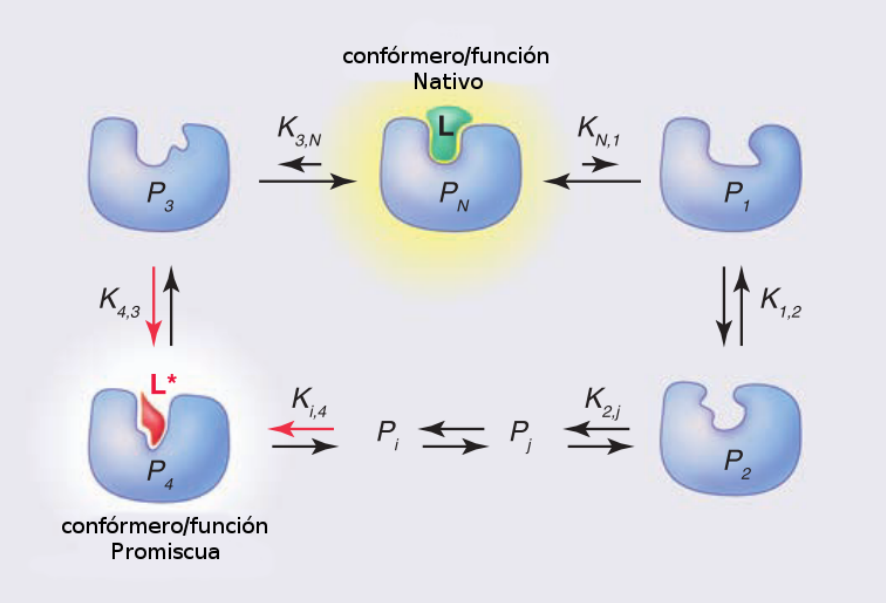

Figura 1.16: Dinámica de la función y estructura proteica y su evolución. Este modelo asume que una proteína existe como un conjunto de confórmeros en equilibrio, siendo el dominante el confórmero P_N, el que interactúa con el ligando nativo L. Los confórmeros alternativos corresponden a variaciones estructurales que van desde rotaciones de las cadenas laterales y reordenamientos del sitio activo hasta otros más profundos como transiciones entre los diferentes plegados. Los confórmeros menos representados (ej. P_4) pueden mediar funciones alternativas, como la interacción promiscua con L^* (donde L^* es un ligando para el cual la proteína no ha evolucionado para unirse al mismo). Las mutaciones pueden gradualmente alterar este equilibrio de tal manera que los confórmeros escasamente poblados se hacen más favorables con efectos sustanciales sobre la correspondiente función promiscua (por ejemplo, un incremento de la ocupación de P_4 desde 0.01 a 0.1 puede producir un aumento de 10-veces en el nivel global de la función promiscua). Por lo tanto, la ocupación relativa del confórmero nativo podría ser gravemente afectada (por ejemplo, desde 0.5 a ≥ 0.41, lo que conduce a la pérdida de $<20\,\%$ de la función nativa). Del mismo modo, una mayor especificidad podría evolucionar a través de mutaciones que reducen la ocupación de determinados confórmeros promiscuos. Este modelo también explica la debilidad de las compensaciones negativas entre las funciones actuales y futuras y el potencial evolutivo de las mutaciones neutrales. Imágen extraída de [Tokuriki and Tawfik, 2009].

1.4. Objetivos

Objetivo General:

El objetivo general de este trabajo de tesis es el estudio de la diversidad conformacional y su relación con los aspectos biológicos, estructurales (dominios, subunidades, interacciones), evolutivos y fisicoquímicos de las mismas. Este trabajo permitirá profundizar el conocimiento en la relación estructura-dinámica-función mediante el análisis computacional de proteínas que exhiben diversidad conformacional y los factores que la modulan. Esta tesis se encuentra enmarcada en el área de bioinformática estructural por lo que se utilizarán herramientas computacionales y bioinformáticas típicas para el análisis de estructuras, como así también

otras que han surgido en los últimos años para la detección de túneles y cavidades. El estudio de la estructura proteica a la luz de la diversidad conformacional permitirá avanzar en el conocimiento actual, no sólo en nuestro conocimiento biológico sino también en el desarrollo de nuevos métodos y herramientas que representen a la proteína como un ensamble de confórmeros en equilibrio.

Objetivos específicos:

1. Generar una base de datos de diversidad conformacional de proteínas completas, que servirá como punto de partida de posteriores análisis. Esta base de datos contendrá una colección redundante de estructuras experimentales de proteínas, datos biológicos, fisicoquímicos y estructurales de las mismas.

2. Desarrollar un servidor en línea de acceso libre y gratuito a la comunidad científica en donde los usuarios puedan efectuar consultas a la base de datos desarrollada y visualizar la información.

3. Estudiar la distribución y la extensión de la diversidad conformacional en proteínas, así como también la caracterización de los factores biológicos y fisicoquímicos que la modulan.

4. Analizar la correlación entre la diversidad conformacional y la divergencia estructural y secuencial en familias de proteínas.

Principalmente los objetivos 1 y 2 se tratan en el capítulo 2, l objetivo 3 en los capítulos 3 y 4 y finalmente el objetivo 4 en el capítulo 5.

Capítulo 2

Desarrollo de una base de datos para el estudio de la Diversidad Conformacional[1]

2.1. Resumen

El estudio de la diversidad conformacional proteica no es un problema trivial como para ser abordado computacionalmente. Si bien se puede estudiar la dinámica de una proteína utilizando diferentes métodos computacionales, muchas veces resultan inviables para ser aplicados a miles de proteínas simultáneamente. Una forma de poder estudiar los confórmeros del estado nativo proteico es utilizar las diferentes estructuras experimentales de una misma proteína que se han obtenido en diferentes condiciones. En este capítulo, utilizando todas las estructuras redundantes que se encuentran en PDB, describiremos el desarrollo e implementación de una base de datos de diversidad conformacional de proteínas denominada CoDNaS ("Conformational Diversity of the Native State"). CoDNaS es la primera base de datos en considerar todas las estructuras disponibles para cada proteína en PDB y estimar su diversidad conformacional a partir de diferentes parámetros estructurales. Además, se encuentra ampliamente anotada para incluir no sólo la información biológica y fisicoquímica de cada proteína, sino también las condiciones experimentales en las que se obtuvo cada uno de sus confórmeros. Por otra

[1]Este capítulo está basado en las publicaciones: [Monzon et al., 2013, Monzon et al., 2016].

parte, CoDNaS posee un servidor web que está disponible en forma gratuita para toda la comunidad científica, permitiendo consultar la base de datos por diferentes criterios de búsqueda y explorar la diversidad conformacional de cada una de sus proteínas.

2.2. Introducción

La caracterización de la diversidad conformacional en el estado nativo proteico resulta un gran desafío debido al tamaño y complejidad del espacio conformacional. No sólo son múltiples las posibles conformaciones de una cadena polipeptídica, sino que además el paisaje energético es dinámico haciendo que los sub-estados sean diferentes en distintas condiciones. El paisaje energético define la amplitud y la escala de tiempo en la que se realizan los movimientos, y estos se pueden detectar con distintas técnicas (Figure 2.1).

Experimentalmente, el estudio de la dinámica proteica ha logrado buenos resultados gra-

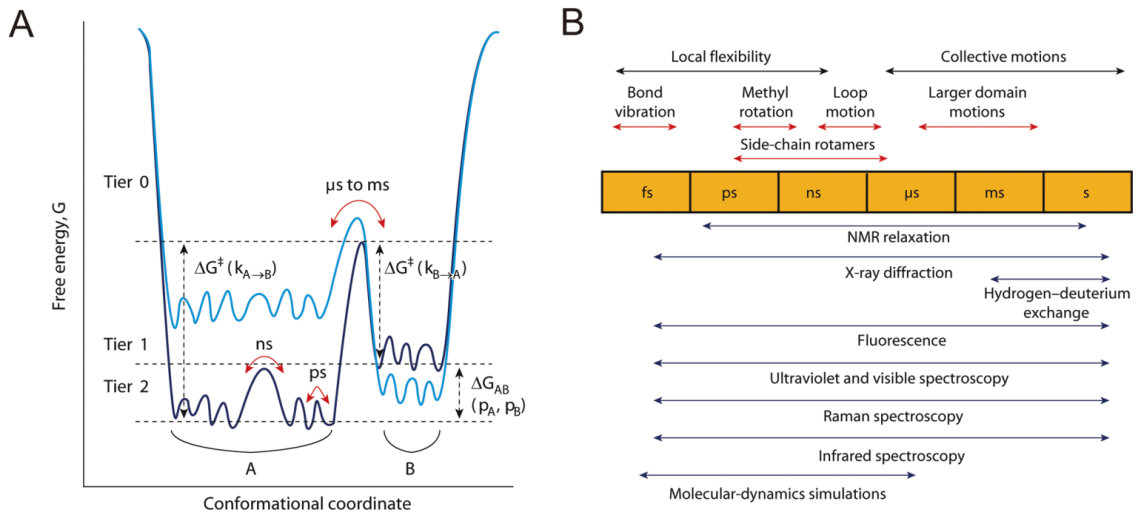

Figura 2.1: El paisaje energético define la amplitud y la escala de tiempo en la que se realizan los movimientos proteicos. A) Corte transversal unidimensional a través del paisaje energético multidimensional de una proteína, que muestra la jerarquía en la dinámica proteica y sus barreras energéticas. Cada nivel se clasifica siguiendo la descripción introducida por Frauenfelder, Sligar and Wolynes and co-workers [Frauenfelder et al., 1991]. Un estado es definido como un mínimo en la superficie de energía, donde el estado de transición es el máximo entre los pozos. Las poblaciones de los estados A y B en el nivel-0 (ρ_A,ρ_B) son definidas como distribuciones de Boltzmann basadas en su diferencia de energía libre(ΔG_{AB}). La barrera entre estos estados ($\Delta G^{\ddagger}$) determina la tasa de interconversión k. Los niveles inferiores describen fluctuaciones más rápidas entre un gran número de subestados estrechamente relacionados dentro de cada estado de nivel-0.Un cambio en el sistema (por ejemplo, la presencia de un ligando) podría alterar el paisaje de energía y desplazar el equilibrio entre los estados (pasar del azul oscuro al azul claro, o vice versa). B) Escala de tiempo de los movimientos proteicos y los métodos experimentales que pueden detectar fluctuaciones en cada escala de tiempo. Imágen extraída y adaptada de [Henzler-Wildman and Kern, 2007].

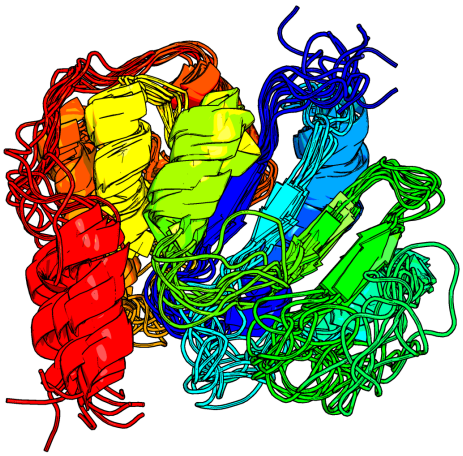

Figura 2.2: Conformaciones obtenidas por RMN de la proteína ALG13 (código PDB: 2JZC).

cias a los avances en el campo de la Resonancia Magnética Nuclear (RMN), que permite inferir las coordenadas espaciales de los átomos de una proteína en diferentes configuraciones pertenecientes a su espacio conformacional en solución [Lindorff-Larsen et al., 2005, Weikl and Paul, 2014] (Figura 2.2). Además, existen otras técnicas experimentales como la denominada "pre-cinética del estado estacionario" o *"pre-steady state binding kinetics"*, que consiste en estudiar los milisegundos posteriores a la mezcla de una enzima con el substrato, para poder identificar posibles intermediarios o conformaciones, antes de que se alcance el estado estacionario con el producto final de la reacción. Sin embargo, sólo permite la identificación de las poblaciones de confórmeros, pero no la caracterización de su estructura tridimensional [Fersht and Requena, 1971, Lancet and Pecht, 1976].

Otra forma de abordar el estudio de la dinámica de una proteína es mediante simulaciones computacionales. Usando dinámicas moleculares considerando todos los átomos y el solvente en forma explícita o utilizando lo que denomina *replica exchange*, podemos muestrear algunas conformaciones nativas y los resultados se acercan a los obtenidos por RMN [Prabhu et al., 2003]. Dado el costo computacional de una simulación usando todos los átomos en algunos casos se utiliza una versión simplificada de la proteína en la técnica llamada dinámica molecular de grano grueso o *"Coarse Grained"* [Chng and Yang, 2008]. Es un método capaz de explorar los movimientos lentos, en el orden de los 10μs, donde se producen algunos cambios conformacionales en una proteína. Sin embargo, debido a que los grandes movimientos se dan en el orden de los milisegundos, requiere mucho tiempo de simulación y por lo tanto una

gran capacidad de cómputo. Otro método de grano grueso utilizado para estudiar grandes movimientos relacionados con la función biológica son los *Elastic Network Models (ENM)*, donde se representa a la proteína como una red elástica de carbonos alfa unidos por resortes que describen las restricciones armónicas sobre los desplazamientos desde la estructura de equilibrio de la proteína. [Bahar and Rader, 2005, Atilgan et al., 2001]. La desventaja de los métodos de grano grueso es que al utilizarse una versión simplificada de la estructura de la proteína, se pierde mucha información que puede ser biológicamente relevante.

La diversidad conformacional puede estudiarse también a partir de diferentes estructuras experimentales de la misma proteína obtenidas en distintas condiciones experimentales. Cada estructura representa un confórmero del estado nativo del sub-estado más estable para la condición experimental en que se obtuvo [Meng and McKnight, 2009, Burra et al., 2009, Best et al., 2006, Karplus and Schulz, 1985, Zoete et al., 2002]. A continuación se detalla ésta metodología y la razón por la cual se la ha empleado en este trabajo.

2.2.1. Estimación de la diversidad conformacional a partir de estructuras redundantes

En el presente trabajo se utilizó un enfoque completamente diferente de los métodos computacionales comúnmente empleados para abordar el estudio de la diversidad conformacional y sus aspectos fisicoquímicos. Básicamente este método fue el utilizado por Perutz en la cristalización diferencial de los confórmeros de la hemoglobina a principios de la década de los 60' [Perutz et al., 1960, Perutz and Mathews, 1966]. El método consiste en tomar en consideración la colección de todas las estructuras cristalográficas y de RMN de la misma proteína en diferentes condiciones, como una descripción de los confórmeros pre-existentes en el equilibrio que caracterizan al estado nativo. Se ha demostrado que diferentes instancias experimentales de la misma proteína obtenidas en distintas condiciones, confinan a la estructura de la misma en sub-estados estables que son representantes del ensamble nativo. Este concepto se encuentra sustentado por las diferencias estructurales entre distintos modelos de la misma proteína resueltos por RMN, así como también por las diferencias estructurales entre estructuras cristalográficas de la misma proteína determinadas en distintas condiciones

experimentales [Zoete et al., 2002,Burra et al., 2009]. Al modificar las condiciones de cristalización (por ejemplo agregando un sustrato, modulador alostérico o incluyendo una modificación postraduccional) se favorecen distintos confórmeros por el desplazamiento del equilibrio conformacional. De esta forma, una colección de estructuras de la misma proteína cristalizada en distintas condiciones puede ser tomada como una descripción del ensamble de confórmeros o de la dinámica de la proteína en su estado nativo [Burra et al., 2009, Meng and McKnight, 2009]. Esta metodología fue demostrada al comparar una colección redundante de estructuras cristalográficas con aquellas estructuras obtenidas por RMN [Best et al., 2006, Zoete et al., 2002].

Una de las ventajas de usar estructuras experimentales es que se reducen posibles artefactos resultantes de la simulación o predicción de estructuras computacionalmente. Además, el costo de cálculo para comparar éstas estructuras es de varios órdenes de magnitud menor que el de hacer, por ejemplo, una simulación de dinámica molecular larga con todos lo átomos. Esto nos permite estimar la diversidad conformacional de miles de proteínas y poder estudiar comportamientos globales que surgen de la relación estructura-dinámica-función. Dentro de las desventajas o limitaciones, se pueden mencionar algunos aspectos que son propios de las técnicas experimentales por la cual se obtuvieron las estructuras y pueden afectar a la misma. En el caso de estructuras obtenidas por RMN se trata de evitar compararlas con estructuras obtenidas por DRX ya que pueden diferir en gran magnitud en sus dinámicas, debido a diferencias entre las técnicas y no por una cuestión biológica. En el capítulo 3 analizaremos este aspecto. En RMN la proteína se encuentra en solución, mientras que en DRX se encuentra cristalizada en fase sólida, a su vez, en RMN las coordenadas de las estructuras se obtienen a partir de distancias inter-atómicas que se ajustan a un modelo, mientras que en DRX se obtienen a partir de un mapa de densidad electrónica [Sikic et al., 2010]. Por lo mencionado anteriormente uno trata de evitar mezclar y comparar estructuras RMN con DRX para estimar la diversidad conformacional de una proteína. Sin embargo, disponer de estructuras obtenidas por ambas técnicas proveen una fuente de información complementaria para estudiar la flexibilidad de una proteína. Cabe mencionar que las estructuras obtenidas por DRX también se encuentran afectadas por cuestiones inherentes al proceso de cristalización como los contactos entre las proteínas que forman el cristal o "*crystal packing contacts*" y

el grupo espacial o simétrico del cristal. Sin embargo, se ha demostrado que usando una determinada cantidad de estructuras para estimar la diversidad conformacional del estado nativo, el patrón de flexibilidad obtenido de la comparación de múltiples estructuras de la misma proteína es más robusto que cualquier efecto que puedan ocasionar los artefactos de la cristalización mencionados anteriormente [Best et al., 2006] (ver capítulo 3). Finalmente podríamos mencionar como una desventaja la necesidad de contar con distintas estructuras de la misma proteína. Podríamos suponer que la cantidad de proteínas depositadas en PDB con distintas estructuras no son la mayoría (éste punto lo trataremos en el capítulo 3). Sin embargo, como mostraremos más adelante, CoDNaS cubre más del 70 % de las estructuras depositadas en la PDB con un promedio de 15.09 confórmeros por proteína.

2.2.2. Cuantificación de la diversidad conformacional

Con el objetivo de estimar la diversidad conformacional de una proteína y cuantificarla, debemos comparar estructuralmente sus confórmeros. Para realizar este procedimiento, en primer lugar los confórmeros deben superponerse estructuralmente con el fin de llevar al mismo eje de coordenadas los centro de masas de cada una de las estructuras a comparar. Esto puede hacerse fácilmente dado que los confórmeros comparten la misma secuencia de aminoácidos por lo que para hacer una superposición rígida es suficiente un alineamiento secuencial para asignar los residuos equivalentes entre las dos estructuras, 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 (RMSD por sus siglas en inglés de *"Root Mean Square Deviation"*) entre los átomos equivalentes de las estructuras, y generalmente se utiliza un algoritmo simple de ajuste de mínimos cuadrados o *"least-squares fitting"* [McLachlan, 1982] para realizar este proceso (Figura 2.3). Es común que esta superposición sea efectuada tomando como referencia las coordenadas espaciales de un átomo en particular por cada aminoácido, por lo que generalmente se utilizan los C_α de la proteína. En este trabajo cada vez que nos remitimos a un valor de RMSD, será aquel obtenido utilizando como referencia los C_α. El RMSD puede calcularse para cualquier par de estructuras superpuestas de manera óptima, según la siguiente ecuación (2.1):

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N} d_i^2}{N}} \tag{2.1}$$

Donde d_i es la distancia entre un par átomos equivalentes, normalmente expresada en Angstroms ($\mathring{A}$). Dos estructuras idénticas poseen un RMSD que puede estar entre 0-0.5 Å, que es el valor estimado del error cristalográfico al comparar dos estructuras de la misma proteína en las mismas condiciones experimentales [Burra et al., 2009]. Este valor de RMSD va aumentando a medida que las estructuras difieren entre sí. Esta medida es considerada un indicador confiable de variabilidad estructural cuando se calcula entre confórmeros de una misma proteína y por lo tanto, estas diferencias reflejan de modo directo los movimientos de la proteína en su estado nativo.

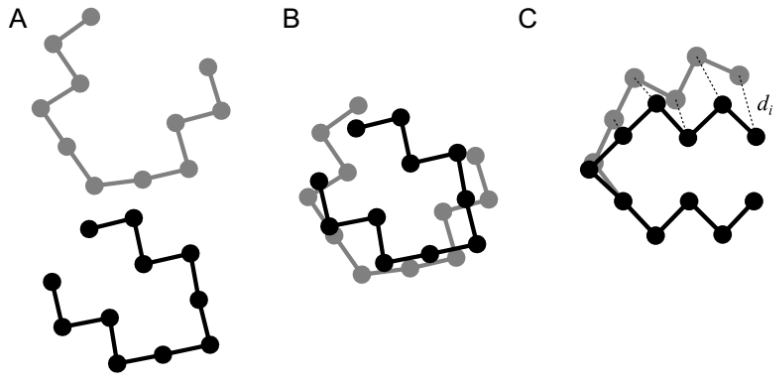

Figura 2.3: En el panel A se representan dos confórmeros de una proteína. Los dos confórmeros se trasladan al mismo centro de masa (B) para luego rotar una de las estructuras hasta minimizar el RMSD de los C_α de los residuos equivalentes (panel C)

2.2.3. Bases de datos de diversidad conformacional

En los últimos años en *Protein Data Bank* (PDB) [Berman, 2000], la base de datos de referencia de estructuras experimentales de proteínas, se han depositado un gran número de estructuras de proteínas de las cuales muchas de ellas son redundantes. Es decir, que tenemos la misma secuencia (proteína o cadena) presente en más de una estructura experimental. Por consiguiente, la extensión y distribución de la diversidad conformacional puede ser explorada en un gran número de proteínas, no accesibles mediante las metodologías mencionadas anteriormente. Esto dio lugar a que se empezaran a generar bases de datos en donde se estudiaba el movimiento de una proteína a partir de estructuras redundantes de la misma. En la tabla

2.1 se listan en orden cronológico, las principales bases de datos de diversidad conformacional de proteínas generadas a partir de estructuras redundantes.

Base de datos	**URL**	**Cantidad de proteínas**
MolMovDB [Gerstein and Krebs, 1998]	http://www.molmovdb.org	230
DynDom [Lee et al., 2003]	http://fizz.cmp.uea.ac.uk/dyndom/	1,580
PCDB [Juritz et al., 2011]	Obsoleta	4,171
PSCDB [Amemiya et al., 2012]	http://idp1.force.cs.is.nagoya-u.ac.jp/pscdb/	839
CoDNaS [Monzon et al., 2013]	http://ufq.unq.edu.ar/codnas/	12,684
CoDNaS 2.0 [Monzon et al., 2016]	http://ufq.unq.edu.ar/codnas/	21,152
PDBFlex [Hrabe et al., 2015]	http://pdbflex.org/	28,939

Tabla 2.1: Bases de datos de estructuras redundantes

Las bases de datos MolMovDB y DynDom han quedado prácticamente obsoletas y se dejaron de actualizar hace bastante tiempo, por lo cual poseen muy pocas proteínas y no permiten hacer estudios a gran escala. Sin embargo, aún mantienen una clasificación del movimiento entre confórmeros que puede ser de gran utilidad. La base de datos PCDB es una base de datos de diversidad conformacional de dominios de proteínas, desarrollada en el grupo de bioinformática estructural de la Universidad Nacional de Quilmes. Esta base de datos, actualmente obsoleta, contenía cerca de 8,000 dominios proteicos derivados de la base de datos CATH [Sillitoe et al., 2015], con un promedio de 4.7 estructuras cristalográficas por dominio representado, lo que da un total de más de 36,000 estructuras. Esta base de datos se utilizó para explorar la extensión y diversidad conformacional de distintos dominios proteicos. Un año más tarde, surge la base de datos PCSCB donde se incluyeron 839 pares de confórmeros de distintas proteínas y se describe el movimiento asociado a la unión del ligando. Esta base de datos a pesar de ser única en su tipo hasta ese momento, no incluía proteínas

que experimentaban cambios conformacionales por otros factores biológicos como podrían ser, cambios de pH, mutaciones, modificaciones post-traduccionales, etc. Además, no se continuó con su actualización por lo que sus datos son muy escasos. En el año 2012 y ante la necesidad de una base de datos global que abarque todas las proteínas con diversidad conformacional conocida, comenzamos el desarrollo de una base de datos de diversidad conformacional usando todas las estructuras redundantes depositadas en PDB para una misma proteína. El desarrollo e implementación de esta base de datos se encuentra abarcado en esta tesis doctoral, y dio origen a la base de datos *Conformational Diversity in the Native State (CoDNaS)* [Monzon et al., 2013, Monzon et al., 2016].

2.3. Construcción de la base de datos

CoDNaS nos permite estudiar el estado nativo proteico desde un punto de vista dinámico y considerándolo como un conjunto de confórmeros en equilibrio. Para ello se ha desarrollado un servidor que se encuentra disponible en línea en http://http://ufq.unq.edu.ar/codnas/, el cual permite reclutar proteínas con distinto grado diversidad conformacional. Cada proteína que se encuentra en la base de datos está representada como un conjunto de estructuras, las cuales han sido comparadas entre sí mediante diferentes parámetros estructurales que nos permiten identificar similitud y/o diferencia entre cada uno de los confórmeros. En la Figura 2.4 se esquematizan los pasos efectuados para el desarrollo la base de datos, los cuales se van a detallar en este capítulo. Además, CoDNaS se encuentra vinculada con otras bases de datos que permiten relacionar información biológica relevante de cada proteína, así como también, posee información acerca de las condiciones experimentales en la que se obtuvo cada confórmero (estructura). De esta manera, se han identificado diversos factores que afectan a la diversidad conformacional del estado nativo proteico. Así, es posible efectuar distintos análisis teniendo en cuenta el aspecto dinámico de las proteínas y relacionarlo con su función, evolución, estructura, etc.

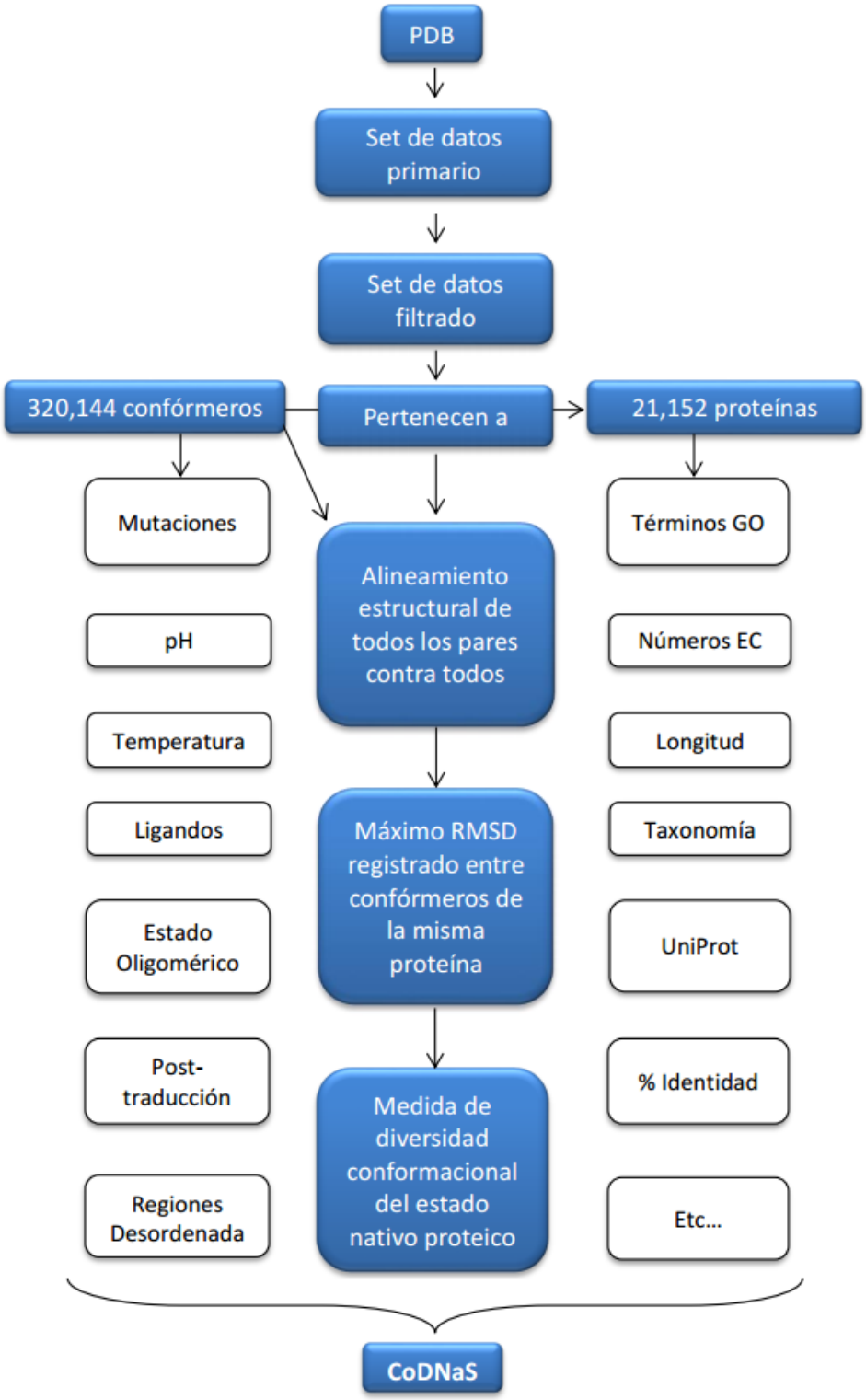

Figura 2.4: Esquema representativo del desarrollo de CoDNaS

2.3.1. Reclutamiento y filtrado de estructuras

La base de datos PDB [Berman, 2000] es un repositorio a nivel mundial de todas las estructuras tridimensionales de macromoléculas orgánicas. En la base de datos se encuentran los archivos con formato "pdb" que contienen las coordenadas de cada uno de los átomos de la estructura, junto con información relevante de esa proteína como el organismo, condiciones experimentales, autores, substratos, entre otros. Dentro de un archivo "pdb" podemos tener estructuras de: proteínas, complejos proteicos, complejos proteína - ADN, etc. Una proteína puede estar obtenida en diferentes estados oligoméricos (homos y heteros) e incluso en complejo con otra proteína, por lo que dentro de un mismo archivo podemos tener una o más cadenas polipeptídicas iguales o distintas.

Para identificar aquellas cadenas de proteínas que poseen estructuras redundantes depositadas en PDB, se utilizó el algoritmo BLASTclust [Altschul et al., 1990] que permite encontrar *clusters* secuenciales a un determinado corte de porcentaje de identidad secuencial. Partiendo de un archivo formato fasta con todas las secuencias proteicas presentes en PDB se generaron *clusters* al 95 % de identidad de secuencia local, con un *coverage* mínimo del 90 % entre todas las secuencias de un determinado clúster. La elección del 95 % de identidad en lugar del 100 % fue para incluir aquellos confórmeros que pueden tener variaciones mínimas o mutaciones puntuales en la secuencia de la proteína. Sin embargo, para evitar la inclusión de confórmeros de proteínas homólogas en cada cluster generado al 95 % de identidad, se verificó que todas las estructuras dentro del clúster posean el mismo código UniProt. Recordemos brevemente que UniProt asigna un código específico a cada producto génico identificando a una determinada proteína de un determinado organismo. En aquellos casos donde al menos una estructura difería en su código UniProt, se removieron esos *clusters* y se buscaron si esas estructuras que los componían formaban *clusters* al 100 % de identidad, para poder abarcar la mayor cantidad de estructuras de PDB.

Una vez generados e identificados los *clusters*, solo se consideraron aquellos que poseían al menos dos cadenas proteicas (de ahora en adelante, confórmeros) provenientes de dos archivos "pdb" diferentes. Es decir que en al menos dos instancias experimentales independientes se depositó esa secuencia en PDB. Además, se removieron de cada clúster los confórmeros que

poseían una resolución <4.5 $\mathring{A}$, y para aquellos obtenidos por RMN se utilizaron los primeros 20 modelos (en caso que hubiera más de 20 informados en PDB). Cada clúster representa una entrada en la base de datos CoDNaS y corresponde a una única secuencia proteica con sus respectivos confórmeros.

2.3.2. Estimación de las diferencias estructurales entre confórmeros

Para estimar la extensión de la diversidad conformacional de cada proteína de la base de datos se calculó el RMSD para todos los pares de confórmeros posibles de cada una de ellas. Se utilizaron tres algoritmos diferentes que realizan la superposición, el alineamiento estructural y el cálculo del RMSD. Se detallan a continuación:

- MAMMOTH (*"Matching Molecular Models Obtained from Theory"*): el programa fue originalmente desarrollado para comparar modelos estructurales estimados teóricamente por métodos computacionales, con aquellos obtenidos experimentalmente [Ortiz et al., 2002]. Se basa en un método heurístico para encontrar, en un modo independiente de la secuencia y por superposición rígida, el subconjunto máximo entre dos proteínas con el mismo *backbone* y conformación tridimensional. El algoritmo consiste en descomponer la estructura de la proteína en péptidos cortos, de siete residuos, que se comparan con los heptapéptidos de la otra proteína. El score de similitud entre los heptapéptidos de las dos estructuras se calcula utilizando un vector unidad de RMSD. Estos scores son almacenados en una matriz de similitud, y mediante programación dinámica híbrida (local-global) se calcula el alineamiento óptimo de cada residuo. La salida del programa permite obtener entre otras cosas el RMSD y su significancia estadística, la cantidad de residuos alineados y el alineamiento secuencial a partir del estructural. Es un programa de rápida ejecución, ampliamente utilizado en el área de bioinformática estructural y con una excelente precisión.

- TM-Score (*"Template Modeling Score"*): es un algoritmo que permite calcular la similitud de topologías entre dos estructuras proteicas [Zhang and Skolnick, 2004]. Se puede

utilizar para cuantificar la calidad de una estructura proteica obtenida por predicción, contra otra homóloga obtenida experimentalmente, entre otras aplicaciones. Nos brinda un valor de TM-score, el cual da mayor peso a las coincidencias estructurales que son cercanas espacialmente antes que a las que se encuentran distantes, dándonos una idea de similitud entre dos estructuras. Por lo que resulta menos sensible que el RMSD para detectar disimilitudes. El rango del TM-score es entre 0 y 1, indicando de menor a mayor el grado de similitud topológica entre el par de estructuras. Además, el programa permite calcular otros parámetros estructurales como el *"Global Distant Test"* (GDT) y en RMSD. El GDTTS (*"GDT total score"*) se utiliza para comparar estructuras que comparten la misma secuencia de aminoácidos, buscando el número máximo de residuos que sepuedan superponer a diferentes umbrales de distancia (1, 2, 4 y 8 $\mathring{A}$). Usando umbrales más bajos de distancia (0, 5, 1, 2 y 4 $\mathring{A}$) puede calcularse el GDTHA (*"GDT high accuracy"*) que es más sensible a pequeños cambios en la estructura [Keedy et al., 2009]. Este programa se utilizó debido a que proporciona otros parámetros estructurales, diferentes al RMSD, los cuales pueden ser de utilidad para futuros análisis, además de ser *scores* muy utilizados en el área de modelado de estructuras de proteínas.

- ProFit (*"Protein least-squares fitting"*): es un programa que se basa en el algoritmo de superposición de estructuras de McLachlan [McLachlan, 1982], la característica que posee a diferencia de los otros programas es que permite ajustar diversos parámetros a criterio del usuario, para realizar el alineamiento estructural. Entre ellos, permite definir regiones de residuos que se quieren alinear, especificar residuos puntuales, calcular el RMSD por posición de cada par de residuos alineados, entre otros. Se decidió utilizarlo ya que nos permitía calcular el RMSD por posición para cada par de confórmeros. El RMSD por posición se utiliza en CoDNaS para derivar el correspondiente Z-score posición, el cual indica cómo varía ese residuo respecto al valor de RMSD promedio de los demás residuos alineados en ese par de estructuras. Para cada valor de RMSD entre el par de aminoácidos alineados (i, j) (RMSD_{ij}) se utiliza la siguiente ecuación 2.2 para calcular el Z-score:

$$Z_{ij} = \frac{RMSD_{ij} - \mu}{\sigma} \tag{2.2}$$

Donde μ es el valor de RMSD promedio y σ el desvío estándar de los valores de RMSD por cada par de residuos alineados en entre dos confórmeros.

Para cada proteína de CoDNaS se identificó el par de confórmeros que maximiza el RMSD, al que denominamos "par máximo" o "par de máximo RMSD" y es una medida de la extensión de la diversidad conformacional de esa proteína. A lo largo de este trabajo mencionaremos en varias oportunidades éste par, ya que define la magnitud de la diversidad conformacional en una secuencia en particular.

Adicionalmente, utilizando los RMSD de todas las comparaciones posibles entre confórmeros, se convirtieron a una matriz de distancias y utilizando un algoritmo de agrupamiento jerárquico se identificó en cada proteína sub-poblaciones de confórmeros ("estados conformacionales") que difieren entre sí. Para ello se utilizó el paquete *hclust* del lenguaje de programación R [Team, 2017].

2.3.3. Asignación de las condiciones experimentales de cada estructura

Cada confórmero en CoDNaS está caracterizado de acuerdo a las condiciones experimentales en la que se obtuvo su estructura: el pH, temperatura, presencia de ligandos y mutaciones, estado oligomérico, modificaciones post-traduccionales y presencia de desorden. Esta información fue asignada utilizando fuentes de datos externas y *scripts* desarrollados en el lenguaje de programación PERL. La temperatura, pH y variaciones en la secuencia fueron extraídas del encabezado del archivo "pdb". Los ligandos fueron asignados usando la información de los heteroátomos dentro del archivo "pdb" y con la base de datos BioLip [Yang et al., 2013]. BioLip es una base de datos curada semi-manualmente que identifica y diferencia aquellos ligandos que son biológicos respecto de los que se utilizan en muchos casos para estabilizar la proteína dentro del cristal. Las modificaciones post-traduccionales fueron identificadas usando la entrada MODRES que poseen los archivos "pdb", mientras que para el estado oligomérico se utilizó el proporcionado por PISA [Krissinel and Henrick, 2007] y el que informa el autor.

Finalmente la presencia de desorden en cada estructura se asignó a partir de los residuos faltantes o *"Missing Residues"* presentes en la estructura (esto se discutirá más en profundidad en el capítulo 4).

2.3.4. Vinculación con otras fuentes de información biológica

Las proteínas comprendidas en CoDNaS han sido vinculadas con otras bases de datos biológicas y datos de programas externos con el fin de enriquecer la información disponible al usuario; además de permitir correlacionar los parámetros estructurales con información biológica y físico-química de las proteínas y sus confórmeros. Cada confórmero se vinculó con otras bases de datos utilizando la plataforma SIFTS (*"Structure integration with function, taxonomy and sequence"*) [Velankar et al., 2013], que mapea cada una de las estructuras de PDB a diferentes bases de datos. Entre ellas las que se detallan a continuación:

- UniProt (`http://www.uniprot.org/`) [Consortium, 2017]: es la base de datos de referencia que almacena todas las secuencias de proteínas conocidas, junto con un amplia variedad de información funcional. Contiene una gran cantidad de información biológica para cada proteína que se deriva de la literatura y mediante algoritmos de anotación automática. Se encuentra subdividida en cuatro secciones, UniProtKB (Swiss-Prot y TrEMBL), UniParc, UniRef, and UniMes. Con el fin de reducir la redundancia y maximizar la fidelidad de las secuencias, todas las proteínas codificadas por el mismo gen son agrupadas en una misma entrada UniProt. Las proteínas que se encuentran en CoDNaS han sido asociadas con su correspondiente código de UniProt. De esta manera el usuario puede vincular toda la información que contiene UniProt, con los datos estructurales de diversidad conformacional.

- *Enzyme Comission number* (EC): aquellas proteínas de CoDNaS que poseen actividad enzimática se las asoció con su correspondiente número EC que proviene de la clasificación de la *Enzyme Commission* [Kotera et al., 2004]. Es un esquema de clasificación numérico para enzimas, basado en las reacciones químicas que catalizan. Cada código de enzimas consiste en 4 números separados por puntos. Estos números representan una clasificación progresivamente más específica.

- *Gene Ontology* (GO) [Ashburner et al., 2000]: el proyecto de ontología génica (`http://www.geneontology.org/`) provee un vocabulario controlado que describe el gen y los atributos del producto génico en cualquier organismo. Consta básicamente de tres ontologías, donde cada una representa un concepto de biología molecular: la función molecular de los productos génicos, su rol en los procesos biológicos de múltiples direcciones y su localización en componentes celulares. Las proteínas de la base de datos CoDNaS han sido asociadas con sus correspondientes términos GO en las tres ontologías mencionadas anteriormente. Además se incluyeron los identificadores numéricos de los términos GO.

- CATH (*"Class - Architecture - Topology - Homology"*) [Sillitoe et al., 2015]: identifica en las estructuras depositadas en PDB dominios estructurales relacionados evolutivamente. Estos dominios se encuentran clasificados dentro de la jerarquía estructural que propone CATH: la clase (C), donde los dominios son asignados de acuerdo a el contenido de estructura secundaria que poseen; la arquitectura (A) como los elementos de estructura secundaria se disponen en la estructura tridimensional; la topología (T), como los elementos de estructura secundaria se conectan y disponen entre sí: la superfamilia de homólogos (H), si hay evidencia que esos dominios están relacionados evolutivamente es decir que sean homólogos. Cada proteína en CoDNaS posee asignado el/los código/s de la/s superfamilia/s de CATH a la que pertenecen.

Adicionalmente, a cada confórmero se le calculó el área superficial accesible al solvente (ASA) con NACCESS [Lee and Richards, 1971]. No sólo se informa el cambio de ASA global entre cada par de confórmeros, sino también la diferencia de ASA relativo.

2.4. Servidor web de CoDNaS

2.4.1. Implementación del servidor

La base de datos CoDNaS se encuentra disponible en línea en (`http://ufq.unq.edu.ar/codnas`) y es de acceso libre y gratuito para su uso académico.

El servidor web fue diseñado con el fin de que el usuario pueda obtener proteínas con diversidad conformacional en el estado nativo y relacionar las mismas con un gran espectro de parámetros e información biológica. Permite realizar la búsqueda en forma personalizada, permitiéndole al usuario restringir los resultados de la misma en función de diversos parámetros: extensión de diversidad conformacional, factores que afectan a la diversidad conformacional, método experimental de obtención de los confórmeros, porcentaje de identidad de secuencia, etc. Además, permite buscar proteínas por su código PDB, su código Uniprot, utilizando una secuencia y por el nombre de la proteína.

El servidor fue implementado en una arquitectura denominada LAMP (Linux, Apache, MySQL y PHP) (Figura 2.5). La mayoría de los datos están almacenados en una base de datos relacional, compuestas por cinco tablas, la cual fue construida usando el servidor de bases de datos MySQL 5.5.43 (`www.mysql.com`). Además, algunos datos son obtenidos usando los servicios web RESTful de la base de datos UniProt. La interfaz gráfica fue desarrollada en HTML, CSS y JavaScript, usando los *frameworks* JQuery y Bootstrap. En el servidor las programación se realizó en PHP 5.3.3 y la comunicación con el cliente muchas veces se hizo mediante AJAX.

Las consultas se envían a la base de datos a través del sitio web y se generan dinámicamente del lado del servidor de acuerdo a los criterios definidos por el usuario. Luego, son enviadas al motor de base de datos MySQL, para que se ejecute la consulta y devuelva una tabla resultante. El resultado es procesado en el lado del servidor y mostrada en un formato accesible al usuario. Generalmente los datos están organizados en diferentes secciones que permiten ver fácilmente la información de una determinada proteína.

La base de datos, estructuras de los confórmeros y demás datos que visualizan los usuarios en CoDNaS se encuentran alojados en un servidor propio del grupo de Bioinformática Estructural en la Universidad Nacional de Quilmes. El mantenimiento del servidor se realiza en conjunto con el personal de la universidad y la administración y puesta en funcionamiento, lo realizamos los miembros del grupo de investigación.

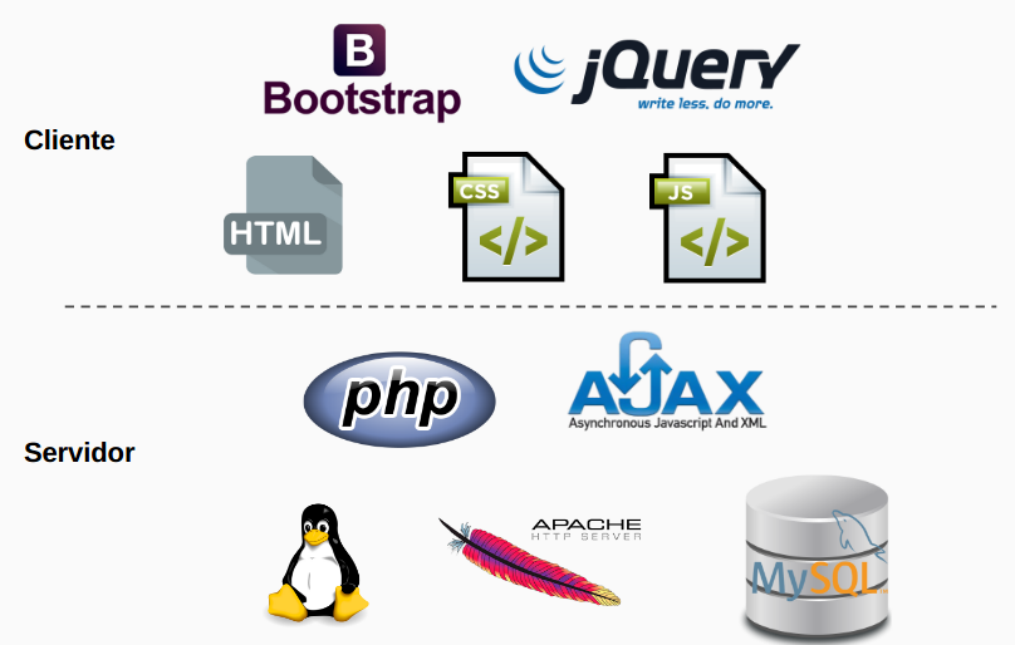

Figura 2.5: Tecnologías utilizadas en implementación del servidor web de CoDNaS

2.4.2. Búsqueda de una proteína en CoDNaS

Para facilitar la búsqueda en la base de datos, el servidor web ofrece diferentes métodos de búsqueda. El método más rápido y simple es buscar una proteína mediante su código PDB o UniProt o el nombre de la misma, en el cuadro de búsqueda que se encuentra en la página principal. La segunda forma de buscar en la base de datos es utilizando la página de "Búsqueda" (accesible desde el menú principal). En esta sección podemos buscar una proteínas filtrando por los siguientes criterios (Figura 2.6):

- Extensión de la diversidad conformacional de una proteína: se pueden fijar los límites del RMSD máximo que se obtiene al comparar todos los confórmeros entre sí de cada una de las proteínas de la base de datos.

- Causas de la diversidad conformacional: se pueden filtrar proteínas asociadas con una o varias causas de diversidad conformacional como presencia de ligandos, mutaciones, cambios de pH, etc.

- Método experimental de obtención de los confórmeros: uno puede obtener proteínas cuyos confórmeros sean obtenidos por DRX o RMN.

- Búsqueda secuencial: el usuario puede ingresar una secuencia proteica o de ADN. El

servidor realizará un BLAST contra todas las secuencias de las proteínas presentes en CoDNaS y listar aquellas proteínas que resulten como *hit* con un valor E mejor que $1E^{-04}$.

Además, la base de datos posee un *browser* que permite navegar la base de datos de acuerdo a los niveles de jerarquía de la base de datos CATH. Permitiendo así encontrar proteínas homólogas y/o que compartan un determinado plegamiento.

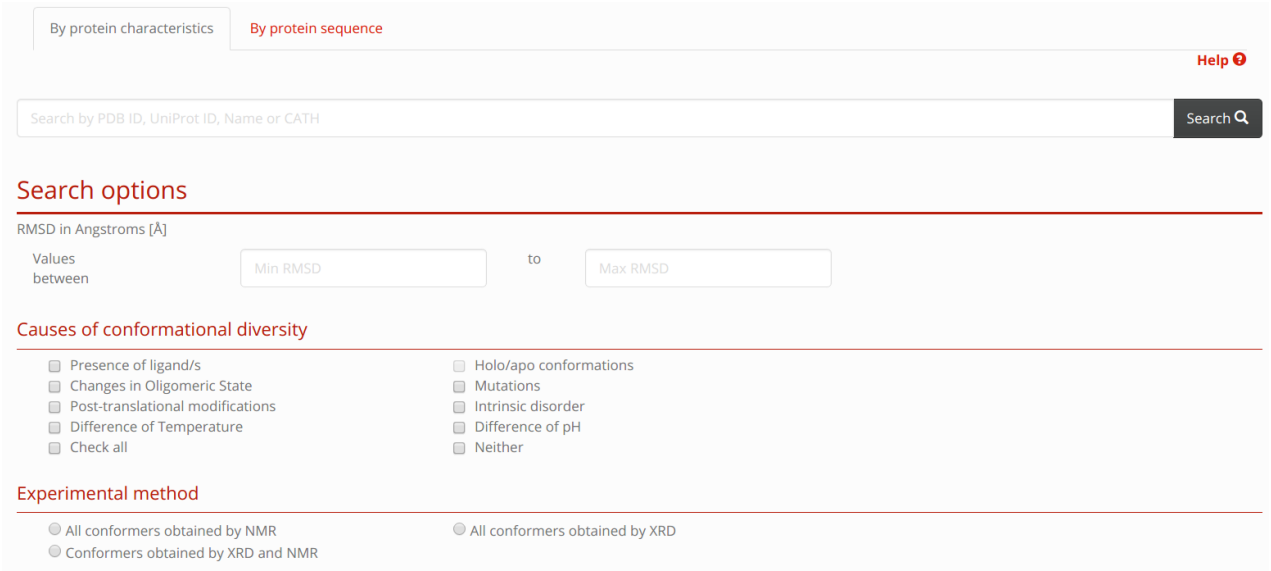

Figura 2.6: Sección de búsqueda avanzada en el servidor de CoDNaS. Posee dos pestañas que permiten elegir entre hacer una búsqueda por diferentes filtros y características o realizar una búsqueda secuencial usando BLAST.

2.4.3. Página de resultados

Una vez realizada la búsqueda por alguno de los métodos mencionados anteriormente, en la página de resultados se muestra una tabla con las proteínas que han sido encontradas (Figura 2.7). La tabla contiene siete columnas fijas: el ID de CoDNaS, el código UniProt de la proteína, la cantidad de confórmeros y el valor de RMSD mínimo, promedio y máximo de la comparación entre todos los confórmeros de esa proteína. Dependiendo del método utilizado para realizar la búsqueda, puede aparecer una columna extra con la jerarquía de CATH (en el caso de que se haya utilizado el *browser*), o una columna con el valor E de BLAST. La tabla permite ser ordenada por cualquier columna que el usuario seleccione y posee un cuadro de búsqueda para encontrar rápidamente alguna proteína por su nombre. De esta manera se

puede encontrar y seleccionar fácilmente una proteína y acceder a la página principal de la misma presionando en la fila correspondiente.

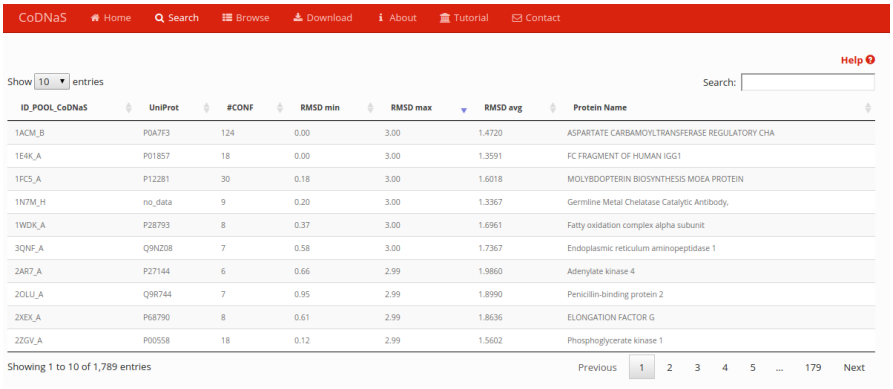

Figura 2.7: En este ejemplo se han obtenido proteínas con un RMSD máximo entre 2 y 3 $\mathring{A}$, y que todos sus confórmeros hayan sido obtenidos por DRX. La tabla muestra los 10 primeros resultados y por defecto ordena las proteínas por la primer columna que es el ID de CoDNaS.

2.4.4. Página principal para una proteína en CoDNaS

Luego que se selecciona e ingresa a la página principal de una proteína, el servidor muestra toda la información que posee de la misma en la base de datos como se muestra en la Figura 2.8 para la Adelinato Kinasa 4. Esta página incluye una serie de cajas que pueden ser colapsables para optimizar el espacio visual disponible en la página. La caja de información general (Figura 2.8a) contiene dos secciones; una con información general de esa proteína (código Uniprot, nombre, organismo, términos GO, número EC y función) y otra con información estructural de los alineamientos entre confórmeros y sus estructuras (resolución, método experimental, cantidad de confórmeros y porcentajes de similitud secuencial entre los confórmeros). La figura 2.8b muestra la sección de confórmeros, donde en una tabla se listan todos los confórmeros correspondientes a esa proteína y las condiciones experimentales en la que se obtuvo cada estructura. El usuario puede seleccionar los confórmeros que le interesen y obtener toda la información de las comparaciones estructurales entre ellos (detalles adicionales se describirán en la sección 2.4.5).

La siguiente caja (Figura 2.8c) muestra los dendrogramas resultantes del *clustering* jerárquico realizado con los valores de RMSD obtenidos de comparar todos los pares de confórmeros de esa proteína. De ésta manera se puede visualizar cuáles conformaciones son similares entre si y cuales no. Finalmente, la última caja contiene información relevante acerca del par de

confórmeros que poseen el máximo RMSD (Figura 2.8d). Incluye una tabla con información estructural, anotaciones biológicas y las condiciones experimentales de ambos confórmeros. Para ver mas detalles de éste par, el usuario puede presionar el botón *"View details"* en el extremo superior de esta caja (detalles adicionales se describirán en la sección 2.4.5).

2.4.5. Comparación entre pares de confórmeros

Esta nueva sección del servidor web de CoDNaS 2.0 permite al usuario explorar, analizar y visualizar todos los pares de confórmeros de una proteína en particular. La página está compuesta de cinco secciones que se detallarán a continuación (Figura 2.9):

- La sección *"Comparison of selected conformers"* (Figura 2.9a), incluye una tabla con todos los pares de confórmeros posibles, entre los que han sido seleccionados, con diferente información estructural como el RMSD, número de residuos alineados y GDT-TS. Cuando el usuario presiona en alguna de las filas de ésta tabla, el servidor calcula los RMSD por posición entre ese par de confórmeros, mapea los valores en las estructuras que se van a observar en el visualizador y muestra las cuatro secciones siguientes.

- La sección *"Structures"* (Figura 2.9b) incluye la herramienta de visualización JSmol [Hanson et al., 2013], la cual permite al usuario ver las estructuras tridimensionales de los confórmeros superpuestas. En el panel derecho contiene diferentes opciones de visualización que permiten, por ejemplo, colorear las estructuras de acuerdo el Z-score del RMSD por posición o visualizar el/los ligando/s en caso que posean. Además, las estructuras superpuestas se pueden descargar usando los links que se encuentran en la esquina superior derecha de esta sección.

- La sección *"Z-score profile by position"* que se observa en la Figura 2.9c muestra un gráfico interactivo con los perfiles de Z-score derivados del RMSD por posición y del RMSD entre los factores-B como dos medidas de la flexibilidad local de una proteína [Smith et al., 2003]. El usuario puede posicionarse sobre cualquier posición en el gráfico para ver el valor numérico exacto.

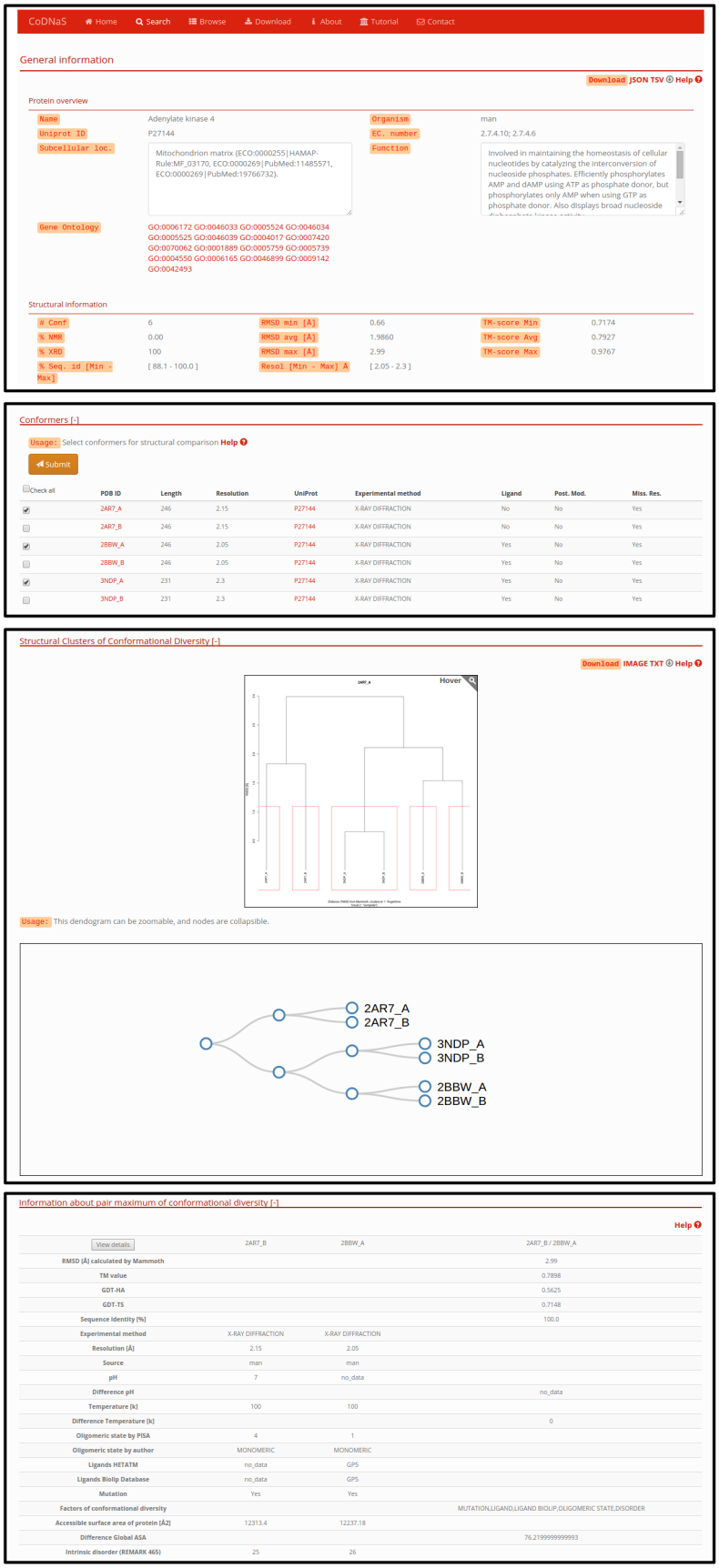

Figura 2.8: Página principal de la proteína Adenilato quinasa 40. a)Información general de la proteína, anotaciones biológicas e información estructural. b) Información detallada de cada confórmero, el usuario puede seleccionar los confórmeros que desee para recuperar todos los pares que resultan de su comparación. c) *Clustering* jerárquico utilizando los valores de RMSD entre todos los pares de confórmeros. Los dendrogramas muestran los *clusters* de confórmeros similares, es posible hacer zoom y contraer los nodos en el dendrograma en la parte inferior de esta sección. La información de los clusters y la imagen se pueden descargar con los enlaces en la parte superior derecha de esta caja. d) Descripción del par de máxima diversidad conformacional de la proteína en cuestión.

- *"Sequence alignment"* (Figura 2.9d) muestra el alineamiento de secuencia resultante de la superposición estructural entre ese par de confórmeros.

- La última sección *"Information about pair of conformers"*, es una tabla con información estructural y biológica del par de confórmeros seleccionado. Incluye el TM-score, GDT, resolución, condiciones experimentales, presencia de ligando y desorden intrínseco. Esta tabla permite al usuario explorar fácilmente las diferencias entre los confórmeros y las implicaciones biológicas de la diversidad conformacional.

2.4.6. Descarga de datos en CoDNaS

CoDNaS posee una sección de descargas (Figura 2.10) donde los usuarios pueden buscar por uno o varios códigos PDB y recuperar todos los pares confórmeros incluidos en la base de datos donde estén estas estructuras. El constructor de descarga permite al usuario generar un archivo descargable que contiene toda la información de interés y el formato de salida de los datos es un archivo separado por tabuladores (TSV). La mayoría de las secciones en diferentes vistas de la base de datos permiten la descarga de resultados de búsqueda en formato TSV o JSON (*"JavaScript Object Notation"*).

2.5. Conclusiones

Estudiar la función proteica implica necesariamente tener en cuenta los movimientos de las proteínas. CoDNaS es una base de datos de conformaciones de proteínas bien curada, con datos experimentales y relacionada con las bases de datos de referencia en proteínas. La evaluación de las diferentes conformaciones de una misma proteína permite extraer información clave en función de los movimientos proteicos y motiva la exploración de las relaciones entre la dinámica de las proteínas, las condiciones experimentales y sus propiedades biológicas.

CoDNaS incluye 73 % de todas las estructuras de proteínas disponibles en PDB, se actualiza periódicamente de acuerdo al crecimiento de la redundancia en PDB. La última versión de la base de datos posee 21,152 cadenas de proteínas, un total de 320,144 confórmeros y más de 20 millones de alineamientos estructurales de a pares. Las diferencias estructurales

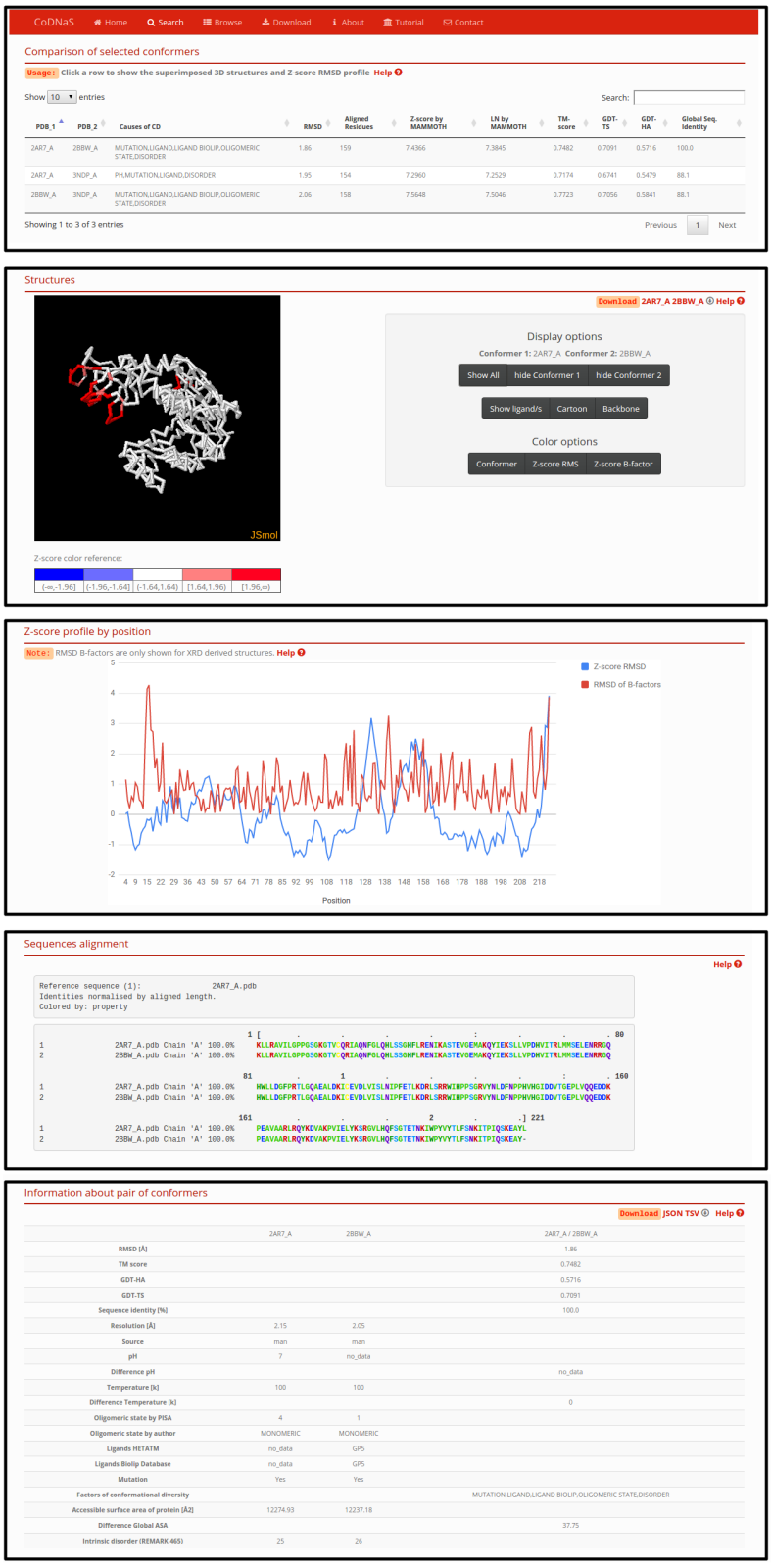

Figura 2.9: Página de comparación entre pares de confórmeros. A) Todas las comparaciones entre los confórmeros seleccionados se muestran en una tabla con las correspondientes medidas estructurales. B) S visualizan las estructuras 3D del par de confórmeros elegido. Es posible cambiar el método de coloración de las estructuras (confórmeros, Z-score del RMSD por posición y Z-score del RMSD entre los factores-B), como se observa por Z-score del RMSD por posición. C) Gráfico del Z-score por posición del RMSD y de la diferencia entre los factores-B. D) Tabla descriptiva y comparativa del par de confórmeros seleccionado.

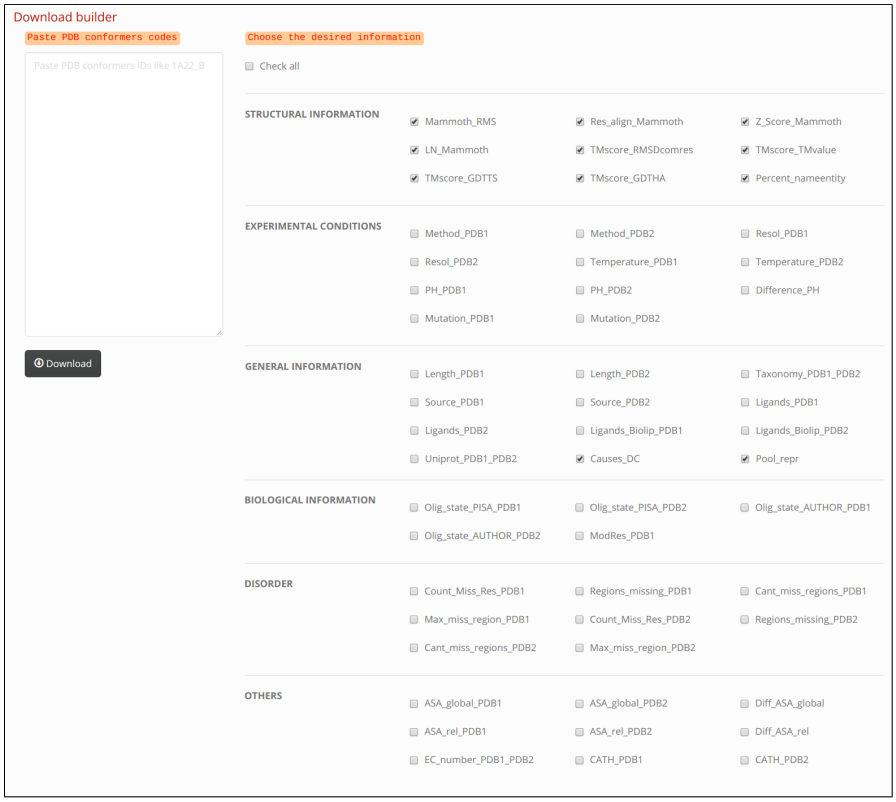

Figura 2.10: Generador de descargas en CoDNaS.

del *backbone* y de la superficie proteica están muy bien caracterizadas en CoDNaS, y estamos planeando incorporar parámetros que caracterizan la diversidad conformacional en gran parte independiente de los movimientos del *backbone*, tales como los relacionados con la apertura y cierre de túneles y los asociados con cavidades o bolsillos en la proteína. Una futura actualización incorporará información importante para estos movimientos, que son significativos para la función biológica de las proteínas. CoDNaS no sólo es parte fundamental de esta tesis doctoral, sino que además posibilitó avanzar en las diferentes líneas de investigación que posee el grupo en diversidad conformacional de proteínas, dando lugar a diferentes trabajos con colaboradores nacionales e internacionales (promiscuidad, reconstrucción ancestral y origen de nuevas funciones, estudio de enfermedades y el impacto de mutaciones puntuales, etc).

Capítulo 3

Descripción de los datos presentes en CoDNaS

3.1. Resumen

En este capítulo describiremos brevemente los datos presentes en la última versión en línea de la base de datos CoDNaS. Se mostrará la distribución de RMSD para todas las proteínas presentes en la base de datos y cómo este RMSD puede estar o no afectado por factores biológicos y no biológicos. Describiremos información de relevancia acerca de las estructuras que componen CoDNaS como lo son su resolución y condiciones experimentales, como así también la cantidad de confórmeros por proteína y taxonomía. Finalmente, comentaremos cómo está representada la redundancia secuencial en PDB, ya que es la principal fuente de datos de CoDNaS.

3.2. Distribución de la diversidad conformacional

La distribución de la extensión de la diversidad conformacional de cada proteína en CoDNaS, puede estudiarse observando los pares de máximo RMSD de cada una de ellas. En la Figura 3.1 se observa la distribución de máximo RMSD de todas las proteínas presentes en CoDNaS (21,152 cadenas de proteínas presentes en la última versión). El valor máximo de RMSD recordemos que se obtiene comparando estructuralmente todos los confórmeros para

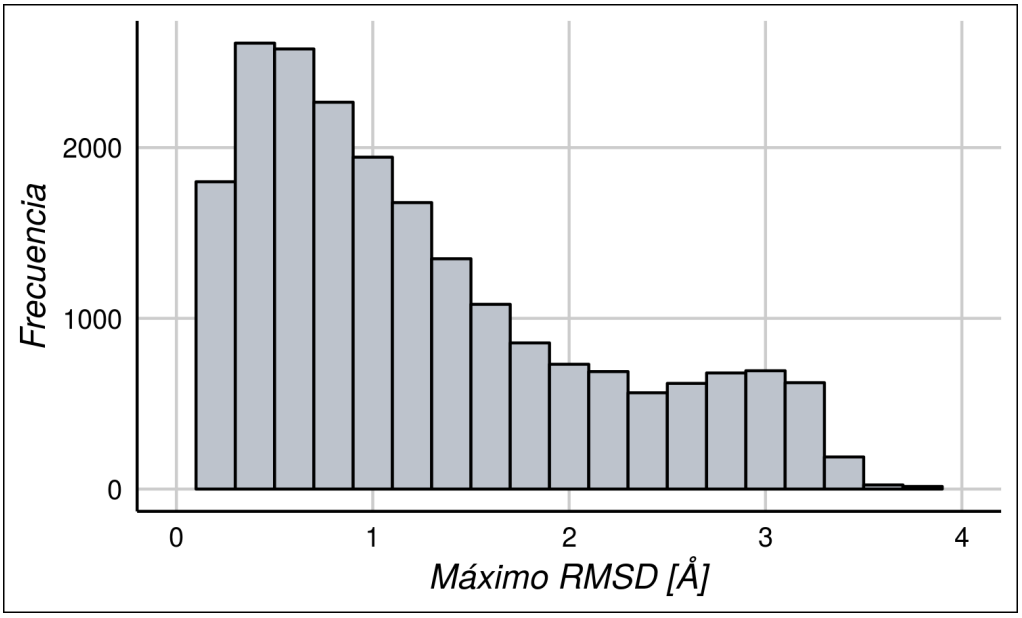

Figura 3.1: Distribución de la extensión de la diversidad conformacional de todas las proteínas en CoDNaS.

una determinada proteína entre sí y registrando el valor máximo de RMSD. Se puede ver que una gran cantidad de proteínas presentan valores de máximo RMSD entre 0 y 1 $\mathring{A}$, siendo la mediana de esta distribución 1.02 $\mathring{A}$ y presentando un pico en valores entre 0.4-0.5 $\mathring{A}$. Esto coincide con la distribución encontrada anteriormente por Burra [Burra et al., 2009] y que la analizaremos con mayor profundidad en el capítulo 4, donde veremos que diferentes relaciones estructura-dinámica-función emergen de la misma.

La distribución mostrada anteriormente puede dividirse entre aquellas proteínas que poseen todos sus confórmeros obtenidos por DRX o por RMN, y por las que tienen confórmeros obtenidos por ambas técnicas. El 5 % de las proteínas tiene todos sus confórmeros obtenidos por RMN, el 89 % todos por DRX y el 6 % por DRX y RMN. Podemos ver en la Figura 3.2 que las proteínas obtenidas por RMN poseen valores de RMSD significativamente mayores (*Kolmogorov-Smirnov test* y *Wilcoxon rank-sum test*, P-valor $<< 0.01$) respecto de las DRX. A partir de esto podemos concluir que uno de los factores que determinan los valores de RMSD obtenidos en cada proteína, es básicamente el método de obtención de sus confórmeros. Basta con que uno de los confórmeros del par sea resuelto por NMR para mover la distribución hacia valores mayores de RMSD. Esto es lógico dado a las diferencias inherentes a cada técnica y que fueron detalladas en la sección 2.2.1. Esta diferencia en los valores de RMSD obtenidos entre confórmeros de proteínas resueltas por un método o el otro, sumado a las

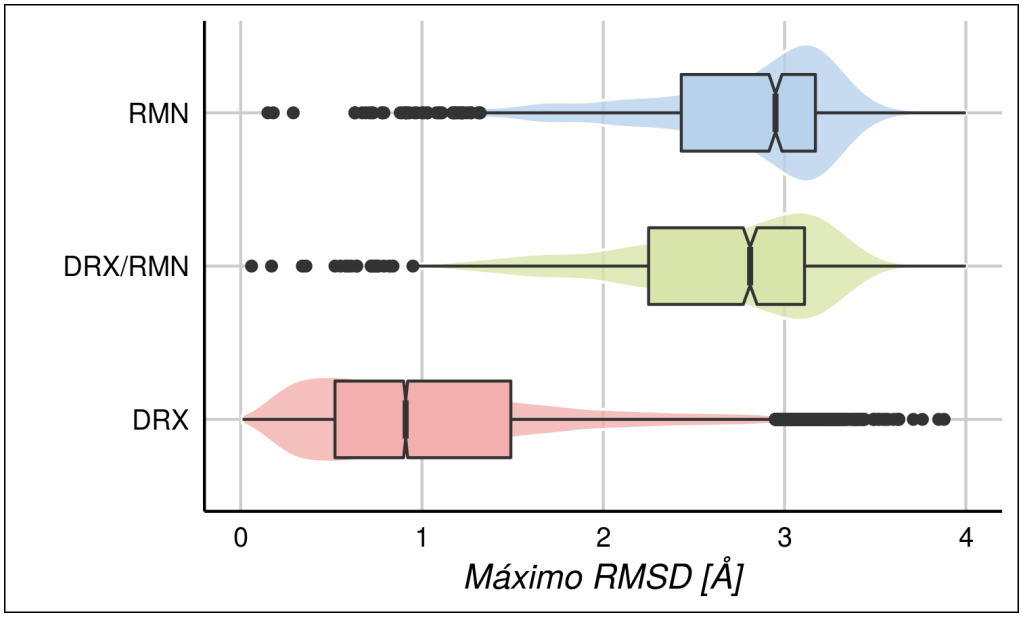

Figura 3.2: Distribución de la diversidad conformacional según el método utilizado para la obtención de los confórmeros. En azul observamos los valores de máximo RMSD de aquellas proteínas que tienen todos sus confórmeros obtenidos por RMN. En verde proteínas con confórmeros RMN y DRX y en rosado proteínas con todos sus confórmeros obtenidos por DRX.

diferencias estructurales que se evidencian entre confórmeros resueltos por métodos diferentes, hacen difícil analizar la diversidad conformacional usando más de un método de determinación estructural a la vez. Dado que la base de datos contiene un 89 % de las estructuras resueltas por DRX, es conveniente excluir de algunos análisis las estructuras resueltas por RMN.

La distribución de la Figura 3.1 puede ser analizada desde el punto de vista de los factores que afectan a la diversidad conformacional de las proteínas. Estos factores como la presencia de ligandos, mutaciones, cambio de estado oligomérico, modificaciones post-traduccionales, diferencia de pH, presencia de desorden y temperatura de cristalización, causan cambios en la conformación y por lo tanto afectan a los valores de RMSD obtenidos. Nuestro análisis sobre un grupo de 17,276 proteínas presentes en la base de datos CoDNaS muestra una distribución similar, con el pico cerca de los $0.4\mathring{A}$, como se ve en el gráfico de violín de la Figura 3.3 para todas las proteínas. Esta figura muestra además la distribución de RMSD en la base de datos CoDNaS para las diferentes condiciones o factores que causan cambios en la conformación proteica. Se ve que las mutaciones entre confórmeros generan una distribución bimodal del RMSD, con un pico cerca de los 0.8 $\mathring{A}$ y el otro cerca de los 3.0 $\mathring{A}$, presentando una mediana más alta que la del resto de las distribuciones. Como se observa en las distribuciones, la mayoría de las proteínas muestran pocos cambios en la posición de sus Cα, sin embargo

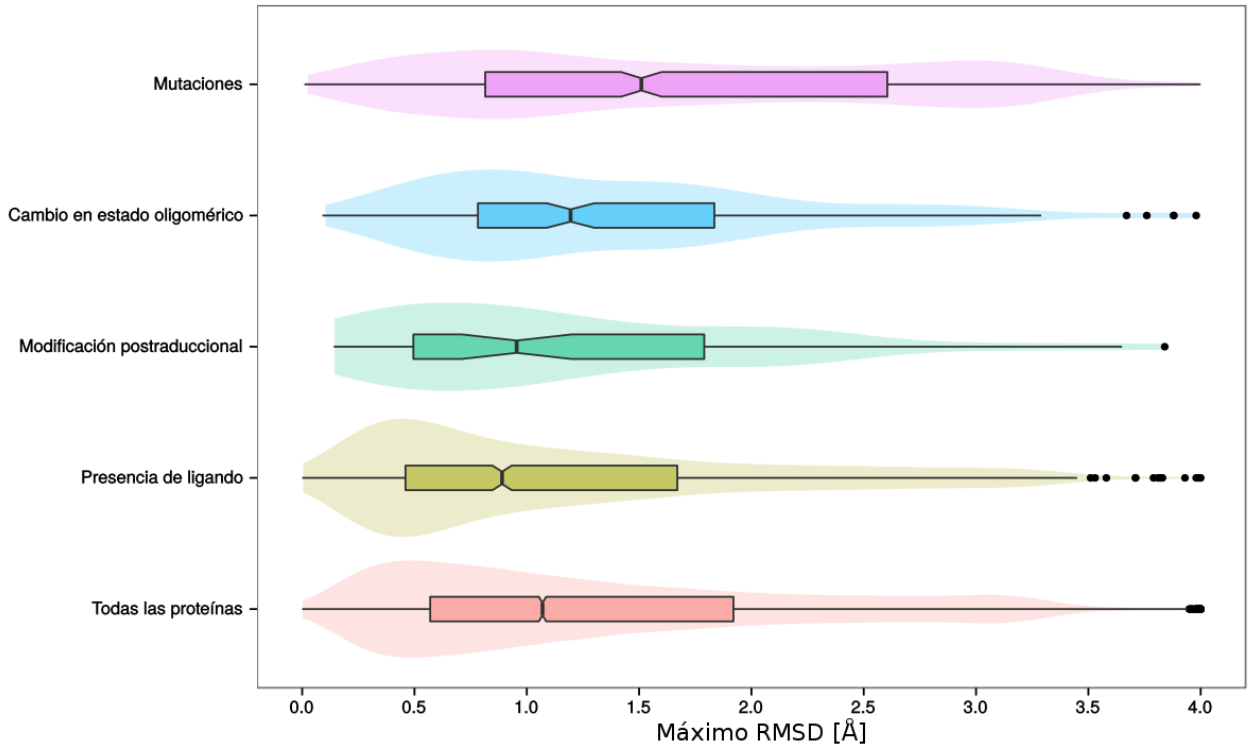

Figura 3.3: Distribución de diversidad conformacional según las condiciones que cambian en el par de confórmeros de máximo RMSD. Cada par es homogéneo para el factor estudiado, es decir que no son pares de confórmeros que podrían estar siendo afectados por más de un factor simultáneamente. "Todas las proteínas" son las 17.276 proteínas en la base de datos CoDNaS. "Presencia de ligando" es el subconjunto de 1.981 proteínas de CoDNaS considerando sólo pares de las formas apo y holo, sin otras condiciones que pudieran alterar la conformación proteica. De manera similar se forman los siguientes grupos. "Modificación postraduccional" incluye sólo 70 proteínas, 246 proteínas forman parte del subconjunto que muestra cambios en el estado oligomérico, y finalmente 987 proteínas muestran un cambio en la secuencia primaria debido a mutaciones.

simples rotaciones en las cadenas laterales podrían modificar el tamaño de cavidades internas o favorecer la apertura o cierre de túneles y bolsillos. Estos pequeños cambios, aunque biológicamente importantes, son difíciles de cuantificar y se estudiaran en el capítulo 4. La ausencia y presencia de ligando es una de las condiciones más importantes a fin de analizar cambios conformacionales, ya que es fácil su interpretación en la función biológica y son los pares de confórmeros mayoritarios los resueltos en formas *apo* y *holo*.

3.3. Descripción de los datos

En esta sección mostraremos algunos datos estadísticos descriptivos de cada una de las proteínas presentes en CoDNaS, como su taxonomía, cantidad de confórmeros y resolución de las estructuras.

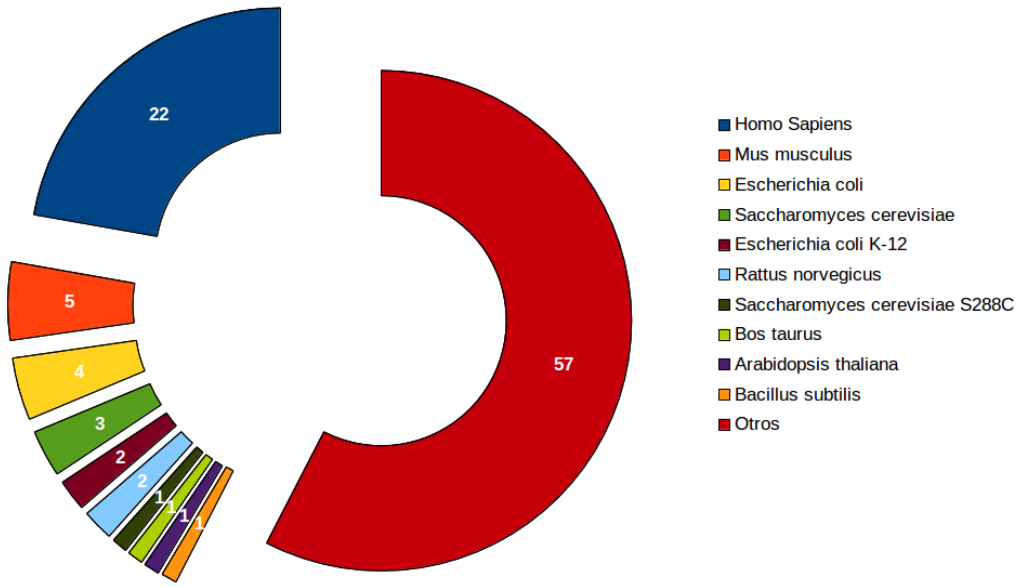

Figura 3.4: Los diez organismos más representados en la base de datos CoDNaS.

3.3.1. Representación taxonómica

Utilizando la clasificación taxonómica provista por NCBI estudiamos la distribución taxonómica de las proteínas contenidas en CoDNaS. Encontramos alrededor de 3000 diferentes organismos, provenientes de los distintos reinos vivos, representados en la base de datos. En la Figura 3.4, se observan los diez organismos más representados en las proteínas de CoDNaS. Podemos ver que las proteínas humanas son las más predominantes, representando un 22 % del total, seguidas en menor medida por *Mus musculus*, *Escherichia coli* y *Saccharomyces cerevisiae*. Además, hay una gran proporción de proteínas, que representan el 57 % del total, y están incluidos todos los organismos que se encuentran en proporciones menores al 1 %.

Dentro de los reinos vivos encontramos que el 52 % de las proteínas pertenecen a Eucariota, el 30 % a Bacteria, el 4 % a Archaea y el 6 % a virus y viroides.

3.3.2. Cantidad de confórmeros

Observando la distribución de la cantidad de confórmeros presentes en cada proteína de CoDNaS (Figura 3.5), vemos una distribución que presenta una mediana de 6 y un promedio de de 15.09 confórmeros. El 12.6 % de las proteínas tiene sólo dos confórmeros, el 10.75 % tiene 3 y el 14 % poseen 4. Un aspecto interesante es que el 62.7 % de las proteínas tiene al

menos 5 o más confórmeros. Este dato es de interés ya que para tener una correcta estimación de la diversidad conformacional es necesario tener al menos 5 estructuras diferentes de la misma proteína [Best et al., 2006], por ello en los análisis que realicemos en esta tesis, se considerarán proteínas con al menos 5 confórmeros. Otro aspecto a destacar, es que la cantidad de confórmeros va a ser mayor en aquellas proteínas obtenidas por RMN, ya que al poseer varios modelos estructurales por cada estructura hace que se incremente considerablemente la cantidad de confórmeros en estas proteínas. La media de confórmeros para las proteínas RMN es de 40, mientras que la media de las DRX es de 6.

La disponibilidad de varias estructuras para una misma proteína puede estar relacionada a fenómenos sin correlato biológicos como la facilidad de su cristalización, la longitud de la proteína en el caso de RMN, el interés farmacéutico o industrial, su importancia en investigación médica, entre otros. Esto también se ve reflejado en la dependencia que existe entre el número de confórmeros y el máximo RMSD obtenido por proteína, cuyo coeficiente de correlación de Spearman *rho* es de 0.46 con un P-valor $<< 0.01$, para todas las proteínas de la base de datos. Sin embargo, al tomar sólo proteínas con al menos 5 o más confórmeros y que sean obtenidas por DRX, esta correlación disminuye a un *rho* de 0.10 con un P-valor $<< 0.10$. Esto ocurre debido a que las chances de encontrar dos confórmeros con un máximo RMSD es similar en todas las proteínas al considerar al menos 5 confórmeros. Además, al quitar las proteínas RMN que introducen un sesgo no sólo en la cantidad de confórmeros sino también en el RMSD obtenido, el efecto del número de confórmeros en el máximo RMSD obtenido se reduce considerablemente.

3.3.3. Resolución de los confórmeros

A fin de evaluar la calidad de las estructuras presentes en CoDNaS, se realizó la distribución de resoluciones de todos los confórmeros obtenidos por DRX que integran la base de datos. La resolución de una estructura cristalográfica está relacionada con la capacidad de discernir los distintos detalles a nivel atómico de la misma. En la Figura 3.6 se muestra la distribución de las resoluciones de los confórmeros obtenidos por DRX en CoDNaS. Esta distribución presenta una mediana de 2.2 $\AA$, donde el 66.6 % de las estructuras tiene una resolución igual o menor

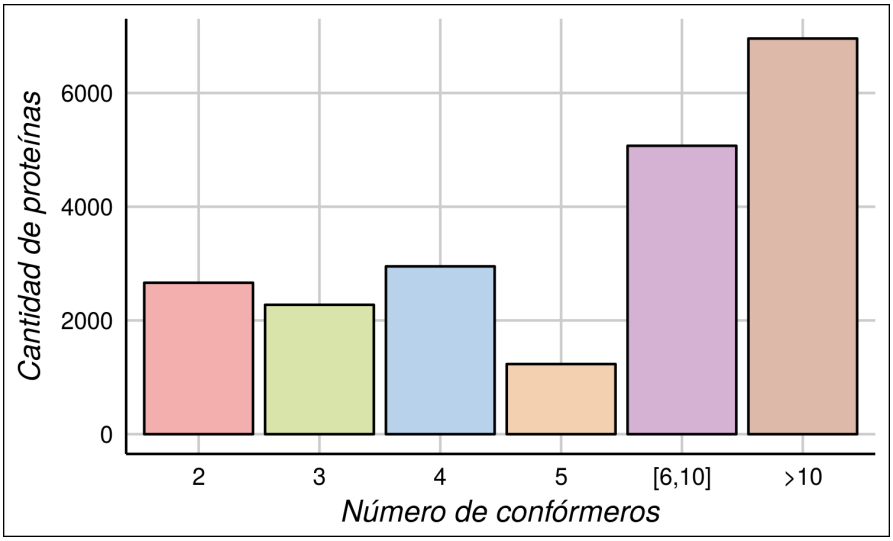

Figura 3.5: Distribución del número de confórmeros por cada proteína de CoDNaS.

a 2.5 $\mathring{A}$. Para tener una idea del nivel de detalle que podemos obtener a una determinada resolución, proporcionamos a continuación información extraída de `http://proteopedia.org/wiki/index.php/Resolution`:

- 1.2 $\mathring{A}$ Excelente: el *backbone* y las cadenas laterales se encuentran muy bien resueltas.
- 2.5 $\mathring{A}$ Buena: el *backbone* y muchas de sus cadenas laterales están bien definidas.
- 3.5 $\mathring{A}$ OK: el *backbone* y las cadenas laterales voluminosas están mayormente claras.
- 5.0 $\mathring{A}$ Pobre: el *backbone* está mayormente claro, sin embargo las cadenas laterales no.

El reclutamiento de estructuras en CoDNaS se hizo filtrando por aquellas que tenían una resolución igual o inferior a 4.5 $\mathring{A}$. Por lo que en base a los datos extraídos de la distribución de la Figura 3.6, podemos decir que la totalidad de las estructuras de CoDNaS poseen el *backbone* bien definido (que es utilizado para estimar el RMSD) y alrededor del 66 % de las mismas poseen una resolución igual o superior a 2.5 $\mathring{A}$ (valor de corte utilizado en la mayoría de los trabajos donde se desea contar con estructuras de alta calidad).

3.3.4. Longitud de las proteínas

CoDNaS contiene proteínas con una amplia variedad de longitudes. Observando la distribución de longitudes de las proteínas que componen la base de datos obtenemos una mediana

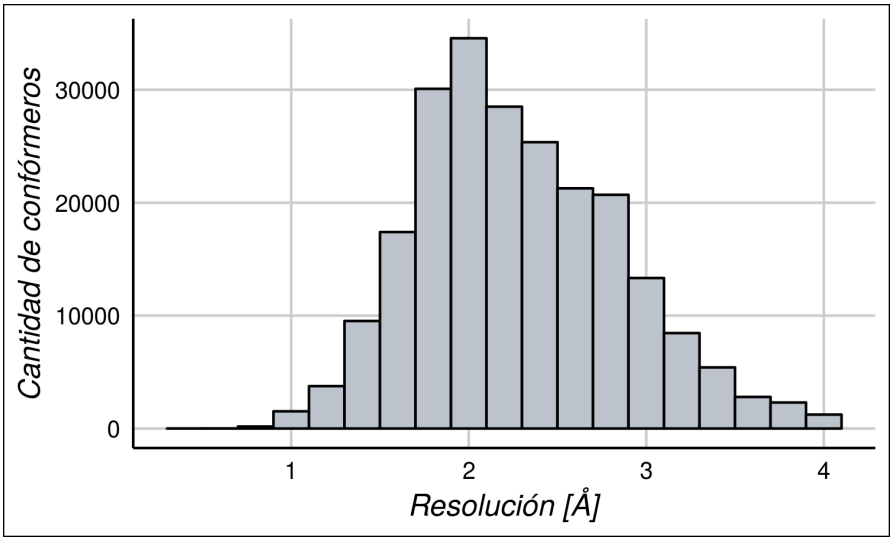

Figura 3.6: Distribución de la resolución de las estructuras presentes en CoDNaS.

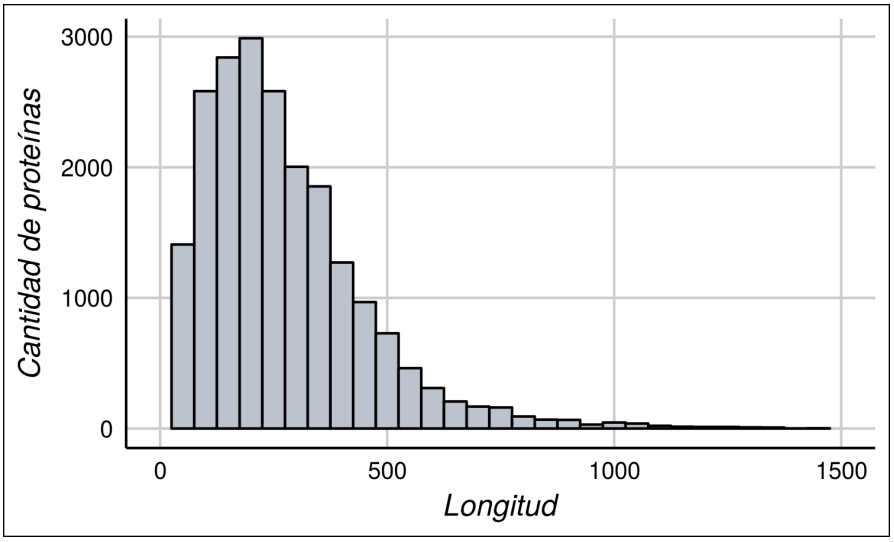

Figura 3.7: Distribución de las longitudes de las proteínas presentes en CoDNaS.

de 236 aminoácidos y un promedio de 276 (Figura 3.7). Con un 75 % percentil de 358 aminoácidos. Uno podría pensar que la longitud puede ser un factor que influya en el RMSD obtenido al comparar dos confórmeros. Al estudiar la correlación entre la longitud de la proteína y el máximo RMSD, obtenemos un coeficiente de correlación de Spearman *rho* de 0.2 con un P-valor $<< 0.10$. A pesar de no ser insignificante, es una correlación muy baja como para tomarla en cuenta a la hora de analizar y filtrar nuestros datos, y se encuentra en correspondencia con la correlación encontrada por *Burra et al.* [Burra et al., 2009].

3.3.5. Contactos cristalográficos en las estructuras

Los contactos cristalográficos o *"crystal packing contacts"* suelen ser considerados una fuente de error en la estructura cristalina de la proteína, sobre todo en aquellos residuos que se encuentran afectados por este tipo de contactos y forman loops [Rapp and Pollack, 2005, Jacobson et al., 2002]. Recordemos que los contactos cristalográficos son aquellos que se producen entre dos cadenas proteicas que se encuentran en celdas unitarias vecinas dentro de la red cristalina. Con el fin de estudiar si estos contactos podrían afectar al RMSD obtenido entre dos conformaciones, utilizamos un subset de proteínas monoméricas y con el mismo grupo simétrico de CoDNaS, es decir una sola cadena por celda unitaria y con la misma simetría, para evitar tener contactos proteína-proteína dentro de la celda unidad. Obtuvimos 392 pares de confórmeros y a cada uno de ellos le estimamos el número de contactos cristalinos utilizando el programa UCSF Chimera [Pettersen et al., 2004]. Este programa genera las repeticiones correspondientes de la celda unidad respetando la simetría del cristal, de manera tal de generar la red cristalina. Luego, utilizamos 4.5 $\mathring{A}$ como distancia para definir contactos entre un residuo de la celda unitaria principal y otro de alguna de sus vecinas. Para cada par de confórmeros calculamos el número de contactos cristalinos promedio de las dos estructuras y lo correlacionamos con el RMSD de ese par, con el fin de ver si había algún efecto del número promedio de contactos y el RMSD obtenido. Encontramos un coeficiente de correlación de Spearman *rho* despreciable de 0.048, por lo que podemos concluir que no hay correlación entre estas dos variables (Figura 3.8). Este resultado está en concordancia con lo encontrado por otros autores, donde se observa que el patrón de flexibilidad observado al comparar estructuras redundantes de la misma proteínas es robusto a efectos que puedan ser ocasionados por los contactos cristalinos o el *crystal packing* [Best et al., 2006, Zoete et al., 2002, Hrabe et al., 2015, Burra et al., 2009]. Además, Sikic y colaboradores encontraron que la flexibilidad de un loop es independiente a los contactos cristalográficos [Sikic et al., 2010].

3.4. Análisis de la redundancia secuencial en PDB

En esta sección trataremos brevemente el tema de la redundancia de secuencias presentes en la PDB que fue abordada en el trabajo *"On the dynamical incompleteness of the Protein*

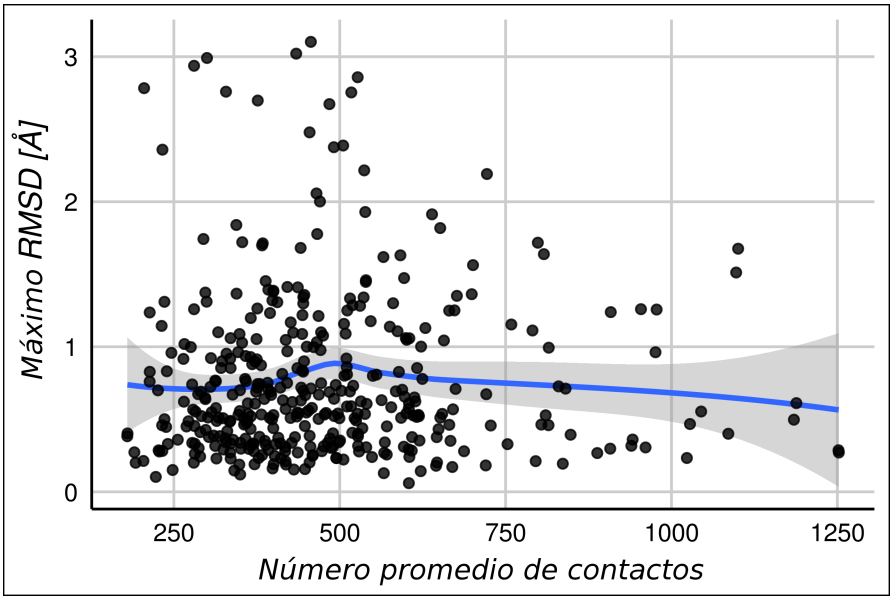

Figura 3.8: Relación entre el RMSD y el número de contactos cristalográficos promedio en el par de confórmeros.

Data Bank" [Marino-Buslje et al., 2017]. Como sabemos PDB es una base de datos redundante, es decir que podemos tener la estructura de una misma secuencia en diferentes estructuras. Si bien la redundancia presente en PDB no está directamente relacionada con los datos que constituyen CoDNaS (ya que hay filtrados previos), si es la fuente de información primaria para que la base de datos crezca en el número de proteínas. Además, consideramos que el tener estructuras redundantes de la misma proteínas en PDB es de gran importancia para poder estudiar la diversidad conformacional y dinámica proteica. Con el fin estudiar que tan redundante es PDB y cómo esta redundancia ha ido evolucionando al pasar los años, efectuamos un análisis en todas las secuencias presentes PDB. La redundancia puede estudiarse como el número de *clusters* secuenciales al 100 % de identidad que hay en la base de datos, y que se obtienen de forma similar a como explicamos en el capítulo anterior en la sección 2.3.1. Cada *cluster*, para ser considerado redundante, va a contener al menos dos cadenas proteicas provenientes de dos instancias de cristalización independientes (dos archivos "pdb" diferentes). En la Figura 3.9A se observa la fracción de cadenas nuevas depositadas en un año en particular en PDB que forman nuevos *clusters* redundantes. Esta es una medida de como una nueva proteína en PDB incrementa la redundancia, disminuyendo el número de *clusters* singulares. Esta fracción ha sido más o menos constante (en promedio 29 %) en los últimos 10 años pero sin embargo muestran un decrecimiento de alrededor del 5 % en los último 5

años. Por otra parte, la fracción de cadenas totales que van a *clusters* viejos o que ya son redundantes (una medida de como una nueva entrada de PDB incrementa la redundancia que ya existe en la base de datos) ha estado aumentando constantemente. Este resultado no es algo desconocido, ya que es sabido que PDB es muy redundante en ciertas proteínas que son de gran interés biológico y farmacológico.

En nuestro análisis encontramos 71,057 *clusters*, que es el número de secuencias no redundantes que hay en PDB. El número de cadenas de proteínas presentes en cada *cluster* se observa en la Figura 3.9B. Entre estos *clusters*, el 66 % contiene solo una cadena (son no redundantes), mientras que el resto (34 %) tiene al menos dos cadenas de proteínas provenientes de archivos "pdb" diferentes. Como hemos mencionado anteriormente, se necesitan al menos dos estructuras para poder estudiar los movimientos de una proteína a partir de estructuras redundantes, sin embargo y como analizamos anteriormente en este capítulo, es necesario un mínimo número de estructuras para poder estimar la diversidad conformacional en forma confiable. Teniendo en cuenta esto, encontramos que el 15 % de las proteínas redundantes en PDB tiene más de 5 estructuras, y solo el 3.2 % tienen más de 20 estructuras. En la Figura 3.9C se observa que el porcentaje de cadenas de PDB acumuladas en los diferentes *clusters* crece muy rápidamente. Por lo tanto, el 50 % de las cadenas de proteínas depositadas en PDB se encuentran en 5600 *clusters* que representan solo el 7 % de las secuencias en PDB, revelando que la redundancia está dominada por sólo algunas proteínas. Las cinco proteínas más redundantes en PDB son la lisozima de *Gallus gallus*, la β-2 microglobulina y la anhidrasa carbónica de humanos, la endothiapepsina de *Cryphonectria parasitica* y la tripsina catiónica de *Bos taurus*. Además, encontramos que esta redundancia está desigualmente distribuida en el espacio secuencial. En la Figura 3.9D mostramos que solo el 31 % de las familias de PFAM tiene al menos una proteína con redundancia estructural, el 17 % tiene solo una estructura asociada a esa familia y alrededor del 50 % de las familias de PFAM no poseen estructuras.

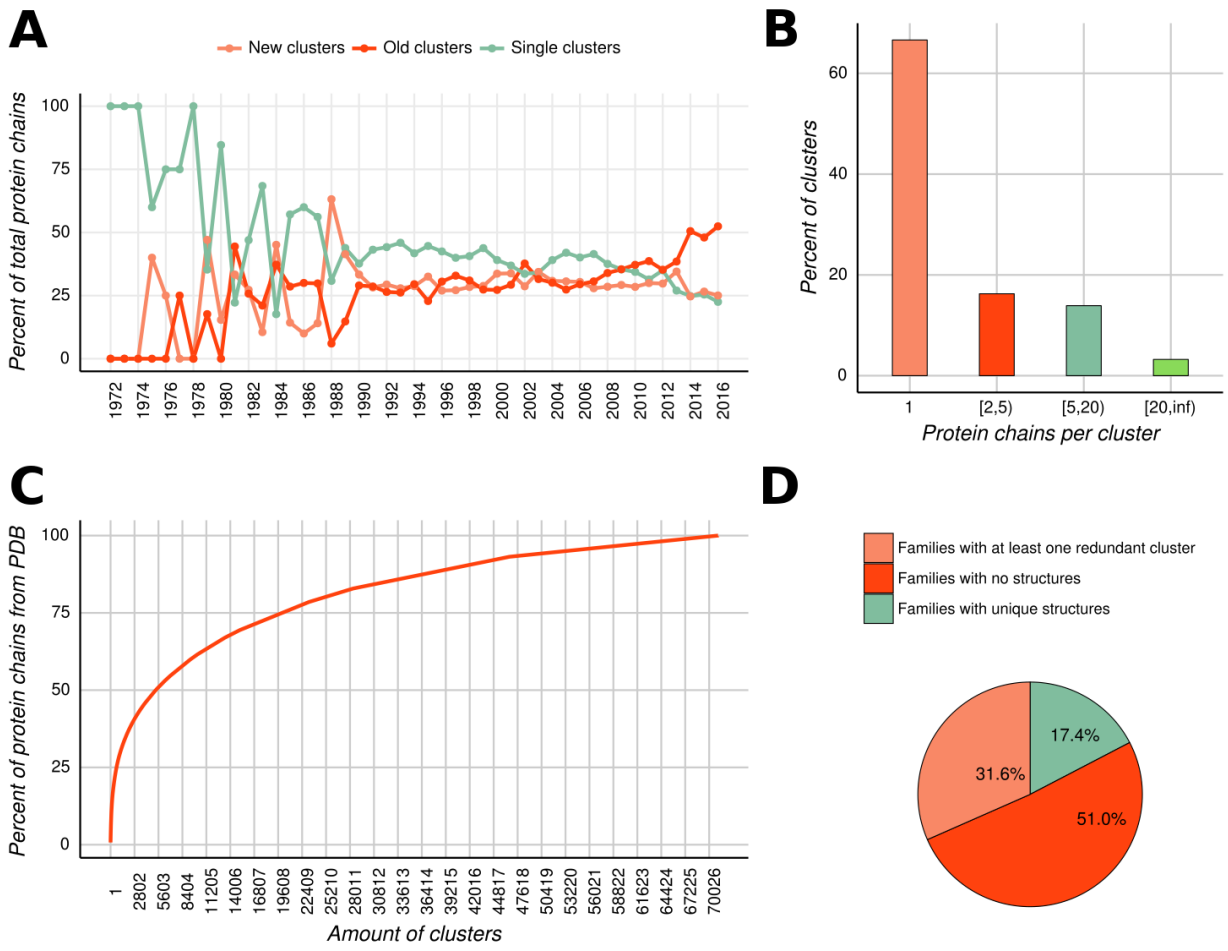

Figura 3.9: Análisis de la redundancia secuencial en PDB. La redundancia se analizó para todas las cadenas de proteínas en PDB depositadas entre el 8 de noviembre de 1972 (primera estructura depositada) y el 31 de diciembre de 2016. Un clúster redundante se compone de al menos dos cadenas de proteínas idénticas de diferentes entradas en PDB. A) Porcentaje de cadenas de proteínas depositadas en PDB en cada año. La línea "new clusters" representa el porcentaje de las estructuras depositadas en un año particular que son idénticas a una cadena única anterior previamente depositada (se forma un nuevo clúster con esta adición)y los "old clusters" corresponden al porcentaje de estructuras depositadas en un año particular que son idénticas a otras que ya están en un clúster redundante (con al menos dos estructuras) y la línea "single clusters" son aquellos formados por una sola estructura (sin redundancia) en PDB. B) Porcentaje de *clusters* con una sola estructura, 2-5, 5-20 o >20 estructuras. C) Porcentaje acumulado de la cantidad de cadenas de proteínas del total PDB que forman cada *cluster*. D) Porcentaje de familias PFAM (versión 31.0) sin estructuras asociadas (estructuralmente sin resolver), familias con estructuras únicas (sin redundancia) y familias con al menos una *cluster* redundante.

3.5. Conclusiones

En el estudio de la diversidad conformacional proteica, es posible usar los confórmeros determinados experimentalmente por RMN o DRX que se encuentran en CoDNaS. Dado que es difícil analizar datos que provienen de la comparación estructural de confórmeros obtenidos

mediante técnicas distintas y sumado a que la mayoría de las estructuras de CoDNaS están resultas por DRX, en los estudios posteriores se usarán sólo confórmeros cristalográficos.

Por lo expuesto anteriormente en este capítulo, es necesario tener algunas consideraciones a la hora de seleccionar las proteínas que formarán parte de los set de datos utilizados en los análisis posteriores. Las diferencias estructurales estimadas por el RMSD pueden tener sesgos dados por: el método experimental utilizado para la determinación de la estructura del confórmero, la presencia de mutaciones en la secuencia, la resolución cristalográfica de la estructura y el número de confórmeros de esa proteína. En particular se observa que las proteínas obtenidas por RMN presentan valores más altos de RMSD y más cantidad de confórmeros que las proteínas provenientes de DRX.

Respecto a la redundancia en PDB, vemos que ha surgido por la acumulación de estructuras de proteínas que son muy estudiadas o de gran interés biotecnológico lo que hace que esté altamente sesgada a este tipo de moléculas, además de aquellas que son fáciles de obtener experimentalmente. La redundancia estructural es de gran importancia para poder estudiar la diversidad conformacional proteica y consecuentemente mejorar y optimizar métodos computacionales que necesitan de la estructura de una proteína. Creemos que es necesario proyectos a futuro que promuevan la obtención de estructuras de proteínas ya conocidas en diferentes condiciones experimentales.

Capítulo 4

El análisis de la diversidad conformacional y su relación con los mecanismos funcionales en proteínas[1]

4.1. Resumen

Los movimientos de las proteínas son una característica clave para comprender la función biológica. Un análisis a gran escala de la distribución de diversidad conformacional de proteínas mostró una distribución sesgada positivamente con un pico en RMSD = 0.5 Å. Para entender esta distribución en términos de relaciones estructura-función, estudiamos un conjunto de datos bien curado ∼5,000 proteínas con diversidad conformacional determinada experimentalmente. En este capítulo, utilizamos distintas características derivadas del ensamble conformacional el cual nos permitió describir la distribución de RMSD en términos de al menos cuatro subconjuntos de distintas relaciones estructura-función. El mayor de estos subconjuntos de proteínas (∼60 %), que llamamos "rígidas" (RMSD promedio = 0.83 Å), no tiene regiones desordenadas, muestra baja diversidad conformacional, túneles más grandes y cavidades más pequeñas y enterradas en la estructura de la proteína. Los dos subconjuntos adicionales contienen regiones desordenadas, pero con una composición de secuencia y un comportamiento diferencial. Las proteínas "parcialmente desordenadas" tienen en promedio

[1]Este capítulo está basado en la publicación: [Monzon et al., 2017a].

el 67 % de sus confórmeros con regiones desordenadas, RMSD promedio = 1.1 Å, el mayor número de regiones hinges o bisagras y las regiones desordenadas más largas. Por el contrario, las proteínas "maleables" tienen en promedio solo el 25 % de confórmeros desordenados y un RMSD promedio = 1.3 Å, cavidades flexibles afectadas en tamaño por la presencia de las regiones desordenadas y muestran la mayor diversidad de ligandos afines. Finalmente, un conjunto minoritario (menos del 15 % de las proteínas del set de datos) lo constituyen las proteínas altamente móviles, que se caracterizan por no poseer regiones desordenadas, alto número de dominios y bisagras, y muestran una gran diversidad conformacional.

Las proteínas en cada conjunto son en su mayoría no homólogas entre sí, no comparten una clase de plegamiento particular, ni tienen similitudes funcionales, pero sí comparten características derivadas de su población de confórmeros. Estas características compartidas podrían representar mecanismos conformacionales relacionados con sus funciones biológicas.

4.2. Introducción

Los primeros estudios de cristalografía sobre la mioglobina no encontraron una forma evidente por la cual el oxígeno pudiera entrar en la molécula y unirse al grupo hemo [Perutz and Mathews, 1966]. Tomó más de una década descubrir que los movimientos de la mioglobina eran esenciales para que el oxígeno entrara y saliera de la misma, demostrando que la dinámica era fundamental para explicar la biología de la mioglobina [Frauenfelder and McMahon, 1998,Case and Karplus, 1979]. Después de estos primeros hallazgos, estos conceptos se generalizaron a casi todas las proteínas ya que se ha acumulado una gran cantidad de información relacionada con el movimiento de las proteínas y su función biológica. Se ha explorado una amplia gama de movimientos en proteínas, desde grandes movimientos relativos entre dominios [Gerstein et al., 1994], reacomodamiento de elementos de estructura secundaria y terciaria [Gerstein and Krebs, 1998], desplazamientos de loops [Gu et al., 2015] hasta pequeños reordenamientos que involucran sólo a los residuos [Gora et al., 2013]. Como mencionamos en el capítulo 1, el caso más extremo de movilidad en proteínas ocurre en las IDRs o IDPs caracterizadas por su alta flexibilidad en su estado nativo, movimientos que se han demostrado íntimamente relacionados con su función biológica [van der Lee et al., 2014, Janin and Sternberg, 2013].

Recientemente se ha publicado un análisis a gran escala, estudiando la extensión de la diversidad conformacional en proteínas, utilizando estructuras redundantes de la misma proteína obtenidas en diferentes condiciones experimentales [Burra et al., 2009]. Como se explica en la sección 2.2.1, estas estructuras son consideradas como posibles confórmeros del ensamble nativo y su diferencias estructurales pueden ser estimadas utilizando el RMSD. El RMSD así como otras medidas de disimilitud estructural, miden las diferencias entre las partes ordenadas de las proteínas (que poseen estructura) y analizando su distribución entre pares de confórmeros de distintas proteínas, es posible inferir que la mayoría de las proteínas tienen movimientos alrededor de los 0.5 $\mathring{A}$; compatible con el error aceptado cuando se comparan dos estructuras de la misma proteína obtenidas por DRX en las mismas condiciones [Burra et al., 2009]. Otro estudio de cambios conformacionales en 60 enzimas entre sus conformaciones *apo* y *holo*, muestra que el 75 % de los pares tiene un RMSD menor que 1 $\mathring{A}$, y el 91 % menor a 2 $\mathring{A}$, con un RMSD promedio de 0.7 $\mathring{A}$ [Gutteridge and Thornton, 2005]. Es decir que las comparaciones entre dos confórmeros en las mismas condiciones experimentales muestran un valor de RMSD solo un poco por debajo de la comparación entre confórmeros *apo* y *holo*. Estos resultados expresan que muchas proteínas no necesitan grandes movimientos para cumplir su función biológica, y coinciden con otras observaciones donde se describen que aún pequeños cambios entre confórmeros podrían afectar en gran medida el comportamiento de los parámetros catalíticos y biológicos en enzimas [Mesecar et al., 1997,Koshland, 1998]. Además, algunos trabajos sugieren que algunas propiedades biológicas claves para las proteínas [Gunasekaran et al., 2004], tales como el alosterismo o el cooperativismo, podrían ocurrir sin la necesidad de grandes cambios conformacionales en la estructura promedio de la proteína [Tsai et al., 2008]. Sin embargo, la distribución de diversidad conformacional presentada por *Burra et al.* muestra además una tendencia muy marcada hacia valores de RMSD altos que indican que una fracción menor del conjunto de proteínas analizadas requieren cambios conformacionales grandes para cumplir su función (ver Figura 4.1) [Burra et al., 2009]. Estos valores de RMSD altos, son comúnmente observados en proteínas multidominio donde los movimientos del tipo bisagra producen grandes desplazamientos como los conocidos de cuerpo rígido [Lesk and Chothia, 1984].

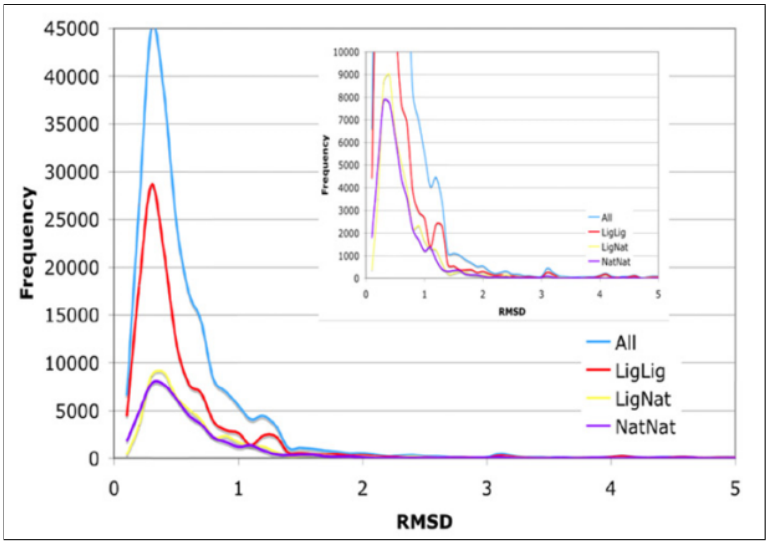

Figura 4.1: Distribución global de RMSD por pares de confórmeros de todas las proteínas del set de datos de trabajo de *Burra et al.*. Imágen extraída de [Burra et al., 2009].

Como señaló Fraunfelder, la búsqueda de conceptos generales entre la dinámica y la función de la proteína es una cuestión clave en la biología estructural [Frauenfelder and McMahon, 1998]. Conocer el grado de los cambios conformacionales podría contribuir enormemente a nuestra comprensión sobre cómo estos movimientos modulan la función proteica. El análisis a gran escala de diferentes grados y tipos de movimientos de proteínas también podría permitirnos inferir reglas generales sobre diferentes relaciones estructura-función en el espacio estructural de proteínas. En este capítulo utilizamos un set grande y curado de proteínas extraídas de la base de datos CoDNaS [Monzon et al., 2013,Monzon et al., 2016] para reproducir la distribución presentada por *Burra et al.*. Esperamos con este estudio derivar conceptos generales para comprender la relación estructura-función en proteínas. Para tal fin, utilizamos diferentes análisis estructurales y dinámicos sobre la población de confórmeros de cada proteína y encontramos que de la distribución de diversidad conformacional emergen al menos cuatro grupos de proteínas con una relación estructura-función bien característica. Las proteínas en cada uno de estos grupos son mayormente no-homólogas y no muestran preferencia alguna por un determinado tipo de plegamiento o función. Sin embargo, estas proteínas comparten características similares que emergen del análisis de su comportamiento dinámico a partir de su población de confórmeros y podrían representar mecanismos conformacionales relacionados con la función proteica.

4.3. Resultados

4.3.1. Distribución general y factores moduladores

Como mencionamos en el capítulo anterior, CoDNaS es una base de datos que contiene una colección redundante de estructuras tridimensionales para la misma secuencia donde cada estructura se puede tomar como una instantánea del conjunto conformacional de la proteína. En la Figura 4.2, se muestra la distribución general de RMSD entre confórmeros para todas las cadenas de proteínas contenidas en CoDNaS obtenidas por DRX (∼16,000 cadenas de proteínas diferentes al momento de este análisis). Sin embargo, la presencia de mutaciones, estructuras de baja resolución y el número de estructuras por proteína (número de confórmeros) podría afectar la estimación del grado de diversidad conformacional [Parisi et al., 2015]. Además, dado que menos del 6 % de las entradas CoDNaS tienen estructuras obtenidas por RMN y debido a su comportamiento de flexibilidad diferencial en comparación con las estructuras cristalográficas [Sikic et al., 2010], se eliminaron los confórmeros RMN para evitar sesgos introducidos mediante la mezcla de estructuras obtenidas por diferentes métodos. En este sentido, utilizamos un conjunto de datos reducido (4,791 cadenas de proteínas diferentes con 74,417 confórmeros y 1,186,312 pares de confórmeros). Todas las distribuciones en la Figura 4.2A muestran la tendencia previamente identificada [Burra et al., 2009].

En nuestro análisis, también es posible estudiar la diversidad conformacional máxima mostrada por una cadena o proteína dada (el par de confórmeros que muestra el máximo RMSD entre todas las comparaciones de a pares para una proteína determinada). Esta distribución (Figura 4.2B) nuevamente sigue la tendencia general, con una mediana de 0.83 $\mathring{A}$, un promedio de 0.99 $\mathring{A}$ y un pico grande cerca de 0.4 $\mathring{A}$. Esta distribución es la que continuaremos analizando en el desarrollo del capítulo cuando nos remitimos a la diversidad conformacional de las proteínas. A partir de esta figura, es posible inferir que la mayoría de las proteínas requieren pequeños movimientos entre confórmeros para cumplir sus funciones biológicas.

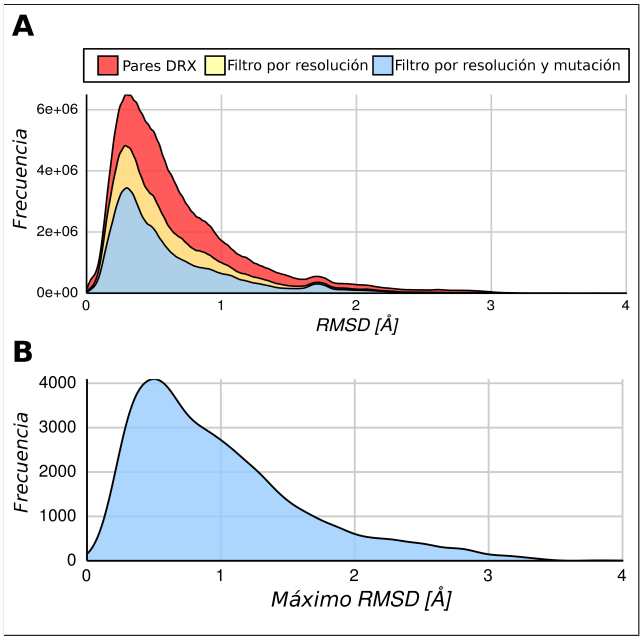

Figura 4.2: A) Todos los valores de RMSD por pares de confórmeros de proteínas con confórmeros obtenidos por DRX. La distribución de color amarillo claro considera sólo pares de confórmeros que se obtuvieron a una resolución igual o inferior a 2.5 Å y la distribución de rojo claro considera además solo pares de confórmeros 100 % idénticos en secuencia (sin mutaciones). B) Distribución de la diversidad conformacional máxima por proteína. Esta distribución se utilizará en el resto del capítulo como medida de la diversidad conformacional del espacio de proteínas estudiado.

4.3.2. Análisis de la distribución

Recientemente, hemos encontrado que las proteínas que experimentan transiciones orden-desorden entre sus confórmeros muestran valores de RMSD mayores que aquellas que no poseen transiciones [Zea et al., 2016]. En este estudio se seleccionaron los pares de confórmeros, en las formas *apo* y *holo*, que muestran la transición más grande en una proteína, y se estimó el RMSD a partir de las zonas estructuradas en ambos confórmeros (ver Figura 4.3). Esta observación nos condujo a explorar el rol de la regiones desordenadas (IDRs) en la diversidad conformacional de las zonas estructuradas en las proteínas. A partir de nuestra definición de IDRs (ver Métodos 4.6), identificamos y mapeamos estas regiones en todos los confórmeros de cada proteína del set de datos. Esto nos permitió separar la distribución de la Figura 4.2B en dos nuevas distribuciones que se muestran en la Figura 4.4. Una de las distribuciones de máximo RMSD (roja clara en la Figura 4.4A) posee un promedio de 0.87 $\mathring{A}$ y corresponde a proteínas ordenadas sin IDRs (en ninguno de sus confórmeros), mientras que la otra distribución (amarilla clara en la Figura 4.4A) posee un RMSD promedio de 1.20 $\mathring{A}$ y corresponde a proteínas que contienen IDRs (en al menos uno de todos sus confórmeros). Estas distri-

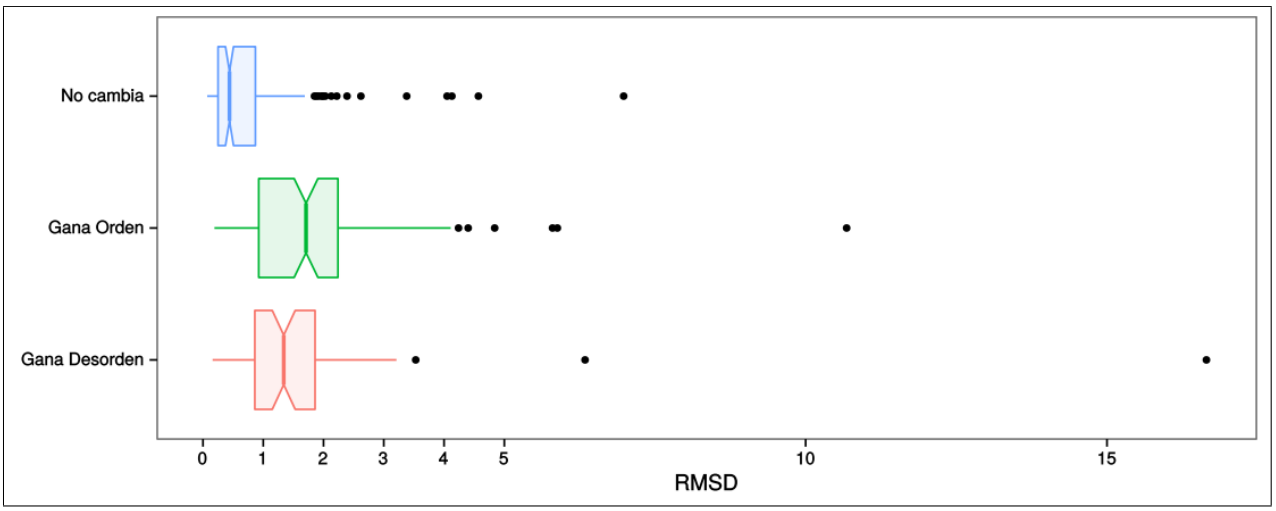

Figura 4.3: Distribuciones de RMSD para los pares de confórmeros en donde se estudiaron transiciones orden/desorden. Las distintas categorías indican la proporción de orden o desorden que predomina al final de la transición (ganan desorden, orden o no cambia). "No cambia" sin embargo puede representar la ausencia de transiciones o la existencia de transiciones compensatorias. Imágen extraída y adaptada de [Zea et al., 2016].

buciones difieren estadísticamente en su forma y en su mediana (*Kolmogorov-Smirnov test* y *Wilcoxon rank-sum test* respectivamente, P-valor $<< 0.001$).

Como las proteínas con IDRs tienen un RMSD mayor respecto de las que no poseen regiones, este valor podría usarse para predecir si una proteína es desordenada (en al menos un confórmero con IDRs) u ordenada (Figura 4.4A). El Área Bajo la curva ROC (*"Receiver Operating Characteristic"*) o AUC para el máximo RMSD como predictor de proteínas desordenadas es 0.67. Adicionalmente, la regresión logística para predecir proteínas desordenadas usando los valores de máximo RMSD muestra un coeficiente estadísticamente significativo (P-valor < 0.001). Utilizando un valor de corte de máximo RMSD igual a 0.9 $\mathring{A}$, la precisión para predecir la presencia de desorden es de 64 % y para RMSD > 1.2 la precisión es de 66 %. Por encima de un valor de máximo RMSD de 0.9 $\mathring{A}$ de diversidad conformacional, encontramos 3.03 veces más proteínas con regiones desordenadas que por debajo de este valor. Estos resultados indican que las proteínas con IDRs muestran una diversidad conformacional mayor (mayor máximo RMSD) en sus partes estructuradas que las proteínas completamente ordenadas.

Además, la presencia de regiones desordenadas en los pares que maximizan la diversidad conformacional puede utilizarse para separar la distribución de proteínas con IDRs. Encontramos que las proteínas en cuyos pares de máximo RMSD al menos uno de los confórmeros tienen

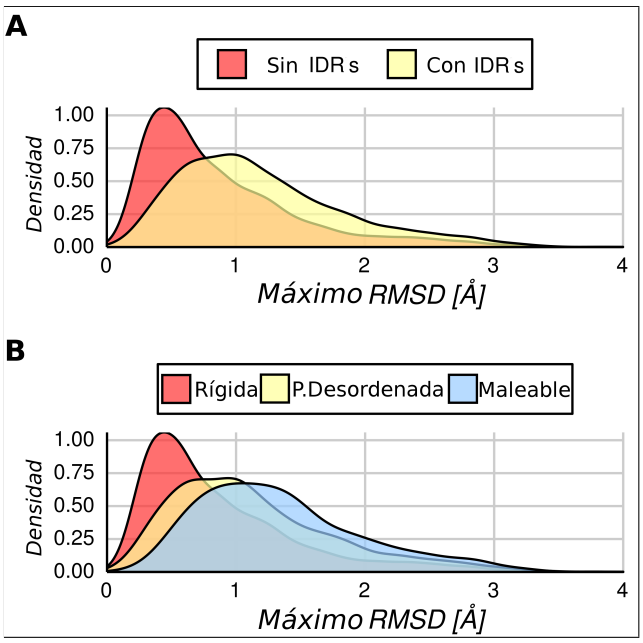

Figura 4.4: A) Pares de confórmeros que poseen el máximo RMSD para proteínas con/sin IDRs considerando todos los confórmeros disponibles por proteína. En rojo claro, mostramos la distribución de proteínas sin IDRs y en color amarillo claro con IDRs en al menos uno de sus confórmeros. La distribución de proteínas sin IDRs posee valores de RMSD globales significativamente más bajos en comparación con las proteínas con IDRs (*Kolmogorov-Smirnov test*, P-valor $<<$ 0.001). B) La distribución máxima de la diversidad conformacional puede representarse mediante tres conjuntos mayoritarios de proteínas: rígida (proteína sin IDRs en todos sus confórmeros), parcialmente desordenada (con IDRs en al menos un confórmero, incluso en aquellos que componen el par máximo de RMSD) y maleable (con IDRs en al menos un confórmero, sin embargo el par de confórmeros con máximo RMSD no posee IDRs). Estas tres distribuciones poseen medianas significativamente diferentes (*Kruskal—Wallis rank sum test*, P $<<$ 0.001 con *Nemenyi post-hoc test*).

IDRs, presentan un RMSD promedio de 1.14 $\mathring{A}$, mientras que aquellas que sus confórmeros no presentan regiones en el par de máximo RMSD tiene un promedio de 1.35 $\mathring{A}$ (difieren significativamente por *Kolmogorov-Smirnov test* y *Wilcoxon rank-sum test*, P-valor $<<$ 0.01). Veíamos que esta propiedad encontrada para el par de máximo RMSD, reflejaba una propiedad del ensamble de confórmeros: las proteínas que contenían regiones desordenadas en su par de máximo RMSD mostraban que la mayoría de sus confórmeros contenían esas regiones en forma desordenada. Por el contrario, las proteínas que no mostraban regiones desordenadas en su par máximo, no poseían regiones desordenadas en la mayoría de sus confórmeros (ver ejemplo esquemático en la Figura 4.5).

En orden creciente de acuerdo al grado de diversidad conformacional que presentan las proteínas en cada grupo, las nombraremos de la siguiente manera: RÍGIDAS con un máximo RMSD promedio de 0.85 $\mathring{A}$ (roja clara en la Figura 4.4B), PARCIALMENTE DESORDENADA con un máximo RMSD promedio de 1.1 $\mathring{A}$ (amarilla clara en la Figura 4.4B) y MALEABLE

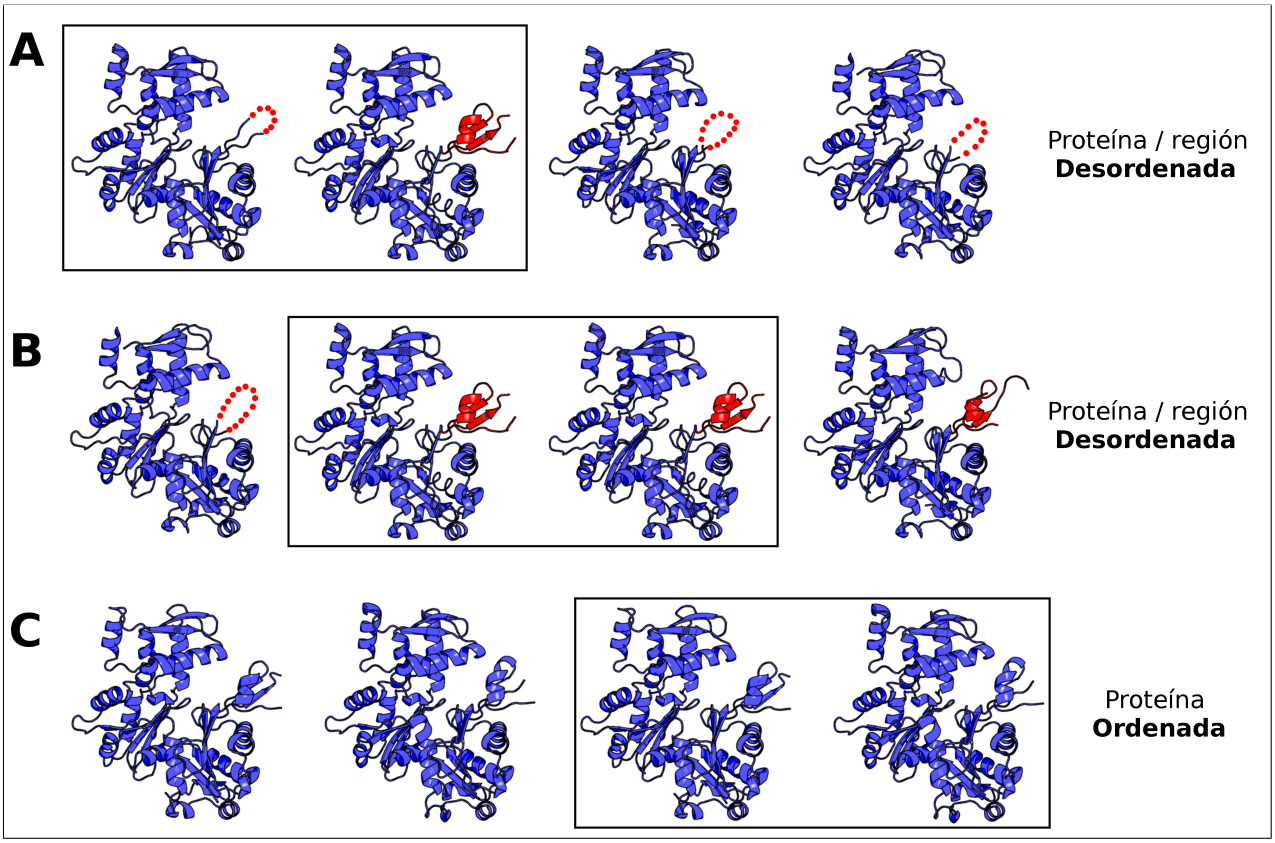

Figura 4.5: Figura esquemática de la clasificación de desorden utilizada en este capítulo. Las regiones coloreadas en rojo indican que esos residuos se encuentran desordenados (son *missing residuos*) en algún confórmero de la proteína, esto se puede observar en los confórmeros que en esa región poseen una línea de puntos. El cuadro abarcando dos confórmeros hace referencia al par de máximo RMSD. A) Proteína desordenada con IDRs en tres de sus confórmeros. El par de máximo RMSD está dado entre un confórmero con IDR y otro ordenado (Proteína parcialmente desordenada). B) Proteína desordenada con IDR en uno de sus confórmeros. El par de máximo RMSD esta dado entre dos confórmeros ordenados (Proteína Maleable). C) Proteína con todos sus confórmeros ordenados (sin IDRs) (Proteína ordenada).

con un máximo RMSD promedio de 1.3 $\mathring{A}$ (azul clara en la Figura 4.4B). Estas tres distribuciones poseen medianas que son estadísticamente diferentes (*Kruskal—Wallis rank sum test*, P << 0.001 con *Nemenyi post-hoc test*). En el transcurso de este capítulo, mostraremos que en base a características dinámicas y estructurales, estos tres grupos (sumaremos un cuarto grupo minoritario en la discusión que se desprende de las proteínas ordenadas) de proteínas podrían representar diferentes mecanismos conformacionales.

Es importante destacar que se analizó si había algún sesgo en la estimación del máximo RMSD, ya sea su dependencia respecto de la longitud de la proteína o la cantidad de confórmeros. En ambos casos se obtuvo un coeficiente de correlación de Spearman despreciable. Además, se estudió la correlación entre el porcentaje de confórmeros con regiones desordenadas respecto del número de confórmeros total por proteína y de nuevo se obtuvo un coeficiente de correlación de Spearman despreciable.

4.3.3. Caracterización estructural de las distribuciones de diversidad conformacional

La Tabla Sup. A.1 resume varias caracterizaciones estructurales y composición de los set de proteínas rígidas, parcialmente desordenadas y maleables. Contiene la media, mediana y la desviación estándar de las distribuciones de las variables comparadas entre los grupos.

Las proteínas rígidas poseen el RMSD promedio más bajo y ninguno de sus confórmeros tiene regiones desordenadas. Este es el grupo que posee más proteínas de los tres analizados (64.18 % del total). Encontramos que las proteínas parcialmente desordenadas y maleables difieren significativamente en el porcentaje de confórmeros con IDRs por proteína. Las parcialmente desordenadas tienen en promedio el 69.1 % de sus confórmeros con IDRs, mientras que en las maleables solo el 24.86 %. Evidentemente las proteínas parcialmente desordenadas tienen regiones "desordenadas" más flexibles que las proteínas maleables. Esta diferencia también se evidencia en la composición de aminoácidos de estas regiones. Comparamos la composición de aminoácidos de las regiones desordenadas de cada confórmero y de la base de datos DisProt [Piovesan et al., 2016] (una base de datos de proteínas con anotación experimental de sus regiones desordenadas), relativas a la composición de aminoácidos de PDB

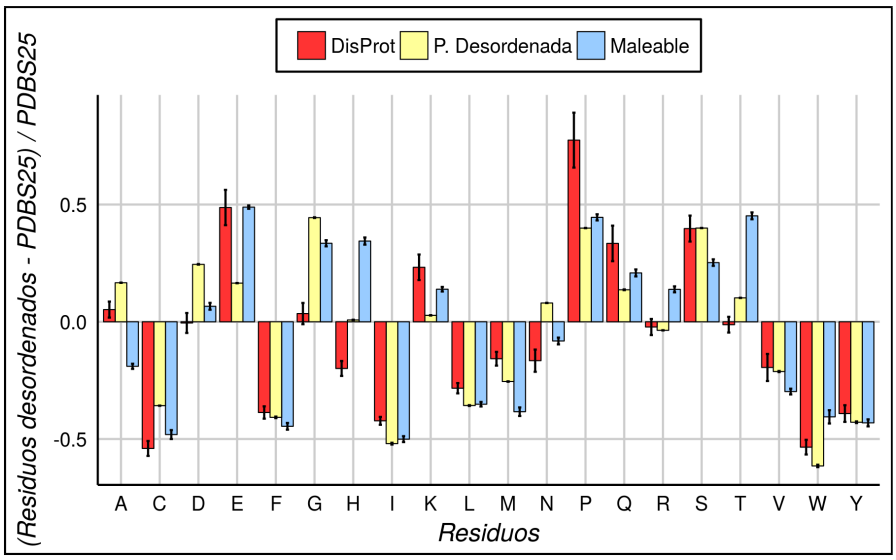

Figura 4.6: La composición de aminoácidos de las IDRs que se encuentran en todos los confórmeros de proteínas maleables (azul claro) y parcialmente desordenadas (amarillo) en relación con PDB Select 25. DisProt se usa como referencia de regiones y proteínas desordenadas experimentales.

Select 25 [Hobohm and Sander, 2008] (proteínas susceptibles a ser cristalizadas mayormente compuestas por residuos ordenados). PDB Select 25 se utilizó como distribución *background* de tal manera de observar que residuos de nuestras regiones están enriquecidos o no, respecto de los residuos que siempre tienden a estar ordenados, y DisProt se utilizó como control de desorden experimental. Encontramos que las IDRs en las parcialmente desordenadas y maleables están enriquecidas en aminoácidos que se caracterizan por participar en regiones de alta flexibilidad y contrariamente reducidas en aminoácidos que participan en regiones ordenadas o globulares. Esta observación puede ser derivada de la distribución mostrada en la Figura 4.6, donde las proporciones de los aminoácidos R, G, Q, S, N, P, D, E y K se encuentran enriquecidas y corresponden a aminoácidos flexibles de acuerdo al índice de flexibilidad de Vihinen [Vihinen et al., 1994]. Además, es posible observar que las proteínas parcialmente desordenadas y maleables a pesar de tener composiciones muy similares, difieren en algunos residuos. Las maleables muestran una mayor proporción de His, Thr y Lys en comparación con las parcialmente desordenadas, que por el contrario muestran una alta proporción de Prolina.

Las diferencias de composición en las IDRs entre los grupos, impactan directamente en las distribuciones observadas de RMSD. Como se puede derivar de la Tabla Sup. A.1, las proteínas parcialmente desordenadas muestran un mayor porcentaje de desorden e IDRs más largas en comparación con las maleables. Las proteínas maleables no muestran IDRs en su par

de máxima diversidad conformacional, esto ocurre porque sus IDRs tienden a estar mayormente ordenadas en el ensamble de confórmeros. Sin embargo, estas regiones que se ordenan en las maleables son altamente flexibles por lo que introducen valores de RMSD altos. Por el contrario, estas regiones en las parcialmente desordenadas se encuentran mayormente desordenadas en todos los confórmeros y no impactan en el cálculo del RMSD (ya que los átomos no se encuentran en la estructura). Una consecuencia directa de esta observación se refleja cuando estimamos el RMSD de las regiones de loops (Ver Métodos 4.6). Encontramos que nuevamente las proteínas maleables muestran valores de RMSD significativamente superiores a las parcialmente desordenadas (ver Figura sup. B.1.A), mientras que en otros elementos de estructura secundaria esto no ocurre (ver Figura sup. B.1.B-C). Además, cuando removemos éstas regiones que se ordenan y desordenan en los confórmeros que muestran la máxima diversidad conformacional y re-calculamos el RMSD; no encontramos diferencias significativas entre las distribuciones de RMSD entre las proteínas maleables y parcialmente desordenadas (*Wilcoxon rank-sum test*, P = 0.74 y *Kolmogorov-Smirnov test, P = 0.83*). En efecto, el RMSD promedio entre los tres grupos resulta muy similar cuando re-calculamos el RMSD eliminando las regiones identificadas como loops: 0.86 $\mathring{A}$, 0.993 $\mathring{A}$ y 0.995 $\mathring{A}$ para rígidas, parcialmente desordenadas y maleables, respectivamente. Esto soporta la visión de que la diversidad conformacional de estos tres grupos mayormente difiere en la conformación de las regiones que se ordenan y que están involucradas en las transiciones orden/desorden pero con diferente tasa de transición entre estos estados.

Encontramos otras diferencias estructurales entre estos tres grupos de proteínas. Por ejemplo, las proteínas parcialmente desordenadas contienen dos veces más regiones *hinges* o bisagras que las proteínas maleables y las rígidas (Figura sup. B.2) y también en promedio un mayor radio de giro normalizado (Figura sup. B.3). La presencia de *hinges* podría estar relacionada con un incremento en la diversidad conformacional debido al movimiento relativo entre dominios, esto además podría fundamentarse en que las proteínas parcialmente desordenadas presentan en promedio un mayor número de dominios (1.61 comparado con 1.50 en maleables y 1.49 en rígidas). Por otro lado, el incremento en el radio de giro en las proteínas parcialmente desordenadas podría evidenciar plegamientos más voluminosos y menos compactados que en los otros dos grupos.

Los confórmeros en nuestro set de datos pueden diferir por distintas causas asociadas a las condiciones experimentales de cristalización (por ejemplo, modificaciones post-traduccionales, diferente estado oligomérico, presencia de ligandos, entre otros). Sin embargo es difícil establecer una correlación exacta entre las diferencias estructurales de los confórmeros y función proteica. Por ello, también estudiamos el comportamiento de los tres grupos de proteínas en aquellas que el máximo RMSD estaba dado por un par de confórmeros *apo* y *holo* (con y sin ligando). Obtuvimos nuevamente la misma tendencia (ver Figura sup. B.4) que en las distribuciones mostradas en la Figura 4.4B, siendo las mismas significativamente diferentes (*Kruskal—Wallis rank sum test*, $P << 0.001$ con *Nemenyi post-hoc test*).

Finalmente, utilizando la base de datos CATH asignamos a cada proteína la/s superfamilia/s a la que pertenecen los dominios que las componen. En las proteínas rígidas más del 84 % corresponden a superfamilias diferentes. Este mismo comportamiento se observa en las proteínas parcialmente desordenadas y maleables con 74 % y 78 %, respectivamente. Además, se intentó agrupar las proteínas de los diferentes grupos de acuerdo a su porcentaje de similitud secuencial y hallamos que en la mayoría de los casos, las proteínas no se agrupan en los diferentes grupos. Estos resultados demuestran que las diferencias estructurales de estos grupos está dada por sus características dinámicas propias, y no por una sobre representación de algún plegamiento o familia de proteínas en los mismos.

4.3.4. Caracterización de los movimientos independientes del *backbone*

Mencionamos anteriormente que las proteínas no siempre requieren grandes cambios conformacionales para cumplir su función biológica. Cuando obtenemos valores de RMSD bajos (en promedio $0.85\mathring{A}$) como los observados en las proteínas rígidas no significa que éstas proteínas carezcan de cambios conformacionales que puedan ser importantes para su función, un caso son los anteriormente mencionados en proteínas alostéricas [Nussinov and Tsai, 2015]. En esta sección estudiamos las presencia de cavidades y túneles, y su variación entre los confórmeros de cada proteína, con el fin de detectar cambios conformacionales localizados que no se observan a nivel de diferencias en los desplazamientos de los carbonos alfa que forman

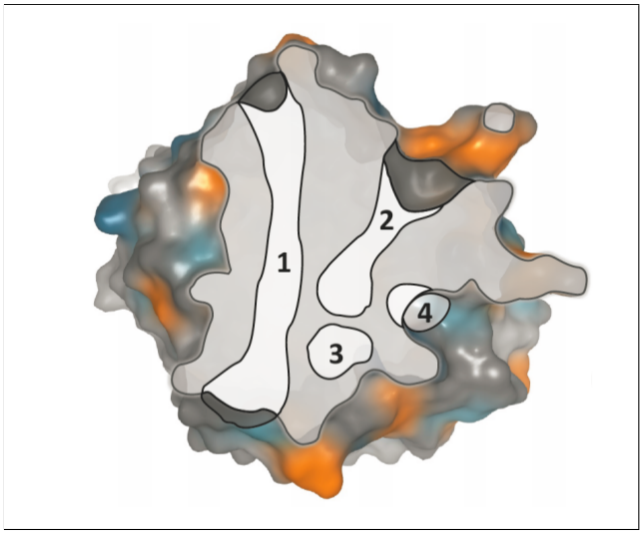

Figura 4.7: Los canales o poros (1) están representados por una vía que atraviesa la estructura de la proteína, sin ninguna interrupción por una cavidad interna, con ambos lados abiertos al solvente. Los túneles (2) están representados por una vía que conduce desde un sitio específico enterrado en el interior de una proteína (la mayoría de las veces desde un sitio activo o de unión) al entorno externo de la proteína (solvente). Las cavidades pueden ser definidas como *voids* (3) cuando se encuentran enterradas en el interior de la proteína y no son directamente accesibles desde la superficie, o como *pockets* o bolsillos (4) cuando se encuentran en la superficie de la proteína y son fácilmente accesibles por un ligando que se une a la misma. Imágen extraída y adaptada de [Strnad, 2014].

el *backbone*. Los túneles y cavidades son estructuras que conectan la superficie de la proteínas con las regiones más enterradas donde se puede encontrar un sitio activo o de unión, y que son esenciales en la mayoría de las proteínas [Gora et al., 2013, Pravda et al., 2014a]. En la Figura 4.7, podemos observar estas diferentes clases de mecanismos o estructuras funcionales que tienen las proteínas.

Con el objetivo de identificar y caracterizar las cavidades (las de mayor volúmen) en cada confórmero utilizamos el programa Fpocket (ver Métodos 4.6), y observamos que las cavidades presentan mayor volúmen en las proteínas parcialmente desordenadas y en las maleables respecto de las rígidas (Figura 4.8A). Vemos que mientras que en las parcialmente desordenadas y maleables las cavidades presentan volúmenes similares, hay una diferencia en el comportamiento de las mismas. Las regiones altamente flexibles (mayormente asociadas a IDRs) en las maleables definen cavidades de grandes volúmenes y además se observa gran variación entre sus confórmeros. Cuando las IDRs son removidas en maleables y parcialmente desordenadas, observamos que las cavidades en las maleables reducen significativamente su volumen, mientras que las cavidades de las parcialmente desordenadas no se ven afectadas (ver Tabla Sup. A.1). Las IDRs forman y definen, al menos en parte, las cavidades de las proteínas maleables,

posiblemente modulando la interacción con distintos sustratos, los cuales suelen ser muy diversos en este grupo de proteínas. Las cavidades en las proteínas parcialmente desordenadas y maleables además se encuentran más expuestas al solvente por lo que son más hidrofílicas respecto de las que se observan en las rígidas.

El tamaño y variación de los túneles entre los confórmeros se estimó utilizado el programa MOLE 2.0 (ver Métodos 4.6). Teniendo en cuenta los túneles de mayor longitud identificados por el MOLE, evaluamos la variación en la longitud de los mismos entre los confórmeros del par máximo para cada proteína de los tres grupos. Observamos que las proteínas rígidas presentan los valores más grandes en las medianas respecto de la variación de los túneles entre sus confórmeros de mayor movilidad (Figura 4.8B). Es interesante destacar que si bien anteriormente mostramos que estas proteínas casi no presentaban movimientos a nivel del *backbone*, presentan aperturas y cierres de túneles que inequívocamente están relacionados con la entrada y salida de substratos, y por lo tanto con su función biológica [Gora et al., 2013]. Esto se refleja al analizar las distribuciones de RMSD por posición de los residuos que rodean los túneles y los que no, donde vemos que los residuos involucrados en la formación del túnel presentan valores mayores que aquellos que no lo están (*Kolmogorov-Smirnov test* y *Wilcoxon rank sum tests*, $P < 0.01$). Además, las proteínas rígidas presentan los túneles de mayor longitud (normalizando por la longitud de la proteína) respecto de las maleables y parcialmente desordenadas (*Kruskal—Wallis rank sum test* $P = 0.01$ con *Nemenyi post-hoc test*) (Figura 4.8C). Cabe mencionar que cuando comparamos este parámetro entre maleables y parcialmente desordenadas no encontramos diferencias significativas.

Finalmente en esta sección de análisis de movimientos locales independientes de *backbone*, analizamos la red de contactos entre residuos en los confórmeros de máximo RMSD en los tres grupos de proteínas. Utilizando el método RING (*"Residue Interaction Network Generator"*) generamos las redes de contacto y calculamos el grado de interconectividad promedio de la red, es decir el número de contactos promedio de los residuos de esa estructura. Encontramos que las proteínas rígidas poseen en promedio valores más pequeños que las proteínas parcialmente desordenadas y maleables (grado promedio: 5, 5.21 y 5.16 respectivamente, *Kruskal—Wallis rank sum test* $P << 0.01$ con *Nemenyi post-hoc test*) (Figura 4.8D). Una mayor densidad de contactos en las regiones ordenadas de proteínas que experimentan transiciones

orden/desorden, puede estar asociada a un proceso evolutivo adaptativo para compensar la presencia de regiones altamente flexibles.

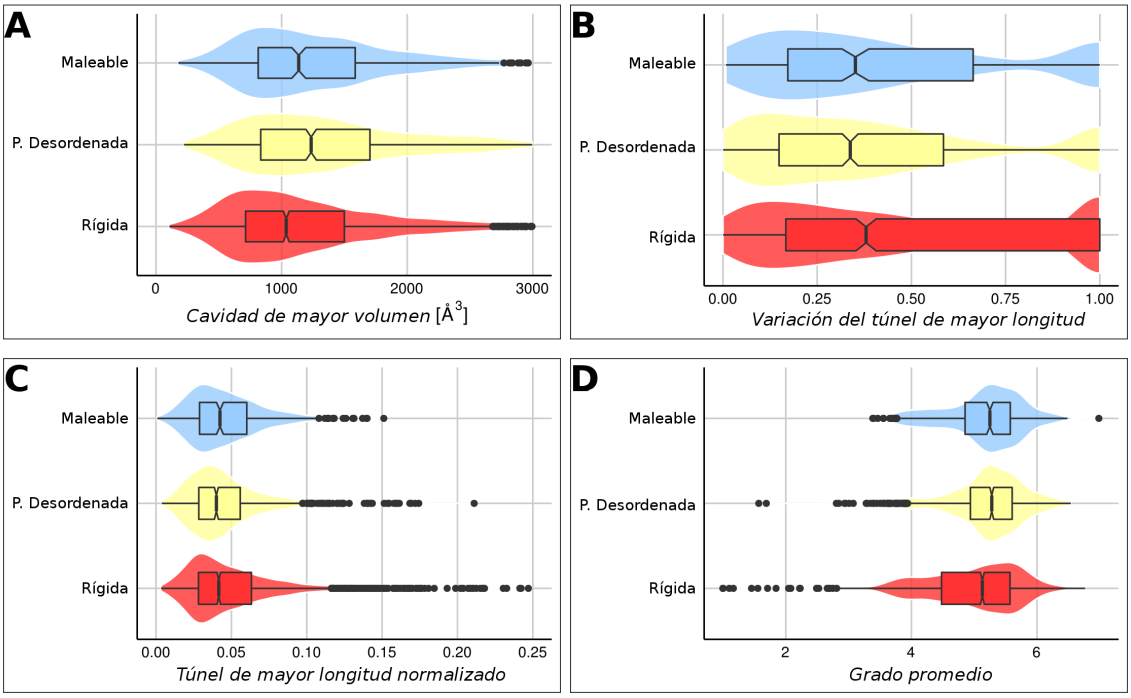

Figura 4.8: En cada *boxplot*, los extremos de la caja corresponden al primer y al segundo cuartil de la distribución, la línea dentro de la caja indica la mediana (segundo cuartil) y los *notches* muestran la desviación media absoluta (M.A.D). Además, el gráfico estilo violín debajo del *boxplot* muestra la densidad de probabilidad de la variable. A) Distribuciones de los volúmenes de las cavidades más grandes en los confórmeros del par de máxima diversidad conformacional. Las cavidades son significativamente más grandes en las proteínas maleables y parcialmente desordenadas que en las rígidas (*Kruskal—Wallis rank sum test* P << 0.01 con *Nemenyi post-hoc test*). B) Distribución de la variación del túnel de mayor longitud entre los confórmeros del par de máxima diversidad conformacional. Esta variable expresa la proporción de variación de la longitud de los túneles más largos entre los confórmeros del par). La mediana en las proteínas rígidas es significativamente mayor (*Wilcoxon rank sum test*, P < 0.001). C) Distribución de la proporción de los túneles más largos respecto de la longitud total de los confórmeros. La longitud del túnel se encuentra normalizada por la longitud en número de aminoácidos de esa estructura. D) Distribuciones del grado promedio de la red de contactos entre residuos de los confórmeros del par de máxima diversidad conformacional.

4.4. Discusión

Hemos encontrado que la distribución de la diversidad conformacional proteica puede interpretarse mediante al menos tres grupos de proteínas. Las diferencias entre éstos grupos no subyacen en una clasificación estructural común entre las proteínas de los mismos, ni en su actividad biológica o historia evolutiva. El comportamiento de las proteínas que componen cada grupo emerge de características estructurales y dinámicas que comparten cuando analizamos sus ensambles conformacionales. Estas características compartidas por las proteínas de

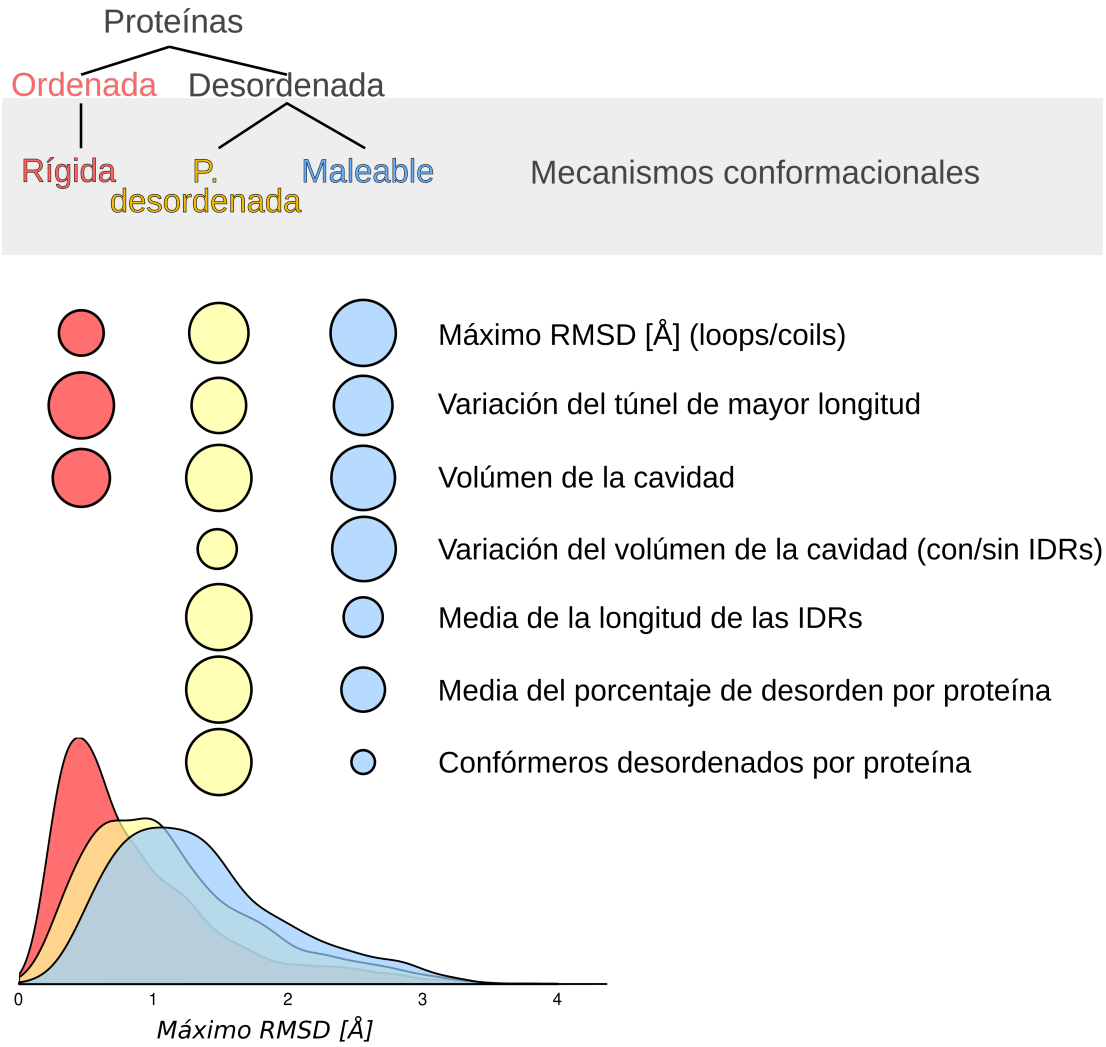

Figura 4.9: Comparación visual de las principales características estructurales de las proteínas que componen los tres grupos. El área de cada círculo es proporcional al promedio de la correspondiente medida cuantitativa.

un determinado grupo pueden verse como distintos mecanismos conformacionales asociados a la función proteica (Figura 4.9).

En general, todas las proteínas ordenadas tienen poca diversidad conformacional y representan el grupo de mayor tamaño (∼60 % de todo el set de datos). Este grupo se caracteriza por presentar leves diferencias a nivel del *backbone* entre sus confórmeros (máximo RMSD promedio = 0.83 Å) evidenciando que grandes cambios conformacionales no son un pre-requisito para que haya función biológica. Hemos llamado a éstas proteínas rígidas, sin embargo presentan movimientos significativos a nivel local (movimientos de residuos independientes del *backbone*) y presentan túneles de mayor longitud en comparación con las proteínas de los otros grupos. Aparentemente, éstas proteínas poseen mecanismos que son independientes de gran-

des movimientos en el *backbone*, como lo son la apertura y cierre de túneles que son requeridos para su función. Como vimos, los túneles son estructuras que permiten el tránsito de ligandos o substratos desde la superficie hasta un sitio de unión interno o el sitio activo de la proteína. Además, los túneles conectan las cavidades internas de las proteínas y promueven la salida de los productos [Gora et al., 2013]. Como describimos anteriormente, los residuos que delimitan los túneles en las proteínas rígidas poseen mayores valores de RMSD por posición que el resto de las posiciones. Esto indica que los pequeños movimientos que se producen en éstas proteínas, están asociados con el movimiento de estructuras funcionales como los túneles. Otro dato interesante, es que las proteínas rígidas poseen en promedio ligandos de menor peso molecular que las proteínas de los otros dos grupos, las cuales presentan cavidades de volúmenes mayores comparadas con las rígidas. Estas observaciones sugieren que la utilización de túneles para el transporte de substratos hacia el interior podría estar limitada sólo a moléculas de un tamaño en particular. La existencia de túneles abiertos y cerrados en los confórmeros como así también su variación en longitud, por ejemplo, en la dinámica de cuello de botella [Chovancova et al., 2012] o en compuertas conformacionales [Zhou et al., 1998], podrían definir diferentes constantes de unión en la proteína que expliquen su función biológica [Biedermannová et al., 2012], la especificidad enzimática [Pravda et al., 2014a] y un importante proceso regulatorio como el alosterismo [Gunasekaran et al., 2004]. Durante los últimos años, el alosterismo en ausencia de un cambio conformacional ha sido explicado en términos de los efectos entrópicos, que surgen a partir de los cambios en frecuencia y amplitud de las fluctuaciones térmicas alrededor de la conformación promedio de la proteína [del Sol et al., 2009, Cui and Karplus, 2008, Tsai et al., 2008]. Esta descripción surge como una alternativa a la consideración clásica de los cambios conformacionales implicados en el alosterismo [Monod et al., 1965, Koshland et al., 1966]. La contribución de movimientos pequeños como en la apertura y cierre de túneles y cavidades es lo que hoy se conoce como alosterismo entrópico [Cooper and Dryden, 1984] y que aún no se ha podido cuantificar con exactitud. Se conocen muchas proteínas donde los túneles y cavidades tienen una gran importancia en su función como ser: la citocromo P450 de humanos [Skopalík et al., 2008], la lumazina sintetasa de *A. aeolicus* [Pravda et al., 2014a] con más de 5 canales de una longitud superior a los 15 $\AA$ y la imidazol glicerol fosfato sintetasa de *T. maritima* [Myers et al., 2005] donde el movimiento de los residuos, como si fueran una

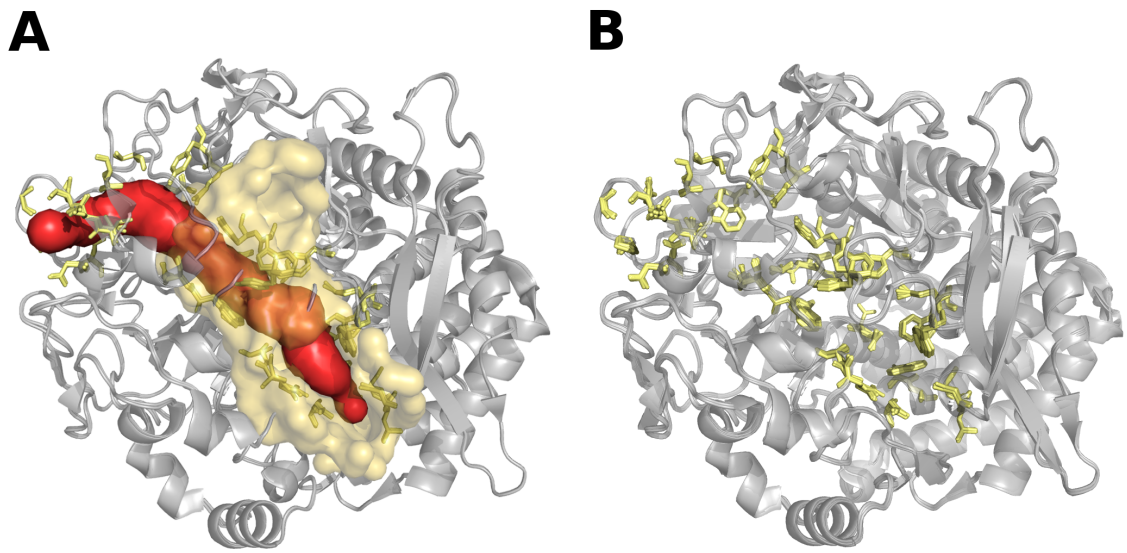

Figura 4.10: Superposición estructural de las conformaciones de la Celulosa cel48F de *Clostridium cellulolyticum*. Esta enzima es una endocelulasa procesiva con un gran sitio activo y de unión, donde se une una cadena de celulosa que ingresa a la proteína a través de un túnel localizado en la superficie de la proteína. En nuestro set de datos, esta proteína contiene 8 confórmeros (códigos PDB: 1F9D_A, 1F9O_A, 1G9G_A, 1FAE_A, 1FBO_A, 1FBW_A, 1FCE_A, 2QNO_A) con un máximo RMSD de 0.21 $\mathring{A}$ entre todas las comparaciones posibles. A) Podemos ver que casi no hay diferencia estructural significativa en el *backbone* entre los confórmeros del par de máximo RMSD (códigos PDB: 1F9O_A, 1G9G_A), sin embargo, el túnel (en rojo) de 65 $\mathring{A}$ de longitud (identificado por MOLE) aparece en un solo confórmero del par (código PDB: 1F9O_A) que es el que contiene ligandos unidos, mientras que en el otro confórmero éste túnel está ausente. B) La superposición de todos los confórmeros sólo muestra ligeras rotaciones y movimientos mínimos en los residuos que limitan el túnel (en amarillo), produciendo la apertura y el cierre del mismo. Además, simulaciones computacionales de dinámica molecular también confirman la rigidez de esta proteína [Vital de Oliveira, 2014].

especie de compuertas, regulan la actividad de los dos sitios activos que posee. Además, otras proteínas que surgen de observar nuestro set de datos como la celulosa cel48F con un túnel de 65 $\mathring{A}$ de (ver Figura 4.10).

Las proteínas desordenadas (que poseen IDRs en al menos un confórmero) se dividen en dos grupos: parcialmente desordenadas y maleables, ambos con mayor diversidad conformacional que las proteínas rígidas desarrolladas anteriormente (ver Figura 4.4B). En un trabajo previo hemos descrito que la presencia de regiones que experimentan transiciones orden/desorden incrementan la diversidad conformacional de la proteína [Zea et al., 2016]. Como ya sabemos que la función está relacionada con la dinámica de la proteína, la presencia de IDRs podría influir en la función biológica afectando al comportamiento de las regiones ordenadas/estructuradas de las proteínas [Schulenburg and Hilvert, 2013]. Estos dos grupos de proteínas con IDRs difieren entre ellos en la tendencia que tienen sus regiones a aparecer ordenadas o desordenadas en el ensamble de confórmeros. La composición de las IDRs entre los dos grupos es

diferente, y posiblemente este relacionado con la capacidad de ordenarse y desordenarse, esto explica la diferencia en el porcentaje de confórmeros con regiones entre parcialmente desordenadas y maleables. Como mencionamos anteriormente, la principal razón de que las maleables poseen mayor diversidad conformacional que las parcialmente desordenadas, está dada porque sus IDRs se ordenan en la estructura dando lugar a regiones altamente flexibles y por consecuencia introduciendo valores de RMSD altos cuando sus confórmeros son comparados.

Las proteínas parcialmente desordenadas y maleables poseen cavidades más grandes y túneles más cortos en comparación a las proteínas rígidas (ver Figura 4.8). Esto puede indicar diferentes formas de interaccionar con los substratos y ligandos. Sin embargo, estos dos grupos presentan además diferencias entre sus cavidades, las cuales son más expuestas al solvente e hidrofílicas que las de las proteínas rígidas. Cuando los residuos involucrados en una IDRs son removidos en las maleables, éstas reducen significativamente el volúmen de sus cavidades casi al nivel de las rígidas. Por el contrario, en las proteínas parcialmente desordenadas sus cavidades no resultan alteradas al remover estos residuos. Estas regiones de alta flexibilidad, aparecen desordenadas en una minoría de los confórmeros, y es una de las principales características que define a las proteínas maleables. Un típico ejemplo de éstas proteínas es la Calmodulina, la cual puede interaccionar con alrededor de otras 350 proteínas, que a pesar de tener regiones desordenadas es comúnmente considerada como una proteína ordenada [Tompa, 2010, Ikura and Ames, 2006] (ver Figura 4.11). Otras proteínas que se caracterizan por tener sitios activos flexibles son: la undecaprenyl-pirofosfato sintetasa de *E. coli* donde loops flexibles regulan el tamaño del producto [Ko et al., 2001], la quinasa dependiente de ciclina 2, la protrombina y la tripsina.

Al contrario de las proteínas maleables, en las parcialmente desordenadas las cavidades no se encuentran afectadas por las IDRs. Este grupo de proteínas muestran en promedio un mayor número de regiones *hinges* o bisagras y de dominios, de aquí asumimos que las IDRs podrían formar conexiones altamente flexibles entre dominios [Mittag et al., 2010]. Además, este grupo de proteínas posee en promedio regiones desordenadas más largas que las maleables (ver Tabla Sup. A.1) y alrededor del 5 % de las mismas tiene un porcentaje de desorden que podría abarcar dominios completos. Debido al alto contenido de desorden, la mayoría de los confórmeros de

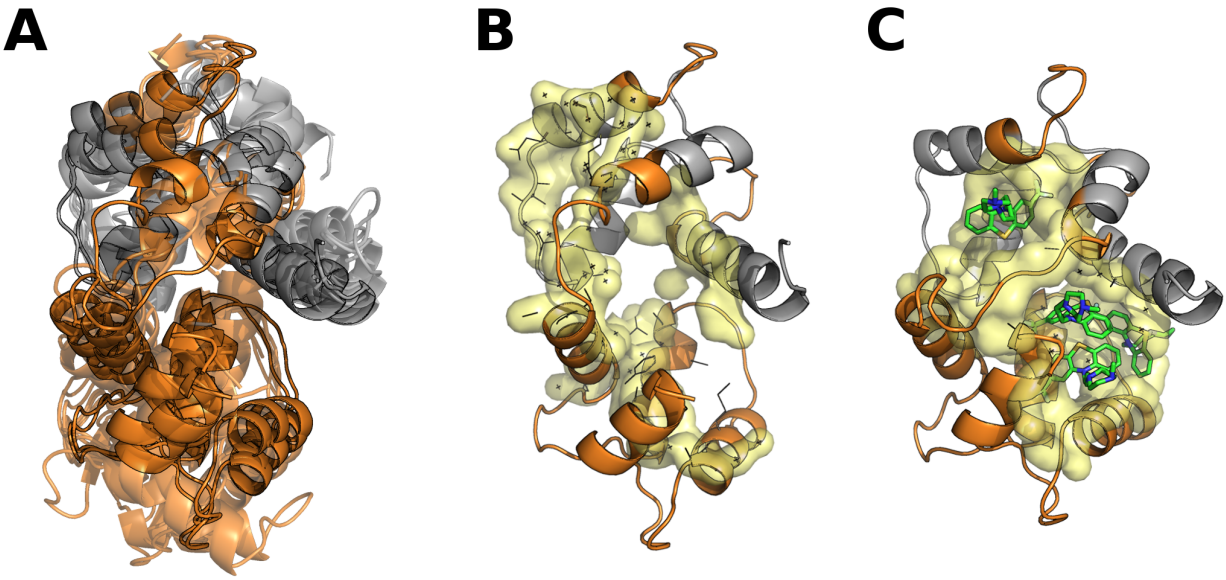

Figura 4.11: En nuestro set de datos la calmodulina tiene 79 confórmeros (8 con IDRs) con un máximo RMSD = 3.2$\AA$ y una región desordenada máxima que involucra el 34.5 % de sus residuos. La calmodulina contiene cuatro motivos denominados *EF-hand* donde un par de ellos forman un dominio globular. Cada uno de estos dominios globulares contiene un bolsillo que es lo suficientemente flexible como permitir la unión de diferentes proteínas *target* [Ikura and Ames, 2006]. Además, cada dominio globular puede adaptar su orientación relativa a través de conectores flexibles aumentando la capacidad de la calmodulina para interactuar con cientos de proteínas y substratos. A) Se muestran varios confórmeros de la calmodulina alineados estructuralmente (códigos PDB: 1LIN_A, 1NIW_E, 3G43_A, 4DCK_B, 2FOT_A, 2X0G_B, 2BE6_A, 2O60_A, 1CDL_A, 3GP2_A). Los residuos involucrados en IDRs, que están mayormente ordenados en la diferentes conformaciones, se encuentran representados en color naranja. B-C) Se observan los confórmeros del par de máximo RMSD, podemos ver que en un confórmero (B, código PDB: 1NIW_E) es más extendido y no presenta bien definida su cavidad, mientras que el otro confórmero que se encuentra unido a trifluoperazina (C, código PDB: 1LIN_A) es más compacto y tiene una cavidad bien definida y de gran volúmen. En este ejemplo podemos observar como la presencia de las regiones desordenadas que se ordenan y desordenan, modulan el volúmen de la cavidad permitiéndole a la proteína mayor plasticidad para la unión de substratos.

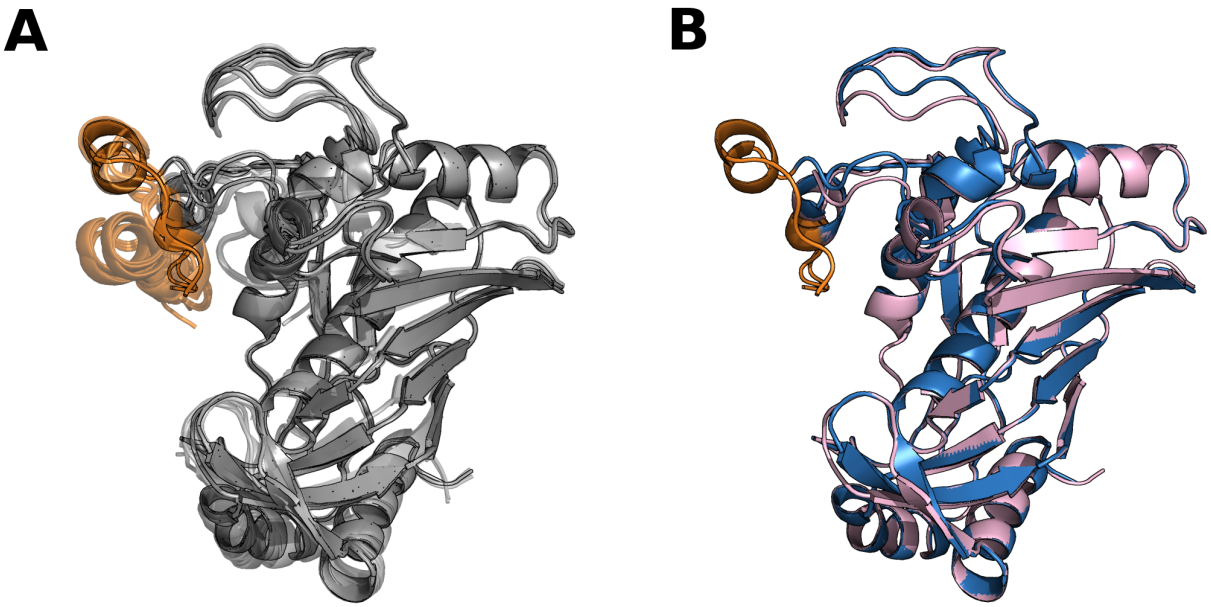

Figura 4.12: La timidilato sintetasa humana (TS) es la responsable de catalizar la metilación reductiva del dUMP a dTMP. Esta reacción es esencial para mantener el conjunto de nucleótidos necesarios durante el crecimiento celular. Esta enzima es un *target* valioso para los fármacos citotóxicos utilizados en la quimioterapia de cáncer, debido a su importancia en la replicación del ADN en las células que se dividen activamente [Peña et al., 2006, Lovelace et al., 2005]. La TS posee 15 confórmeros en nuestro set de datos (11 con IDRs), con máximo RMSD de 1.34 $\AA$ y un máximo porcentaje de desorden de 9 %. La región desordenada de esta proteína está compuesta por residuos de uno de los insertos que presenta la TS humana que la hace diferir a la de la TS bacteriana. A) Podemos ver todas las conformaciones de la TS superpuestas estructuralmente (códigos PDB: 1HW3_A, 1HZW_A, 1HZW_B, 1I00_A, 1I00_B, 1YPV_A, 3EDW_X, 3EHI_X, 3GH2_X, 3N5E_A, 3N5E_B, 3N5G_A, 4G2O_X, 4G6W_X, 4GD7_A). La región desordenada cambia de desorden a orden en varios confórmeros (residuos en naranja), por ejemplo, en la conformación activa (código PDB: 1HZW_A esta región se encuentra ordenada, mientras que en la inactiva (código PDB: 1YPV_A) se encuentra desordenada. B) Par de confórmeros de máximo RMSD (1YPV_A en rosa, 4GD7_A en azul claro). La región desordenada aparece parcialmente ordenada en uno de los dos confórmeros, presentando en la estructura 4GD7_A 11 residuos más ordenados (formando una hélice alfa) que la misma región en el otro confórmero.

las proteínas parcialmente desordenadas presentan IDRs (en promedio el 70 %), por lo que creemos que éstas proteínas están formadas por IDRs que han sido bien caracterizadas en la literatura y que coinciden con el estado del arte en el campo de IDPs [van der Lee et al., 2014, Tompa, 2010, Tompa, 2012]. En este grupo encontramos muchos ejemplos de IDPs donde sus transiciones orden/desorden han sido ampliamente caracterizadas y asociadas a la función biológica como: las proteínas L15, L11, L10 y L32e del ribosoma 50S, serina/treonina quinasa PLK1, la proteína G y la timidilato sintetasa (Figura 4.12).

Históricamente, la diversidad conformacional ha sido asociada con los grandes cambios conformacionales gracias a los trabajos pioneros de Chothia, Lesk y Gerstein [Lesk and Chothia, 1984, Gerstein et al., 1994]. Consecuentemente muchos ejemplos de proteínas que muestran

grandes cambios conformacionales fueron recolectados y caracterizados en la base de datos de movimientos de proteínas creada por Gerstein [Gerstein et al., 1994]. Ahora, ¿dónde se encuentran éstas proteínas en los grupos que hemos descrito anteriormente? La mayor parte de éste tipo de proteínas caracterizadas años atrás no poseen IDRs y las podemos encontrar en la cola de la distribución de las proteínas rígidas. Para cuantificar la presencia de las mismas en el grupo de proteínas rígidas, observamos el valor de 75 percentil de la distribución de RMSD y encontramos que el 75 % de las mismas está por debajo de 1.13 $\mathring{A}$. Por encima de éste valor encontramos 769 proteínas sin IDRs que poseen grandes cambios conformacionales. La mayoría de los cambios están dados por movimientos entre dominios o regiones de loops, movimientos de cuerpo rígido o en otros casos movimientos más complejos [Rashin et al., 2010]. Este tipo de proteínas que se caracterizan por realizar este tipo de movimientos representan un 15 % del set de datos que se estudio en este capítulo. En la Figura 4.13 podemos ver un ejemplo de proteína altamente móvil, la cual no posee regiones desordenadas y sin embargo experimenta grandes cambios conformacionales. Esta enzima conocida como *3-phosphoshikimate 1-carboxyvinyltransferase* (AroA) posee 12 confórmeros en nuestro set de datos y un máximo RMSD de 3.34 $\mathring{A}$. Su movimiento es mayormente dado por una región de *hinge* que se ubica entre los dos dominios estructurales que posee.

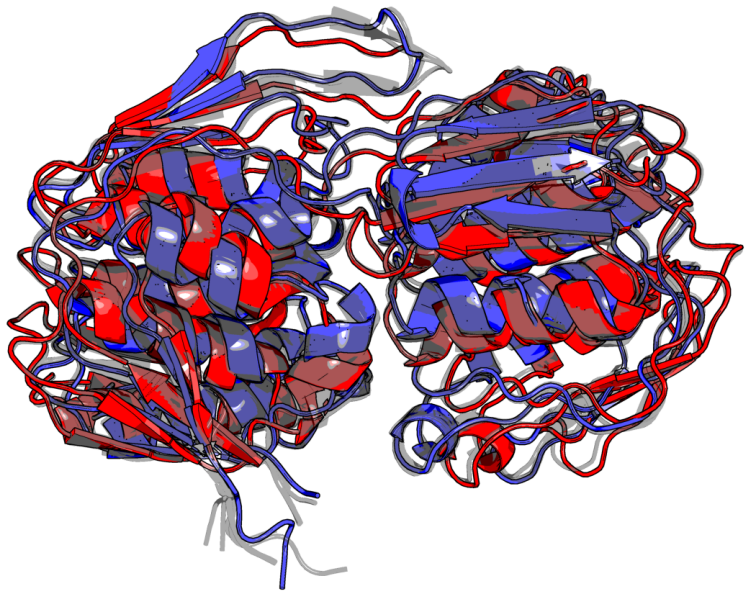

Figura 4.13: Estructuras superpuestas de la enzima *3-phosphoshikimate 1-carboxyvinyltransferase* (AroA) perteneciente al set de proteínas ordenadas muy dinámicas. En rojo y azul se encuentran los confórmeros que componen el par de máxima diversidad conformacional, mientras que en gris transparente los otros diez confórmeros restantes.

4.5. Conclusiones destacadas

El sostenido avance de la biología estructural en los últimos años se ha centrado en dilucidar la relación estructura-función para mejorar la anotación de proteínas que muestran estructuras similares y/o secuencias (por ejemplo [Loewenstein et al., 2009]). Esta tendencia práctica y útil muchas veces oculta el concepto que la relación estructura-función es casi un accidente evolutivo en vista de la gran redundancia de funciones similares sostenidas por estructuras de proteínas diferentes [Elias and Tawfik, 2012, Orengo et al., 1994], o por el contrario cuando estructuras similares realizan funciones diferentes (por ejemplo en *superfolds* [Liu et al., 2003, Orengo et al., 1994]). En este trabajo hemos encontrado al menos cuatro mecanismos o relaciones estructura-función expresados como relaciones estructura-dinámica que representan diferentes formas de lograr una gran diversidad de funciones biológicas. Creemos que estas categorías; proteínas rígidas, ordenadas muy dinámicas, parcialmente desordenadas y maleables, son algunas representantes de las distintas que podrían llegar a existir en el espacio estructural de las proteínas. Por ejemplo, no hemos incluido en esta clasificación las proteínas completamente desordenadas (Figura 4.14) que poseen ensambles de confórmeros sumamente complejos, difíciles de obtener y analizar mediante las técnicas actuales.

Creemos que las características derivadas de los ensembles son lo suficientemente generales como para describir el comportamiento biológico de las proteínas, como lo sugirió H. Frauenfelder en su búsqueda de conceptos a partir de la dinámica de proteínas [Frauenfelder and McMahon, 1998].

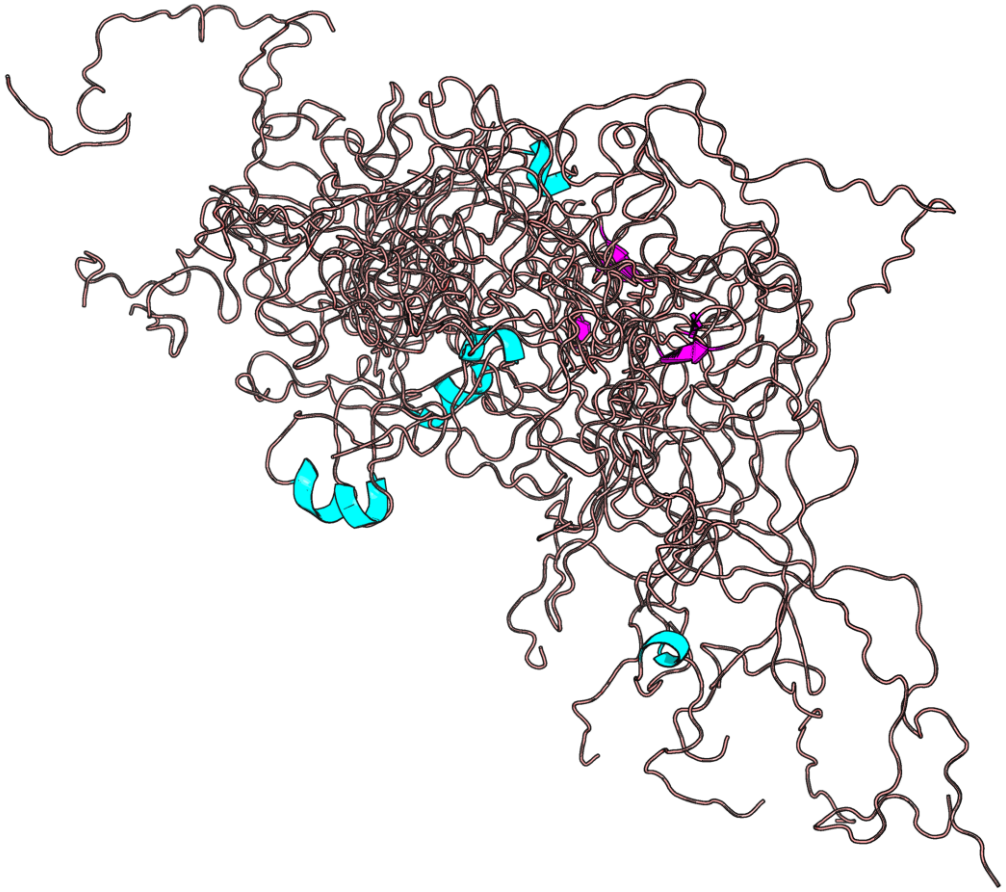

Figura 4.14: Ensamble de confórmeros correspondientes a la proteína intrínsecamente desordenada Alfa-sinucleína extraído de la base de datos PED [Varadi et al., 2014].

4.6. Métodos específicos del capítulo

4.6.1. Generación del set de datos

Utilizando la base de datos CoDNaS, cuya creación fue desarrollada en el capítulo 2, reclutamos proteínas con diversidad conformacional utilizando los siguientes criterios:

- Proteínas cuyos confórmeros hayan sido obtenidos por DRX con una resolución igual o mejor a 2.5 $\AA$.
- Proteínas que tengan al menos 5 confórmeros.
- 100 % de identidad secuencial entre los confórmeros de cada proteína.

De esta forma se obtuvo un set de datos compuesto por 4,791 cadenas de proteínas con un total de 74,417 confórmeros y de 1,186,312 de comparaciones estructurales entre confórmeros.

Por cada proteína se identificó el par de confórmeros que maximizan el RMSD entre todas las comparaciones posibles, el cual se utiliza como medida de la extensión de la diversidad conformacional de esa proteína. Además, el hecho de que sólo se tomaron proteínas con estructuras de buena resolución y al menos 5 confórmeros, garantiza una correcta estimación de la diversidad conformacional de esa proteína. Estas condiciones están respaldadas por el trabajo de Best et al. y nuestras observaciones detalladas en el capítulo 3. Donde se sugiere que es necesario considerar un cierto número de estructuras para estimar la diversidad conformacional a partir de estructuras cristalográficas redundantes [Best et al., 2006] (alrededor de 5 para la flexibilidad del *backbone* y 20 para la heterogeneidad de las cadenas laterales), como así también evitar la mezcla de confórmeros obtenidos por diferentes técnicas experimentales y la utilización de confórmeros de buena resolución. En nuestro set de datos tenemos en promedio 15.5 confórmeros por proteína.

4.6.2. Identificación de regiones intrínsecamente desordenadas (IDRs)

La identificación y definición de IDRs en las estructuras de los confórmeros se realizó utilizando la información que provee PDB acerca de los residuos faltantes o *missing residues* en la estructura. Se sabe que aquellos residuos que no se pueden observar en el mapa de densidad electrónica, que surge de la difracción de Rayos-X, son desordenados y por lo tanto no se puede resolver la posición de los mismos en la estructura [Dunker et al., 2001]. De esta manera definimos una región desordenada donde hay al menos 5 o más residuos faltantes en la estructura y que los mismos no se encuentren en los extremos amino o carboxilo terminales de la estructura (primeros y últimos 20 residuos), ya que estas zonas tienden a ser faltantes debido a que las proteínas se mueven mucho en sus extremos.

La composición de aminoácidos de las regiones desordenadas se estudió utilizando el programa *Composition Profiler* [Vacic et al., 2007], el cual permite de una manera muy rápida investigar el sesgo en la composición de aminoácidos entre dos conjuntos de secuencias de proteínas. Un conjunto de secuencias de estudio (en este caso las secuencias de las IDRs) se puede analizar contra una muestra representativa de proteínas que proporcione una distribución de fondo. Nosotros utilizamos como distribución *background* PDB Select 25 [Hobohm and

Sander, 2008], que es una base de datos de secuencias de proteínas capaces de ser cristalizadas, y por lo tanto su composición está enriquecida en residuos que se suelen encontrar ordenados en la estructura. Además, el programa estima la significancia asociada al enriquecimiento o no de cada residuo realizando un *boostraping* de 10,000 iteraciones.

4.6.3. Caracterizaciones estructurales

Además de las caracterizaciones estructurales e información biológica que proporciona CoDNaS y que se describieron anteriormente, en este capítulo se realizaron algunas caracterizaciones particulares.

- Cálculo del RMSD global por tipo de estructura secundaria: en primer lugar se asignó la estructura secundaria a cada confórmero del par de máximo RMSD, utilizando el programa DSSP [Kabsch and Sander, 1983]. A los 8 estados de estructura secundaria que informa el DSSP los agrupamos en tres estados (alfa-hélices, láminas-beta y *loops* como se explica en los Métodos 5.5. Utilizando el paquete de R *Bio3d* [Grant et al., 2006] se identificaron los pares de residuos equivalentes que compartían la misma clasificación de estructura secundaria en el par de confórmeros y se utilizaron esos residuos para calcular nuevamente el RMSD global de aquellos involucrados en alfa-hélices, láminas-beta y loops.

- Identificación de regiones bisagras o *hinges*: la detección de estas regiones generalmente se realiza utilizando programas de alineamiento estructural flexible. Para lograr la mayor superposición entre dos estructuras, si es necesario, estos programas le permiten a ciertos residuos adquirir flexibilidad. Estos residuos suelen encontrarse en regiones bisagra de la estructura y participan en el movimiento de la proteína. Nosotros utilizamos el programa FlexProt [Shatsky et al., 2002] que se encuentra optimizado para detectar regiones de *hinges* y no requiere un conocimiento a priori de las regiones bisagras para realizar el alineamiento. La cantidad de regiones bisagras o *hinges* se obtuvo contando la cantidad de regiones identificadas por el FlexProt que maximizan la superposición de los residuos de los confórmeros del par máximo.

- Radio de giro (R_g): el radio de giro es una medida de compactitud de la estructura proteica, se refiere a la distribución de los componentes de un objeto alrededor de un eje (átomos de los residuos alrededor del centro de masa de la estructura). El cálculo del R_g se le realizó a cada confórmero del par de máximo RMSD, utilizando el grupo de herramientas de MMTSB (disponible en `http://blue11.bch.msu.edu/mmtsb/`) que cuentan con el programa *rgyr* que realiza este cálculo. Debido a que el R_g depende del tamaño de la proteína, en nuestro análisis usamos un R_g normalizado por el R_g de una esfera ideal del mismo volúmen de la estructura que se usó para calcularlo, esto se hizo siguiendo la metodología aplicada por *Lobanov et al.* [Lobanov et al., 2008].

- Detección de cavidades: las cavidades en los confórmeros del par de máximo RMSD se identificaron utilizando el programa FPocket [Le Guilloux et al., 2009]. FPocket puede identificar e informar más de una cavidad en la estructura, en estos casos utilizamos como referencia en cada estructura la cavidad de mejor puntuación (de acuerdo a la evaluación que hace FPocket) y de mayor volúmen identificada por FPocket (que llamaremos cavidad máxima). Luego, se calculó la variación de los volúmenes de las cavidades máximas con/sin los residuos involucrados en IDRs utilizando la siguiente ecuación 4.1:

$$cavVolVar = \frac{\mid max(pV_{1,IDR}, pV_{2,IDR}) - max(pV_1, pV_2) \mid}{max(max(pV_{1,IDR}, pV_{2,IDR}, max(pV_1, pV_2))} \tag{4.1}$$

  Donde pV_i es el volúmen (con o sin los residuos involucrados en IDRs) en el confórmero correspondiente del par (1 o 2).

- Detección de túneles: la identificación y caracterización de los túneles en las estructuras de los confórmeros del par de máximo RMSD se realizó utilizando el programa MOLE 2.0 [Berka et al., 2012]. Se configuró el programa con los siguientes valores en algunos de sus parámetros: *Probe Radius* y *Origin Radius* 3 $\mathring{A}$, *Interior Threshold* 1.25 $\mathring{A}$ y valores por defectos en los parámetros restantes [Pravda et al., 2014b]. Del archivo de salida XML se extrajo la cantidad y longitudes de el/los túnel/es identificado/s, residuos que delimitan cada túnel, entre otros. La proporción de la variación entre los túneles más largos en cada confórmero del par se calculó utilizando la siguiente ecuación 4.2:

$$tunLenVar = \frac{\mid L_1 - L_2 \mid}{max(L_1, L_2)} \tag{4.2}$$

Donde L_i es la longitud del túnel más largo en el confórmero correspondiente del par. Cuando éste número tiende a 0 significa que pasamos de tener un túnel abierto de longitud L en una conformación a otro de menor longitud o cerrado.

- Cálculo de redes de interacción entre residuos o *Residue Interaction Networks* (RINs): en una RIN, cada residuo de la estructura se representa como una nodo en la red, y las conexiones entre los mismos están dadas por la interacción entre los residuos en la estructura. Las interacciones consideradas pueden ser puentes de disulfuro, puentes salinos, puentes de hidrógeno e interacciones aromáticas. Las redes de los confórmeros del par de máximo RMSD fueron estimadas utilizando el programa RING [Martin et al., 2011] y analizadas con los *plug-in* de Cytoscape denominados *NetworkAnalyzer* y *RINalyzer* [Doncheva et al., 2012].

  El grado de un nodo (d_i) en la red es el número de conexiones que tiene ese nodo con los demás nodos de la red. Por lo que el grado promedio de la red (d) esta dado por:

$$\langle d \rangle = \frac{\sum d_i}{N} \tag{4.3}$$

  Donde N es el número de nodos (residuos) de la red. Un valor promedio grande indica que esta red puede estar más interconectada entre los distintos nodos que posee. Lo que en términos de una estructura proteica es que sus residuos se encuentren en contacto.

En cada paso se fueron utilizando diferentes *scripts* tanto para ejecutar los programas en miles de estructuras como para así también extraer la información de las salidas de los mismos y asignarla a las estructuras de los confórmeros, entre otras tareas. Los *scripts* fueron mayormente desarrollados en el lenguaje de programación *Perl* y los gráficos y análisis estadísticos se hicieron utilizando el paquete R [Team, 2017].

Capítulo 5

Relación entre la diversidad conformacional y la divergencia secuencial y estructural en familias de proteínas homólogas[1]

5.1. Resumen

La alta correlación entre la divergencia secuencial y la divergencia estructural es un concepto clave en el modelado por homología. La consecuencia práctica de esta relación es que proteínas homólogas que posean secuencias similares también van a tener estructuras similares, ya que la estructura proteica es más conservada que la secuencia en la evolución. Sin embargo, la diversidad conformacional de la proteína reduce esta correlación entre secuencia y estructura ya que podemos encontrar variación estructural sin cambios en la secuencia. En este capítulo exploramos el impacto que tiene la diversidad conformacional en la relación entre divergencia estructural y secuencial. Primeramente, encontramos que la extensión de la diversidad conformacional de una proteína puede ser tan alta como la divergencia estructural entre proteínas homólogas de una misma familia. Además, y como esperábamos, la diversidad conformacional debilita la correlación entre la divergencia estructural y secuencial, la cual es

[1]Este capitulo está basado en la publicación: [Monzon et al., 2017b].

más compleja y ruidosa de lo que se ha sugerido en otros trabajos. Sin embargo, encontramos que podemos utilizar información a priori proveniente de la relación estructura-función para mejorar esta correlación. En este capítulo mostraremos que familias de proteínas con baja diversidad conformacional muestran una mejor correlación entre la divergencia estructural y secuencial, que aquellas que poseen diversidad conformacional alta. Esta pérdida de correlación al aumentar la diversidad conformacional de la proteína podría perjudicar los resultados obtenidos por el modelado por homología en proteínas que son muy móviles. Finalmente, mostramos que la presencia de transiciones de orden/desorden puede proveer información de cierta importancia para mejorar el desempeño del modelado por homología en proteínas muy dinámicas.

5.2. Introducción

El modelado basado en un molde o *template-based modeling* (TBM) es la técnica más confiable, precisa y rápida para predecir la estructura terciaria de una proteína [Baker and Sali, 2001, Zhang, 2008, Qu et al., 2009]. TBM abarca las técnicas de *threading* y de modelado por homología [Fiser, 2010]. En los últimos años el crecimiento de PDB ha incrementado el espacio de plegamientos que se conocen [Khafizov et al., 2014], los cuales en combinación con los métodos de detección de moldes o *templates* y métodos de TBM, permiten la predicción de aproximadamente el 50 % de las estructuras del proteoma humano y de casi el 70 % para algunos proteomas eucariotas [Schwede, 2013, Anand et al., 2011].

TBM se basa en el hecho de que proteínas homólogas, con un porcentaje de identidad moderado, van a tener estructuras 3D similares. En un trabajo pionero realizado por Lesk y Chothia describieron que la divergencia estructural (DE) aumenta con la distancia evolutiva, medida como el porcentaje de identidad, y sigue una relación no-lineal (ver Figura 5.1) [Chothia and Lesk, 1986]. Esto en otras palabras quiere decir que las secuencias muy similares muestran sólo algunas diferencias muy pequeñas entre sus estructuras, las cuales empiezan a aumentar drásticamente cuando el porcentaje de identidad secuencial cae por debajo del 30 %. Esta tendencia fue luego confirmada por otros trabajos posteriores [Russell et al., 1997, Wilson et al., 2000, Panchenko et al., 2005]; sin embargo en algunos estudios recientes encuentran una

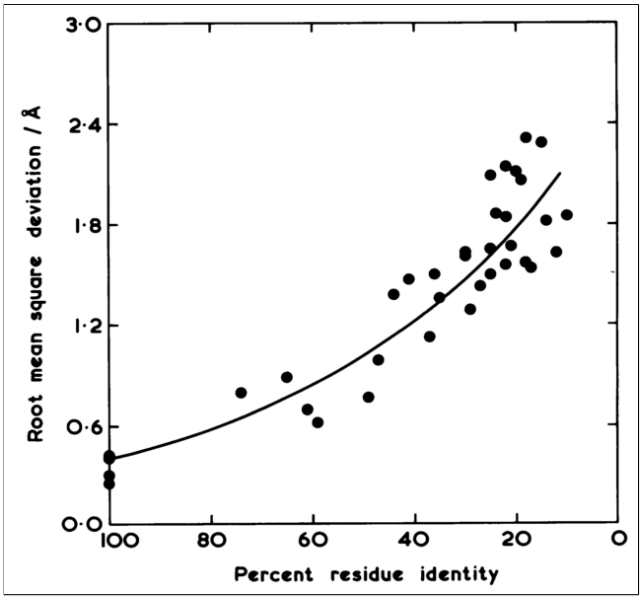

Figura 5.1: Relación entre el porcentaje de identidad de secuencia y el RMSD del core de átomos conservado, para 32 pares de proteínas homólogas. Imágen extraída de [Chothia and Lesk, 1986].

relación lineal entre la divergencia estructural y secuencial [Illergård et al., 2009b, Wood and Pearson, 1999a].

Utilizando la relación entre identidad de secuencia y distancia estructural, el primer paso en TBM comprende la búsqueda de un molde adecuado para realizar el modelado. De acuerdo con los estudios mencionados anteriormente acerca de las relaciones estructura y secuencia, el mejor molde será aquel con mayor identidad de secuencia a la secuencia de la proteína que se desea modelar o *target*, y el que posea una estructura determinada experimentalmente conocida. Además, la resolución de la estructura que se usará de molde, la presencia de substratos en la misma y la historia evolutiva, mejoran la selección de los mismos [Fiser, 2010]. Este paso es seguido por el alineamiento entre la secuencia *target* y la secuencia del molde para detectar las regiones conservadas y variables, y finalmente se realiza el modelado y refinamiento de la estructura. Usando el porcentaje de identidad de secuencia como una medida de la distancia evolutiva entre la secuencia a modelar y el molde, se ha encontrado que los modelos estructurales difieren 1-2 $\mathring{A}$ de RMSD de la estructura nativa determinada en forma experimental, en el caso de moldes con más del 50 % de identidad de secuencia. En el caso de moldes entre 30 y 50 % de identidad, la distancia entre el modelo obtenido y la estructura nativa es de alrededor 4 $\mathring{A}$, mientras que para moldes por debajo de 30 % de identidad, los métodos *ab-initio* superan a las técnicas TBM [Baker and Sali, 2001].

A pesar de las destacadas contribuciones que tienen y han tenido los métodos TBM en muchas áreas de las ciencias biológicas [Zhang, 2009], aún es difícil obtener modelos 3D de alta calidad. Los errores derivados del alineamiento entre la secuencia del molde y el *target* [Kopp and Schwede, 2004], como así también los que se producen en el refinamiento del modelo inicial para obtener estructuras más cercanas a las nativas, son algunos de los mayores problemas a resolver para mejorar la calidad del modelo 3D. Sin embargo, queda aún un concepto clave que no ha sido tenido en cuenta en las técnicas de TBM y que mejoraría sus predicciones. Este concepto está relacionado con la naturaleza del estado nativo proteico, compuesto por un ensamble de confórmeros en equilibrio y que hemos desarrollado en profundidad en esta tesis. En este sentido, las técnicas de TBM deberían avanzar hacia la predicción del ensamble nativo y no simplemente de la estructura de la secuencia *target*. Esta cuestión ha sido señalada por otros autores, que destacan el impacto de la diversidad conformacional en las técnicas de TBM [Burra et al., 2009], básicamente porque un determinado molde (con una determinada distancia evolutiva a la secuencia *target*) puede tener diferentes confórmeros que muestrean el espacio conformacional [Illergård et al., 2009a, Kosloff and Kolodny, 2008]. Como hemos mencionado a lo largo de los distintos capítulos, estos confórmeros pueden presentar diferentes tipos de movimientos en un amplio rango de valores de RMSD, por lo que teniendo en cuenta este espectro de diversidad estructural que le confiere la diversidad conformacional a una misma secuencia, los métodos TBM deberían ser re-evaluados. Generalmente los protocolos de evaluación a ciegas de modelos, por ejemplo, las evaluaciones del tipo CASP (*"Critical Assessment of Techniques for Protein Structure Prediction"*) [Cozzetto et al., 2009], utilizan solo un confórmero de la secuencia molde. Claramente el rendimiento y calidad del modelo es altamente dependiente de esa selección como se ha demostrado anteriormente [Hrabe et al., 2015, Palopoli et al., 2016, Marks et al., 2017].

Sin embargo, aparte del efecto de la diversidad conformacional en las técnicas de TBM, el problema más complejo de resolver es cómo se codifica la información estructural en la estructura primaria de la proteína [Wood and Pearson, 1999b]. El llamado modelo local sostiene que unas pocas posiciones en la proteína definen el arreglo estructural global. El comportamiento no lineal en la relación estructura-secuencia respalda esta hipótesis debido a la observación de que se requiere una gran cantidad de variación en la secuencia para cambiar drásticamente la

estructura (principalmente por debajo del 20-25 % de identidad de secuencia). Por el contrario, el modelo global respalda la idea de que varias posiciones distribuidas a lo largo de la proteína definen la disposición estructural. Una relación lineal entre cambio estructural y divergencia de secuencia apoyaría este modelo mostrando un cambio proporcional entre esas variables. Sin embargo, considerando que una sola secuencia puede adoptar varias conformaciones, hace aún más complicado predecir cómo las sustituciones no sinónimas se correlacionan con divergencia estructural.

Como un concepto clave a ser tenido en cuenta en las técnicas de TBM, en este capítulo exploramos el impacto que tiene la diversidad conformacional en la relación entre la divergencia estructural y secuencial. Para ello utilizamos un set de datos curado compuesto por 2024 proteínas con diversidad conformacional proveniente de CoDNaS, agrupadas en 524 familias de homólogos (>30 % de identidad de secuencia local y un 90 % de cobertura). Estas proteínas están representadas en las cuatro clases estructurales de la base de datos CATH siendo el 17 % mayormente alfa, 25 % mayormente betas, el 57 % alfa betas y 1 % proteínas con poca estructura secundaria. Cada una de las familias de homólogos fueron analizadas para derivar las similitudes estructurales y secuenciales entre las distintas proteínas que las componen. Encontramos que al usar un set de datos altamente redundante desde el punto de vista estructural (para considerar la diversidad conformacional proteica) se desdibuja la relación entre divergencia estructural y secuencial mostrada en estudios previos. Sin embargo, vemos que esta tendencia puede ser abordada usando información *a priori* proporcionada por la relación estructura-función de esa proteína. Mostraremos que familias que contienen proteínas con baja diversidad conformacional (las que denominamos rígidas en el capítulo anterior), muestran una buena correlación entre la divergencia estructural y secuencial. Por el contrario, esta correlación se reduce drásticamente en familias con gran diversidad conformacional en sus proteínas. Esta pérdida de correlación podría influenciar los resultados obtenidos por las técnicas de TBM en proteínas muy móviles. Finalmente, mostraremos que la presencia de regiones orden/desorden puede ser de utilidad como información previa para mejorar el rendimiento de los métodos de TBM.

5.3. Resultados

5.3.1. La diversidad conformacional proteica puede ser tan grande como la divergencia estructural en una familia

En primer lugar, realizamos los alineamientos estructurales de todas las estructuras contra todas dentro de cada una de las 524 familias, las cuales contenían 2024 proteínas con diversidad conformacional experimental extraídas de CoDNaS y con un total de 37,775 estructuras. Para cada par de proteínas homólogas dentro de una familia, el porcentaje de identidad de secuencia global fue obtenido con el algoritmo de Needleman-Wunch [Needleman and Wunsch, 1970]. Cada proteína dentro de una familia está representada por un ensamble de confórmeros, por lo que la divergencia estructural entre dos proteínas homólogas va a estar dada por la comparación de todos sus confórmeros entre si, y denominaremos MSD (*"Maximum structural divergence"*) al par de confórmeros provenientes de dos proteínas homólogas que posean en máximo RMSD entre todas las comparaciones posibles, es decir que maximizan la divergencia estructural de ese par de proteínas homólogas. Además, cada proteína en las distintas familias tiene asociada su diversidad conformacional que será definida como lo hemos hecho a lo largo de esta tesis, el par de confórmeros que dan el máximo RMSD de todas las comparaciones posibles (de ahora en mas lo denominaremos CD por *"Conformational Diversity"*). En la Figura 5.2 podemos ver un esquema del protocolo explicado anteriormente.

En la Figura 5.3A podemos observar la relación entre MSD (puntos verdes) y porcentaje de identidad de secuencia, además podemos observar los valores de CD de cada proteína (puntos rojos). Los puntos verdes muestran el comportamiento típico que se ha mostrado en otros trabajos previos entre la divergencia secuencial y estructural [Lesk and Chothia, 1984, Wood and Pearson, 1999b], con un precipitado decrecimiento en la similitud estructural a bajos porcentajes de identidad (aproximadamente por debajo de 30 %). Sin embargo, en nuestro set de datos nosotros observamos grandes variaciones de divergencia estructural a altos porcentajes de identidad de secuencia, esto es una consecuencia de proteínas con una diversidad conformacional equivalente a la divergencia estructural de esa familia, en términos de valores de RMSD.

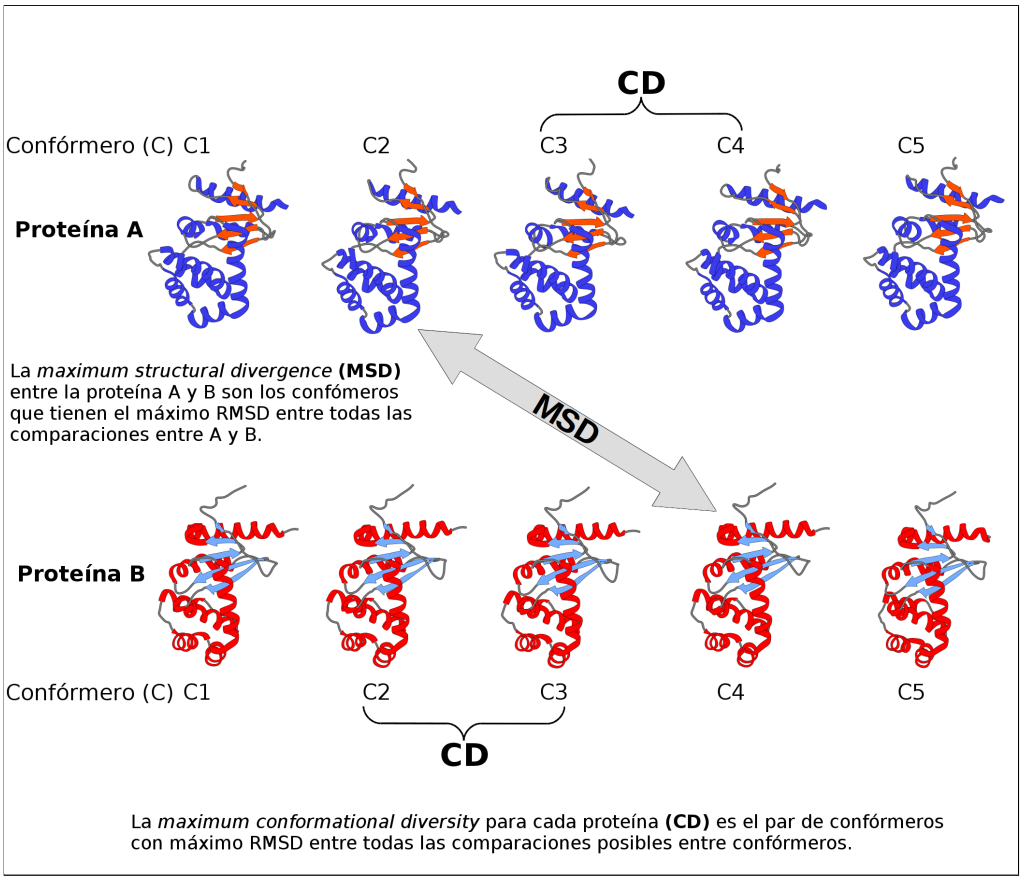

Figura 5.2: Representación esquemática de cómo se derivan la máxima divergencia estructural (MSD) entre proteínas homólogas y la máxima diversidad conformacional (CD) de una proteína.

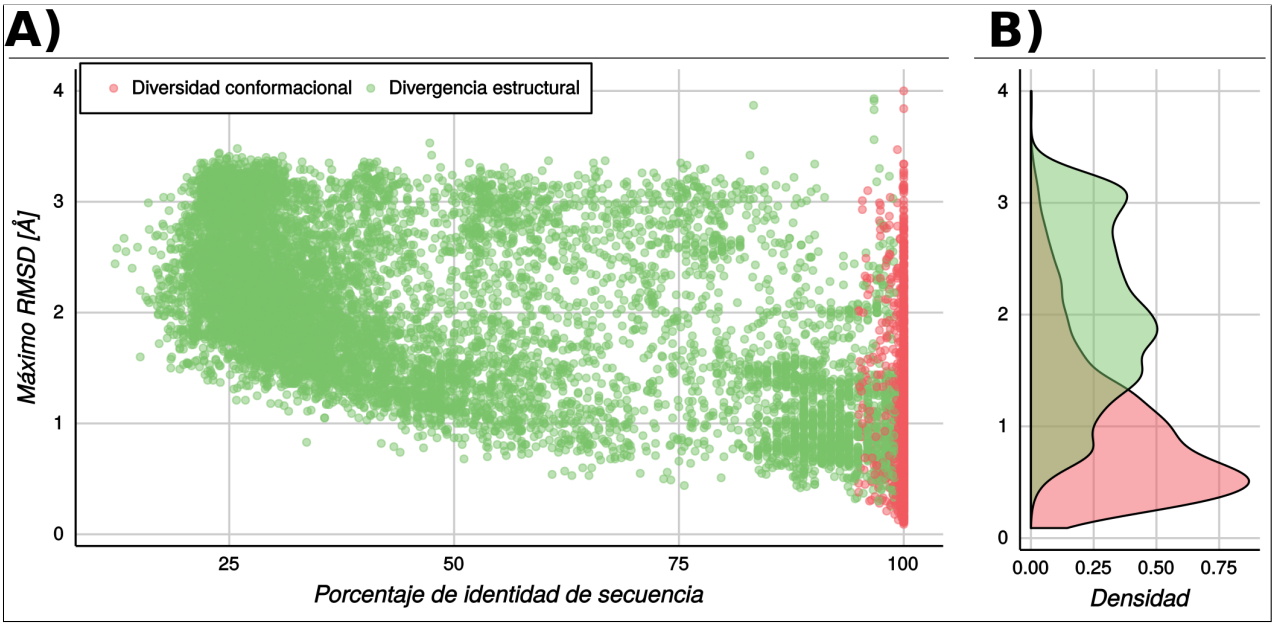

Figura 5.3: Máximo RMSD (MSD y CD) versus el porcentaje de identidad de secuencia. Cada punto representa el máximo RMSD obtenido al comparar todos contra todos los confórmeros provenientes de dos proteínas homólogas (MSD), o provenientes de la misma proteína (CD). A) Los puntos verdes son comparaciones entre dos proteínas homólogas y los puntos rojos son comparaciones entre confórmeros de la misma proteína. B) Distribuciones de los valores de máximo RMSD entre dos proteínas homólogas (verde) y entre confórmeros de la misma proteína (roja).

La distribución de la Figura 5.3B muestra valores mayormente moderados de RMSD con un promedio de 1 $\mathring{A}$ para la distribución de diversidad conformacional (color roja). Esto está en concordancia con trabajos previos [Burra et al., 2009]. El 90 percentil de la distribución de CD muestra un RMSD por debajo de 2 $\mathring{A}$, y el 10 % de las proteínas poseen una diversidad conformacional tan grande como la MSD, que se obtiene de comparar proteínas homólogas (aproximadamente 3 $\mathring{A}$). Este resultado es interesante, ya que nos indica que una secuencia dada puede existir en un espacio conformacional tan grande como la divergencia estructural que se obtiene por la acumulación de sustituciones (proceso evolutivo). Usando el concepto de la diversidad conformacional es posible entender que proteínas homólogas cercanas (por ejemplo, por encima de 80 % de identidad de secuencia) pueden tener distintos valores de RMSD (bajos o altos) dependiendo de que confórmeros de esas proteínas se estén comparando. Por lo tanto, la diversidad conformacional puede conducir a grandes valores de RMSD entre proteínas homólogas en períodos de tiempos evolutivos cortos, en lugar de alcanzar estos valores a través de un largo proceso de acumulación de mutaciones en la secuencia (que por ejemplo, puede tomar millones de años). Además del RMSD, se estimaron otras dos medidas de divergencia estructural como la fracción de estructura secundaria que no se conserva ($diffSS$) y el cambio entre enterrado/expuesto en la accesibilidad relativa al solvente ($diffRSA$) (ver Métodos 5.5) que se pueden observar en las Figuras sup. B.5 y B.6, respectivamente. Podemos ver que ambas medidas siguen la misma tendencia que se observa en la Figura 5.3A para el RMSD. Otra vez es interesante destacar la alta dispersión de variación en RSA y estructura secundaria a valores altos de identidad de secuencia.

La otra consecuencia importante de lo mostrado anteriormente es que la correlación entre la divergencia secuencial y estructural es más débil a la encontrada en trabajos previos [Panchenko et al., 2005,Illergård et al., 2009b,Wood and Pearson, 1999a]. En otras palabras, debido a la diversidad conformacional, una secuencia dada puede adoptar diferentes conformaciones, por lo tanto el cambio estructural debido a sustituciones no sinónimas en un proceso de evolución divergente hará que la relación entre secuencia y estructura sea mas ruidosa. De hecho, la relación entre MSD y porcentaje de identidad de secuencia (puntos verdes en la Figura 5.3A) da un coeficiente de correlación *rho* de Spearman de -0.52. Esta relación entre secuencia y estructura veremos que resulta más clara en términos de la diversidad conformacional, cosa

que abordaremos más adelante en este capítulo.

5.3.2. Selección del molde y diversidad estructural

Como mencionamos anteriormente, las técnicas TBM requieren el uso de una proteína con estructura conocida como molde. La identificación de este molde puede realizarse mediante una gran variedad técnicas con distinta sensibilidad [Qu et al., 2009, Yan et al., 2013, Xiang, 2006]. El punto clave en este paso es la selección del mejor molde, el cual según la relación entre la divergencia estructural y secuencial será el que maximice la cobertura y el porcentaje de identidad con la secuencia *target* [Rost, 1999]. Sin embargo, y como podemos ver en la Figura 5.3, este criterio no es tan simple como lo que se ha establecido anteriormente. En la Figura 5.4, mostramos la distribución de MSD entre pares de proteínas homólogas en diferentes intervalos de porcentaje de identidad de secuencia. Se puede observar la gran variación de los valores de RMSD en los diferentes intervalos. Además, todos los intervalos presentan un promedio en valor máximo de RMSD de 3.54 $\mathring{A}$ con una desviación estándar de 0.23 $\mathring{A}$. En consecuencia, la selección del molde no es tan sencillo y no basta simplemente con seleccionar una estructura en un intervalo de identidad dado, ya que no se sabe cómo esas estructuras pertenecen al ensamble conformacional de la secuencia que se modelará [Palopoli et al., 2016].

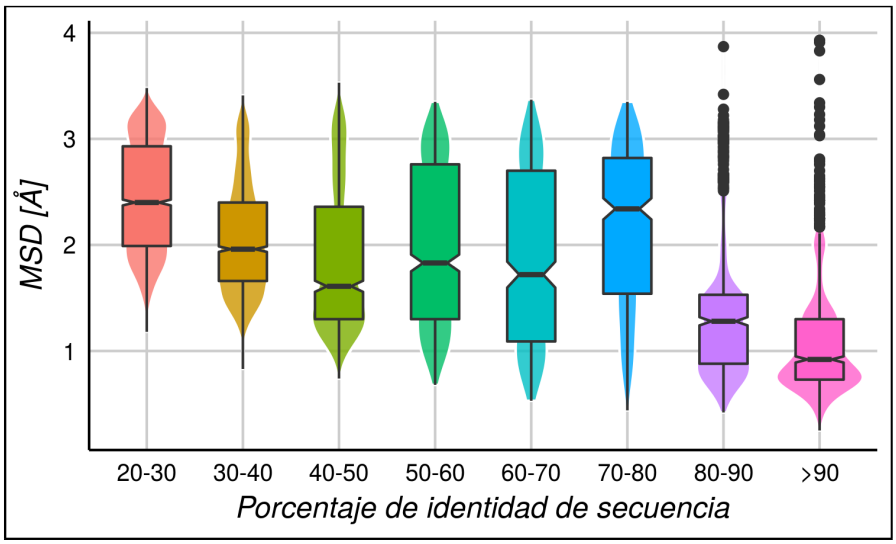

Figura 5.4: Distribuciones de MSD en diferentes intervalos de 10 % de identidad.

No obstante, las distribuciones por intervalos de identidad de secuencia mostradas en la Figura 5.4 podrían estar influenciadas por algunas familias que posean una diversidad es-

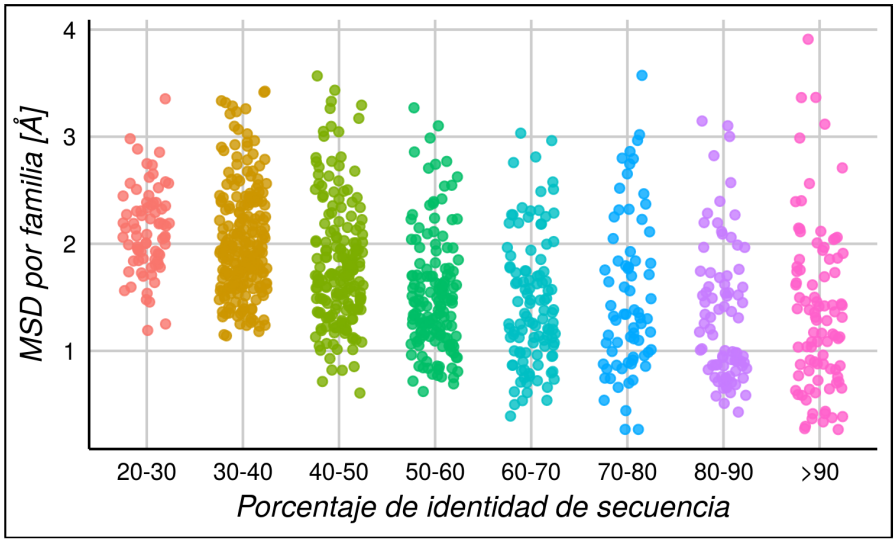

Figura 5.5: Distribuciones de MSD en diferentes intervalos de 10 % de identidad agrupado por familia. Cada punto representa el promedio de MSD para todos los pares de proteínas homólogas de una familia en un determinado intervalo de identidad de secuencia.

tructural excesivamente grande o pequeña. Por este motivo, en la Figura 5.5 mostramos el promedio de MSD para cada familia del set de datos, en los diferentes intervalos de identidad secuencial. Los valores de RMSD promediados se encuentran entre 1.34 y 2.10 $\mathring{A}$ para los diferentes intervalos (con una desviación estándar entre 0.19 y 0.29), mostrando que la dispersión de valores de MSD no está sólo relacionada con el porcentaje de identidad de secuencia entre las proteínas comparadas. En la Figura Sup. B.7, se muestra que la mediana de MSD no depende de que tan poblada esté ésa familia (cantidad de pares proteínas homólogas) en nuestro set de datos.

Teniendo en cuenta estos resultados, la selección del molde adecuado dependerá fuertemente de si la diversidad conformacional de esa proteína es grande o pequeña. Entonces, a éste punto nos preguntamos ¿Cuáles serían las recomendaciones generales para seleccionar un buen molde? Veremos a continuación que la relación entre la divergencia estructural y la diversidad conformacional será de ayuda.

5.3.3. ¿Cómo se relacionan la divergencia estructural y la diversidad conformacional?

En un escenario evolutivo simple de una familia de proteínas, podemos pensar que la diversidad conformacional es un rasgo conservado entre sus miembros y que si todas las proteínas

muestran entre sí valores de RMSD alrededor de 0.5 $\mathring{A}$ (valor equivalente al error experimental de DRX) al comparar todos sus confórmeros, esto indicaría que la población de conformaciones accesibles para esas proteínas es casi idéntica. Es importante marcar que las diferencias estructurales entre estos confórmeros son imperceptibles a nivel del *backbone*, pero sabemos por lo que hemos mostrado en el capítulo anterior y en otros trabajos, que hay diversidad conformacional a nivel de pequeños cambios en las cadenas laterales de los residuos acompañando la apertura de túneles y cavidades [Eyal et al., 2005, Daily and Gray, 2007].

Considerando que una familia tiene una presión selectiva para mantener su diversidad conformacional (por ejemplo, por una restricción funcional), la mayor parte de la divergencia estructural en esta familia habría sido originada por la acumulación de mutaciones no-sinónimas. Por el contrario, en el caso que una familia haya sido originada por una ancestro común que poseía alta diversidad conformacional, el proceso de divergencia debido a la acumulación de sustituciones no-sinónimas se vería incrementada a causa de la cantidad de conformaciones accesibles por esas proteínas y eventualmente se incrementaría la divergencia estructural de esa familia. Nosotros encontramos que la dispersión de la diversidad conformacional en una familia (medida en término del desvío estándar de los valores de diversidad conformacional de sus proteínas) es baja, lo que posiblemente indique que la diversidad conformacional en una familia tienda a conservarse (ver Figura sup. B.8). Sin embargo, es necesario realizar un estudio más exhaustivo para poder responder la pregunta de si la diversidad conformacional se conserva o no dentro de una familia, y si lo hace, en qué magnitud.

Nuestra hipótesis central es que la diversidad conformacional de una proteína estará correlacionada con la MSD que alcance esa familia. Para demostrar esto, estudiamos la correlación entre la diversidad conformacional promedio de las proteínas que conforman una familia y la MSD promedio de esa familia. Encontramos un coeficiente de correlación de Pearson de 0.75 (P-valor < 0.07) y podemos observar esta relación en la Figura 5.6. Como la MSD está relacionada con la extensión de divergencia secuencial que posee cada familia, en la Figura sup. B.9 mostramos que la relación encontrada anteriormente se mantiene en familias con diferentes rangos de identidad de secuencia.

En vista a los resultados mostrados en la Figura 5.6, estudiamos la relación entre la diver-

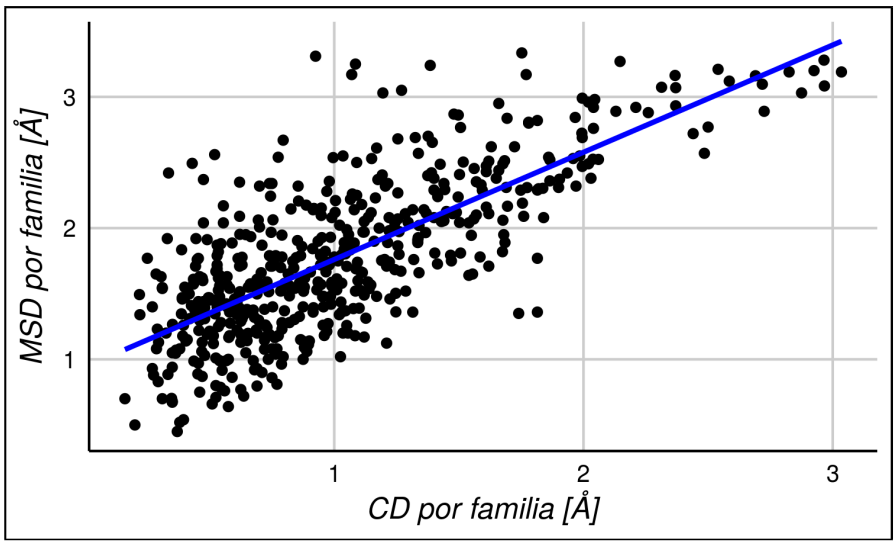

Figura 5.6: Relación entre la MSD y la CD. Cada punto representa el RMSD promedio para la MSD y la CD en una determinada familia de proteínas. Los datos muestran una correlación de Pearson de 0.75.

gencia estructural y secuencial dividiendo el set de datos entre pares de proteínas homólogas con baja y alta diversidad conformacional ($\leq$0.5 y >0.5 $\AA$, respectivamente, obtenida como la diversidad conformacional promedio entre un par de proteínas homólogas). Este valor de corte seleccionado es cercano al error que se obtiene en DRX al comparar estructuras de la misma proteína en las mismas condiciones experimentales [Burra et al., 2009]. Es interesante ver que el coeficiente de correlación de Spearman *rho* entre divergencia estructural (MSD) y secuencial es -0.83 (con un P-valor $<$ 0.01), en el subgrupo de proteínas homólogas con baja diversidad conformacional. Por otro lado, el coeficiente de correlación de Spearman *rho* en proteínas con alta diversidad conformacional es -0.51 (con un P-valor $<$ 0.01) (ver Figura 5.7). Estos resultados indican que la conocida correlación entre secuencia y estructura [Lesk and Chothia, 1984] es fuerte sólo en el subgrupo de proteínas homólogas con una baja diversidad conformacional. En estas proteínas homólogas, la función biológica puede ser realizada por confórmeros casi idénticos a nivel del *backbone*. Esto tiene como consecuencia que la relación entre la secuencia y la estructura sea directa, es decir, el cambio en la estructura es proporcional al cambio observado en el nivel de secuencia. Por el contrario, en el subgrupo de proteínas homólogas con alta diversidad conformacional, dos escenarios pueden ser posibles: la función biológica es menos estricta con una sola conformación, o inversamente, la función requiere una gran plasticidad de la estructura. En este sentido, en el subgrupo con diversidad

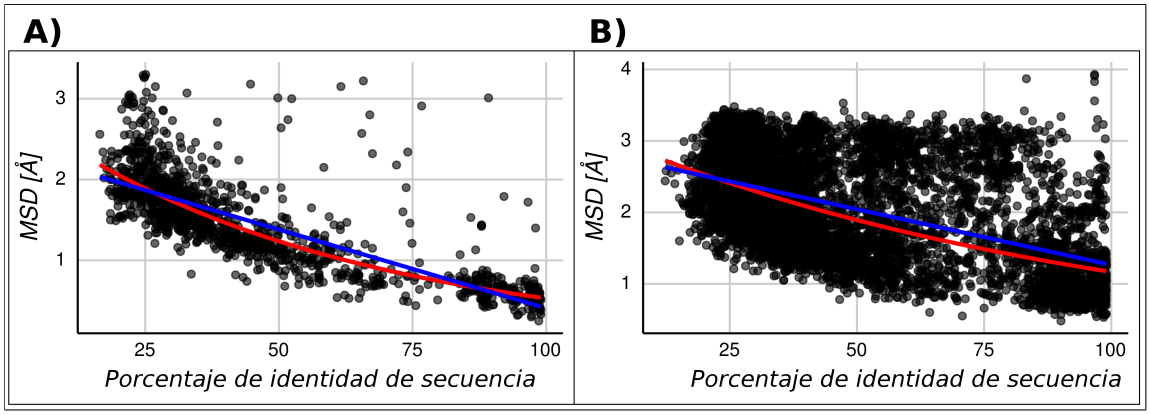

Figura 5.7: MSD versus porcentaje de identidad de secuencia entre pares de proteínas homólogas. Las regresiones lineales (roja) y exponenciales (azul) se observan por encima del gráfico de puntos. A) Pares de proteínas homólogas con una diversidad conformacional promedio menor o igual a 0.5 $\AA$. La expresiones de las regresiones lineales y exponenciales son $RMSD = 2{,}354 - 0{,}019SEQID$ y $RMSD = exp^{1{,}06}exp^{-0{,}017SEQID}$, respectivamente. B) Pares de proteínas homólogas con una diversidad conformacional promedio mayor a 0.5 $\AA$. La expresiones de las regresiones lineales y exponenciales son $RMSD = 2{,}823 - 0{,}015SEQID$ y $RMSD = exp^{1{,}121}exp^{-0{,}010SEQID}$, respectivamente.

conformacional por debajo de 0.5 $\AA$ de RMSD, la relación entre secuencia y estructura tiene un ajuste lineal y exponencial con valores de R^2 de 0.54 ± 0.15 y 0.66 ± 0.12, respectivamente. Mientras que para el subgrupo con alta diversidad conformacional con RMSD por encima de 0.5 $\AA$, la relación entre secuencia y estructura tiene un ajuste lineal y exponencial con valores de R^2 de 0.23 ± 0.21 y 0.28 ± 0.18, respectivamente. Estos valores de R^2 son la media obtenida mediante el testeo de los subgrupos realizando validación cruzada en 5 oportunidades. Además, encontramos que dividiendo en subgrupos a diferentes intervalos de diversidad conformacional, los coeficientes de correlación encontrados cambian monotónicamente a medida que la diversidad conformacional se incrementa (ver Tabla sup. A.2).

De acuerdo a estos resultados, las técnicas TBM serán más precisas en familias de proteínas con baja diversidad conformacional, ya que el cambio esperado en la estructura será proporcional al cambio en la secuencia. En estas familias, la elección del molde va a estar dada por aquél que posea el mayor porcentaje de identidad y cobertura, el cual será el que defina la precisión de las técnica de TBM. Como podemos ver en la Figura 5.7A, ambas relaciones lineales y exponenciales dan un RMSD de aproximadamente 0.45 $\AA$ para 100 % de identidad de secuencia. Por el contrario, en las familias con alta diversidad conformacional (Figura 5.7B), esta relación pierde predictibilidad debido a la gran variabilidad para la misma secuencia al 100 % de identidad (aproximadamente, un RMSD 1.3 $\AA$ para ambas relaciones, lineal y ex-

ponencial). Las diferencias entre si es mejor el ajuste lineal y exponencial para los subgrupos no son tan importantes, ya que podrían asociarse con diferentes causas (como las variaciones intra-familia o la dependencia no lineal de RMSD con la longitud de la proteína, entre otras). Sin embargo, las interpolaciones de las regresiones a 100 % identidad, son informativas sobre la importancia de la DC y la variabilidad estructural de los posibles moldes.

Ahora bien, ¿cómo podemos utilizar esta información y traducirla en consejos prácticos para el uso de los métodos de TBM?. Si bien sabemos que es muy difícil conocer *a priori* o predecir la diversidad conformacional a partir de una secuencia, podemos usar nuestros trabajos previos sobre la mayor diversidad conformacional que muestran las proteínas con IDRs [Zea et al., 2016, Monzon et al., 2017b]. En la siguiente sección estudiaremos la relación entre divergencia estructural y secuencial en proteínas ordenadas/desordenadas.

5.3.4. ¿Cómo se relaciona el desorden con la divergencia estructural?

Sabemos que las IDRs y las IDPs están involucradas en un gran repertorio de importantes funciones biológicas [Xie et al., 2007, Oldfield and Dunker, 2014] y en el capítulo anterior mostramos como la presencia de estas regiones en los confórmeros del ensamble nativo incrementa la diversidad conformacional de las proteínas. En vista a los resultados obtenidos, donde proteínas con mayor diversidad conformacional exhiben mayor divergencia estructural, estudiaremos la relación analizada en la sección anterior pero ahora entre pares de proteínas con regiones desordenadas en al menos un confórmero (que las llamaremos "Desordenadas") y proteínas sin regiones desordenadas en ninguno de sus confórmeros (que las llamaremos "Ordenadas"). Este fue el mismo criterio que utilizamos en el capítulo anterior. Encontramos que los pares de proteínas homólogas clasificadas como desordenadas muestran valores de MSD mayores que aquellas ordenadas (Figura 5.8). Estas distribuciones son estadísticamente diferentes con los tests de *Wilcoxon* y *Kolmogorov Smirnov*, P-valor < 0.01.

Dado entonces que los pares de proteínas homólogas con IDRs muestran una mayor MSD, esperamos que la correlación entre divergencia estructural y secuencial se vea afectada, así como mostramos anteriormente en la Figura 5.7. Encontramos que el coeficiente de correlación

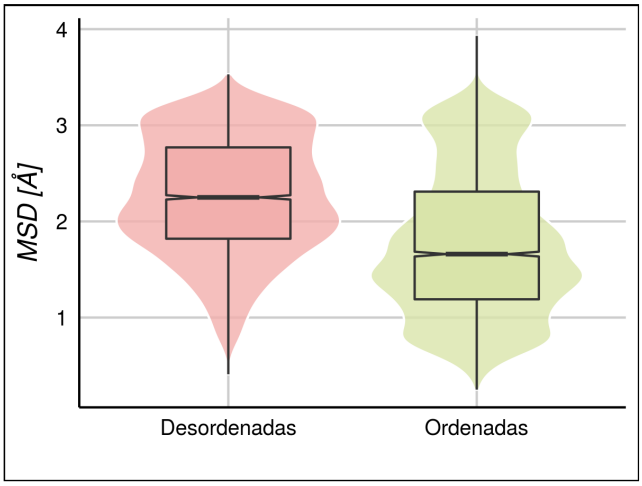

Figura 5.8: Distribuciones de MSD en pares de proteínas homólogas "ordenadas" y "desordenadas". El set desordenado posee 4,439 pares de proteínas homólogas y el set ordenado 5,034.

de Spearman *rho* es -0.36 y -0.71 para los pares de proteínas homólogas desordenadas y ordenadas, respectivamente. Estos resultados muestran que la presencia de IDRs en la estructura molde y/o *target* podría indicar una mayor diversidad conformacional y consecuentemente que la relación entre secuencia y estructura sea menos predecible. Ahora bien, si repetimos las regresiones realizadas en la sección anterior pero separando en pares de proteínas homólogas ordenadas y desordenadas, obtenemos las relaciones que se observan en la Figura 5.9.

En base a los coeficientes de correlación obtenidos, podemos decir que la presencia de regiones desordenadas tiene una capacidad moderada de predecir el ruido que se genera en la relación entre divergencia estructural y secuencial en proteínas muy móviles. Por otra parte, los resultados muestran que la mayoría de las proteínas ordenadas muestran valores bajos-moderados de diversidad conformacional, como hemos descrito anteriormente [Monzon et al., 2017b]. Teniendo en cuenta estas consideraciones, y la capacidad de predecir fácilmente y en forma confiable las regiones desordenadas en las proteínas, esta información podría utilizarse como guía en las técnicas de TBM.

5.4. Discusión

El análisis de las relaciones estructura-secuencia resulta fundamental en varias áreas que se centran en el estudio de las proteínas, como predicción de la función biológica [Sadowski and Jones, 2009], proteómica estructural [Drew et al., 2011], evolución proteica [Xia and Levitt,

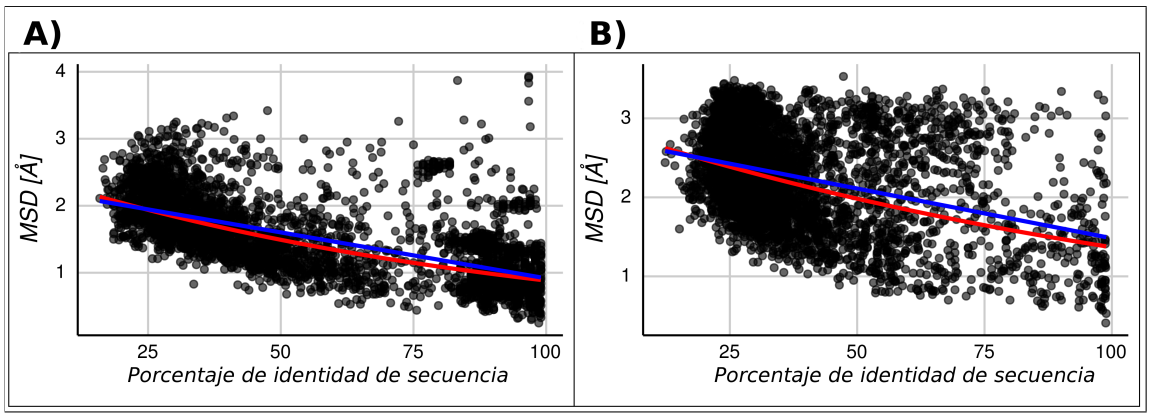

Figura 5.9: MSD versus porcentaje de identidad de secuencia entre pares de proteínas homólogas. Las regresiones lineales (roja) y exponenciales (azul) se observan por encima del gráfico de puntos. A) Pares de proteínas homólogas sin regiones desordenadas. La expresiones de las regresiones lineales y exponenciales son $RMSD = 2{,}288 - 0{,}014SEQID$ y $RMSD = exp^{0{,}092}exp^{-0{,}01SEQID}$, respectivamente. B) Pares de proteínas homólogas con regiones desordenadas. La expresiones de las regresiones lineales y exponenciales son $RMSD = 2{,}741 - 0{,}012SEQID$ y $RMSD = exp^{1{,}055}exp^{-0{,}007SEQID}$, respectivamente.

2004] y modelado por homología [Qu et al., 2009], sólo por mencionar algunos ejemplos. Entre todas las formas en que la relación estructura-secuencia podría adoptar, nos hemos centrado en cómo la presencia de la diversidad conformacional proteica podría influir en la relación entre secuencia y cambio estructural, y por lo tanto, afectar a las técnicas TBM. Un punto central en esta relación, fue establecido por el trabajo de Lesk y Chothia, donde mostraron que el éxito en la predicción de la estructura 3D de una proteína dependerá de la identidad entre el molde seleccionado y la secuencia *target* [Lesk and Chothia, 1984]. Básicamente, a partir del comportamiento entre RMSD y el porcentaje de identidad establecieron una relación en la cual la similitud estructural se mantenía a pesar de decrecer la identidad en secuencia. Estos resultados luego fueron verificados y soportados por numerosos estudios [Flores et al., 2007, Russell et al., 1997, Panchenko et al., 2005, Illergård et al., 2009b, Koehl and Levitt, 2002, Hubbard and Blundell, 1987, Russell and Barton, 1994]. Estos estudios encuentran una correlación de moderada a alta entre diferentes parámetros equivalentes a la similitud secuencial y estructural, como distancia evolutiva, significancia estadística del RMSD, entre otros. Además, encuentran ajustes lineales y no-lineales en el comportamiento de ésta relación, y una baja variación estructural al 100 % de identidad de secuencia (aproximadamente de 0.5 $\AA$). Sin embargo, estos métodos utilizan una única estructura por proteína como representante de su estado nativo, justificando que una secuencia al 100 % de identidad no presenta variación

estructural.

En este capítulo, encontramos que el grado de diversidad conformacional está relacionado con la máxima divergencia estructural (MSD) observada en una familia, y que la relación entre divergencia estructural y secuencial es más compleja que lo que se pensaba en trabajos previos. Primero, encontramos que la diversidad conformacional de una proteína puede ser tan grande como la divergencia estructural en términos de RMSD (Figura 5.3). Las distribuciones de la Figura 5.3 también evidencian que muchas de las proteínas de nuestro set de datos tienen una diversidad conformacional de moderada a baja, indicando que podrían funcionar con pequeños movimientos localizados que son imperceptibles al nivel del *backbone* [Monzon et al., 2017b, Mesecar et al., 1997], tema tratado en el capítulo anterior.

Algunos estudios destacaron la importancia de la diversidad conformacional en las técnicas de TBM [Burra et al., 2009, Kosloff and Kolodny, 2008, Palopoli et al., 2016], sin embargo la consideración de la CD en el estudio de la relación secuencia-estructura a menudo era considerada como una fuente de sesgo o "ruido" estructural [Illergård et al., 2009b, Wood and Pearson, 1999a]. Por un lado, evitando introducir ruido por información redundante (por ejemplo, considerar la DC) permitiría asumir que los cambios estructurales serían proporcionales a los cambios en la secuencia. No obstante, la existencia de una misma secuencia en múltiples conformaciones define una "degeneración" de la información estructural codificada en una secuencia dada, introduciendo un comportamiento no lineal en el espacio proteico. Así, como vimos, esta no linealidad podría debilitar el desempeño de técnicas como TBM [Rackovsky, 2015].

Sin embargo, teniendo en cuenta la notable importancia de la diversidad conformacional para explicar diferentes procesos biológicos en proteínas, resulta casi imposible ignorar su presencia y su efecto en la estructura proteica. Como podemos derivar de la Figura 5.4, existe una gran dispersión de los valores de RMSD, incluso en valores altos de similitud secuencial, y cabe destacar que esta dispersión también se mantiene en las distintas familias (Figura 5.5). Estas figuras muestran la gran incerteza que existe al elegir el mejor molde, incluso a porcentajes de similitud secuencial altos. Además, hemos encontramos que la diversidad conformacional es proporcional a la divergencia estructural que alcanza una determinada familia (Figura 5.6) y es independiente a la divergencia secuencial de esa familia (Figura B.7).

En este punto es donde resulta interesante dividir nuestro set de datos en dos grupos de baja y alta diversidad conformacional para estudiar la relación estructura-secuencia. Observamos que en proteínas que evolucionan bajo una presión selectiva de mantener una DC baja, la correlación entre divergencia estructural y secuencial es alta (coeficiente de correlación de Spearman *rho* = -0.83). Previamente, hemos caracterizado esta clase de proteínas a las que denominamos rígidas y cuyas características se explicaron en el capítulo anterior. Para esta clase de proteínas sería posible una buena predicción de modelos estructurales usando técnicas de TBM (Figura 5.7A). Por el contrario, encontramos que proteínas con gran diversidad conformacional también van a exhibir gran divergencia estructural, y por lo tanto una mayor variedad de estructuras accesibles por la misma secuencia, incluso en porcentajes de identidad secuencial altos (Figura 5.7B). Encontramos que la correlación entre secuencia-estructura se ve disminuida en esta clase de proteínas (coeficiente de correlación de Spearman *rho* = -0.51), dificultando la idea general de que secuencias similares van a tener estructuras similares. Por ejemplo, usando la regresión lineal, al 100 % de identidad secuencial esperamos un valor de aproximadamente 1.3 $\mathring{A}$ de RMSD.

Nuestros resultados indican que el éxito de obtener un buen modelo estructural con las técnicas de TBM, esta íntimamente relacionado con la diversidad conformacional que posea la proteína que queremos modelar o *target*. Como es aún muy difícil predecir la CD de una proteína dada a partir de su secuencia, sugerimos que utilizar la definición de desorden y buscar regiones desordenadas en la secuencia, podría indicar que esa proteína tendrá mayor o menor CD. En este capítulo mostramos que cuando analizamos proteínas sin regiones desordenas (en todos sus confórmeros) y con regiones (en al menos uno de sus confórmeros), la correlación entre secuencia-estructura presenta valores más altos en aquellas proteínas ordenadas (coeficiente de correlación de Spearman *rho* = -0.71). Esto confirma nuevamente la alta correlación entre la divergencia estructural y secuencial en proteínas con baja CD. Gracias a la existencia de numerosos y efectivos métodos para la predicción del desorden a partir de una secuencia [He et al., 2009], su evaluación sería una buena manera de saber *a priori* qué dispersión de valores de RMSD esperaríamos para la selección de la estructura molde. Sin embargo, es necesario realizar estudios futuros y con datos experimentales para poder establecer cómo se conserva la diversidad conformacional a través de la evolución de una familia de proteínas.

En resumen, la relación entre la divergencia estructural y secuencial es un proceso más complejo de lo que se pensaba y se mostraba en trabajos previos. La diversidad conformacional proteica desafía el concepto que se ha establecido a partir de esta relación, el cual es un punto fundamental para el éxito de las técnicas de TBM. Es necesario un mayor trabajo para profundizar el conocimiento en muchas áreas asociadas con el estudio de las proteínas, teniendo en cuenta la característica dinámica de las mismas, así como también mejorar los métodos actuales para obtener y evaluar modelos 3D de proteínas.

5.5. Métodos específicos del capítulo

5.5.1. Selección de las familias de proteínas con diversidad conformacional

La selección de las proteínas con diversidad conformacional de la base de datos CoDNaS se realizó siguiendo el mismo procedimiento y utilizando los mismos criterios que los mencionados en el capítulo anterior, en la sección de Métodos 4.6.1. Con el objetivo de identificar familias de proteínas homólogas con diversidad conformacional, utilizamos el programa BLASTClust [Altschul et al., 1990] para obtener *clusters* de proteínas con una identidad local mínima de 30 % y una cobertura de 90 % entre todas las secuencias de ese clúster. De cada proteína con diversidad conformacional utilizamos la secuencia de referencia PDB SEQRES (de un confórmero representativo por cada proteína) para realizar el *clustering*. Además, nos aseguramos de que en cada clúster haya al menos dos proteínas distintas, es decir que posean diferente código UniProt. De esta manera, las familias van a estar compuestas por proteínas homólogas cercanas (al menos 30 % de identidad local). El set final de datos analizado en este capítulo está compuesto por 2024 proteínas con un total de 37,775 confórmeros (representando un 25 % de las proteínas incluidas en la última versión de CoDNaS). Estas proteínas están agrupadas en 524 familias con un promedio de cuatro proteínas por familia (un mínimo de 2 y un máximo de 61) y de acuerdo a la clasificación de la base de datos CATH, estas familias representan 250 plegamientos diferentes. La longitud promedio de las proteínas en las distintas familias es 283 residuos con una desviación estándar de 141, y el rango de identidad secuencial

de las familias está comprendido entre 20 y 98 %, con una mediana de alrededor del 47 %.

5.5.2. Estimación de la similitud secuencial y estructural

Los valores de RMSD para la diversidad conformacional de cada proteína se extrajeron directamente de la base de datos CoDNaS. La divergencia estructural se estimó para cada par de proteínas homólogas en cada familia. Calculamos el RMSD de carbonos alfa utilizando la herramienta MAMMOTH, entre todos los pares de confórmeros posibles, provenientes de dos proteínas homólogas. Identificamos el par de máxima divergencia estructural (MSD) que es el par de confórmeros, de dos proteínas homólogas, que maximiza la divergencia estructural.

La similitud secuencial se midió calculando el porcentaje de identidad entre las secuencias de dos proteínas homólogas. Para ello utilizamos el algoritmo de alineamiento global de Needleman-Wunch [Needleman and Wunsch, 1970], que a pesar de que es un método computacionalmente costoso (ya que analiza todos los alineamientos posibles para encontrar el de óptimo *score*) es altamente confiable.

El total de comparaciones entre todas las combinaciones posibles de confórmeros, provenientes de cada par de proteínas homólogas en las diferentes familias da un total de 3.5 x 10^6 alineamientos de a pares. Adicionalmente, para cada par de MSD y CD se calcularon dos parámetros adicionales como la fracción de estructura secundaria que no se conserva ($diffSS$) y el cambio entre enterrado/expuesto en la accesibilidad relativa al solvente ($diffRSA$). Para calcular la $diffSS$, en primer lugar se asignó la estructura secundaria utilizando la herramienta DSSP [Kabsch and Sander, 1983] a cada estructura del par MSD. Los 8 estados de estructura secundaria que define DSSP fueron agrupados en tres estados, alfa-hélices (H y G), láminas-beta (E) y *loops* (T, S, B, I, ' '), como se ha realizado en otros trabajos [Linding et al., 2003]. Para cada par de residuos alineados del par de MSD, comparamos los elementos de estructura secundaria definidos anteriormente con la ecuación 5.1:

$$Identidad_{SS} = \frac{(HH + SS + LL)}{HH + SS + LL + SH + HS + LH + HL + LS + SL)} \tag{5.1}$$

Donde HH, SS y LL son las cantidades de pares de residuos alineados con la misma estructura secundaria en ambas estructuras, es decir alfa-hélices (H), láminas-beta (S) y loops,

respectivamente. HS y SH es cuando se alinea un residuo que esta en alfa-hélice con uno que está en lámina-beta, y lo mismo aplica a las demás combinaciones de estructura secundaria. Finalmente, para estimar las diferencias locales entre elementos de estructura secundaria en el par de estructuras, calculamos la fracción de estructura secundaria que no se conserva como: $diffSS = 1 - Identidad_{SS}$.

Con el objetivo de calcular la $diffRSA$, en primer lugar calculamos la Accesibilidad relativa al solvente (RSA) para cada residuo de las estructuras del par. Para ello utilizamos un programa estándar para este cálculo como el Naccess 2.1.1 [Hubbard and Blundell, 1987] con un radio de prueba de 1.4 $\mathring{A}$. De acuerdo a los valores de RSA que varían de 0 a 100 para enterrado y expuesto respectivamente, definimos dos categorías para clasificar cada residuo: enterrado o *buried* (B) para valores de RSA $\leq$25 % y expuesto (E) para valores >25 %. Utilizando una metodología similar a la empleada con estructura secundaria, comparamos las categorías B y E en cada par de residuos alineados de las estructuras del par de MSD y obtuvimos $Identidad_{RSA}$. Luego, utilizamos la diferencia para calcular la fracción de cambio entre enterrado/expuesto en la accesibilidad relativa al solvente: $diffRSA = 1 - Identidad_{RSA}$.

En la sección donde se estudia la relación entre secuencia-estructura en proteínas ordenadas y desordenadas, utilizamos la definición de desorden explicada en el capítulo anterior en la sección de Métodos 4.6.2.

5.5.3. Análisis estadísticos

Los coeficientes de correlación mostrados en este capítulo se calcularon utilizando la función *cor.test* del paquete R [Team, 2017], con su correspondiente prueba de hipótesis de dos colas (la hipótesis nula es que el coeficiente de correlación es igual a 0). Mayormente se utilizaron correlaciones de Spearman ya que no asume linealidad en los datos.

La validación cruzada de los coeficientes de determinación se realizó utilizando el lenguaje Julia, con las librerías *LqsFit* y *MLBase*. Todas las regresiones fueron pesadas, y cada punto (MSD versus porcentaje de identidad) tiene un peso de 1 sobre el número de pares de proteínas homólogas de la familia correspondiente. Esto se realizó con el fin de evitar que familias muy pobladas predominen en los resultados y puedan sesgar los valores obtenidos. El R^2 informado

fue calculado como la media de los R^2 que se obtuvieron del testeo de cada subgrupo en una validación cruzada de 5 repeticiones.

Capítulo 6

Conclusiones Generales

Las proteínas son las macromoléculas más diversas que podemos encontrar y que adquieren una estructura tridimensional la cual está íntimamente relacionada con la función que cumplen. A pesar de que la estructura proteica se encuentra codificada en forma lineal por los 20 aminoácidos conocidos, éstas moléculas presentan una gran diversidad estructural y funcional, participando y mediando gran cantidad de procesos biológicos y fisicoquímicos dentro de la célula. El estudio de la relación estructura-función de proteínas ha ocupado un rol muy importante dentro de la biología estructural, desde el modelo de llave-cerradura propuesto por Emil Fischer en 1894 pasando por el de Monod en 1965, han pasado más de cien años hasta la actualidad [Fischer, 1894, Koshland et al., 1958, Monod et al., 1965]. Este concepto fue evolucionando hasta lo que conocemos hoy en día; las proteínas no poseen una estructura única en su estado nativo sino que este es un ensamble nativo compuesto por confórmeros que se encuentran en equilibrio dinámico, agregándole aún más dificultad a la relación estructura-dinámica-función [James and Tawfik, 2003]. La estructura de una proteína no es estática sino que necesita de diferentes grados de movilidad para poder realizar su función biológica. Entender la dinámica de la estructura proteica es de vital importancia para poder comprender la relación estructura-dinámica-función y por consecuencia, poder mejorar los métodos y algoritmos actuales que mayormente utilizan una estructura estática para realizar sus cálculos y/o predicciones.

Durante estos 6 años en el grupo de investigación de Bioinformática Estructural en la Universidad Nacional de Quilmes, hemos desarrollado herramientas computacionales y nume-

rosos trabajos de investigación básica para poder comprender más en profundidad la relación estructura-dinámica-función de las proteínas. El diseño, desarrollo e implementación de CoDNaS (*"Conformational Diversity of the Native State"*) [Monzon et al., 2013, Monzon et al., 2016] presentado en este trabajo de tesis fue uno de los principales motores que dieron continuidad a esta línea de investigación. CoDNaS posee características únicas que la hacen una herramienta innovadora y de gran utilidad para aquellos interesados en estudiar la diversidad conformacional de proteínas. El proceso de curado y filtrado de la información fue uno de los más extensos dentro del proyecto, donde el objetivo principal fue generar datos confiables y de buena calidad para los posteriores análisis que se iban a realizar. Se tuvieron en cuenta muchas variables inherentes a la técnica experimental por la que se obtiene la estructura tridimensional de la proteína, de manera tal de detectar posibles artefactos dentro de la misma que puedan estar causando perdida en la señal biológica de los parámetros estructurales calculados. Además, se asociaron los datos estructurales de cada una de las proteínas con diferentes bases de datos biológicas para el estudio de la diversidad conformacional. CoDNaS permitió ir más allá del estado del arte en el campo de la diversidad conformacional de proteínas permitiéndonos abordar diferentes preguntas biológicas acerca de la relación estructura-dinámica-función de las mismas. CoDNaS ha sido el hito más importante de esta tesis, posibilitando el desarrollo de diferentes trabajos en el área de bioinformática estructural y biofísica, además de ser referente entre las bases de datos de este estilo que hay hoy disponibles. Consecuentemente, CoDNaS propuso nuevos desafíos tanto personales, como para el grupo de investigación, donde nuevas líneas de estudio fueron surgiendo a partir del análisis de la base de datos. Dentro de los trabajos de importancia, podemos mencionar: el estudio de la relación entre la velocidad de evolución y la diversidad conformacional [Zea et al., 2013], identificación de posiciones de importancia que definen la diversidad conformacional proteica [Saldaño et al., 2016], estudio de la relación entre las transiciones orden/desorden y la diversidad conformacional [Zea et al., 2016], evaluación de los métodos de TBM en múltiples conformaciones [Palopoli et al., 2013], entre otros trabajos que fueron abarcados directamente en esta tesis. Desde su primer versión publicada en 2013, la base de datos se ha ido actualizando y mejorando logrando una segunda versión en 2016. En el transcurso de este tiempo la base de datos se ha posicionado dentro de las herramientas de su estilo, adquiriendo visibilidad a nivel internacional y siendo utilizada

en una gran variedad de trabajos del área de bioinformática y biología estructural. Finalmente en año 2017, CoDNaS ha sido relacionada con uno de los principales repositorios de proteínas desordenadas como lo es MobiDB 3.0 [Piovesan et al., 2018]. Esperamos que esto incremente aún más la visibilidad y utilización de la base de datos por parte de la comunidad científica.

Mediante el uso de nuestra base de datos hemos podido avanzar en el estudio de la diversidad conformacional proteica, pudiendo entender y explicar la distribución global de diversidad conformacional que había sido presentada pero no estudiada en profundidad en el trabajo de *Burra et al.* [Burra et al., 2009, Monzon et al., 2017a]. En el capítulo 4 mostramos que hemos encontrado al menos cuatro mecanismos o relaciones estructura-función expresados como relaciones estructura-dinámica que representan diferentes formas de lograr una gran diversidad de funciones biológicas. Estos grupos de proteínas emergen del análisis de la características de sus ensambles conformacionales. Encontramos que mayormente tenemos proteínas que denominamos rígidas, las cuales que no requieren grandes movimientos apreciables en el *backbone* para cumplir su función, pero sin embargo realizan pequeños movimientos a nivel de las cadenas laterales de sus aminoácidos que le permiten la apertura y cierre de túneles y cavidades [Pravda et al., 2014b]. Además, en orden de menor a mayor diversidad conformacional, encontramos proteínas que tienen al menos un confórmero con IDRs (desordenadas) que las clasificamos en proteínas parcialmente desordenadas y maleables. Estos dos grupos además de tener diferencias estructurales y funcionales que fueron descritas en el capitulo 4, tienen IDRs con diferente propensidad a estar desordenadas en el ensamble conformacional. Creemos que este trabajo resulta de gran importancia para entender las relaciones estructura-función teniendo en cuenta la variable dinámica de las proteínas, ya que hoy en día la mayoría de los métodos computacionales en el área de la bioinformática estructural, solo utilizan una estructura como representativa del estado nativo, perdiendo muchas veces la sensibilidad necesaria para profundizar nuestro conocimiento de los mecanismos que subyacen a la biología de las proteínas.

Utilizando el concepto de diversidad conformacional, nos propusimos re-examinar la relación bien conocida y establecida desde hace muchos años, como lo es la relación entre divergencia estructural y secuencial, que data del año 1984 estudiada por primera vez por Lesk y Chothia [Lesk and Chothia, 1984]. La importancia de esta relación subyace en que la estruc-

tura proteica se conserva mucho más que la secuencia a lo largo de la evolución y en que sentó las bases para los métodos de modelado por homología que se desarrollaron posteriormente. Analizando familias de proteínas con diversidad conformacional extraídas de CoDNaS, encontramos que esta relación es más compleja de lo que se pensaba y que al tener en cuenta la diversidad conformacional, podemos tener grandes valores de divergencia estructural a niveles altos de identidad de secuencia. Esto ocurre principalmente por la presencia de la diversidad conformacional, esto es, la variabilidad estructural de una misma secuencia. Este trabajo remarca la importancia de tener en cuenta la diversidad conformacional de la proteína a la hora de obtener modelos estructurales por homología, donde éstas técnicas van a ser muy poco eficientes en proteínas con mucha movilidad.

Consideramos que el concepto de ensamble conformacional está subrepresentado en todas las técnicas y métodos computacionales que estudian la estructura y función de las proteínas. Como hemos mostrado a lo largo de este trabajo de tesis, poder explicar la función de las proteínas a partir de su ensamble conformacional y las propiedades del mismo, resulta de gran importancia para poder abordar a conclusiones biológicas más precisas. El ensamble conformacional proporciona descriptores globales que van más allá del tipo de plegado, estructura secundaria, secuencia o familias/superfamilias de proteínas homólogas. Conociendo los ensambles proteicos podríamos en un futuro agrupar y caracterizar la función proteica por las características dinámicas que poseen, la cuales son más generales que cualquier clasificación ya conocida. En los últimos 20 años, el paradigma de estructura función ha sido desafiado por la presencia de regiones o proteínas que carecen de una estructura tridimensional definida en condiciones fisiológicas (las IDPs o IDRs). Estas proteínas participan en una gran cantidad de procesos biológicos, mediando diversas funcionas dentro de la célula. El concepto se ensamble conformacional toma aún mayor fuerza en estas proteínas que poseen una dinámica conformacional extrema, la que no sólo dificulta su estudio por métodos computacionales, sino también experimentales. Brevemente, creemos que existe un continuo entre proteínas ordenadas como la hemoglobina que requieren unos pocos confórmeros para describir su estado nativo, y las IDPs que requieren decenas o incluso cientos de ellos.

Consideramos que en los próximos años la biología estructural evolucionará y permitirá la obtención experimental de los ensambles nativos con la ayuda de nuevas técnicas para

resolución de estructuras de alta calidad, como lo es la técnica de Cryo-EM. Las caracterización de ensambles conformacionales de proteínas ordenadas y desordenadas hoy en día es un desafío debido a la complejidad que poseen estos ensambles para ser analizados y poder cuantificar sus propiedades estructurales. Información clave como las distribuciones de las poblaciones de confórmeros, las propensiones de la estructura secundaria, los contactos nativos, las energías relativas y los estados de población conformacional, son difíciles de obtener para ensambles de estructuras y más aún de IDPs. Estos ensambles requerirán el desarrollo de nuevas técnicas para describir con precisión los movimientos de grandes amplitudes y escalas de tiempo largas, típicas de los ensambles conformacionales de IDPs, reemplazando la dicotomía entre orden y desorden de proteínas por una descripción cuantitativa y continua entre estos dos extremos. Creemos que el estudio de los ensambles conformacionales será un aspecto clave para poder avanzar en el entendimiento de la función proteica y su dinámica conformacional.

6.1. Perspectivas a futuro

A futuro y en continuación con la línea de trabajo presentada en esta tesis, estamos estudiando cómo la dinámica de la proteína se conserva en la evolución, es decir, poder entender si la diversidad conformacional se conserva o no en proteínas homólogas. Por otro lado, tenemos dos proyectos a futuro en base a continuar con el desarrollo de CoDNaS. Uno, es el desarrollo de la versión número tres de CoDNaS, donde se incluirían cálculos de túneles y cavidades para cada confórmero, se mejoraría la usabilidad e interfaz de la base de datos, entre otras mejoras. El otro proyecto consiste en desarrollar una base de datos de diversidad conformacional de movimientos cuaternarios, es decir aquellos que se producen en los oligómeros entre sus subunidades. Creemos que el estudio de este tipo de movimientos será de gran importancia para poder continuar con el estudio de la diversidad conformacional y su relación con la función proteica.

Apéndice

Apéndice A

Tablas suplementarias

Variables	ORDERED			DISORDERED					
	Rigid			P. Disordered			Malleable		
	Mean ± CI 95%	Std. Dev	Median	Mean ± CI 95%	Std. Dev	Median	Mean ± CI 95%	Std. Dev	Median
Amount of conformers By each protein	14.39 ± 0.81	22.89	8	14 ± 1.15	20.19	9	25.76 ± 4	46.58	12
Amount of conformers with IDRs By each protein	-	-	-	9.09 ± 0.73	12.93	6	5.58 ± 1.34	15.55	2
Maximum RMSD [Å]	0.87 ± 0.02	0.61	0.68	1.14 ± 0.04	0.64	1.02	1.35 ± 0.05	0.62	1.24
Conformer's clusters	1.69 ± 0.05	1.47	1	2.24 ± 0.10	1.73	2	3.10 ± 0.29	3.37	2
Conformers with IDRs By each protein [%]	-	-	-	69.10 ± 1.75	30.87	75	24.87 ± 1.59	18.49	20
Maximum disorder per protein [%]	-	-	-	6.95 ± 0.39	6.93	5.05	4.64 ± 0.37	4.26	3.47
Maximum disorder region lengths	-	-	-	17.76 ± 1.14	20.11	12	10.63 ± 0.85	9.92	8
Amount of hinges	0.21 ± 0.03	0.72	0	1.35 ± 0.07	1.31	1	0.52 + 0.08	0.95	0
Maximum pocket volume [Å]	1287.74 ± 32.05	866.33	1063.19	1469.53 ± 50.04	866.89	1277.34	431.84 ± 83.7	964.48	1167.3
Max. pocket volume variation[1] (with/without IDRs)	-	-	-	11.8 ± 0.84	14.56	6.031	20.45 ± 1.62	18.47	15.08
Maximum tunnel length variation[2]	0.49 ± 0.01	0.36	0.4	0.41 ± 0.02	0.32	0.35	0.44 ± 0.03	0.33	0.35
Maximum RMSD [Å] (loops/coils)	0.93 ± 0.02	0.63	0.76	1.19 ± 0.03	0.6	1.12	1.32 ± 0.05	0.57	1.23
Maximum RMSD [Å] (alpha-helix)	0.51 ± 0.02	0.48	0.36	0.71 ± 0.03	0.57	0.5	0.75 ± 0.05	0.62	0.53
Maximum RMSD [Å] (beta-strand)	0.39 ± 0.02	0.4	0.26	0.49 ± 0.03	0.48	0.33	0.48 ± 0.04	0.47	0.32
Degree in the residue interaction network	5 ± 0.03	0.94	5.13	5.16 ± 0.04	0.62	5.28	5.21 ± 0.05	0.62	5.25
Normalized radius of gyration	1.21 ± 0.01	0.26	1.14	1.22 ± 0.02	0.21	1.17	1.19 ± 0.01	0.23	1.14
Total proteins	3075			1194			522		

Tabla A.1: Resumen de diferentes descriptores analizados en el capítulo 4. Los valores corresponden a la media (con el intervalo de confianza al 95 %), la desviación estándar y la mediana de cada distribución comparada. Vea la sección de Métodos 4.6 para detalles sobre los descriptores.

Structural measures	Average CD [Å]			
	(0,0.5]	*(0.5,1]*	*(1,1.5]*	*(1.5,4]*
RMSD	-0.83	-0.78	-0.55	-0.32
Fraction of unconserved SS	-0.75	-0.76	-0.74	-0.71
Fraction of unconserved RSA	-0.86	-0.87	-0.81	-0.80

Tabla A.2: Coeficientes de correlación de Spearman *rho* entre MSD y CD, calculados para diferentes intervalos de diversidad conformacional.

Apéndice B

Figuras suplementarias

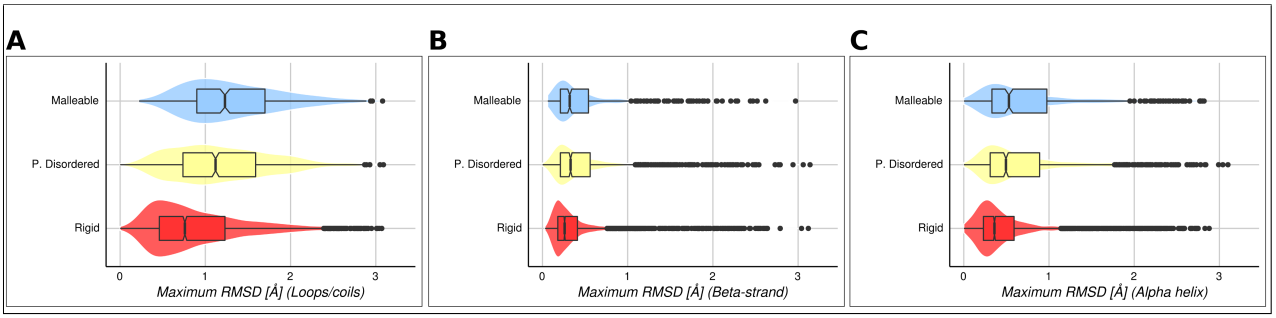

Figura B.1: (A-C) Distribuciones de los valores de RMSD en los pares de máxima diversidad conformacional en cada grupo discriminados por los elementos de estructura secundaria.

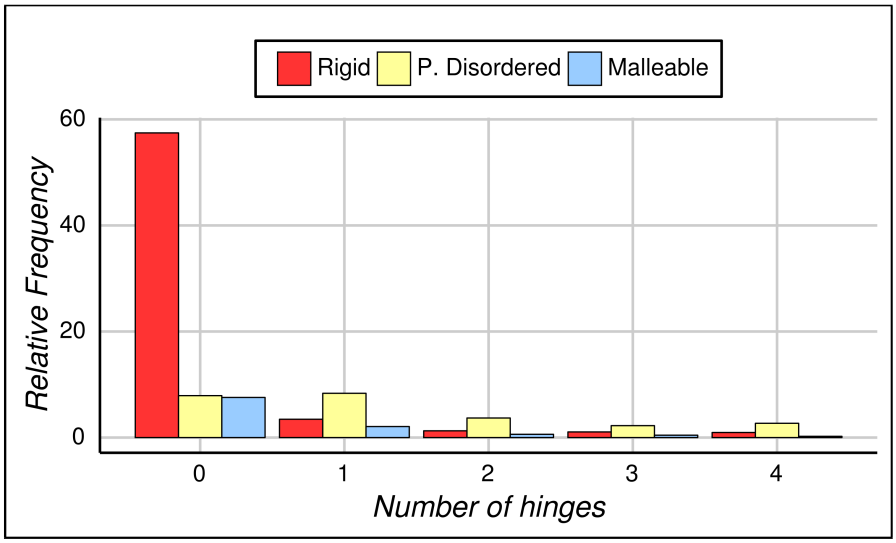

Figura B.2: Distribución de la cantidad de *hinges* en los pares de máxima diversidad conformacional de cada grupo. Cada barra representa la frecuencia relativa del número de *hinges* en cada set.

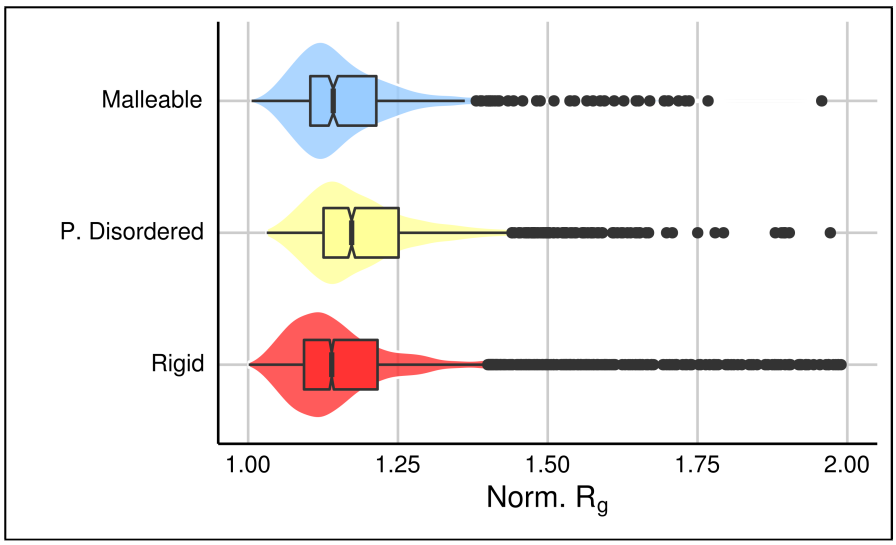

Figura B.3: Distribuciones del radio de giro normalizado promedio de los pares de confórmeros de máximo RMSD en cada grupo de proteínas.

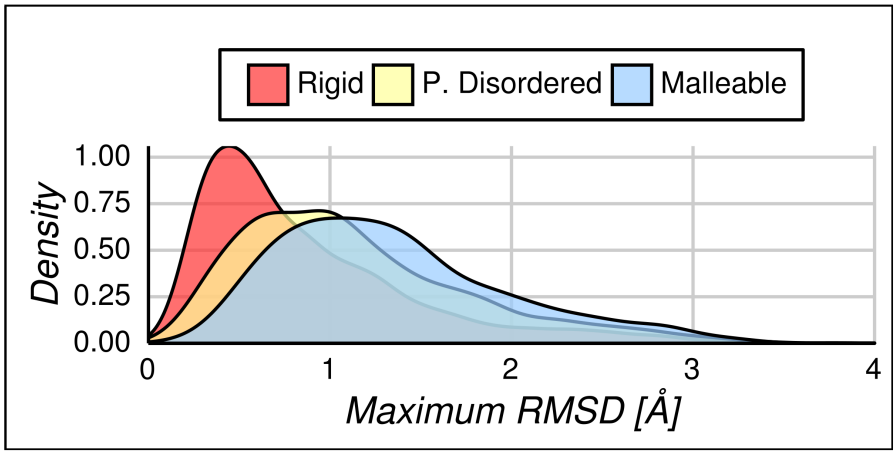

Figura B.4: Distribuciones de la diversidad conformacional máxima en el subset de pares *apo*/*holo*.

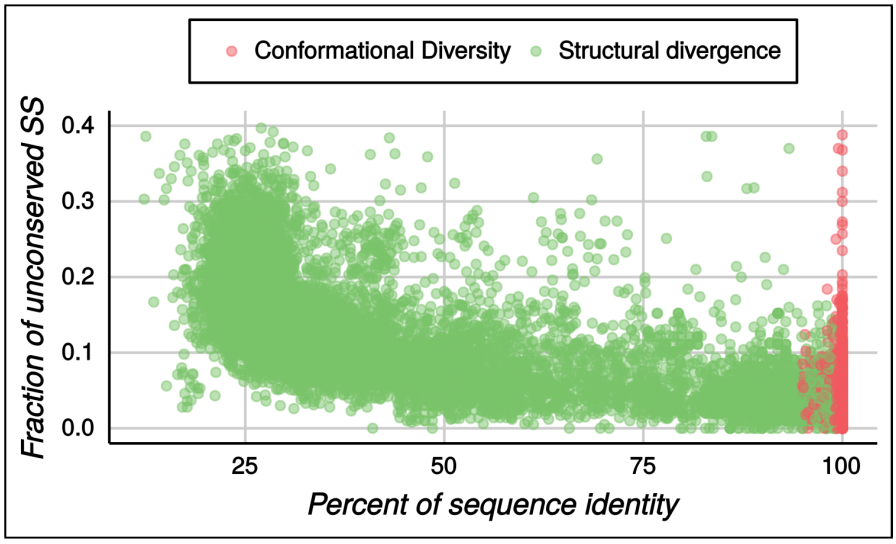

Figura B.5: Fracción de estructura secundaria no conservada versus porcentaje de identidad de secuencia. El eje Y hace referencia a la fracción de residuos que cambian su estructura secundaria en un par de proteínas homólogas (MSD), o entre la misma proteína (CD).

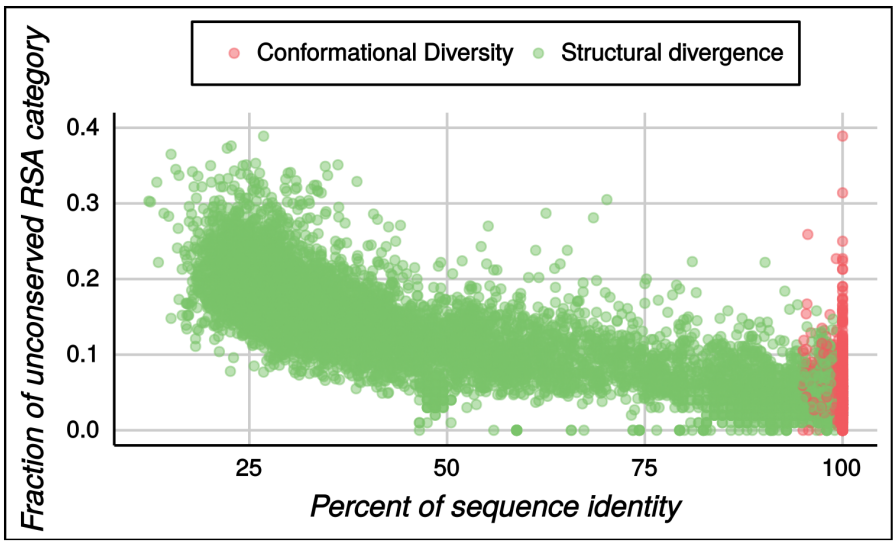

Figura B.6: Fracción de categoría de RSA (expuesto/enterrado) no conservada versus porcentaje de identidad. El eje Y hace referencia a la fracción de residuos que cambian entre expuestos/enterrados en un par de proteínas homólogas (MSD), o entre la misma proteína (CD).

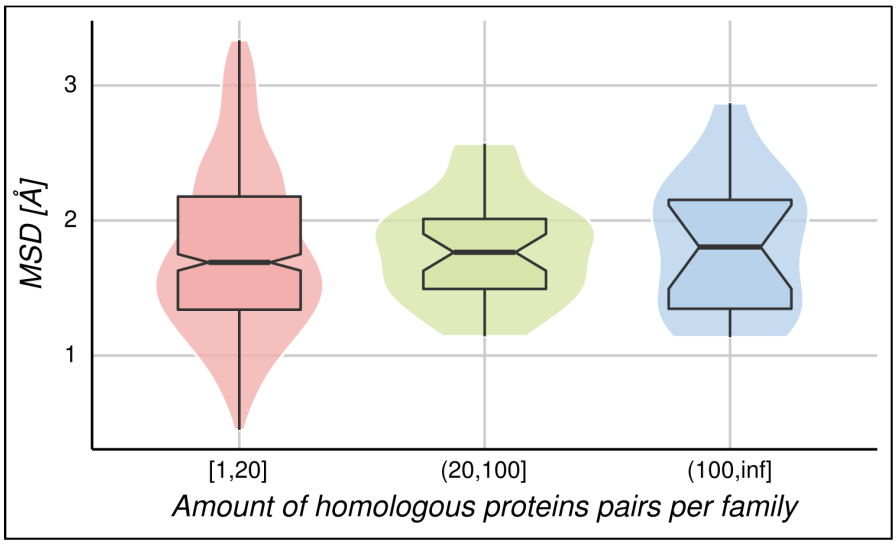

Figura B.7: Distribuciones de MSD en intervalos de cantidad de pares de proteínas homólogas en cada familia.

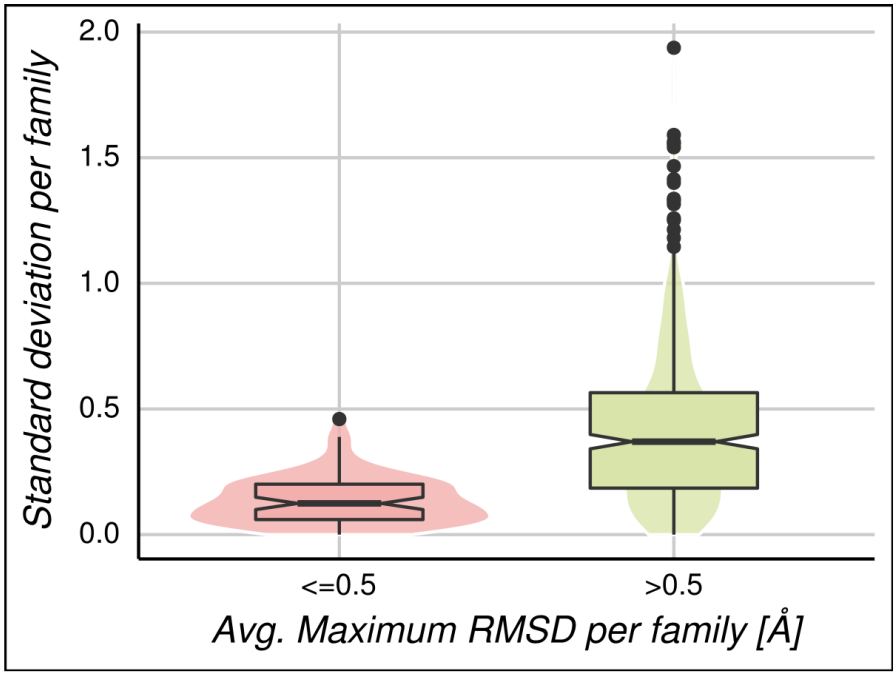

Figura B.8: Distribuciones de la desviación estándar del RMSD de diversidad conformacional por familia. Por cada familia de calculo la desviación estándar de los valores de diversidad conformacional de las proteínas que la componen. Se graficaron agrupando por familias con baja diversidad conformacional (un RMSD promedio menor a 0.5 $\mathring{A}$ entre todas sus proteínas) y alta diversidad conformacional (mayor a 0.5 $\mathring{A}$).

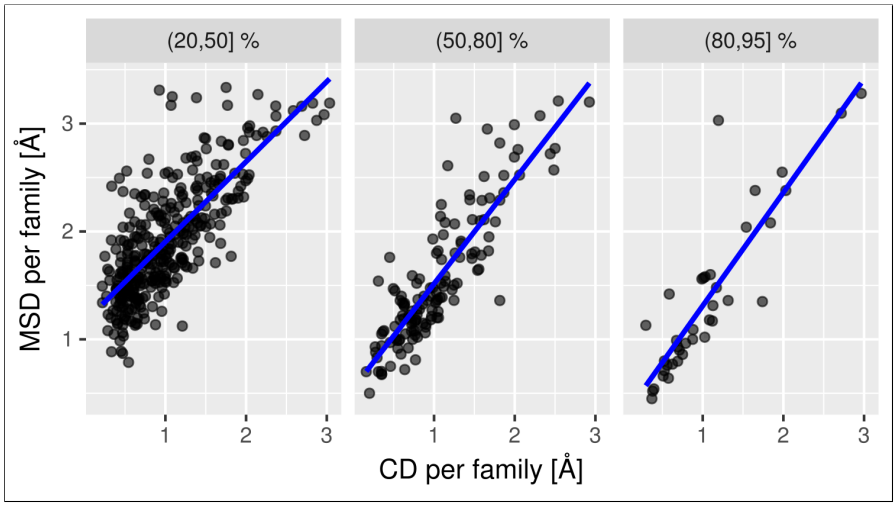

Figura B.9: Relación entre la MSD y la DC en intervalos de porcentaje de identidad de secuencia. Cada punto representa el RMSD promedio para la MSD y la DC de los pares de proteínas homólogas de una familia en particular, en un intervalo de identidad secuencial. El número de familias en cada intervalo son 348, 138 y 38, con unos coeficientes de correlación de Pearson de 0.77, 0.87 y 0.88, respectivamente.

Bibliografía

Abagyan and Batalov, 1997. Abagyan, R. a. and Batalov, S. (1997). Do aligned sequences share the same fold? *Journal of molecular biology*, 273(1):355–368.

Altschul et al., 1990. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–10.

Amemiya et al., 2012. Amemiya, T., Koike, R., Kidera, A., and Ota, M. (2012). PSCDB: a database for protein structural change upon ligand binding. *Nucleic acids research*, 40(Database issue):D554–8.

Anand et al., 2011. Anand, P., Sankaran, S., Mukherjee, S., Yeturu, K., Laskowski, R., Bhardwaj, A., Bhagavat, R., Brahmachari, S. K., Chandra, N., and Chandra, N. (2011). Structural Annotation of Mycobacterium tuberculosis Proteome. *PLoS ONE*, 6(10):e27044.

Ashburner et al., 2000. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1):25–9.

Atilgan et al., 2001. Atilgan, A. R., Durell, S. R., Jernigan, R. L., Demirel, M. C., Keskin, O., and Bahar, I. (2001). Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical journal*, 80(1):505–15.

Bahar and Rader, 2005. Bahar, I. and Rader, A. J. (2005). Coarse-grained normal mode analysis in structural biology. *Current opinion in structural biology*, 15(5):586–92.

Baker and Sali, 2001. Baker, D. and Sali, A. (2001). Protein Structure Prediction and Structural Genomics. *Science*, 294(5540):93–96.

Berka et al., 2012. Berka, K., Hanak, O., Sehnal, D., Banas, P., Navratilova, V., Jaiswal, D., Ionescu, C.-M., Svobodova Varekova, R., Koca, J., and Otyepka, M. (2012). MOLEonline 2.0: interactive web-based analysis of biomacromolecular channels. *Nucleic Acids Research*, 40(W1):W222–W227.

Berman, 2000. Berman, H. M. (2000). The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242.

Best et al., 2006. Best, R. B., Lindorff-Larsen, K., DePristo, M. A., and Vendruscolo, M. (2006). Relation between native ensembles and experimental structures of proteins. *Proceedings of the National Academy of Sciences*, 103(29):10901–10906.

Biedermannová et al., 2012. Biedermannová, L., Prokop, Z., Gora, A., Chovancová, E., Kovács, M., Damborský, J., and Wade, R. C. (2012). A Single Mutation in a Tunnel to the Active Site Changes the Mechanism and Kinetics of Product Release in Haloalkane Dehalogenase LinB. *Journal of Biological Chemistry*, 287(34):29062–29074.

Boehr et al., 2006. Boehr, D. D., McElheny, D., Dyson, H. J., and Wright, P. E. (2006). The dynamic energy landscape of dihydrofolate reductase catalysis. *Science (New York, N.Y.)*, 313(5793):1638–42.

Boehr et al., 2009a. Boehr, D. D., Nussinov, R., and Wright, P. E. (2009a). The role of dynamic conformatinal ensembles in biomolecular recognition. *Nat Chem Biol.*, 5(11):789–796.

Boehr et al., 2009b. Boehr, D. D., Nussinov, R., and Wright, P. E. (2009b). The role of dynamic conformational ensembles in biomolecular recognition. *Nature chemical biology*, 5(11):789–96.

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

Bryngelson and Wolynes, 1989. Bryngelson, J. D. and Wolynes, P. G. (1989). Intermediates and Barrier Crossing in a Random Energy Model (with Applications to Protein Folding). *Journal of Physical Chemical*, 93(10):6902–6915.

Burra et al., 2009. Burra, P. V., Zhang, Y., Godzik, A., and Stec, B. (2009). Global distribution of conformational states derived from redundant models in the PDB points to non-uniqueness of the protein structure. *Proceedings of the National Academy of Sciences of the United States of America*, 106(26):10505–10.

Case and Karplus, 1979. Case, D. and Karplus, M. (1979). Dynamics of ligand binding to heme proteins. *Journal of Molecular Biology*, 132(3):343–368.

Changeux, 2011. Changeux, J. P. (2011). 50th anniversary of the word "allosteric".

Chiti and Dobson, 2006. Chiti, F. and Dobson, C. M. (2006). Protein Misfolding, Functional Amyloid, and Human Disease. *Annual Review of Biochemistry*, 75(1):333–366.

Chng and Yang, 2008. Chng, C.-P. and Yang, L.-W. (2008). Coarse-grained models reveal functional dynamics–II. Molecular dynamics simulation at the coarse-grained level–theories and biological applications. *Bioinformatics and biology insights*, 2:171–85.

Chothia and Lesk, 1986. Chothia, C. and Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *The EMBO journal*, 5(4):823–6.

Chovancova et al., 2012. Chovancova, E., Pavelka, A., Benes, P., Strnad, O., Brezovsky, J., Kozlikova, B., Gora, A., Sustr, V., Klvana, M., Medek, P., Biedermannova, L., Sochor, J., and Damborsky, J. (2012). CAVER 3.0: A Tool for the Analysis of Transport Pathways in Dynamic Protein Structures. *PLoS Computational Biology*, 8(10):23–30.

Choy and Forman-Kay, 2001. Choy, W. Y. and Forman-Kay, J. D. (2001). Calculation of ensembles of structures representing the unfolded state of an SH3 domain. *Journal of molecular biology*, 308(5):1011–32.

Consortium, 2017. Consortium, U. (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 45(D1):D158–D169.

Cooper and Dryden, 1984. Cooper, A. and Dryden, D. T. F. (1984). Allostery without conformational change. *European Biophysics Journal*, 11(2):103–109.

Cordes et al., 2000. Cordes, M. H., Burton, R. E., Walsh, N. P., McKnight, C. J., and Sauer, R. T. (2000). An evolutionary bridge to a new protein fold. *Nature structural biology*, 7(12):1129–32.

Cozzetto et al., 2009. Cozzetto, D., Kryshtafovych, A., Fidelis, K., Moult, J., Rost, B., and Tramontano, A. (2009). Evaluation of template-based models in CASP8 with standard measures. *Proteins: Structure, Function and Bioinformatics*, 77(SUPPL. 9):18–28.

Csermely et al., 2010. Csermely, P., Palotai, R., and Nussinov, R. (2010). Induced fit, conformational selection and independent dynamic segments: An extended view of binding events. *Trends in Biochemical Sciences*, 35(10):539–546.

Cui and Karplus, 2008. Cui, Q. and Karplus, M. (2008). Allostery and cooperativity revisited. *Protein science : a publication of the Protein Society*, 17(8):1295–1307.

Daily and Gray, 2007. Daily, M. D. and Gray, J. J. (2007). Local motions in a benchmark of allosteric proteins. *Proteins: Structure, Function and Genetics*, 67(2):385–399.

Dean and Thornton, 2007. Dean, A. M. and Thornton, J. W. (2007). Mechanistic approaches to the study of evolution: the functional synthesis. *Nature reviews. Genetics*, 8(9):675–88.

DeForte and Uversky, 2016. DeForte, S. and Uversky, V. (2016). Order, Disorder, and Everything in Between. *Molecules*, 21(8):1090.

del Sol et al., 2009. del Sol, A., Tsai, C.-J., Ma, B., and Nussinov, R. (2009). The Origin of Allosteric Functional Modulation: Multiple Pre-existing Pathways. *Structure*, 17(8):1042–1050.

Doncheva et al., 2012. Doncheva, N. T., Assenov, Y., Domingues, F. S., and Albrecht, M. (2012). Topological analysis and interactive visualization of biological networks and protein structures. *Nature Protocols*, 7(4):670–685.

Drew et al., 2011. Drew, K., Winters, P., Butterfoss, G. L., Berstis, V., Uplinger, K., Armstrong, J., Riffle, M., Schweighofer, E., Bovermann, B., Goodlett, D. R., Davis, T. N., Shasha, D., Malmström, L., and Bonneau, R. (2011). The Proteome Folding Project: Proteome-scale prediction of structure and function. *Genome Research*, 21(11):1981–1994.

Dunker et al., 2001. 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., and Obradovic, Z. (2001). Intrinsically disordered protein. *Journal of molecular graphics & modelling*, 19(1):26–59.

Dunker et al., 2008. Dunker, A. K., Silman, I., Uversky, V. N., and Sussman, J. L. (2008). Function and structure of inherently disordered proteins.

Dyson and Wright, 2005. Dyson, H. J. and Wright, P. E. (2005). Intrinsically unstructured proteins and their functions. *Nature Reviews Molecular Cell Biology*, 6(3):197–208.

Eisenmesser et al., 2005. Eisenmesser, E. Z., Millet, O., Labeikovsky, W., Korzhnev, D. M., Wolf-Watz, M., Bosco, D. A., Skalicky, J. J., Kay, L. E., and Kern, D. (2005). Intrinsic dynamics of an enzyme underlies catalysis. *Nature*, 438(7064):117–121.

Elias and Tawfik, 2012. Elias, M. and Tawfik, D. S. (2012). Divergence and Convergence in Enzyme Evolution: Parallel Evolution of Paraoxonases from Quorum-quenching Lactonases. *Journal of Biological Chemistry*, 287(1):11–20.

Eyal et al., 2005. Eyal, E., Gerzon, S., Potapov, V., Edelman, M., and Sobolev, V. (2005). The limit of accuracy of protein modeling: Influence of crystal packing on protein structure. *Journal of Molecular Biology*, 351(2):431–442.

Fersht and Requena, 1971. Fersht, A. R. and Requena, Y. (1971). Equilibrium and rate constants for the interconversion of two conformations of α-chymotrypsin. *Journal of Molecular Biology*, 60(2):279–290.

Fischer, 1894. Fischer, E. (1894). Einfluss der Configuration auf die Wirkung der Enzyme. *Berichte der deutschen chemischen Gesellschaft*, 27(3):2985–2993.

Fiser, 2010. Fiser, A. (2010). Template-Based Protein Structure Modeling. In *Methods in molecular biology (Clifton, N.J.)*, volume 673, pages 73–94.

Flock et al., 2014. Flock, T., Weatheritt, R. J., Latysheva, N. S., and Babu, M. M. (2014). Controlling entropy to tune the functions of intrinsically disordered regions. *Current Opinion in Structural Biology*, 26(1):62–72.

Flores et al., 2007. Flores, S. C., Lu, L. J., Yang, J., Carriero, N., and Gerstein, M. B. (2007). Hinge Atlas: relating protein sequence to sites of structural flexibility. *BMC Bioinformatics*, 8(1):167.

Foote and Milstein, 1994. Foote, J. and Milstein, C. (1994). Conformational isomerism and the diversity of antibodies. *Proceedings of the National Academy of Sciences of the United States of America*, 91(22):10370–4.

Frauenfelder and McMahon, 1998. Frauenfelder, H. and McMahon, B. (1998). Dynamics and function of proteins: the search for general concepts. *Proceedings of the National Academy of Sciences of the United States of America*, 95(9):4795–7.

Frauenfelder et al., 1991. Frauenfelder, H., Sligar, S. G., and Wolynes, P. G. (1991). The energy landscapes and motions of proteins. *Science (New York, N.Y.)*, 254(5038):1598–603.

Gardino et al., 2009. Gardino, A. K., Villali, J., Kivenson, A., Lei, M., Liu, C. F., Steindel, P., Eisenmesser, E. Z., Labeikovsky, W., Wolf-Watz, M., Clarkson, M. W., and Kern, D. (2009). Transient non-native hydrogen bonds promote activation of a signaling protein. *Cell*, 139(6):1109–18.

Gerhart and Pardee, 1962. Gerhart, J. C. and Pardee, A. B. (1962). The enzymology of control by feedback inhibition. *The Journal of biological chemistry*, 237:891–6.

Gerstein and Krebs, 1998. Gerstein, M. and Krebs, W. (1998). A database of macromolecular motions. *Nucleic acids research*, 26(18):4280–90.

Gerstein et al., 1994. Gerstein, M., Lesk, A. M., and Chothia, C. (1994). Structural Mechanisms for Domain Movements in Proteins. *Biochemistry*, 33(22):6739–6749.

Gora et al., 2013. Gora, A., Brezovsky, J., and Damborsky, J. (2013). Gates of enzymes. *Chemical Reviews*, 113(8):5871–5923.

Grant et al., 2006. Grant, B. J., Rodrigues, A. P. C., ElSawy, K. M., McCammon, J. A., and Caves, L. S. D. (2006). Bio3d: An R package for the comparative analysis of protein structures. *Bioinformatics*, 22(21):2695–2696.

Gu et al., 2015. Gu, Y., Li, D.-W., and Brüschweiler, R. (2015). Decoding the Mobility and Time Scales of Protein Loops. *Journal of Chemical Theory and Computation*, 11(3):1308–1314.

Gunasekaran et al., 2004. Gunasekaran, K., Ma, B., and Nussinov, R. (2004). Is allostery an intrinsic property of all dynamic proteins? *Proteins*, 57(3):433–43.

Gutteridge and Thornton, 2005. Gutteridge, A. and Thornton, J. (2005). Conformational changes observed in enzyme crystal structures upon substrate binding. *Journal of Molecular Biology*, 346(1):21–28.

Hanson et al., 2013. Hanson, R. M., Prilusky, J., Renjian, Z., Nakane, T., and Sussman, J. L. (2013). JSmol and the Next-Generation Web-Based Representation of 3D Molecular Structure as Applied to Proteopedia. *Israel Journal of Chemistry*, 53(3-4):207–216.

He et al., 2009. He, B., Wang, K., Liu, Y., Xue, B., Uversky, V. N., and Dunker, A. K. (2009). Predicting intrinsic disorder in proteins: An overview.

Henzler-Wildman and Kern, 2007. Henzler-Wildman, K. and Kern, D. (2007). Dynamic personalities of proteins. *Nature*, 450(7172):964–972.

Hobohm and Sander, 2008. Hobohm, U. and Sander, C. (2008). Enlarged representative set of protein structures. *Protein Science*, 3(3):522–524.

Hollfelder et al., 1996. Hollfelder, F., Kirby, A. J., and Tawfik, D. S. (1996). Off-the-shelf proteins that rival tailor-made antibodies as catalysts. *Nature*, 383(6595):60–2.

Hrabe et al., 2015. Hrabe, T., Li, Z., Sedova, M., Rotkiewicz, P., Jaroszewski, L., and Godzik, A. (2015). PDBFlex: exploring flexibility in protein structures. *Nucleic acids research*, 44(D1):D423–428.

Hubbard and Blundell, 1987. Hubbard, T. J. and Blundell, T. L. (1987). Comparison of solvent-inaccessible cores of homologous proteins: Definitions useful for protein modelling. *Protein Engineering, Design and Selection*, 1(3):159–171.

Ikura and Ames, 2006. Ikura, M. and Ames, J. B. (2006). Genetic polymorphism and protein conformational plasticity in the calmodulin superfamily: Two ways to promote multifunctionality. *Proceedings of the National Academy of Sciences*, 103(5):1159–1164.

Illergård et al., 2009a. Illergård, K., Ardell, D. H., and Elofsson, A. (2009a). Structure is three to ten times more conserved than sequence - A study of structural response in protein cores. *Proteins: Structure, Function and Bioinformatics*, 77(3):499–508.

Illergård et al., 2009b. Illergård, K., Ardell, D. H., and Elofsson, A. (2009b). Structure is three to ten times more conserved than sequence-A study of structural response in protein cores. *Proteins: Structure, Function, and Bioinformatics*, 77(3):499–508.

Jacobson et al., 2002. Jacobson, M. P., Friesner, R. a., Xiang, Z., and Honig, B. (2002). On the Role of the Crystal Environment in Determining Protein Side-chain Conformations. *Journal of Molecular Biology*, 320(3):597–608.

James et al., 2003. James, L. C., Roversi, P., and Tawfik, D. S. (2003). Antibody multispecificity mediated by conformational diversity. *Science (New York, N.Y.)*, 299(5611):1362–7.

James and Tawfik, 2001. James, L. C. and Tawfik, D. S. (2001). Catalytic and binding poly-reactivities shared by two unrelated proteins: The potential role of promiscuity in enzyme evolution. *Protein science : a publication of the Protein Society*, 10(12):2600–7.

James and Tawfik, 2003. James, L. C. and Tawfik, D. S. (2003). Conformational diversity and protein evolution–a 60-year-old hypothesis revisited. *Trends in biochemical sciences*, 28(7):361–8.

Janin and Sternberg, 2013. Janin, J. and Sternberg, M. J. E. (2013). Protein flexibility, not disorder, is intrinsic to molecular recognition. *F1000 Biology Reports*, 5(January):2.

Juritz et al., 2011. Juritz, E. I., Alberti, S. F., and Parisi, G. D. (2011). PCDB: a database of protein conformational diversity. *Nucleic acids research*, 39(Database issue):D475–9.

Kabsch and Sander, 1983. Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–637.

Karplus and Schulz, 1985. Karplus, P. A. and Schulz, G. E. (1985). Prediction of chain flexibility in proteins. *Naturwissenschaften*, 72(4):212–213.

Karush, 1950. Karush, F. (1950). Heterogeneity of the Binding Sites of Bovine Serum Albumin ¹. *Journal of the American Chemical Society*, 72(6):2705–2713.

Keedy et al., 2009. Keedy, D. A., Williams, C. J., Headd, J. J., Arendall, W. B., Chen, V. B., Kapral, G. J., Gillespie, R. A., Block, J. N., Zemla, A., Richardson, D. C., Richardson, J. S., and Richardson, J. S. (2009). The other 90 % of the protein: assessment beyond the Calphas for CASP8 template-based and high-accuracy models. *Proteins*, 77 Suppl 9(Suppl 9):29–49.

Kern et al., 2005. Kern, D., Eisenmesser, E. Z., and Wolf-Watz, M. (2005). Enzyme dynamics during catalysis measured by NMR spectroscopy. *Methods in Enzymology*, 394:507–524.

Keskin et al., 2000. Keskin, O., Jernigan, R. L., and Bahar, I. (2000). Proteins with similar architecture exhibit similar large-scale dynamic behavior. *Biophysical journal*, 78(4):2093–106.

Khafizov et al., 2014. Khafizov, K., Madrid-Aliste, C., Almo, S. C., and Fiser, A. (2014). Trends in structural coverage of the protein universe and the impact of the Protein Structure Initiative. *Proceedings of the National Academy of Sciences*, 111(10):3733–3738.

Khersonsky et al., 2006. Khersonsky, O., Roodveldt, C., and Tawfik, D. S. (2006). Enzyme promiscuity: evolutionary and mechanistic aspects. *Current opinion in chemical biology*, 10(5):498–508.

Ko et al., 2001. Ko, T. P., Chen, Y. K., Robinson, H., Tsai, P. C., Gao, Y. G., Chen, A. P. C., Wang, A. H. J., and Liang, P. H. (2001). Mechanism of Product Chain Length Determination and the Role of a Flexible Loop in Escherichia coli Undecaprenyl-pyrophosphate Synthase Catalysis. *Journal of Biological Chemistry*, 276(50):47474–47482.

Koehl and Levitt, 2002. Koehl, P. and Levitt, M. (2002). Sequence variations within protein families are linearly related to structural variations. *Journal of Molecular Biology*, 323(3):551–562.

Kopp and Schwede, 2004. Kopp, J. and Schwede, T. (2004). Automated protein structure homology modeling: a progress report. *Pharmacogenomics*, 5(4):405–16.

Koshland, 1994. Koshland, D. E. (1994). The Key-Lock Theory and the Induced Fit Theory Introduction of the Induced Fit Theory. *Angewandte Chemie International Edition*, 33(510):2375–2378.

Koshland, 1998. Koshland, D. E. (1998). Conformational changes: How small is big enough? *Nat Med*, 4(10):1112–1114.

Koshland et al., 1966. Koshland, D. E., Némethy, G., and Filmer, D. (1966). Comparison of experimental binding data and theoretical models in proteins containing subunits. *Biochemistry*, 5(1):365–85.

Koshland et al., 1958. Koshland, D. J., Ray, W. J., and Erwin, M. (1958). Protein structure and enzyme action. *Fed Proc*, 17(4):1145–1150.

Kosloff and Kolodny, 2008. Kosloff, M. and Kolodny, R. (2008). Sequence-similar, structure-dissimilar protein pairs in the PDB. *Proteins: Structure, Function and Genetics*, 71(2):891–902.

Kossel, 1898. Kossel, A. (1898). Ueber die Constitution der einfachsten Eiweissstoffe. *XX*, 25(3-4):165–189.

Kotera et al., 2004. Kotera, M., Okuno, Y., Hattori, M., Goto, S., and Kanehisa, M. (2004). Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions. *Journal of the American Chemical Society*, 126(50):16487–98.

Krissinel and Henrick, 2007. Krissinel, E. and Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. *Journal of molecular biology*, 372(3):774–97.

Kumar et al., 2000. Kumar, S., Ma, B., Tsai, C. J., Sinha, N., and Nussinov, R. (2000). Folding and binding cascades: dynamic landscapes and population shifts. *Protein science : a publication of the Protein Society*, 9(1):10–9.

Lancet and Pecht, 1976. Lancet, D. and Pecht, I. (1976). Kinetic evidence for hapten-induced conformational transition in immunoglobin MOPC 460. *Proceedings of the National Academy of Sciences of the United States of America*, 73(10):3549–53.

Le Guilloux et al., 2009. Le Guilloux, V., Schmidtke, P., and Tuffery, P. (2009). Fpocket: an open source platform for ligand pocket detection. *BMC bioinformatics*, 10(1):168.

Lee and Richards, 1971. Lee, B. and Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *Journal of molecular biology*, 55(3):379–400.

Lee et al., 2003. Lee, R. A., Razaz, M., and Hayward, S. (2003). The DynDom database of protein domain motions. *Bioinformatics (Oxford, England)*, 19(10):1290–1.

Lesk and Chothia, 1984. Lesk, A. M. and Chothia, C. (1984). Mechanisms of domain closure in proteins. *Journal of molecular biology*, 174(1):175–91.

Linding et al., 2003. Linding, R., Jensen, L. J., Diella, F., Bork, P., Gibson, T. J., and Russell, R. B. (2003). Protein disorder prediction: implications for structural proteomics. *Structure (London, England : 1993)*, 11(11):1453–9.

Lindorff-Larsen et al., 2005. Lindorff-Larsen, K., Best, R. B., Depristo, M. A., Dobson, C. M., and Vendruscolo, M. (2005). Simultaneous determination of protein structure and dynamics. *Nature*, 433(7022):128–32.

Liu et al., 2003. Liu, X., Fan, K., and Wang, W. (2003). The number of protein folds and their distribution over families in nature. *Proteins: Structure, Function, and Bioinformatics*, 54(3):491–499.

Lobanov et al., 2008. Lobanov, M. I., Bogatyreva, N. S., and Galzitskaia, O. V. (2008). [Radius of gyration is indicator of compactness of protein structure]. *Molekuliarnaia biologiia*, 42(4):701–6.

Loewenstein et al., 2009. Loewenstein, Y., Raimondo, D., Redfern, O. C., Watson, J., Frishman, D., Linial, M., Orengo, C., Thornton, J., and Tramontano, A. (2009). Protein function annotation by homology-based inference. *Genome Biology*, 10(2):207.

Lovelace et al., 2005. Lovelace, L. L., Minor, W., and Lebioda, L. (2005). Structure of human thymidylate synthase under low-salt conditions. *Acta crystallographica. Section D, Biological crystallography*, 61(Pt 5):622–7.

Ma et al., 2002. Ma, B., Shatsky, M., Wolfson, H. J., and Nussinov, R. (2002). Multiple diverse ligands binding at a single protein site: a matter of pre-existing populations. *Protein science : a publication of the Protein Society*, 11(2):184–97.

Macol et al., 2001. Macol, C. P., Tsuruta, H., Stec, B., and Kantrowitz, E. R. (2001). Direct structural evidence for a concerted allosteric transition in Escherichia coli aspartate transcarbamoylase. *Nature Structural Biology*, 8(5):423–426.

Maguid et al., 2006. Maguid, S., Fernández-Alberti, S., Parisi, G., and Echave, J. (2006). Evolutionary conservation of protein backbone flexibility. *Journal of molecular evolution*, 63(4):448–57.

Marino-Buslje et al., 2017. Marino-Buslje, C., Monzon, A. M., Zea, D. J., Fornasari, M. S., and Parisi, G. (2017). On the dynamical incompleteness of the Protein Data Bank. *Briefings in Bioinformatics*, (June):1–4.

Marks et al., 2017. Marks, C., Shi, J., and Deane, C. M. (2017). Predicting loop conformational ensembles. *Bioinformatics*.

Martin et al., 2011. Martin, A. J. M., Vidotto, M., Boscariol, F., Di Domenico, T., Walsh, I., and Tosatto, S. C. E. (2011). RING: networking interacting residues, evolutionary information and energetics in protein structures. *Bioinformatics (Oxford, England)*, 27(14):2003–5.

McLachlan, 1982. McLachlan, A. (1982). Rapid comparison of protein structures.

Meng and McKnight, 2009. Meng, J. and McKnight, C. J. (2009). Heterogeneity and dynamics in villin headpiece crystal structures. *Acta crystallographica. Section D, Biological crystallography*, 65(Pt 5):470–6.

Mesecar et al., 1997. Mesecar, A. D., Stoddard, B. L., and Koshland, D. E. (1997). Orbital steering in the catalytic power of enzymes: small structural changes with large catalytic consequences. *Science (New York, N.Y.)*, 277(5323):202–6.

Mirsky and Paulin, 1936. Mirsky, A. E. and Paulin, L. (1936). On the Structure of Native, Denatured, and Coagulated Proteins. *Proc. Natl. Acad*, 22(7):439–447.

Mittag et al., 2010. Mittag, T., Kay, L. E., and Forman-Kaya, J. D. (2010). Protein dynamics and conformational disorder in molecular recognition. *Journal of Molecular Recognition*, 23(2):105–116.

Monod et al., 1965. Monod, J., Wyman, J., and Changeux, J. P. (1965). On the Nature of Allosteric Transitions: a Plausible Model. *Journal of molecular biology*, 12(December):88–118.

Monzon et al., 2013. Monzon, A. M., Juritz, E., Fornasari, M. S., Parisi, G., Fornasari, S., Parisi, G., Fornasari, M. S., and Parisi, G. (2013). CoDNaS: A database of conformational diversity in the native state of proteins. *Bioinformatics*, 29(19):2512–2514.

Monzon et al., 2016. Monzon, A. M., Rohr, C. O., Fornasari, M. S., and Parisi, G. (2016). CoDNaS 2.0: a comprehensive database of protein conformational diversity in the native state. *Database : the journal of biological databases and curation*, 2016(0):baw038–.

Monzon et al., 2017a. Monzon, A. M., Zea, D. J., Fornasari, M. S., Saldaño, T. E., Fernandez-Alberti, S., Tosatto, S. C. E., and Parisi, G. (2017a). Conformational diversity analysis reveals three functional mechanisms in proteins. *PLoS Computational Biology*, 13(2):1–18.

Monzon et al., 2017b. Monzon, A. M., Zea, D. J., Marino-Buslje, C., and Parisi, G. (2017b). Homology modeling in a dynamical world. *Protein Science*, 26(11):2195–2206.

Myers et al., 2005. Myers, R. S., Amaro, R. E., Luthey-Schulten, Z. A., and Davisson, V. J. (2005). Reaction coupling through interdomain contacts in imidazole glycerol phosphate synthase. *Biochemistry*, 44(36):11974–85.

Needleman and Wunsch, 1970. Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453.

Nussinov and Ma, 2012. Nussinov, R. and Ma, B. (2012). Protein dynamics and conformational selection in bidirectional signal transduction. *BMC biology*, 10(1):2.

Nussinov et al., 2013. Nussinov, R., Ma, B., Tsai, C. J., and Csermely, P. (2013). Allosteric conformational barcodes direct signaling in the cell. *Structure*, 21(9):1509–1521.

Nussinov and Tsai, 2015. Nussinov, R. and Tsai, C. J. (2015). Allostery without a conformational change? Revisiting the paradigm. *Current Opinion in Structural Biology*, 30:17–24.

Oldfield and Dunker, 2014. Oldfield, C. J. and Dunker, A. K. (2014). Intrinsically Disordered Proteins and Intrinsically Disordered Protein Regions. *Annual Review of Biochemistry*, 83(1):553–584.

Onuchic and Wolynes, 2004. Onuchic, J. N. and Wolynes, P. G. (2004). Theory of protein folding. *Current opinion in structural biology*, 14(1):70–5.

Orengo et al., 1994. Orengo, C. A., Jones, D. T., and Thornton, J. M. (1994). Protein superfamilles and domain superfolds. *Nature*, 372(6507):631–634.

Ortiz et al., 2002. Ortiz, A. R., Strauss, C. E. M., and Olmea, O. (2002). MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein science : a publication of the Protein Society*, 11(11):2606–21.

Palopoli et al., 2013. Palopoli, N., Lanzarotti, E., and Parisi, G. (2013). BeEP Server: Using evolutionary information for quality assessment of protein structure models. *Nucleic acids research*, 41(Web Server issue):398–405.

Palopoli et al., 2016. Palopoli, N., Monzon, A. M., Parisi, G., and Fornasari, M. S. (2016). Addressing the role of conformational diversity in protein structure prediction. *PLoS ONE*, 11(5):1–14.

Panchenko et al., 2005. Panchenko, A. R., Wolf, Y. I., Panchenko, L. A., and Madej, T. (2005). Evolutionary plasticity of protein families: Coupling between sequence and structure variation. *Proteins: Structure, Function, and Bioinformatics*, 61(3):535–544.

Papaleo et al., 2016. Papaleo, E., Saladino, G., Lambrughi, M., Lindorff-Larsen, K., Gervasio, F. L., and Nussinov, R. (2016). The Role of Protein Loops and Linkers in Conformational Dynamics and Allostery. *Chemical Reviews*, 116(11):6391–6423.

Parisi et al., 2015. Parisi, G., Zea, D. J., Monzon, A. M., and Marino-Buslje, C. (2015). Conformational diversity and the emergence of sequence signatures during evolution. *Current Opinion in Structural Biology*, 32:58–65.

Pauling, 1940. Pauling, L. (1940). A Theory of the Structure and Process of Formation of Antibodies [*]. *Journal of the American Chemical Society*, 62(10):2643–2657.

Peña et al., 2006. Peña, M. M. O., Xing, Y. Y., Koli, S., and Berger, F. G. (2006). Role of N-terminal residues in the ubiquitin-independent degradation of human thymidylate synthase. *The Biochemical journal*, 394(Pt 1):355–63.

Perutz and Mathews, 1966. Perutz, M. and Mathews, F. (1966). An X-ray study of azide methaemoglobin. *Journal of Molecular Biology*, 21(1):199–202.

Perutz et al., 1960. Perutz, M. F., ROSSMANN, M. G., CULLIS, A. F., MUIRHEAD, H., WILL, G., and NORTH, A. C. T. (1960). Structure of Hæmoglobin: A Three-Dimensional Fourier Synthesis at 5.5-A. Resolution, Obtained by X-Ray Analysis. *Nature*, 185(4711):416–422.

Pettersen et al., 2004. Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004). UCSF Chimera - A visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25(13):1605–1612.

Piovesan et al., 2016. Piovesan, D., Tabaro, F., Mičetić, I., Necci, M., Quaglia, F., Oldfield, C. J., Aspromonte, M. C., Davey, N. E., Davidović, R., Dosztányi, Z., Elofsson, A., Gasparini, A., Hatos, A., Kajava, A. V., Kalmar, L., Leonardi, E., Lazar, T., Macedo-Ribeiro, S., Macossay-Castillo, M., Meszaros, A., Minervini, G.,

Murvai, N., Pujols, J., Roche, D. B., Salladini, E., Schad, E., Schramm, A., Szabo, B., Tantos, A., Tonello, F., Tsirigos, K. D., Veljković, N., Ventura, S., Vranken, W., Warholm, P., Uversky, V. N., Dunker, A. K., Longhi, S., Tompa, P., and Tosatto, S. C. (2016). DisProt 7.0: a major update of the database of disordered proteins. *Nucleic Acids Research*, 45(D1):gkw1279.

Piovesan et al., 2018. Piovesan, D., Tabaro, F., Paladin, L., Necci, M., Mičetić, 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., and Tosatto, S. C. E. (2018). MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins. *Nucleic Acids Research*, 46(D1):D471–D476.

Prabhu et al., 2003. Prabhu, N. V., Lee, A. L., Wand, A. J., and Sharp, K. A. (2003). Dynamics and Entropy of a Calmodulin?Peptide Complex Studied by NMR and Molecular Dynamics [?]. *Biochemistry*, 42(2):562–570.

Pravda et al., 2014a. Pravda, L., Berka, K., Svobodová Vařeková, R., Sehnal, D., Banáš, P., Laskowski, R. a., Koča, J., and Otyepka, M. (2014a). Anatomy of enzyme channels. *BMC Bioinformatics*, 15(1):1–8.

Pravda et al., 2014b. Pravda, L., Berka, K., Svobodová Vařeková, R., Sehnal, D., Banáš, P., Laskowski, R. A., Koča, J., and Otyepka, M. (2014b). Anatomy of enzyme channels. *BMC Bioinformatics*, 15(1):379.

Qu et al., 2009. Qu, X., Swanson, R., Day, R., and Tsai, J. (2009). A guide to template based structure prediction. *Current protein & peptide science*, 10(3):270–85.

Rackovsky, 2015. Rackovsky, S. (2015). Nonlinearities in protein space limit the utility of informatics in protein biophysics. *Proteins: Structure, Function and Bioinformatics*, 83(11):1923–1928.

Rapp and Pollack, 2005. Rapp, C. S. and Pollack, R. M. (2005). Crystal packing effects on protein loops. *Proteins*, 60(1):103–9.

Rashin et al., 2010. Rashin, A. A., Rashin, A. H. L., and Jernigan, R. L. (2010). Diversity of function-related conformational changes in proteins: coordinate uncertainty, fragment rigidity, and stability. *Biochemistry*, 49(27):5683–704.

Rost, 1999. Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Engineering, Design and Selection*, 12(2):85–94.

Russell and Barton, 1994. Russell, R. B. and Barton, G. J. (1994). Structural features can be unconserved in proteins with similar folds. An analysis of side-chain to side-chain contacts secondary structure and accessibility.

Russell et al., 1997. Russell, R. B., Saqi, M. A., Sayle, R. A., Bates, P. A., and Sternberg, M. J. (1997). Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation 1 1Edited by F. E. Cohen. *Journal of Molecular Biology*, 269(3):423–439.

Sadowski and Jones, 2009. Sadowski, M. I. and Jones, D. T. (2009). The sequence-structure relationship and protein function prediction. *Current Opinion in Structural Biology*, 19(3):357–362.

Saldaño et al., 2016. Saldaño, T. E., Monzon, A. M., Parisi, G., and Fernandez-Alberti, S. (2016). Evolutionary Conserved Positions Define Protein Conformational Diversity. *PLoS computational biology*, 12(3):e1004775.

Schulenburg and Hilvert, 2013. Schulenburg, C. and Hilvert, D. (2013). Protein Conformational Disorder and Enzyme Catalysis. In *Topics in current chemistry*, volume 337, pages 41–67.

Schwede, 2013. Schwede, T. (2013). Protein Modeling: What Happened to the "Protein Structure Gap"? *Structure*, 21(9):1531–1540.

Shatsky et al., 2002. Shatsky, M., Nussinov, R., and Wolfson, H. J. (2002). Flexible protein alignment and hinge detection. *Proteins: Structure, Function and Genetics*, 48(2):242–256.

Sikic et al., 2010. Sikic, K., Tomic, S., and Carugo, O. (2010). Systematic comparison of crystal and NMR protein structures deposited in the protein data bank. *The open biochemistry journal*, 4:83–95.

Sillitoe et al., 2015. Sillitoe, I., Lewis, T. E., Cuff, A., Das, S., Ashford, P., Dawson, N. L., Furnham, N., Laskowski, R. A., Lee, D., Lees, J. G., Lehtinen, S., Studer, R. A., Thornton, J., and Orengo, C. A. (2015). CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Research*, 43(D1):D376–D381.

Sinha and Nussinov, 2001. Sinha, N. and Nussinov, R. (2001). Point mutations and sequence variability in proteins: redistributions of preexisting populations. *Proceedings of the National Academy of Sciences of the United States of America*, 98(6):3139–44.

Skopalík et al., 2008. Skopalík, J., Anzenbacher, P., and Otyepka, M. (2008). Flexibility of human cytochromes P450: molecular dynamics reveals differences between CYPs 3A4, 2C9, and 2A6, which correlate with their substrate preferences. *The journal of physical chemistry. B*, 112(27):8165–73.

Smith et al., 2003. Smith, D. K., Radivojac, P., Obradovic, Z., Dunker, A. K., and Zhu, G. (2003). Improved amino acid flexibility parameters. *Protein science : a publication of the Protein Society*, 12(5):1060–72.

Strnad, 2014. Strnad, O. (2014). *Algorithms for Detecting Pathways in Large Protein Structures and Their Ensembles*. PhD thesis.

Team, 2017. Team, R. C. (2017). R: A language and environment for statistical computing.

Tokuriki and Tawfik, 2009. Tokuriki, N. and Tawfik, D. (2009). Protein dynamism and evolvability. *Science*, 324(5924):203–207.

Tompa, 2002. Tompa, P. (2002). Intrinsically unstructured proteins. *Trends in biochemical sciences*, 27(10):527–33.

Tompa, 2010. Tompa, P. (2010). *Structure and function of intrinsically disordered proteins.* Chapman & Hall/CRC Press.

Tompa, 2012. Tompa, P. (2012). Intrinsically disordered proteins: a 10-year recap. *Trends in biochemical sciences*, 37(12):509–16.

Tsai et al., 1999. Tsai, C., Kumar, S., Ma, B., and Nussinov, R. (1999). Folding funnels, binding funnels, and protein function. *Protein Science*, 8:1181–1190.

Tsai et al., 2008. Tsai, C.-J., del Sol, A., and Nussinov, R. (2008). Allostery: Absence of a Change in Shape Does Not Imply that Allostery Is Not at Play. *Journal of Molecular Biology*, 378(1):1–11.

Tsai and Nussinov, 2014. Tsai, C. J. and Nussinov, R. (2014). A Unified View of "How Allostery Works". *PLoS Computational Biology*, 10(2).

Umbarger and Brown, 1957. Umbarger, H. E. and Brown, B. (1957). Threonine deamination in Escherichia coli. II. Evidence for two L-threonine deaminases. *Journal of bacteriology*, 73(1):105–112.

Vacic et al., 2007. Vacic, V., Uversky, V. N., Dunker, a. K., and Lonardi, S. (2007). Composition Profiler: a tool for discovery and visualization of amino acid composition differences. *BMC bioinformatics*, 8:211.

van der Lee et al., 2014. 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., and Babu, M. M. (2014). Classification of Intrinsically Disordered Regions and Proteins. *Chemical Reviews*, 114(13):6589–6631.

Varadi et al., 2014. 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., and Tompa, P. (2014). PE-DB: A database of structural ensembles of intrinsically disordered and of unfolded proteins. *Nucleic Acids Research*, 42(D1):326–335.

Velankar et al., 2013. Velankar, S., Dana, J. M., Jacobsen, J., van Ginkel, G., Gane, P. J., Luo, J., Oldfield, T. J., O'Donovan, C., Martin, M.-J., and Kleywegt, G. J. (2013). SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic acids research*, 41(Database issue):D483–9.

Velyvis et al., 2007. Velyvis, A., Yang, Y. R., Schachman, H. K., and Kay, L. E. (2007). A solution NMR study showing that active site ligands and nucleotides directly perturb the allosteric equilibrium in aspartate transcarbamoylase. *Proceedings of the National Academy of Sciences of the United States of America*, 104(21):8815–20.

Vihinen et al., 1994. Vihinen, M., Torkkila, E., and Riikonen, P. (1994). Accuracy of protein flexibility predictions. *Proteins: Structure, Function, and Genetics*, 19(2):141–149.

Vital de Oliveira, 2014. Vital de Oliveira, O. (2014). Molecular Dynamics and Metadynamics Simulations of the Cellulase Cel48F. *Enzyme Research*, 2014:1–7.

Volkman et al., 2001. Volkman, B. F., Lipson, D., Wemmer, D. E., and Kern, D. (2001). Two-state allosteric behavior in a single-domain signaling protein. *Science (New York, N.Y.)*, 291(5512):2429–33.

Wei et al., 2016. Wei, G., Xi, W., Nussinov, R., and Ma, B. (2016). Protein Ensembles: How Does Nature Harness Thermodynamic Fluctuations for Life? the Diverse Functional Roles of Conformational Ensembles in the Cell. *Chemical Reviews*, 116(11):6516–6551.

Weikl and Paul, 2014. Weikl, T. R. and Paul, F. (2014). Conformational selection in protein binding and function. *Protein Science*, 23(11):1508–1518.

Wilson et al., 2000. Wilson, C. A., Kreychman, J., and Gerstein, M. (2000). Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *Journal of Molecular Biology*, 297(1):233–249.

Wolf-Watz et al., 2004. Wolf-Watz, M., Thai, V., Henzler-Wildman, K., Hadjipavlou, G., Eisenmesser, E. Z., and Kern, D. (2004). Linkage between dynamics and catalysis in a thermophilic-mesophilic enzyme pair. *Nature Structural & Molecular Biology*, 11(10):945–949.

Wood and Pearson, 1999a. Wood, T. C. and Pearson, W. R. (1999a). Evolution of protein sequences and structures. *Journal of molecular biology*, 291(4):977–995.

Wood and Pearson, 1999b. Wood, T. C. and Pearson, W. R. (1999b). Evolution of protein sequences and structures. *Journal of molecular biology*, 291(4):977–995.

Wright and Dyson, 1999. Wright, P. E. and Dyson, H. J. (1999). Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *Journal of molecular biology*, 293(2):321–31.

Xia and Levitt, 2004. Xia, Y. and Levitt, M. (2004). Simulating protein evolution in sequence and structure space.

Xiang, 2006. Xiang, Z. (2006). Advances in homology protein structure modeling. *Curr. Protein Pept. Sci.*, 7(3):217–27.

Xie et al., 2007. Xie, H., Vucetic, S., Iakoucheva, L. M., Oldfield, C. J., Dunker, A. K., Uversky, V. N., and Obradovic, Z. (2007). Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. *Journal of Proteome Research*, 6(5):1882–1898.

Yan et al., 2013. Yan, R., Xu, D., Yang, J., Walker, S., and Zhang, Y. (2013). A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction. *Scientific reports*, 3:2619.

Yang et al., 2013. Yang, J., Roy, A., and Zhang, Y. (2013). BioLiP: A semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Research*, 41(D1):1096–1103.

Zea et al., 2013. Zea, D. J., Monzon, A. M., Fornasari, M. S., Marino-Buslje, C., and Parisi, G. (2013). Protein conformational diversity correlates with evolutionary rate. *Molecular Biology and Evolution*, 30(7):1500–1503.

Zea et al., 2016. Zea, D. J., Monzon, A. M., Gonzalez, C., Fornasari, M. S., Tosatto, S. C., and Parisi, G. (2016). Disorder transitions and conformational diversity cooperatively modulate biological function in proteins. *Protein Science*, 25(6):1138–1146.

Zhang, 2008. Zhang, Y. (2008). Progress and challenges in protein structure prediction. *Current Opinion in Structural Biology*, 18(3):342–348.

Zhang, 2009. Zhang, Y. (2009). Protein structure prediction: when is it useful? *Current Opinion in Structural Biology*, 19(2):145–155.

Zhang and Skolnick, 2004. Zhang, Y. and Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins*, 57(4):702–10.

Zhou et al., 1998. Zhou, H.-X., Wlodek, S. T., and McCammon, J. a. (1998). Conformation gating as a mechanism for enzyme specificity. *Proceedings of the National Academy of Sciences*, 95(16):9280–9283.

Zoete et al., 2002. Zoete, V., Michielin, O., and Karplus, M. (2002). Relation between sequence and structure of HIV-1 protease inhibitor complexes: a model system for the analysis of protein flexibility. *Journal of molecular biology*, 315(1):21–52.