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

Universidad
Nacional
de Quilmes

Barletta, Patricio

Dinámica de cavidades proteicas : flexibilidad y cambios conformacionales

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

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

Cita recomendada:
Barletta, P. (2020). Dinámica de cavidades proteicas: flexibilidad y cambios conformacionales. (Tesis de doctorado). Universidad Nacional de Quilmes, Bernal, Argentina. Disponible en RIDAA-UNQ Repositorio Institucional Digital de Acceso Abierto de la Universidad Nacional de Quilmes
http://ridaa.unq.edu.ar/handle/20.500.11807/2183

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

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

Patricio Barletta, Repositorio Institucional Digital de Acceso Abierto,
Abril del 2020, pp. 171,
http://ridaa.unq.edu.ar,
Universidad Nacional de Quilmes, Secretaría de Posgrado,
Doctorado en Ciencia y Tecnología

Dinámica de cavidades proteicas: flexibilidad y cambios conformacionales

TESIS DOCTORAL

Patricio Barletta

pbarletta@gmail.com

Resumen

La flexibilidad y dinámica de proteínas pueden considerarse como el puente que permite conectar su estructura y función biológica. Actualmente está bien establecido que la formafuncional de una proteína, normalmente conocida como el estado nativo, no puede ser representada por una única estructura. El estado nativo de una proteína se define como una distribución no uniforme de poblaciones de confórmeros en equilibrio dinámico. El paisaje energético de proteínas puede describirse como un equilibrio de poblaciones preexistentes, de estados separados por barreras energéticas que pueden superarse mediante fluctuaciones térmicas. Estos estados corresponden a diferentes conformaciones cuyas poblaciones siguen distribuciones de acuerdo con la termodinámica estadística. Las alturas de las barreras energéticas que separan las conformaciones definen la escala de tiempo del intercambio conformacional. Esta visión emergente de la estructura y dinámica de las proteínas proporciona un marco teórico para estudiar embudos de plegamiento de proteínas, selección conformacional, alosterismo, unión al ligando, evolución y por lo tanto, el objetivo final de comprender el mecanismo de la función biológica de las proteínas.

La flexibilidad y dinámica vibracional asociada a cada conformación garantiza las transiciones conformacionales que, debido a su importancia funcional, poseen rasgos conservados evolutivamente. Las técnicas de Análisis de Modos Normales (NMA), Dinámica Molecular (MD) y Análisis de Componentes Principales (PCA) permiten su análisis según los distintos grados de detalles requeridos o el enfoque del estudio realizado. Estas metodologías, junto a las desarrolladas a lo largo de esta tesis, fueron utilizadas para el estudio de la flexibilidad y dinámica de cavidades.

La dinámica de las proteínas altera también la geometría y el tamaño sus cavidades. Los túneles y cavidades de proteínas suelen jugar un rol crítico en su función biológica. Tanto el reconocimiento y la unión a ligandos, como las actividades de transporte y catálisis enzimática requieren, comúnmente, reorganizaciones de las cavidades. Debido a este rol funcional, la flexibilidad y dinámica de las cavidades debe estar garantizada por la dinámica vibracional de las proteínas.

En esta tesis, presentamos un método novedoso para caracterizar la dinámica de las cavidades de proteínas en términos de su vector de gradiente de volumen. Para este propósito, hacemos uso de algoritmos para el cálculo del volumen de la cavidad que resultan robusitos para las diferenciaciones numéricas. El vector de gradiente de volumen se expresa de coordenadas colectivas, tanto modos de PCA o NMA. Analizamos las contribuciones de las distintas coordenadas colectivas al vector de gradiente de volumen de acuerdo con su

frecuencia y grado de deslocalización. En todos nuestros casos de prueba, encontramos que los modos de baja frecuencia juegan un papel crítico junto con contribuciones menores de los modos de media frecuencia. Estos modos implican, principalmente, movimientos concertados de los residuos que recubren la cavidad estudiada. Mostramos que las proteínas cuyas fluctuaciones colectivas de baja frecuencia contribuyen más a los cambios en el volumen de la cavidad exhiben cavidades más flexibles. En resumen, el método desarrollado en esta tesis permite realizar estudios de la relación entre las fluctuaciones de la proteína y los cambios en el volumen de sus cavidades, analizar y comparar flexibilidades de cavidades y estudiar su dinámica ante cambios conformacionales de la proteína.
Esto implicó el desarrollo completo de un nuevo software, ya que ninguno de los métodos preexistentes nos permitió darle este enfoque al trabajo de tesis.
En la segunda parte de la tesis, aplicamos la metodología desarrollada al estudio del receptor del factor de crecimiento epidérmico (EGFR). El EGFR es un receptor prototípico de la superficie celular que desempeña un papel clave en la regulación de la señalización celular, la proliferación y la diferenciación. Las mutaciones de su dominio de quinasa se han asociado con el desarrollo de una variedad de cánceres y, por lo tanto, ha sido el objetivo del diseño del fármaco. Se ha comprobado que las sustituciones de aminoácidos individuales (SAS) en este dominio alteran el equilibrio de las poblaciones preexistentes.
A pesar de los avances en las descripciones estructurales de sus llamadas conformaciones activas e inactivas, los aspectos dinámicos asociados que los caracterizan aún no se han estudiado a fondo. Como los comportamientos dinámicos y los movimientos moleculares de las proteínas son importantes para una comprensión completa de sus relaciones estructura-función, presentamos un procedimiento novedoso, utilizando el Análisis de Modos Normales (NMA), para identificar la dinámica colectiva compartida entre diferentes confórmeros en EGFR quinasa. El método permite la comparación de patrones de modos vibracionales de baja frecuencia que definen direcciones representativas de movimientos.
Nuestro procedimiento puede enfatizar las principales similitudes y diferencias entre las dinámicas colectivas de los diferentes confórmeros. En el caso de la quinasa de EGFR, se han encontrado dos direcciones representativas de los movimientos como huellas digitales dinámicas de los confórmeros activos. El movimiento de la proteína en ambas direcciones revela tener un impacto significativo en el volumen de la cavidad del bolsillo principal del sitio activo. De lo contrario, los confórmeros inactivos exhiben una distribución más heterogénea de los movimientos colectivos.
Otro caso de aplicación de los métodos desarrollados fueron las proteínas de unión a lípidos (LBP). Las LBPs son proteínas solubles responsables de la absorción, transporte y almacenamiento de una gran variedad de moléculas lipofílicas, como ácidos grasos, esteroides y otros lípidos en el entorno celular. Entre las LBPs, las proteínas de unión a ácidos grasos (FABP) presentan afinidades de unión preferenciales por los ácidos grasos de cadena larga. Mientras que la mayoría de las FABPs en vertebrados e invertebrados presentan estructuras similares de barril β con ligandos alojados en su cavidad central, los gusanos nematodos parásitos exhiben proteínas adicionales de unión a retinol y ácidos grasos ricas en hélice α que son inusuales (FAR). En esta tesis, informamos la comparación de simulaciones de dinámica molecular extendida realizadas en los estados de enlace de ácido libre de ligando y palmítico del Necator americanus FAR-1 (Na-FAR-1) con respecto a otras FABPs clásicas de barril β. El análisis de componentes principales (PCA) se ha utilizado para identificar las diferentes conformaciones adoptadas por cada sistema durante las simulaciones MD. La estructura compuesta por hélices α abarca una compleja cavidad interna de unión a ligando con una notable plasticidad conformacional que permite el cambio

Repositorio Institucional Digital de Acceso Abierto, Universidad Nacional de Quilmes

reversible entre estados distintos en el holo-Na-FAR-1. La cavidad puede cambiar hasta un tercio de su tamaño por cambios conformacionales del complejo proteína-ligando.
Además, el ligando dentro de la cavidad no está fijo sino que experimenta grandes cambios conformacionales entre conformaciones plegadas y estiradas. Estos cambios en la conformación del ligando siguen a los cambios en el tamaño de la cavidad dictados por la conformación transitoria de la proteína. Por el contrario, el complejo proteína-ligando en las FABP de barril β fluctúa alrededor de una conformación única. La cavidad ligando holo-Na-FAR-1 significativamente más flexible explica su multiplicidad de ligando más grande con respecto a las FABP de barril β.
Para todo esto, en el marco de la presente tesis doctoral, se han desarrollado técnicas que permiten identificar y caracterizar aspectos de la dinámica vibracional de proteínas, directamente vinculados a la flexibilidad y dinámica de cavidades y su conexión con la flexibilidad y dinámica del resto de la estructura proteica. También se ha estudiando la manera de comparar subespacios de modos normales específicos para distintos confórmeros de una proteína. También hemos avanzado en el desarrollo del método de ANA (Analysis of Null Areas) que permite establecer una conexión entre las fluctuaciones térmicas y cambios en las cavidades de las proteínas que permite analizar el impacto que tienen los desplazamientos de la estructura proteica en direcciones predefinidas sobre el volumen de las cavidades. Estas direcciones predefinidas pueden ser coordenadas colectivas de relevancia para la función biológica de la proteína. El método de ANA se facilita a la comunidad científica por medio de un software curado, documentado y un sitio oficial de soporte ANA.

Repositorio Institucional Digital de Acceso Abierto, Universidad Nacional de Quilmes

Dinámica de cavidades proteicas: flexibilidad y cambios conformacionales.

Alumno: Patricio Barletta
Director: Sebastián Fernandez-Alberti

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

Unidad de Fisicoquímica
Departamento de Ciencia y Tecnología
Universidad Nacional de Quilmes

Febrero 2020

Table of Contents

Abstract **i**

Abreviaturas

1 Estado nativo y diversidad conformacional **1**

1.1 Definiciones históricas del estado nativo 1

1.2 La interacción proteína - ligando 2

1.2.1 El modelo de llave-cerradura de Fischer, *key-lock* 2

1.2.2 El modelo de encaje inducido de Koshland, *induced fit* 2

1.2.3 El equilibrio conformacional de Monod, *pre-equilibrium* 4

1.3 La pregunta por el plegamiento 7

1.3.1 La paradoja de Levinthal 7

1.3.2 *Folding pathways* 7

1.3.3 *Folding funnels* 8

1.4 El estado nativo, el fondo del embudo 10

1.5 Diversidad conformacional y función proteica 11

1.6 Desorden en proteínas 12

2 Dinámica Molecular **13**

2.1 Introducción histórica 13

2.2 Métodos clásicos y métodos cuánticos 14

2.3 Fundamentos de la Mecánica Molecular 15

2.3.1 Campos de fuerza 15

2.3.2 Condiciones periódicas de contorno 20

2.3.2.1 Truncamiento de potenciales de corto alcance 21

2.3.3 Algoritmo básico de la Dinámica Molecular 22

2.3.4 Condiciones iniciales y preparación del sistema 24

2.3.4.1 Minimización 25

2.3.4.2 Calentamiento: asignación de velocidades 26

2.3.4.3 Equilibración 26

2.3.5 Termostatos y Barostatos 27

2.3.5.1 Termostato 27

2.3.5.2 Barostato 28

2.4 Practicas metodológicas de la Dinámica Molecular observadas en esta tesis 29

2.4.1 Preparación del sistema 29
2.4.2 Minimización 30
2.4.3 Calentamiento 30
2.4.4 Equilibración 30

3 Análisis de movimientos colectivos 31
3.1 Análisis Modos Normales 31
3.1.1 Introducción histórica 31
3.1.2 Ecuaciones del movimiento vibracional 32
3.1.3 Solución de las ecuaciones de movimiento 35
3.1.3.1 Modos normales de vibración 39
3.1.4 Modelo de Red Anisotrópica 40
3.1.4.1 Fundamentos 40
3.1.4.2 Aplicaciones de los modos normales 46
3.1.4.3 Adaptaciones utilizadas en esta tesis: constante de fuerza asociada al tipo de interacción 48
3.2 Dinámica esencial: Análisis de Componentes Principales 50
3.2.1 Análisis cuasiarmónico 52

4 Cavidades: clasificación, dinámica y los programas y métodos para estudiarlas 55
4.1 Introducción 55
4.2 Dinámica de cavidades 56
4.2.1 Encaje inducido o selección conformacional 58
4.3 Programas de detección de cavidades 59
4.3.1 Métodos de grilla 59
4.3.1.1 Cálculo de volumen 59
4.3.1.2 Definición de la cavidad 61
4.3.2 Esfera *gap* inscripta 61
4.3.3 Esfera rodante 63
4.3.4 Superficie 65
4.3.5 Teselación 69
4.3.5.1 *Convex Hull* (CH) 70
4.3.5.2 *Voronoi Diagram* (VD) 73
4.3.5.3 *Delaunay Triangulation* (DT) 74

5 ANA: Analysis of Null Areas 77
5.1 Introducción 77
5.2 Definición de la cavidad: *Included Area* 77
5.2.1 Estrategias utilizados por otros programas 77
5.2.2 Estrategia utilizada por ANA: *Convex Hull* como *Included Area* . 79

5.3 Cálculo del volumen vacío: *Delaunay Triangulation* 79
5.4 Dinámica y flexibilidad de cavidades: *Non-Delaunay Dynamics* 83
5.4.1 Gradiente del volumen: *Volume Gradient Vector* 84
5.4.2 Continuidad en el cálculo del volumen y *Non-Delaunay Dynamics* (NDD) 86
5.4.3 Casos de estudio 88
5.4.3.1 Dinámicas Moleculares y PCA 88
5.4.4 Resultados 89
5.4.4.1 Frecuencias de los modos relevantes 89
5.4.4.2 Deslocalización de los modos que afectan a la cavidad . . 91
5.4.4.3 Residuos de la pared 93
5.4.4.4 Flexibilidad de la cavidad 95

6 Receptor del Factor de Crecimiento Epidermal (EGFR) 98
6.1 Introducción 98
6.2 Métodología 100
6.2.1 Conjunto de estructuras de quinasas activas e inactivas 100
6.2.2 Comparación ponderada de subespacios de NMA 100
6.2.3 SVD y vectores representativos 102
6.3 Resultados 103

7 Lipid Binding Proteins 113
7.1 Introducción 113
7.2 Métodos 115
7.2.1 Dinámica Molecular y su análisis 115
7.2.2 Cavidad del ligando: definición, volumen y flexibilidad 116
7.3 Resultados 116
7.3.1 LBPs de hélices α 116
7.3.2 LBPs de hebras β 124
7.3.3 Flexibilidad relativa de las cavidades 126

8 Conclusiones 129

Apéndice A: Protocolos de equilibración utilizados 132

Apéndice B: Publicaciones 136

Apéndice C: Información Suplementaria de las publicaciones incluídas 137
8.1 Capítulo 6 137
8.1.1 PDBs del conjunto de estructuras quinasas 137
8.1.2 Residuos de la cavidad activa 138
8.2 Capítulo 7 139

8.2.1 Tablas de residuos de la cavidad 139
8.2.1.1 I-FABP apo (1URE) 139
8.2.1.2 I-FABP apo (1IFB) 139
8.2.1.3 I-FABP holo (2IFB) 140
8.2.1.4 Na-FAR-1 apo (4UET) 140
8.2.1.5 Na-FAR-1 holo (4XCP) 141
8.2.1.6 Ce-FAR-7 apo (2W9Y) 141
8.2.2 Fluctuaciones cuadráticas medias de raíz (RMSF) durante nuestras simulaciones MD 142
8.2.3 Histograma de la proyección del conjunto de instantáneas MD de holo-I-FABP en su tercer modo PCA 143
8.2.4 Primero (rojo) y segundo (azul) modos de PCA de (a) apo- y (b) holo-I-FABP 144
8.2.5 Superposición de los cuatro conformadores (A, B, C y D) de apo-I-FABP 145
8.2.6 Gráficos de densidad de contorno de la proyección del conjunto de instantáneas MD de holo-I-FABP en sus modos PCA primero y segundo correspondientes 146

Referencias **147**

Abreviaturas

PDB	**P**rotein **D**ata **B**ank
pdb	extensión correspondiente al formato de los archivos extraídos de la PDB
NMR	***N**uclear **M**agnetic **R**esonance*
	(Resonancia Magnética nuclear)
XRD	***X**- **R**ay **D**iffraction*
	(Difracción de Rayos X)
$C\alpha$	**C**arbono **A**lfa
RMSD	Cα ***R**oot **M**ean **S**quare **D**eviation*
	(Raíz de la Desviación Cuadrática Media)
SEP	**S**uperficie de **E**nergía **P**otencial
IDP	***I**ntrinsically **D**isordered **P**rotein*
	(Proteína Intrínsicamente Desordenada)
IDR	***I**ntrinsically **D**isordered **R**egion*
	(Región Intrínsicamente Desordenada)
Å	*angstroms*
MM	***M**ecánica **M**olecular*
CG	***C**oarse **G**rained*
	(Grano Grueso)
PBC	**P**eriodic **B**oundary **C**onditions
	(Condiciones Periódicas de Contorno)
PCA	***P**rincipal **C**omponent **A**nalysis*
	(Análisis de Componentes Principales)
ED	***E**ssential **S**ubspace*
	(Subespacio Esencial)
ED	***E**ssential **D**ynamics*
	(Dinámica Esencial)
NMA	***N**ormal **M**odes **A**nalysis*
	(Análisis de Modos Normales)
ENM	***E**lastic **N**etwork **M**odel*
	(Modelo de Red Elástica)
ANM	***A**nisotropic **N**etwork **M**odel*

	(Modelo de Red Anisotrópica)
GNM	***G**aussian **N**etwork **M**odel*
	(Modelo de Red Gausiana)
MD	***M**olecular **D**ynamics*
	(Dinámica Molecular)
QM/MM	***Q**uantum **M**echanics **M**olecular **M**echanics*
	(Mecánica cuántica / Mecánica Clásica)
LB	***L**igand **B**ound*
	(Unida a Ligando)
LF	***L**igand **F**ree*
	(Libre de Ligando)
SSE	***S**econdary **S**structure **E**lement*
	(Elemento de estructura secundaria)
SVdW	***S**uperficie de **V**an **d**er **W**alls*
SAS	***S**olvent **E**xcluded **S**urface*
	(Superficie Excluída del Solvente)
SES	***S**olvent **A**ccessible **S**urface*
	(Superficie Accesible al Solvente)
CH	***C**onvex **H**ull*
	(Caparazón Convexo)
DT	***D**elaunay **T**riangulation*
	(Triangulación de Delaunay)
VD	***V**oronoi **D**iagram*
	(Diagrama de Voronoi)
ANA	***A**nalysis of **N**ull **A**reas*
	(Análisis de Espacios vacíos)
NDD	***N**on-**D**elaunay **D**ynamics*
	(Dinámicas no Delaunay)
VGV	***V**olume **G**radient **V**ector*
	(Vector Gradiente del volumen)
PN	***P**articipation **N**umber*
	(Número de Participación)
EGFR	***E**pidermal **G**rowth **F**actor **R**eceptor*
	(Receptor del Factor de Crecimiento Epidermal)
SAS	***S**ingle **A**minoacid **S**ubstitution*
	(Sustitución de un Único Aminoácido o Mutación Puntual)
LBP	***L**ipid **B**inding **P**rotein*
	(Proteína de Unión a Lipido)
KS	***K**olmogorov **S**mirnov*
FAR	***F**atyy **A**cid and **R**etinol binding protein*

	(Proteína de Unión a Ácidos Grasos y Retinol)
FABP	***F****atyy* ***A****cid* ***B****inding* ***P****rotein*
	(Proteína de Unión a Ácidos Grasos)

Chapter 1

Estado nativo y diversidad conformacional

En este capítulo se introducen los conceptos básicos de la biología estructural de proteínas utilizados en esta tesis: unión a ligando, plegamiento y estado nativo, entendiendo que estos 3 conceptos están intrínsicamente acoplados.

1.1 Definiciones históricas del estado nativo

En 1936, Mirsky y Pauling definen al estado nativo de una proteína como una o más cadenas polipeptídicas, contínuas, no ramificadas, sin interrupción, que adoptan un plegado con una conformación única y estabilizada mediante puentes de hidrógeno. Esta conformación única sería la responsable de las propiedades de las proteínas en su estado nativo. En este trabajo también definen el estado desnaturalizado de una de proteína, caracterizado como la pérdida de la conformación nativa(Mirsky & Pauling 1931). Cuatro años más tarde, Pauling propone que los anticuerpos pueden adoptar diferentes conformaciones de energías similares para poder asumir la conformación que sea complementaria al antígeno(Pauling 1940).

En 1950 Fred Karush extendió este concepto al proponer que el estado nativo de las proteínas podría contener diferentes confórmeros con energías similares y en equilibrio dinámico para explicar la heterogeneidad de unión de la seroalbúmina(Karush 1950). Él definió como adaptabilidad configuracional al hecho de que diferentes confórmeros puedan tener diferentes afinidades para los ligandos. Este temprano trabajo fue mayormente ignorado.

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, pero con energías similares. Se consideraba que la proteína se

encontraba en un estado funcional cuando estaba plegada y no funcional cuando perdía su estructura, es decir, se desnaturalizaba.

1.2 La interacción proteína - ligando

1.2.1 El modelo de llave-cerradura de Fischer, *key-lock*

A partir de los años 50 empezaron a surgir diferentes modelos para explicar la relación entre la estructura-función de las proteínas. Pero todos estos modelos tuvieron un precedente. Probablemente el primer modelo que intentó dar una base estructural a un comportamiento observado en proteínas haya sido el propuesto por Emil Fischer y fue conocido como **key-lock** (llave y cerradura) (Fischer 1894). Motivado por reciente evidencia experimental de la alta especificidad de muchas enzimas, 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. Como muestra la Figura 1.1, considera a la estructura de la proteína como una cerradura y al sustrato como a una llave que encaja de forma perfecta en esa cerradura.

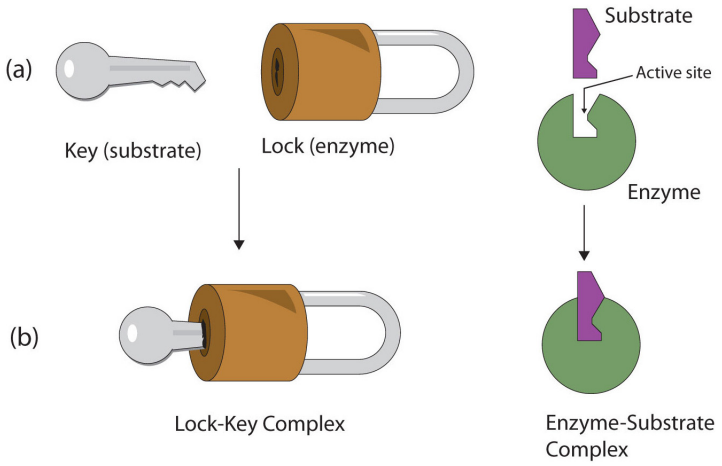

Figure 1.1: Unión a ligando según el mecanismo de *key-lock*.

1.2.2 El modelo de encaje inducido de Koshland, *induced fit*

Años más tarde, en 1958, Daniel Koshland(Koshland 1958) observó que el modelo de *key-lock* no explicaba ciertas discrepancias que se daban entre algunos tipos de reacciones químicas tales como la inhibición no competitiva y, principalmente, la unión de ligandos de diferentes tamaños y formas. De esta manera surge el modelo conocido como *induced fit* (ajuste inducido) bajo las siguientes premisas:

- La acción enzimática requiere la orientación precisa de los grupos catalíticos.
- El substrato causa un cambio apreciable en las posiciones relativas de los aminoácidos del sitio activo.
- Los cambios en la estructura de la proteína causados por la unión con el substrato posicionarán a los grupos catalíticos en la alineación adecuada, mientras que si no es el sustrato indicado, este proceso no tendrá lugar.

Con la imagen de la Figura 1.2 Koshland ilustró estos conceptos.

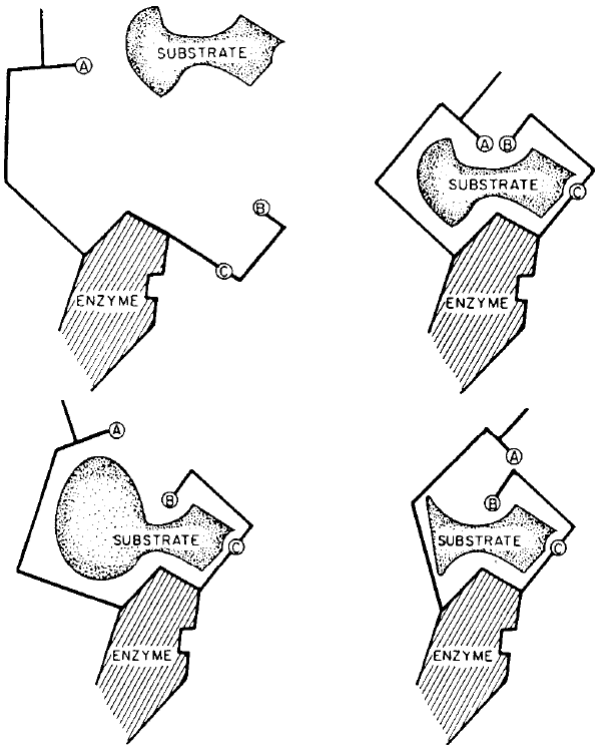

Figure 1.2: Esquema del modelo de *induced fit*.

Este modelo implica una nueva característica que el modelo de *key-lock* omitía: la flexibilidad. Para demostrar su importancia para la función enzimática, Koshland eligió la enzima fosfoglucomutasa. 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. La Figura 1.3 resume el experimento: la presencia de tioles libres titulables luego de la unión de sustrato manifiesta un cambio conformacional "inducido" por el mismo.

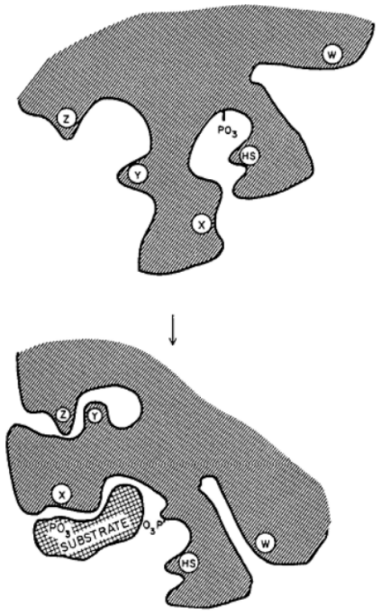

Figure 1.3: *induced fit* en la fosfomutasa. Luego de la unión al ligando, los grupos tioles **W**,**X**,**Y** y **Z** se vuelven inaccesibles para el solvente.

El modelo de *key-lock* era incapaz de explicar estos resultados experimentales. De esta manera, el modelo de *induced fit*, 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). Así fue como la idea de 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.

1.2.3 El equilibrio conformacional de Monod, *pre-equilibrium*

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 que estudió la L-treonina deaminasa (Umbarger & Brown 1957), su actividad en presencia de urea y su desnaturalización por temperatura. También en los resultados sobre la aspartato transcarbamoilasa obtenidos por Gerhart en 1962 (Gerhart & Pardee 1962). Finalmente, el conocimiento de los confórmeros de la hemoglobina obtenidos por Max Perutz(Perutz 1970) fue la última pieza. Monod empezo a considerar el estado nativo de las proteínas como un equilibrio conformacional independiente de la existencia del sustrato.

A partir de estos trabajos surge el modelo de Monod, Wyman y Changeux (MWC) de *pre-equilibrium* (preequilibrio), basado en un equilibrio entre al menos dos confórmeros preexistentes al agregado del ligando. Según este modelo, no es el ligando el que modifica el estado nativo de la proteína, generando un nuevo **confórmero**; sino que este confórmero

ya formaba parte del *pool* de confórmeros que es el estado nativo (*pre-equilibrium*). El rol del ligando, según MWC, no sería la modificación de la estructura, sino la selección del confórmero que mejor se ajusta a su estructura. Este último paso fue llamado, posteriormente, *conformational selection*, (selección conformacional). Se entiende por confórmero de una proteína a una de 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 modelo de MWC puede ser considerado como uno de los primeros en tomar en cuenta explícitamente la preexistencia de un conjunto de confórmeros en equilibrio. A diferencia del modelo de Koshland, como ya se ha dicho, en este modelo los confórmeros existen con anterioridad a la aparición del ligando, reconociéndose así la existencia de un estado nativo más complejo (Figura 1.4). Entendemos por confórmeros 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.

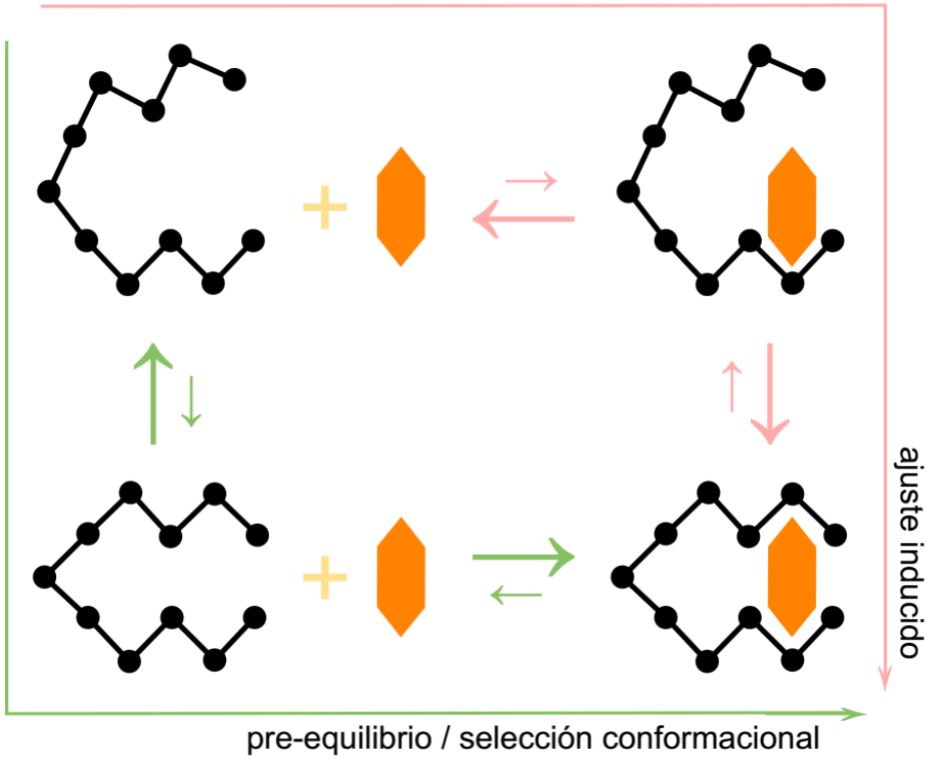

Figure 1.4: Los caminos alternativos en la unión del ligando: *induced fit* y *conformational selection*.

Gracias a los trabajos de Koshland y Monod, la dinámica de la estructura de las proteínas comenzó a tenerse en cuenta como uno de los factores determinantes de su función biológica. La mioglobina, hayada en varias células animales, se convirtió en el caso arquetípico de esta nueva visión(Kendrew et al. 1958). Esta proteína posee un grupo hemo con un átomo de hierro y su función es la de transportar oxígeno a los músculos. Esto requiere la captación y posterior liberación de oxígeno en las células somáticas. Para ello, debe ser capaz de alterar su constante de disociación de oxígeno. Son los cambios conformacionales

los que permiten este fenómeno.

Se concluye, de este apartado histórico, que el modelo del estado nativo de las proteínas siempre estuvo ligado al modelo de la interacción proteína-ligando. También se podría decir, haciendo una interpretación laxa, que los 3 grandes modelos constituidos para comprender la interacción proteína-ligando, se han mantenido hasta el día de hoy. Si bien esta interpretación debe ser particularmente liviana con el modelo de *key-lock*, el más primitivo de los 3, el fenómeno de *conformational selection* —comprendido en el modelo de MWC—, es reminiscente al modelo de *key-lock* con más de una *lock*. Dicho así, el modelo de MWC sería una alternativa que reconcilia el antiguo modelo de *key-lock* de Fischer, con los fenómenos que Koshland notó en su trabajo experimental. El modelo de este último autor también se agrega a la visión actual de la interacción proteína ligando, que combina la *conformational selection* con el *induced fit*, como ilustra la Figura 1.5(Okazaki & Takada 2008).

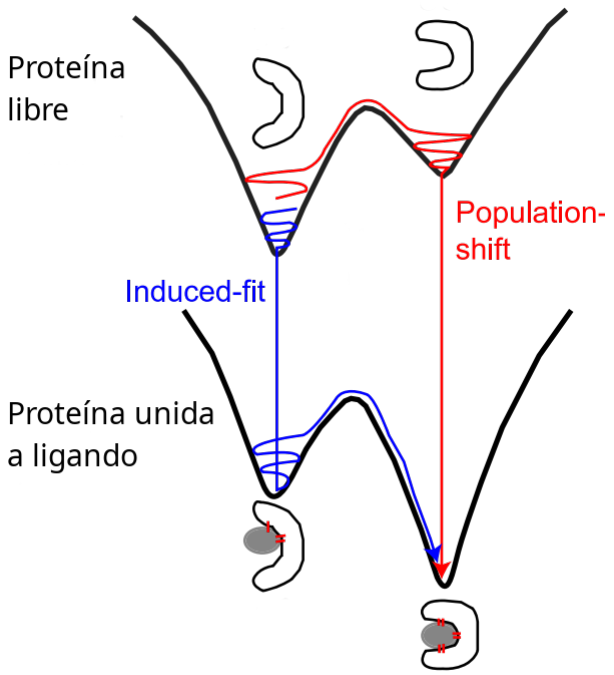

Figure 1.5: Ambos fenómenos —*induced fit* y *conformational selection*—, participan de la interacción proteína - ligando, según Nussinov.

Así como se detalló la interacción proteína-ligando, en el próximo apartado —antes de definir el estado nativo—, se desarrollará el tema del plegamiento de una proteína.

1.3 La pregunta por el plegamiento

1.3.1 La paradoja de Levinthal

Karplus separa al problema del plegamiento en 2 preguntas: la predicción de la estructura a partir de la secuencia y la cinética del proceso del plegamiento(Karplus 1997). De este último trata la llamada paradoja de Levinthal. Luego de que en 1961 Anfinsen comprobara la reversibilidad del plegamiento de las proteínas, demostrando que el estado nativo se encuentra en mínimos de energía libre(Anfinsen et al. 1961) (Anfinsen 1973), en 1969 Levinthal hizo un agregado: este proceso de búsqueda del estado nativo, no puede ser aleatorio. Dado el elevado número de grados de libertad de un polipéptido de longitud media, si éste explorara sus confórmeros aleatoriamente, el proceso de plegado duraría más que cualquier proceso biológico, cuando en realidad se sabe que las proteínas se pliegan en cuestión de milisegundos a segundos. Esto implica que las proteínas exploran un reducido conjunto de confórmeros hasta llegar al estado nativo. A este conjunto, Levinthal llamó *folding pathway*. La Superficie de Energía Potencial (SEP) de una hipotética proteína que se pliega siguiendo caminos aleatorios se encuentra en la Figura 1.6.

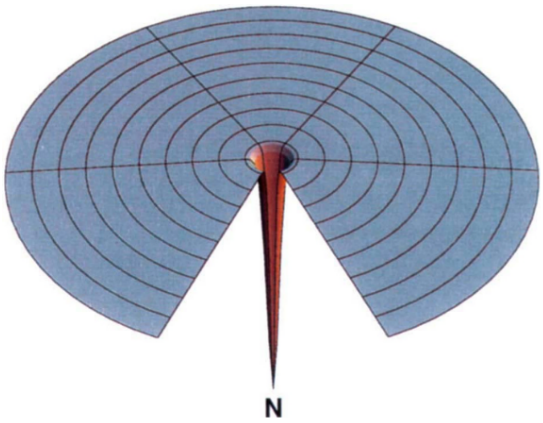

Figure 1.6: *Folding funnel* si la premisa de la paradoja de Levinthal fuera cierta. El punto **N** marca el estado nativo.

1.3.2 *Folding pathways*

En los 70 se entendió que si se encontraban los estados intermedios del proceso de plegado, se podía caracterizar al *folding pathway*. Así comenzaron los experimentos de cinética de plegado(Tsong et al. 1971). Que rápidamente evidenciarion intermediarios incorrectamente plegados que se encontrarían fuera de la ruta de plegado(Ikai & Tanford 1971). La Figura 1.7 muestra la SEP de una proteína que sigue un camino único al plegarse.

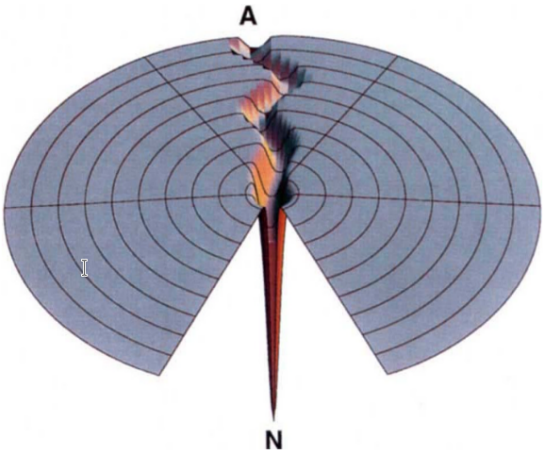

Figure 1.7: *Folding funnel* según Levinthal. **A** marca la estructura de partida que sigue una única ruta de plegamiento, con sus intermediarios, hasta la estructura nativa.

Por otro lado, mientras los experimentos eran incapaces de obtener un detalle atomístico de estos confórmeros intermedios, y algunos investigadores se preguntaban si el camino de plegado realmente era único (Harrison & Durbin 1985), Wolynes y sus colaboradores construían el marco teórico que domina hasta hoy.

1.3.3 *Folding funnels*

Wolynes y sus colaboradores, reemplazan el concepto de un camino único de eventos secuenciales (*pathway*) por uno de eventos paralelos. Entiende al plegado como un proceso de difusión que puede tomar caminos alternativos y aún así llegar a la misma estructura nativa (Bryngelson & Wolynes 1987). Este proceso se expresa con un *folding funnel* (embudo de plegado), como el que se muestra en la Figura 1.8 (Bryngelson & Wolynes 1989).

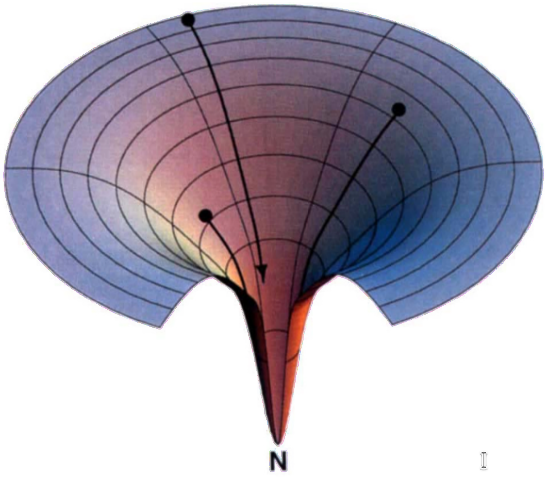

Figure 1.8: *Folding funnel* simplificado de Wolynes. La energía libre de estructura inicial puede tomar numerosos valores posibles, y el camino de plegamiento —es decir, sus distintos intermediarios—, también puede variar; pero a medida que se van plegando, la fracción de contactos nativos aumenta y finalmente coinciden en la misma estructura nativa.

Al igual que el *folding pathway*, el *folding funnel*, "resuelve" la supuesta paradoja de Levinthal al mostrar que conformaciones previas condicionan a las futuras, reduciendo el universo de confórmeros posibles; pero el embudo agrega también, entre otras cosas, la posibilidad de que la proteína explore varios caminos alternativos en simultáneo. Por otra parte, al considerar numerosos puntos de inicio en el proceso de plegado, a diferencia del modelo de *pathway*, entiende al estado no plegado como una multitud de estados posibles. Esta multiplicidad de puntos de partida y caminos posibles se reduce a medida que avanza el plegamiento. El embudo es la mejor forma de representar esta reducción del número de conformaciones posibles que la proteína puede alcanzar en condiciones y tiempos biológicos. A su vez, el plegado ocurre, simultáneamente, en varios puntos de la cadena y estos plegamientos parciales son capaces de generar nuevas interacciones favorables. Éste último punto está basado en los resultados de Go (Go & Abe 1981), quien observó que en la estructura nativa no hay compromisos: las primeras interacciones favorables dan lugar a estructuras secundarias, que a su vez tienen interacciones favorables entre ellas.

Ahora bien, la Figura 1.8 es una simplificación de la más precisa Figura 1.9. La primera implicaría un proceso de plegado totalmente cooperativo, prácticamente instantáneo, de un solo paso, en donde la proteína forma sus contactos nativos sin detenerse hasta llegar a su estructura nativa.

El verdadero proceso de plegado implica la formación de contactos transitorios que luego se rompen para formar contactos nativos. O más aún, contactos nativos prematuros que deben romperse para luego formarse. Esto ocurre hasta en la proteína de más rápido plegado, la caja de triptófano. Una proteína de 20 residuos con el tiempo de plegado más rápido conocido (4μs), pero que aún así forma, prematuramente, un puente salino que da lugar a su principal intermediario de plegado. Este puente salino debe romperse, para luego volver a formarse en la estructura nativa final (Zhou 2003).

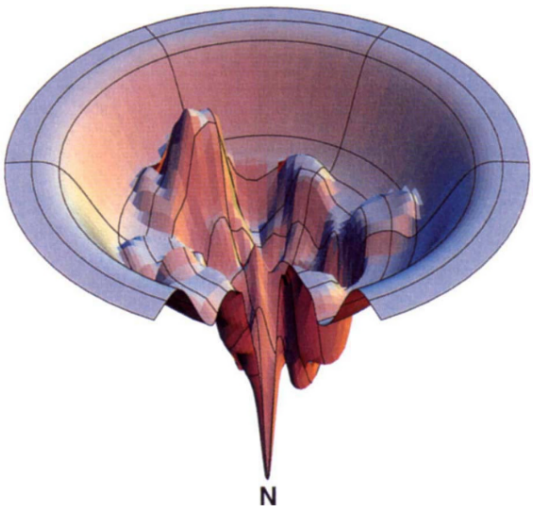

Figure 1.9: *Folding funnel* según Wolynes. Los máximos implican estados de transición entre intermediarios de plegado. También existen valles de intermediarios particularmente estables pero que aún no han terminado su plegamiento.

1.4 El estado nativo, el fondo del embudo

Los trabajos de Wolynes y sus colaboradores, enfocados en el problema del plegamiento proteico que buscaban resolver la paradoja de Levinthal (Dill & Chan 1997), junto a los estudios de la interacción proteína-ligando constituyen la base de la definición actual del estado nativo. Y a diferencia del pasado, no es tanto la bioquímica la que define el estado nativo, sino la termodinámica y la mecánica estadística.

El embudo de plegado, no es un embudo llano, sino más bien "fractal". Así como durante el proceso de plegado se encuentran embudos montados sobre otros embudos, es decir, el plegado no es perfectamente cooperativo, sino que existen estados intermedios e intermadiarios; luego del plegado, a la proteína la esperan otros embudos. Es decir, nuevos estados intermedios y nuevos intermediarios, sólo que a estos intermediarios los llamamos confórmeros. En condiciones biológicas de funcionamiento normal —plegadas y en equilibrio termodinámico—, las fluctuaciones energéticas de la proteína, sumadas a las del agua, llevan a la proteína a alternar estos confórmeros en distintas escalas de tiempo. La cinética de estos **cambios conformacionales** está determinada por las energías de activación de estos procesos. Esta es la *Energy Landscape Theory* y esto que, en un exceso de informalidad, llamamos embudos, es el paisaje energético de la proteína que predice el conjunto de confórmeros en equilibrio dinámico que serán explorados. A este conjunto de conformeros nos referiremos como *native ensemble*, o más simplemente, estado nativo (Onuchic et al. 1994) (Bryngelson et al. 1995) (Frauenfelder et al. 1991) (Tsai et al. 1999) (Wei et al. 2016). El grado de diversidad conformacional estará determinado por la distribución y la altura de las barreras energéticas entre confórmeros, es decir, por la topología del fondo del "embudo energético". En cualquier momento dado, las moléculas

de la proteína se encontrarán en distintos confórmeros y la distribución de sus poblaciones estará determinada por las estabilidades de los confórmeros. La Figura 1.10, tomada de (Wei et al. 2016), ilustra "el fondo del embudo".

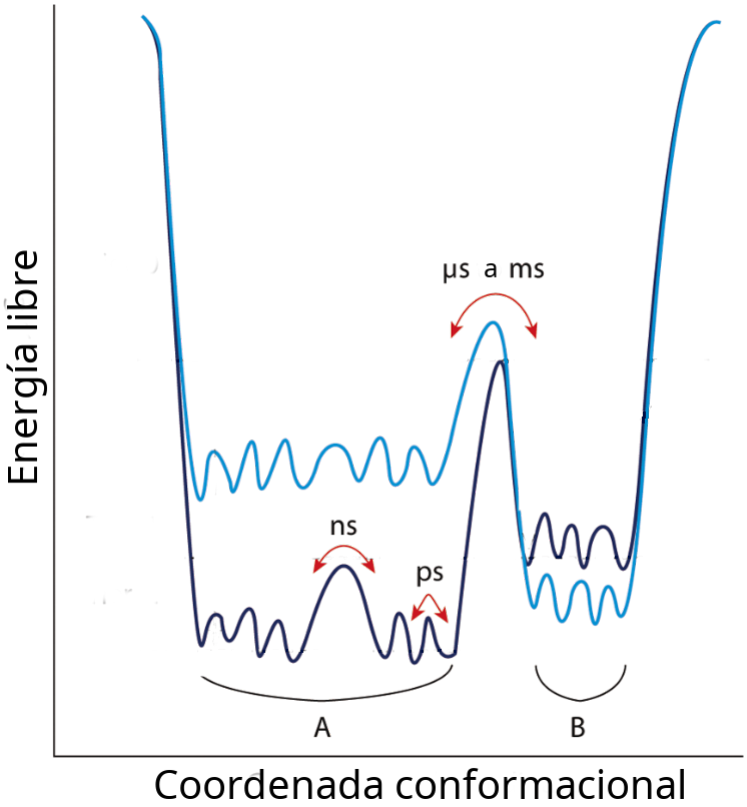

Figure 1.10: Paisaje conformacional de una proteína. Numerosos confórmeros están separados por barreras de energía libre que implican una constante de tiempo para su intercambio. **A** y **B** marcan 2 confórmeros distintos. Se entiende que cada confórmero tiene una cierta flexibilidad y admite fluctuaciones en su estructura.

1.5 Diversidad conformacional y función proteica

La idea inicial de paisajes energéticos estáticos, fue extendida posteriormente a la noción de paisajes energéticos dinámicos para incluir los efectos del solvente o la presencia de otras moléculas (Figura 1.10) (Kumar et al. 2008). Esta última idea era la necesaria para terminar de combinar los trabajos de Wolynes con el modelo de MWC y el *induced fit* de Koshland, ya que permite explicar la súbita presencia de el/los confórmero/s asociado/s al ligando. El proceso modelo sería el siguiente: en una población de confórmeros, el confórmero de mayor afinidad al ligando no sería necesariamente el de mayor estabilidad. Lo que explicaría su relativa ausencia en el estado nativo. Pero si la superficie de energía libre es dinámica, la unión del ligando podría alterarla, estabilizando a esta conformación de mayor afinidad. Así, esta conformación se volvería la mayoritaria. Esta última adenda al panorama actual fue llamada *population shift* y está caracterizada en la Figura 1.8, tomada de (Tsai & Nussinov 2014).

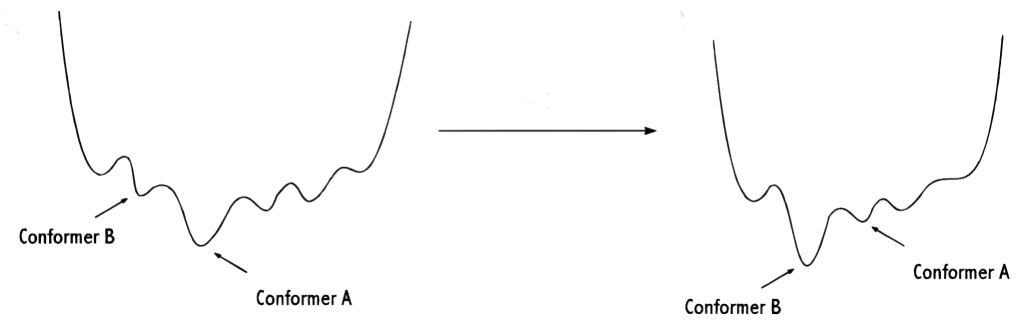

Figure 1.11: La visión termodinámica aportada por Wolynes, la multiplicidad conformacional de MWC y el efecto del ligando descripto por Koshland. Todos estos fenómenos son usados en la idea de *population shift*.

Todos estos conceptos sobre el estado nativo resultan fundamentales para la comprensión de procesos como el cooperativismo y el alosterismo (Tsai & Nussinov 2014) (Cui & Karplus 2008), la función biológica (Eisenmesser et al. 2005), el reconocimiento molecular (Boehr et al. 2009) (Nussinov et al. 2013), la catálisis enzimática y el origen de nuevas funciones, entre otros ejemplos (Tokuriki & Tawfik 2009).

1.6 Desorden en proteínas

Hoy en día, un capítulo sobre el estado nativo de las proteínas que no mencione a las proteínas intrínsicamente desordenadas (*Intrinsically Disordered Proteins*, IDPs), es un capítulo incompleto. Si bien no fueron objeto de estudio de esta tesis, cabe aclarar que el desorden en proteínas modificó la relación estructura-función y lo que se entiende como estado nativo. Estas proteínas poseen largos segmentos desplegados y funcionales, lo que antes se pensaba era una contradicción. No tienen estructura tridimensional definida, su ensamble nativo es amplio y de rápido intercambio y aún así son funcionales. Más aún, esta flexibilidad conformacional es fundamental para su función.

Chapter 2

Dinámica Molecular

Habiendo establecido la noción de diversidad conformacional de la proteína y que estos confórmeros son esenciales para su función, se introducirá uno de los principales métodos computacionales para hallar estos confórmeros y así poder estudiar su función: la dinámica molecular. Como se verá, este método produce una gran cantidad de información, de la cual buena parte es poco informativa. En el próximo capítulo se ahondará en uno de los tantos métodos creados para analizar dinámicas moleculares, el análisis de los movimientos esenciales o PCA, por sus siglas en inglés. Como los modos normales, este método resulta en un conjunto de vectores que representarán los movimientos biológicamente relevantes de una proteína.

2.1 Introducción histórica

La dinámica molecular (MD por sus siglas en inglés) es un conjunto de técnicas computacionales creadas para simular el comportamiento de átomos y moléculas basándose en las leyes de la física. Comenzó a formarse a fines de los años 50, en el campo de la física teórica, con el estudio de las interacciones entre esferas rígidias (Alder & Wainwright 1957). Este trabajo fue extendido para estudiar líquidos simples, comenzando por los gases nobles (Rahman 1964), para luego ser aplicado en el campo de las ciencias materiales. Pero antes de eso, el mismo grupo que simuló al Argón, publicó la primer simulación con interacciones moleculares, Stillinger simuló agua líquida en 1971 (Rahman & Stillinger 1971) (Stillinger & Rahman 1974). La primera proteína en ser simulada por dinámica molecular sería la *Bovine Pancreatic Trypsin Inhibitor* (BTI) en 1977 (McCammon et al. 1977). La proteína fue simulada en fase gaseosa por sólo 3ps. La Figura 2.1 fue extraída de esta publicación.

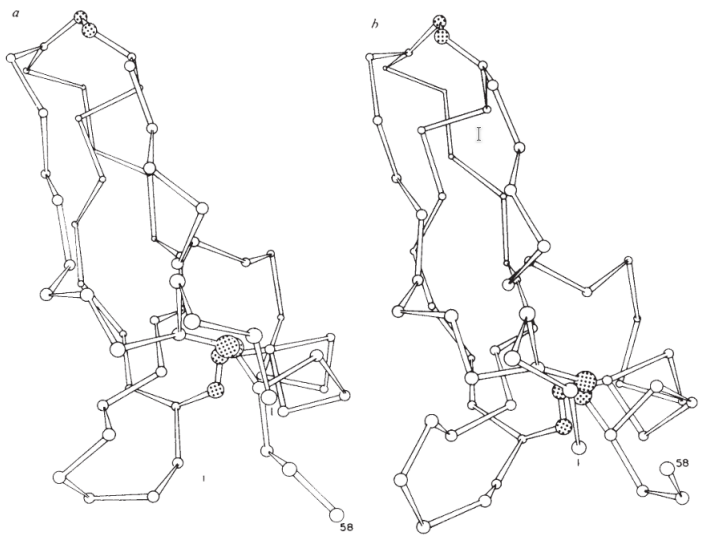

Figure 2.1: Las 2 estructuras corresponden al *backbone* de la BTI antes y después de la simulación.

La base teórica de la MD se apoya en la mecánica clásica y la mecánica estadística de Boltzmann. Es la energía térmica la que mueve a los átomos de la proteína y son las fórmulas de la mecánica newtoniana las utilizadas para predecir estos movimientos. El resultado es un conjunto de estructuras (conformaciones, *snapshots*) a lo largo del tiempo, que representan la dinámica de la proteína. Y por la hipótesis ergódica se asume que el promedio de esta distribución de estructuras a lo largo del tiempo, representa el promedio del ensamble de estructuras de la proteína en solución. Esto presenta la primer dificultad de la MD: para satisfacer las condiciones de esta hipótesis, se debe simular al sistema por el período de tiempo suficiente como para explorar todo el espacio conformacional.

Por su pretensión de predecir el comportamiento futuro de un sistema basándose solamente en el conocimiento previo y en las leyes de la física, se dice que la MD intenta realizar la visión Laplaciana de la mecanica newtoniana. Pierre-Simon de Laplace teorizó, en 1820, que quien conociera todas las "posiciones y fuerzas que animan a los entes del universo", sería capaz de calcular la posición de todas estas entidades, en el pasado y el futuro. Pero es sabido que este sueño ya estaba muerto incluso antes del nacimiento de la MD.

2.2 Métodos clásicos y métodos cuánticos

Fueron Poincaré y Heisenberg con el problema de los tres cuerpos y el Principio de Incertidumbre los que terminaron con el sueño Laplaciano de encontrar una conexión inequívoca y analítica entre el pasado y el futuro. Estas son las principales limitaciones de la MD, que se agravan cuando esta se aplica a sistemas complejos como lo son las proteínas en solución. Los efectos caóticos son inevitables y pueden hacer a las simulaciones de MD extremadamente sensibles a las condiciones iniciales, lo que garantiza la irreproducibilidad

de la corrida de simulación (Braxenthaler et al. 1997) (Elofsson & Nilsson 1993). Idealmente, se debería calcular la energía potencial de la molecula mediante su Hamiltoneano completo (incluyendo núcleos y electrones). En la práctica, el altísimo costo computacional de los métodos cuánticos limita su aplicación a sistemas pequeños. La mecánica molecular (MM), también llamada método de campo de fuerza empírico, calcula la energía potencial de un sistema ignorando el movimiento de los electrones y sólo utiliza las coordenadas nucleares (una extensión de la aproximación de Bohr-Oppenheimer). Para esto emplea ecuaciones sencillas con parámetros ajustables, lo que permite tratar sistemas con un número de átomos considerable. La MM es empleada para calcular energías de interacción, geometrías en equilibrio y frecuencias vibracionales. Si bien también existen métodos híbridos de mecánica cuántica/clásica (*QM/MM*, por sus siglas en inglés), las "regiones cuánticas" se limitan a pequeñas zonas del sistema, en donde se preveen rotura y formación de enlaces, o se estima que el efecto de la polarización es determinante. Este último fenómeno es modelado en algunos campos de fuerza (Albaugh et al. 2016), pero la mayoría de los trabajos existentes de MD no tienen en cuenta los efectos de la polarización de lo átomos, ya que esto también acarrea un significativo costo computacional (Cui & Karplus 2003) (Guvench & MacKerell 2008).

2.3 Fundamentos de la Mecánica Molecular

2.3.1 Campos de fuerza

Los campos de fuerza de mecánica molecular clásica obtienen el valor en una superficie de energía potencial de una conformación utilizando funciones simples de la forma:

$$E = E_{enlazados} + E_{no\ enlazados} \tag{2.1}$$

La definición exacta de las funciones de energía potencial de enlace y no enlace varían según el campo de fuerza específico. (Allen & Tildesley 2017) los clasifica en 3 clases, la primer clase es la utilizada en esta tesis y la de mayor aplicación. Estos campos de fuerza modelan a las interacciones entre pares de átomos y a éstos como esferas rígidas. La segunda clase de campos de fuerza altera lo primero: añade términos anarmónicos a las interacciones entre átomos al considerar el efecto que una interaccion tiene sobre otra. Mientras que la tercera clase de campos de fuerza tiene en cuenta la deformación de la nube electrónica que los átomos sufren por su entorno. Esto mejora la reproducción de las interacciones electrostáticas. Ejemplos de las 3 clases se encuentran en la Tabla 2.1

Table 2.1: Extraída de (Allen & Tildesley 2017).

Campo de fuerza	Clase	Campo de aplicación
CHARMM22	1	Proteínas en agua
AMBER ff99	1	péptidos, moléculas orgánicas
GAFF	1	moléculas orgánicas
COMPASS	2	moléculas pequeñas, polímeros
AMBER ff02	3	Átomos polarizables
AMOEBA	3	Multipolos polarizables, diseño de drogas (Ponder et al. 2010)

Las siguientes ecuaciones representan a la generalidad de los campos de fuerza de la primera clase (P. Kollman, R. Dixon, W. Cornell, T. Fox, C. Chipot 1997) (Frenkel & Smit 1996).

El término de los átomos enlazados se compone de otros 3 términos:

$$E_{enlazados} = E_{covalente} + E_{angulo} + E_{dihedro} \tag{2.2}$$

A continuación, se describen estos 3 términos. Las interacciones covalentes son aproximadas como interacciones armónicas:

$$E_{covalente} = \sum_{enlaces} K_x (x - x_0)^2 \tag{2.3}$$

Donde:

- K_x: constante de fuerza de la interacción entre partículas i y j
- x: longitud del enlace actual entre partículas i y j
- x_0: longitud de equilibrio del enlace entre partículas i y j

Este potencial armónico reemplaza, y aproxima, al potencial de Morse de un enlace equivalente, como se ve en la Figura 2.2. Existen aproximaciones de órdenes superiores, pero el aumento de precisión no compensa el aumento de coste computacional que implican, por lo que no suelen utilizarse.

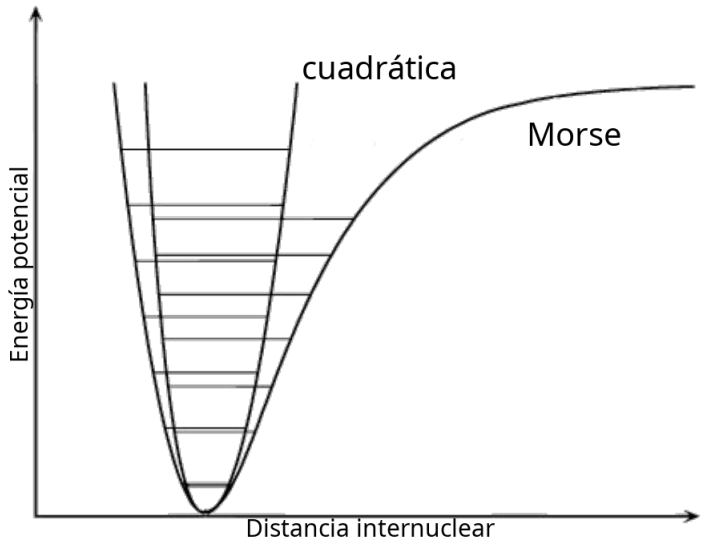

Figure 2.2: Aproximación cuadrática al potencial de Morse. Así se previene la ruptura de enlaces y se acelera el cálculo energético para las posiciones cercanas al equilibrio.

Lo mismo sucede con los ángulos entre 3 partículas:

$$E_{angulo} = \sum_{angulos} K_\theta(\theta - \theta_0)^2 \tag{2.4}$$

Donde:

- K_θ: constante de fuerza del ángulo entre partículas i y j
- θ: ángulo actual entre partículas i, j y k
- θ_0: ángulo de equilibrio entre partículas i, j y k

Los ángulos dihedros, o torsionales, son los ángulos que se forman entre los 2 planos que forman 4 partículas, o 2 enlaces. En este grupo entran los ángulos ϕ y ψ. Para representar el cambio en la energía potencial por los ángulos dihedros, se usan términos de Fourier.

$$E_{dihedro} = \sum_{dihedros} V[1 + cos(n\chi - \sigma)] \tag{2.5}$$

Donde:

- V: altura máxima de la barrera de energía
- n: multiplicidad
- χ: ángulo dihedro entre partículas i, j, k y m
- σ: ángulo de fase

La eq. 2.5 utiliza una función trigonométrica para representar los cambios en la estabilidad

de un compuesto al rotar sus ángulos torsionales, lo que en la isomería conformacional se denomina como posiciones anti, *gauche*, eclipse. La Figura 2.3 ilustra esto.

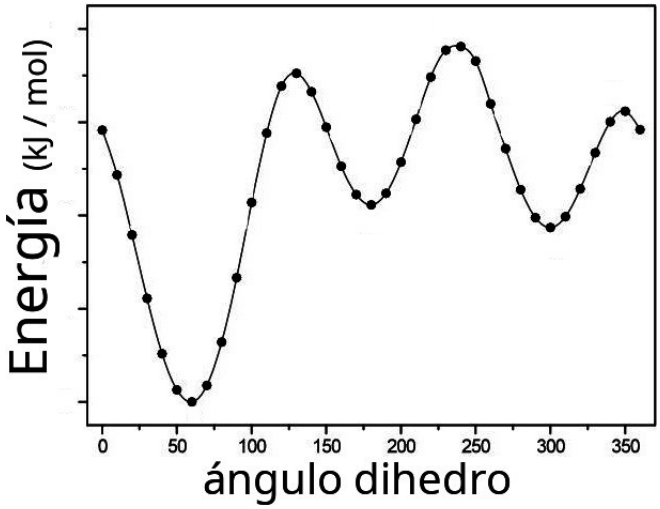

Figure 2.3: Función de energía potencial modelo para un ángulo dihedro. Éste puede ser un ángulo ϕ, ψ o los dihedros de cadenas laterales χ

Por otro lado, las interacciones de no enlace representan a las interacciones a distancia. Éstas se calculan entre pares de átomos que se encuentran en distintas moléculas o en la misma molécula pero separados por, al menos, tres enlaces. El primero de ellos es el término de las interacciones de Van der Waals y el segundo es el electroestático.

$$E_{no\ enlazados} = E_{VanderWaals} + E_{electroestaticas} \tag{2.6}$$

Una suma de términos de Lennard-Jones para representar atracciones y repulsiones de corto alcance:

$$E_{VanderWaals} = \sum_{pares\ ij\ no\ enlazados} \left(\frac{A}{r_{ij}^{12}} - \frac{B}{r_{ij}^{6}} \right) \tag{2.7}$$

Donde:

- A: término de rechazo. $A = 4\epsilon\sigma^{12}$
- B: término de atracción. $B = 4\epsilon\sigma^{6}$
- ϵ: mínimo de energía para la interacción entre los partículas i y j según potencial de Lennard-Jones
- σ: distancia finita a la que se anula la interacción entra las partículas i y j según potencial de Lennard-Jones

Y una suma de términos de Coulomb, representando interacciones electrostáticas de largo alcance. Se las llama así por que a diferencia de las interacciones de Van der Waals,

el exponente de su cociente es lineal, lo que representa una caída de su magnitud más paulatina con respecto a su distancia.

$$E_{electroestaticas} = \sum_{pares\ ij\ no\ enlazados} \left(\frac{q_i q_j}{r_{ij}} \right) \tag{2.8}$$

Donde:

- $q_{i/j}$: carga parcial de la partícula i o j
- r_{ij}: distancia entre las partículas i o j

Estas sumatorias no abarcan la totalidad de átomos del sistema, sino que para ambas interacciones existen distancias máximas (*cutoffs*) a partir de las cuales se anulan estas interacciones. Una razón para hacer esto es la de bajar el costo computacional, pero otra razón aún más importante se verá en el siguiente apartado.

Todos estos parámetros son ajustados para reproducir mediciones experimentales (Monticelli & Tieleman 2013). La Figura 2.4 grafica las interacciones vistas:

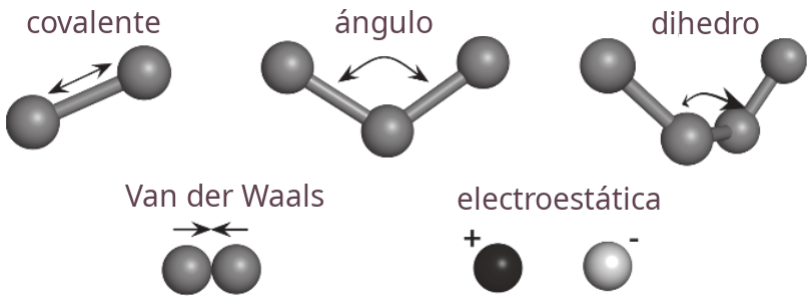

Figure 2.4: Ilustración de las interacciones repasadas. Las 3 interacciones de enlace y las 2 de no enlace que consisten en atracción y repulsión de Lennard-Jones y culómbica.

Nótese también que se hace referencia a "partículas" y no átomos. Esto se debe a que existen campos de fuerza de grano grueso (CG, por sus siglas en inglés) que se abstraen del átomo y representan a las moléculas en términos de partículas, que agrupan varios átomos. Su objetivo es reducir los tiempos de cálculo y captar nuevas conformaciones que resultan muy difíciles de alcanzar con modelos atomísticos por las altas barreras de energías que deben ser superadas. En el siguiente capítulo se introducirá uno de estos modelos CG utilizado, no para MD, sino para el cálculo de modos normales (NMA, por sus siglas en inglés).

2.3.2 Condiciones periódicas de contorno

Así como se intenta aproximar la química de las interacciones biológicas, también se intenta replicar las condiciones *in vivo* o *in vitro* al momento de simular una proteína. En principio, esto implicaría un alto número de particulas en nuestro sistema, lo que aumentaría exponencialmente el tiempo de ejecución de la dinámica. Por lo tanto se busca utilizar la menor cantidad de átomos posibles para realizar la simulación. Existe un límite a cuán pequeño puede hacerse un sistema dado que la simulación explícita de una proteína en un solvente requiere de un número mínimo de moléculas que garantice la correcta simulación de sus propiedades termodinámicas como líquido. Si el sistema se hiciera demasiado pequeño, la mayoría de las moléculas del solvente se encontrarían en la superficie del sistema, donde no experimentarán las mismas fuerzas que experimentarían si estuvieran en el seno del líquido. La solución a este problema consiste en utilizar condiciones periódicas de contorno (PBC, por sus siglas en inglés) (Born and von Karman, 1912).

En esta técnica se supone que el cubo que contiene al sistema, la celda primaria, está rodeado por réplicas exactas de sí mismo en todas las direcciones, las celdas imágenes, formando una red infinita. Estas celdas imágenes contienen los mismos átomos que la celda primaria y, durante una simulación, cada uno de los átomos de las celdas imágenes se mueve de la misma forma que los átomos de la celda primaria. Así, si un átomo de la celda primaria la abandona por una de sus caras, su imagen de la cara opuesta entrará en la celda primaria. De este modo ya no existen superficies limitantes de sistema y podemos imaginar a la celda primaria replicada periódicamente en todas las direcciones formando una muestra macroscópica de la sustancia de interés. La Figura 2.5 ilustra este concepto.

Figure 2.5: Ilustración de las Condiciones Periódicas de Contorno en 2 dimensiones, en una caja cuadrada y con sus celdas imágenes más próximas.

Si bien la caja cúbica es la más utilizada en simulaciones computacionales, existen otras geometrías como la de dodecahedro rómbico y octahedro truncado (Figura 2.6). Esta

última es la geometría más utilizada en simulaciones de proteínas por aproximar mejor su geometría (exceptuando los casos de proteínas en membranas, donde aún se utilizan cajas cúbicas). En comparación con la caja cúbica, el octahedro truncado necesita menos solvente para lograr la misma distancia entre las imágenes del soluto. Por lo tanto, reduce el costo computacional de la simulación.

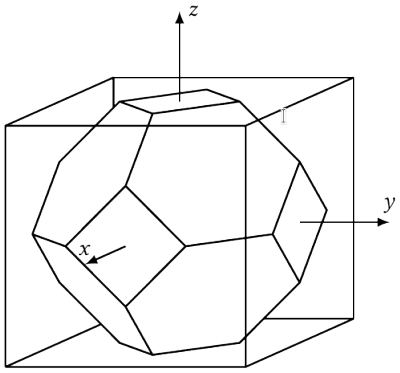

Figure 2.6: Ilustración de la caja de octahedro truncado y su cubo contenedor. Esta es la geometría de caja utilizada en esta tesis.

Si el *cutoff* para el cálculo de interacciones de no enlace era algo deseable, con la técnica de PBC esto se vuelve una necesidad, ya que no utilizarlo implicaría el cálculo de infinitas interacciónes por existir, potencialmente, infinitas imágenes. Para evitar esto, es necesario truncar los potenciales.(Allen & Tildesley 2017)

2.3.2.1 Truncamiento de potenciales de corto alcance

Utilizando el ejemplo de la Figura 2.7, se puede definir:

2.3.2.1.1 Imagen mínima: considerar la partículas 1 que se encuentra en el centro de una región con el mismo tamaño y misma geometría que la caja de simulación. la mólecula 1 interaccionará sólo con las partículas cuyos centros estén dentro de esta región. Esta metodología se denomina Convención de la Imagen Mínima (Metropolis et al. 1953). Dentro de la región de imagen mínima se encontrarán las partículas de posible interacción, pero sólo las que se encuentren dentro de una distancia de *cutoff* serán tenidas en cuenta. Este *cutoff* será menor a la mitad del tamaño de la celda para evitar que las partículas interactuen consigo mismas. Para simulaciones de macromoléculas, este *cutoff* varía entre 7Å y 14Å.

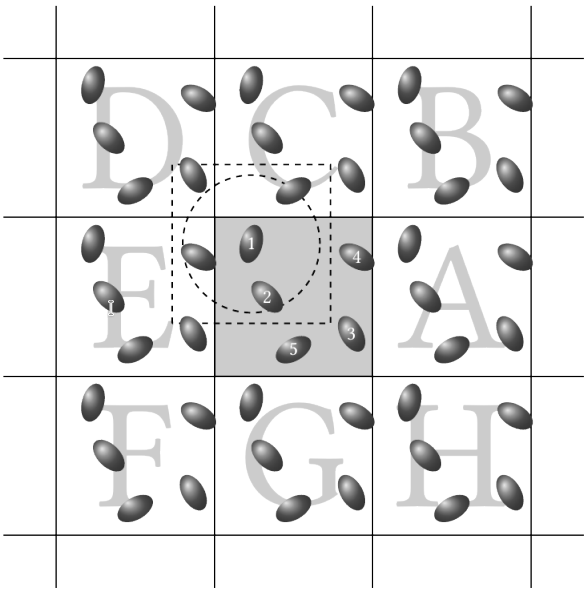

Figure 2.7: Imagen mínima en 2 dimensiones con caja cuadrada. Las cajas imagen son identificadas con letras A-H. La imagen mínima limita las partículas interactuantes posibles a N-1, donde N es el número de partículas. La partícula 1 interactuará con las partículas 2 (de su propia caja), 4E y 5C

Todo esto es aplicable en interacciones de no enlace de corto alcance (Van der Waals), pero no para las de largo alcance (electrostáticas), ya que su truncamiento introduciría demasiado error en el cálculo. Estas interacciones son aproximadas con el método de Ewald que reemplaza la sumatoria original de la eq. 2.8 —que se volvería infinita—, por otras 2 sumatorias que convergen más rápidamente, lo que permite su aproximación.(Frenkel & Smit 1996)

2.3.3 Algoritmo básico de la Dinámica Molecular

La trayectoria de un conjunto de partículas se determina resolviendo numéricamente el sistema de ecuaciones diferenciales obtenido al aplicar la segunda ley de Newton para cada partícula. Se empieza con el conjunto de coordenadas de las partículas y un campo de fuerza. Con estas 2 entradas se determina el potencial del sistema en ese punto. Luego se deriva este potencial para obtener las fuerzas sobre cada partícula:

$$F_i = \frac{\delta V(r_1, ..., r_N)}{\delta r_i} \qquad i = 1, 2, ..., N \tag{2.9}$$

Con la fuerza sobre cada partícula se pueden obtener las posiciones actualizadas:

$$F_i = m_i a_i = m_i \frac{\delta v_i}{\delta t} = m_i \frac{\delta^2 r_i}{\delta t^2} \qquad i = 1, 2, ..., N \tag{2.10}$$

Donde:

- m_i: masa de la iésima partícula
- r_i: posición de la iésima partícula
- v_i: velocidad de la iésima partícula
- a_i: aceleración de la iésima partícula
- F_i: fuerza ejercida en la iésima partícula

A la resolución de estas ecuaciones se denomina, paso de integración. Existen casos sencillos como el de una partícula sujeta a una fuerza constante o el de un oscilador armónico, en donde la solución es analítica. Sin embargo, en la MD, los sistemas son más complejos por lo que es necesario recurrir a algún algoritmo para hallar numéricamente las trayectorias de cada una de las partículas que componen el sistema. La forma habitual de resolver este tipo de ecuaciones para sistemas grandes es utilizando los denominados métodos de diferencias finitas(Allen & Tildesley 2017) reduciéndose el problema a uno de condiciones iniciales tal que, dadas las posiciones y velocidades a un tiempo t, debemos obtenerlas a un tiempo $t + dt$. La base del método consiste en sustituir el intervalo de tiempo infinitesimal dt por un intervalo finito Δt, durante el cual se supone que las fuerzas que actúan sobre las partículas son constantes. De este modo, las ecuaciones de movimiento se resuelven paso a paso, integrándolas a cada intervalo Δt. La Figura 2.8 resume este proceso.

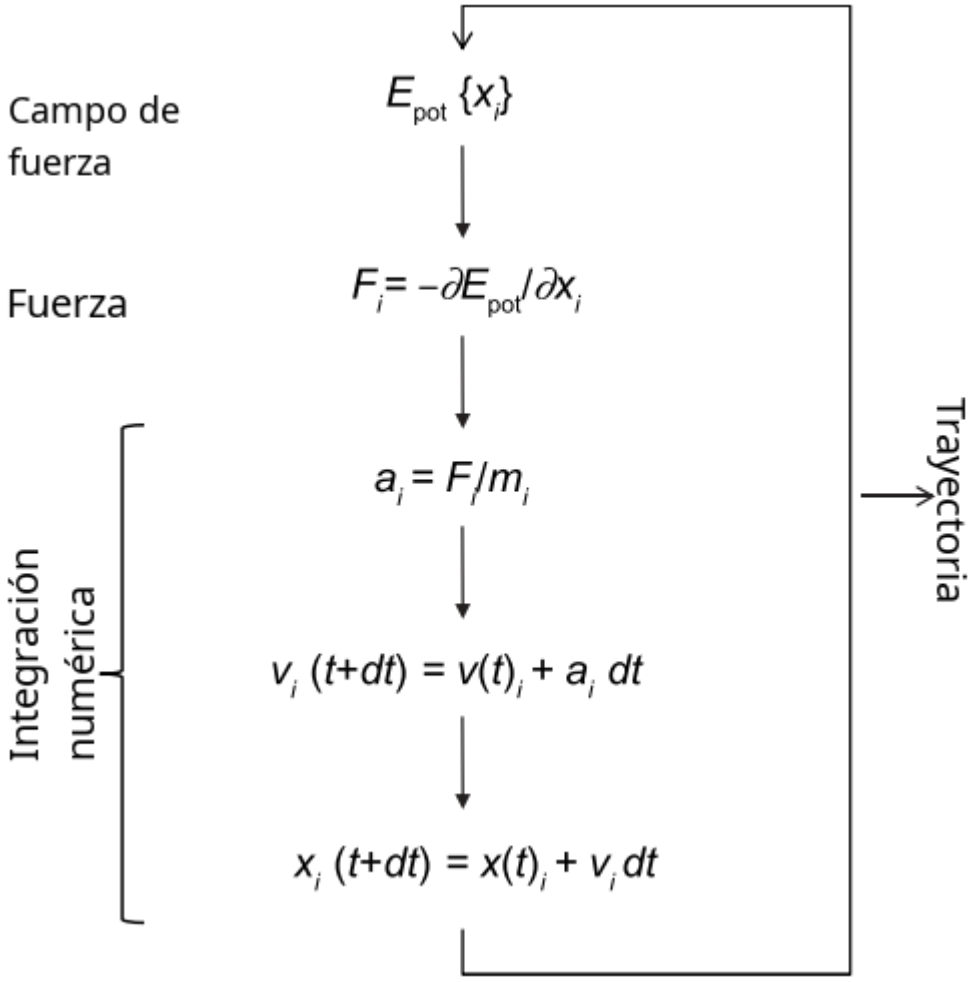

Figure 2.8: Diagrama de flujo de la MD. El primer paso es el cálculo del campo escalar de la energía potencial del sistema. Derivando este campo se obtienen las fuerzas en las 3 dimensiones para cada partícula, que según su masa percibirá cierta aceleración. Dado un paso de tiempo, la velocidad de la partícula en el siguiente paso será actualizada, así como su posición, la cual afectará el cálculo de la energía potencial en el siguiente paso de integración.

Este método es determinista (dentro de su propia teoría y más allá del margen de incertidumbre que puede haber en los cálculos) ya que si se saben las posiciones y las velocidades en un momento dado, se puede predecir el comportamiento futuro y pasado del sistema (P. Kollman, R. Dixon, W. Cornell, T. Fox, C. Chipot 1997). Al encontrar la trayectoria de cada una de las partículas, se describe cómo cambia en el tiempo, la posición y la velocidad de cada una de ellas. También es posible, a partir de los valores promedios, obtener datos de diferentes magnitudes termodinámicas (Leroy 2013).

2.3.4 Condiciones iniciales y preparación del sistema

Para llevar a cabo el proceso descripto en la sección anterior se debe realizar una serie de pasos previos. La ya mencionada caja de simulación debe ser creada, su geometría y su tamaño deben ser definidos. En una caja demasiado pequeña, la proteína podría sufrir un cambio conformacional y salir de la misma; una caja demasiado grande implicará un exceso

de solvente y un alto costo computacional. Para proteínas en agua, una separación de 10Å es la regla general (Leach 2001). Una vez decidido el tamaño de la caja, los programas de MD agregan el solvente, agua en este caso, sin buscar una configuración de equilibrio para sus moléculas. Si la MD comenzara, esta configuración inicial inestable provocará la aparición de valores locales de energía de interacción muy elevados, lo que inducirá una situación física irreal, de elevadas fuerzas locales y, consecuentemente, trayectorias irreales. Lo que suele considerarse una "explosión". Para solucionar esto está la minimización.

2.3.4.1 Minimización

La función de la minimización es la de encontrar un mínimo local en la superficie de energía potencial del sistema. La minimización comienza con las partículas estáticas e itera la busqueda un mínimo de energía del sistema (Leach 2001). El método de búsqueda es lo que diferencia a los distintos protocolos de minimización. Los 2 más populares para la mecánica molecular son el método de *steepest descent* (gradiente) y *conjugate gradient* (gradientes conjugados).

2.3.4.1.1 ***steepest descent*** simplemente mueve al sistema en la dirección del gradiente de la superficie de energía potencial. Es decir, en la dirección de la fuerza. Es útil al principio de la minimización, suele reducir la energía en unos pocos pasos, pero luego se encierra en mínimos locales, las sucesivas iteraciones no afectan mucho al potencial y no permite explorar otros mínimos cercanos, como muestra la Figura 2.9:

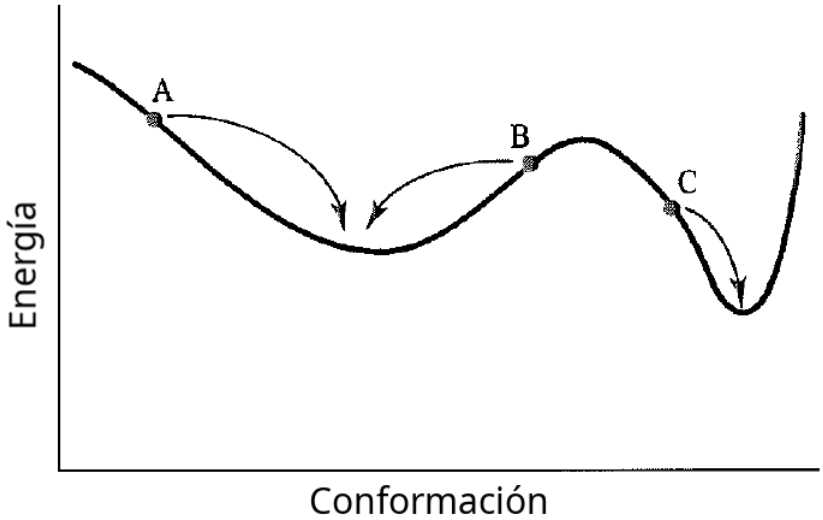

Figure 2.9: Representación cualitativa del paisaje conformacional de un sistema de proteína y agua. Si la configuración inicial se encontrará en las coordenadas **A** o **B**, el mínimo global no sería alcanzado.

2.3.4.1.2 ***conjugate gradient*** comienza su primer paso siguiendo el gradiente del sistema para la configuración actual (al igual que el método de gradiente), pero a partir

del segundo paso en adelante, cada nuevo paso estará determinado por una combinación lineal entre el gradiente del paso anterior y el actual. Este método suele lograr mayores estabilizaciones del sistema una vez que el protocolo anterior se estabiliza (Braun et al. 2019).

Así, la minimización lleva al sistema a una configuración a partir de la cual se pueden integrar las ecucaciones de movimiento y obtener la trayectoria. Pero además de posiciones, son necesarias también las velocidades.

2.3.4.2 Calentamiento: asignación de velocidades

Calentar el sistema hasta llevarlo a temperaturas en el rango biológico consiste en asignar velócidades a las partículas respetando la distribución de Maxwell-Boltzmann en el sistema. Si bien esto no es necesario, ya que el sistema alcanzará naturalmente esta distribución, respetarla desde un inicio acelerará el protocolo de calentamiento. Un aumento de temperatura es un aumento de energía cinética y por lo tanto de energía total, por eso debe ser paulatino y así el sistema podrá adaptarse a estos nuevos valores, sin que la proteína experimente cambios conformacionales violentos que no se corresponden con la realidad biológica (Braun et al. 2019).

2.3.4.3 Equilibración

Finalmente se buscará simular al sistema en condiciones de volumen y temperatura constante y así obtener sus valores de equilibrio. La determinación exacta de equilibrio implica conocer todo el mapa configuracional del sistema, lo cual es imposible de realizar en sistemas tan grandes. Por eso se busca que ciertos parametros se mantengan en un equilibrio dinámico para poder dar por finalizada la fase de equilibrio. Estos parámetros suelen ser: el volumen de la caja, su densidad, las energías del sistema y el *RMSD* (Raíz de la Desviación Cuadrática Media de las particulas) de la proteína. Esta es una medida de comparación de 2 estructuras que se obtiene de la siguiente forma:

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N} d_i^2}{N}} \tag{2.11}$$

Donde:

- N: número de partículas
- d_i: distancia de la i-ésima particula entre las 2 conformaciones

El *RMSD* es la principal medida de disimilitud entre 2 proteínas. La Figura 2.10 ilustra

su uso en 2 sistemas en proceso de equilibración.

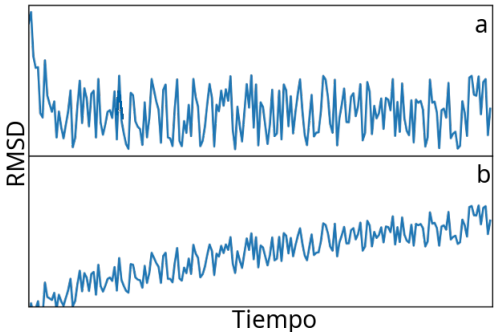

Figure 2.10: Distintos sistemas necesitarán distintos plazos para equilibrarse. La figura muestra una representación cualitativa de los RMSDs de distintas proteínas a cada momento de su equilibración, respecto a su estructura original. El sistema en **a** fue rápido en su equilibración y fluctúa alrededor de un valor constante de RMSD, mientras que el **b** continua variando sistemáticamente, por lo que su equilibración no puede darse por finalizada.

2.3.5 Termostatos y Barostatos

El ensamble que sampleará nuestro sistema estará dado por el número de partículas (N), la temperatura (T), la presión (P) y el volumen (V). El número de partículas está dado por nuestro sistema de partida, mientras que el volumen es fácil de controlar, aumentando y disminuyendo el tamaño de la caja de simulación. Son la temperatura y la presión las que necesitan de rutinas específicas para ser controladas.

2.3.5.1 Termostato

Usualmente, la MD de macromoléculas biológicas busca reproducir las condiciones experimentales de temperatura constante para poder obtener conformaciones del ensamble canónico (Zuckerman 2011); definiendo la temperatura con el teorema de equipartición (Leroy 2013):

$$\frac{3}{2}Nk_BT = \langle\sum_{i=1}^{N}\frac{1}{2}m_iv_i^2\rangle \tag{2.12}$$

Donde:

- N: número de partículas
- k_B: constante de Boltzmann
- T: temperatura promedio
- m_i: masa de la i-ésima particula
- v_i: velocidad de la i-ésima particula

Los corchetes angulares $\langle\rangle$ indican que esta es una medición promediada en el tiempo. Para mantener constante esta temperatura promedio, casi siempre es necesario un termostato que mantenga una temperatura biológica *objetivo* (Hünenberger 2005). Es importante diferenciar esta temperatura objetivo que se mantendrá constante a lo largo de la dinámica, de la temperatura instantántea que fluctuará alrededor de la instantánea, ya que casi ningún termostato, por más agresiva que sea su metodología, mantendrá constante a la temperatura instantánea (Basconi & Shirts 2013) (Hünenberger 2005), siendo el termostato gaussiano (*gaussian thermostat*) la excepción a esta regla. Los termostatos actúan alterando el cálculo de la trayectoria por parte de las ecuaciones newtonianas —que son intrínsicamente microcanónicas (mantienen constante la energía canónica)–, alterando la velocidad de las partículas para alcanzar la temperatura objetivo. El termostato de *velocity rescaling* es el ejemplo más directo: escala las magnitudes de las velocidades cada cierto número de pasos, para que la temperatura instantánea global coincida con la temperatura objetivo. De cualquier manera, se sabe que este termostato introduce artefactos en la trayectoria y no es recomendable para estas simulaciones (Braun et al. 2018). El termostato usado en esta tesis es el de Langevin.

2.3.5.1.1 Termostato de Langevin Suplementa las ecucaciones microcanónicas de Newton con dinámicas Brownianas, incluyendo la viscosidad y las colisiones aleatorias propias de un líquido (Schneider & Stoll 1978). Esto lo hace agregando 2 términos al cálculo de fuerza:

$$F = F_{interaccion} + F_{friccion} + F_{aleatoria}\rangle \tag{2.13}$$

Donde: - F: fuerza neta modificada por termostato de Langevin - $F_{interaccion}$: fuerza tradicional, obtenida con el campo de fuerza. - $F_{friccion}$: fuerza de amortiguación que simula la viscosidad de un líquido implícito. - $F_{aleatoria}$: fuerza aleatoria proveniente de las colisiones con las moléculas del líquido implícito.

El impacto de estas fuerzas estarán dadas por la elección de una constante específica de este termostato llamada frecuencia de colisión (γ). Una constante muy baja aproximará al sistema a un ensamble microcanónico y una demasiado alta a uno de puro movimiento Browniano.

2.3.5.2 Barostato

La otra variable a controlar (en principio) para reproducir las condiciones experimentales, es la presión. La MD de macromoléculas biológicas suelen ser llevadas a cabo en condiciones isotérmicas e isobáricas, por lo que se hace necesario un barostato. La presión es calculada usando el teorema del virial, utilizando a las posiciones de las partículas y

las fuerzas sobre ellas, y al igual que la temperatura es una magnitud promediada en el tiempo(Leach 2001). Otra similitud con la temperatura es que su medición instantánea variará, incluso cuandos se la mantenga constante, ya que es la magnitud promedio la que se controla y no así la instantánea (Braun et al. 2019). El barostato usado en esta tesis fue el de Berendsen.

2.3.5.2.1 Barostato de Berendsen Suele ser llamado el barostato de acoplamiento débil (*weak coupling*) en relación al barostato de método más directo: termostato de escalado de volumen (*simple volume rescaling*). Éste último escala la presión del sistema para igualarla a la presión objetivo en un sólo paso, lo cual genera artefactos en la trayectoria. En cambio, Berendsen, somete al sistema a un "baño de presión", adecuando el volumen de la caja periódicamente y aproximando, paulatinamente, la presión del sistema a la presión objetivo (Berendsen et al. 1984).

Aún así, el sistema no sampleará el ensamble isotérmico-isobárico y su uso está limitado a trayectorias de equilibración y no así para las de producción (Rizzi et al. 2019).

Si bien el uso de un barostato parece ser la opción correcta para generar trayectorias de relevancia biológica, se sabe que una vez que el sistema alcance la densidad deseada las conformaciones obtenidas manteniendo la presión constante no diferirán de las obtenidas a volumen constante. En efecto, uno puede mantener el tamaño de la caja constante, ahorrandose así costo computacional y posibles artefactos introducidos por un barostato incorrecto (Braun et al. 2019). En esta tesis se utilizó el barostato de Berendsen sólamente para la preparación del sistema, hasta alcanzar la densidad deseada de $1\frac{g}{ml}$

2.4 Practicas metodológicas de la Dinámica Molecular observadas en esta tesis

2.4.1 Preparación del sistema

Todas las simulaciones fueron realizadas con el paquete de software de simulación AMBER(Case et al. 2005). Cada sistema fue solvatado con agua TIP3P(Jorgensen et al. 1983) en una caja periódica de octahedro truncado suficientemente grande como para contener la proteína y 10Å de solvente en cada dirección. Luego se agregaron iones cloro o sodio para neutralizar. Las interacciones de los residuos fueron modeladas con el campo de fuerza ff14SB(Maier et al. 2015) y los ligandos fueron parametrizados con el *(*General Amber Force Field) (*GAFF*)(Junmei Wang et al. 2004) usando el método de asignación de cargas AM1-BCC(Jakalian et al. 2002).

2.4.2 Minimización

La minimización consistió en 100 pasos de *steepest-descent* y 400 de *conjugate gradient* para relajar el solvente, aplicando restricciones al movimiento de los átomos del soluto. Luego se simularon otros 400 pasos de *conjugate gradient* sin restricciones.

2.4.3 Calentamiento

Los sistemas fueron calentados gradualmente durante 400ps hasta llegar a los 300K, usando el termostato de Langevin y a volumen constante. Durante el calentamiento, se aplican restricciones de 25 kcal/(mol · Å2) a los átomos de la cadena principal de la proteína y todos los átomos del ligando, si hubiere.

2.4.4 Equilibración

Los sistemas fueron equilibrados a presión y temperatura constantes usando el barostato de Berendsen y termostato de Langevin, respectivamente. Las restricciones de 25 kcal/(mol · Å2) fueron relajándose hasta quitarlas completamente. Una vez quitadas, se deja de controlar la presión del sistema y se mantiene el volumen de la caja constante. La Tabla 1.2 del Apéndice A muestra el protocolo utilizado para la mayoría de los sistemas y la 2.2 el utilizado para sistemas que se creía podían necesitar un protocolo más paulatino.

Chapter 3

Análisis de movimientos colectivos

3.1 Análisis Modos Normales

3.1.1 Introducción histórica

Como ya se dijo en el capítulo primero, las proteínas, en su estado nativo, alternan distintas conformaciones en distintas escalas de tiempo según la barrera energética del cambio conformacional y la estabilidad relativa de los confórmeros involucrados. Los cambios conformacionales de mayor relevancia biológica, como el requerido para la unión a otra molécula, suelen ser alcanzados mediante desplazamientos colectivos de grandes grupos de residuos o dominios estructurales. Estos desplazamientos involucran movimientos concertados de baja frecuencia dentro de la proteína. Su relevancia funcional ha sido demostrada en numerosos trabajos anteriores (Case 1994) (Brooks et al. 1995) (Janežič & Brooks 1995) (Tama & Sanejouand 2001).

La dinámica vibracional en torno a la estructura nativa de una proteína puede ser descripta, en buena aproximación, por modelos armónicos como el análisis de modos normales (NMA, por sus siglas en inglés). Esta idea surgió inicialmente hace más de 40 años (McCammon et al. 1976). En el estudio de un movimiento tipo *bending* de la lisozima, se encontró que el potencial era aproximadamente parabólico, y se trató el movimiento relativo de dos dominios como un oscilador armónico angular. La frecuencia obtenida para la oscilación coincidió con la frecuencia del primer modo normal calculada previamente por otro método —llamado minimización energética adiabática— (Brooks & Karplus 1985). Desde entonces el NMA se ha utilizado para explorar la dinámica conformacional de las proteínas (Brooks & Karplus 1983) y ha demostrado ser un método físicamente plausible y eficiente como herramienta matemática y computacional. El interés en el uso de NMA se renovó en los últimos 20 años, principalmente debido a la introducción de modelos simplificados, basados en la mecánica de redes de polímeros (Tirion 1996) (Bahar et al.

1997) (Hinsen 1998) (Atilgan et al. 2001) (Li & Cui 2002)

3.1.2 Ecuaciones del movimiento vibracional

Si consideramos una molécula poliatómica formada por N átomos que vibran en torno a cierta posición de equilibrio, para describir sus movimientos vibracionales conviene utilizar un sistema de coordenadas fijo en la molécula, un sistema de referencia que se traslade y gire con ella, eliminando la traslación y rotación del centro de masa. Sean x_α, y_α, z_α las coordenadas cartesianas del átomo α respecto al sistema fijo en la molécula y sean x_α^0, y_α^0, z_α^0 los valores en la configuración de equilibrio. La energía cinética vibracional viene dada por la expresión:

$$K = \frac{1}{2}\sum_{\alpha=1}^{N} m_\alpha \left[\left(\frac{dx_\alpha}{dt}\right)^2 + \left(\frac{dy_\alpha}{dt}\right)^2 + \left(\frac{dz_\alpha}{dt}\right)^2\right] \tag{3.1}$$

En este punto, conviene introducir las denominadas **coordenadas cartesianas ponderadas en masa** definidas como:

$$q_1 = \sqrt{m_1}(x_1 - x_1^0), \quad q_2 = \sqrt{m_1}(y_1 - y_1^0), \quad q_3 = \sqrt{m_1}(z_1 - z_1^0)$$

$$q_4 = \sqrt{m_2}(x_2 - x_2^0), \quad q_5 = \sqrt{m_2}(y_2 - y_2^0), \quad q_6 = \sqrt{m_2}(z_2 - z_2^0)$$

$$\ldots$$

$$q_{3N-2} = \sqrt{m_N}(x_N - x_N^0), \quad q_{3N-1} = \sqrt{m_N}(y_N - y_N^0), \quad q_{3N} = \sqrt{m_N}(z_N - z_N^0) \tag{3.2}$$

Donde:

- m_i: masa de la partícula i-ésima.
- q_i: coordenada cartesiana ponderada en masa de la partícula i-ésima.

Derivando estas expresiones con respecto al tiempo y sustituyendo los resultados en eq. 3.1 la energía cinética puede escribirse como:

$$K = \frac{1}{2}\sum_{i=1}^{3N}\left(\frac{dq_i}{dt}\right)^2 \tag{3.3}$$

Los movimientos vibracionales de la molécula están controlados por su superficie de energía potencial V. Esta superficie de potencial depende de las $3N$ coordenadas cartesianas o de las $3N$ coordenadas de desplazamiento ponderadas q_i:

$$V = V(q_1, ..., q_{3N}) \tag{3.4}$$

Entonces podemos expresar la energía potencial como un desarrollo en serie de Taylor de las coordenadas q_i en torno a sus valores en la configuración equilibrio:

$$V = V_0 + \sum_{i=1}^{3N} \left(\frac{\delta V}{\delta q_i}\right)_0 q_i + \frac{1}{2!} \sum_{i=1}^{3N} \sum_{j=1}^{3N} \left(\frac{\delta^2 V}{\delta q_i \delta q_j}\right)_0 q_i q_j + \frac{1}{3!} \sum_{i=1}^{3N} \sum_{j=1}^{3N} \sum_{k=1}^{3N} \left(\frac{\delta^3 V}{\delta q_i \delta q_j \delta q_k}\right)_0 q_i q_j q_k + ... \tag{3.5}$$

El subíndice 0 hace referencia a la conformación de equilibrio, que corresponde a un mínimo en la energía potencial, por lo tanto su derivada es nula:

$$\left(\frac{\delta V}{\delta q_i}\right)_0 = 0 \tag{3.6}$$

Además, V_0 es el potencial de referencia en el equilibrio, al estar definido como una constante se puede definir nulo para la conformación de equilibrio de referencia, por lo cual:

$$V_0 = 0 \tag{3.7}$$

Finalmente, asumiendo que los desplazamientos del sistema son suficientemente pequeños, podemos despreciar los términos de tercer orden y superiores en el desarrollo de Taylor, reduciendo el potencial a la forma cuadrática en las fluctuaciones, llamada **aproximación armónica** (*harmonic approximation*) del potencial:

$$V = \frac{1}{2} \sum_{i=1}^{3N} \sum_{j=1}^{3N} \left(\frac{\delta^2 V}{\delta q_i \delta q_j}\right)_0 q_i q_j \tag{3.8}$$

definiendo:

$$k_{ij} = \left(\frac{\delta^2 V}{\delta q_i \delta q_j}\right)_0 \tag{3.9}$$

donde $k_{ij} = k_{ji}$, nos queda:

$$V = \frac{1}{2}\sum_{i=1}^{3N}\sum_{j=1}^{3N} k_{ij} q_i q_j \tag{3.10}$$

Las constantes k_{ij} son denominadas constantes de fuerza armónica para las coordenadas de desplazamiento ponderadas.

Para averiguar cómo cambian las coordenadas con el tiempo, es decir como vibra la molécula, hemos de resolver las ecuaciones del movimiento vibracionales. Derivando la energía potencial para obtener la fuerza sobre las partículas y utilizando las ecuaciones de Newton, podemos escribir las ecuaciones del movimiento en función de la energía potencial de la siguiente manera:

$$\begin{aligned} F_{x,\alpha} &= m_\alpha \frac{d^2 x_\alpha}{dt^2} = \frac{\delta V}{\delta x_\alpha} \\ F_{y,\alpha} &= m_\alpha \frac{d^2 y_\alpha}{dt^2} = \frac{\delta V}{\delta y_\alpha} \\ F_{z,\alpha} &= m_\alpha \frac{d^2 z_\alpha}{dt^2} = \frac{\delta V}{\delta z_\alpha} \end{aligned} \tag{3.11}$$

Donde:

- $F_{xyz,\alpha}$: componentes cartesianas de la fuerza que actúa sobre el átomo α.

Cambiemos ahora en estas ecuaciones, las coordenadas cartesianas por las coordenadas de desplazamiento ponderadas. Derivando 2 veces respecto al tiempo a la eq. 3.2 —y tomando como ejemplo a la ecuación del movimiento de la coordenada x_1—, obtenemos:

$$\frac{d^2 q_1}{dt^2} = \sqrt{m_1}\frac{d^2 x_1}{dt^2} \tag{3.12}$$

y usando la regla de la cadena en la derivada del potencial nos queda:

$$\frac{\delta V}{\delta x_1} = \frac{\delta V}{\delta q_1}\frac{dq_1}{dx_1} = \sqrt{m_1}\frac{\delta V}{\delta q_1} \tag{3.13}$$

Usando estas dos ultimas ecuaciones en eq. 3.11 obtenemos:

$$\frac{d^2 q_1}{dt^2} = \frac{\delta V}{\delta q_1} \tag{3.14}$$

Las ecuaciones de movimiento restantes se transforman de un modo análogo y podemos escribirlas todas de forma compacta como:

$$\frac{d^2 q_u}{dt^2} + \frac{\delta V}{\delta q_u} = 0 \quad u = 1, ..., 3N \tag{3.15}$$

Ahora debemos sustituir en la eq. 3.15, la expresión obtenida para la energía potencial eq. 3.10. Pero primero desarrollamos la derivada de V con respecto a q_u:

$$\begin{aligned}
\frac{\delta V}{\delta q_u} &= \frac{1}{2}\sum_{i=1}^{3N}\sum_{j=1}^{3N} k_{ij}\left(\frac{\delta q_i}{\delta q_u} q_j + \frac{\delta q_j}{\delta q_u} q_i\right) \\
&= \frac{1}{2}\sum_{i=1}^{3N}\sum_{j=1}^{3N} k_{ij}\left(\delta_{iu} q_j + \delta_{ju} q_i\right) \\
&= \frac{1}{2}\sum_{i=1}^{3N}\sum_{j=1}^{3N} k_{ij}\delta_{iu} q_j + k_{ij}\delta_{ju} q_i \\
&= \frac{1}{2}\sum_{j=1}^{3N} k_{uj} q_j + \frac{1}{2}\sum_{i=1}^{3N} k_{iu} q_i \\
&= \sum_{j=1}^{3N} k_{uj} q_j
\end{aligned} \tag{3.16}$$

donde hemos tenido en cuenta que las coordenadas q_u son independientes entre sí. Sustituyendo esta expresión en la eq. 3.15 obtenemos finalmente:

$$\frac{d^2 q_u}{dt^2} + \sum_{j=1}^{3N} k_{uj} q_j = 0 \qquad u = 1, ..., 3N \tag{3.17}$$

3.1.3 Solución de las ecuaciones de movimiento

La eq. 3.17 representa un sistema de ecuaciones diferenciales acopladas cuyas incógnitas son las coordenadas de desplazamiento ponderadas q_u. Cada una de estas ecuaciones diferenciales contiene a todas las variables q_u, es decir la derivada segunda de cada coordenada q_u depende, no sólo de q_u sino también del resto de las coordenadas q. Estos acoplamientos complican la resolución del sistema de ecuaciones diferenciales por lo que es necesario un cambio de variables para resolverlo.

Ahora se reescribirá la eq. 3.8 como un sistema de ecuaciones desacopladas. Llamemos Q_i a las nuevas variables que buscamos y supongamos que las coordenadas de desplazamiento ponderadas q_u estén relacionadas con ellas mediante las combinaciones lineales:

$$q_u = \sum_{i=1}^{3N} l_{ui} Q_i \tag{3.18}$$

Donde:

- l_{ui}: coeficientes que especificaremos mas adelante.

En notación matricial estas ecuaciones de transformación se expresan de la forma:

$$\boldsymbol{q} = \boldsymbol{L}\boldsymbol{Q} \tag{3.19}$$

Donde:

- $\boldsymbol{q}$: vector columna que contiene las coordenadas de desplazamiento ponderadas q_u
- $\boldsymbol{Q}$: vector columna que contiene las nuevas coordenadas Q_i
- $\boldsymbol{L}$: matriz de coeficientes l_{ui}.

La eq. 3.8 para la función de energía potencial en notación matricial viene dada, a su vez por:

$$V = \frac{1}{2}\tilde{\boldsymbol{q}}\boldsymbol{U}\boldsymbol{q} \tag{3.20}$$

Donde:

- $\tilde{\boldsymbol{q}}$: transpuesta del vector columna $\boldsymbol{q}$, es decir, el vector fila.
- $\boldsymbol{U}$: matriz cuadrada de constantes de fuerza k_{ij}.

$\boldsymbol{U}$ se obtiene según eq. 3.9 :

$$\boldsymbol{U} = \begin{pmatrix} \frac{\delta^2 V}{\delta q_1^2} & \dots & \frac{\delta^2 V}{\delta q_1 \delta q_N} \\ \dots & \dots & \dots \\ \frac{\delta^2 V}{\delta q_N \delta q_1} & \dots & \frac{\delta^2 V}{\delta q_{3N}^2} \end{pmatrix} \tag{3.21}$$

Combinando eq. 3.20 y eq. 3.21 obtenemos la expresión:

$$V = \frac{1}{2}\tilde{\boldsymbol{q}}\boldsymbol{U}\boldsymbol{q} = \frac{1}{2}\widetilde{\boldsymbol{L}\boldsymbol{Q}}\boldsymbol{U}\boldsymbol{L}\boldsymbol{Q} = \frac{1}{2}\widetilde{\boldsymbol{Q}}\tilde{\boldsymbol{L}}\boldsymbol{U}\boldsymbol{L}\boldsymbol{Q} \tag{3.22}$$

Debemos ahora encontrar la matriz de coeficientes de transformación $\boldsymbol{L}$ que apareció por primera vez en eq. 3.19.

En álgebra lineal a esta se la define como la matriz de autovectores de $\boldsymbol{U}$ y se obtiene a partir de la diagonalización de la misma. La ecuación matricial de valores propios de $\boldsymbol{U}$ esta dada por:

$$\boldsymbol{U}\boldsymbol{L} = \boldsymbol{L}\boldsymbol{\Lambda} \tag{3.23}$$

donde $\boldsymbol{\Lambda}$ es la matriz diagonal que contiene los autovalores λ_i . Multiplicando por la izquierda por la matriz inversa de $\boldsymbol{L}$ obtenemos:

$$\boldsymbol{L}^{-1}\boldsymbol{U}\boldsymbol{L} = \boldsymbol{L}^{-1}\boldsymbol{L}\boldsymbol{\Lambda} = \boldsymbol{I}\boldsymbol{\Lambda} = \boldsymbol{\Lambda} \tag{3.24}$$

Donde:

- $\boldsymbol{I}$: matriz unidad.

Puesto que la matriz $\boldsymbol{U}$ es una matriz real y simétrica, su matriz de autovectores es ortonormal. Una matriz ortonormal es aquellas cuya inversa es igual a su traspuesta. En nuestro caso por tanto tenemos que $\boldsymbol{L}^{-1} = \tilde{\boldsymbol{L}}$, y podemos escribir eq. 3.24 de la forma:

$$\tilde{\boldsymbol{L}}\boldsymbol{U}\boldsymbol{L} = \boldsymbol{\Lambda} \tag{3.25}$$

Usando esta expresión en la eq. 3.22 queda:

$$V = \frac{1}{2}\widetilde{\boldsymbol{Q}}\boldsymbol{\Lambda}\boldsymbol{Q} \tag{3.26}$$

y desarrollando los productos matriciales, teniendo en cuenta que la matriz $\boldsymbol{\Lambda}$ es diagonal, se obtiene:

$$V = \frac{1}{2}\sum_{i=1}^{3N}\lambda_i Q_i \tag{3.27}$$

Conseguimos así escribir la función de energía potencial como una sumatoria de las nuevas variables Q_i. Que pueden ser escritas en función de las coordenadas q_u como:

$$Q_i = \sum_{u=1}^{3N} l_{ui} q_u \tag{3.28}$$

Ahora debemos reescribir las ecuaciones de movimiento eq. 3.17 respecto a las nuevas coordenadas Q_i . Para esto primero derivamos dos veces respecto al tiempo la eq. 3.18, obteniendo:

$$\frac{d^2 q_u}{dt^2} = \sum_{i=1}^{3N} l_{ui} \frac{d^2 Q_i}{dt^2} \tag{3.29}$$

y usamos, en segundo lugar, la regla de la cadena para desarrollar las derivadas de la energía potencial como:

$$\frac{\delta V}{\delta q_u} = \sum_{i=1}^{3N} \frac{\delta V}{\delta Q_i} \frac{\delta Q_i}{\delta q_u} = \sum_{i=1}^{3N} \frac{\delta V}{\delta Q_i} l_{ui} \tag{3.30}$$

donde hemos utilizado las eq. 3.28 para determinar las derivadas. Sustituyendo las eq. 3.29 y eq. 3.30 en las ecuaciones del movimiento (eq. 3.17) y sacando como factor común los coeficientes l_{ui}, queda:

$$\sum_{i=1}^{3N} l_{ui} \left(\frac{d^2 Q_i}{dt^2} + \frac{\delta V}{\delta Q_i} \right) = 0 \quad u = 1, ..., 3N \tag{3.31}$$

Puesto que los coeficientes l_{ui} son, diferentes de cero, esta sumatoria sólo se anula cuando lo hacen los factores que multiplican a dichos coeficientes, es decir, cuando se cumple:

$$\frac{d^2 Q_i}{dt^2} + \frac{\delta V}{\delta Q_i} = 0 \quad i = 1, ..., 3N \tag{3.32}$$

Estas son las ecuaciones del movimiento que buscamos en función de las coordenadas Q_i. Derivando la energía potencial dada por la eq. 3.27 respecto a Q_i obtenemos

$$\frac{\delta V}{\delta Q_i} = \sum_{k=1}^{3N} \lambda_k Q_k \frac{\delta Q_k}{\delta Q_i} = \sum_{k=1}^{3N} \lambda_k Q_k \delta_{ki} = \lambda_i Q_i \tag{3.33}$$

y sustituyendo este resultado en las ecuaciones del movimiento (eq. 3.32) queda:

$$\frac{d^2 Q_i}{dt^2} + \lambda_i Q_i = 0 \quad i = 1, ..., 3N \tag{3.34}$$

Estas ecuaciones están completamente desacopladas, es decir la derivada segunda de cada variable Q_i con respecto al tiempo depende únicamente de ella misma. Las coordenadas Q_i

introducidas para desacoplar las ecuaciones vibracionales del movimiento de la molécula se denominan **coordenadas normales de vibración**.

3.1.3.1 Modos normales de vibración

Las ecuaciones de movimiento eq. 3.34 son similares a las de un oscilador armónico unidimensional de masa reducida y constante de fuerza λ_i . Las soluciones clásicas de la eq. 3.34 pueden escribirse de la forma

$$Q_i(t) = B_i sen(\lambda_i^{1/2} t + b_i) \quad i = 1, ..., 3N \tag{3.35}$$

Donde:

- B_i: amplitud del i-ésimo modo.
- b_i: fase del i-ésimo modo. Depende de las condiciones iniciales.

Una vez obtenidas las coordenadas normales Q_i podemos determinar las coordenadas de desplazamiento ponderadas q_u sustituyendo la eq. 3.35 en la eq. 3.18, quedando:

$$q_u = \sum_{i=1}^{3N} A_{ui} sen(\lambda_i^{1/2} t + b_i) \quad i = 1, ..., 3N \tag{3.36}$$

donde $A_{ui} = l_{ui} B_i$. Estas son las soluciones generales de las ecuaciones del movimiento vibracional para las coordenadas de desplazamiento ponderadas. Puesto que dichas ecuaciones forman un sistema de $3N$ ecuaciones diferenciales de segundo orden, sus soluciones generales admiten un total de $6N$ constantes arbitrarias a especificar, que son las $3N$ amplitudes B_i y los $3N$ factores de fase b_i . Los valores de estas constantes se calculan fijando las condiciones iniciales del sistema, es decir, a partir de los $3N$ valores iniciales (a tiempo $t = 0$) de las coordenadas q_u y los $3N$ valores iniciales de las correspondientes velocidades.

3.1.3.1.1 Vibración bajo un único modo normal Si suponemos que las condiciones iniciales son tales que todas las amplitudes B_i son iguales a cero salvo una, la B_m, la única constante A_{ui} que no se anula en la eq. 3.36 es la A_{um}, de modo que las soluciones para las coordenadas $q_u(t)$ se reducen a

$$q_u = A_{um} sen(\lambda_m^{1/2} t + b_m) \tag{3.37}$$

En este caso todos los átomos vibran en fase con la misma frecuencia ν_m asociada a la constante de fuerza normal λ_m, es decir

$$\nu_m = \frac{\sqrt{\lambda_m}}{2\pi} \tag{3.38}$$

Usando la eq. 3.35, estas soluciones particulares pueden expresarse como

$$q_u = l_{um} Q_m(t) \quad u = 1, ..., 3N \tag{3.39}$$

Las vibraciones de este tipo son los denominados modos normales de vibración.

3.1.4 Modelo de Red Anisotrópica

Si bien el modelo atomístico original de NMA fue aplicado con éxito, resulta computacionalmente costoso y presenta ciertas limitaciones, especialmente en su conjunto de premisas. En primer lugar, la dinámica de una proteína no es armónica; ya en las primeras, y cortas, dinámicas moleculares se pudo apreciar que las proteínas superan pequeñas barreras energéticas con frecuencia (Elber & Karplus 1987). Los campos de fuerza utilizados por el NMA tradicional, similares a los de MD, reflejan esta anarmonicidad. Otro de los requisitos es que la estructura se encuentre en un mínimo energético y se entiende que las estructuras obtenidas del PDB no lo están. Por lo que es necesario un primer paso de minimización de la estructura, el cual dependerá del campo de fuerza y la rutina de minimización empleados. El uso de NMA en modelos de redes elásticas proponen una alternativa válida que permite reducir significativamente el costo computacional, solucionando a su vez, problemas de rugosidades locales de la superficie de energía potencial de las proteínas(Yang & Chng 2008).

Por eso los modelos de redes elásticas intentan modificar el NMA tradicional, simplificando el campo de fuerza utilizado, lo que al mismo tiempo evita la necesidad de una minimización previa. Si bien estos modelos utilizan aún más simplificaciones que el NMA tradicional, obtienen aún mejores resultados, por respetar las condiciones iniciales de armonicidad y de estructura en un mínimo de energía.

3.1.4.1 Fundamentos

En esencia, hay dos tipos diferentes de modelos de redes elásticas (ENM, por sus siglas en inglés), que difieren en su dimensionalidad de las coordenadas de movimiento por cada partícula de la molécula analizada. El Modelo de Red Gausiana (GNM, por sus siglas en inglés), propuesto por Ivet Bahar (Haliloglu et al. 1997), es un modelo de 1 coordenada

por partícula. Mientras que el modelo de Tirion, más tarde llamado Modelo de Red Anisotrópica (ANM, por sus siglas en inglés) (Atilgan et al. 2001) (Tirion 1996), resulta en 3 coordenadas por partícula. De aqui en adelante nos referiremos, exclusivamente, al modelo de ANM.

ANM es una de las variantes de bajo costo computacional a los cálculos de NMA que considera a la proteína como una red de partículas (o nodos) interconectadas por resortes armónicos (Figura 3.1). En su versión original, estos últimos están afectados por una única constante de fuerza γ, la cual determina un potencial armónico universal para el modelo (Tirion 1996).

El número de grados de libertad del sistema se reduce al considerar solo los $C\alpha$ provenientes de la estructura cristalográfica obtenida de la Protein Data Bank (PDB). Es decir, se tiene en cuenta un solo nodo por residuo. Por otro lado, el modelo considera a la estructura cristalográfica como la correspondiente estructura de equilibrio. Por tal motivo el método no requiere la minimización del sistema para la obtención de la matriz Hessiana.

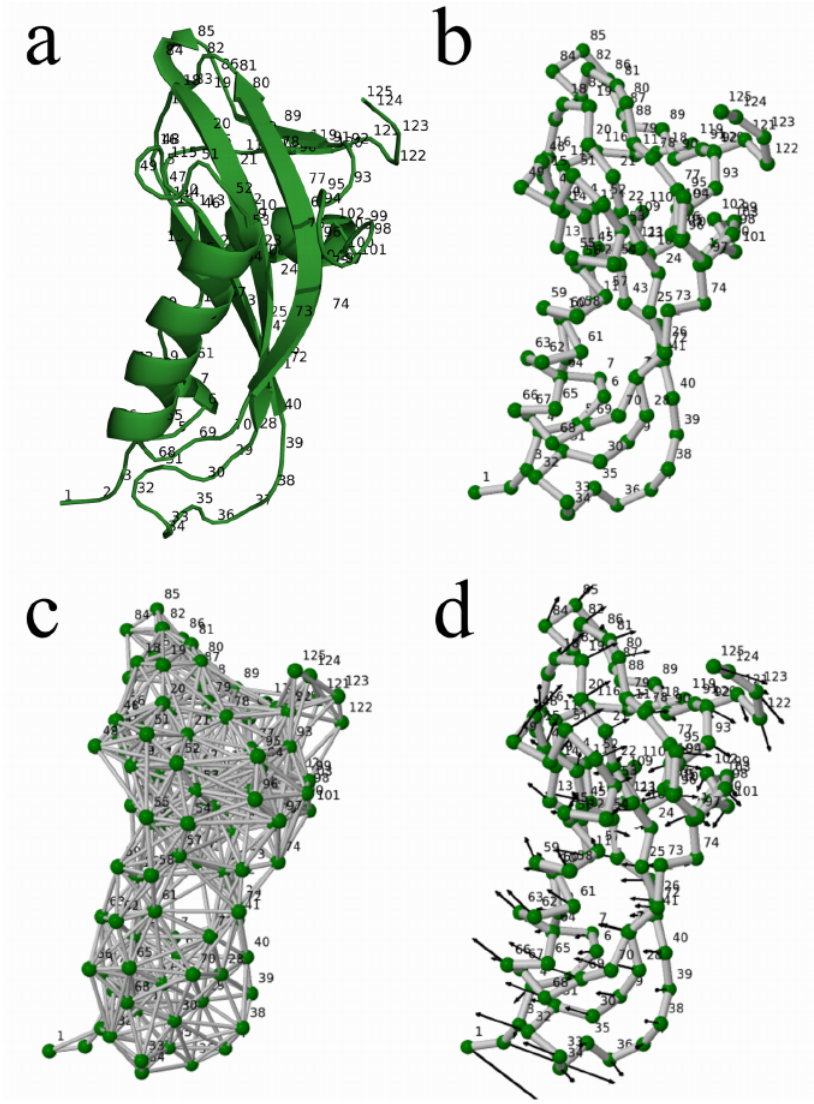

Figure 3.1: **a)** Estructura sin ligando de la proteína Fosfohistidina Fosfatasa 1. **b)** Posiciones de los $C\alpha$ de la cadena principal. **c)** Modelo de red elástica que considera a todos los $C\alpha$ dentro de un radio determinado como nodos interactuantes. **d)** Dirección de desplazamiento de los $C\alpha$ correspondiente al primer modo normal derivado del ANM

De este modo se cumple con una de las condiciones iniciales de NMA —ya que ahora la estructura se encuentra en punto de energía mínimo—, y se logra una gran reducción en el costo computacional de los cálculos.

Para determinar si 2 $C\alpha$ deben estar unidos por un resorte, se fija un radio de corte r_c tal que:

$$\begin{aligned} si \quad & |r_j^0 - r_i^0| \leqslant r_c \Rightarrow k_{ij} = \gamma, \\ si \quad & |r_j^0 - r_i^0| \geqslant r_c \Rightarrow k_{ij} = 0 \end{aligned} \tag{3.40}$$

Donde:

- $r_{i(j)}^0$: posición de equilibrio de la partícula $i(j)$ (Figura 3.2).

- k_{ij}: constante de fuerza del resorte entre partículas i y j.
- γ: constante de fuerza del resorte única (parámetro a definir).
- r_c: radio de corte.

Según este planteo, entonces, se contemplan tanto uniones entre dos residuos adyacentes como entre cualquier par de residuos dentro del radio r_c definido. Esto significa que no se discierne entre uniones del tipo covalente, puente de hidrógeno, interacciones iónicas o interacciones de Van der Waals.

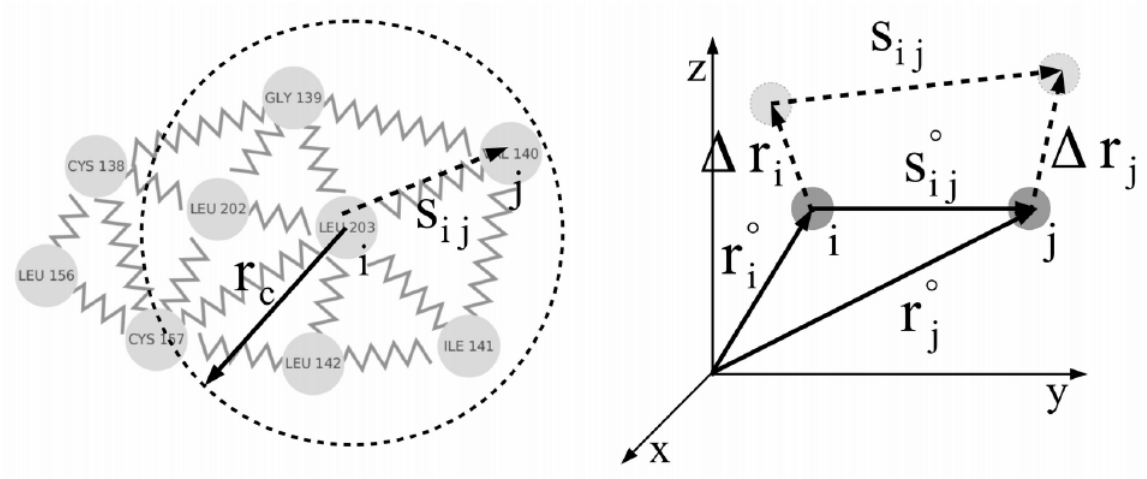

Figure 3.2: Representación esquemática de los nodos en el ANM. Cada nodo se conecta a sus vecinos mediante resortes, dependiendo del radio de corte r_c . $\mathring{s}_{ij}$ es la distancia de equilibrio entre los sitios i y j, cuyas posiciones de equilibrio son $\mathring{r}_i$ y $\mathring{r}_j$. Los vectores de fluctuación Δr_i , Δr_j y la distancia entre los residuos s_{ij}, se grafican con línea punteada.

Si consideramos dos residuos j y i unidos por un resorte, el potencial elástico armónico al cual están sujetos será:

$$V = \frac{1}{2}k_{ij}(s_{ij} - s^0_{ij})^2 = \frac{1}{2}k_{ij}\left[[(x_j - x_i)^2 + (y_j - y_i)^2 + (z_j - z_i)^2]^2 - s^0_{ij}\right]^2 \qquad (3.41)$$

Donde:

- $s_{ij} = r_j - r_i$: distancia instantánea entre los átomos i y j.
- $s^0_{ij} = r^0_j - r^0_i$: distancia de equilibrio entre los átomos i y j.
- $r_{i(j)}$: posición instantánea de la partícula $i(j)$.

Es decir, el potencial variará cuanto más se alejen las partículas de su posición inicial. Ésta es la razón por la cual no es necesaria una minimización para el cálculo de NMA con ENM: al modelar a la proteína como un sistema de resortes en reposo se obtiene un sistema en su mínimo absoluto.

Siguiendo con el modelo de ENM, lo que nos interesa es la segunda derivada del potencial. La primera y segunda derivadas del potencial V con respecto a las componentes de r_i son:

$$\frac{\delta V}{\delta x_i} = -\frac{\delta V}{\delta x_j} = -k_{ij}(x_j - x_i)\left(1 - \frac{s_{ij}^0}{s_{ij}}\right) \tag{3.42}$$

$$\frac{\delta^2 V}{\delta^2 x_i} = -\frac{\delta^2 V}{\delta^2 x_j} = -k_{ij}\left(1 + s_{ij}^0 \frac{(x_j - x_i)^2}{s_{ij}^3} - \frac{s_{ij}^0}{s_{ij}}\right) \tag{3.43}$$

Ecuaciones análogas se obtienen para las componentes y_i y z_i de r_i.

Cuando la molécula está en posición de equilibrio ($s_{ij} = s_{ij}^0$), el potencial elástico es $V_{s_{ij}^0} = 0$ y las eq. 3.42 y eq. 3.43 se reducen a:

$$\frac{\delta V}{\delta x_i} = 0 \tag{3.44}$$

$$\frac{\delta^2 V}{\delta^2 x_i} = -k_{ij}\frac{(x_j - x_i)^2}{s_{ij}^2} \tag{3.45}$$

Y las segundas derivadas cruzadas:

$$\frac{\delta^2 V}{\delta x_i \delta y_j} = \frac{\delta^2 V}{\delta x_j \delta y_i} = -k_{ij}\frac{(x_j - x_i)(y_j - y_i)}{s_{ij}^2} \tag{3.46}$$

Para el caso general de N residuos unidos por M resortes, las ecuaciones de las segundas derivadas del potencial pueden ser organizadas en una matriz Hessiana de $3N$ x $3N$, H:

$$\boldsymbol{H} = \begin{pmatrix} H_{11} & ... & H_{1N} \\ ... & ... & ... \\ H_{12} & ... & H_{NN} \end{pmatrix} \tag{3.47}$$

Donde cada superelemento $\boldsymbol{H_{ij}}$ (tal que $i \neq j$) es:

$$\boldsymbol{H_{ij}} = \begin{pmatrix} \frac{\delta^2 V}{\delta x_i x_j} & \frac{\delta^2 V}{\delta x_i y_j} & \frac{\delta^2 V}{\delta x_i z_j} \\ \frac{\delta^2 V}{\delta y_i x_j} & \frac{\delta^2 V}{\delta y_i y_j} & \frac{\delta^2 V}{\delta y_i z_j} \\ \frac{\delta^2 V}{\delta z_i x_j} & \frac{\delta^2 V}{\delta z_i y_j} & \frac{\delta^2 V}{\delta z_i z_j} \end{pmatrix} \tag{3.48}$$

Finalmente, los modos normales del ANM se obtienen diagonalizando la matriz $\boldsymbol{H}$:

$$\widetilde{\boldsymbol{Q}}\boldsymbol{H}\boldsymbol{Q} = \boldsymbol{\Lambda} \tag{3.49}$$

Los autovalores λ_k son los elementos de la matriz diagonal $\boldsymbol{\Lambda}$ y representan los cuadrados de las frecuencias ω_k de los modos individuales:

$$\lambda_k = \omega_k^2 \tag{3.50}$$

Y los autovectores Q_k están definidos como:

$$Q_k = B_k sen(\lambda_i^{1/2} t) \quad i = 1, ..., 3N \tag{3.51}$$

Donde:

- B_k: amplitud del autovector k

La eq. 3.51 es similar a la eq. 3.35, sólo que la primera representa a modos normales usando el modelo de ANM. De las frecuencias obtenidas en la diagonalización, 6 de ellas serán nulas; están asociadas con las tres direcciones de traslación pura de un cuerpo rígido y los tres ángulos que definen una rotación en un cuerpo rígido. Entonces existen 6 direcciones de movimiento (de las $3N$) que corresponden a características de cuerpo rígido exclusivamente, y no aportan información sobre las vibraciones internas.

3.1.4.1.1 Sobre el radio de corte r_c Éste es el principal parámetro a ajustar. Un r_c demasiado grande conectará átomos muy distantes y la red elástica dejará de representar la topología de la proteína, la cual es esencial para el resultado del NMA por ANM. En cambio, un r_c demasiado pequeño generará subredes locales dentro de la proteína, desconexas entre sí, lo que impedirá calcular movimientos globales (Sanejouand 2013). El r_c de una red elástica de una proteína suele encontrarse entre los 7Å y 16Å(Kondrashov et al. 2007)(Kundu et al. 2002). Estos valores se han ajustado luego de correlacionar resultados experimentales y teóricos mediante una metodología que se detallará más adelante.

Los modos son vectores ortonormales, por lo tanto el conjunto de los $3N$ modos normales forman una base de coordenadas que pueden representar todos los posibles desplazamientos del sistema en torno de la conformación de equilibrio. Cada modo normal representa una fluctuación concertada de los átomos del sistema que vibran con la misma frecuencia. Los perfiles de los modos de baja frecuencia revelan los mecanismos de movimientos cooperativos o globales que involucran el desplazamiento de grandes bloques o dominios de las proteínas (Mahajan & Sanejouand 2015) (Tama & Sanejouand 2001). Los residuos

con movimiento más restringido en dichos modos (sus mínimos) suelen tener un comportamiento crítico, por ejemplo como ejes de movimientos tipo *bending* (visagra), que gobiernan el movimiento relativo entre dominios completos (Keskin et al. 2000) (Bahar et al. 1998) (Y. Wang et al. 2004).

Trabajos anteriores evidencian que el ANM reproduce adecuadamente las amplitudes de movimiento relativo de los residuos descriptas por los modos de baja frecuencia (Hinsen & Kneller 1999) (Micheletti et al. 2004). Estos modos involucran el desplazamiento global de un gran número de residuos y por lo tanto experimentan un potencial efectivo insensible al detalle de las interacciones individuales específicas de cada par de residuos involucrados (Zheng et al. 2006). Por el contrario, los modos de alta frecuencia definen fluctuaciones más localizadas y su correcta descripción requiere de un tratamiento diferencial de las interacciones entre pares de residuos específicos(Bahar et al. 1998).

3.1.4.2 Aplicaciones de los modos normales

3.1.4.2.1 Factores B de temperatura Las fluctuaciones cuadráticas medias $\langle \Delta r_i^2 \rangle$ de los carbonos alfa de los residuos pueden determinarse experimentalmente mediante los *B-factors* de temperatura asociados a la determinación cristalográfica por RX o por diferencias cuadráticas medias entre los diferentes modelos de RMN. Estas suelen ser utilizadas para verificar la validez del modelo ANM en la proteína de estudio ya que acumulando el desplazamiento de un residuo bajo cada modo normal puede predecirse, de manera teórica, el *B-factor* de ese residuo. Si estos *B-factors* teóricos concuerdan con los experimentales, entendemos que el modelo de ANM refleja la dinámica de la proteína (Bahar et al. 1997) (Atilgan et al. 2001).

El *B-factor* o también llamado factor cristalográfico de Debye-Waller B_i del átomo i, está relacionado con las fluctuaciones de sus coordenadas atómicas a través de (Sanejouand 2013):

$$B_i^{exp} = \frac{8\pi^2}{3} \langle \Delta r_i^2 \rangle \tag{3.52}$$

Donde:

- $\Delta r_i^2 = \Delta x_i^2 + \Delta y_i^2 + \Delta z_i^2$. Es la magnitud de las fluctuaciones cuadráticas del residuo i

Los modos normales obtenidos con el modelo de ANM predicen las fluctuaciones de los átomos, por lo que pueden utilizarse para predecir los *B-factors*. Como el desplazamiento de los átomos bajo cada modo normal es independiente del resto de los modos, la simple

suma de los desplazamientos que cada modo genera en un átomo i —ponderada por la frecuencia del modo—, es suficiente para calcular el desplazamiento cuadrático medio (Cui & Bahar 2006):

$$\langle \Delta r_i^2 \rangle = \frac{3k_B T}{\gamma} \sum_k^{3N-6} \frac{Q_{ik}^2}{\lambda_k} \tag{3.53}$$

Donde:

- k_B: constante de Boltzmann.
- T: temperatura.

Estos desplazamientos cuadráticos medios pueden utilizarse en la eq. 3.52 y así obtener los *B-factors* teóricos B_i^{teo} y su ajuste con los B_i^{exp} permite obtener una nueva γ, la γ' (Yang et al. 2009):

$$\gamma' = \frac{\sum_i^N B_i^{exp}}{\sum_i^N B_i^{teo}} \tag{3.54}$$

Donde:

- γ': nueva constante de resorte.
- B_i^{teo}: *B-factor* teórico del carbono alfa del residuo i.

Si esta nueva γ' se usa en la eq. 3.53, se obtendrán *B-factors* teóricos de escala apropiada.

La correlación de Pearson entre los valores teóricos y los experimentales permite evaluar la validez del modelo de ANM para la proteína de estudio:

$$\rho^{teo-exp} = \frac{\sum_{i=1}^N (B_i^{teo} - \langle B^{teo} \rangle)(B_i^{exp} - \langle B^{exp} \rangle)}{\sqrt{\sum_{i=1}^N (B_i^{teo} - \langle B^{teo} \rangle)^2} \sqrt{(B_i^{exp} - \langle B^{exp} \rangle)^2}} \tag{3.55}$$

Donde:

- $\rho^{teo-exp}$: correlación entre *B-factors* teóricos y experimentales.

Se considera que una correlación superior a 0.5 valida el uso del modelo de ANM (Yang et al. 2009) (Rueda et al. 2007) (Cui & Bahar 2006) (Bahar et al. 1997) (Atilgan et al. 2001).

3.1.4.2.2 Grado de colectividad en el movimiento de los residuos El grado de colectividad κ de un modo normal $\boldsymbol{Q}_k$ es una medida del número de residuos que son desplazados significativamente en la vibración descripta por ese modo(Tama & Sanejouand 2001) . Para cada modo normal $\boldsymbol{Q}_k$, de longitud 3N (recordando que N es el número de partículas) con elementos Q_{ik} , el grado de colectividad κ_k, se define como:

$$\kappa_k = \frac{1}{N} exp(-\sum_{i=1}^{N} (Q_{ik}^r)^2 ln(Q_{ik}^r)^2) \tag{3.56}$$

Donde:

- $(Q_{ik}^r)^2 = (Q_{ik}^x)^2 + (Q_{ik}^y)^2 + (Q_{ik}^z)^2$
- Q_{ik}^{xyz}: componente x, y o z del i-ésimo Carbono alfa en el modo normal k

Una colectividad κ que tienda a $\frac{1}{N}$ significa que el modo normal representa el movimiento de un bajo numero de residuos, mientras que una colectividad cercana a la unidad implica que el modo normal representa el desplazamiento de un gran número de residuos.

3.1.4.3 Adaptaciones utilizadas en esta tesis: constante de fuerza asociada al tipo de interacción

Como se mencionó en la sección el método de ANM utiliza una única constante de fuerza γ para todas las interacciones. Sin embargo, en ésta tesis se ha modificado parcialmente el método de manera de poder obtener una visión biológica más adecuada que contemple la información de tipos de interacciones entre residuos. Esta información es incorporada en el método de ANM mediante constantes de fuerza diferenciales (Barletta et al. 2018). Esto es, se utilizaron constantes de fuerza entre residuos que representaran aproximadamente el orden de magnitud relativo a cada tipo de unión, tomando como referencia la unión covalente. Para cada par de aminoácidos —representados por sus carbonos alfa y separados por una distancia menor a r_c—, se aplicaron las siguientes reglas para determinar el valor de la constante según el tipo de interacción:

γ	Covalente	Puente de hidrógeno	Otras
Valor	1.0	0.1	0.01

Las interacciones covalentes responden a carbonos alfa consecutivos y puentes disulfuro. Estos últimos son fácilmente detectables según las distancias entre los átomos de sulfuro. Mientras tanto, los puentes de hidrógeno fueron detectados con el software HBPLUS (McDonald & Thornton 1994). Las interacciones del resto de los carbonos alfa que se

encuentran a una distancia menor al radio de corte r_c fueron asignadas con el valor de interacción más débil.

El otro parámetro a ajustar es la distancia de *cutoff* r_c. Su valor óptimo se obtiene variándolo entre 7Å y 20Å y seleccionando el r_c que permite obtener el máximo valor de $\rho^{teo-exp}$, como se detalló en la anterior sección.

3.2 Dinámica esencial: Análisis de Componentes Principales

Una trayectoria de MD provee una multitud de conformaciones de la proteína con cuantiosa información de su dinámica en una ventana de tiempo dado. Una vez que se decide dar por terminada una trayectoria, el problema no es la escasez de información, sino la abundancia de esta. Rearreglos conformacionales colectivos y de gran reelevancia biológica están enmascarados por fluctuaciones térmicas de todas las partículas que carecen de interés biológico. El Análisis de Componentes Principales (o PCA por sus siglas en inglés) (García 1992) (Amadei et al. 1993) (aplicado al análisis de la dinámica de proteínas), PCA permite identificar los movimientos colectivos asociados a cambios conformacionales de interés biológico (Orellana et al. 2010) (Rueda et al. 2007) (Ichiye & Karplus 1991) (Groot et al. 1996).

Es una técnica que permite recuperar patrones representativos de datos con importante ruido estadístico. La idea es mapear al sistema estudiado de un espacio multidimensional complejo a un espacio reducido representado por unos pocos componentes principales (PCs), identificando así las características principales subyacentes a los datos obtenidos. Esto permite analizar las deformaciones estructurales observadas en términos de contribuciones individuales (los PCs) ordenadas según la varianza original que describen.

Para realizar PCA es necesario contar un conjunto de conformaciones, provengan estas de fuentes experimentales como el NMR o de una MD. Luego, el proceso consta de 3 pasos; el primero consiste en superponer las conformaciones con un ajuste de mínimos cuadrados(Kabsch 1978) a una conformación de referencia para eliminar los desplazamientos y rotaciones que la macromolécula pudo haber sufrido, como centro de masa, durante el experimento. Luego, se calcula la estructura promedio del ensamble y con ella, la matriz de covarianza de fluctuaciones según la siguiente ecuación (Ichiye & Karplus 1991):

$$\boldsymbol{C}_{ij} = \langle (x_i(t) - \langle x_i \rangle)(x_j(t) - \langle x_j \rangle) \rangle \tag{3.57}$$

Donde:

- $\boldsymbol{C}$: matriz 3Nx3N simétrica de covarianza, donde N es el número de átomos.
- $\langle x_{i(j)} \rangle$: coordenada i(j)ésima de la estructura promedio del ensamble.
- $x_{i(j)}(t)$: coordenada i(j)ésima de una estructura del ensamble.

Las coordenadas x se denotan en función del tiempo t por claridad, aunque esto no tiene que ser necesariamente así. $\boldsymbol{C}$ es una matriz simétrica que contiene, como elementos, las covarianzas entre los desplazamientos atómicos de todos los átomos. Átomos con movimientos correlacionados resultarán en covarianzas positivas, y aquellos con movimientos anticorrelacionados tendrán covarianzas negativas. Átomos sin desplazamientos lin-

ealmente relacionados resultarán en covarianzas cercanas a cero (0). Los elementos de la diagonal se corresponderán a la covarianza entre los desplazamientos de un mismo átomo, por lo que tendrán el valor de uno (uno) y carecerán de interés alguno.

Esta matriz es luego diagonalizada para así obtener un set de autovectores y autovalores. Los primeros serán los componentes principales y los últimos sus magnitudes.

$$\boldsymbol{C} = \boldsymbol{T}\boldsymbol{\Lambda}\boldsymbol{T} \tag{3.58}$$

Donde:

- $\boldsymbol{T}$: matriz de autovectores como columnas.
- $\boldsymbol{\Lambda}$: matriz diagonal de autovalores.

Los autovalores λ_i son los desplazamientos cuadráticos medios de los autovectores, por lo tanto, cuanto más altos sean, más alta es la magnitud de desplazamiento de la proteína a lo largo de su correspondiente autovector. 6 de ellos corresponderán a los desplazamientos y rotaciones del centro de masa en las 3 coordenadas, por los que serán cercanos a 0. Estos autovalores son descartados junto a sus autovectores correspondientes.

El método separa la trayectoría de las partículas en movimientos colectivos (que involucran a gran parte de las partículas) y movimientos más bien aleatorios, que son asociados a las fluctuaciones térmicas de toda molécula(Grubmüller 1995) (Groot et al. 1996). Esto permite analizar los movimientos colectivos por separado o en un subconjunto que agrupe a los movimientos más relevantes, ya que al desacoplar y eliminar las fluctuaciones no informativas de la proteína, el PCA resulta en un subespacio de baja dimensionalidad llamado Subespacio Esencial (ES, por sus siglas en inglés) que contiene los movimientos que suelen estar asociados a la función biológica. Por eso a esta dinámica también se la denomina *Essential Dynamics*, Dinámica Esencial(Amadei et al. 1993).

En contraste con NMA, el PCA de una dinámica molecular no asume que los movimientos de las proteínas son armónicos. Más aún, los principales modos obtenido del análisis de PCA representan movimientos anarmónicos (Hayward et al. 1995)(Amadei et al. 1993). Mientras que en NMA los principales modos son los de menor frecuencia, en el PCA puro los principales modos son los de mayor desplazamiento.

La Figura 3.1, extraída de (Hub & Groot 2009), muestra los 3 primeros vectores de PCA de la lisozima T4.

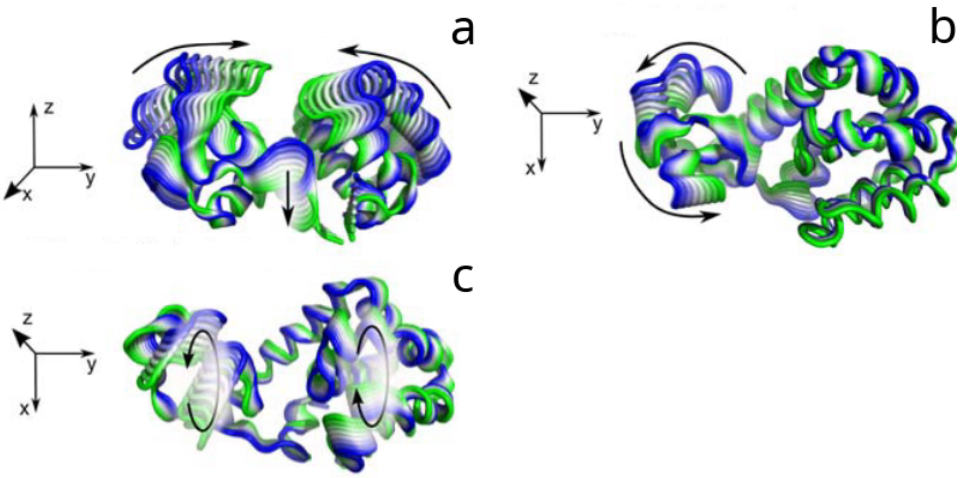

Figure 3.3: Los 3 modos de PCA de menor frecuencia (mayor desplazamiento) de la lisozima del viruts T4.**a** representa el modo de flexión, **b** el de giro y **c** el de torsión.

3.2.1 Análisis cuasiarmónico

El análisis cuasiarmónico es un método para obtener los modos de vibración de una macromolécula a partir de una dinámica molecular. Si además de los desplazamientos de los átomos se consideraran sus masas, se contaría con un campo de fuerza implícito y se podrían obtener los modos vibracionales de la macromolécula; cómo la dinámica molecular de la que provienen los desplazamientos utiliza un campo de fuerza anarmónico, los modos vibracionales obtenidos por ese método son llamados cuasiarmónicos. Si se toma la eq. 3.53 y en vez de calcular los desplazamientos cuadráticos medios de los átomos se calcularan los productos entre los desplazamientos medios de todos los pares de átomos, se obtendría: (Teeter & Case 1990)

$$\langle \Delta r_i \Delta r_j \rangle = \frac{3k_B T}{\gamma} \sum_k^{3N-6} \frac{Q_{ik} Q_{jk}}{\lambda_k} \tag{3.59}$$

Ahora la sumatoria representa un elemento de la inversa de la matriz de fuerzas:

$$\boldsymbol{F}_{ij}^{-1} = \sum_k^{3N-6} \frac{Q_{ik} Q_{jk}}{\lambda_k} \tag{3.60}$$

Donde (eq. 3.21):

$$\boldsymbol{F}_{ij} = \frac{\delta^2 V}{\delta x_i \delta x_j} \tag{3.61}$$

Así, la matriz de fuerzas y la matriz de fluctuaciones son inversas.
NMA partía de la matriz hessiana y calculaba los modos normales de vibración. Luego con

estos modos obtiene las fluctuaciones de los átomos. Si estas fluctuaciones se conocieran por otra vía (por ejemplo, una dinámica molecular), se podrían obtener nuevos modos. Estos son los modos cuasiarmónicos. Como las matrices de fuerzas y fluctuaciones tienen los mismos autovalores, los autovalores de la matriz $\boldsymbol{T}$ (eq. 3.58) serán las frecuencias cuasiarmónicas de estos modos.

Entonces, si las coordenadas de los átomos de la eq. 3.57 son ponderadas en masa, estos autovalores no serán los desplazamientos medios, sino las frecuencias de los autovectores y se tratará a estos desplazamientos de la proteína a lo largo de estos autovectores como armónicos y la frecuencia de estas oscilaciones provendrán de los autovalores(Brooks et al. 1995):

$$\omega_i = \sqrt{\frac{k_B T}{\lambda_i}} \tag{3.62}$$

Donde:

- ω_i: frecuencia del modo de PCA i.

Éste análisis se denomina análisis cuasiharmónico (Karplus & Kushick 1981) (Levy et al. 1984) (Teeter & Case 1990), en contraste con el análisis de modos normales que es puramente armónico. Esto se debe a que si bien permite modelar el movimiento de la proteína como una suma de osciladores armónicos, en ningún momento aproxima la superficie de energía potencial de la proteína a una parábola, ya que la matriz $\boldsymbol{C}$ fue llenada con información extraída de una MD con un campo de fuerza anarmónico (Ichiye & Karplus 1991) (Amadei et al. 1993) (Brooks et al. 1995).

Estos modos cuasiarmónicos pueden ser utilizados como los modos normales y comparados entre ellos. Pueden, o no, ser similares a los modos normales según la dinámica de la que provengan. Si ésta se mantuvo en un mismo mínimo, entoces los modos armónicos y cuasiarmónicos serán similares. Si la dinámica atravesó múltiples mínimos de energía, entonces la matriz de fuerzas (F) será más "suave" (frecuencias más bajas, las oscilaciones no son tan rígidas), correspondiendose a un movimiento más amplio (Teeter & Case 1990).

Por otro lado, si bien en principio se pueden utilizar las coordenadas de todos los átomos, también se pueden utilizar sólamente las coordenadas de los carbonos α. Esto resulta en una matriz $\boldsymbol{C}$ considerablemente más pequeña y, consecuentemente, con una diagonalización menos costosa. Aún así, los autovectores y autovalores de los movimientos colectivos principales extraídos de estas matrices reducidas serán similares a los obtenidos utilizando todos los átomos(Amadei et al. 1993) (Van Aalten et al. 1996). Por estas razones, el PCA suele considerar únicamente los desplazamientos de los carbonos α y ésta es la metodología seguida en esta tesis.

Así, los primeros autovectores —los de frecuencias más bajas—, conformaran un reducido subespacio esencial que abarcarán los movimientos de mayor relevancia biológica(Amadei et al. 1993). Estos vectores pueden ser comparados entre si, o contra los vectores de NMA, o contra los vectores PCA de otras dinámicas, por medio de productos escalares, o pueden ser comparados en conjunto los subespacios que conforman, por medio de productos matriciales, etc…(De Groot et al. 1996) (Hinsen 1998). Encapsular el concepto de dinámica de una proteína en conjuntos de vectores pone a disposición todas las herramientas del álgebra lineal para su estudio.

Chapter 4

Cavidades: clasificación, dinámica y los programas y métodos para estudiarlas

4.1 Introducción

En proteínas, la ausencia de átomos es tan importante como su presencia. A lo largo de su vida útil, toda proteína debe interaccionar con otros solutos para llevar a cabo su actividad y en toda interacción está envuelta al menos una cavidad. Las cavidades suelen ser clasificadas según su geometría y número de entradas (o salidas). Cavidades abiertas con más de una entrada son llamadas canales o poros, suelen encontrarse en proteínas transmembrana de todos los tamaños y a través de estos canales fluyen iones, agua y otras moléculas de mayor tamaño. De similar geometría pero con una sola entrada son los túneles, que suelen llevar hacia un sitio activo enterrado en el interior de la proteína; el túnel que lleva al oxígeno hacia el grupo hemo de las globinas bacterianas es quizás el ejemplo más conocido. No tan profundos son los bolsillos que también pueden alojar sitios activos, pero expuestos al solvente; ahí suelen catalizarse reacciones que necesitan del agua. Bolsillos ocluídos cumplen la función opuesta, excluyen al solvente y catalizan reacciones que necesitan de un ambiente hidrofóbico. Bolsillos muy superficiales son llamados surcos o hendiduras, estos suelen participar en interacciones proteína-proteína y como sitio de interacción son quizás los de más dificil identificación ya que su presencia en las superficies de las proteínas es ubicua(Brezovsky et al. 2013). La Figura 4.1 grafica estos conceptos.

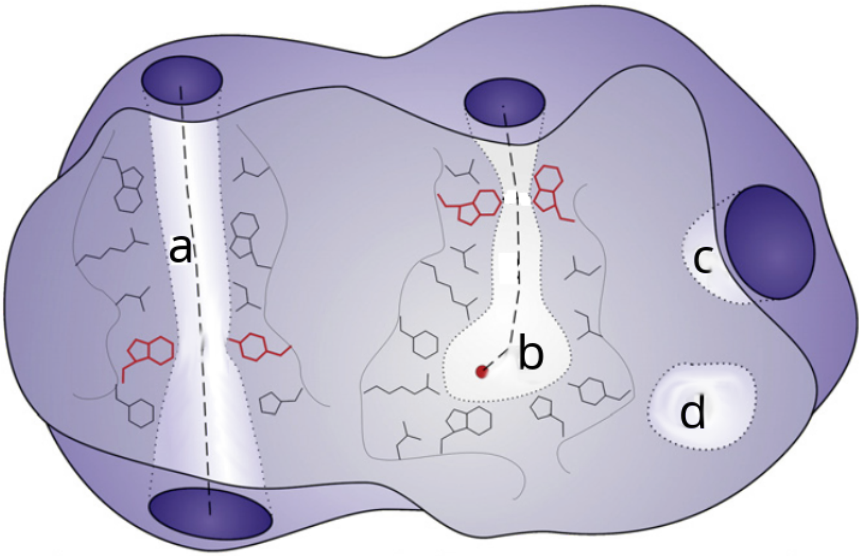

Figure 4.1: Figura esquemática con las posibles cavidades de una proteína. **a** señala el canal, **b** el túnel, **c** el bolsillo ocluído y **d** el bolsillo expuesto al solvente. Túneles y canales, por ser angostos, están sujetos a dinámicas particulares: pueden ser ocluídos con el movimiento de pocos residuos. En rojo se destacan los residuos ocupados del *gating* del poro y el túnel. Este tipo de cierre y apertura, Gora et al. denomina de "puerta vaivén" y suele ser llevada a cabo por residuos aromáticos.

4.2 Dinámica de cavidades

Las cavidades no son estáticas sino dinámicas. Los bolsillos ocluídos de ciertas enzimas hacen eviente esta diversidad estructural, la flexibilidad es necesaria en todos los tipos de cavidades. Los túneles de las porinas regulan su función por medio de su dinámica, la Figura 4.1 subraya en rojo a los residuos que participan de mecanismos de (*gating*), fenómeno que no sólo regula la función, sino que también puede aumentar la especificidad de una enzima por su sustrato (Zhou et al. 1998). Gora et al. define *gating* como un sistema dinámico que puede estar compuesto por unos pocos residuos o hasta dominios enteros que alternan conformaciones cerradas y abiertas, regulando el ingreso y egreso de moléculas pequeñas dentro y fuera de la proteína(Gora et al. 2013). Esta amplísima definición incluye la obstrucción de un túnel o un canal por parte de la rotación de un enlace de una cadena lateral, hasta la dinámica lenta de "respiración" que las porinas suelen tener, abriendo y cerrando el poro principal, movimiento que implica el desplazamiento de prácticamente todos sus residuos.

Stank et al. intentó otra clasificación en 5 tipos de dinámicas distintas, como figura en la tabla 4.1

Table 4.1: Clasificación de dinámica de cavidades de Stank et al..

Dinámica	Descripción
sub-pocket	Aparición o desaparición de un sub bolsillo a uno ya existente
adjacent pocket	Aparición o desaparición de un bolsillo aledaño a uno ya existente en la superficie de la proteína
breathing motion	Movimiento coordinado de los residuos de las paredes, ya sea por rotaciones de cadenas laterales, vibraciones de la cadena principal o vibraciones interdominio
channel	*gating* de residuos en un cuello de botella entre una cavidad y el solvente
allosteric pocket	Aparición o desaparición de un bolsillo alostérico, es decir, alejado del bolsillo original al que afecta

Quizás la última clasificación sea la más problemática, ya que requeriría la comprobación de que las dinámicas de los huecos está acoplada funcionalmente. Una clasificación simplificada sería:

Table 4.2: Variación de la clasificación de dinámica de cavidades de Stank et al., sin incluir alosterismo.

Dinámica	Descripción
sub-pocket	Aparición o desaparición de un sub bolsillo a uno ya existente
adjacent pocket	Aparición o desaparición de un bolsillo aledaño a uno ya existente en la superficie de la proteína
distant pocket	Aparición o desaparición de un bolsillo alejado de cualquier otro bolsillo que pueda o no existir
breathing motion	Movimiento coordinado de los residuos de las paredes, ya sea por rotaciones de cadenas laterales, vibraciones de la cadena principal o vibraciones interdominio

Dinámica	Descripción
channel	*gating* de residuos en un cuello de botella entre una cavidad y el solvente

4.2.1 Encaje inducido o selección conformacional

La dinámica de las proteínas se hace evidente en la cinética de unión a ligando. El capítulo primero resumió el modelo de paisaje energético conformacional en el que ambos modelos de unión a ligando están presentes: si el estado nativo de una proteína se entiende como una población de conformaciones que reparte su proporción según las estabilidades energéticas de las conformaciones, entonces la diversidad conformacional necesaria para la selección conformacional existe previamente a la aparición del ligando. Pero a la vez, este paisaje energético es perturbado por la cercanía del ligando, alterando así la dinámica de la proteína; así como lo sugiere el modelo de encaje inducido. Ambos modelos se utilizan para entender la cinética de unión a ligando(Okazaki & Takada 2008).

Aún así, la coexistencia de ambos mecanismos no evita la preponderancia de uno u otro. Modelos teóricos(Zhou 2010) y métodos experimentales(Gianni et al. 2014) se han desarrollado para determinar la relevancia de cada uno, encontrándose casos extremos en donde la cinética de unión es explicada por uno solo de estos modelos, como la unión de geldanamicina a la chaperona HSP90 (encaje inducido) (Gooljarsingh et al. 2006) y la unión de glucosa a la glucoquinasa humana (selección conformacional) (Kim et al. 2007). El primer caso revela la importancia de conocer el mecanismo de unión ya que su desconocimiento condujo a subestimar en 2 órdenes de magnitud la constante de unión entre la geldanamicina y HSP90, por llevar a cabo ensayos *in vitro* de muy corta duración que no permitieron alcanzar el equilibrio unión/disociación del ligando(Copeland 2011). En efecto, la escala de tiempo de unión está relacionada con el mecanismo, así como los tamaños relativos entre proteína y ligando, y las concentraciones de ambos(Hammes et al. 2009). Ligandos pequeños (en relación a la proteína), en baja concentración y proteínas de cambios conformacionales lentos, parecen favorecer la selección conformacional por sobre el encaje inducido (Greives & Zhou 2014).

Cualquiera sea el mecanismo de unión, la dinámica de la proteína —y por lo tanto, de la cavidad—, es fundamental para su función; esto se hace aún más relevante cuanto más flexible sea la proteína(Marsh et al. 2012). Como ya se ha dicho, la dinámica molecular (MD), el análisis de componentes principales (PCA) y los modos normales (NMA) son herramientas que aportan esta información en relación a la proteína. Pero para hacer foco en sus cavidades, son necesarias también otras herramientas.

4.3 Programas de detección de cavidades

Existe ya una gran variedad de programas para detectar cavidades y calcular su volumen, algunos también aportan datos fisicoquímicos de los aminoácidos que las recubren o los caminos que un potencial ligando recorrería dentro de la proteína para llegar a ella. Aunque la tarea parezca trivial, la definición de una cavidad no es unívoca entre los programas. La correcta definición de la cavidad es fundamental para determinar sus propiedades como volumen, accesibilidad al solvente, hidrofobicidad y flexibilidad, entre otras. Por eso es útil repasar los diferentes métodos existentes para la definición de cavidades y determinación de sus propiedades. Estos se pueden clasificar en 5 categorías, como Figura en la tabla 4.3. Esta clasificación intenta ser exhaustiva, pero no así los ejemplos dados, que lejos están de cubrir el abanico de herramientas disponibles (Stank et al. 2017) (Weisel et al. 2007) (Huang 2009).

Table 4.3: Clasificación de métodos utilizados para detectar y analizar cavidades.

Método	Ejemplo
Grilla	POVME
Esfera *gap* inscripta	SURFNET
Esfera rodante	Roll
Superficie	MSPocket (Pocasa)
Teselación	MOLE

4.3.1 Métodos de grilla

4.3.1.1 Cálculo de volumen

El método de grilla es quizás el más intuitivo. Como se verá, fué aplicado en numerosas herramientas(Laurent et al. 2014) y tiene como ejemplo a uno de los programas más citados en el área y que ya acumula su tercer versión, POVME(Wagner et al. 2017). También cabe aclarar que en esta sección nos referiremos a los métodos de grilla puros. Otros programas usan grillas en alguna etapa, pero no forman parte de su algoritmo principal de detección de cavidades, sino que cumplen funciones auxiliares.

Estos métodos dividen el espacio utilizando celdas cúbicas. El tamaño de las celdas es

definido como la **resolución** del método y sus centros suelen ser llamados **puntos**. Luego de divir todo el espacio en celdas, determina cuales están libres y cuales ocupadas. Esto se hace recorriendo todas las celdas de la grilla y evaluando lo siguiente: por cada celda, compara la distancia del punto (su centro) al átomo más cercano y el radio de Van der Waals de éste; si la distancia es menor al radio, la celda se dará por ocupada; si la distancia es mayor al radio, la celda se dará por libre. Luego de evaluar todas las celdas, el volumen de la cavidad se obtiene a partir del número de celdas libres.

Esta forma de evaluar la ocupación de las celdas plantea la primer desventaja del método: la grilla de puntos es equidistante, por lo que las celdas son cúbicas, pero el método para determinar si estas celdas están libres usa sólamente las coordenadas del centro de la celda y el centro del átomo, por lo que no podemos asegurar que la celda entera esté libre, sino sólamente su **esfera circunscripta**, cuyo centro coincide con el de la celda y cuyo radio es la mitad del tamaño de la celda (la resolución de la grilla). Este problema de la diferencia entre el volumen de una sección cúbica y el volumen de su esfera circunscripta suele llamarse como *sphere packing* y se ilustra, en 2 dimensiones, en la Figura 4.2.

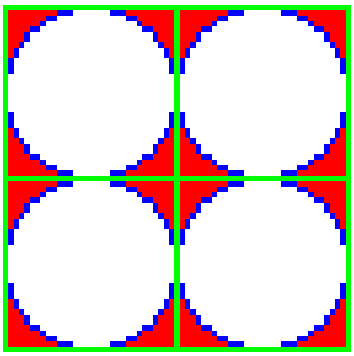

Figure 4.2: El problema de *sphere packing* en 2 dimensiones. Una grilla en 2 dimensiones en líneas verdes y círculos azules representan los átomos. En rojo, el área que no es evaluada.

Entonces, la disyuntiva está entre calcular el volumen libre utilizando la fórmula del volumen de un cubo o el de una esfera. La mayoría de los programas de grilla se limitan a calcular el volumen libre total, como la suma de los volúmenes de las esferas circunscriptas:

$$V_{total} = N\frac{4}{3}\pi(G/2)^3 \tag{4.1}$$

Donde:

- N: número de celdas libres.
- G: resolución de la grilla (tamaño de las celdas).

Naturalmente, la solución al problema de *sphere packing* es aumentar la resolución de la grilla, pero esto conlleva un alto costo computacional ya que la complejidad de este

algoritmo en su versión *naive* (sin optimizaciones) es cuadrática (Radhakrishnan et al. 2007): $\Theta(mn)$, donde m es el número de celdas de la grilla y n el número de átomos de la proteína. Ya que se deben calcular todas las combinaciones posibles entre todos los átomos y todas las celdas.

Además del problema de *sphere packing*, los métodos de grilla cuentan con otra desventaja: el volumen calculado depende de la orientación de la molécula. Los programas de grilla, por simplicidad, aplican la grilla en un sistema de referencia XYZ fijo en el espacio, si la proteína rotase, el volumen obtenido variará. Este problema es consecuencia de la discretización del espacio y la forma de mitigarlo sigue siendo aumentar la resolución.

4.3.1.2 Definición de la cavidad

Ninguno de los programas actualmente disponibles que utilicen métodos de grilla cuentan con métodos de detección de cavidades, por lo que el usuario deberá definirla por su cuenta. Esto puede parecer un inconveniente, pero en términos prácticos, permite un mayor control en la definición de la cavidad respecto a otros métodos. La forma predilecta de definir el área de búsqueda de la cavidad es por medio de figuras geométricas como esferas, prismas y cilíndros. Por ejemplo, si el usuario decide utilizar una esfera como definición de la cavidad, configurará su radio y posición y el volumen abarcado por esta esfera será dividido en celdas por la grilla y se dará comienzo al proceso ya descripto en la sección anterior. Sólo en el volumen abarcado por esta esfera se obtendrán cavidades. La Figura 4.3 fue extraída de (Durrant et al. 2011) y muestra el proceso entero llevado a cabo por éste tipo de herramientas.

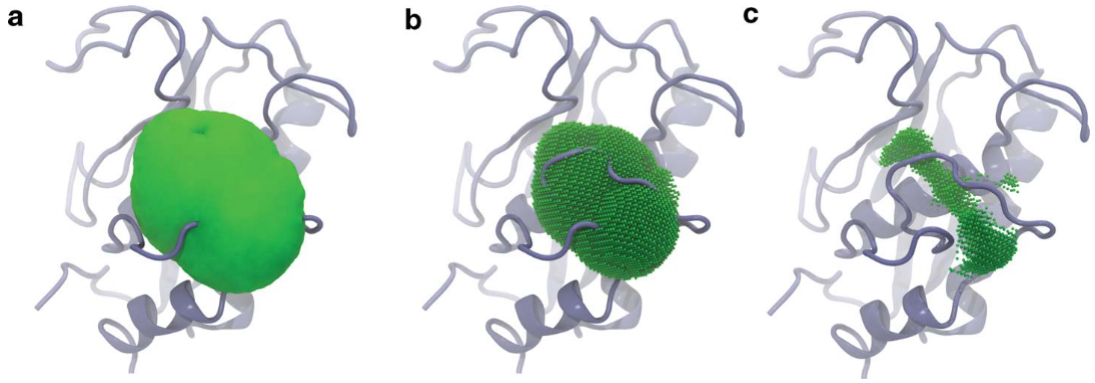

Figure 4.3: Los pasos de un programa de grilla puro. En **a** se utilizan 4 esferas (en verde claro) superpuestas para definir la cavidad. **b** muestra la grilla resultante (en verde obscuro) antes de eliminar las celdas ocupadas por átomos.**c** muestra el resultado final.

4.3.2 Esfera *gap* inscripta

El proceso seguido por este método está resumido en la Figura 4.4. Un buen ejemplo lo constituye la implementación en el software SURFNET (Laskowski 1995). Consiste en

generar esferas *gap* entre todos los posibles pares de átomos de la proteína, ocupando el espacio vacío entre los radios de Van der Waals. Luego se miden las distancias a los átomos vecinos y se reduce el radio de estas esferas gap para evitar superposiciones. Es decir, el tamaño de estas esferas *gap* se adapta de tal modo que no se superpongan con ninguna de las esferas centradas en cada átomo y con radio correspondiente a sus respectivos radios de Van der Waals. Si el radio de estas esferas es menor a un cierto mínimo predefinido, se descartan. Así, el espacio vacío en la proteína quedará recubierto por estas esferas *gap* que representarán a las cavidades. Finalmente, se superpone el conjunto de esferas *gap* en una grilla de celdas cúbicas y aquellas celdas de la grilla cuyos centros se encuentren dentro de las esferas *gap* representarán la cavidad (ver Figura 4.4.). Estos puntos serán "contorneados" (triangulados) para generar una superficie de fácil visualización. La Figura 4.4 fue extraída de (Laskowski 1995) y clarifica este proceso.

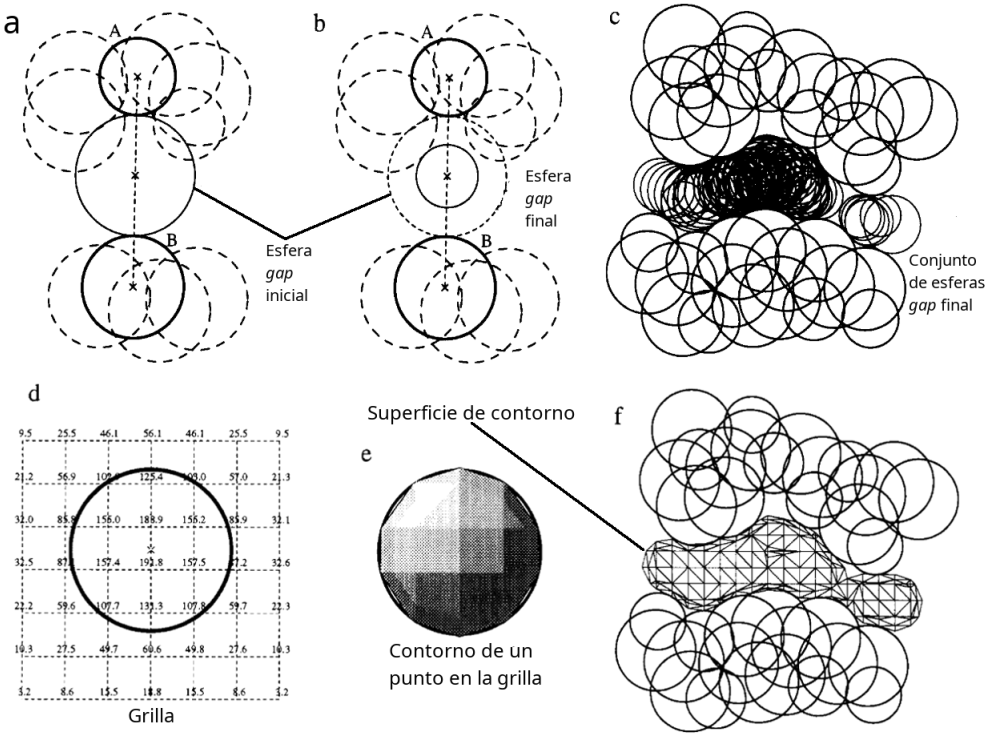

Figure 4.4: Los pasos de un programa de esferas *gap* inscriptas. **a** Cada esfera *gap* se coloca entre un par de átomos, como se muestra para los átomos A y B, a medio camino entre sus superficies de Van der Waals y tocando su perímetro. **b** Si algún átomo vecino penetra en esta esfera *gap*, su radio se reduce hasta que apenas toca el átomo intruso. Si el radio de la esfera cae por debajo de un mínimo (*threshold*) (por ejemplo, 1.0 A), se rechaza. De lo contrario, la esfera final se guarda. **c** Cuando se han considerado todos los pares de átomos, las esferas de espacio guardadas llenan la región de espacio. **d** Cada esfera se usa para actualizar puntos en una grilla 3D (aquí en 2D). **e** Cuando se contornea, las densidades en la matriz dan. Una superficie de contorno 3D hecha de polígonos que se aproxima a la esfera original. La superficie de contorno es un límite que delimita el tamaño y la forma de la región de espacio. Por otro lado, nótese el alto grado de superposición entre las esferas *gap*. Esto implica cálculos innecesarios.

La gran debilidad de estos métodos es el costo computacional. Como se advierte en la Figura 4.4 **c**, como consecuencia del algoritmo, el mismo espacio vacío es analizado repetidas veces y lo mismo sucede con el espacio ocupado por los átomos. Esto aumenta, innecesariamente, el costo computacional.

4.3.3 Esfera rodante

La mayoría de los métodos de esfera rodante (Voss & Gerstein 2010)(Yu et al. 2009)(Smart et al. 1996), se basan en modificaciones introducidas al método original de (Lee & Richards 1971) . A diferencia de los métodos anteriores, este método no fue desarrollado para estudiar cavidades sino para definir y estudiar la superficie de las proteínas. Consecuentemente no es la mejor opción para lidiar con bolsillos ocluídos, pero obtiene a cambio una alta eficacia al capturar la geometría de los bolsillos poco profundos llamados surcos.

El método de (Lee & Richards 1971) entiende a la proteína como un conjunto de esferas centradas en los átomos y de radio igual al radio de Van der Waals de su respectivo átomo. Luego, define 3 superficies: la **Superficie de Van der Waals** (SVdW) está conformada por la frontera del conjunto de esferas, es decir, sigue la línea exacta del perímetro de las esferas expuestas al solvente. También define una **sonda** como una esfera de cierto radio (1.4Å para el agua) que representa al solvente. Esta sonda recorre toda la SVdW definiendo en su camino a las otras 2 superficies: la **Superficie Expuesta al Solvente** (SAS, por sus siglas en inglés) será la traza del centro de la sonda y la **Superficie Excluída del Solvente** (SES, por sus siglas en inglés) será la traza del punto de contacto entre la sonda y las esferas centradas en los átomos. La **traza** se define como la línea que describe la trayectoria seguida por un punto dado de la sonda al recorrer la proteína. Como ya se dijo, si ese punto es el centro de la sonda, la traza describirá a la SAS; en cambio, si el punto es el punto de contacto entre la sonda y los átomos de la proteína, la traza describirá a la SES. La Figura 4.5 ilustra la diferencia entre estas 2 últimas superficies:

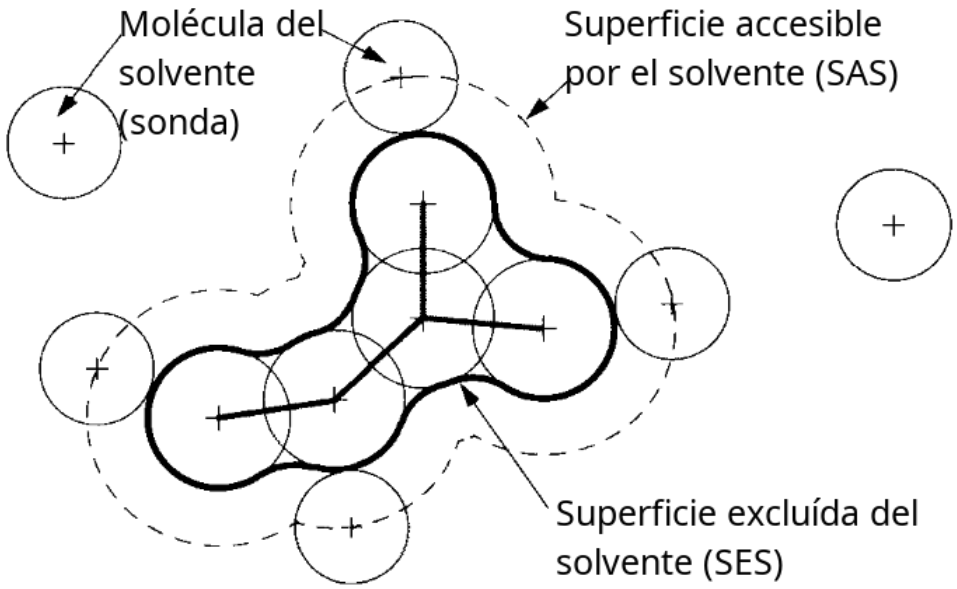

Figure 4.5: La SAS en línea punteada y la SES en contínua.

Los programas de esfera rodante obtienen la SAS, buscando los puntos en el espacio donde la sonda pueda ser colocada y mantenga contacto superficial con 3 átomos de la proteína,

sin entrar en conflicto con otros átomos, como muestra la Figura 4.6. Estos puntos y su traslado (al rodar la sonda) definirán la SAS.

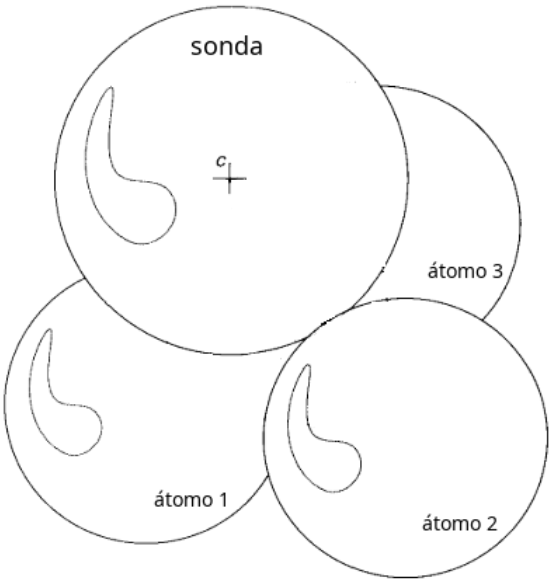

Figure 4.6: El punto **c** será uno de los puntos de la SAS.

El resultado final será la superficie de contacto entre la sonda y la proteína; superficie generada a lo largo de todo el recorrido de la sonda. Naturalmente, cuanto más grande sea la sonda más espacio quedará entre esta y la superficie de la proteína, por la misma razón, si la sonda se reduce a un infinitésimo, descartará todo surco posible ya que recorrerá toda la superficie de todo átomo, como muestra la Figura 4.7. En este límite, las 3 superficies: SVdW, SAS y SES, son idénticas.

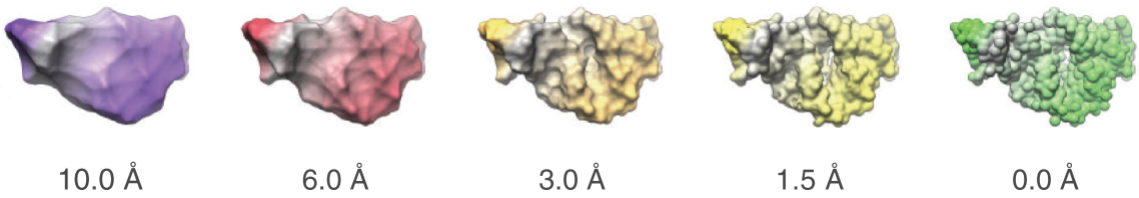

Figure 4.7: Las distintas SASs obtenidas al variar el radio de la sonda (en angstroms).

Por esta razón se suelen alternar distintos tamaños de sonda. La Figura 4.8 muestra el caso del programa 3V (Voss & Gerstein 2010): una sonda de 1.4 Å que simula al agua y otra sonda de mayor tamaño (llamda sonda *shell*) que diferenciará el volumen cercano a la proteína del resto del solvente.

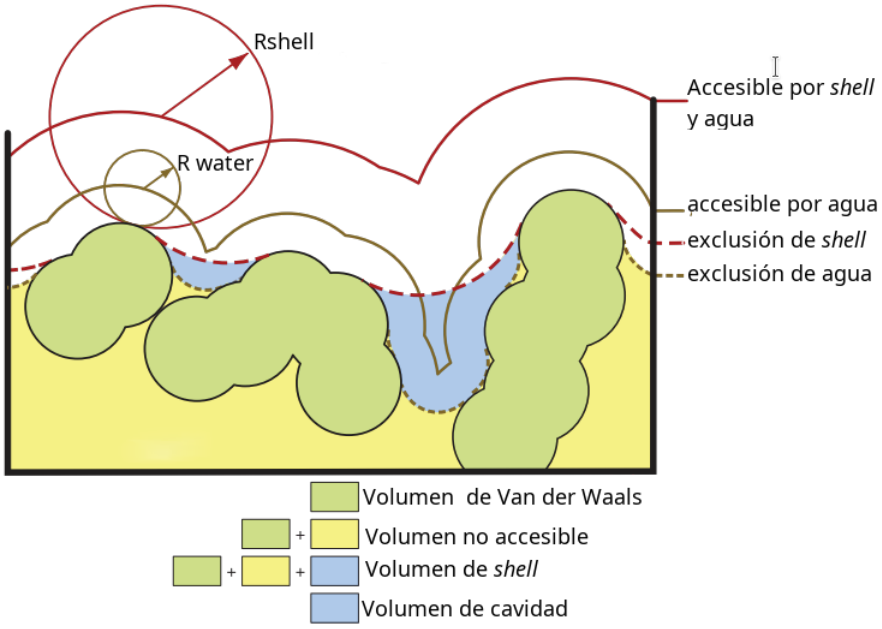

Figure 4.8: Figura extraída del manuscrito de la herramienta 3V. Define a la cavidad como la zona que queda entre el perímetro de la sonda *shell* y la sonda de agua.

En la versión del método implementada en el programa Roll(Yu et al. 2009) (Figura 4.9) también se superpone una grilla para obtener la geometría de la cavidad. Los puntos que no se superponen con la proteína pero si lo hacen con la sonda, serán los puntos de la cavidad.

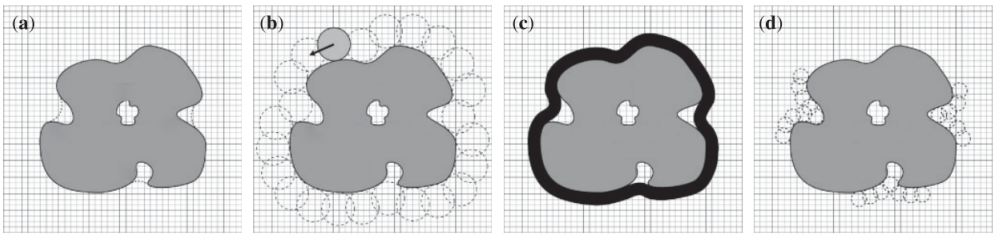

Figure 4.9: **a** Roll superpone una grilla sobre el sistema y luego comienza con el sondeo. **b** muestra todas las posiciones que adopta la primera sonda, mientras que **c**, la superficie expuesta al solvente. **d** muestra las cavidades alcanzadas por una segunda sonda más pequeña que la anterior.

Si bien estos programas no son los más eficientes y tienen un campo de aplicación acotado, los surcos (bolsillos poco profundos), son excelentes opciones para analizar este tipo de cavidades ya que aproximan la verdadera forma de la cavidad con buena precisión. Por otro lado, la sonda puede ser utilizada para modelar un ligando de interés y obtener una aproximación de las posibles interacciones entre la proteína y sus ligandos.

4.3.4 Superficie

Las definiciones de SVdW, SAS y SES de (Lee & Richards 1971) fueron utilizadas por (Sanner et al. 1996) para desarrollar un método que no sólo calcule la SAS, sino también

la superficie excluída del solvente (SES, por sus siglas en inglés). Como se ve en la Figura 4.5, cuando la sonda recorre la proteína, la SAS queda definida por la traza del centro de la sonda que recorre la superficie de la molécula, mientras que la zona de exclusión del agua (el solvente) queda definida por la traza del punto de contacto entre ésta sonda y las esferas que representan a los átomos de la proteína.

(Sanner et al. 1996) define la **superficie reducida** que es usada para obtener la SAS y la SES. Esta superficie reducida es una triangulación de la proteína. El proceso es el siguiente: Cada vez que la sonda, en su recorrido por la superficie de la proteína, se pone en simultáneo contacto con 3 átomos, estos 3 forman un **triángulo** (como se mostró en la Figura 4.6). Este triángulo será una cara de la superficie reducida. Al igual que en los métodos de sonda puros, el centro de la sonda será un punto de la SAS. Sin embargo, para obtener los puntos de la SES se realiza un nuevo procedimiento: se utilizará el triángulo definido por los 3 átomos y estos 3 centros más el centro de la sonda formarán un **tetrahedro**. La intersección entre la esfera de la sonda y este tetrahedro es una superficie esférica y está superficie es parte de la SES. Y así es como el método obtiene:

- una cara de la superficie reducida (un triángulo)
- un punto de la SAS
- una cara de la SES (un parche esférico)

La Figura 4.10 ilustra todo esto:

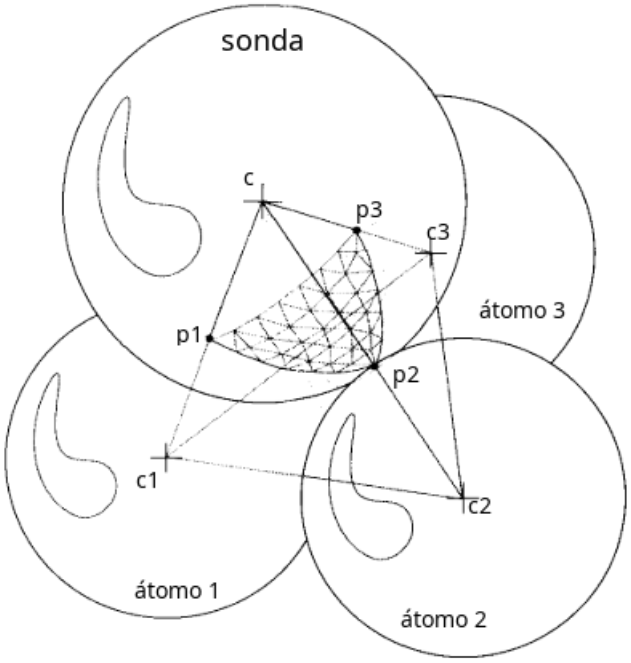

Figure 4.10: $\boldsymbol{c1}$, $\boldsymbol{c2}$ y $\boldsymbol{c3}$ son los centros de los átomos 1, 2 y 3, que definen un triángulo de la superficie reducida. Junto al centro de la sonda $\boldsymbol{c}$ y con la intersección de la misma, definen la superficie esférica $(\boldsymbol{p1}, \boldsymbol{p2}, \boldsymbol{p3})$. Ésta pertenece a la SES, mientras que el punto $\boldsymbol{c}$ pertenece a la SAS.

Ahora bien, cuando la sonda ruede por encima de 2 átomos (una arista de la superficie reducida, es decir, un lado del triángulo) a un nuevo triángulo, la traza de su centro dibujará un arco, este arco conecta 2 puntos de la SAS. A su vez, uno de los átomos perderá contacto con la sonda y la superficie esférica que pertenecía a la SES pasará a ser un arco, este arco dibujará un parche de una superficie toroidea. Este parche también pertenece a la SES. Así, el método obtuvo:

- un arco de la SAS
- una cara de la SES (superficie toroidea)

La Figura 4.11 intenta ilustrar lo que ocurre cuando la sonda se desplaza:

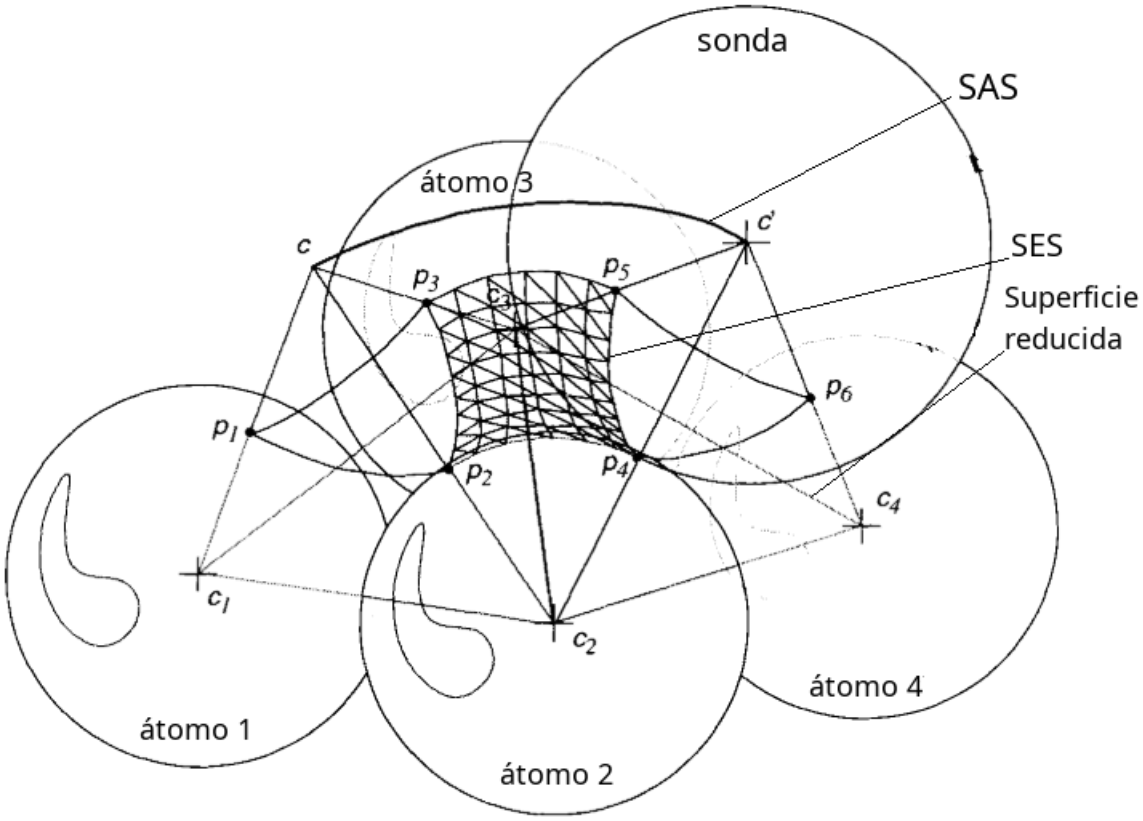

Figure 4.11: La sonda comienza en el triángulo de la superficie reducida $(\boldsymbol{c1}, \boldsymbol{c2}, \boldsymbol{c3})$ y luego se desplaza al triángulo $(\boldsymbol{c2}, \boldsymbol{c3}, \boldsymbol{c4})$. En su camino define el arco $(\boldsymbol{c}, \boldsymbol{c'})$, perteneciente a la SAS. También define la superficie esférica $(\boldsymbol{p1}, \boldsymbol{p2}, \boldsymbol{p3})$ antes de su desplazamiento y durante la transición también define la superficie toroidea $(\boldsymbol{p2}, \boldsymbol{p3}, \boldsymbol{p4}, \boldsymbol{p5})$, ambas pertenecientes a la SES.

Esta serie de triángulos conforma una triangulación y esta es la superficie reducida. Como ejemplo, la Figura 4.12 muestra la superficie reducida de la crambina.

Figure 4.12: Superficie reducida de la crambina. Las esferas son los átomos, conectados por ejes blancos. La superficie reducida se muestra en bordó.

MSPocket(Zhu & Pisabarro 2011) es una aplicación que usa éste método de superficie para detectar y reconstruir las cavidades. MSPocket integra la rutina de triangulación y sondeo descripta para detectar cavidades. Y en esa triangulación, buscará los vértices que pertenezcan a una cavidad. El proceso para descubrir una cavidad es el siguiente:

1. Se calcula el vector normal a la superficie de cada triángulo obtenido mediante el procedimiento ya descripto (ver Figura 4.10).
2. Los vértices de cada triángulo se desplazan en la dirección del vector normal.
3. Si el desplazamiento del paso **2** acercó los vértices de 2 triángulos distintos, entonces esos triángulos son parte de una cavidad. Estos triángulos, que forman parte de alguna cavidad, serán conservados, mientras que el resto serán descartados.

Luego sigue el proceso para calcular el volumen de la cavidad encontrada. Una vez que se dispone de los triángulos (que incluyen a sus vértices) que conforman una cavidad, se calcula el centro de masa de los vértices (átomos) que conforman la cavidad y se definen tetrahedros entre este centro de masa y los vértices de los triángulos. La suma de los volúmenes de estos tetrahedros, es el volumen de la cavidad. La Figura 4.13 resume estos últimos pasos.

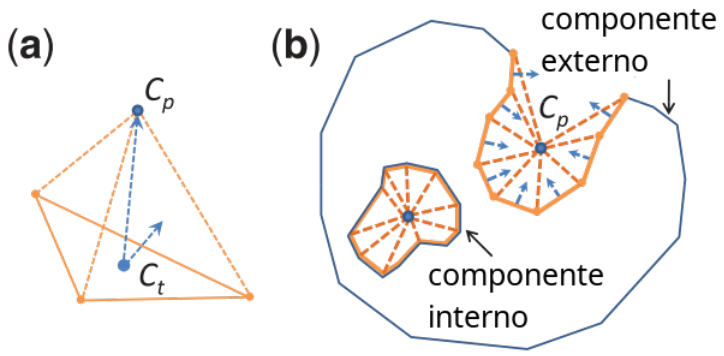

Figure 4.13: **a** C_p es el centro de masa de los vértices (átomos) que conforman una cavidad, junto a los vértices de uno de los triángulos conforma un tetrahedro. **b** intenta mostrar, en 2 dimensiones, la conformación de una cavidad y su centro de masa C_p. Si se suman las superficies de los triángulos formados por C_p y los vértices, se obtendría la superficie de la cavidad. Lo mismo ocurre en 3 dimensiones, sólo que en vez de calcular superficies de triángulos, se calculan volúmenes de tetrahedros. MSPocket separa los componentes externos (bolsillos), de los internos (bolsillos ocluídos).

MSPocket reporta la superficie de cada cavidad como la suma de las superficies de los triángulos de la superficie reducida y el volumen de la cavidad como el volumen de los tetrahedros que se forman entre los triángulos que pertenezcan a cavidades y el centro de masa de las mismas (C_p, ver Figura 4.13). Esta metodología esta sujeta a errores por la aproximación de superficies y volúmenes irregulares a triángulos y tetrahedros.

4.3.5 Teselación

La teselación de una superficie es una subdivisión en figuras geométricas (polígonos), sin superposición ni espacios vacíos. En el caso de una proteína esto implica, como en métodos anteriores, modelarla como un conjunto de esferas centradas en sus átomos y de radio igual a su radio de Van der Waals respectivo. La teselación de los centros de estas esferas será un conjunto de polihedros que permitirán inferir información estructural de la proteína. Si bien ya se habían utilizado algunos métodos de teselación para estudiar la superficie de las proteínas(Sanner et al. 1996), fue Edelsbrunner quien utilizó los métodos de la geometría computacional intensivamente para detectar y analizar cavidades (Edelsbrunner et al. 1998) (J Liang, H Edelsbrunner & Woodward 1998) (Jie Liang et al. 1998) (J Liang, H Edelsbrunner, P Fu, et al. 1998). Luego fueron desarrollados CAVER(Petrek et al. 2006), Fpocket(Le Guilloux et al. 2009) —junto a su versión destinada a analizar resultados de MD, MDPocket (Schmidtke et al. 2011)—, y MOLE(Petřek et al. 2007) (Sehnal 2013) (Berka et al. 2012). 2 de estos, CAVER y MOLE, se especializan en túneles y canales, y fueron creados por el mismo desarrollador y tienen principios similares. CAVER, al igual que tantos, utiliza una grilla pero también agrega el algoritmo de *Convex Hull* (caparazón convexo) para diferenciar el interior y el exterior de la proteína.

4.3.5.1 *Convex Hull* (CH)

El caparazón convexo (CH, por sus siglas en inglés) de un conjunto de puntos es el polígono convexo más pequeño —es decir, definido por la menor cantidad de puntos—, que contiene a todos los puntos. La Figura 4.14 muestra un ejemplo de lo que esto representa en el caso de puntos en el espacio bidimensional.

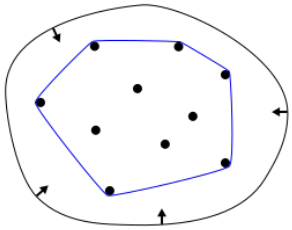

Figure 4.14: Analogía de los clavos y la banda elástica para el CH. Si los puntos de la imagen fueran clavos en una mesa vistos desde arriba, la banda elástica azul representaría el CH de este conjunto de puntos. La serie de clavos (puntos) que están en contacto con la banda elástica (CH) y el orden en el que deben ser conectados son suficientes para definir el CH. Esto es equivalente a definir los segmentos del CH.

El CH puede obtenerse para puntos en cualquier número de dimensiones y en el caso de las proteínas, es una primera y rudimentaria aproximación a su topología, como muestra la Figura 4.15, extraída de (Berka et al. 2012):

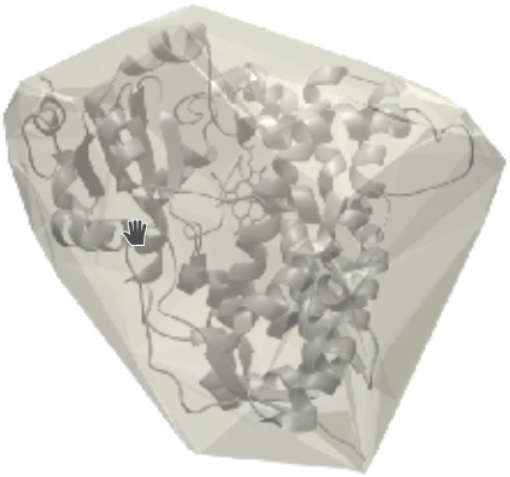

Figure 4.15: CH de una proteína. Las caras triangulares del CH dejan ver la proteína a partir de la cual fue calculado.

Como ya se ha dicho, CAVER combina el CH con el uso de una grilla, lo que le permite a CAVER hallar cavidades, evaluar cuáles de éstas tienen un acceso al solvente y luego determinar el camino óptimo hacia el interior de cada una (ver Figura 4.16). Pero antes de entender el funcionamiento de CAVER, se repasan algunas definiciones:

4.3.5.1.1 Grafo Un grafo $G = (V, X)$ es un par de conjuntos, donde V es un conjunto de puntos o **vértices** y X es un subconjunto del conjunto de pares no ordenados de

elementos distintos de V. Los elementos de X se llaman **ejes**, aristas o arcos y relacionan a 2 vértices de V "conectándolos". Dicho de otra forma, si v y w son 2 vértices de V y existe un eje $e = (v, w)$ que los conecta, se dice que v y w son adyacentes. Por no ser orientado, se dice que $e = (v, w) = (w, v)$. Si los ejes fueran orientados, estaríamos hablando de un grafo orientado y a éste tipo de grafo se lo llama **digrafo** (Radhakrishnan et al. 2007). En la Figura 4.16 se encuentran ejemplos de grafos y digrafos.

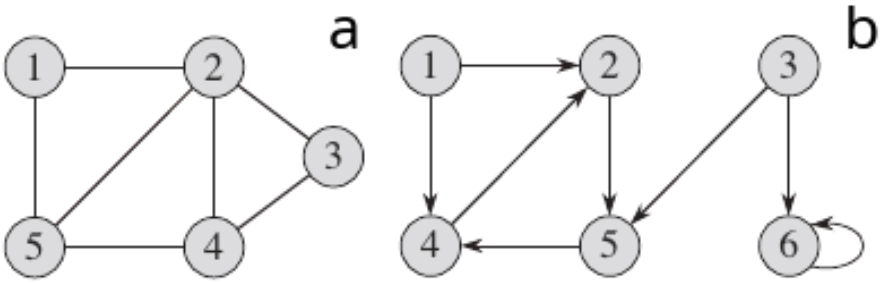

Figure 4.16: **a** grafo con 5 vértices conectados por sus ejes. **b** digrafo de 6 vértices conectados por sus ejes. Notar que los vértices **1** y **2** son mutuamente adyacentes en el grafo de **a**, pero en el digrafo de **b**, **1** es adyacente al nodo **2**, pero éste no es adyacente al nodo **1**.

El programa CAVER determina las cavidades con aceso al solvente modelando la grilla como un digrafo donde las celdas libres son vértices conectados entre sí por medio de ejes ponderados y se utiliza el algoritmo de Dijsktra para encontrar el/los camino/s más corto/s entre el exterior e interior de la proteína (Pavelka et al. 2016).

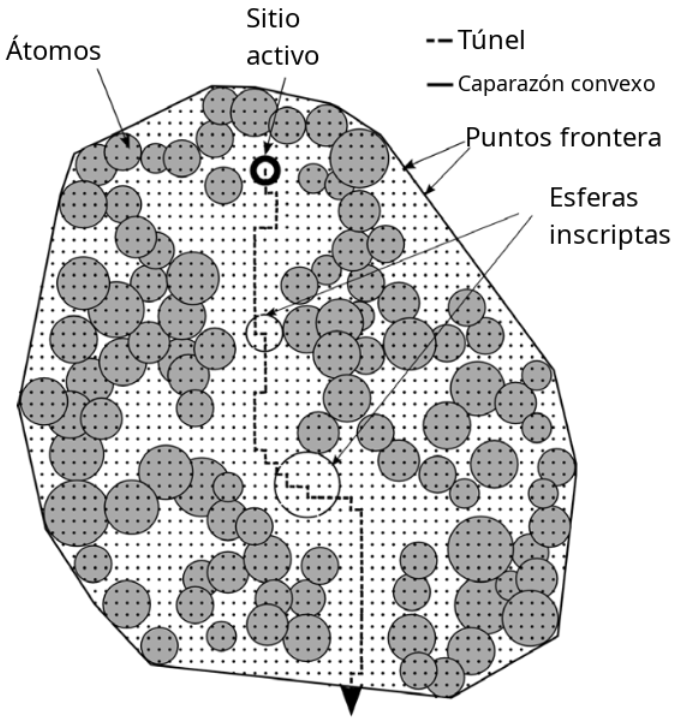

Figure 4.17: Extraído de la publicación original de CAVER. Los puntos de la grilla que se encontraban en el exterior del CH fueron eliminados y sus puntos contiguos son adjetivados *frontera*, el camino óptimo que conecta uno de estos con el interior de la proteína será el túnel reportado por CAVER. A cada punto del camino se determina el número máximo de puntos vecinos libres, y así el radio de las esferas inscriptas; el radio de la esfera más pequeña será el *bottleneck radius*, el punto más angosto del túnel. Esta información puede ser utilizada por el usuario para determinar si un ligando es capaz de ingresar a la proteína por ese túnel.

La ponderación de los ejes (sus pesos) corresponde al radio de la esfera inscripta del punto de la grilla al que conducen. Así, los puntos que tracen un camino más amplio serán los seleccionados como parte del camino, como se puede ver en el ejemplo de la Figura 4.18.

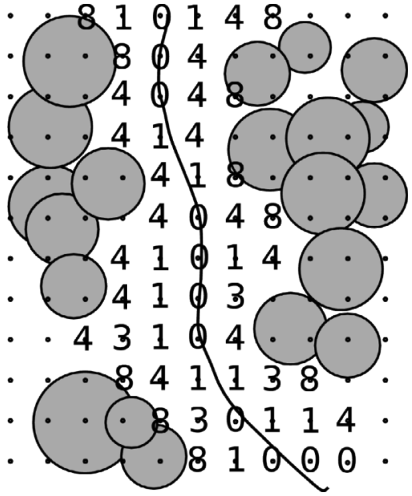

Figure 4.18: Extraído de la publicación original de CAVER. Las esferas de una proteína, la grilla de CAVER y los pesos (costos) de cada eje. La línea recorre el camino óptimo que el algoritmo de Dijkstra trazaría. Se entiende que éste es el camino que seguiría un ligando y por lo tanto representa al eje axial del túnel.

Ahora bien, la alta densidad de puntos de una grilla resulta en una redundancia de puntos

para representar cada cavidad, por lo que se detectan varios caminos alternativos cercanos, que en realidad no son más que el mismo túnel. Esto implica un alto costo computacional para el algoritmo de Dijsktra. Para solucionar esto se desarrolló MOLE, que representa los mismos caminos pero en vez de utilizar celdas de una grilla, utiliza puntos del Diagrama de Voronoi (VD, por sus siglas en inglés) de la proteína (Medek et al. 2007).

4.3.5.2 *Voronoi Diagram* (VD)

El VD de un conjunto de puntos $\boldsymbol{P}$, a los que llamaremos **nodos**, es una subdivisión del espacio tal que cada partición (cara de Voronoi) contenga a 1 nodo de $\boldsymbol{P}$ y a todo el espacio cuyo nodo de $\boldsymbol{P}$ más cercano sea ese mismo nodo. La Figura 4.18 contiene un simple ejemplo en 2 dimensiones.

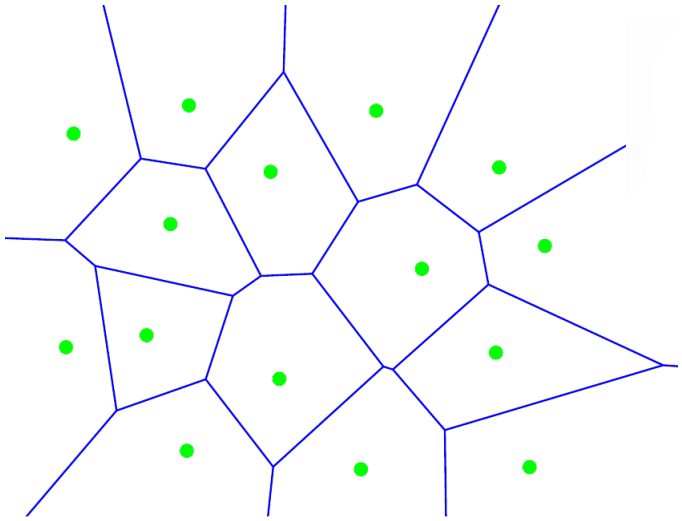

Figure 4.19: El VD del conjunto de puntos verdes. Observar que el número de aristas de una cara de Voronoi no es definido; estos polígonos pueden estar compuestos por 3 o más aristas.

El VD de un conjunto de puntos (nodos) en 3 dimensiones estará especificado por las aristas y caras de los diagramas de Voronoi. Como se observa en la Figura 4.19, cada arista de un VD pasa entre 2 nodos del diagrama, y por la definición del VD, a cada punto de esta arista, las distancias a esos nodos, serán iguales. Es decir, la arista pasa por el centro exacto de estos 2 nodos.

Para el caso de una proteína, estos nodos serán sus átomos y las aristas serán los posibles segmentos que componen los distintos túneles. Así, una vez calculada la cavidad de la proteína y el VD, se puede estimar el camino óptimo de un ligando entre el exterior e interior de una proteína. Al igual que CAVER, MOLE modelará el problema como un digrafo, las aristas del VD serán los ejes y sus nodos, los vértices. Una vez más, los ejes serán ponderados por el espacio vacío que los rodea, teniendo en cuenta los radios de Van der Waals de los átomos. Esto se ejemplifica en la Figura 4.19:

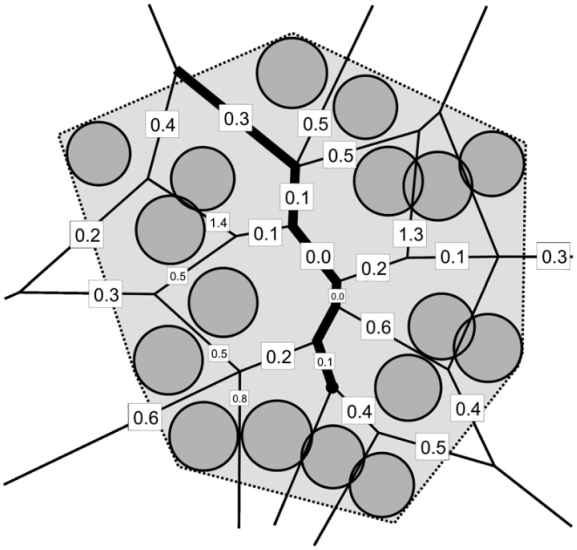

Figure 4.20: Extraído del manuscrito de MOLE. Una vez más, átomos representados como esferas, la zona gris representa el interior de la proteína y las líneas negras son las aristas de extraídas de Voronoi, con sus pesos (obtenidos según los radios de Van der Waals de los átomos que separan). Como el VD desconoce esferas y sólo trata con puntos, aparecerán aristas que se superponen con los átomos y no representan espacio vacío alguno; estas aristas tendrán un peso (costo) alto, que las excluirá de cualquier túnel posible. La linea negra gruesa marca el camino óptimo desde el exterior hacia el interior. Este camino será hallado por el algoritmo de Dijkstra.

Finalmente, y al igual que en CAVER, se obtendrán las esferas inscriptas en cada nodo que se encuentra en el camino del túnel. En CAVER estos nodos eran puntos de la grilla, mientras que en MOLE son vértices del VD. A continuación, se ahondará en el método de obtención del VD usado en MOLE, llamado Triangulación de Delaunay. Como MOLE carece de grilla, este método también es utilizado para obtener los espacios vacíos dentro de la proteína y es el método fundamental dentro de todos los programas de detección de cavidades por medio de teselaciones.

4.3.5.3 *Delaunay Triangulation* (DT)

Si un VD se entiende como un grafo, la Triangulación de Delaunay (DT, por sus siglas en inglés) de un conjunto de puntos será el grafo dual de ese VD.

4.3.5.3.1 Grafo dual El grafo dual H de un grafo G es un grafo con un vértice por cada cara del grafo G y un eje por cada vez que 2 caras del grafo G compartan un lado (eje). En estos grafos planos, la dualidad es simétrica, por lo que G será también el grafo dual de H (Radhakrishnan et al. 2007).

La Figura 1.20 agrega la DT a la Figura 1.18:

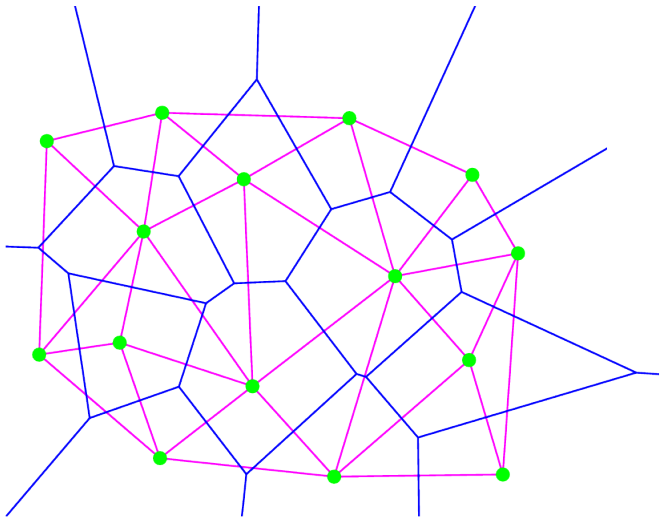

Figure 4.21: VD y DT del conjunto de puntos verdes. En azul el VD y en violeta la DT.

Así, todos los algoritmos de teselación empiezan realizando una DT del conjunto de puntos (átomos) y luego, a partir de la DT, obtienen el VD. La DT de un conjunto de puntos $\boldsymbol{P}$ ($DT(\boldsymbol{P})$) es una triangulación tal que no existen puntos de $\boldsymbol{P}$ dentro de los círculos inscriptos en los tríangulos de $DT(\boldsymbol{P})$, como se ve en la Figura 1.21.

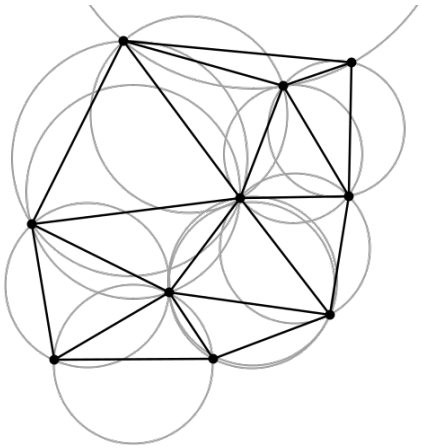

Figure 4.22: DT de ejemplo, con sus círculos inscriptos delineados en gris. 3 puntos son necesarios para especificar una circunferencia, estos son los puntos de cada triángulo. Notar que cada circunferencia está vacía de todo punto

Luego de obtener la DT, el paso siguiente para calcular el VD es obtener los centros de los círculos inscriptos, como aparecen en la Figura 4.22. Estos centros serán los vértices de Voronoi. Como ya se dijo, al obtener el grafo dual de un grafo, cada cara del grafo (tríangulo en este caso) será reemplazada por un vértice.

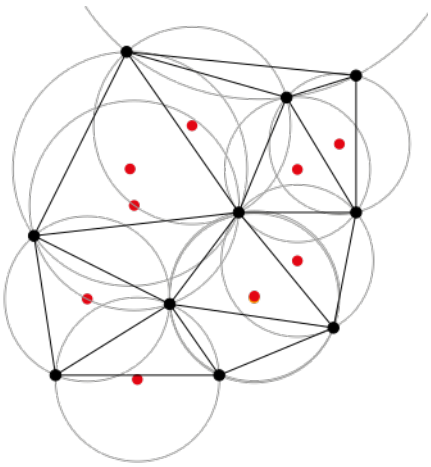

Figure 4.23: La misma triangulación de la figura anterior, pero con los centros de los círculos especificados como puntos rojos.

Finalmente, los vértices de Voronoi serán conectados entre sí por ejes, si provienen de triángulos que compartían un lado (Figura 1.23)

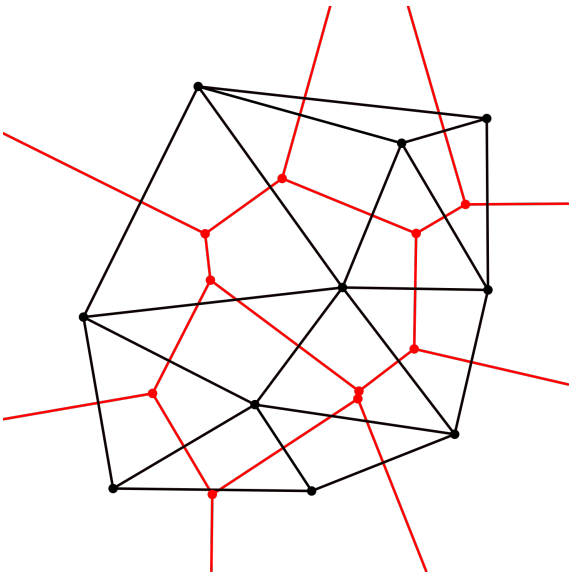

Figure 4.24: En negro, la DT y el VD en rojo.

Esta triangulación es el paso fundamental de todos los métodos de teselación, donde lo que se obtienen no son triángulos, sino ya tetrahedros. Tetrahedros vacíos que serán el punto de partida para obtener las cavidades. Átomos muy próximos darán tetrahedros pequeños que pueden ser descartados rápidamente, mientras que aquellos más grandes y agrupados en cercanía, serán indicios de cavidades proteicas.

Chapter 5

ANA: Analysis of Null Areas

5.1 Introducción

En el capítulo anterior se familiarizó al lector con el estado actual de los programas de detección y cálculo de cavidades, y los métodos en los q se basan. Si bien éste área se desarrolla activamente desde los años 90, al día de hoy se siguen publicando nuevos métodos (Chen et al. 2019) (Krone et al. 2016) (Simões et al. 2017) y algunas de las deficiencias se sostienen. Estas deficiencias se deben, en su mayoría, a la consideración de la proteína como una estructura única. Pocos programas son capaces de analizar cavidades a lo largo de trayectorias de dinámica molecular y si bien el análisis de modos normales y el estudio de las cavidades son 2 campos de mucha actividad, ningun programa es capaz de combinar ambas herramientas. La baja eficiencia en la detección de cavidades es otra consecuencia de analizar estructuras únicas. Con el advenimiento de los programas de dinámica molecular adaptados a placas gráficas, las trayectorias largas se han vuelto ubicuas y los programas de detección de cavidades deben también acelerar su procesamiento para lidiar con la gran cantidad de información alojada en una trayectoria. En este contexto, y con la intención de solucionar algunos de estos problemas, se desarrolló ANA (por sus siglas en inglés, *Analysis of Null Areas*).

5.2 Definición de la cavidad: *Included Area*

5.2.1 Estrategias utilizados por otros programas

ANA fue desarrollada para su aplicación en el campo de la dinámica molecular y el análisis de modos normales. En estas áreas las cavidades de interés suelen ser conocidas con antelación y la comunidad prefiere tener pleno control sobre la definición de la cavidad; POVME es uno de los programas más utilizados en estos casos y utiliza una definición

manual(Wagner et al. 2017). Epock, más reciente, también emplea una estrategia manual(Laurent et al. 2014). Estos métodos manuales ofrecen figuras geométricas predeterminadas para definir las cavidades. Esferas, cubos y cilindros son superpuestos para definir la cavidad, como muestra la Figura 5.1. Como no existe interfaz apropiada para este método, el usuario debe ingresar las coordenadas de estas figuras geométricas y su tamaño, luego observar el resultado y ajustar los parámetros. Este proceso debe iterarse hasta lograr una definición de la cavidad aceptable. El usuario debe, además, conformarse con una definición de la cavidad que se base en 1 o 2 esferas, cubos o cilindros. Por otro lado, MDPocket (Schmidtke et al. 2011), proveniente de Fpocket(Le Guilloux et al. 2009) utiliza la estrategia híbrida de proponer cavidades posibles circunscriptas a una zona de búsqueda definida por el usuario. Para esto, el usuario define una zona de interés ajustando parámetros de la búsqueda automática y sólo dentro de esta zona se buscarán cavidades. De esta forma se tiene un control parcial de la definición de la cavidad de interés. Otras herramientas hacen todo esto automáticamente y ofrecen al usuario un conjunto de parámetros para ajustar su decisión. Con tiempo y experiencia, los usuarios aprenden el efecto que el valor de cada parámetro tiene sobre la definición de la cavidad, aunque el resultado final depende siempre del algoritmo.

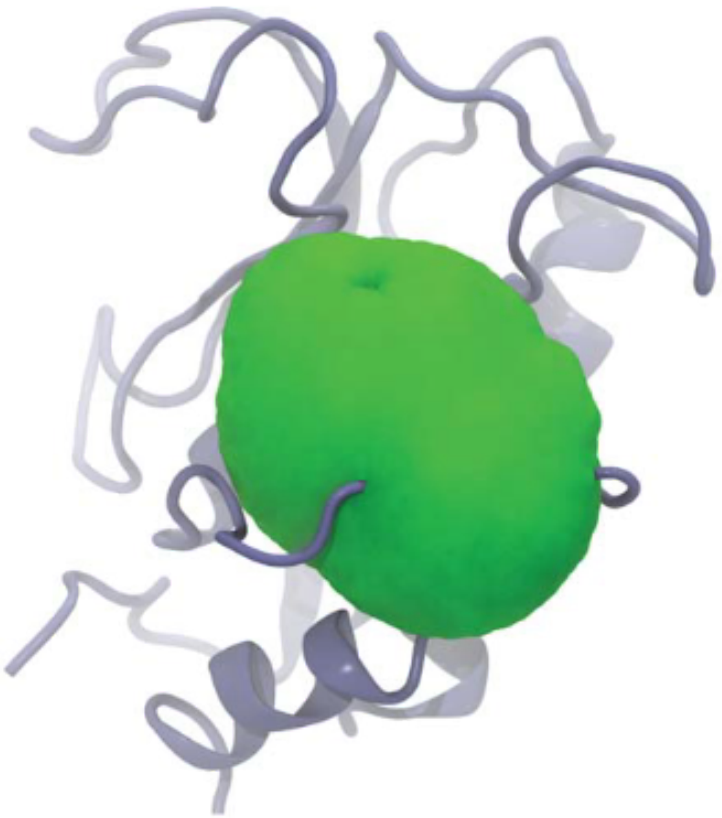

Figure 5.1: Definición de *Included Area* con POVME. 4 esferas se superponen para definir la cavidad.

Por último, cabe mencionar que todos los programas de cálculo de volúmenes de huecos utilizados en MD utilizan una definición estática de la cavidad. Esto es, no permiten su redefinición durante la simulación. Sin embargo, en muchas ocasiones, cambios conforma-

cionales experimentados por la proteína durante la trayectoria de MD hacen necesaria la redefinición de la cavidad.

5.2.2 Estrategia utilizada por ANA: *Convex Hull* como *Included Area*

ANA hace uso del algoritmo de *Convex Hull* (CH, caparazón convexo) para definir la cavidad. El usuario ingresa el listado de residuos que conforman la pared de la cavidad y las coordenadas de los carbonos alfa son utilizadas como los puntos de entrada para construir un CH y dentro de este caparazón convexo se buscarán las cavidades. Como alternativa, el usuario también puede ingresar un listado de átomos que conforman la pared de la cavidad. La Figura 5.2 muestra el resultado de aplicar este procedimiento en una proteína ejemplo.

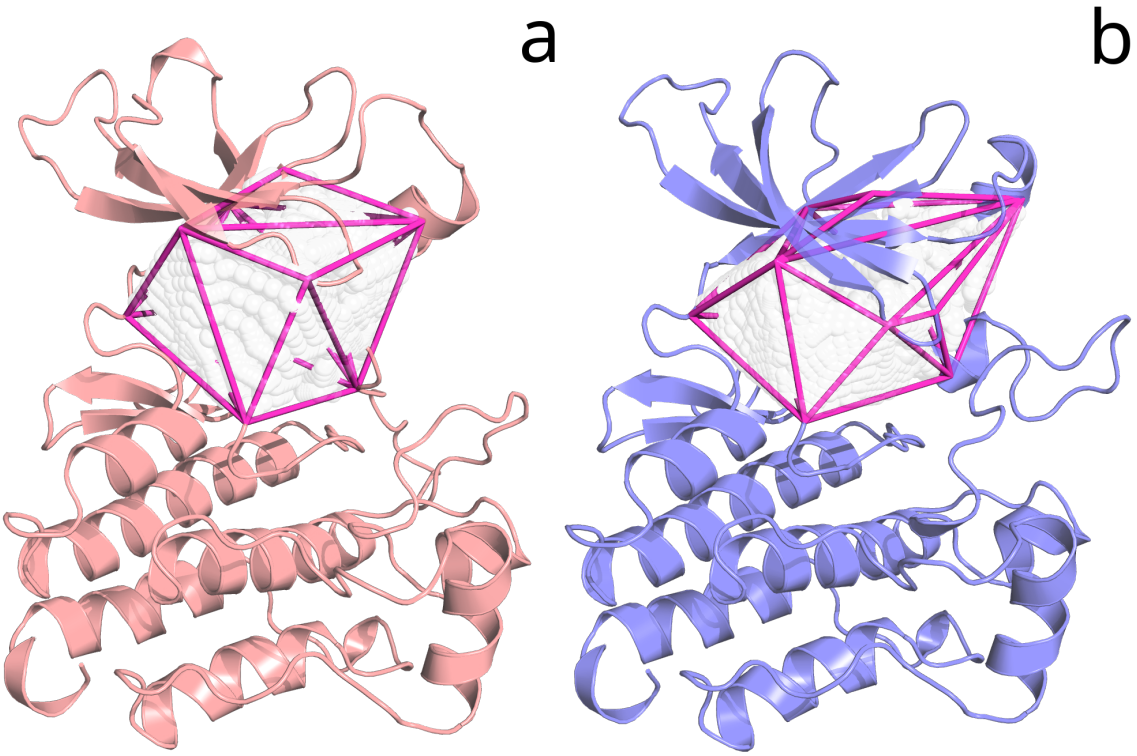

Figure 5.2: Comparación entre 2 confórmeros de la subunidad quinasa del EGFR. La cavidad se muestra como esferas blancas y la *Included Area* en fucsia. El cambio conformacional implicó un achatamiento y estiramiento de la cavidad, cambio que fue acompañado por la *Included Area*.

5.3 Cálculo del volumen vacío: *Delaunay Triangulation*

ANA utiliza el algoritmo de triangulación de Delaunay (DT) para determinar los huecos dentro de la proteína. Una vez triangulada la molécula(Boissonnat et al. 2002), ANA

obtiene una serie de tetrahedros que representan el espacio vacío dentro de la molécula. La triangulación trabaja sobre el conjunto de puntos que representa a los átomos, sin considerarlos como esferas, por lo que muchos de estos tetrahedros son de muy pequeño tamaño y no representan un verdadero espacio vacío. ANA descarta aquellos que tengan un volumen menor al de una molécula de agua (aunque éste es un parámetro modificable) y luego descarta aquellos que no estén dentro del CH previamente calculado (Barber et al. 1996) (Overmars 1996) (Fabri et al. 1998). Sólo quedan los tetrahedros que atraviesan el área definida por el CH.

Además, ANA presenta opciones de uso de baja precisión (LP, por sus siglas en inglés) y alta precisión (HP, por sus siglas en inglés). En modo LP ANA conserva todos los tetrahedros que tienen al menos una sección de su volumen dentro del área definida por el CH. En el modo HP, ANA además computa las intersecciones entre el CH y los tetrahedros, descartando los volúmenes de los tetrahedros que estén por fuera del área definida por el CH. La diferencia entre estos 2 métodos se puede ver en la Figura 5.3. Este paso extra conlleva un costo computacional que se verá en la Figura 5.5.

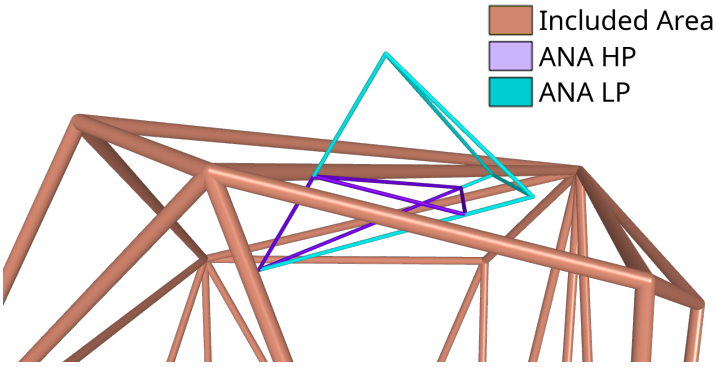

Figure 5.3: *Included Area*, definida por el CH (salmón). Comparación entre los modos de baja precision (**LP**, en celeste) y alta precisión (**HP**, en violeta) de ANA. El modo HP implica un leve aumento en el costo computacional.

La Figura 5.4 resume el método de ANA.

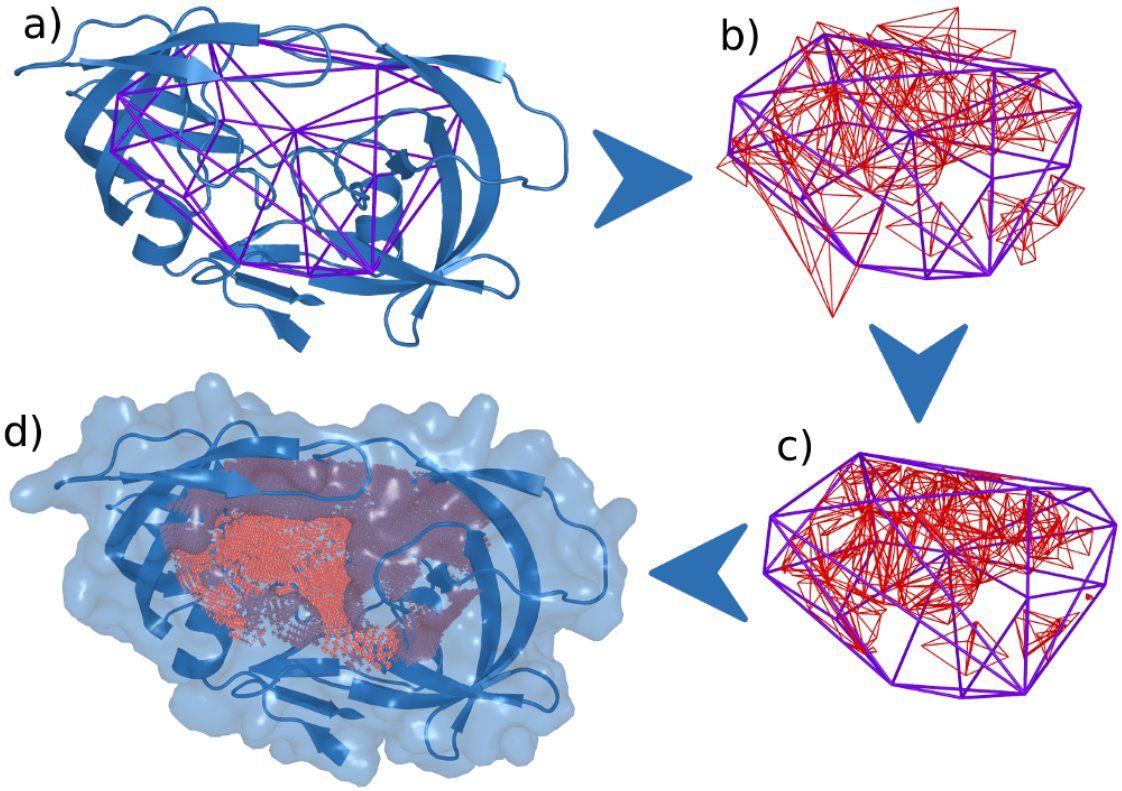

Figure 5.4: Ejemplo de aplicación de ANA en la cavidad principal de la proteasa del virus de HIV (HIVP). **a** muestra a la HIVP con su *Included Area* en violeta. **b** descarta a la HIVP y muestra los tetrahedros de la triangulación que no fueron descartados por su reducido tamaño y que se encuentran, al menos en parte, dentro de la *Included Area*. **c** muestra el paso extra del modo de alto precisión, calculando la intersección entre los tetrahedros y la *Included Area*. **d** muestra el resultado final.

El uso de GPUs (*Graphical Processing Units*) en la simulación de MD ha reducido sustancialmente los tiempos de cálculo y, por lo tanto, las trayectorias han alcanzado las escalas de tiempo real de microsegundos y hasta milisegundos. Por este motivo, la eficiencia de ANA en el cálculo de volúmenes de cavidades fuev comparada con herramientas similares como Epock, POVME y MDPocket usando como *benchmark* 600 configuraciones de una trayectoria de una porina de Rhodopseudomonas Blastica (PDB ID: 1PRN), que consiste en un cilíndro β con una cavidad relativamente grande (5000 ^{3}) (Figuras 5.5). ANA resultó el programa más rápido de los comparados, en sus 2 modalidades, y superando el límite de las 30 configuraciones por segundo en su modo de baja precisión.

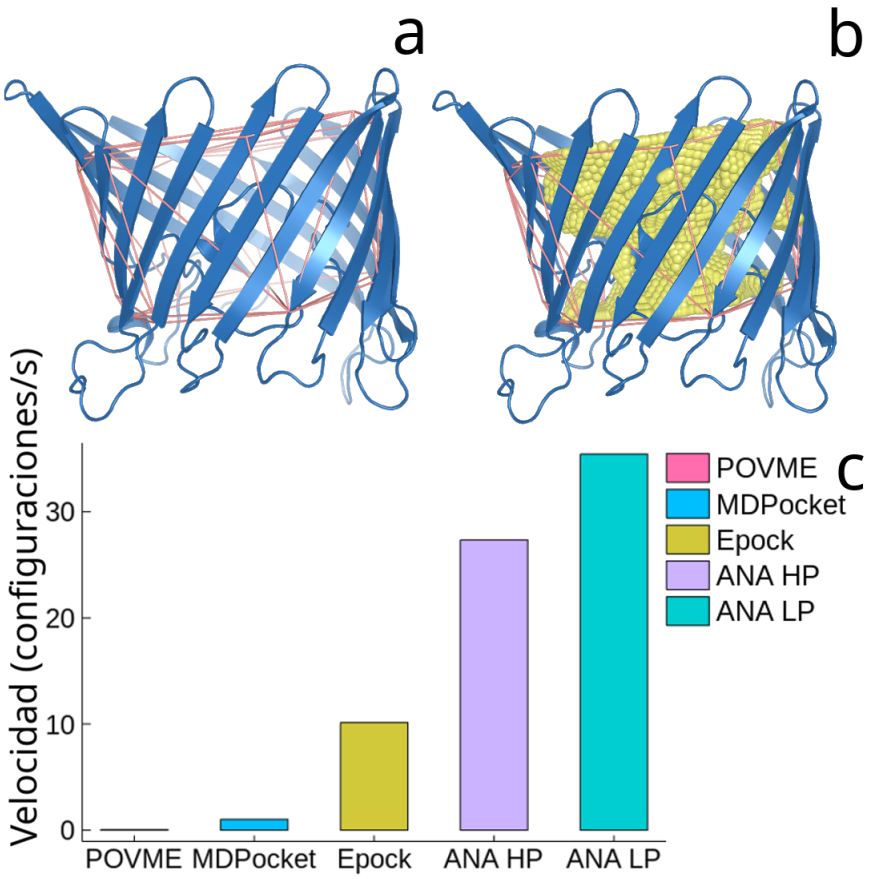

Figure 5.5: *Benchmark* de la porina de Rhodopseudomonas blastica. **a** muestra el túnel de la porina delineado (en color salmón) por la *Included Area* de ANA y **b** también muestra su cavidad representada por esferas amarillas. **c** Comparación entre las velocidades de cálculo de cada programa. POVME se ejecutó en 4 *threads* en paralelo. *Hardware*: DELL XPS 15, RAM: 32GB DDR4 2400 MHz, CPU: Intel i7-7700HQ 2.80GHz, SSD: PM961 NVMe SAMSUNG 1024GB

También se compararon los volúmenes obtenidos entre los distintos métodos, para comprobar la precisión de ANA. La Figura 5.6 muestra una comparación entre los volúmenes calculados entre los distintos métodos. Como puede verse, la penalidad en el costo computacional del modo de alta precisión está bien justificada por la mayor precisión en el cálculo de volumen y su reducida fluctuación en el valor del volumen calculado entre distintas configuraciones.

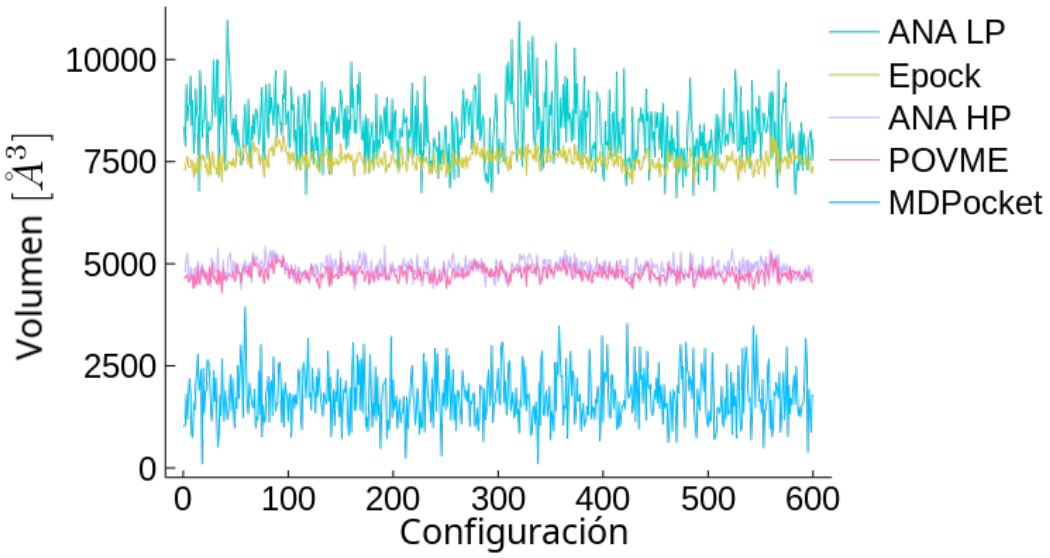

Figure 5.6: Comparación de los volúmenes obtenidos de los 600 configuraciones de la trayectoria de la porina de Rhodopseudomonas blastica para los 4 programas distintos. Si bien los valores absolutos varían, las correlaciones entre los programas son las apropiadas, salvo en el caso de MDPocket. MDPocket reportó los volumenes más alejados y los más variables, con saltos mayores a 1000 ^{3} entre *frames* consecutivos. Mientras tanto, los volúmenes obtenidos por ANA obtuvieron una correlación de 0.6 respecto a POVME y 0.56 respecto a ANA.

5.4 Dinámica y flexibilidad de cavidades: *Non-Delaunay Dynamics*

Como se dijo en la sección anterior, las cavidades proteicas no son objetos estáticos. Se reconoce ampliamente que la flexibilidad y la dinámica de las cavidades son cruciales para la función (Stank et al. 2017). Los cambios conformacionales con frecuencia implican cambios en los volúmenes de las cavidades con impactos posteriores sobre la afinidad y especificidad del ligando y, por lo tanto, la regulación de la función de una proteína (Kokh et al. 2013) (Pravda et al. 2014). Si bien la descripción estructural detallada de las cavidades es útil para predecir el tipo de ligandos que interactúan con la proteína, sus características dinámicas contribuyen a analizar la potencial multiplicidad de ligando (Pabon & Camacho 2017) (Barril & Fradera 2006). Además, la forma, el tamaño, y la flexibilidad de las cavidades se ha asociado recientemente con el grado de diversidad conformacional y regiones de desorden, entre otras características estructurales y dinámicas de las proteínas (Monzon et al. 2017). La flexibilidad de las cavidades es proporcionada por la dinámica vibracional de las proteínas que van desde movimientos lentos y colectivos hasta movimientos rápidos y localizados. Las simulaciones de MD combinadas con el análisis de componentes principales (PCA) proporcionan un marco para descomponer la complejidad de los movimientos de proteínas en contribuciones individuales desacopladas. Asimismo, el análisis de modos normales también puede cumplir una función similar. MD y PCA, se han aplicado recientemente para desarrollar un procedimiento que revela la

existencia de correlaciones entre la dinámica de las cavidades y las estructuras (Feig et al. 2015).

ANA puede ser utilizada para analizar la dinámica de las cavidades proteicas. La caracterizamos en términos de su vector de gradiente de volumen (∇V). Para este propósito, utilizamos algoritmos para el cálculo de volúmenes de cavidades que resultan robustos para diferenciaciones numéricas. ∇V se expresa en términos de modos PCA o NMA. Como resultado, las características dinámicas de las cavidades se pueden conectar directamente a las fluctuaciones térmicas en las proteínas.

5.4.1 Gradiente del volumen: *Volume Gradient Vector*

El gradiente del volumen (∇V) es un vector que permite caracterizar la flexibilidad de una cavidad. Los elementos de ∇V representan los cambios de volumen debidos al desplazamiento a lo largo de cada modo de PCA o NMA. Se obtiene por derivadas parciales del volumen de la cavidad en base a los modos de PCA $\boldsymbol{Q}_i$ $(i = 1, 3N)$, siendo N el número de residuos de proteínas.

$$\nabla \boldsymbol{V} = \sum_{i=1}^{3N} c_i \boldsymbol{Q}_i = \sum_{i=1}^{3N} \frac{\delta V}{\delta \boldsymbol{Q}_i} \boldsymbol{Q}_i \tag{5.1}$$

Donde:

- $\nabla \boldsymbol{V}$: vector gradiente del volumen.
- $\frac{\delta V}{\delta \boldsymbol{Q}_i}$: derivada parcial del volumen a lo largo del modo de PCA $\boldsymbol{Q_i}$.

Se calcula la derivada parcial del volumen a lo largo del modo $\boldsymbol{Q}_i$ por diferencia finita:

$$\frac{\delta \boldsymbol{V}}{\delta \boldsymbol{Q}_i} = \frac{V_i - V_{eq}}{\Delta \boldsymbol{Q}_i} \tag{5.2}$$

Donde:

- V_{eq}: volumen de la estructura de referencia.
- V_i: volumen de la estructura desplazada en la dirección del modo de PCA $\boldsymbol{Q_i}$.
- $\Delta \boldsymbol{Q}_i = \boldsymbol{Q}_i - \boldsymbol{Q}_{i,eq}$: desplazamiento a lo largo del modo $\boldsymbol{Q_i}$.

La magnitud del desplazamiento de cada modo estará dada por la amplitud térmica que le corresponde. Para encontrarla, se parte de la energía potencial promedio en la coordenada Q_i:

$$\langle V_i \rangle = \frac{1}{2}\lambda_i^{NMA}\langle Q_i \rangle = \frac{1}{4}\lambda_i^{NMA}A_i^2 \tag{5.3}$$

Donde:

- $\langle V_i \rangle$: energía potencial media debida al modo i-ésimo.
- λ_i^{NMA}: autovalor del modo normal $\boldsymbol{Q}_i$.
- $\langle Q_i \rangle$: desplazamiento medio a lo largo del modo $\boldsymbol{Q}_i$.
- A_i: amplitud del modo i-ésimo.

Y utilizando el teorema de equipartición $\langle V_i \rangle = \frac{1}{2}k_BT$, se obtiene la amplitud térmica:

$$A_i = \sqrt{\frac{2k_BT}{\lambda_i^{NMA}}} \tag{5.4}$$

Ahora bien, como no disponemos de modos normales sino de modos de PCA, trabajando con el análisis cuasiharmónico:

$$\lambda_i^{NMA} = \frac{k_BT}{\lambda_i} \tag{5.5}$$

Donde:

- λ_i: autovalor del modo de PCA $\boldsymbol{Q}_i$.

Así se llega a la ecuación final para obtener la amplitud correspondiente al modo $i$ésimo de PCA:

$$A_i = \sqrt{(2\lambda_i)} \tag{5.6}$$

Si se utilizan los autovalores obtenidos mediante el paquete de dinámica molécular de Amber (Case et al. 2005), debido a las unidades que este utiliza, la ecuación tomará esta última forma:

$$A_i = \frac{\sqrt{2}}{\lambda_i^{amber}} \tag{5.7}$$

Para evitar conflictos estéricos entre los átomos debido a los desplazamientos realizados durante el cálculo de las derivadas parciales por diferencias finitas, las distorsiones estructurales se restringieron por debajo de una diferencia cuadrática media (RMSD) de 0.05 Å.

5.4.2 Continuidad en el cálculo del volumen y *Non-Delaunay Dynamics* (NDD)

En la Figura 5.6 puede apreciarse el problema del cálculo dinámico de cavidades, los volúmenes calculados varían fuertemente entre configuración y configuración. Este problema es particularmente grave en los programas que utilizan triangulaciones, como MD-Pocket y ANA. Incluso los programas de grilla como POVME y Epock, limitados en su precisión por la resolución de la grilla y, consecuentemente, el costo computacional, tienen problemas para el cálculo robusto de diferencias finitas. Ninguna de las herramientas existentes es capaz de calcular la perturbación en el volumen de manera estable. ANA resuelve este problema por una vía alternativa: *Non-Delaunay Dynamics* (NDD).

En vez de retriangular todo el sistema cada vez que se lo desplaza a lo largo de un modo $\boldsymbol{Q}_i$, ANA —luego de realizar la primera triangulación de Delaunay y descartar los tetrahedros innecesarios (Figura 5.4b)—, desplaza los atomos de la proteína en dirección de cada modo $\boldsymbol{Q}_i$. Esto produce una deformación en los tetrahedros y, consecuentemente, un cambio contínuo en sus volúmenes, sin introducir artefactos en el volumen final. La Figura 5.7 muestra el beneficio de evitar una retriangulación del sistema.

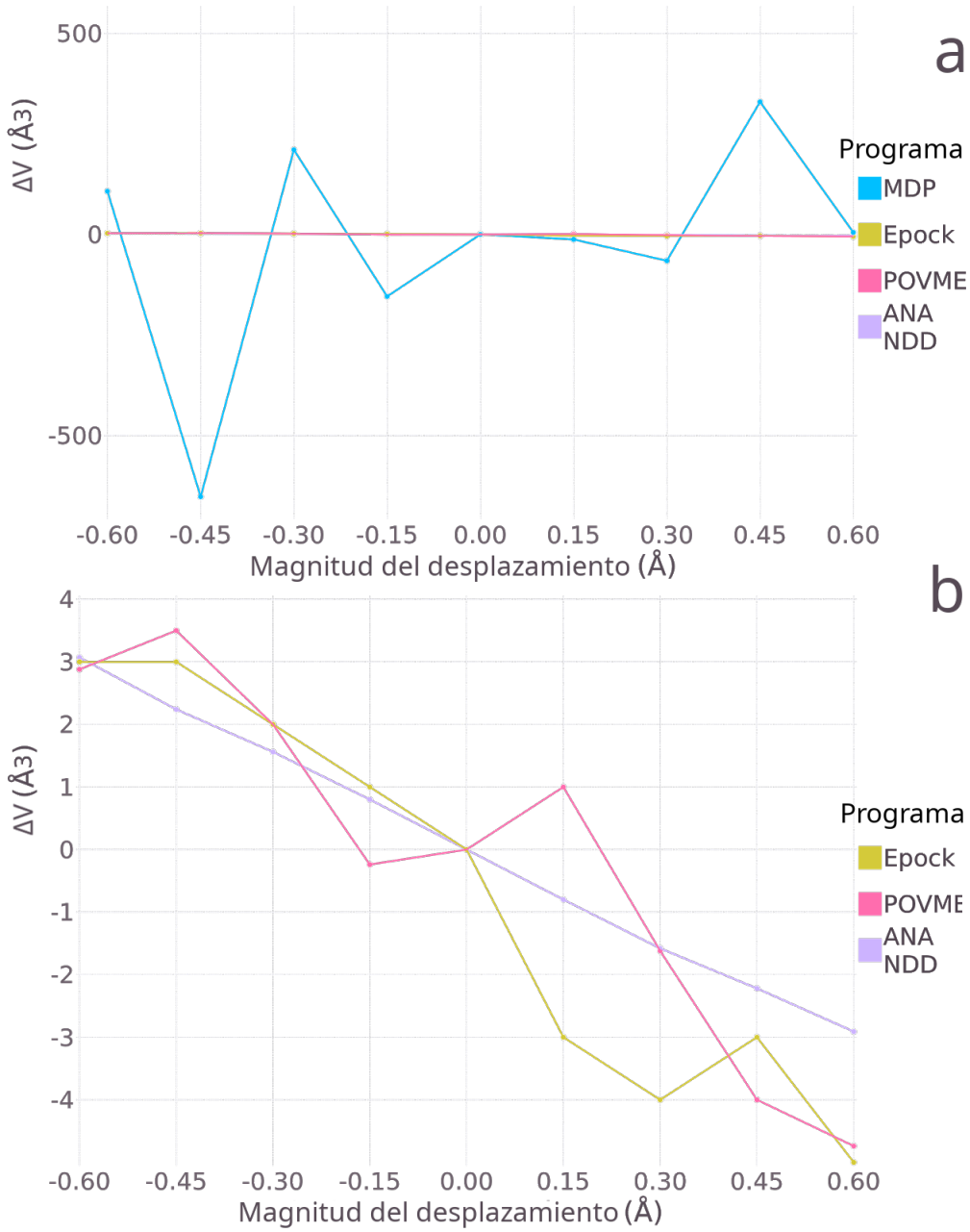

Figure 5.7: Comparación de la continuidad en el cálculo del volumen utilizando distintas herramientas. Se muestra la variación de volumen en el túnel principal del cilíndro β de la porina de Rhodopseudomonas blastica al variar su estructura a lo largo del principal modo de PCA. **a** comparación con MDPocket, Epock and POVME. **b** comparación con Epock and POVME. ANA NDD es el único programa que logra una continuidad aceptable bajo mínimos desplazamientos, y por consiguiente, valores correctos de derivadas parciales.

El nombre del método proviene de la violación a la propiedad de Delaunay que este método, en un principio, estaría haciendo. El más mínimo desplazamiento de un vértice de una triangulación de Delaunay implica el recálculo de toda la triangulación para mantener la propiedad de Delaunay (que ningún vertice se encuentre dentro de la esfera circunscripta por los vértices de los tetrahedros, ver capítulo anterior). En la práctica, esto se debe al desplazamiento de los vértices de los tetrahedros más pequeños, que carecen de todo interés práctico. Como ANA descarta estos tetrahedros, puede realizar leves desplazamientos en el resto de los tetrahedros minimizando la incertidumbre en el cálculo del volumen.

5.4.3 Casos de estudio

5.4.3.1 Dinámicas Moleculares y PCA

Para demostrar el funcionamiento del método, se lo aplicó en 4 sistemas de estudio que fueron simulados por MD. Las estructuras de partida, 1HVR (proteasa de HIV), 1M14 (activo conformero del EGFR), y 256L (lisozima T4), fueron extraídas de la base de datos de PDB. Los sistemas fueron construidos, minimizados, llevados a temperatura biológica y equilibrados, siguiendo los protocolos descriptos en el capítulo 2. Luego se produjeron trayectorias de 600ns de cada uno y se extrajeron configuraciones equilibradas a intervalos de 10ps. El análisis de PCA fue realizado siguiendo los métodos del capítulo 3.

1. HIVP-APO: proteinasa de tipo I del HIV. Es un homodímero con una cavidad afectada principalmente por un movimiento de apertura y cierre de baja frecuencia. El sitio activo está formado simétricamente por los 2 monómeros y su acceso está controlado por 2 giros de hebras β (como puede apreciarse en la figura 5.8), que cumplen la función de *gating*. Esta proteína es un típico ejemplo de movimiento de bisagra que abre o cierra el acceso a su cavidad involucrando el desplazamiento colectivo de muchos de sus residuos.

2. HIVP-HOLO: estructura de la proteinasa unida a un inhibidor diseñado basado en urea ciclada(Patrick et al. 1991). Se intentará encontrar diferencias en su dinámica respecto a la estructura apo.

3. EGFR: El dominio quinasa del Receptor del Factor Epidermal de Crecimiento (EGFR, por sus siglas en inglés) presenta un sitio activo limitado por un lóbulo en el N terminal y otro en el C terminal, conectados por una región bisagra (Taylor & Kornev 2011). Se han identificado y cristalizado confórmeros inactivos y activos de este dominio y se utilizan características estructurales equívocas para diferenciarlos, por lo que también se han buscado características dinámicas que los diferencien(Wan & Coveney 2011). Se entiende que un movimiento de apertura y cierre es responsable por la transición del confórmero activo inestable hacia el confórmero inactivo más estable(Shan et al. 2013).

4. T4L: la lisozima T4 ha sido largamente estudiada por métodos computacionales y experimentales (Matthews & Remington 1974) (Bruccoleri et al. 1986). Estudios previos han conectado movimientos de bisagra, giro y torsiones relativas entre los lóbulos que rodean la hendidura donde se une su ligando (Hub & Groot 2009). Debido a esto, puede ser un buen ejemplo de contribuciones de varios modos de baja frecuencia afectando su cavidad.

5.4.4 Resultados

5.4.4.1 Frecuencias de los modos relevantes

5.4.4.1.1 Número de participación: Para conservar solamente los modos relevantes en la dinámica de la cavidad, se seleccionan los P_q modos más relevantes según la fórmula (Bell & Dean 1970):

$$P_q = \frac{1}{\sum_{i=1}^{3N} c_i^4} \tag{5.8}$$

Donde:

- P_q: número de participación.
- c_i: componente del ∇V.
- N: número de partículas.

P_q representa la delocalización de ∇V sobre la base de los modos de PCA o NMA. Su valor (redondeado) indica el mínimo número de modos necesarios para representar ∇V. Un valor de P_q cercano a $3N$ significaría que la flexibilidad de la cavidad depende de todos los modos de PCA, sin que ninguno adquiera mayor relevancia. Mientras que un valor cercano a 1 indicaría que es un único modo el responsable de su dinámica. Sólo los primeros P_q modos ordenados por su c_i^2 fueron utilizados en el análisis, lo que permite descartar contribuciones menores y hacer foco en fluctuaciones fundamentales. Valores de $\frac{P_q}{3N}$ de 0.25, 0.16, 0.20, y 0.26 fueron obtenidos para HIVP-APO, HIVP-HOLO, EGFR, y T4L, respectivamente. Es decir, sólo una fracción reducida de modos de PCA contribuyen significativamente al desplazamiento de las cavidades respecto a la dirección de sus gradientes de volumen (esto es, la dinámica de mayor cambio a sus volúmenes).

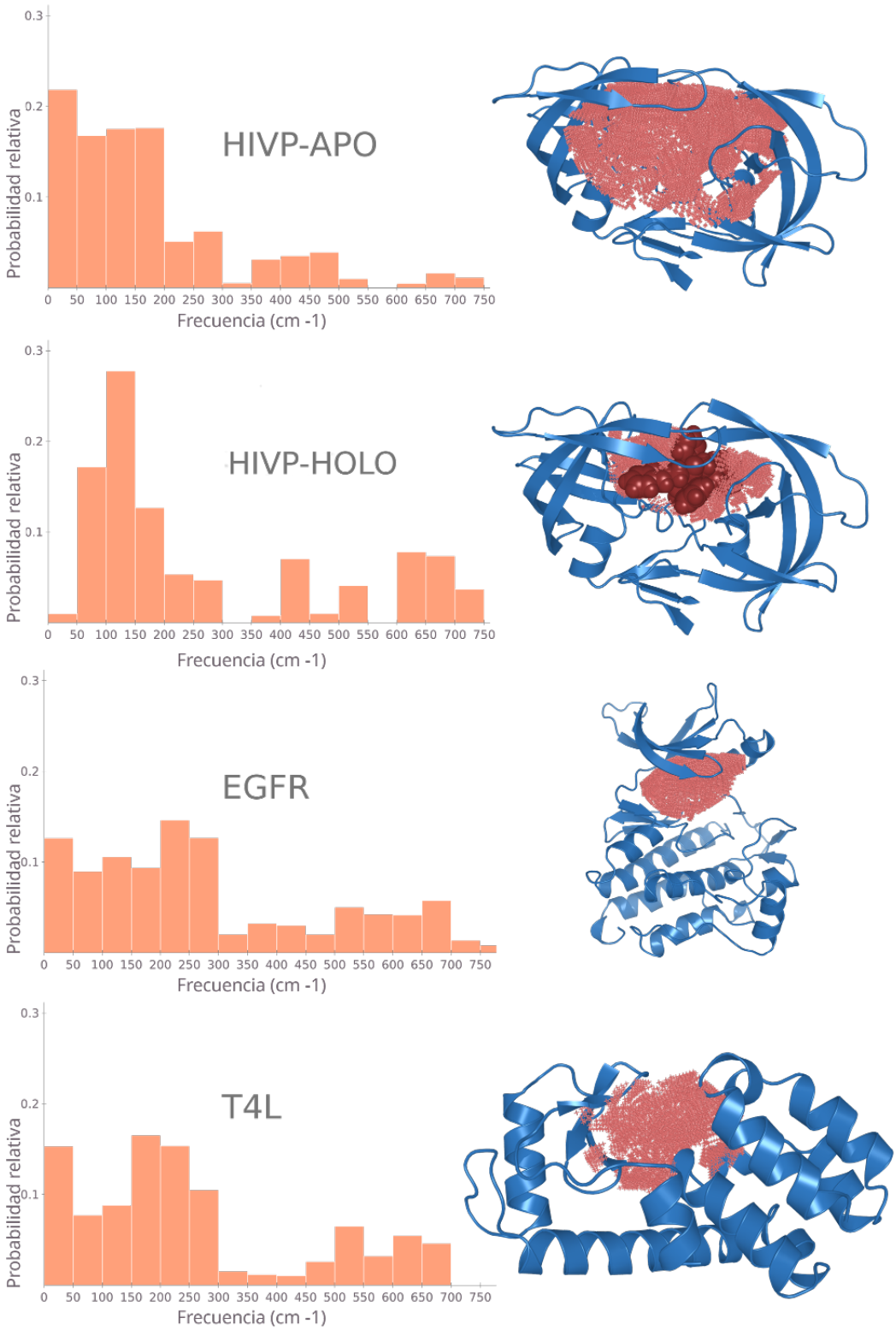

Figure 5.8: **izquierda** Histograma de la contribución a ∇V por parte de los modos de PCA, segun la frecuencia. **derecha** Los sistemas de estudio en azul y sus cavidades en salmón.

La Figura 5.8 muestra la distribución de la contribución de los modos de PCA al ∇V. Para todos los sistemas, modos de baja frecuencia (menores a $300cm^{-1}$) representan la mayor contribución. Hay tambien una notable, y esperable, diferencia en el caso de HIVP, debida a la presencia de ligando. Este parece rigidizar la cavidad, eliminando casi por completo la contribución de los modos de más baja frecuencia a la cavidad y favoreciendo

fuertemente a los intermedios.

5.4.4.2 Deslocalización de los modos que afectan a la cavidad

El mismo concepto del número de participación puede utilizarse para determinar el grado de localización o deslocalización espacial de los modos:

$$P_i = \frac{1}{\sum_{i=1}^{N} Q_{ij}^2} \tag{5.9}$$

Donde:

- P_i: número de participación del modo i.
- $Q_{ij} = (Q_{ij}^x)^2 + (Q_{ij}^y)^2 + (Q_{ij}^z)^2$: fluctuación del átomo i bajo el modo j.
- N: número de partículas.

Los valores de $P_i \approx N$ describen fluctuaciones distribuidas equitativamente en todos los residuos de la proteína, y los $P_i \approx 1$ corresponden a las fluctuaciones de un solo residuo. En la Figura 5.9a se muestra la distribución de la fracción de residuos involucrados en el movimiento de los modos PCA ponderados por su contribución al gradiente de volumen ∇V. En todos nuestros casos de estudio, las fluctuaciones que contribuyen a maximizar el volumen de la cavidad pueden involucrar hasta aproximadamente la mitad de los residuos totales de la proteína. Esto se espera debido a las principales contribuciones de los modos colectivos de baja frecuencia a los cambios en la cavidad mostrados anteriormente en la Figura 5.8. Para analizar más a fondo esta característica, reducimos nuestro análisis a los residuos que recubren las cavidades. Para este propósito, los modos de PCA se recortaron y renormalizaron reteniendo solo elementos que involucran estos residuos. La figura 5.9b muestra la distribución correspondiente de la fracción de residuos en la superficie de las cavidades que participan juntos en movimientos concertados debidos a modos colectivos relacionados con los cambios de cavidad. En todos los casos, hasta el 70% de los residuos de las cavidades participan en fluctuaciones colectivas con un impacto directo en el tamaño de las cavidades. Es decir, la mayoría de los residuos que recubren las cavidades se pueden desplazar de forma sincronizada para maximizar los efectos en los tamaños y formas de las cavidades. Estos modos estructurales de baja frecuencia, aunque implican movimientos de residuos localizados en toda la proteína (es decir, hasta aproximadamente la mitad de los residuos totales de la proteína), están particularmente más localizados en los residuos que recubren las cavidades. Por lo tanto, se espera que tengan un impacto potencial en la función de la proteína.

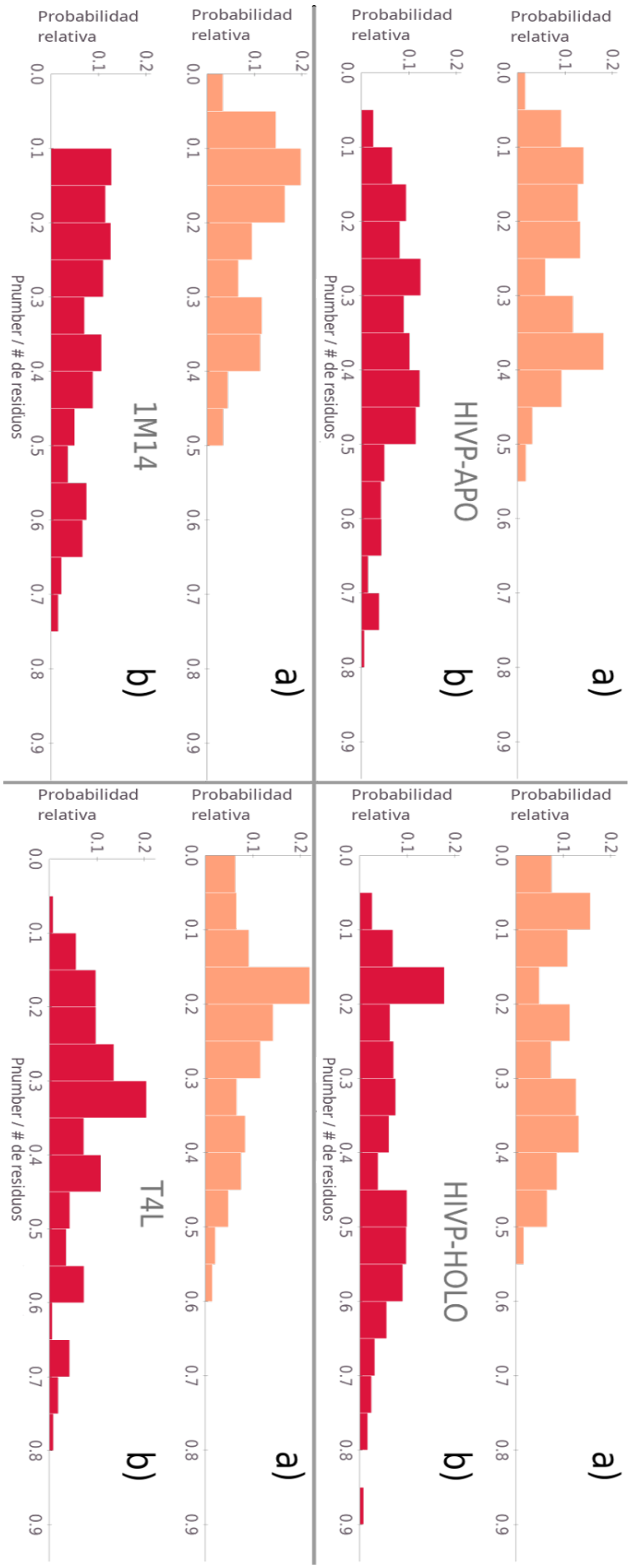

Figure 5.9: Histograma de la contribución al ∇V según número de participación (P_i). El P_i es relativo al número total de residuos de la proteína para comparar los 4 casos. **a** Histogramas de los modos completos. **b** histogramas de los modos recortados, conservando sólo los desplazamientos en los residuos de las paredes de las cavidades.

5.4.4.3 Residuos de la pared

El ∇V es un vector expresado en la base de los modos normales. Disponiendo de esta base, puede obtenerse el ∇V en coordenadas cartesianas. Este vector tendrá una dimensión de $3N$, donde N es el número de partículas y se referirá a las componentes x, y y z centradas en los $C\alpha$ de los residuos. Para obtener el ∇V en coordenadas cartesianas se reemplaza en eq. 5.1 a las coordenadas normales por las coordenadas de los residuos:

$$\nabla \boldsymbol{V} = \sum_{i=1}^{3N} \frac{\delta \boldsymbol{V}}{\delta \boldsymbol{Q}_i} \boldsymbol{Q}_i = \sum_{i=1}^{3N} \frac{\delta \boldsymbol{V}}{\delta \boldsymbol{Q}_i} \left(\sum_{j=1}^{3N} Q_{ij} q_j \right) \tag{5.10}$$

que resulta en:

$$\nabla \boldsymbol{V}_{xyz} = \sum_{j=1}^{3N} \frac{\delta \boldsymbol{V}}{\delta q_j} \boldsymbol{q}_j \tag{5.11}$$

ya que:

$$Q_{ji} = \frac{\delta \boldsymbol{Q}_i}{\delta q_j} \tag{5.12}$$

y:

$$\frac{\delta \boldsymbol{V}}{\delta \boldsymbol{q_j}} = \sum_{i=1}^{3N} \frac{\delta \boldsymbol{V}}{\delta \boldsymbol{Q_i}} \frac{\boldsymbol{Q}_i}{\delta q_j} \tag{5.13}$$

Operativamente, una vez obtenido el ∇V, esto implica convertirlo en coordenadas cartesianas, utilizando la matriz de modos:

$$\boldsymbol{\nabla V}_{xyz} = \boldsymbol{Q} \ \ \boldsymbol{\nabla V} \tag{5.14}$$

Donde:

- $\boldsymbol{\nabla V}_{xyz}$: VGV en coordenadas cartesianas.
- $\boldsymbol{Q}$: matriz de modos.

La visualización de este vector $\boldsymbol{\nabla V}_{xyz}$, para cada sistema, se encuentra en la Figura 5.10. Naturalmente, los residuos que afectan a la cavidad son los residuos de sus paredes, el resto de los residuos presentan apenas mínimas fluctuaciones que están dentro del error del método. Esto se debe a que si bien los movimientos colectivos que afectan a las cavidades no sólo mueven a residuos de la pared, sino que también a otros, a la hora de calcular

el efecto en la cavidad, las fluctuaciones de los residuos que no pertenecen a la pared se cancelan y sólo se conservan las fluctuaciones de los residuos que rodean a la cavidad.

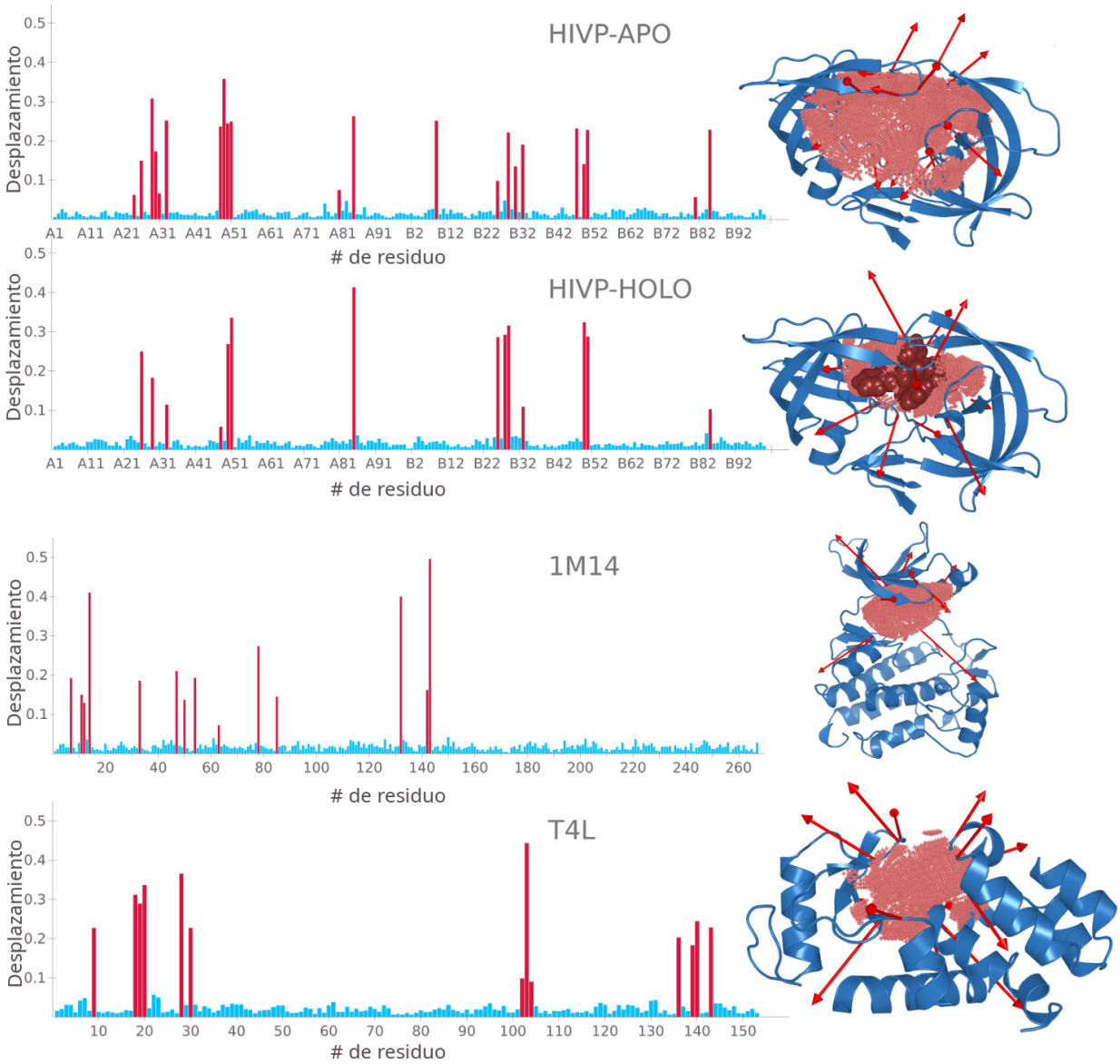

Figure 5.10: **izquierda** Desplazamiento relativo de los residuos bajo el vector del gradiente en coordenadas cartesianas ($\boldsymbol{\nabla V}_{xyz}$). Los desplazamientos en rojo son los de magnitud considerable y se considera a esos residuos como los residuos de la pared de la proteína. En celeste, el resto de los residuos, con valores despreciables. **derecha** Los sistemas analizados, con sus cavidades en salmón y los *porcupine plot* (flechas en rojo) de los $\boldsymbol{\nabla V}_{xyz}$, indicando los residuos más relevantes y su dirección de desplazamiento para el máximo cambio en el volumen.

Los desplazamientos más significativos corresponden a residuos localizados en la superficie de las cavidades. Si bien las fluctuaciones colectivas de baja frecuencia, que implican la movilidad de hasta el 50% de los residuos de las proteínas, contribuyen en gran medida a los cambios en el volumen de la cavidad (ver Figuras 5.8 y 5.9); se observan cancelaciones efectivas entre las contribuciones de diferentes modos a la movilidad de los residuos fuera de la cavidad. Por lo tanto, los cambios en los tamaños de las cavidades en realidad no implican grandes cambios conformacionales visibles en toda la estructura de la proteína, sino distorsiones estructurales diferenciales y una reorganización localizada de los residuos

que recubren sus superficies.

En el caso de HIVP-APO, $\boldsymbol{\nabla V}_{xyz}$ implica grandes desplazamientos de residuos pertenecientes al dominio del *flap* y ubicados frente a la cavidad (43%), así como residuos pertenecientes al dominio central (43%). La unión del ligando reduce ligeramente las contribuciones de los residuos en el dominio del *flap* (38%). Mientras que $\boldsymbol{\nabla V}_{xyz}$ de HIVP-HOLO se distribuye equitativamente entre ambos monómeros, se observan asimetrías dinámicas para HIVP-APO. Esto se debe a la localización asimétrica de los modos de baja frecuencia que más contribuyen a $\boldsymbol{\nabla V}_{xyz}$.

El confórmero activo EGFR presenta un $\boldsymbol{\nabla V}_{xyz}$ principalmente localizado en las cadenas β y la hélice αC del lóbulo N que contienen residuos con cadenas laterales que apuntan hacia la cavidad (46%). También implica movimientos relativos de residuos cerca de la región de la bisagra entre los lóbulos N y C (45%). La conformación activa de EGFR es relativamente inestable (Shan et al. 2013). Estudios previos muestran que la transición conformacional activa-inactiva implica la apertura de los lóbulos N y C para permitir el despliegue local en la región de la bisagra, antes del cierre de los lóbulos para reestabilizarse en su conformación inactiva (Li et al. 2014). Observamos que estos movimientos no contribuyen significativamente a $\boldsymbol{\nabla V}_{xyz}$. EGFR presenta contribuciones significativas de los modos de frecuencia de rango medio a los cambios de cavidad de volumen, una característica que conduce a una relativa rigidez de su cavidad.

Los modos de baja frecuencia que involucran movimientos entre los lóbulos de la proteína T4L son de flexión de bisagra, torsión y corte y representan contribuciones significativas a $\boldsymbol{\nabla V}_{xyz}$ en la lisozima T4. Esto está de acuerdo con los estudios de PCA que han informado previamente (Hub & Groot 2009) el gran impacto de estos movimientos en la clave catalítica.

5.4.4.4 Flexibilidad de la cavidad

Como ya se dijo en el capítulo 4, en el marco del análisis cuasiharmónico se entiende que la proteína desplaza sus residuos armónicamente, a lo largo de cada modo de PCA. Así, se puede obtener la perturbación energética que cada modo genera en la cavidad:

$$\Delta E_{Q_i} = \frac{1}{2} k_i c_i^2 \Delta X^2 \tag{5.15}$$

Donde:

- ΔE_{Q_i}: Cambio de energía potencial de la cavidad debido al modo $\boldsymbol{Q_i}$.
- $k_i = \frac{k_B T}{\lambda_i}$.
- k_B: constante de Boltzmann.

- T: temperatura.
- λ_i: autovalor del modo de $\boldsymbol{Q_i}$.
- c_i: contribución del modo $\boldsymbol{Q_i}$ al $\boldsymbol{\nabla V}$.
- ΔX: desplazamiento a lo largo del modo $\boldsymbol{Q_i}$, respecto a la estructura de referencia utilizada para el cálculo de PCA.

Y así obtener el cambio en la energía potencial de la proteína por un desplazamiento a lo largo del gradiente $\boldsymbol{\nabla V}$

$$\Delta E_{\nabla V} = \sum_{i=1}^{3N} \Delta E_{Q_i} \tag{5.16}$$

Donde:

- ΔE_{Q_i}: Cambio de energía potencial de la cavidad.

Así, dado cierto desplazamiento ΔX, se puede obtener $\Delta E_{\nabla V}$, y usar este valor como una medida de la flexibilidad de la cavidad.

El orden de flexibilidad de las cavidades da como resultado HIVP-APO > T4L > EGFR > HIVP-HOLO. Como se espera, la cavidad del HIVP se vuelve más rígida después de la unión del ligando. Para dar una idea de los valores de las flexibilidades relativas, la Figura 5.11b muestra los valores acumulativos de las contribuciones de los modos PCA (c_i^2) a $\boldsymbol{\nabla V}$ en función de las frecuencias de los modos de PCA (ver eq. 5.1). Las proteínas cuyas fluctuaciones colectivas de baja frecuencia participan más en los cambios del volumen de la cavidad exhiben cavidades más flexibles. A pesar de la gran contribución de estos modos a los desplazamientos en la dirección de $\boldsymbol{\nabla V}$ (ver Figura 5.8), no representan el costo principal de energía en esta dirección. Esto se puede ver en la Figura 5.11c, donde se muestran los valores acumulativos de $\frac{\Delta E_{\boldsymbol{Q_i}}}{\Delta E_V}$ en función de las frecuencias del modo PCA. Los modos de baja frecuencia, es decir, modos con frecuencias de hasta $300cm^1$, solo contribuyen menos del 40% al $\Delta E_{\nabla V}$ total en el caso de HIVP-APO, mientras que sus contribuciones se atenúan en gran medida en el HIVP-HOLO por la presencia del ligando. Comparando nuestros cuatro casos de estudio, podemos concluir que esas cavidades relativamente más flexibles (Figura 5.11a) se caracterizan por contribuciones comparativamente mayores de modos de bajas frecuencias a $\boldsymbol{\nabla V}$ (Figura 5.11b) que representan una fracción mayor del costo total de energía (Figura 5.11c).

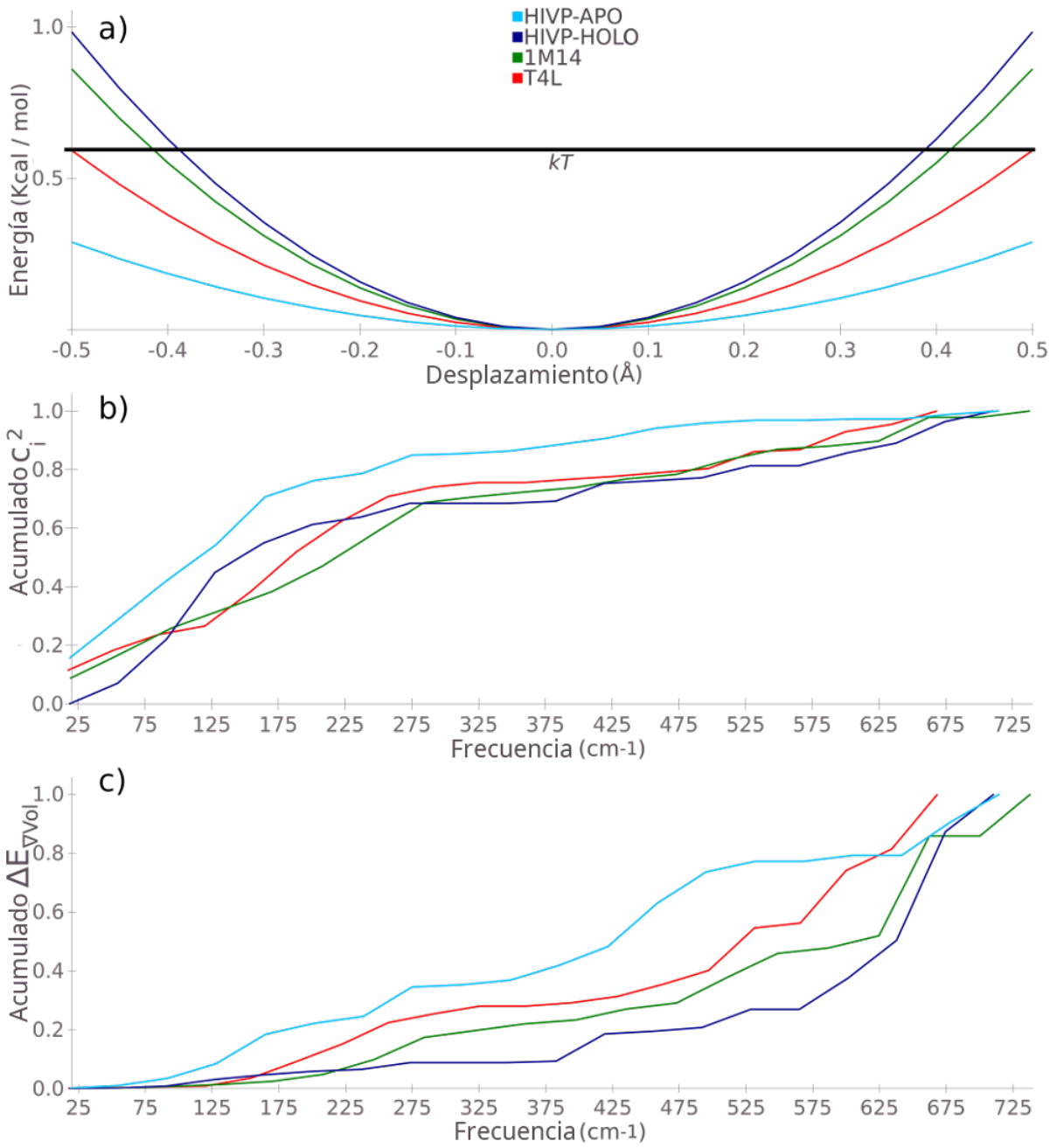

Figure 5.11: **a** Energía potencial a lo largo del VGV. **b** Acumulado de las contribuciones de los modos al VGV. **c** Acumulado de las contribuciones de los modos a $\Delta E_{\nabla V}$.

Todos estos análisis fueron posibles gracias al método de desplazamientos finitos de *Non-Delaunay Dynamics* que implementa ANA, y que permiten analizar la dinámica de las cavidades, que es uno de los parámetros a considerar para entender su función biológica.

Chapter 6

Receptor del Factor de Crecimiento Epidermal (EGFR)

6.1 Introducción

El Receptor del Factor de Crecimiento Epidermal (EGFR, por sus siglas en inglés) es una proteína clave en la señalización y regulación de la proliferación celular(Taylor & Kornev 2011). Está compuesto por un receptor extracelular (residuos 25 a 641, siguiendo la numeración de UniProtKB, P00533), una hélice transmembrana (residuos 642 a 668) y una región citoplasmática que se divide en un segmento próximo a la membrana (residuos 669 a 711), un dominio quinasa (residuos 712 a 979) y una cola desordenada en su C terminal de 231 residuos (980 a 1210) a la que se unen las moléculas de señalización cuando ésta se encuentra fosforilada. Mutaciones en el dominio quinasa y sobreexpresión de EGFR fueron asociadas a distíntos tipos de cáncer(Arteaga & Engelman 2014) por aumentar la actividad quinasa, razón por la cual, este dominio es investigado extensamente como un blanco de drogas (Wang et al. 2014). Sustituciones (SASs, *Single Aminoacid Substitution*), deleciones e inserciones en este dominio pueden alterar el preequilibrio de los confórmeros, aumentando la preponderancia del confórmero activo necesario para la autofosforilazión de la cola desordenada del C terminal, lo que desencadena toda la cascada de señalización (Gajiwala et al. 2013). El mecanismo de activación del EGFR empieza con la unión del EGF al receptor extracelular, que promueve la dimerización del EGFR, con la interacción asimétrica de los dominios quinasa. Una quinasa (activadora) acomodará su lóbulo C (ver Figura 6.1) contra el segmento próximo a la membrana y el lóbulo N de la otra quinasa (la activada). Así, la quinasa activada adopta el confórmero activo y fosforila las tirosinas de la cola desordenada en el C terminal(Endres et al. 2013). Las características que diferencian a los confórmeros activo e inactivo es uno de los temas de interés. Hasta ahora, estas diferencias se han limitado a características estructurales (ver Figura 6.1), que separan al confórmero activo del inactivo: **a)** la hélice α C orientada hacia el interior,

lo que permite la formación de un puente salino entre el glutamato 762 (E762) y la lisina 745 (K745) y **b)** una conformación extendida del llamado *activation loop* que orienta al aspartato 837 de la tríada HRD (por la histidina 835, arginina 836 y el aspartato 837) hacia el sitio catalítico donde se une el ATP(Jura et al. 2011). Pero, no todos los confórmeros se ajustan a estas características y existen numerosas conformaciones ambiguas cuya clasificación, segun estos criterios, no es posible(Endres et al. 2013).

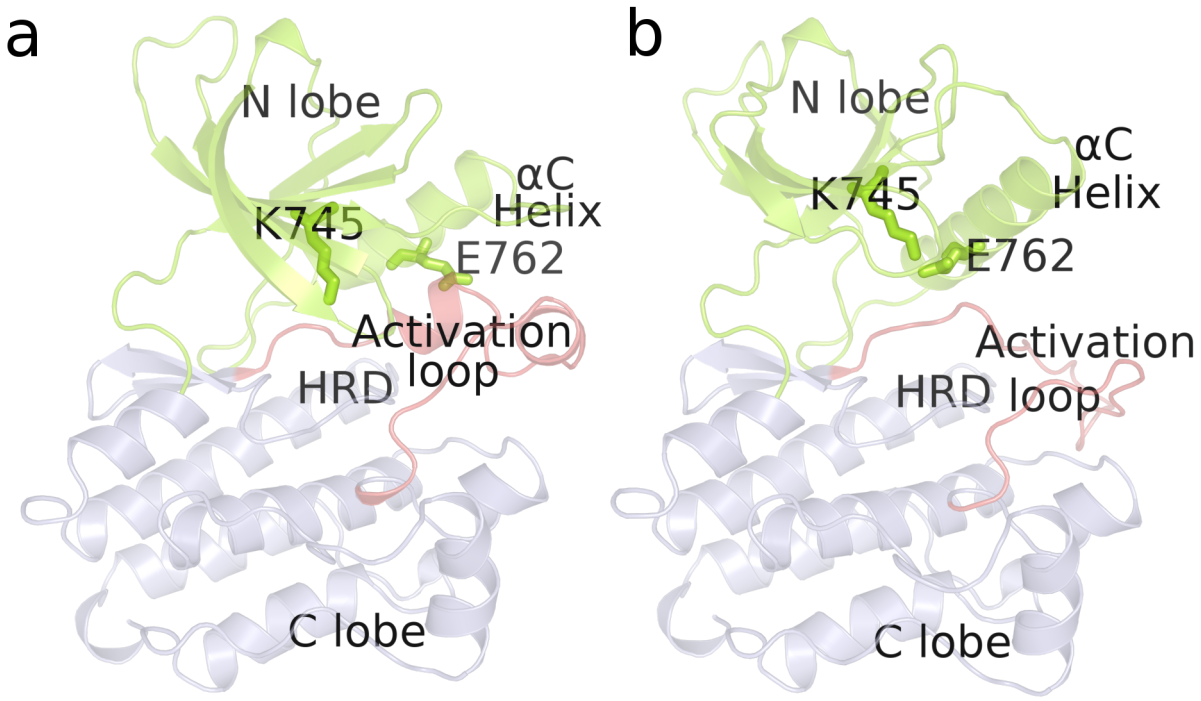

Figure 6.1: Conformerós inactivo (**a**) y activo (**b**). El lóbulo N está coloreado en verde y el C en violeta, mientras que el *activation loop* está en salmón. También se indica la hélice C, los residuos que forman el puente salino (K745 y E762) y la tríada HRD, histidina, arginina y aspartato

En nuestro grupo se han caracterizado los parámetros estructurales de un amplio set de dominios quinasa para su clasificación en confórmeros activos o inactivos(Hasenahuer et al. 2017). Se encontró que haciendo foco en los cambios de RMSD de la cavidad de la quinasa y realizando una clusterización jerárquica, se puede mejorar la clasificación. Lo que no abunda, es una caracterización del comportamiento dinámico de estos confórmeros. Shan et al., utilizando simulaciones de MD, exploraron las transiciones entre estos confórmeros para un único mutante de EGFR (Shan et al. 2013). Otros trabajos han utilizado PCA para caracterizar al confórmero activo como uno de mayor flexibilidad (Wan & Coveney 2011), mientras que otros se enfocaron en el efecto de las SASs en la estabilidad de los confórmeros(Shan et al. 2012). En el presente capítulo, utilizamos el análisis de modos normales (NMA) para caracterizar la dinámica que diferencia a los confórmeros activos e inactivos del EGFR, aprovechando que estos suelen ser robustos frente a las variaciones de secuencia (Maguid et al. 2008) (Zheng et al. 2006). Presentamos también un procedimiento para identificar, y comparar, las variables dinámicas colectivas más relevantes de cada confórmero de la quinasa EGFR. Esto permite definir características

dinámicas de los confórmeros activos que los identifican.

6.2 Métodología

6.2.1 Conjunto de estructuras de quinasas activas e inactivas

El conjunto de estructuras fue obtenido de la base de datos de confórmeros CoDNAS(Monzon et al. 2016) que correspondían a la secuencia canónica del EGFR humano (UniProtKB ID: P00533). Estructuras con residuos faltantes fueron modeladas utilizando MODELLER siguiendo las prácticas recomendadas para su módulo de *loop optimization* (Šali & Blundell 1993). Si los residuos faltantes eran más de 16, las estructuras fueron descartadas. La mayoría de estos residuos faltantes se encontraba en el *activation loop* y un *loop* entre las hebras β del lóbulo N. Cada modelo fue evaluado usando los métodos de molpdf, GA341 y DOPE(Šali & Blundell 1993) para asegurar que cada residuo modelado tenga un valor energético negativo. *B-factors* teóricos fueron obtenidos utilizando NMA (siguiendo lo descripto en el capítulo 3) y comparados con los *B-factors* experimentales. Aquellas estructuras que obtuvieron una correlación entre estos *B-factors* menor a 0.6 fueron removidas. El conjunto de estructuras final fue de 41 estructuras, 26 que habían sido clasificadas como activas y otras 15 como inactivas. El listado completo está en el Apéndice C.

6.2.2 Comparación ponderada de subespacios de NMA

Los modos vibracionales de los confórmeros fueron comparados con especial énfasis en las vibraciones de los residuos cercanos a la cavidad de la quinasa y utilizando el método de la matriz gramiana (Krzanowski 1979) (Grosso et al. 2015). La comparación del conjunto de modos de 2 confórmeros **A** y **B** sería:

$$\boldsymbol{p}_j^{AB} = \sum_{k=1}^{3N-6} (\boldsymbol{q}_j^{A'} \cdot w_k \boldsymbol{q}_k^B) \boldsymbol{q}_k^B \tag{6.1}$$

Donde:

- $\boldsymbol{q}_j$: modo $j$ésimo del confórmero **A**.
- $\boldsymbol{q}_k$: modo $k$ésimo del confórmero **B**.
- N: número de partículas.
- w_k: peso asociado al modo $k$ésimo.
- $\boldsymbol{p}_j^{AB}$: vector de la proyección entre $\boldsymbol{q}_j$ y $\boldsymbol{q}_k$, a lo largo de $\boldsymbol{q}_k$.

El peso w_k asociado al modo $\boldsymbol{q}_k$ es definido como la acumulación, de las contribuciones normalizadas del modo k al *B-factor* de los residuos cercanos a la cavidad. Es decir, cuanto más mueva un modo a los residuos cercanos a la cavidad, mayor relevancia tendrá en el cálculo de la similitud de subespacios. Para obtener este peso:

$$w_k = \frac{\sum_{i\epsilon pocket} \left[\lambda_k^{-1} \boldsymbol{q}_k \boldsymbol{q}_k^T\right]_{ii}}{\sum_{j=1}^{3N-6} \sum_{i\epsilon pocket} \left[\lambda_j^{-1} \boldsymbol{q}_j \boldsymbol{q}_j^T\right]_{ii}} \tag{6.2}$$

Donde:

- $\sum_{i\epsilon pocket}$: sólo se consideran los residuos cercanos a la cavidad.
- λ_k: autovalor del modo $k$ésimo.

Habiendo obtenido todos los vectores proyección $\{\boldsymbol{p}_j^{AB}\}_{1,N}$, se obtienen todos los productos escalares entre ellos para obtener la matriz del gramiano ($\boldsymbol{G}\epsilon\mathbb{R}^{3N-6,3N-6}$) con elementos:

$$G_{kl} = (w_k \boldsymbol{p}_k^{AB} \cdot w_l \boldsymbol{p}_l^{AB}) \tag{6.3}$$

Donde:

- G_{kl}: obtenidos del producto escalar entre los vectores proyección $\boldsymbol{p}_k$ y $\boldsymbol{p}_l$, ponderado por sus pesos.

$\boldsymbol{G}$ es diagonalizada tal que:

$$\boldsymbol{L}_G^T \boldsymbol{G} \boldsymbol{L}_G = \boldsymbol{\Omega}_G \tag{6.4}$$

Donde:

- $\boldsymbol{L}_G$: autovectores de $\boldsymbol{G}$.
- $\boldsymbol{\Omega}_G$: matriz diagonal de autovalores de $\boldsymbol{G}$.

Los autovectores representarán las direcciones en común entre los 2 espacios **A** y **B**, pero son los autovalores los que darán cuenta de la similitud de los subespacios.

Los autovalores de $\boldsymbol{G}$ varían entre 0 y 1 (Krzanowski 1979). Cuanto más común sea la dirección de $\boldsymbol{L}_{Gk}$ a ambos espacios **A** y **B**, mayor será su valor de Ω_{Gk} correspondiente. Este valor, normalizado al número de vectores, es utilizado como medida de similitud entre los 2 subespacios:

$$\zeta^{AB} = \frac{\sum_{k=1}^{3N-6} \Omega_k}{3N-6} \tag{6.5}$$

Donde:

- ζ^{AB}: similitud entre subespacios **A** y **B**.
- $3N-6$: número de modos en cada subespacio, siendo N el número de partículas.

De los autovalores también se puede obtener el índice de Kirkpatrick (Kirkpatrick 2009). Éste número, redondeado al entero mayor más próximo, es el número de dimensiones en común entre los 2 subespacios:

$$n_D^{AB} = \frac{\sum_{k=1}^{3N-6} \Omega_k}{\Omega_1} \tag{6.6}$$

Donde:

- n_D^{AB}: número de dimensiones en común.

6.2.3 SVD y vectores representativos

La similitud entre direcciones de fluctuaciones compartidas por diferentes pares de estructuras puede analizanada de la siguiente manera. Las matrices $\boldsymbol{A}^k$ de dimensión $3NxM$ se construyen con columnas que representan las direcciones $\boldsymbol{L_{Gk}}$ de cada uno de los M pares de espacios de NMA, comparados como se describió en la sección anterior:

($\boldsymbol{L_{Gk}}$, par # 1) ($\boldsymbol{L_{Gk}}$, par # 2) ... ($\boldsymbol{L_{Gk}}$, par # M)

$$A^k = \begin{Bmatrix} residuo\ \#1\ ,\ x \\ residuo\ \#1\ ,\ y \\ residuo\ \#1\ ,\ z \\ residuo\ \#2\ ,\ x \end{Bmatrix} \tag{6.7}$$

Se realiza la *Singular Value Decomposition* (SVD) (Wall et al. 2005) de cada matriz $A_{\boldsymbol{k}}$. Es decir, cada $A_{\boldsymbol{k}}$ se escribe como el producto de una matriz de columnas ortogonales $3NxM$ $\boldsymbol{U}^k$, una matriz diagonal MxM $\boldsymbol{W}^k$ con elementos de 0 a 1 (los valores singulares) y la transpuesta de una matriz ortogonal MxM $\boldsymbol{V}^k$:

$$\boldsymbol{A}^k = \boldsymbol{U}^k \begin{Bmatrix} w_1^k & & \\ & w_i^k & \\ & & w_L^k \end{Bmatrix} (\boldsymbol{V}^k)^T \tag{6.8}$$

Así, cada elemento a_{ij}^k de la matriz $\boldsymbol{A}^k$ puede ser expresado como la suma entre los productos de una columna de $\boldsymbol{U}^k$ y una fila de $(\boldsymbol{V}^k)^T$, "ponderados" por un valor singular de la matriz diagonal $\boldsymbol{W}_k$:

$$a_{ij}^k = \sum_{z=1}^{l} w_z^k u_{iz}^k v_{jz}^k \tag{6.9}$$

A estos vectores representativos los llamaremos $\boldsymbol{U}^k$. Así, el vector $\boldsymbol{U}_1^k$, tiene el valor singular más alto (w_1^k) y es considerado el modo más representativo de la matriz $\boldsymbol{A}^k$.

6.3 Resultados

La cavidad de la quinasa, definida por Hasenahuer et al., se muestra en la Figura 6.1. Esta definición incluye a todos los átomos que están dentro de un radio de 5Å de cualquier átomo del análogo de ATP unido a la quinasa de la estructura 2GS6(Hasenahuer et al. 2017). Son 53 residuos en total y están listados en el Apéndice C. También se incluye el listado de PDBs de confórmeros activos e inactivos, aunque esta clasificación no es definitiva.

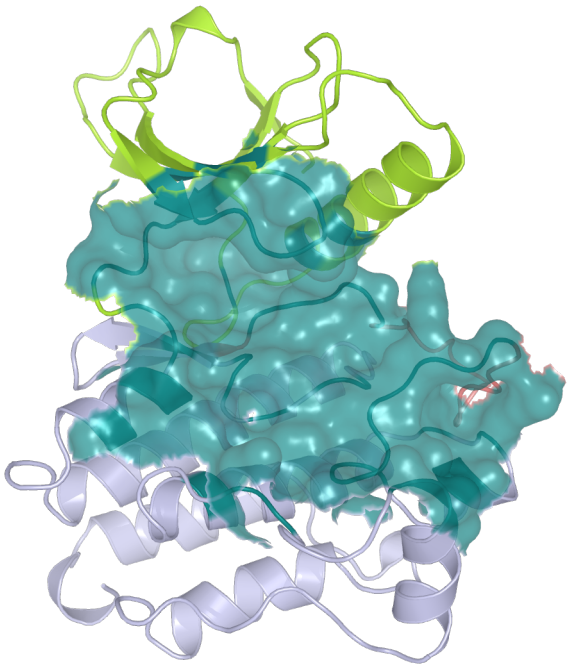

Figure 6.2: Dominio quinasa del EGFR humano. Residuos cercanos a la cavidad están representados en verde, según su superficie.

En esta sección, se explorarán dinámicas comunes entre los confórmeros activos que permiten identificarlos. Analizamos las contribuciones a las fluctuaciones de los residuos de la cavidad. En primer lugar, se utilizó la eq. 6.2 para calcular los pesos de los modos, según

su efecto en los residuos de la cavidad. La Figura 6.3 muestra las distribuciones relativas de estos pesos asociados a cada modo normal $\boldsymbol{q}_k$. El modo de más baja frecuencia ($\boldsymbol{q}_1$) es uno de los que más contribuyen a las fluctuaciones de los residuos de la cavidad. Por eso $\frac{w_k}{w_1}$ representa la contribución relativa del modo k. Valores pequeños de $\frac{w_k}{w1}$ indican que la contribución del modo normal $\boldsymbol{q}_k$ a las fluctuaciones térmicas de los residuos de la cavidad principal se puede despreciar en comparación con la contribución correspondiente del primer modo.

La Figura 6.3 muestra que para los confórmeros activos, las fluctuaciones térmicas están mayormente representadas por los 2 primeros modos de menor frecuencia. Lo opuesto sucede en los confórmeros inactivos, donde al menos los 6 modos de más baja frecuencia son relevantes en su dinámica. Es decir, las fluctuaciones de la cavidad principal de confórmeros activos están restringidas a un menor número de movimientos colectivos de baja frecuencia que los confórmeros inactivos.

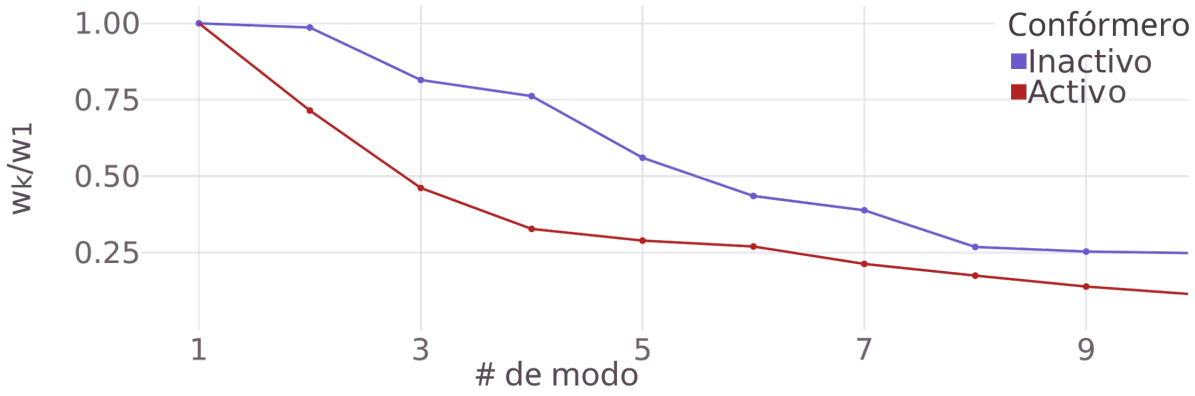

Figure 6.3: Distribuciones relativas de las contribuciones de los modos a los *B-factors* de los residuos de la cavidad.

En segundo lugar se investigaron las diferencias en la naturaleza de estos modos de baja frecuencia. Para hacerlo, se utilizó la ecuación del *Participation Number* (PN), del capítulo anterior, que aquí se repite:

$$P_i = \frac{1}{\sum_{j=1}^{N} q_{ij}^4} \tag{6.10}$$

Donde:

- P_i: número de participación del modo i, PN.
- $q_{ij}^2 = (q_{ij}^x)^2 + (q_{ij}^y)^2 + (q_{ij}^z)^2$: componente del átomo j en el modo i.
- N: número de partículas.

y se obtuvo número de participación de los modos de más baja frecuencia de los confórmeros activos e inactivos. Como ya se ha dicho, el PN describe el grado de localización

de un modo, dando el número de residuos que se mueven significativamente bajo un cierto modo normal. Las distribuciones del PN de los primeros 2 modos aparecen en la Figura 6.4, calculadas como $\frac{P_1}{N}$ y $\frac{P_2}{N}$, donde N es el número total de residuos.

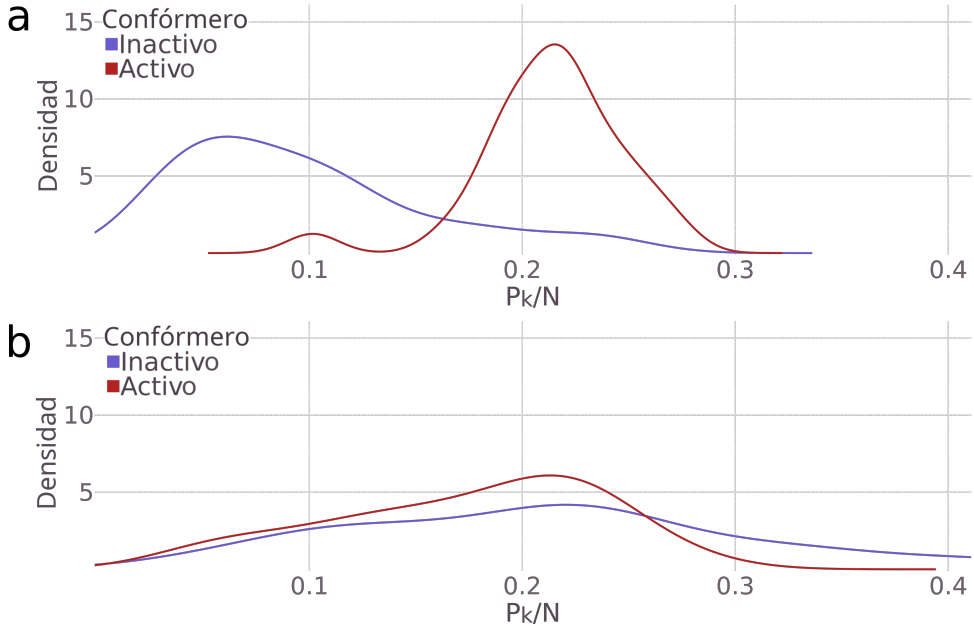

Figure 6.4: Distribuciones de las fracciones de residuos **a** Número de participación **b)**

En promedio, los primeros modos de los confórmeros activos son más deslocalizados ($\langle PN_1 \rangle_A = 0.21 \pm 0.03$) en comparación a los primeros modos de los confórmeros inactivos ($\langle PN_1 \rangle_I = 0.10 \pm 0.05$), involucrando más del doble de residuos. En cambio, los segundos modos tienen valores de PN mucho más dispersos ($\langle PN_2 \rangle_A = 0.17 \pm 0.06$ para los activos y $\langle PN_2 \rangle_I = 0.10 \pm 0.06$ para los inactivos). Es decir, en términos de deslocalización, es el primer modo normal el que separa a los confórmeros inactivos de los activos. Es decir, $\langle PN_1 \rangle_I$ indica un gran impacto de mutaciones no activadoras en la colectividad original del modo normal más bajo observado en las conformaciones activas. Por lo tanto, se espera que los aspectos funcionales de la EGFR quinasa impliquen movimientos coordinados entre los residuos que se reflejan principalmente en el modo normal de conformaciones activas de menor frecuencia.

Se ha demostrado que la conservación de un modo correlaciona con su colectividad (Maguid et al. 2008), lo que indicaría una dinámica más conservada en los confórmeros activos que en los inactivos. Para determinar si la flexibilidad en los residuos de la cavidad está conservada se calcularon las correlaciones de Pearson entre los *B-factors* de los carbonos alfa de los residuos de la cavidad. En la Figura 6.5 se muestran las distribuciones obtenidas para confórmeros activos e inactivos con correlaciones promedio de $\rho_B = 0.92$ y $\rho_B = 0.81$, respectivamente. Los confórmeros activos presentan una flexibilidad más conservada. Las distribuciones de la Figura 6.5 presentan un Kolmogorov-Smirnov (KS) de 0.57, por lo que sus diferencias son estadísticamente significativas, si bien como puede

verse en la Figura 6.5 estos patrones de flexibilidad no son lo suficientemente precisos como para diferenciar a los confórmeros.

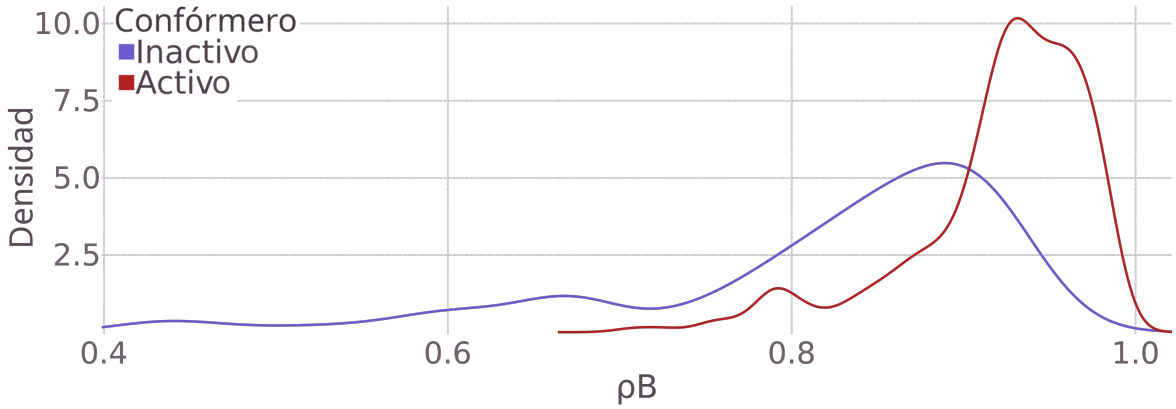

Figure 6.5: Correlación entre *B-factors* de cada modo

Los movimientos vibratorios asociados a fluctuaciones estructurales de los confórmeros de quinasa EGFR pueden compararse adicionalmente mediante el cálculo de la matriz gramiana $\boldsymbol{G}$ ponderada por los pesos de los modos w_k (ver eq. 6.1, eq. 6.2, eq. 6.3). La diagonalización de $\boldsymbol{G}$ proporciona un conjunto de nuevas direcciones de movimiento $\{\boldsymbol{L}_{\boldsymbol{G}k}\}$ (donde $\{k = 1, 3N - 6\}$), dadas en orden decreciente de sus autovalores correspondientes $\{\Omega_k\}$ (donde $\{k = 1, 3N - 6\}$). Los valores de Ω_k varían en un rango de 0 a 1. Un valor de $\Omega_k \approx 1$ indicaría que la dirección de movimiento definida por el vector $\boldsymbol{L}_{\boldsymbol{G}k}$ correspondiente es compartida totalmente por ambos confórmeros. Por otro lado, un valor de $\Omega_k \approx 0$ indicaría que la dirección $\boldsymbol{L}_{\boldsymbol{G}k}$ correspondiente es específica para un solo confórmero. El valor de ζ^{AB}, (ver eq. 6.5), proporciona nuestra medida de similitud final entre espacios vibracionales ponderados de los confórmeros A y B. La Figura 6a muestra la distribución de los valores de ζ^{AB} obtenidos sobre todos los pares de conformaciones activas e inactivas. La diferencia entre ambas distribuciones se valida estadísticamente mediante el valor estadístico KS de 0.97.

Esto significa que ζ^{AB} separa las distribuciones activas e inactivas mejor que los coeficientes de correlación lineal de Pearson ρ_B entre los *B-factors* de carbonos alfa. Es decir, la información dinámica sobre la direccionalidad de los movimientos colectivos permite una mejora significativa en la distinción entre conformaciones activas e inactivas con respecto a resultados anteriores utilizando patrones de flexibilidad (ver Figura 6.5). Además, el número efectivo de dimensiones compartidas por pares de confórmeros, dados por el índice Kirkpatrick n_D^{AB} (eq. 6.9), también permite una clara diferenciación entre confórmeros activos e inactivos. La Figura 6.6b muestra las distribuciones de valores de n_D^{AB} para los confórmeros activos e inactivos. Un valor promedio de 1.7 entre pares de confórmeros activos indica que las similitudes dinámicas entre ellos se pueden reducir eficientemente a 1 o 2 direcciones comunes de movimientos. Por el contrario, las similitudes dinámicas entre los confórmeros inactivos están dispersos entre 3 y 4 (valor promedio de 3.4) diferentes direcciones, que son menos compartidas entre los pares de confórmeros.

El valor estadístico KS correspondiente entre ambas distribuciones es 0.89. Es decir, n_D^{AB} resulta ser ligeramente menos efectivo que ζ^{AB} para separar las dinámicas activas e inactivas.

La identidad de las dos direcciones comunes $\boldsymbol{L}_{G1}$ y $\boldsymbol{L}_{G2}$ para pares de confórmeros activos puede explorarse analizando sus proyecciones sobre la base de los modos normales originales. La contribución promedio del primer modo normal de más baja frecuencia a $\boldsymbol{L}_{G1}$ es $0.73 \approx 0.41$ y la contribución de los segundos modos normales de menor frecuencia a $\boldsymbol{L}_{G2}$ es $0.6 \approx 0.41$. Por lo tanto, los confórmeros activos comparten una dirección de movimiento representada principalmente por los modos normales de menor frecuencia. Esto está de acuerdo con los resultados previos obtenidos por Coveney et al. (Wan & Coveney 2011) (Wan et al. 2012) usando MD y PCA, que muestra que el primer modo PCA distingue entre estados activos e inactivos. Por el contrario, valores de $0,59 \pm 0.35$ y 0.35 ± 0.30 se obtienen para las contribuciones del primero y segundo modos normale a $\boldsymbol{L}_{G1}$ y $\boldsymbol{L}_{G2}$ de los confórmeros inactivos. Es decir, mientras que la dinámica de los pares de confórmeros activos revela una dirección de movimiento común que corresponde a su frecuencia natural más baja de vibración, la dinámica de los pares de confórmeros inactivos no presenta patrones únicos que puedan asignarse a modos normales originales individuales.

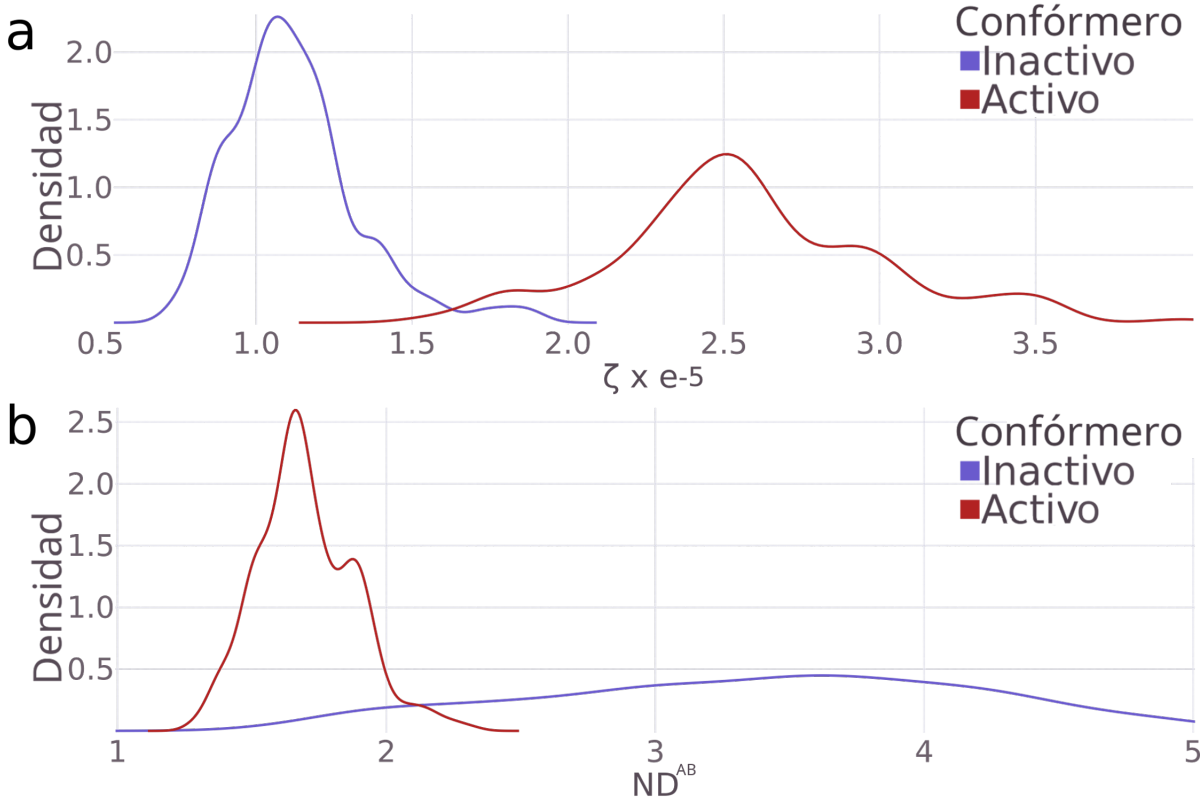

Figure 6.6: Histogramas de valores de **a** medida de similitud ζ^{AB} entre pares de subespacios vibracionales ponderados, y **b** el número efectivo de dimensiones n_D^{AB} compartidas por pares de confórmeros.

Como se ha mostrado, los pares de confórmeros activos comparten direcciones comunes de movimiento representadas por la dirección $\boldsymbol{L}_{G1}$. Para obtener una huella que los carac-

terice entre el conjunto completo de confórmeros activos, se requiere un análisis adicional. SVD, como técnica de compresión de datos, destaca las principales características comunes de las direcciones originales de $\boldsymbol{L}_{Gk}$ dentro de unos pocos modos representativos $\boldsymbol{U}_i^k$. La Figura 6.7 muestra las distribuciones de la superposición entre modos representativos $\boldsymbol{U}_1^k$, y direcciones $\boldsymbol{L}_{Gk}$ originales. Como puede verse, nuestro conjunto completo de confórmeros activos comparte direcciones comunes representadas por el modo representativo SVD correspondiente. En contraste, el correspondiente $\boldsymbol{U}_1^1$ en confórmeros inactivos no puede usarse como vector representativo del conjunto. Finalmente, las direcciones $\boldsymbol{L}_{G2}$ y $\boldsymbol{L}_{G3}$ son diferentes entre pares de confórmeros activos o inactivos. En resumen, todos los confórmeros activos de nuestro conjunto de datos comparten dinámicas comunes que finalmente pueden asociarse a su modo de frecuencia más baja. Por el contrario, la dinámica de los confórmeros inactivos resulta heterogénea y no se pueden encontrar patrones dinámicos únicos que los incluyan a todos.

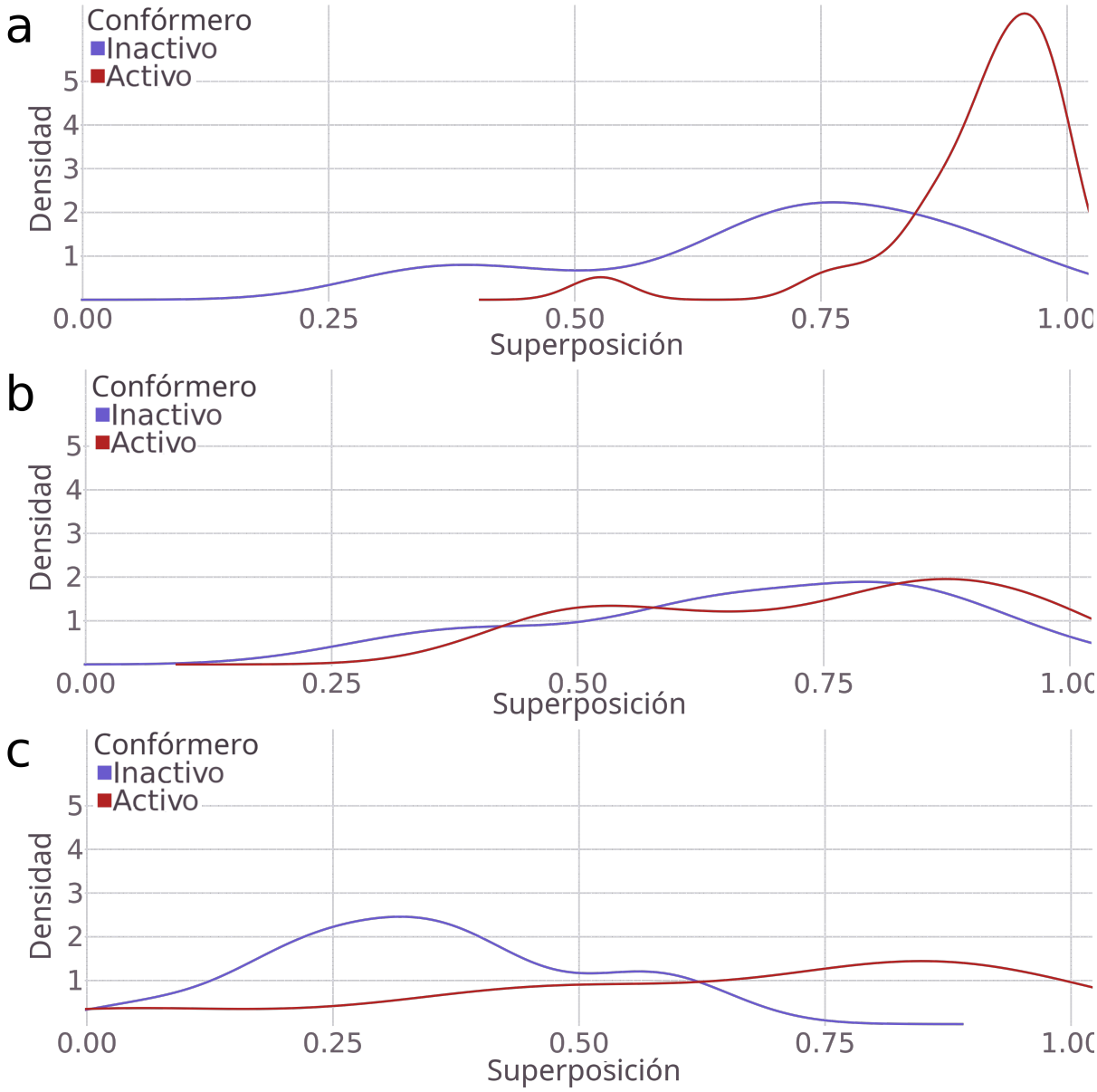

Figure 6.7: Histogramas de superposición entre el primer vector SVD representativo de modos ($\boldsymbol{U}_1^k$), y las direcciones originales **a** $\boldsymbol{L}_{G1}$ **b** $\boldsymbol{L}_{G2}$ y **c)** $\boldsymbol{L}_{G3}$.

La Figura 6.8 muestra ese modo "huella" $\boldsymbol{U}_1^1$, compartido por todos los confórmeros activos, y también el modo menos representativo $\boldsymbol{U}_1^2$. Ambos modos describen desplazamientos relativos entre los dos lóbulos (lóbulos N y C) del dominio de la quinasa. Particularmente, movimientos de lóbulo N y la hélice αC. Los reordenamientos locales de la hélice αC se han relacionado previamente con el cambio conformacional de la quinasa, asociado a la unión del ligando (Morando et al. 2016). Además, una supresión relativa de las deformaciones del *loop* activo también caracteriza los modos $\boldsymbol{U}_1^1$ y $\boldsymbol{U}_1^2$ con respecto a los movimientos de los modos correspondientes de conformaciones inactivas, de acuerdo con los resultados MD anteriores que revelan una hélice αC relativamente más rígida y un *loop* más flexible en el estados activos (Wan & Coveney 2011) (Wan et al. 2012) (Li et al. 2014). En aras de la comparación, también mostramos los modos correspondientes para los confórmeros inactivos.

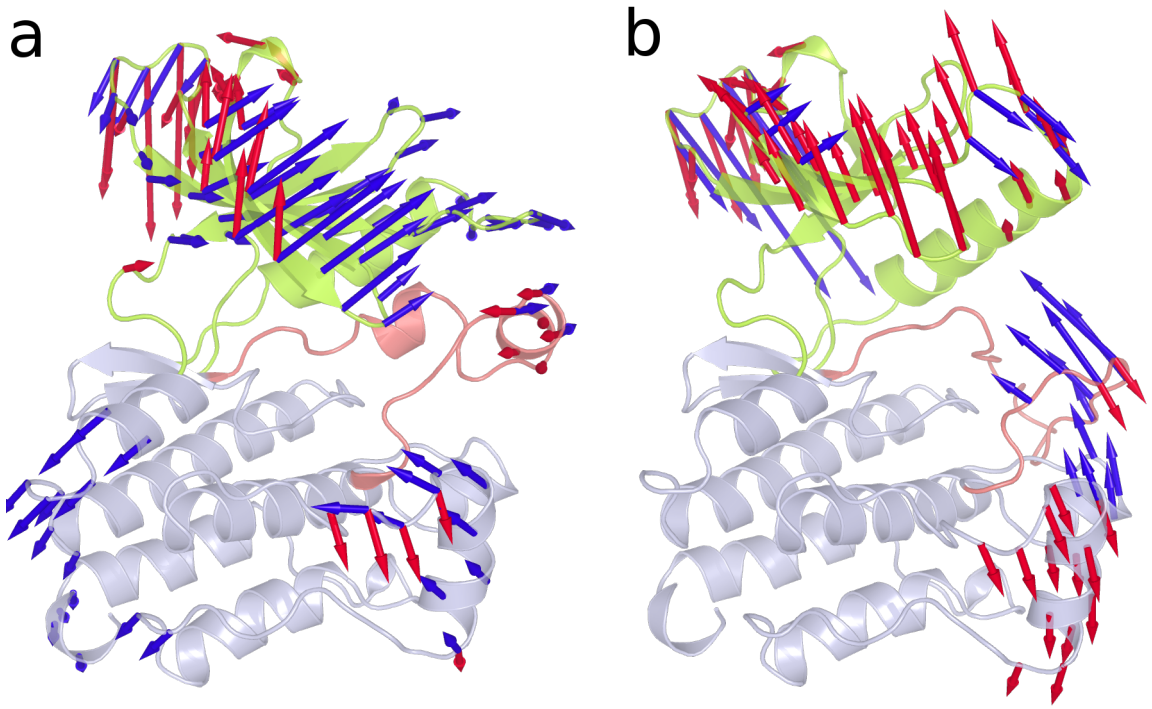

Figure 6.8: Modos SVD representativos $\boldsymbol{U}_1^1$ (rojo) y $\boldsymbol{U}_1^2$ (azul). En violeta, el lóbulo C, en verde, el lóbulo N y en rojo, el *activation loop*. Confórmeros inactivos (**a**) y activos (**b**).

Para evaluar los efectos de los modos $\boldsymbol{U}_1^1$ y $\boldsymbol{U}_1^2$ sobre la conformación activa del dominio de quinasa EGFR, calculamos los cambios de radio de giro (ΔR_g) que estos vectores generan. La Figura 6.9 muestra la distribución de valores de ΔR_g debido a desplazamientos en la dirección de estos modos. Como puede verse, su efecto sobre los confórmeros activos difiere significativamente del efecto de los desplazamientos de los modos correspondientes de los confórmeros inactivos. Los modos $\boldsymbol{U}_1^1$ y $\boldsymbol{U}_1^2$ de los confórmeros activos conducen a conformaciones más extendidas que implican una separación tipo bisagra de los lóbulos N y C.

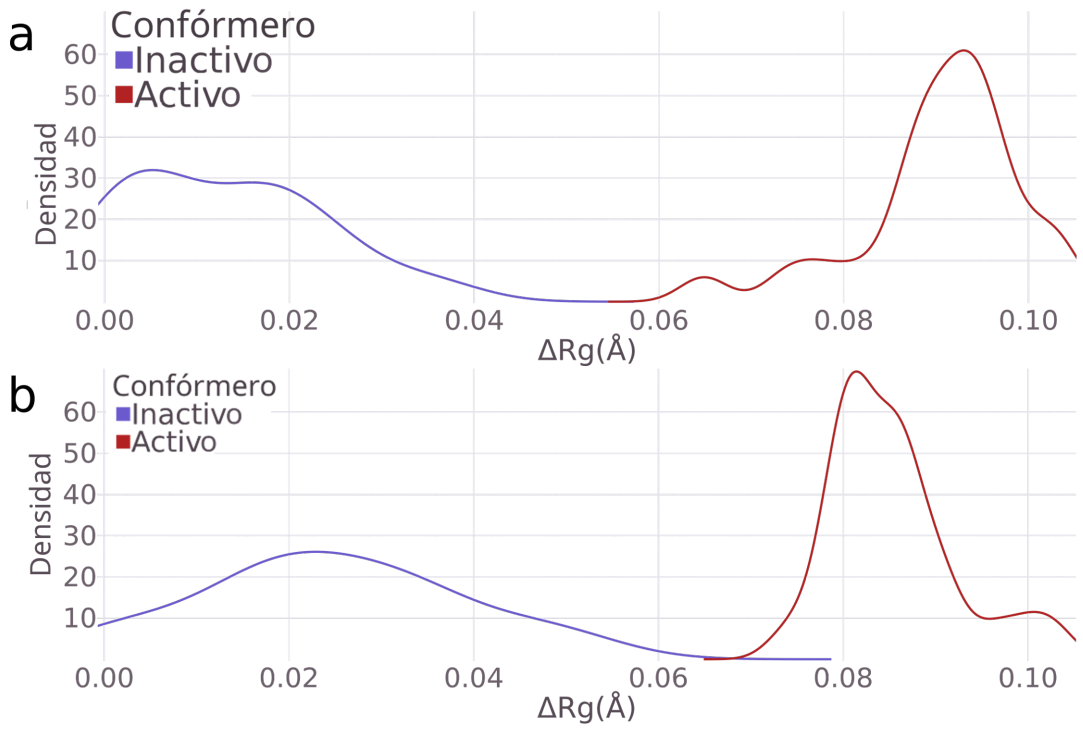

Figure 6.9: Histogramas de cambios en el radio de giro ΔR_g de la quinasa EGFR debido a desplazamientos a lo largo de los modos representativos de SVD **a** $\boldsymbol{U}_1^1$ **b** $\boldsymbol{U}_1^2$, tanto para los confórmeros activos como los inactivos.

Los movimiento en las direcciones de los modos $\boldsymbol{U}_1^1$ y $\boldsymbol{U}_1^2$ en confórmeros activos son consistentes con los movimientos previamente reportados (Shan et al. 2012) (Shan et al. 2013) utilizando simulaciones de dinámica molecular, que describen la transición conformacional activa-inactiva. Shan et al. informaron una apertura de los dos lóbulos para permitir el despliegue local en la región de la bisagra (una dinámica llamada *cracking* (Miyashita et al. 2003) (Ansari et al. 1985)), antes del cierre de los lóbulos para reestabilizar su conformación inactiva. Además, Coveney et al. (Wan & Coveney 2011) (Wan et al. 2012) informaron movimientos de bisagra y cizalladura (*shear*) entre los lóbulos C y N asociados a los modos PCA primero y segundo. Nuestro análisis sugiere que estos hallazgos son en realidad un sello distintivo de la dinámica de los confórmeros activos cuando se trata de diferenciarlos de los confórmeros inactivos.

Finalmente exploramos aspectos funcionales de los movimientos en la dirección de los modos $\boldsymbol{U}_1^1$ y $\boldsymbol{U}_1^2$. Para este propósito, calculamos los cambios en el volumen de la cavidad del bolsillo principal del sitio activo debido a los desplazamientos estructurales en la dirección de estos modos. Los volúmenes de la cavidad se calcularon utilizando ANA. La Figura 6.10 muestra la distribución de las diferencias en los volúmenes de la cavidad entre las estructuras de los confórmeros activos e inactivos desplazados previamente una magnitud que se obtiene:

$$Ai = \sqrt{\frac{2kBT}{\lambda_i}} \tag{6.11}$$

Donde:

- A_i: amplitud en Å del modo i a temperatura ambiente.
- k_B: constante de Boltzmann.
- T: temperatura ($300K$).
- λ_i: autovalor del modo normal i.

λ_i fue ajustado para predecir mejor a las fluctuaciones teóricas de los residuos, utilizando los factores de temperatura experimentales correspondientes. Tal como se detalla en el capítulo 3.

Como se puede ver, los movimientos en la dirección de los dos modos normales de baja frecuencia de confórmeros activos conducen a cambios formales que implican cambios más grandes en los volúmenes de la cavidad que los movimientos en la dirección de los modos correspondientes de los confórmeros inactivos. Los cambios en los volúmenes de la cavidad podrían tener un impacto posterior en la afinidad del ligando y, por lo tanto, en la regulación de la función biológica del bolsillo del sitio activo de la proteína quinasa EGFR. Aquí, hemos identificado dinámicas comunes compartidas solo por los confórmeros activos de la quinasa EGFR. Dado que estas dinámicas están asociadas a cambios significativos en el volumen del bolsillo principal del sitio activo de EGFR quinasa, se deben esperar sus impactos funcionales. Trabajos anteriores han demostrado que los inhibidores pueden actuar al obstaculizar y / o cambiar la dirección de los movimientos de proteínas específicas. Además, la identificación de residuos dinámicamente importantes asociados a estos movimientos se puede realizar utilizando métodos previamente desarrollados (Salda et al. 2016) (Zheng et al. 2005). Por lo tanto, se pueden emplear estrategias para el desarrollo de nuevos inhibidores que obstaculicen o modifiquen estos movimientos relevantes que los distinguen. Particularmente, compuestos que interrumpen los desplazamientos relativos entre los dos lóbulos (lóbulos N y C), enfocándose en los movimientos del lóbulo N y la hélice α C. Además, los inhibidores especialmente diseñados para interactuar con el bucle rico en glicina podrían modificar fuertemente estos movimientos concertados.

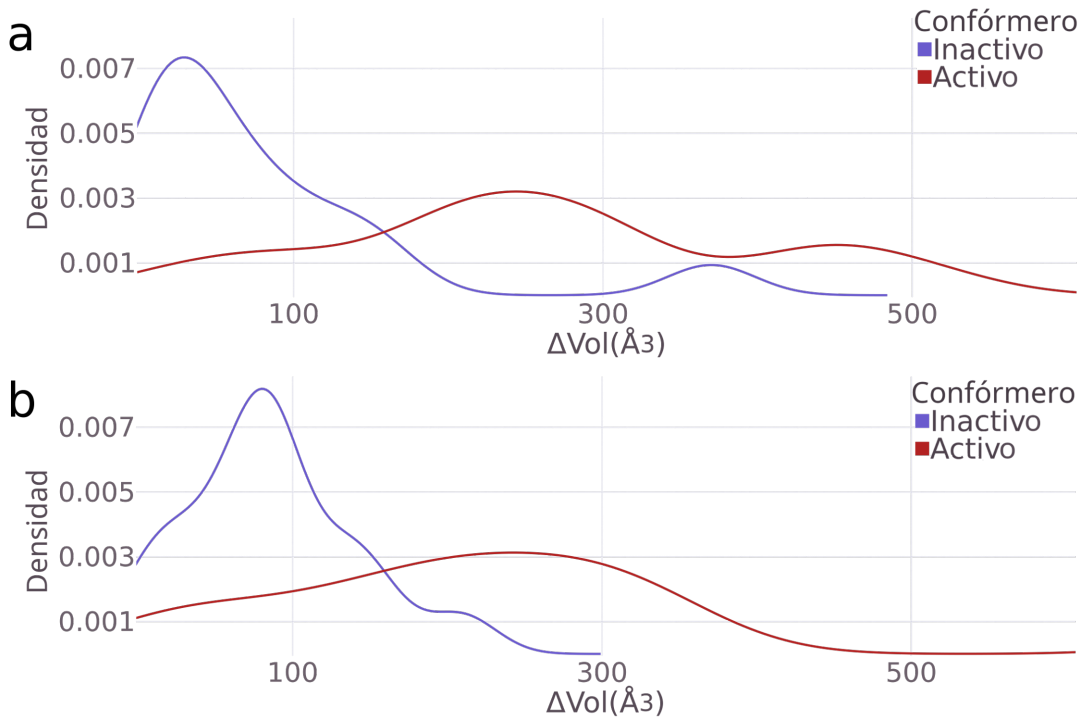

Figure 6.10: **a** Volumen **b** Volumen

Chapter 7

Lipid Binding Proteins

7.1 Introducción

Moléculas lipofílicas hidrofóbicas como ácidos grasos, esteroides, retinoides y sus derivados participan en una gran variedad de funciones dentro de una célula, incluyendo almacenamiento de energía, señalización, regulación de la expresión génica, roles hormonales y regulación de la permeabilidad de la membrana, entre otros. Su insolubilidad en agua y su potencial degradación oxidativa requieren su transporte coordinado y disponibilidad, protección y regulación en todo el entorno hidrofílico intracelular. Las proteínas solubles de unión a lípidos (LBP) son un grupo de proteínas abundantes que son responsables de estas tareas en todo el entorno acuoso dentro de numerosos tipos de células y fluidos corporales de diferentes organismos. Los parásitos helmintos tienen un metabolismo lipídico restringido y deben adquirir lípidos simples y complejos de sus organismos anfitriones, por lo tanto, las LBP desempeñan funciones importantes para el crecimiento y desarrollo de parásitos (Franchini et al. 2015).

Las proteínas de unión a ácidos grasos y retinol (FARs) son LBPs que se han descrito como componentes de los fluidos E / S de los nematodos parásitos y se supone que juegan un papel esencial en la adquisición de lípidos y la distribución de nutrientes, así como la posible amortiguación del sistema inmune del huésped (Geng et al. 2002) (Garofalo et al. 2003). Las FAR junto con las poliproteínas de nematodos / alergenos (NPA) (Solovyova et al. 2003) son proteínas pequeñas (14-20 kDa) ricas en hélice que se unen al retinol y los ácidos grasos y no tienen contrapartidas reconocibles en otros grupos de animales (Franchini et al. 2015). Dadas estas características, se ha demostrado que las FAR son útiles para el serodiagnóstico y las vacunas experimentales (Zhan et al. 2018). Además, existe evidencia de que las FAR de nematodos filariales pueden unir fármacos antihelmínticos (Sani & Vaid 1988).

La enfermedad del anquilostoma es una infección helmíntica altamente debilitante que está

relacionada con la anemia por deficiencia de hierro (IDA) y se situa en países tropicales en desarrollo, con una prevalencia estimada de 451 millones de casos que causan 1,6 millones de años vividos con discapacidad (YLD, por sus siglas en inglés) (Vos et al. 2017). *Necator americanus*, junto con *Ancylostoma duodenale* y *Ancylostoma ceylanicum*, son los agentes causantes de la mencionada "enfermedad del anquilostoma". Es importante tener en cuenta que *N. americanus* es responsable de la mayoría de los casos en todo el mundo. Esta parasitosis ha sido erradicada con éxito de los países desarrollados por los tratamientos farmacológicos masivos y por el desarrollo económico (Loukas et al. 2016). Sin embargo, la presencia de la enfermedad sigue siendo alta en muchos países de bajos ingresos medios como la región norte de Argentina (Zonta et al. 2010).

Las FAR se producen en varias isoformas, y se ha descubierto que Na-FAR-1 es altamente expresada en la forma adulta (Daub et al. 2000). En la actualidad, se resolvieron dos estructuras de FAR ortólogas, una de *Necator americanus* (Na-FAR-1 por RMN y RX; PDBs: 4UET y 4XCP, respectivamente) (Rey-Burusco et al. 2015) y otra de *Caenorhabditis elegans* (Ce-FAR-7, por RX; PDB: 2W9Y) (Jordanova et al. 2009). Ambas presentan estructuras ricas en hélices α similares, con ciertas diferencias estructurales. Particularmente, el tamaño y la forma de sus cavidades internas son diferentes, denotando diferencias en su selectividad de ligando. Na-FAR-1, en sus conformaciones apo y holo, presenta una cavidad interna de unión a ligando más grande y más compleja(Rey-Burusco et al. 2015).

Entre las LBP solubles, otro grupo interesante es la familia de las proteínas de unión a los ácidos grasos (FABPs, por sus siglas en inglés) que presentan afinidades de unión preferenciales para los ácidos grasos de cadena larga (Friedman et al. 2006) (Storch & McDermott 2009). Mientras que las FAR se han encontrado exclusivamente en nematodos, FABP se pueden encontrar en vertebrados e invertebrados (McDermott et al. 1999). A pesar de su baja identidad de secuencia y su divergencia funcional, probablemente relacionada con sus preferencias particulares de unión a lípidos, comparten una estructura terciaria común (Zimmerman & Veerkamp 2002). Todas tienen estructuras de barril β similares, que encierran el ácido graso unido. El volumen de la cavidad interna está determinado por las cadenas laterales de los residuos que definen la superficie molecular que lo encierra. Estos residuos varían entre los diferentes tipos de FABP y determinan la especificidad de ligando de la cavidad. Diversas mutaciones puntuales, realizadas en residuos que recubren la cavidad de diferentes tipos de FABP, han demostrado que modifican la estabilidad conformacional de las proteínas, la especificidad y la afinidad de los ligandos (Sha et al. 1993) (Laulumaa et al. 2018). Varios estudios, basados en el análisis de cristales y soluciones, predijeron la forma en que los ácidos grasos entran y salen del sitio de unión de FABP (Hodsdon & Cistola 1997); esto es fundamental para comprender el mecanismo molecular de la selección y entrega de ligandos en las FABP (Friedman et al. 2006) (Tsfadia et al. 2007) (Guo et al. 2019). Estos trabajos han demostrado la

importancia de ciertos residuos y dominios en la dinámica de la proteína, confirmando observaciones realizadas por diferentes métodos experimentales y permitiendo formular hipótesis sobre las funciones propuestas de estas proteínas en la célula. Si bien los nematodos también producen FABP de barril β, las razones por las que los nematodos se han especializado en el uso de proteínas ricas en hélices α siguen sin estar claras.

Una comprensión de cómo la diversidad conformacional de las FAR contribuye a su multiplicidad de ligandos, variando las afinidades relativas por diferentes moléculas lipofílicas, podría iluminar su papel en el parasitismo y sugerir posibles objetivos para intervenciones terapéuticas. Los ensayos de unión a ligandos basados en fluorescencia y la titulación de Na-FAR-1 con oleato de sodio monitoreado por RMN revelan su alta multiplicidad de ligandos (Rey-Burusco et al. 2015). Estos estudios sugieren que el plegamiento de hélices α tiene una mayor propensión para unirse a una mayor variedad y cantidad de FA y otras clases de lípidos que el plegamiento del barril β. Además, la unión del ligando a Na-FAR-1 induce cambios sustanciales de *chemical shift* en los residuos en toda la proteína, lo que indica cambios conformacionales significativos que permiten que la estructura se expanda.

En este capítulo exploraremos la relación estructura-dinámica-función de Na-FAR-1 usando simulaciones de MD largas combinadas con PCA en sus formas apo y holo. Analizamos su plasticidad y el impacto de las diferentes conformaciones en el volumen de la cavidad de unión al ligando. Nos centramos en las relaciones dinámicas entre las fluctuaciones de proteínas, los cambios en la cavidad y las diferentes conformaciones del ligando incluido. Los resultados obtenidos se comparan con simulaciones MD de la proteína de unión a ácidos grasos intestinales de rata (I-FABP) con el típico plegamiento de barril β FABP, y se realiza el ortólogo Ce-FAR-7. Nuestro análisis revela que Na-FAR-1 abarca una compleja cavidad interna de unión a ligando con una notable plasticidad conformacional que permite el cambio reversible entre estados distintos de acuerdo con las diferentes conformaciones del ligando adjunto.

7.2 Métodos

7.2.1 Dinámica Molecular y su análisis

Se realizaron simulaciones de MD de Na-FAR-1 en su forma apo (PDB: 4UET), holo (PDB: 4XCP) e I-FABP en su forma apo (PDB: 4UET) tanto en sus formas apo , Na-FAR-1, y 1IFB, I-FABP) y holo (PDB: 4XCP, Na-FAR-1, y 2IFB, respectivamente) con palmitato en sus bolsillos de unión, y Ce-FAR-7 en su forma no ligada (PDB: 2W9Y). Los 6 sistemas fueron preparados como se indica en el capítulo 2 y sus trayectorias de producción fueron de 3μs, tomando cuadros de las trayectorias cada 10ps, con estos, se obtuvieron los componentes principales de los sistemas, como detalla el capítulo 3.

7.2.2 Cavidad del ligando: definición, volumen y flexibilidad

Las cavidades de ligando se han definido mediante inspección visual del promedio de estructuras MD equilibradas y conocimiento previo de cada sistema. La lista completa de residuos que recubren la liga principal se proporciona en el Apéndice C. Los volúmenes de las cavidades se calcularon con ANA.

7.3 Resultados

7.3.1 LBPs de hélices α

Mientras que la mayoría de las FABPs presentan un plegamiento de barril β, las FAR revelan un plegamiento de hélices α inusual. En el caso de Na-FAR-1, consiste en una estructura en forma de cuña compuesta de 11 hélices con diferentes longitudes que encierran una cavidad interna de unión a ligando. El cambio conformacional general de unión a ligando implica un RMSD global de 0,98 Å entre confórmeros, calculado a partir de la superposición de carbonos α de las estructuras promedio (a lo largo de la trayectoria) de apo y holo, (ver Figura 7.1b y d). Ambas estructuras holo para Na-FAR-1 e I-FABP están unidos a una sola molécula de palmitato. Es importante tener en cuenta que este es el ligando preferido de Na-FAR-1 en un entorno biológico (Rey-Burusco et al. 2015).

Las principales distorsiones estructurales tras la unión del ligando se localizan en las hélices α2, α7, α10 y los bucles entre α2, α2-α3, α4-α5 y α7-α8. Entre estos Elementos de Estructura Secundaria (SSE, por sus siglas en inglés), α4-α5 y α7-α8 han mostrado las mayores fluctuaciones de la raíz cuadrática media (RMSF, por sus siglas en inglés; ver Apéndice C) durante nuestras simulaciones MD, particularmente los residuos 39-45 en el bucle α4-α5 y los residuos 100−103 en bucle α7-α8 presentan la mayor flexibilidad relativa. El cambio estructural del bucle α4-α5 durante la unión del ligando es esperablef, ya que este bucle es parte de la abertura única de la cavidad de unión del ligando, ubicada entre este bucle y las hélices α6 y α7. Además de esta apertura, el bucle α7-α8 se ha propuesto previamente(Rey-Burusco et al. 2015) como el principal candidato para participar de la entrada del ligando a través de la porción de la cavidad accesible al solvente.

En este punto, es interesante notar que el RMSD entre las estructuras holo y apo promedio de Na-FAR-1 es solo 1.58 Å. Sin embargo, pequeñas distorsiones estructurales pueden implicar grandes cambios en las cavidades de las proteínas (Monzon et al. 2017). Además, en los casos en que las proteínas exploran múltiples confórmeros durante las simulaciones, el promedio estructural no es una buena estadística. Por lo tanto, en lo que sigue, se discute la identificación de diferentes confórmeros y su impacto en la cavidad ligando.

Los histogramas de volúmenes de Na-Far-1 de la Figura 7.1a, c calculados sobre el conjunto de estructuras recogidas durante las simulaciones equilibradas de MD de apo y holo-Na-FAR-1, muestran algunas de sus diferencias. Sus valores promedio son $1353 \approx 254$ y $1397 \approx 266$ ^{3} respectivamente. Estos valores difieren de los 940 y 2170 ^{3} correspondientes calculados en las estructuras experimentales iniciales (Rey-Burusco et al. 2015). Como hemos señalado anteriormente, definimos las cavidades internas de acuerdo con las estructuras promedio obtenidas de nuestras simulaciones MD. Las distribuciones que se muestran en la Figura 7.1a, c son el resultado de las fluctuaciones térmicas de la proteína que pueden involucrar diferentes cambios conformacionales a lo largo de las simulaciones de MD de 3μs. Las fluctuaciones de las hélices que forman la cavidad presentan pequeños rearreglos en la estructura a las proteínas que pueden conducir a cambios significativos en el tamaño de la cavidad interna (Monzon et al. 2017). Los histogramas que se muestran en la Figura 7.1a, c revelan que la cavidad interna puede duplicar su volumen debido a las fluctuaciones de los residuos. Si bien la distribución de volúmenes para apo-Na-FAR-1 corresponde a una distribución gaussiana que puede asociarse con fluctuaciones térmicas alrededor de una conformación proteica única, este no es el caso de holo-Na-FAR-1.

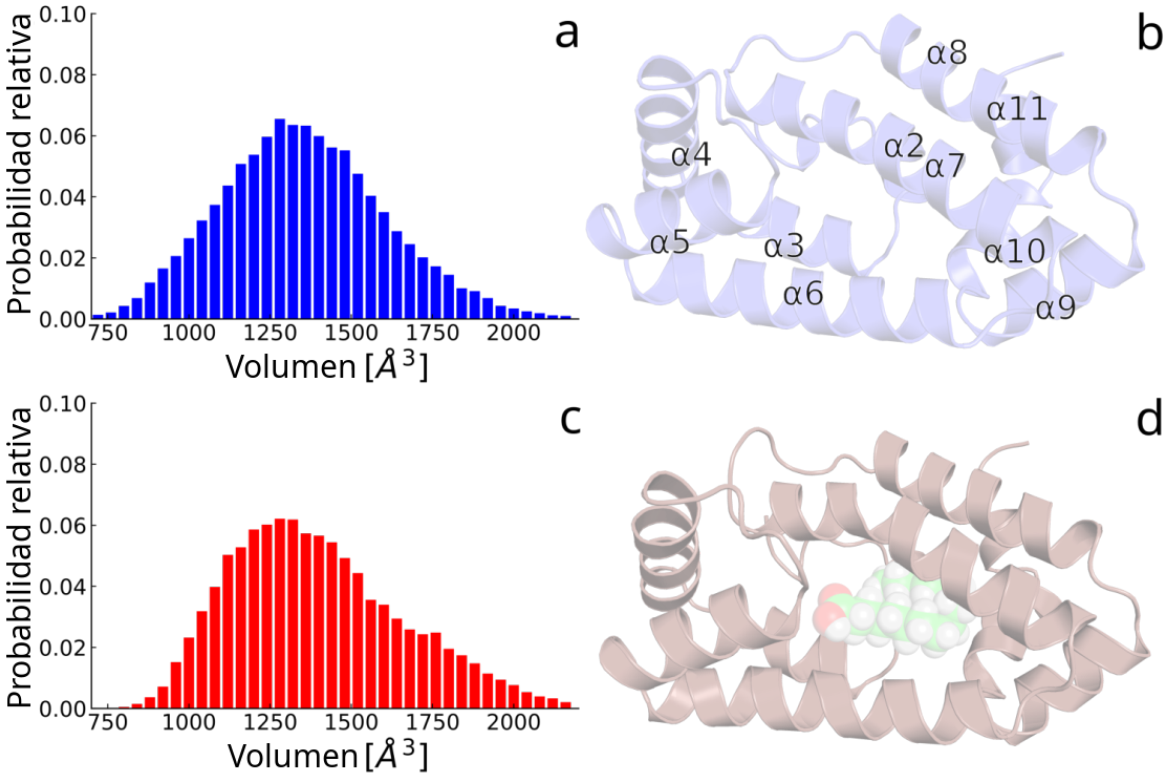

Figure 7.1: Estructuras promedio apo (**b**) y holo (**d**) de Na-Far1, obtenidas de las simulaciones. Se indican los nombres de las hélices. Histogramas de los volúmenes a lo largo de las trayectorias de apo (**a**) y holo (**c**) Na-Far-1. Notar que el volumen de la apo sigue una distribución normal, a diferencia de la holo.

Los cambios en el volumen de la cavidad suelen estar asociados con fluctuaciones de la estructura proteica. Por lo tanto, para dilucidar estas relaciones, las simulaciones MD se analizaron en términos de PCA. Los primeros y segundos modos de PCA de apo y holo-Na-FAR-1 se muestran en la Figura 7.2a, b. En ambos confórmeros, estos 2 modos involucran

el movimiento concertado de los residuos ubicados en las hélices $\alpha 4$, $\alpha 5$, el bucle entre ellas y el término C de la hélice $\alpha 7$. De acuerdo con observaciones experimentales (Rey-Burusco et al. 2015), la última hélice tiene el mayor impacto en el volumen de la cavidad, mientras que la primera forma la puerta de entrada del ligando (puerta $\alpha 4$-$\alpha 5$).

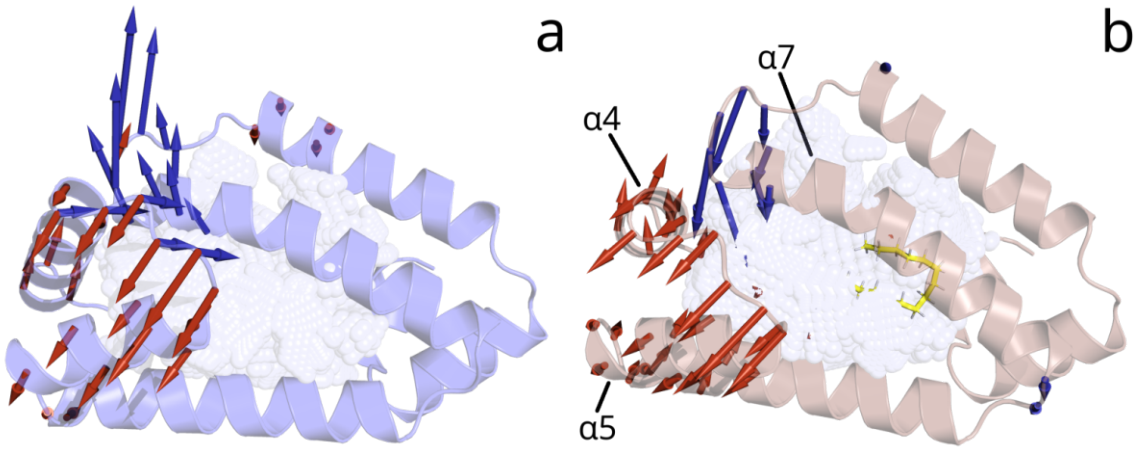

Figure 7.2: Modos de PCA 1(rojo) y 2(azul) de la apo(**a**) y holo (**b**) Na-Far-1. Las hélices que sufren el mayor desplazamiento son indicadas de nuevo.

La Figura 7.3a, b muestra la proyección del conjunto de conformaciones instantáneas de MD de apo y holo-Na-FAR-1 en sus primeros y segundos modos de PCA correspondientes. Las fluctuaciones térmicas de apo-Na-FAR-1 se revelan como combinaciones graduales de ambos modos sin mostrar una prevalencia significativa de distorsiones estructurales en ninguna dirección específica. Es decir, apo-Na-FAR-1 no visita ninguna conformación nueva que persista durante una cantidad significativa de tiempo durante la simulación MD. Por el contrario, podemos observar que holo-Na-FAR-1 evidencia la existencia de tres confórmeros diferentes: dos confórmeros estables que presentan distorsiones estructurales principalmente en ambos sentidos de la dirección del primer modo PCA (confórmeros A y B), y un tercer confórmero C en la dirección del segundo modo PCA. Las proyecciones del conjunto de instantáneas MD de apo y holo-NaFAR-1 en sus correspondientes terceros modos de PCA no muestran la existencia de nuevos confórmeros estables con distorsiones estructurales en la dirección de estos modos (ver Apéndice C).

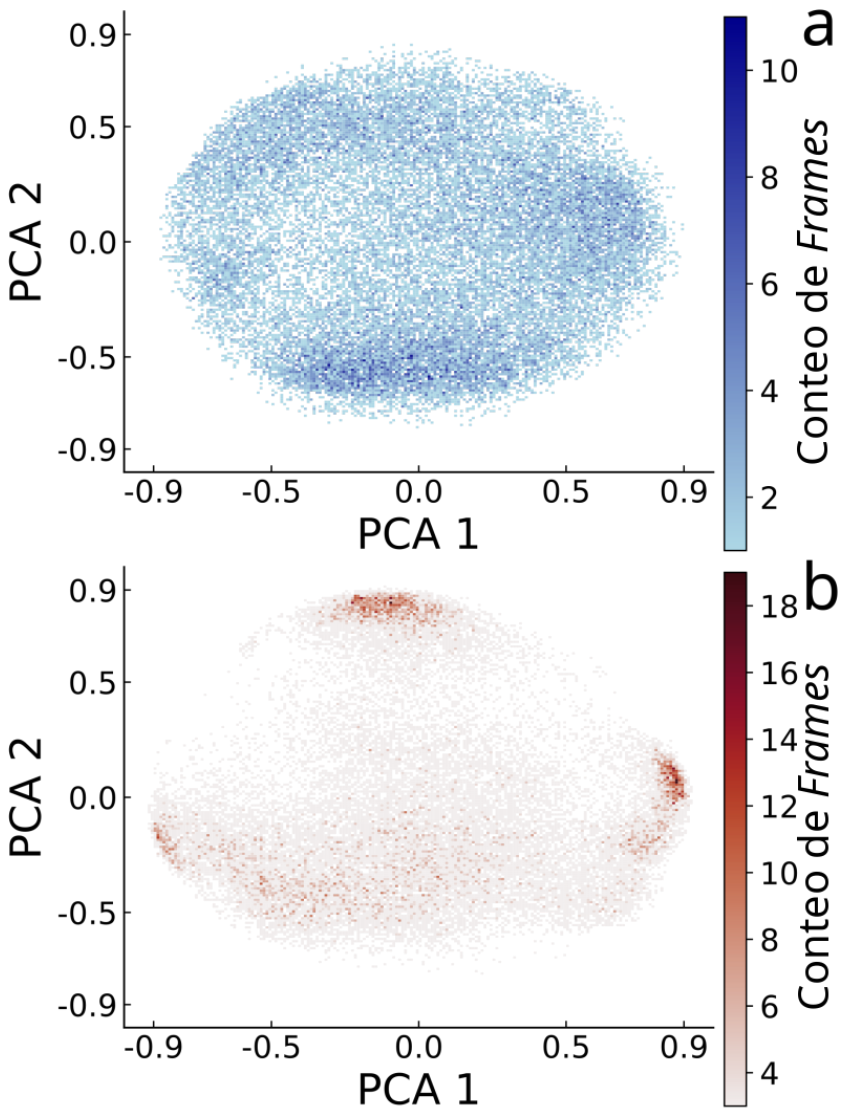

Figure 7.3: Mapa de calor de las proyecciones de las trayectorias de apo (**a**) y holo (**b**) Na-Far-1 sobre sus modos de PCA 1 y 2.

Las principales diferencias entre los confórmeros A, B y C se encuentran en la mentada puerta α4-α5 y la hélice α7. La hélice α7 del confórmero A está relativamente enderezada, lo que permite que la puerta α4-α5 se cierre. La hélice α7 de B tiene un plegamiento al lado de su extremo C terminal alrededor de ILE 104, que desplaza la puerta α4-α5. Esta deformación es aún más pronunciada en el confórmero C. Esta deformación es la razón principal de la disminución del volumen en los confórmeros B y C (ver Figura 7.4a). Por lo tanto, la distribución de los volúmenes de la cavidad interna que se muestra en la Figura 7.1c puede interpretarse como la contribución de tres conformaciones diferentes exploradas por holo-Na-FAR-1 durante la simulación de MD y sería esta la razón por la que su volumen no sigue una distribución normal.

La Figura 7.4b muestra la distribución de los volúmenes de la cavidad para cada uno de los confórmeros de holo-Na-FAR-1. Mientras que dos de los confórmeros holo-Na-FAR-1 (B y C) encierran cavidades internas relativamente pequeñas con volúmenes promedio

de 1130 ± 126 y 1211 ± 150 ^{3}, el otro confórmero (A) presenta una gran cavidad de 1568 ± 222 ^{3}. Estos resultados indican que holo-Na-FAR-1 presenta una notable plasticidad conformacional que impulsa una dinámica compleja de la cavidad interna. Los tres confórmeros identificados están en equilibrio dinámico conectados por cambios conformacionales que involucran el primer y segundo modo PCA. La Figura 7.4c muestra la evolución en el tiempo del volumen de la cavidad mostrando las diferentes contribuciones de cada uno de los tres confórmeros. Se pueden observar interconversiones reversibles entre ellos durante la simulación de MD. Estos resultados están en completo acuerdo con análisis previos de las estructuras de apo y holo de Na-FAR1 empleando espectroscopía de RMN (Rey-Burusco et al. 2015). Los espectros de RMN de holo-Na-FAR-1 en solución, como los de otras proteínas FAR previamente estudiados, se caracterizan por picos de señal amplios indicativos de conformaciones múltiples y / o intercambio conformacional. Sin embargo, apo-Na-FAR-1 dio buenos espectros de RMN de solución que permitieron determinar la estructura de apo-Na-FAR-1. En el mismo trabajo, se siguió el proceso de unión del ligando a través de RMN y mostró que la proteína exhibía un comportamiento de intercambio lento mediante la adición de 1, 2 y 3 equivalentes molares del ligando (oleato), lo que sugeriría que la proteína une tres ligandos con alta afinidad. La mayor plasticidad de la proteína después de la incorporación de una molécula de ligando se revela, en el presente trabajo, a lo largo de la simulación MD, en sus modos PCA primero y segundo.

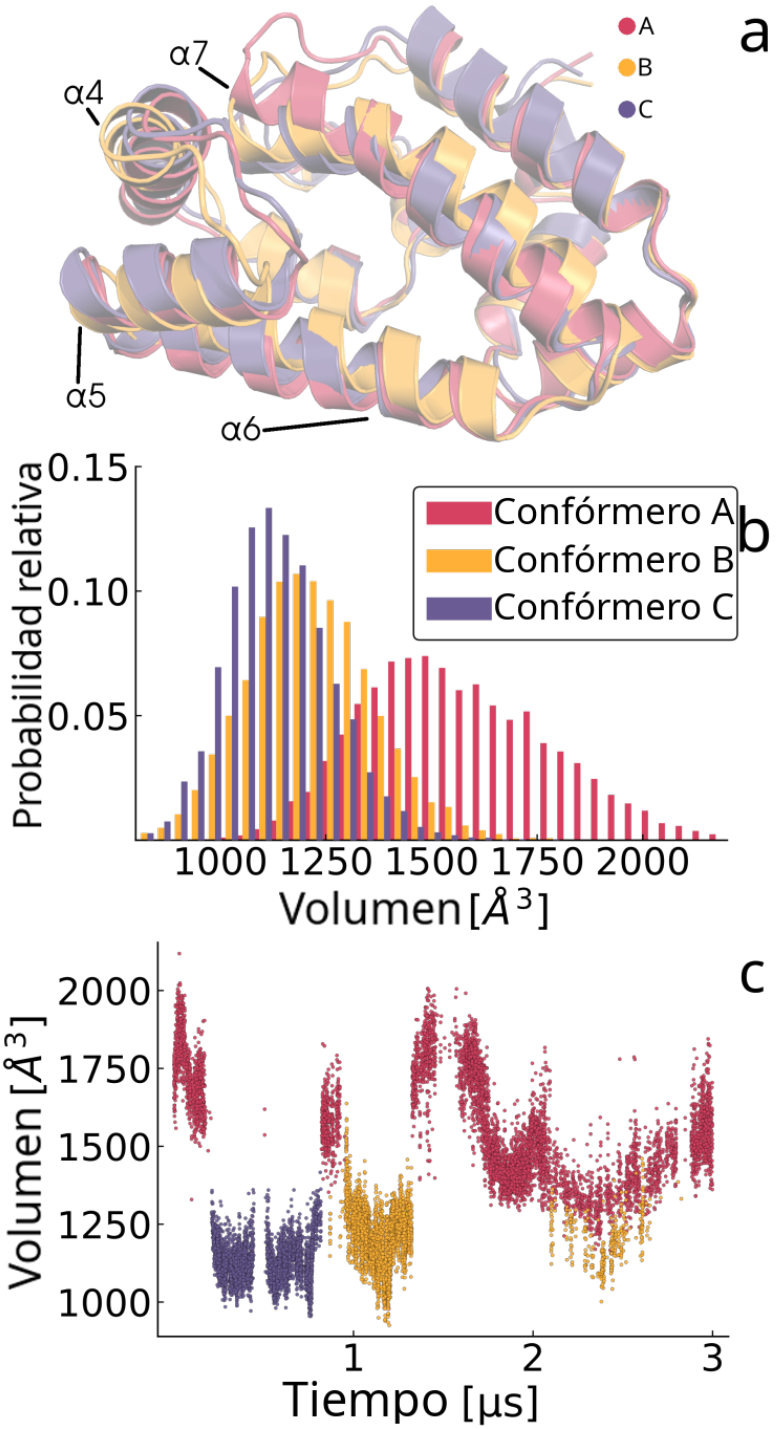

Figure 7.4: **a** Superposición de las estructuras promedio de los 3 confórmeros de holo-Na-FAR-1. **b** Histogramas de los volúmenes de la cavidad en la trayectoria, clasificados por el confórmero. **c** Volumen de la cavidad a lo largo de la trayectoria, clasificado por los confórmeros.

Observamos dos conformaciones distintivas de ligando representadas por las distorsiones estructurales en ambos sentidos de la dirección del primer modo PCA. Estos corresponden a las conformaciones plegadas y estiradas que se muestran en la Figura 7.5a. Como se puede ver en la Figura 7.5b, el ligando fluctúa entre ellas, siendo la conformación estirada la asociada a grandes volúmenes de cavidad, mientras que la plegada se observa dentro de volúmenes de cavidad más pequeños (Figura 7.5c). Es decir, lejos de estar fijo dentro de la cavidad, el ligando experimenta grandes cambios conformacionales asociados con

cambios en el volumen de la cavidad.

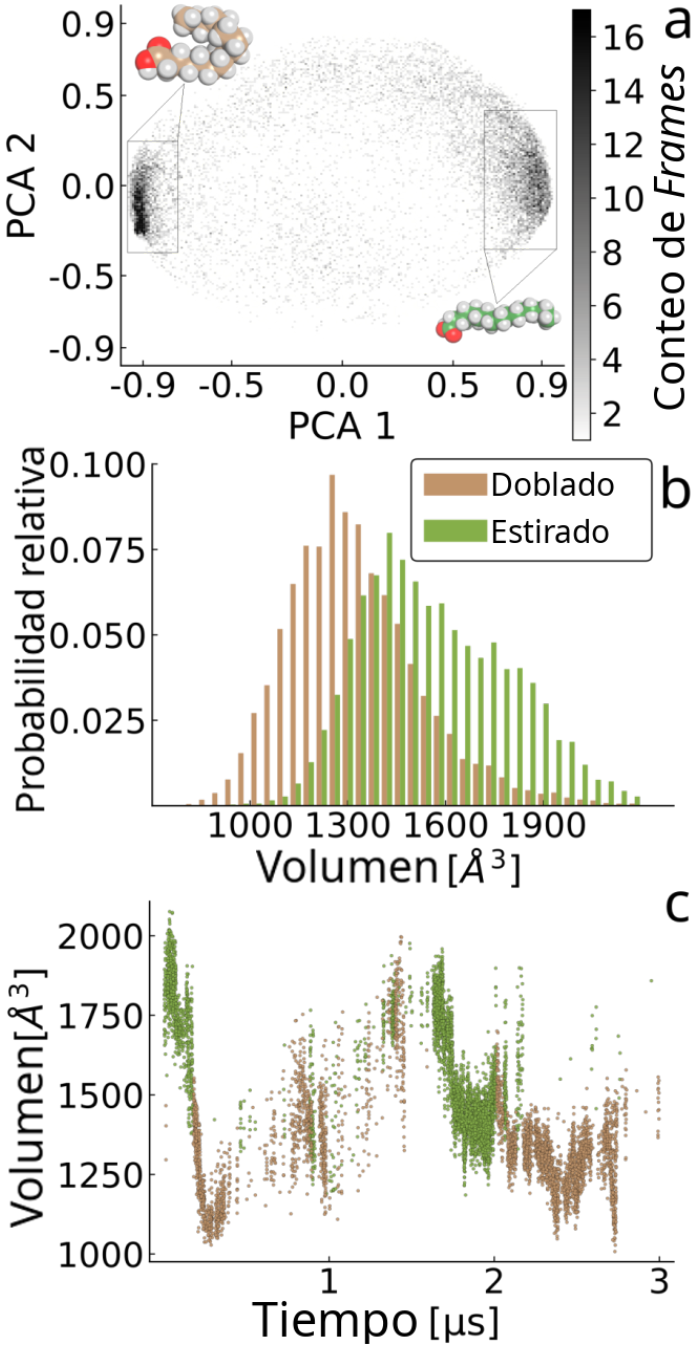

Figure 7.5: **a** Mapa de calor de las proyecciones de las trayectorias del palmitato sobre sus modos de PCA 1 y 2. **b** Histogramas de los volúmenes de la cavidad en la trayectoria, clasificados por la conformación del ligando que albergaba esta cavidad. **c** Volumen de la cavidad a lo largo de la trayectoria, clasificado por la conformación del ligando que albergaba.

La relación entre los diferentes confórmeros de holo-Na-FAR-1, con sus correspondientes cambios asociados en el volumen de la cavidad interna, y las diferentes conformaciones de ligando pueden analizarse representando la distribución de distancias entre los extremos de la molécula de palmitato, es decir, la distancia desde el átomo de C del grupo carboxilo al átomo de C del grupo metilo (ver Figura 7.6). Podemos observar que la conformación de palmitato estirada está asociada con el confórmero holo-Na-FAR-1 (A) con la cavidad

interna más grande y la hélice $\alpha 7$ enderezada para dejar espacio para el ligando, mientras que la conformación plegadaf está presente principalmente en los otros dos confórmeros (B y C). Dado que los tres confórmeros holo-NaFAR-1 están en equilibrio dinámico durante la simulación MD (ver Figura 7.4c), el ligando cambia su conformación de acuerdo con los cambios correspondientes en los tamaños de cavidad asociados con cada cambio conformacional de la proteína.

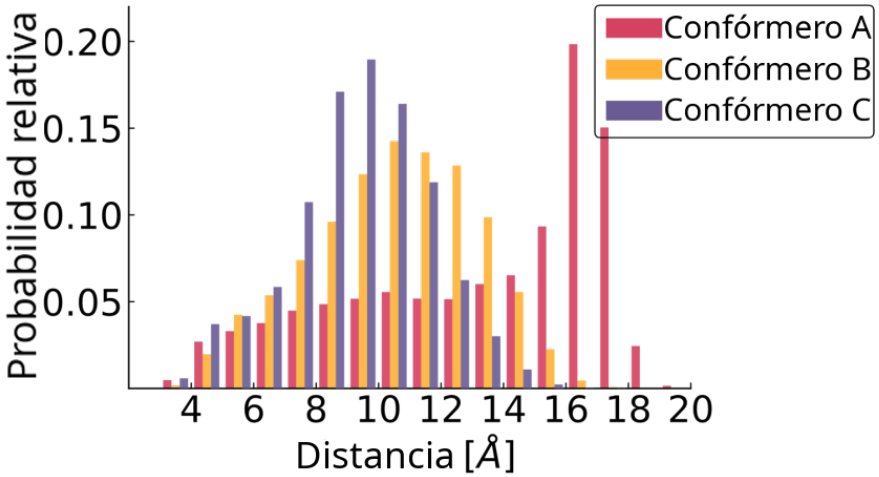

Figure 7.6: Histogramas de las distancias entre los extremos del palmitato clasificados por el confórmero que lo alberga.

Finalmente, se han realizado simulaciones MD con el ortólogo Ce-FAR-7 en su conformación apo. La Figura 7.7b muestra la estructura promedio obtenida de la simulación de MD equilibrada correspondiente. Ce-FAR-7 es un ortólogo de NaFAR-1 que, a pesar de su plegamiento general similar, presenta una cavidad de menor tamaño y distinta forma con respecto a Na-FAR1 (Rey-Burusco et al. 2015) Por lo tanto, una comparación de la flexibilidad relativa de las cavidades para Ce-FAR-7 y Na-FAR-1 pueden aclarar el origen de las diferencias en sus propiedades ligantes y biológicas.

El RMSD entre la apo-Na-FAR-1 promedio y la apo-CeFAR-7 es 2.77 Å. De acuerdo con apo-Na-FAR-1, la Figura 7.7a muestra que la distribución de su volumen de la cavidad interna puede estar asociada con fluctuaciones de proteínas alrededor de una conformación única caracterizada por un paisaje de energía libre con un pozo relativamente profundo. Estos resultados están de acuerdo con las observaciones hechas por Rey-Burusco et al. (Rey-Burusco et al. 2015) donde la cavidad estimada para Ce-FAR-7 reveló un tamaño mucho menor que para ambas formas de Na-FAR-1.

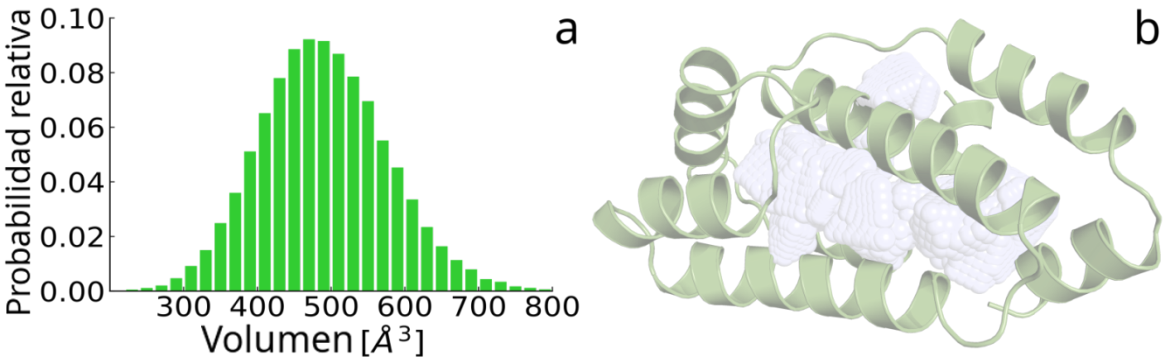

Figure 7.7: **a** Histograma del volumen a lo largo de las trayectorias de apo-CeFAR-7. **b** Estructura promedio de apo-CeFAR-7, obtenida de las simulaciones.

7.3.2 LBPs de hebras β

Mientras que las FARs exhiben plegamientos ricos en hélice α, la mayoría de las FABPs presentan un plegamiento típico de barril β que incluye un núcleo hidrofóbico pequeño y desplazado y una cavidad llena de moléculas de agua. Con el fin de comprender cómo los diferentes plegamientos impactan en las propiedades proteicas asociadas con el transporte de una variedad de ligandos con diferentes formas y tamaños, se han realizado simulaciones MD en la proteína de unión a ácidos grasos intestinales de rata (I-FABP) en sus formas holo y apo. El cambio conformacional de unión a ligando implica una distorsión estructural con un RMSD = 1.00 Å. La cavidad media interna del ligando es significativamente menor que Na-FAR-1, 605 ± 145 y 926 ± 85 ^{3} para apo-I-FABP y holo-I-FABP respectivamente (ver Figura 7.8). Podemos observar que la distribución de los volúmenes de la cavidad para apo-I-FABP puede estar asociada con la contribución de diferentes conformaciones exploradas durante la simulación MD. Por el contrario, holo-I-FABP parece presentar un estado rígido único. Estos resultados concuerdan con las mediciones de RMN previas, realizadas en L-FABP (Hodsdon & Cistola 1997) humano y en I-FABP (Cai et al. 2012) de rata que describen la unión del ligando como una transición de la estructura de la proteína desde un estado apo ligeramente más desordenado y flexible a uno holo más ordenado. Además, los experimentos de proteólisis limitada mostraron que la forma de holo era resistente al tratamiento mientras que la apo estaba completamente degradada (Arighi et al. 2003). Además, la comparación de RMSF obtenida durante nuestras simulaciones MD indica fluctuaciones más grandes para la apo-I-FABP que para la holo-I-FABP (ver Apéndice C). Esto está de acuerdo con los resultados de Matsuoka et al. (Matsuoka et al. 2015), donde los autores muestran que los valores calculados de RMSF fueron inferiores a 1.0 Å para casi todos los residuos, lo que indica que esta proteína es rígida en su forma holo. Este aumento de la movilidad en el estado apo pueden facilitar la entrada del ligando en la cavidad.

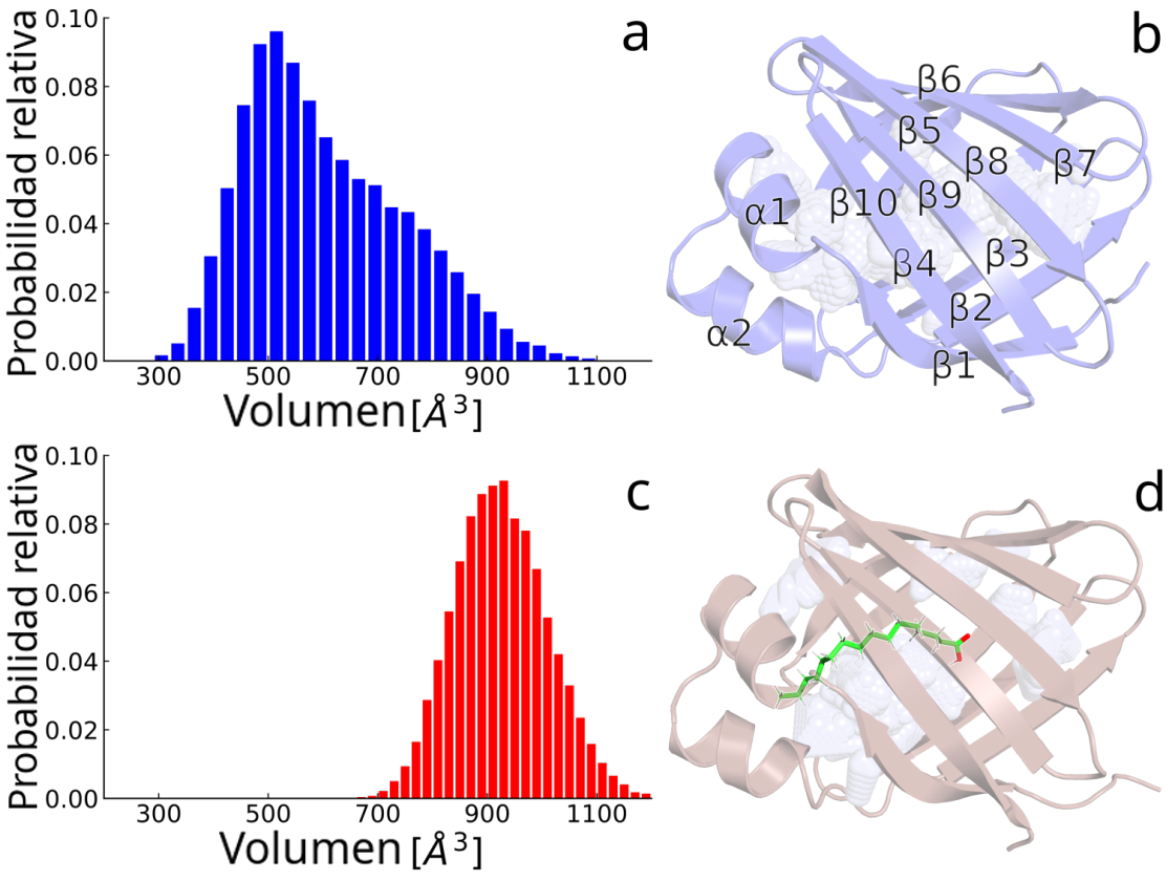

Figure 7.8: Estructuras promedio apo (**b**) y holo (**d**) de I-FABP, obtenidas de las simulaciones. Se indican los nombres de las SSE. Histogramas de los volúmenes a lo largo de las trayectorias de apo (**a**) y holo (**c**) I-FABP. Notar que el volumen de la holo sigue una distribución normal, a diferencia de la apo.

PCA permite la identificación de los diferentes confórmeros apo-I-FABP y su efecto correspondiente en el volumen de la cavidad interna (ver Figura 7.9a, b). Se han identificado cuatro confórmeros diferentes, asociados con diferentes combinaciones de distorsiones estructurales en las direcciones de los modos PCA primero y segundo, (ver también Figura 7.S3 y Figura 7.S4). Dos de ellos (A y B) están asociados con volúmenes de cavidad más pequeños que los otros dos (C y D). La Figura 7.9c muestra que apo-I-FABP experimenta múltiples cambios conformacionales a lo largo de la simulación, lo que indica una barrera de energía relativamente baja entre sus estados. Por el contrario, la proyección del conjunto de instantáneas MD de holo-I-FABP en su primer y segundo modo PCA no revela la existencia de múltiples confórmeros sino un estado rígido único (ver Apéndice C). Esto está de acuerdo con la distribución de sus volúmenes de cavidad, que se muestra en la Figura 7.8c, representada como una distribución normal que puede asociarse a fluctuaciones alrededor de un mínimo único en el espacio conformacional de la proteína. La unión del ligando parece cambiar el equilibrio conformacional de I-FABP a una conformación única con un pozo lo suficientemente profundo como para asegurar que una fracción significativa de moléculas de proteína quede atrapada fluctuando dentro.

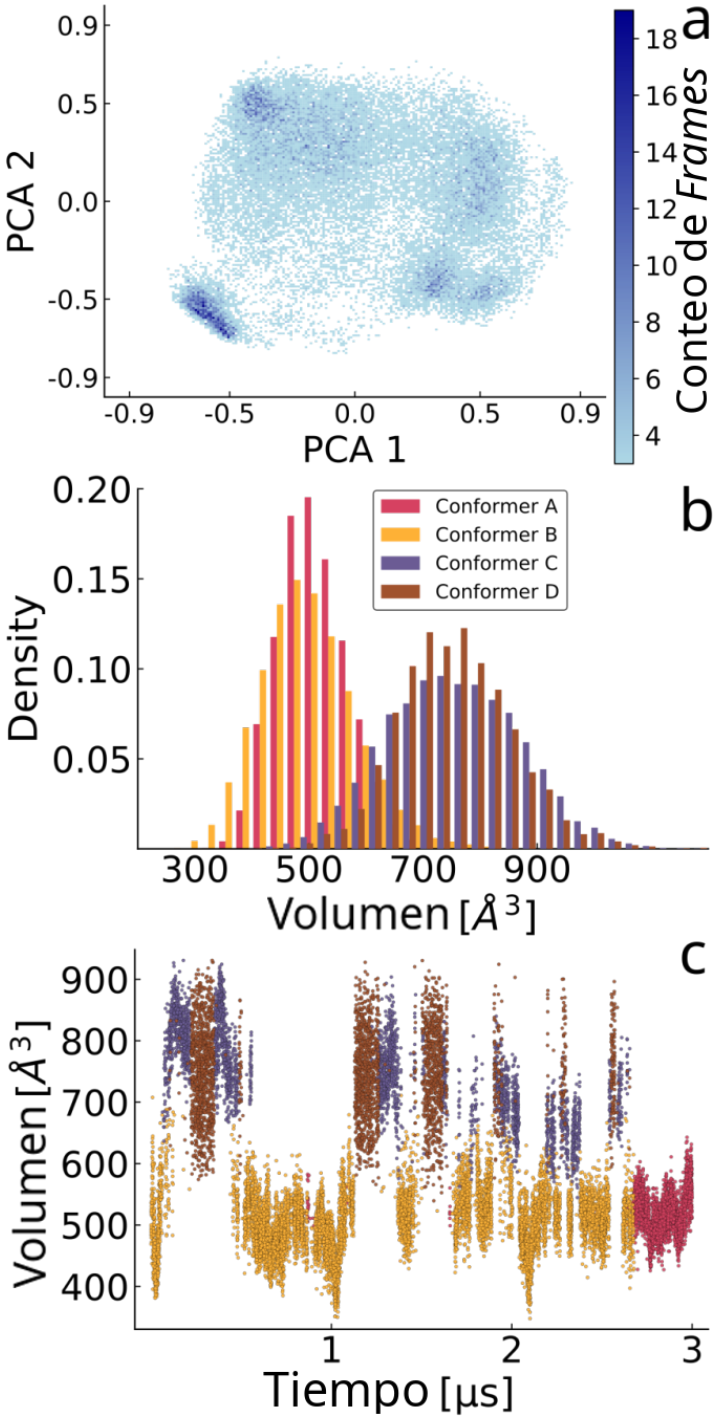

Figure 7.9: **a)** Mapa de calor de las proyecciones de las trayectorias de apo (**a**) y holo (**b**) I-FABP sobre sus modos de PCA 1 y 2. **b** Histogramas de los volúmenes de la cavidad en la trayectoria, clasificados por el confórmero de apo-I-FABP. **c** Volumen de la cavidad a lo largo de la trayectoria, clasificado por los confórmeros para holo-I-FABP.

7.3.3 Flexibilidad relativa de las cavidades

Las diferentes FARs y FABPs analizadas en este estudio han mostrado cavidades de ligando con diferentes formas y cuya dinámica está sujeta a la plasticidad proteica correspondiente. Para analizar qué plegamiento de LBP abarca una cavidad más flexible y, por lo tanto, una cavidad que pueda contribuir a una mayor multiplicidad de ligandos,

calculamos la variación de la energía potencial de la cavidad de cada LBP en la dirección del gradiente F(ver capítulo 4). Los resultados se muestran en la Figura 7.10a. Consideramos la amplitud del desplazamiento en la dirección de ∇V logrado con una energía igual a $kT = 0.593 kcal/mol$ como una medida de flexibilidad de la cavidad. Podemos ver que la cavidad interna de apo-I-FABP resulta la más flexible, seguida de holo y apo NaFAR-1. apo-Ce-FAR-7 presenta una cavidad relativamente más rígida. Además, los dos holo-I-FABP (1URE y 2IFB) encierran las cavidades más rígidas, lo que refuerza la idea de que los I-FABP de barril β siguen una estrategia de unión a ligando que implica un estado holo con libertad de movimiento restringida.

Mientras que tanto holo-Na-FAR-1 como apo-I-FABP abarcan cavidades con diferentes tamaños de acuerdo con la conformación transitoria de proteínas, la Figura 7.10b, c muestra el análisis de los confórmeros individuales correspondientes. Podemos observar que, en ambos casos, cada confórmero resulta relativamente más rígido que el promedio mostrado en la Figura 7.10a, lo que indica que sus contribuciones individuales introducen un componente adicional a la flexibilidad general de la cavidad. Además, los confórmeros holo-Na-FAR-1 son menos rígidos que el promedio (ver Figura 7.10b) en comparación con los confórmeros apo-I-FABP en relación con su promedio correspondiente (ver Figura 7.10c). Es decir, la flexibilidad de holo-Na-FAR-1 parece estar más uniformemente distribuida entre las poblaciones de confórmeros en equilibrio dinámico. Entonces, de la comparación entre las flexibilidades apo y holo de Na-FAR-1 e I-FABP se observa que la unión del ligando rigidiza fuertemente a la última, mientras que flexibiliza, levemente, a la primera. Estos resultados indican una propensión de Na-FAR-1 a unirse no solo a los ácidos grasos sino también a una gama más amplia de clases de lípidos como el retinol y los fosfolípidos. Esta característica está de acuerdo con los experimentos de fluorescencia previos realizados en NaFAR-1 y Ce-FAR-7

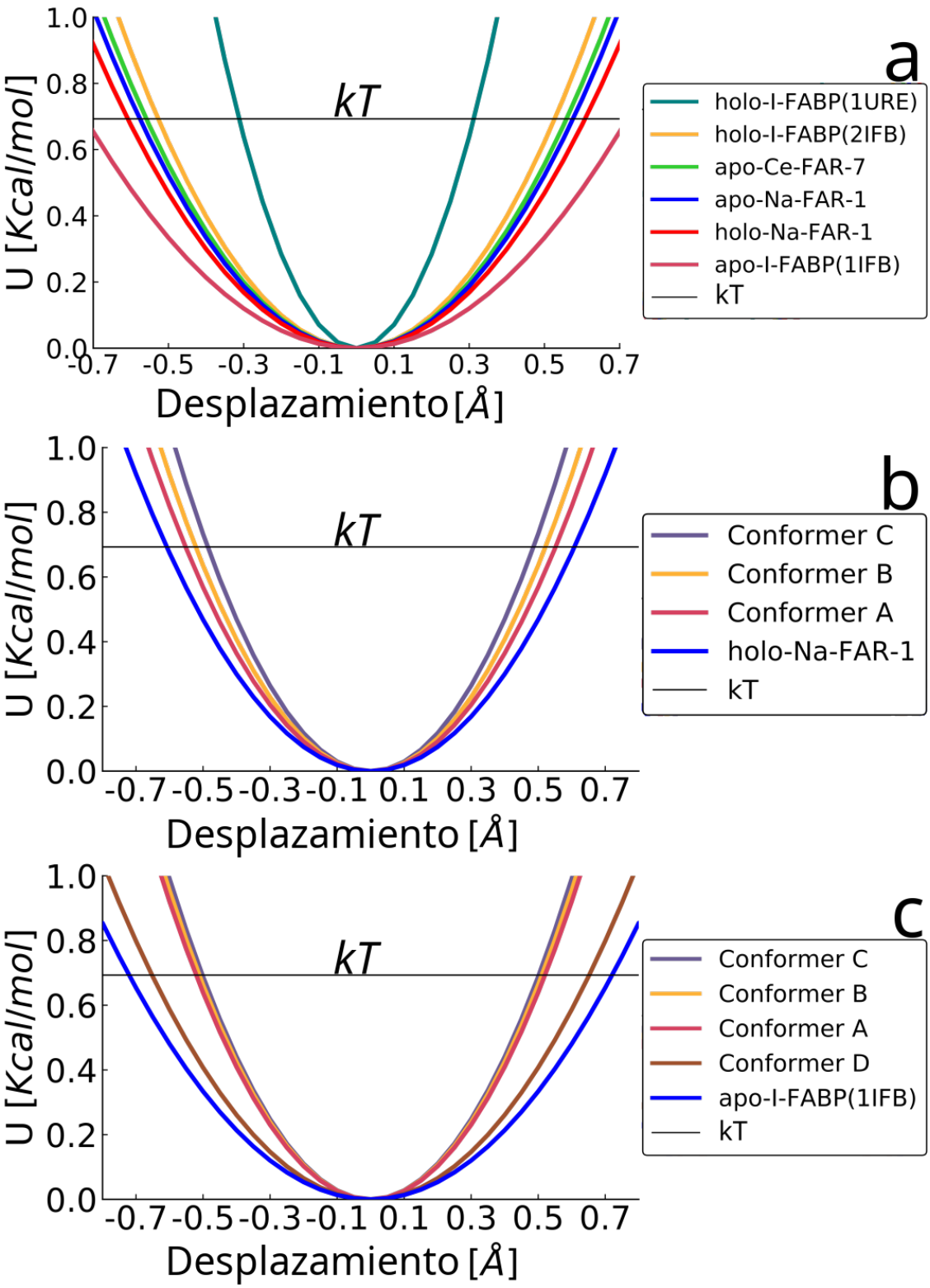

Figure 7.10: **a** Cambio de energía potencial en la dirección del ∇V para cada proteína. **b** confórmeros de holo-Na-FAR-1.**c** confórmeros de apo-I-FABP. La línea negra se agrega como referencia de la energía promedio a 298K (0.593 kcal/mol).

Chapter 8

Conclusiones

En la presente tesis hemos utilizado técnicas de Análisis de Modos Normales (NMA), Dinámica Molecular (MD) y Análisis de Componentes Principales (PCA) aplicadas al estudio de las relaciones entre las fluctuaciones de las proteínas y los cambios en sus cavidades. Esto ha requerido el desarrollo tanto de nuevas técnicas de análisis, como de un software capaz de cuantficar la flexibilidad de cavidades y analizar el cambio de su volumen ante deformaciones estructurales de la proteína en direcciones predefinidas, como ser, en la dirección de sus modos normales o modos de PCA de relevancia biológica conocida. Nuestro estudio proporciona aspectos dinámicos complementarios para una evaluación directa de las fluctuaciones del volumen y la estructura de las proteínas, obtenidas durante las simulaciones MD. Para este propósito, utilizamos una combinación de algoritmos para cálculos de volumen de cavidad lo suficientemente robustos para diferenciaciones numéricas. En todos nuestros casos de estudio, el análisis de las contribuciones de los modos PCA al vector de gradiente de volumen (∇V) revela contribuciones importantes de los modos de baja frecuencia y contribuciones menores de los modos de media frecuencia.

Considerando la variación de la energía potencial de una proteína en la dirección de ∇V como medida de flexibilidad de la cavidad, observamos que las proteínas cuyas fluctuaciones colectivas de baja frecuencia son las que más participan en los cambios del volumen de la cavidad exhiben cavidades más flexibles. Si bien los modos de baja frecuencia representan las principales contribuciones al ∇V, no representan el principal costo energético en dicha dirección.

Este desarrollo metodológico nos ha permitido encarar el estudio de las características dinámicas características de los distintos confórmeros del EGFR. Hemos realizado un extenso análisis comparativo de las características de dinámica global compartidas por los confórmeros existentes de la EGFR quinasa. Hemos identificado direcciones claras de movimientos que se pueden usar como huellas para diferenciar los confórmeros activos e inactivos. Se ha aplicado un nuevo procedimiento que puede usarse para conjuntos de confórmeros en otras proteínas. El método permite la comparación de patrones de vi-

braciones que rescatan direcciones representativas comunes de movimientos compartidos entre el conjunto. Se han identificado y caracterizado dos direcciones representativas de movimientos entre conformaciones activas en EGFR quinasa. Estos movimientos representan huellas de los confórmeros activos que se pueden agregar a las características estructurales reportadas previamente. Su conservación entre el conjunto completo de confórmeros activos nos informa sobre la existencia de mecanismos funcionales mínimos comunes dentro de ellos. Los confórmeros inactivos han mostrado una tendencia general a interrumpir el movimiento colectivo de los residuos involucrados en estas vibraciones. Por lo tanto, se espera que los aspectos funcionales de la EGFR quinasa impliquen movimientos coordinados entre residuos que se reflejan principalmente en los dos modos normales de baja frecuencia de los confórmeros activos. Estos modos permiten que los confórmeros activos alcancen conformaciones más extendidas que implican una separación tipo bisagra de los lóbulos N y C, un requisito previo para lograr una conformación inactiva relativamente más estable. Además, utilizando nuestros métodos ya desarrollados para el estudio de cavidades, observamos que estos modos representativos conducen a cambios más grandes en los volúmenes de la cavidad, en comparación con los modos correspondientes de confórmeros inactivos.

Por otro lado, hemos estudiado la relación entre las fluctuaciones de la estructura de la proteína y cambios en las cavidades de las LBPs de parásitos helmintos. Las LBP de helmintos a menudo son inmunodominantes en la infección. Las FARs se encuentran comúnmente en las secreciones de nematodos parásitos, indicando su posible rol en el parasitismo. Los parásitos necesitan adquirir nutrientes de sus anfitriones y a la vez defenderse contra su respuesta inmune. En este sentido, se presume que interfieren secuestrando los lípidos de señalización producidos por el huésped. Por lo tanto, una gran multiplicidad de ligandos de FAR ayudaría tanto en la adquisición de lípidos como en el secuestro.

Hemos explorado la relación fluctuaciones de la proteína - cambios de su cavidad para diferentes FARs ricas en hélice α y FABP de barril β utilizando simulaciones MD equilibradas de estructuras apo y holo. Encontramos una cavidad Na-FAR-1 que cuando está ocupada es significativamente flexible, lo que puede explicar la mayor multiplicidad de ligandos observada de FARs de hélices α con respecto a las FABPs de barril β. La comparación de la flexibilidad relativa de las cavidades de unión a ligando de Ce-FAR-7 y Na-FAR-1 revela cómo un plegamiento similar puede encerrar cavidades internas con diferencias significativas en su flexibilidad y dinámica. Estas diferencias pueden explicar las diferencias en su multiplicidad de ligandos y, por lo tanto, su función biológica. Además, se han observado diferencias en las capacidades de unión del ligando entre dos isoformas de la misma especie. Hemos informado dos estrategias diferentes de unión a ligando. En particular, holo-Na-FAR-1 presenta una notable plasticidad conformacional que impulsa una dinámica compleja de la cavidad interna involucrando diferentes estados. El tamaño de la cavidad es significativamente afectado por los cambios conformacionales de la pro-

teína. Además, el ligando también cambia su conformación de acuerdo con estos cambios conformacionales. Es decir, lejos de estar fijo dentro de la cavidad, el ligando experimenta grandes cambios conformacionales entre una conformación plegada y estirada. La conformación del ligando cambia de acuerdo con el tamaño de la cavidad dictada por la conformación transitoria de la proteína. Por el contrario, la unión de ligando en I-FABP parece cambiar el equilibrio conformacional a una conformación única. De esta manera, las FARs de hélices α y las FABPs de barril β parecen seguir dos estrategias diferentes para la unión de ligandos. Las FARs involucran un estado holo con alta plasticidad; experimentan cambios conformacionales que impactan significativamente en el volumen de la cavidad y en las conformaciones de ligando que albergan. Por otro lado, las FABPs experimentan una transición inversa de desorden-orden modulada por ligando que conduce a un estado holo de movimiento restricto. Esta información podría arrojar luz sobre las razones biológicas de la existencia de diferentes tipos de LBP en el mismo organismo.

El trabajo de esta tesis representa un aporte al estudio de las conexiones entre las fluctuaciones de proteínas y los cambios en el volumen de la cavidad. La flexibilidad de las cavidades proteicas puede afectar aspectos funcionales como las afinidades a ciertos ligandos y hasta la aparición o desaparición de la cavidad. Las mutaciones que reducen el costo de la energía en la dirección de los cambios máximos en los volúmenes de la cavidad deberían aumentar la flexibilidad de la cavidad y, por lo tanto, pueden conducir a aumentar la afinidad por diferentes sustratos.

Por último, estos métodos se han facilitado a la comunidad científica por medio de un software desarrollado en C++, curado, documentado y con un sitio oficial de soporte ANA. Este programa permite obtener el gradiente del volumen de la cavidad o calcular su flexibilidad, ejecutando una única línea de comando, facilitando y aceleranto el estudio de las cavidades de proteínas.

Apéndice A: Protocolos de equilibración utilizados

Table 8.1: El último paso, el realizado a volumen constante, totaliza 5ns de preparación del sistema, entre calentamiento y equilibración.

Intervalo[ps]	Restricción [kcal/(mol · Å2)]
100	50.0
100	40.0
100	30.0
100	30.0
100	20.0
100	15.0
100	10.0
100	9.0
100	8.0
100	7.0
100	6.0
100	5.0
100	4.0
100	3.0
100	2.0
100	1.5
100	1.0
100	0.8
100	0.6
100	0.4
100	0.3
100	0.2
100	0.17
100	0.14
100	0.11
100	0.08

Intervalo[ps]	Restricción [kcal/(mol · Å2)]
100	0.05
100	0.02
100	0.0
1960	0.0

Table 8.2: El último paso, el realizado a volumen constante, totaliza 20ns de preparación del sistema, entre calentamiento y equilibración.

Intervalo[ps]	Restricción [kcal/(mol · Å2)]
100	25.0
100	20.0
100	15.0
100	10.0
100	9.5
100	9.0
100	8.5
100	8.0
100	7.5
100	7.0
100	6.5
100	6.0
100	5.5
100	5.0
100	4.5
100	4.0
100	3.5
100	3.0
100	2.5
100	2.0
100	1.5
100	1.0
100	0.9
100	0.8
100	0.7
100	0.6
100	0.5
100	0.45
100	0.4
100	0.35
100	0.30
100	0.25
100	0.20
100	0.18
100	0.16
100	0.15
100	0.14

Intervalo[ps]	Restricción [kcal/(mol · Å2)]
100	0.13
100	0.12
100	0.11
100	0.10
100	0.09
100	0.08
100	0.07
100	0.06
100	0.05
100	0.045
100	0.04
100	0.035
100	0.03
100	0.025
100	0.020
100	0.015
100	0.014
100	0.013
100	0.012
100	0.011
100	0.010
100	0.009
100	0.008
100	0.007
100	0.006
100	0.005
100	0.0045
100	0.004
100	0.0035
100	0.003
100	0.0025
100	0.0020
100	0.0015
100	0.0010
100	0.0005
100	0.0000
100	0.0000
12200	0.0000

Apéndice B: Publicaciones

1. Hasenahuer, Marcia Anahi and Barletta, German Patricio and Fernandez-Alberti, Sebastián and Parisi, Gustavo and Fornasari, María Silvina 2017. Pockets as structural descriptors of EGFR kinase conformations. PLoS ONE.

2. Barletta, German P. and Hasenahuer, Marcia Anahí and Fornasari, María Silvina and Parisi, Gustavo and Fernández-Alberti, Sebastián, 2018. Dynamics fingerprints of active conformers of epidermal growth factor receptor kinase. Journal of Computational Chemistry.

3. Barletta, German P. & Fernandez-Alberti, S., 2018. Protein Fluctuations and Cavity Changes Relationship. Journal of Chemical Theory and Computation.

4. Barletta German P., Matias Barletta and Sebastian Fernandez-Alberti. 2019 ANA accurate cavity definition and fast tracking. **Submitted**.

5. Barletta, German P. and Franchini, Gisela and Corsico, Betina and Fernandez-Alberti, Sebastian. 2019 Fatty acid and retinol-binding protein: Unusual protein conformational and cavity changes dictated by ligand fluctuations. **In press**.

Apéndice C: Información Suplementaria de las publicaciones incluídas

8.1 Capítulo 6

8.1.1 PDBs del conjunto de estructuras quinasas

Inactivas	Activas
2GS7_A	1M14_A
2RGP_A	1M17_A
3BEL_A	2EB2_A
3GOP_A	2GS2_A
3GT8_A	2GS6_A
3IKA_A	2ITN_A
3W2R_A	2ITP_A
3W2S_A	2ITU_A
3W32_A	2ITX_A
3W33_A	2ITZ_A
4I1Z_A	3IKA_A
4I22_A	3UG1_A
4I24_A	3VJN_A
4ZJV_A	4G5J_A
5CNN_A	4I23_A
	4LI5_A
	4LQM_A
	4R3P_A
	4R5S_A
	4RJ4_A
	4ZAU_A

Inactivas	Activas
	5C8K_A
	5CAO_A
	5CAP_A
	5CAV_A
	5CZH_A

8.1.2 Residuos de la cavidad activa

Residuo	Residuo
LEU 16	ARG 139
GLY 17	ASN 140
SER 18	LEU 142
GLY 19	THR 152
ALA 20	ASP 153
PHE 21	PHE 154
GLY 22	GLY 155
VAL 24	LEU 156
ALA 41	GLY 171
LYS 43	GLY 172
GLU 60	LYS 173
MET 64	VAL 174
CYS 73	PRO 175
LEU 86	ILE 176
THR 88	LYS 177
GLN 89	TRP 178
LEU 90	MET 179
MET 91	SER 183
PRO 92	ILE 184
GLY 94	ARG 187
CYS 95	TYR 189
LEU 97	GLU 204
ASP 98	LYS 211
ARG 101	PRO 212
ARG 134	ALA 218
ASP 135	SER 219
ALA 137	

8.2 Capítulo 7

8.2.1 Tablas de residuos de la cavidad

8.2.1.1 I-FABP apo (1URE)

Table 8.5: Lista de residuos que recubren la cavidad principal de unión a ligando de 1URE. Los residuos se numeran según su orden en el archivo PDB correspondiente.

Residuo	Residuo	Residuo	Residuo
LYS 7	ASN 45	THR 76	ASN 111
TYR 14	LYS 46	GLU 77	GLU 112
LYS 16	SER 53	LEU 78	LEU 113
MET 18	PHE 55	THR 79	THR 118
GLU 19	ASN 57	GLY 80	TYR 119
ILE 23	ILE 58	THR 81	GLU 120
ASN 24	VAL 60	TRP 82	GLY 121
LYS 27	VAL 61	MET 84	GLU 123
ARG 28	PHE 62	LYS 88	
ALA 32	GLU 63	LYS 94	
HIE 33	LEU 64	ASN 98	
LYS 37	VAL 66	GLU 107	
THR 39	PHE 68	ILE 108	
THR 41	ASP 74	SER 109	
GLY 44	GLY 75	GLY 110	

8.2.1.2 I-FABP apo (1IFB)

Table 8.6: Lista de residuos que recubren la cavidad principal de unión a ligando de 1IFB. Los residuos se numeran según su orden en el archivo PDB correspondiente.

Residuo	Residuo	Residuo	Residuo
ASN 13	SER 53	LEU 78	ILE 108
TYR 14	PHE 55	THR 79	SER 109
GLU 19	ARG 56	GLY 80	GLY 110
ILE 23	ASN 57	THR 81	GLU 112
ASN 24	ILE 58	TRP 82	LEU 113
LYS 27	PHE 62	THR 83	THR 118
ARG 28	GLU 63	MET 84	TYR 119
LYS 37	LEU 64	GLY 86	GLU 120

Residuo	Residuo	Residuo	Residuo
THR 39	VAL 66	ASN 87	GLY 121
THR 41	PHE 68	LYS 88	GLU 123
GLU 43	TYR 70	LYS 94	
GLY 44	ASP 74	ASP 97	
ASN 45	GLY 75	ASN 98	
LYS 46	THR 76	GLU 101	
SER 52	GLU 77	GLU 107	

8.2.1.3 I-FABP holo (2IFB)

Table 8.7: Lista de residuos que recubren la cavidad principal de unión a ligando de 2IFB. Los residuos se numeran según su orden en el archivo PDB correspondiente.

Residuo	Residuo	Residuo	Residuo
ASN 13	SER 53	THR 79	GLU 107
TYR 14	PHE 55	GLY 80	ILE 108
GLU 19	ARG 56	TRP 82	SER 109
ILE 23	ASN 57	THR 83	GLY 110
ASN 24	ILE 58	MET 84	GLU 112
LYS 27	PHE 62	GLU 85	LEU 113
ARG 28	GLU 63	GLY 86	THR 118
LYS 37	LEU 64	ASN 87	TYR 119
THR 39	VAL 66	LYS 88	GLU 120
THR 41	PHE 68	LYS 94	GLY 121
GLU 43	ASP 74	ARG 95	GLU 123
GLY 44	GLY 75	ASP 97	
ASN 45	THR 76	ASN 98	
LYS 46	GLU 77	LYS 100	
SER 52	LEU 78	GLU 101	

8.2.1.4 Na-FAR-1 apo (4UET)

Table 8.8: Lista de residuos que recubren la cavidad principal de unión a ligando de 4UET. Los residuos se numeran según su orden en el archivo PDB correspondiente.

Residuo	Residuo	Residuo	Residuo
TYR 10	LEU 33	ARG 93	SER 125
ASP 12	GLU 35	ILE 95	ASP 126

Residuo	Residuo	Residuo	Residuo
PRO 15	TYR 42	ALA 97	ALA 127
PRO 16	LYS 54	ARG 98	PHE 132
ARG 19	GLU 58	TYR 100	GLN 135
ASP 20	SER 72	THR 101	LEU 139
LEU 22	ALA 75	GLY 102	LYS 141
GLN 23	ALA 76	GLU 104	GLU 144
ASN 24	LEU 77	PRO 105	
LEU 25	GLU 80	THR 106	
SER 26	ALA 81	ASP 108	
ASP 27	ALA 85	ASP 109	
ASP 29	GLU 86	LEU 110	
THR 31	LYS 87	TYR 121	
VAL 32	ALA 92	LYS 122	

8.2.1.5 Na-FAR-1 holo (4XCP)

Residuo	Residuo	Residuo	Residuo
TYR 10	VAL 32	LYS 87	TYR 121
ASP 12	LEU 33	ALA 92	LYS 122
LEU 13	GLU 35	ARG 93	SER 125
PRO 15	TYR 42	GLY 94	ASP 126
PRO 16	LYS 54	ILE 95	ASP 131
ARG 19	GLU 58	ALA 97	PHE 132
ASP 20	SER 72	ARG 98	GLN 135
LEU 22	ALA 75	TYR 100	LEU 139
GLN 23	ALA 76	GLU 104	LYS 141
ASN 24	LEU 77	PRO 105	GLU 144
LEU 25	PRO 79	THR 106	
SER 26	GLU 80	ASP 108	
ASP 27	ALA 81	ASP 109	
ASP 29	ALA 85	LEU 110	
THR 31	GLU 86	LYS 119	

8.2.1.6 Ce-FAR-7 apo (2W9Y)

Table 8.10: Lista de residuos que recubren la cavidad principal de unión a ligando de 2W9Y. Los residuos se numeran según su orden en el archivo PDB correspondiente.

Residuo	Residuo	Residuo	Residuo
PRO 7	GLU 29	GLY 70	THR 101
GLU 8	LYS 30	ASN 71	VAL 102
CYS 9	PRO 31	LYS 73	GLY 103
ASN 12	LEU 33	LEU 75	LYS 104
PRO 15	GLU 35	PRO 80	ILE 106
GLN 18	PHE 37	ILE 90	HID 115
LEU 19	GLN 38	MET 92	PHE 118
GLU 20	CYS 42	VAL 93	GLN 119
PHE 21	PHE 43	THR 94	ALA 128
SER 22	MET 49	THR 95	
SER 23	GLU 52	THR 96	
SER 24	LYS 55	LEU 97	
ILE 25	HIE 57	CYS 98	
ALA 27	PRO 58	SER 99	
ASP 28	LEU 60	LEU 100	

8.2.2 Fluctuaciones cuadráticas medias de raíz (RMSF) durante nuestras simulaciones MD

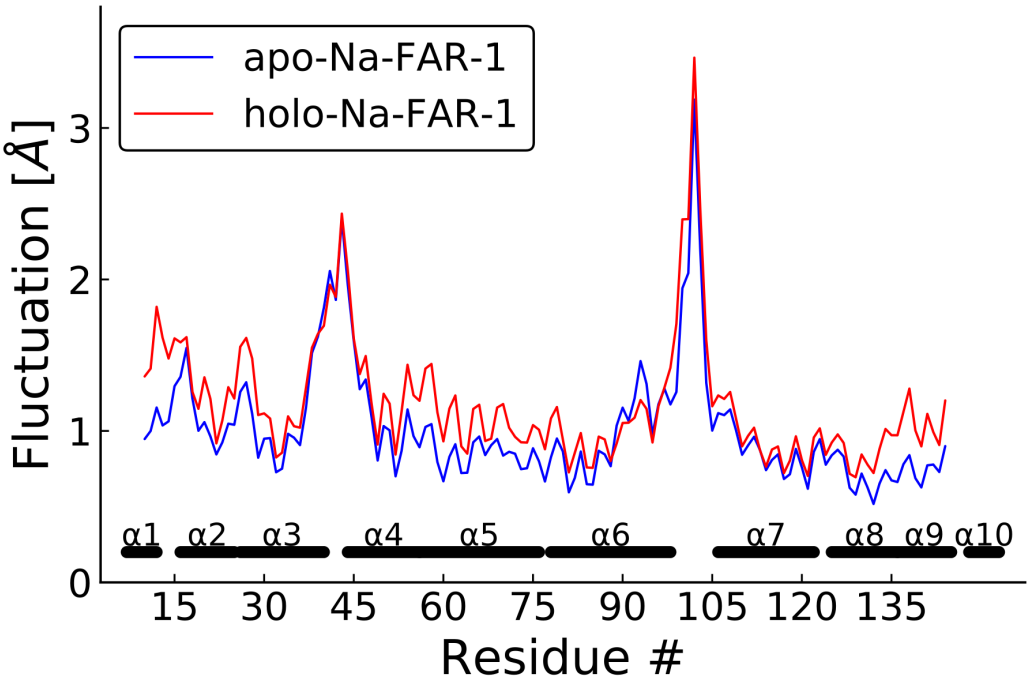

Figure 8.1: **a)**

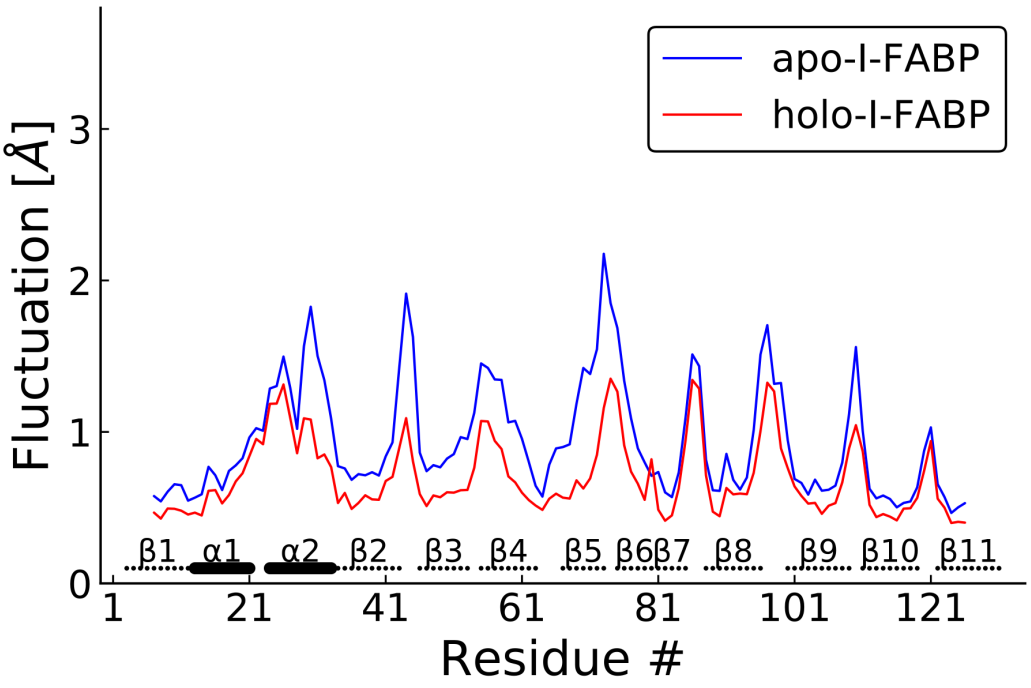

Figure 8.2: **a)**

8.2.3 Histograma de la proyección del conjunto de instantáneas MD de holo-I-FABP en su tercer modo PCA

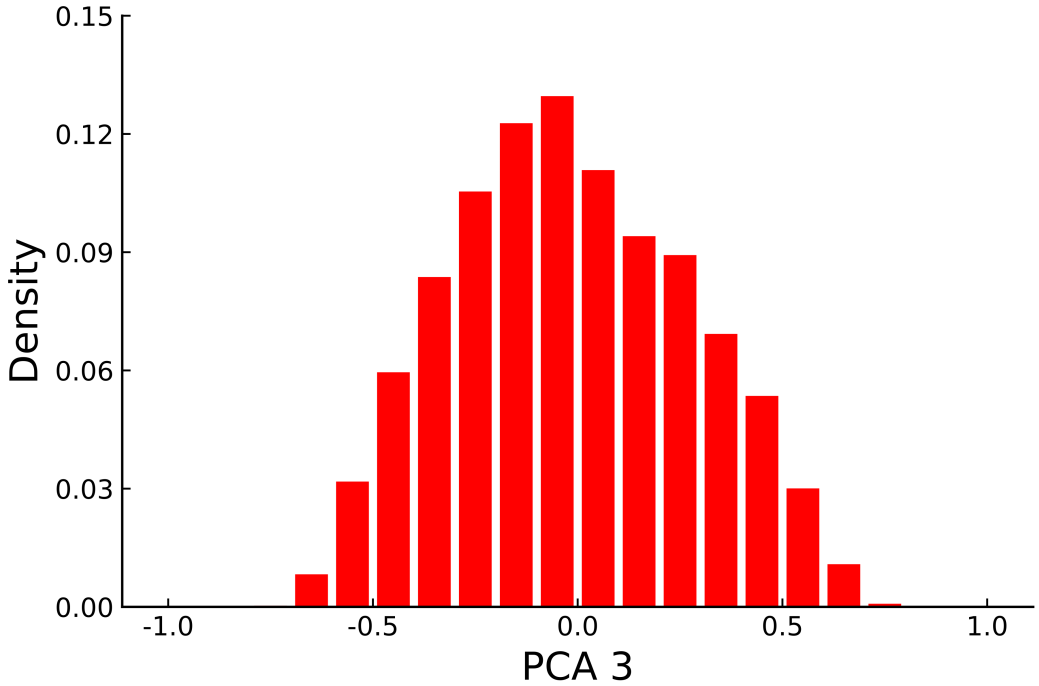

Figure 8.3: **a)**

8.2.4 Primero (rojo) y segundo (azul) modos de PCA de (a) apo- y (b) holo-I-FABP

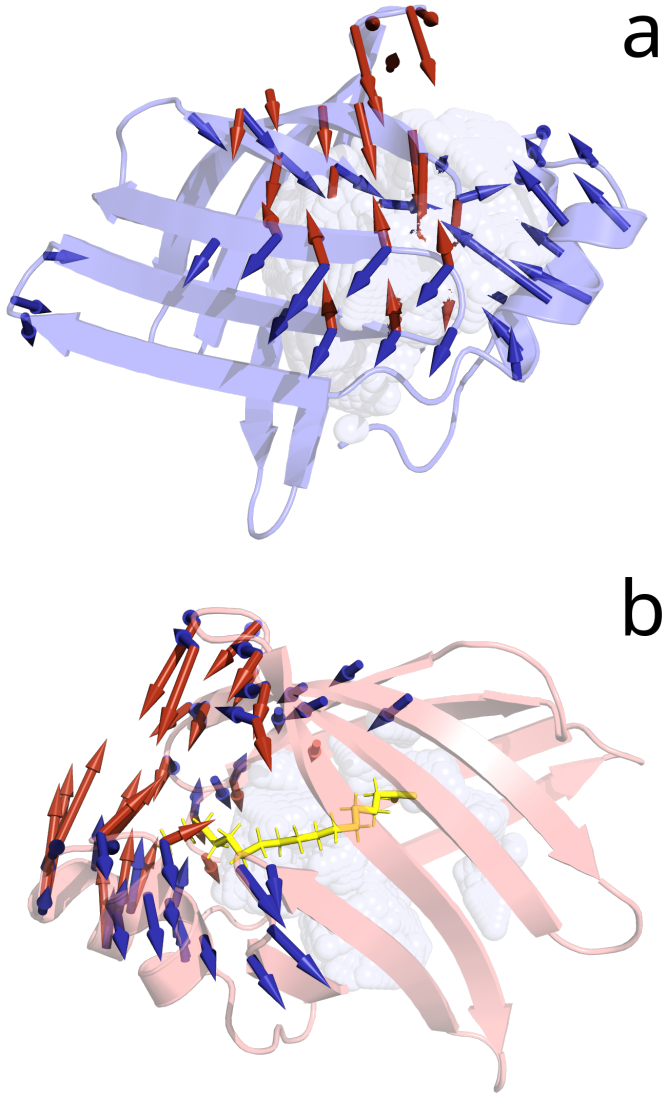

Figure 8.4: **a)**

8.2.5 Superposición de los cuatro conformadores (A, B, C y D) de apo-I-FABP

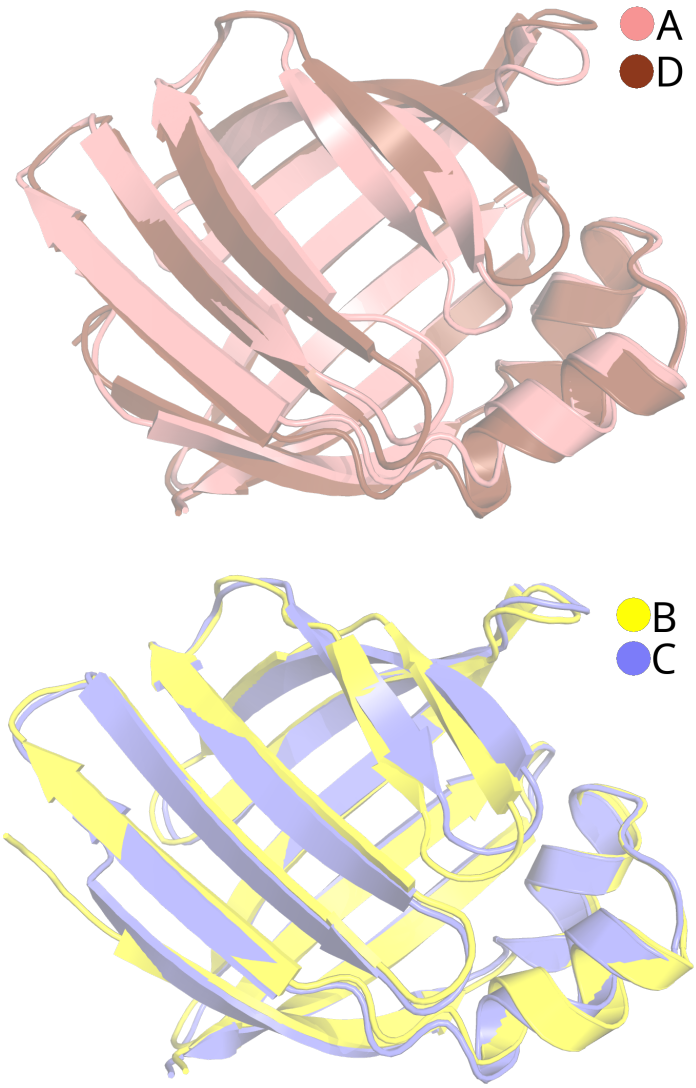

Figure 8.5: **a)**

8.2.6 Gráficos de densidad de contorno de la proyección del conjunto de instantáneas MD de holo-I-FABP en sus modos PCA primero y segundo correspondientes

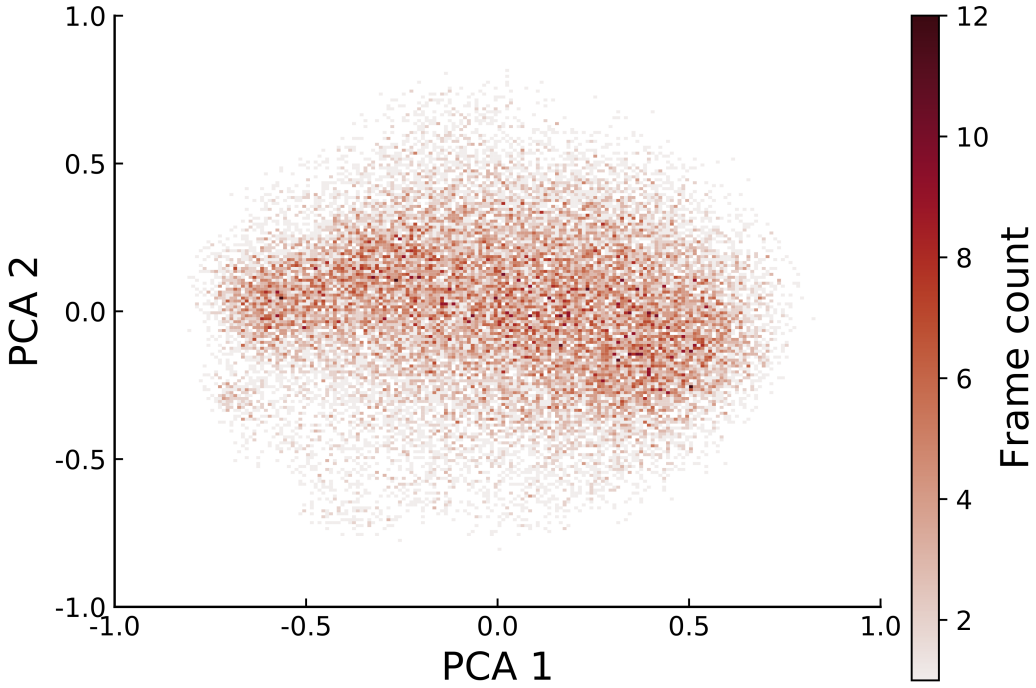

Figure 8.6: **a)**

Referencias

Albaugh, A. et al., 2016. Advanced Potential Energy Surfaces for Molecular Simulation. *Journal of Physical Chemistry B*, 120(37), pp.9811–9832.

Alder, B.J. & Wainwright, T.E., 1957. Phase transition for a hard sphere system. *The Journal of Chemical Physics*, 27(5), pp.1208–1209.

Allen, M.P. & Tildesley, D.J., 2017. Computer simulation of liquids: Second edition. *Computer Simulation of Liquids: Second Edition*, pp.1–626.

Amadei, A., Linssen, A.B.M. & Berendsen, H.J.C., 1993. Essential dynamics of proteins. *Proteins: Structure, Function and Genetics*, 17(4), pp.412–425.

Anfinsen, C.B., 1973. Principles that govern the folding of protein chains. *Science*, 181(4096), pp.223–230.

Anfinsen, C.B. et al., 1961. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proceedings of the National Academy of Sciences of the United States of America*, 47(9), pp.1309–1314.

Ansari, A. et al., 1985. Protein states and proteinquakes (equilibrium fluctuations/functionaily important motions/hierarchy of states/glass-like structure). *Biophysics*, 82(August), pp.5000–5004. Available at: http://www.pnas.org/content/82/15/5000.long.

Arighi, C.N., Rossi, J.P.F. & Delfino, J.M., 2003. Temperature-induced conformational switch in intestinal fatty acid binding protein (IFABP) revealing an alternative mode for ligand binding. *Biochemistry*, 42(24), pp.7539–7551.

Arteaga, C.L. & Engelman, J.A., 2014. ERBB Receptors: From Oncogene Discovery to Basic Science to Mechanism-Based Cancer Therapeutics. *Cancer Cell*, 25(3), pp.282–303. Available at: https://linkinghub.elsevier.com/retrieve/pii/S1535610814000865.

Atilgan, A.R. et al., 2001. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical journal*, 80(1), pp.505–515. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1301

Bahar, I., Atilgan, A.R. & Erman, B., 1997. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding and Design*, 2(3), pp.173–181.

Bahar, I. et al., 1998. Vibrational dynamics of folded proteins: Significance of slow and fast motions in relation to function and stability. *Physical Review Letters*, 80(12), pp.2733–2736.

Barber, C.B., Dobkin, D.P. & Huhdanpaa, H., 1996. The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software*, 22(4), pp.469–483.

Barletta, G.P. et al., 2018. Dynamics fingerprints of active conformers of epidermal growth factor receptor

kinase. *Journal of Computational Chemistry*, pp.1–9. Available at: http://doi.wiley.com/10.1002/jcc.25590.

Barril, X. & Fradera, X., 2006. Incorporating protein flexibility into docking and structure-based drug design. *Expert Opinion on Drug Discovery*, 1(4), pp.335–349.

Basconi, J.E. & Shirts, M.R., 2013. Effects of temperature control algorithms on transport properties and kinetics in molecular dynamics simulations. *Journal of Chemical Theory and Computation*, 9(7), pp.2887–2899.

Bell, R.J. & Dean, P., 1970. Atomic vibrations in vitreous silica. *Discussions of the Faraday Society*, 50, pp.55–61.

Berendsen, H.J. et al., 1984. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81(8), pp.3684–3690.

Berka, K. et al., 2012. MOLEonline 2.0: Interactive web-based analysis of biomacromolecular channels. *Nucleic Acids Research*, 40(W1), pp.222–227.

Boehr, D.D., Nussinov, R. & Wright, P.E., 2009. The role of dynamic conformational ensembles in biomolecular recognition. *Nature Chemical Biology*, 5(11), pp.789–796.

Boissonnat, J.D. et al., 2002. Triangulations in CGAL. *Computational Geometry: Theory and Applications*, 22(1-3), pp.5–19.

Braun, E. et al., 2019. Best Practices for Foundations in Molecular Simulations [Article v1.0]. *Living Journal of Computational Molecular Science*, 1(1), pp.1–28.

Braun, E., Moosavi, S.M. & Smit, B., 2018. Anomalous Effects of Velocity Rescaling Algorithms: The Flying Ice Cube Effect Revisited. *Journal of Chemical Theory and Computation*, 14(10), pp.5262–5272.

Braxenthaler, M. et al., 1997. Chaos in protein dynamics. *Proteins: Structure, Function and Genetics*, 29(4), pp.417–425.

Brezovsky, J. et al., 2013. Software tools for identification, visualization and analysis of protein tunnels and channels. *Biotechnology Advances*, 31(1), pp.38–49. Available at: http://dx.doi.org/10.1016/j.biotechadv.2012.02.002.

Brooks, B.R. & Karplus, M., 1983. Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proceedings of the National Academy of Sciences of the United States of America*, 80(21), pp.6571–5. Available at: http://www.ncbi.nlm.nih.gov/pubmed/6579545{\%}0Ahttp://www.pubm

Brooks, B.R. & Karplus, M., 1985. Normal modes for specific motions of macromolecules: application to the hinge-bending mode of lysozyme. *Proceedings of the National Academy of Sciences of the United States of America*, 82(15), pp.4995–4999.

Brooks, B.R., Janežič, D. & Karplus, M., 1995. Harmonic analysis of large systems. I. Methodology. *Journal of Computational Chemistry*, 16(12), pp.1522–1542. Available at: http://doi.wiley.com/10.1002/jcc.540161209.

Bruccoleri, R.E., Karplus, M. & McCammon, J.A., 1986. The hinge-bending mode of a lysozyme–inhibitor complex. *Biopolymers*, 25(9), pp.1767–1802.

Bryngelson, J.D. & Wolynes, P.G., 1989. Intermediates and barrier crossing in a random energy model

(with applications to protein folding). *Journal of Physical Chemistry*, 93(19), pp.6902–6915.

Bryngelson, J.D. & Wolynes, P.G., 1987. Spin glasses and the statistical mechanics of protein folding (disordered systems/irreversible denaturation/molten-globule state/biomolecular self-assembly). *Proc. Nati. Acad. Sci. USA*, 84(November), pp.7524–7528.

Bryngelson, J.D. et al., 1995. Funnels, pathways, and the energy landscape of protein folding: A synthesis. *Proteins: Structure, Function, and Bioinformatics*, 21(3), pp.167–195.

Cai, J. et al., 2012. Solution structure and backbone dynamics of human liver fatty acid binding protein: Fatty acid binding revisited. *Biophysical Journal*, 102(11), pp.2585–2594.

Case, D.A., 1994. Normal mode analysis of protein dynamics. *Current Opinion in Structural Biology*, 4(2), pp.285–290.

Case, D.a. et al., 2005. The Amber biomolecular simulation programs. *Journal of Computational Chemistry*, 26(16), pp.1668–1688.

Chen, Z. et al., 2019. D3Pockets: A Method and Web Server for Systematic Analysis of Protein Pocket Dynamics. *Journal of Chemical Information and Modeling*, p.acs.jcim.9b00332. Available at: http://pubs.acs.org/doi/10.1021/acs.jcim.9b00332.

Copeland, R.A., 2011. Conformational adaptation in drug-target interactions and residence time. *Future Medicinal Chemistry*, 3(12), pp.1491–1501.

Cui, Q. & Bahar, I., 2006. *Normal Mode Analysis Theory and Applications*,

Cui, Q. & Karplus, M., 2008. Allostery and cooperativity revisited., pp.1295–1307.

Cui, Q. & Karplus, M., 2003. Protein Simulations. *Advances in Protein Chemistry*, 66, pp.315–372. Available at: http://www.sciencedirect.com/science/article/pii/S0065323303660080.

Daub, J. et al., 2000. A survey of genes expressed in adults of the human hookworm, Necator americanus. *Parasitology*, 120(2), pp.171–184.

De Groot, B.L. et al., 1996. An extended sampling of the configurational space of HPr from E. coli. *Proteins: Structure, Function and Genetics*, 26(3), pp.314–322.

Dill, K.A. & Chan, S.H., 1997. From Levinthal to Pathways to Funnels: The "New View" of Protein Folding Kinetics. *Nat. Struct. Biol.*, 4(1), p.10.

Durrant, J.D., De Oliveira, C.A.F. & McCammon, J.A., 2011. POVME: An algorithm for measuring binding-pocket volumes. *Journal of Molecular Graphics and Modelling*, 29(5), pp.773–776. Available at: http://dx.doi.org/10.1016/j.jmgm.2010.10.007.

Edelsbrunner, H., Facello, M. & Liang, J., 1998. On the definition and the construction of pockets in macromolecules. *Discrete Applied Mathematics*, 88(1-3), pp.83–102.

Eisenmesser, E.Z. et al., 2005. Intrinsic dynamics of an enzyme underlies catalysis. *Nature*, 438(7064).

Elber, R. & Karplus, M., 1987. Multiple conformational states of proteins: A molecular dynamics analysis of myoglobin. *Science*, 235(4786), pp.318–321.

Elofsson, A. & Nilsson, L., 1993. How consistent are molecular dynamics simulations? Comparing

structure and dynamics in reduced and oxidized Escherichia coli thioredoxin., 233, pp.766–780.

Endres, N.F. et al., 2013. Conformational coupling across the plasma membrane in activation of the EGF receptor. *Cell*, 152(3), pp.543–556. Available at: http://dx.doi.org/10.1016/j.cell.2012.12.032.

Fabri, A. et al., 1998. On the Design of CGAL, the Computational Geometry Algorithms Library. *Esprit*, 29(February), pp.1–38.

Feig, M. et al., 2015. Principal Component Analysis reveals correlation of cavities evolution and functional motions in proteins., 58, pp.1–9.

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

Franchini, G.R. et al., 2015. The unusual lipid binding proteins of parasitic helminths and their potential roles in parasitism and as therapeutic targets. *Prostaglandins Leukotrienes and Essential Fatty Acids*, 93, pp.31–36. Available at: http://dx.doi.org/10.1016/j.plefa.2014.08.003.

Frauenfelder, H., Sligar, S.G. & Wolynes, P.G., 1991. The Energy Landscapes and Motions of Proteins.

Frenkel, D. & Smit, B., 1996. Understanding molecular simulation: From algorithms to applications.

Friedman, R., Nachliel, E. & Gutman, M., 2006. Fatty acid binding proteins: Same structure but different binding mechanisms? Molecular dynamics simulations of intestinal fatty acid binding protein. *Biophysical Journal*, 90(5), pp.1535–1545.

Gajiwala, K.S. et al., 2013. Insights into the aberrant activity of mutant EGFR kinase domain and drug recognition. *Structure*, 21(2), pp.209–219. Available at: http://dx.doi.org/10.1016/j.str.2012.11.014.

García, A.E., 1992. Large-amplitude nonlinear motions in proteins. *Physical Review Letters*, 68(17), pp.2696–2699.

Garofalo, A. et al., 2003. The FAR protein family of the nematode Caenorhabditis elegans: Differential lipid binding properties, structural characteristics, and developmental regulation. *Journal of Biological Chemistry*, 278(10), pp.8065–8074.

Geng, J. et al., 2002. Secretion of a novel developmentally regulated chitinase (family 19 glycosyl hydrolase) into the perivitelline fluid of the parasitic nematode, Ascaris suum. *Molecular and Biochemical Parasitology*, 124(1-2), pp.11–21.

Gerhart, J.C. & Pardee, A.B., 1962. The enzymology of control by feedback inhibition. *The Journal of biological chemistry*, 237(3), pp.891–6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/13897943.

Gianni, S., Dogan, J. & Jemth, P., 2014. Distinguishing induced fit from conformational selection. *Biophysical Chemistry*, 189, pp.33–39. Available at: http://dx.doi.org/10.1016/j.bpc.2014.03.003.

Go, N. & Abe, H., 1981. Noninteracting local-structure model of folding and unfolding transition in globular proteins. *Biopolymers*, 20, pp.991–1011.

Gooljarsingh, L.T. et al., 2006. A biochemical rationale for the anticancer effects of Hsp90 inhibitors: Slow, tight binding inhibition by geldanamycin and its analogues. *Proceedings of the National Academy of Sciences of the United States of America*, 103(20), pp.7625–7630.

Gora, A., Brezovsky, J. & Damborsky, J., 2013. Gates of enzymes. *Chemical Reviews*, 113(8), pp.5871–

5923.

Greives, N. & Zhou, H.X., 2014. Both protein dynamics and ligand concentration can shift the binding mechanism between conformational selection and induced fit. *Proceedings of the National Academy of Sciences of the United States of America*, 111(28), pp.10197–10202.

Groot, B.L. de et al., 1996. The consistency of large concerted motions in proteins in molecular dynamics simulations. *Biophysical Journal*, 71(4), pp.1707–1713.

Grosso, M. et al., 2015. On the analysis and comparison of conformer-specific essential dynamics upon ligand binding to a protein. *The Journal of chemical physics*, 142(24), p.245101. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26133456.

Grubmüller, H., 1995. Predicting slaw structural transitions in macromolecular systems: Conformational flooding. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 52(3), pp.1–153.

Guo, Y., Duan, M. & Yang, M., 2019. The observation of ligand-binding-relevant open states of fatty acid binding protein by molecular dynamics simulations and a markov state model. *International Journal of Molecular Sciences*, 20(14).

Guvench, O. & MacKerell, A.D., 2008. Comparison of protein force fields for molecular dynamics simulations. *Methods in Molecular Biology*, 443, pp.63–88.

Haliloglu, T., Bahar, I. & Erman, B., 1997. Gaussian Dynamics of Folded Proteins., (1), pp.1–4. Available at: papers://c33b182f-cf88-47e8-a9c5-ad67b5626483/Paper/p438.

Hammes, G.G., Chang, Y.C. & Oas, T.G., 2009. Conformational selection or induced fit: A flux description of reaction mechanism. *Proceedings of the National Academy of Sciences of the United States of America*, 106(33), pp.13737–13741.

Harrison, S.C. & Durbin, R., 1985. Is there a single pathway for the folding of a polypeptide chain? *Proceedings of the National Academy of Sciences of the United States of America*, 82(12), pp.4028–4030.

Hasenahuer, M.A. et al., 2017. Pockets as structural descriptors of EGFR kinase conformations. *PLoS ONE*, 12(12).

Hayward, S., Kitao, A. & Gō, N., 1995. Harmonicity and anharmonicity in protein dynamics: A normal mode analysis and principal component analysis. *Proteins: Structure, Function, and Bioinformatics*, 23(2), pp.177–186.

Hinsen, K., 1998. Analysis of domain motions by approximate normal mode calculations. *Proteins Structure Function and …*, 429(July), pp.417–429. Available at: https://i12r-studfilesrv.informatik.tu-muenchen.de/wiki/images/5/54/Hinsen1998{_}Proteins.pdf.

Hinsen, K. & Kneller, G.R., 1999. A simplified force field for describing vibrational protein dynamics over the whole frequency range. *Journal of Chemical Physics*, 111(24), pp.10766–10769.

Hodsdon, M.E. & Cistola, D.P., 1997. Ligand binding alters the backbone mobility of intestinal fatty acid-binding protein as monitored by 15n nmr relaxation and 1h exchange. *Biochemistry*, 36(8), pp.2278–2290.

Huang, B., 2009. Metapocket: A meta approach to improve protein ligand binding site prediction. *OMICS A Journal of Integrative Biology*, 13(4), pp.325–330.

Hub, J.S. & Groot, B.L. de, 2009. Detection of functional modes in protein dynamics. *PLoS computational biology*, 5(8), p.e1000480. Available at: http://journals.plos.org/ploscompbiol/article?id=10.1371/

journal.pcbi.1000480.

Hünenberger, P.H., 2005. Thermostat algorithms for molecular dynamics simulations. *Advances in Polymer Science*, 173, pp.105–147.

Ichiye, T. & Karplus, M., 1991. Collective motions in proteins: A covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins: Structure, Function, and Bioinformatics*, 11(3), pp.205–217. Available at: http://doi.wiley.com/10.1002/prot.340110305.

Ikai, A. & Tanford, C., 1971. Kinetic evidence for incorrectly folded intermediate states in the refolding of denatured proteins. *Nature*, 26(138), pp.584–585.

Jakalian, A., Jack, D.B. & Bayly, C.I., 2002. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *Journal of Computational Chemistry*, 23(16), pp.1623–1641.

Janežič, D. & Brooks, B.R., 1995. Harmonic analysis of large systems. II. Comparison of different protein models. *Journal of Computational Chemistry*, 16(12), pp.1543–1553.

Jordanova, R. et al., 2009. Fatty acid- and retinoid-binding proteins have distinct binding pockets for the two types of cargo. *Journal of Biological Chemistry*, 284(51), pp.35818–35826.

Jorgensen, W.L. et al., 1983. Comparison of simple potential functions for simulating liquid water., 926(May 2012).

Jura, N. et al., 2011. Catalytic Control in the EGF Receptor and Its Connection to General Kinase Regulatory Mechanisms. *Molecular Cell*, 42(1), pp.9–22. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1097276511001808.

Kabsch, W., 1978. A discussion of the solution for the best rotation to relate two sets of vectors., (9).

Karplus, M., 1997. The Levinthal paradox: Yesterday and today. *Folding and Design*, 2(4), pp.69–75.

Karplus, M. & Kushick, J.N., 1981. Method for estimating the configurational entropy of macromolecules. *Macromolecules*, 14(2), pp.325–332.

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

Kendrew, J.C. et al., 1958. A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. *Nature*, 181(4610), pp.662–666. Available at: http://www.nature.com/articles/181662a0.

Keskin, O., Jernigan, R.L. & Bahar, I., 2000. Proteins with similar architecture exhibit similar large-scale dynamic behavior. *Biophysical journal*, 78(4), pp.2093–2106. Available at: http://dx.doi.org/10.1016/S0006-3495(00)76756-7.

Kim, Y.B., Kalinowski, S.S. & Marcinkeviciene, J., 2007. A pre-steady state analysis of ligand binding to human glucokinase: Evidence for a preexisting equilibrium. *Biochemistry*, 46(5), pp.1423–1431.

Kirkpatrick, M., 2009. Patterns of quantitative genetic variation in multiple dimensions. *Genetica*, 136(2), pp.271–284.

Kokh, D.B. et al., 2013. TRAPP: A tool for analysis of Transient binding Pockets in Proteins. *Journal of Chemical Information and Modeling*, 53(5), pp.1235–1252.

Kondrashov, D.A. et al., 2007. Protein Structural Variation in Computational Models and Crystallo-

graphic Data. *Structure*, 15(2), pp.169–177. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0969212607000299.

Koshland, D.E., 1958. Application of a theory of enzyme specificity to protein synthesis. *Proc. Natl. Acad. Sci.*, pp.98–104.

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), pp.2375–2378.

Krone, M. et al., 2016. Visual Analysis of Biomolecular Cavities: State of the Art. *Computer Graphics Forum*, 35(3), pp.527–551.

Krzanowski, W.J., 1979. Between-Groups Comparison of Principal Components. *Journal of the American Statistical Association*, 74(367), p.703. Available at: https://www.jstor.org/stable/2286995?origin=crossref.

Kumar, S. et al., 2008. Folding and binding cascades: Dynamic landscapes and population shifts. *Protein Science*, 9(1), pp.10–19.

Kundu, S. et al., 2002. Dynamics of proteins in crystals: Comparison of experiment with simple models. *Biophysical Journal*, 83(2), pp.723–732. Available at: http://dx.doi.org/10.1016/S0006-3495(02)75203-X.

Laskowski, R.a., 1995. SURFNET: A program for visualizing molecular surfaces, cavities, and intermolecular interactions. *Journal of Molecular Graphics*, 13(5), pp.323–330.

Laulumaa, S. et al., 2018. Structure and dynamics of a human myelin protein P2 portal region mutant indicate opening of the β barrel in fatty acid binding proteins. *BMC Structural Biology*, 18(1), pp.1–13.

Laurent, B. et al., 2014. Epock: Rapid analysis of protein pocket dynamics. *Bioinformatics*, 31(9), pp.1478–1480.

Le Guilloux, V., Schmidtke, P. & Tuffery, P., 2009. Fpocket: an open source platform for ligand pocket detection. *BMC bioinformatics*, 10, p.168.

Leach, A.R., 2001. *The Use of Molecular modelling and Chemoinformatics to Discover and Design New Molecules*, Available at: http://books.google.de/books?id=kB7jsbV-uhkC.

Lee, B. & Richards, F.M., 1971. The interpretation of protein structures: Estimation of static accessibility. *Journal of Molecular Biology*, 55(3).

Leroy, F., 2013. *Molecular Driving Forces. Statistical Thermodynamics in Biology, Chemistry, Physics, and Nanoscience*,

Levy, R.M. et al., 1984. Evaluation of the Configurational Entropy for Proteins: Application to Molecular Dynamics Simulations of an α-Helix. *Macromolecules*, 17(7), pp.1370–1374.

Li, G. & Cui, Q., 2002. A coarse-grained normal mode approach for macromolecules: An efficient implementation and application to Ca2+-ATPase. *Biophysical Journal*, 83(5), pp.2457–2474. Available at: http://dx.doi.org/10.1016/S0006-3495(02)75257-0.

Li, Y. et al., 2014. Conformational transition pathways of epidermal growth factor receptor kinase domain from multiple molecular dynamics simulations and Bayesian clustering. *Journal of Chemical Theory and Computation*, 10(8), pp.3503–3511.

Liang, J., Edelsbrunner, H. & Woodward, C., 1998. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein science : a publication of the*

Protein Society, 7, pp.1884–1897.

Liang, J. et al., 1998. Analytical shape computation of macromolecules: I. Molecular area and volume through alpha shape. *Proteins: Structure, Function and Genetics*, 33(1), pp.1–17.

Liang, J. et al., 1998. Analytical shape computing of macromolecules II: identification and computation of inaccessible cavities inside proteins. *Proteins*, 33(1), pp.18–29.

Loukas, A. et al., 2016. Hookworm infection. *Nature Reviews Disease Primers*, 2.

Maguid, S., Fernandez-Alberti, S. & Echave, J., 2008. Evolutionary conservation of protein vibrational dynamics. *Gene*, 422(1-2), pp.7–13. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18577430.

Mahajan, S. & Sanejouand, Y.H., 2015. On the relationship between low-frequency normal modes and the large-scale conformational changes of proteins. *Archives of Biochemistry and Biophysics*, 567(January), pp.59–65. Available at: http://dx.doi.org/10.1016/j.abb.2014.12.020.

Maier, J.A. et al., 2015. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of Chemical Theory and Computation*, 11(8), pp.3696–3713.

Marsh, J.A., Teichmann, S.A. & Forman-Kay, J.D., 2012. Probing the diverse landscape of protein flexibility and binding. *Current Opinion in Structural Biology*, 22(5), pp.643–650. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0959440X1200142X.

Matsuoka, D. et al., 2015. Molecular dynamics simulations of heart-type fatty acid binding protein in apo and holo forms, and hydration structure analyses in the binding cavity. *Journal of Physical Chemistry B*, 119(1), pp.114–127.

Matthews, B.W. & Remington, S.J., 1974. The three dimensional structure of the lysozyme from bacteriophage T4. *Proceedings of the National Academy of Sciences of the United States of America*, 71(10), pp.4178–4182.

McCammon, J.A., Gelin, B.R. & Karplus, M., 1977. Dynamics of folded proteins. *Nature*.

McCammon, J.A. et al., 1976. The hinge-bending mode in lysozyme. *Nature*, 262(5566), pp.325–326.

McDermott, L., Cooper, A. & Kennedy, M.W., 1999. Novel classes of fatty acid and retinol binding protein from nematodes. *Molecular and Cellular Biochemistry*, 192(1-2), pp.69–75.

McDonald, I. & Thornton, J., 1994. Satisfying hydrogen bonding potential in Proteins. *Journal of Molecular Biology*.

Medek, P., Benes, P. & Sochor, J., 2007. Computation of Tunnels in Protein Molecules using Delaunay Triangulation. *Journal of WSCG*, 15(1-3), pp.107——114. Available at: http://decibel.fi.muni.cz/download/papers/medek08.pdf.

Metropolis, N. et al., 1953. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6), pp.1087–1092.

Micheletti, C., Carloni, P. & Maritan, A., 2004. Accurate and Efficient Description of Protein Vibrational Dynamics: Comparing Molecular Dynamics and Gaussian Models. *Proteins: Structure, Function and Genetics*, 55(3), pp.635–645.

Mirsky, A.E. & Pauling, L., 1931. On the Structure of Native, denatured, and coagulated proteins.

Scientific Papers Part I. Cf. H. Poincar6, op. cit, 83(3), pp.179–184.

Miyashita, O., Onuchic, J.N. & Wolynes, P.G., 2003. Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 100(22), pp.12570–12575.

Monod, J., Wyman, J. & Changeux, J.P., 1965. On the Nature of Allosteric Transitions : A Plausible Model., (December).

Monticelli, L. & Tieleman, P.D., 2013. Force Fields for Classical Molecular Dynamics. *Essentials of Micro- and Nanofluidics*, 924, pp.447–474.

Monzon, A.M. et al., 2016. CoDNaS 2.0: A comprehensive database of protein conformational diversity in the native state. *Database*, 2016, pp.1–8.

Monzon, A.M. et al., 2017. Conformational diversity analysis reveals three functional mechanisms in proteins. *PLoS Computational Biology*, 13(2), pp.1–18.

Morando, M.A. et al., 2016. Conformational Selection and Induced Fit Mechanisms in the Binding of an Anticancer Drug to the c-Src Kinase. *Scientific Reports*, 6(April), p.24439. Available at: http://www.nature.com/articles/srep24439.

Nussinov, R. et al., 2013. Allosteric conformational barcodes direct signaling in the cell. *Structure*, 21(9), pp.1509–1521. Available at: http://dx.doi.org/10.1016/j.str.2013.06.002.

Okazaki, K.-I. & Takada, S., 2008. Dynamic energy landscape view of coupled binding and protein conformational change: induced-fit versus population-shift mechanisms. *Proceedings of the National Academy of Sciences of the United States of America*, 105(32), pp.11182–11187. Available at: http://www.pnas.org/content/105/32/11182.full.

Onuchic, J.N. et al., 1994. Toward an outline of the topography of a realistic protein-folding funnel. *Proceedings of the National Academy of Sciences*, 92(8), pp.3626–3630.

Orellana, L. et al., 2010. Approaching elastic network models to molecular dynamics flexibility. *Journal of Chemical Theory and Computation*, 6(9), pp.2910–2923.

Overmars, M.H., 1996. Designing the computational geometry algorithms library CGAL. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 1148, pp.53–58.

P. Kollman, R. Dixon, W. Cornell, T. Fox, C. Chipot, A.P., 1997. *Computer Simulation of Biomolecular Systems: Theoretical and Experimental Applications, Volume 3*, Available at: http://books.google.com/books?hl=fr{\&}lr={\&}id=B3iDerOPyoUC{\&}pgis=1.

Pabon, N.A. & Camacho, C.J., 2017. Probing protein flexibility reveals a mechanism for selective promiscuity. *eLife*, 6, pp.1–24.

Patrick et al., 1991. *Rational design of potent, bioavailable, nonpeptide cyclic ureas as HIV protease inhibitors*, Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.8278812.

Pauling, L., 1940. A Theory of the Formation of Antibodies., 372(6). Available at: https://pubs.acs.org/sharingguidelines.

Pavelka, A. et al., 2016. CAVER: Algorithms for Analyzing Dynamics of Tunnels in Macromolecules.

IEEE/ACM Transactions on Computational Biology and Bioinformatics, 13(3), pp.505–517.

Perutz, M.F., 1970. Stereochemistry of cooperative effects in haemoglobin. *Nature*, 225, pp.538–539.

Petrek, M. et al., 2006. CAVER: a new tool to explore routes from protein clefts, pockets and cavities. *BMC bioinformatics*, 7, p.316.

Petřek, M. et al., 2007. MOLE: A Voronoi Diagram-Based Explorer of Molecular Channels, Pores, and Tunnels. *Structure*, 15(11), pp.1357–1363.

Ponder, J.W. et al., 2010. Current status of the AMOEBA polarizable force field. *Journal of Physical Chemistry B*, 114(8), pp.2549–2564.

Pravda, L. et al., 2014. Anatomy of enzyme channels. *BMC Bioinformatics*, 15(1), p.379. Available at: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-014-0379-x.

Radhakrishnan, S., Kolippakkam, D. & Mathura, V.S., 2007. *Introduction to algorithms*,

Rahman, A., 1964. Correlations in the Motion of Atoms in Liquid Argon. *Phys. Rev.*, 136(2A), pp.405–411.

Rahman, A. & Stillinger, F.H., 1971. Molecular Dynamics Study of Liquid Water. *The Journal of Chemical Physics*, 55(7), pp.3336–3359.

Rey-Burusco, M.F. et al., 2015. Diversity in the structures and ligand-binding sites of nematode fatty acid and retinol-binding proteins revealed by Na-FAR-1 from Necator americanus. *Biochemical Journal*, 471(3), pp.403–414. Available at: http://biochemj.org/cgi/doi/10.1042/BJ20150068.

Rizzi, A. et al., 2019. The SAMPL6 SAMPLing challenge: Assessing the reliability and efficiency of binding free energy calculations. *bioRxiv*, p.795005. Available at: https://www.biorxiv.org/content/10.1101/795005v1.full.

Rueda, M., Chacón, P. & Orozco, M., 2007. Thorough Validation of Protein Normal Mode Analysis: A Comparative Study with Essential Dynamics. *Structure*, 15(5), pp.565–575.

Salda, T.E. et al., 2016. Evolutionary Conserved Positions Define Protein Conformational Diversity., pp.1–25.

Sanejouand, Y.-H., 2013. Elastic Network Models: Theoretical and Empirical Foundations. In *Biomolecular simulations: Methods and protocols*. pp. 601–616. Available at: http://link.springer.com/10.1007/978-1-62703-017-5{_}23.

Sani, B.P. & Vaid, A., 1988. Specific interaction of ivermectin with retinol-binding protein from filarial parasites. *Biochemical Journal*, 249(3), pp.929–932.

Sanner, M.F., Olson, A.J. & Spehner, J.C., 1996. Reduced surface: An efficient way to compute molecular surfaces. *Biopolymers*, 38(3), pp.305–320.

Schmidtke, P. et al., 2011. MDpocket: Open-source cavity detection and characterization on molecular dynamics trajectories. *Bioinformatics*, 27(23), pp.3276–3285.

Schneider, T. & Stoll, E., 1978. Molecular-dynamics study of a three-dimensional one-component model

for distortive phase transitions. *Physical Review B*, 17(3), pp.1302–1322.

Sehnal, D., 2013. MOLE 2 . 0 : advanced approach for analysis of biomacromolecular channels., pp.1–13.

Sha, R.S. et al., 1993. Modulation of ligand binding affinity of the adipocyte lipid-binding protein by selective mutation. Analysis in vitro and in situ. *Journal of Biological Chemistry*, 268(11), pp.7885–7892.

Shan, Y. et al., 2013. Transitions to catalytically inactive conformations in EGFR kinase. *Proceedings of the National Academy of Sciences*, 110(18), pp.7270–7275. Available at: http://www.pnas.org/cgi/doi/10.1073/pnas.1220843110.

Shan, Y. et al., 2012. Oncogenic mutations counteract intrinsic disorder in the EGFR kinase and promote receptor dimerization. *Cell*, 149(4), pp.860–870. Available at: http://dx.doi.org/10.1016/j.cell.2012.02.063.

Simões, T. et al., 2017. Geometric Detection Algorithms for Cavities on Protein Surfaces in Molecular Graphics: A Survey. *Computer Graphics Forum*, 00(0), pp.1–41.

Smart, O.S. et al., 1996. HOLE: A program for the analysis of the pore dimensions of ion channel structural models. *Journal of Molecular Graphics*, 14(6), pp.354–360.

Solovyova, A.S. et al., 2003. The polyprotein and FAR lipid binding proteins of nematodes: Shape and monomer/dimer states in ligand-free and bound forms. *European Biophysics Journal*, 32(5), pp.465–476.

Stank, A. et al., 2017. TRAPP webserver: Predicting protein binding site flexibility and detecting transient binding pockets. *Nucleic Acids Research*, 45(W1), pp.W325–W330.

Stillinger, F.H. & Rahman, A., 1974. Improved simulation of liquid water by molecular dynamics. *The Journal of Chemical Physics*, 60(4), pp.1545–1557. Available at: http://aip.scitation.org/doi/10.1063/1.1681229.

Storch, J. & McDermott, L., 2009. Structural and functional analysis of fatty acid-binding proteins. *Journal of Lipid Research*, 50(SUPPL.).

Šali, A. & Blundell, T.L., 1993. Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, 234(3), pp.779–815.

Tama, F. & Sanejouand, Y.H., 2001. Conformational change of proteins arising from normal mode calculations. *Protein Engineering Design and Selection*, 14(1), pp.1–6. Available at: http://peds.oxfordjournals.org/cgi/doi/10.1093/protein/14.1.1.

Taylor, S.S.S. & Kornev, A.A.P., 2011. Protein Kinases: Evolution of Dynamic Regulatory Proteins. *Trends in biochemical sciences*, 36(2), pp.65–77. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?ar

Teeter, M.M. & Case, D.A., 1990. Harmonic and quasiharmonic descriptions of crambin. *Journal of Physical Chemistry*, 94(21), pp.8091–8097.

Tirion, M.M., 1996. Large amplitude elastic motions in proteins from a single parameter atomic analysis. *Phyisical Review Letters*.

Tokuriki, N. & Tawfik, D.S., 2009. Protein Dynamism and Evolvability. *Science*, 324(April), pp.203–207.

Tsai, C.-J. & Nussinov, R., 2014. A unified view of "how allostery works". *PLoS computational biology*, 10(2), p.e1003394.

Tsai, C.J. et al., 1999. Folding funnels, binding funnels, and protein function. *Protein science : a*

publication of the Protein Society, 8(6), pp.1181–90.

Tsfadia, Y. et al., 2007. Molecular dynamics simulations of palmitate entry into the hydrophobic pocket of the fatty acid binding protein. *FEBS Letters*, 581(6), pp.1243–1247.

Tsong, T.Y., Baldwin, R.L. & Elson, E.L., 1971. The sequential unfolding of ribonuclease A: detection of a fast initial phase in the kinetics of unfolding. *Proceedings of the National Academy of Sciences of the United States of America*, 68(11), pp.2712–2715.

Umbarger, H.E. & Brown, B., 1957. Threonine deamination in Escherichia coli. II. Evidence for two L-threonine deaminases. *Journal of bacteriology*, 73(1), pp.105–12.

Van Aalten, D. et al., 1996. A Comparison of Techniques for Calculating Protein Essential Dynamics. *Journal of computational chemistry*, 18(2), pp.169–181.

Vos, T. et al., 2017. Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990-2016: A systematic analysis for the Global Burden of Disease Study 2016. *The Lancet*, 390(10100), pp.1211–1259.

Voss, N.R. & Gerstein, M., 2010. 3V: Cavity, channel and cleft volume calculator and extractor. *Nucleic Acids Research*, 38(SUPPL. 2), pp.555–562.

Wagner, J.R. et al., 2017. POVME 3.0: Software for Mapping Binding Pocket Flexibility. Available at: http://pubs.acs.org.sire.ub.edu/doi/pdf/10.1021/acs.jctc.7b00500.

Wall, M.E., Rechtsteiner, A. & Rocha, L.M., 2005. Singular Value Decomposition and Principal Component Analysis. *A Practical Approach to Microarray Data Analysis*, pp.91–109.

Wan, S. & Coveney, P.V., 2011. Molecular dynamics simulation reveals structural and thermodynamic features of kinase activation by cancer mutations within the epidermal growth factor receptor. *Journal of Computational Chemistry*, 32(13), pp.2843–2852.

Wan, S., Wright, D.W. & Coveney, P.V., 2012. Mechanism of Drug Efficacy within the Epidermal Growth Factor Receptor Revealed by Microsecond Molecular Dynamics Simulation. *Molecular Cancer Therapeutics*, 11(November), pp.2394–2401.

Wang, J. et al., 2004. Development and testing of a general Amber force field. *Journal of Computational Chemistry*, 25(9), pp.1157–1174.

Wang, Q., Zorn, J.A. & Kuriyan, J., 2014. *A structural atlas of kinases inhibited by clinically approved drugs* 1st ed., Elsevier Inc. Available at: http://dx.doi.org/10.1016/B978-0-12-397918-6.00002-1.

Wang, Y. et al., 2004. Global ribosome motions revealed with elastic network model. *Journal of Structural Biology*, 147(3), pp.302–314.

Wei, G. et al., 2016. Protein Ensembles: How Does Nature Harness Thermodynamic Fluctuations for Life? The Diverse Functional Roles of Conformational Ensembles in the Cell. *Chemical Reviews*, p.acs.chemrev.5b00562. Available at: http://pubs.acs.org/doi/abs/10.1021/acs.chemrev.5b00562.

Weisel, M., Proschak, E. & Schneider, G., 2007. PocketPicker: analysis of ligand binding-sites with shape descriptors. *Chemistry Central journal*, 1, p.7.

Yang, L., Song, G. & Jernigan, R.L., 2009. Comparisons of experimental and computed protein anisotropic temperature factors. *Proteins: Structure, Function and Bioinformatics*, 76(1), pp.164–175.

Yang, L.-W. & Chng, C.-P., 2008. Coarse-Grained Models Reveal Functional Dynamics - I. Elastic

Network Models – Theories, Comparisons and Perspectives. *Bioinformatics and Biology Insights*, 2, p.BBI.S460.

Yu, J. et al., 2009. Roll: A new algorithm for the detection of protein pockets and cavities with a rolling probe sphere. *Bioinformatics*, 26(1), pp.46–52.

Zhan, B. et al., 2018. Ligand binding properties of two Brugia malayi fatty acid and retinol (FAR) binding proteins and their vaccine efficacies against challenge infection in gerbils. *PLoS neglected tropical diseases*, 12(10), p.e0006772.

Zheng, W., Brooks, B.R. & Thirumalai, D., 2006. Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations. *Proceedings of the National Academy of Sciences*, 103(20), pp.7664–7669.

Zheng, W. et al., 2005. Network of dynamically important residues in the open/closed transition in polymerases is strongly conserved. *Structure*, 13(4), pp.565–577.

Zhou, H.X., 2010. From induced fit to conformational selection: A continuum of binding mechanism controlled by the timescale of conformational transitions. *Biophysical Journal*, 98(6), pp.L15–L17. Available at: http://dx.doi.org/10.1016/j.bpj.2009.11.029.

Zhou, H.X., Wlodek, S.T. & Mccammon, J.A., 1998. Conformation gating as a mechanism for enzyme specificity. *Proceedings of the National Academy of Sciences of the United States of America*, 95(16), pp.9280–9283.

Zhou, R., 2003. Trp-cage: folding free energy landscape in explicit water. *Proceedings of the National Academy of Sciences of the United States of America*, 100, pp.13280–13285. Available at: http://www.pnas.org/content/100/23/13280.full.pdf{\%}5Cnhttp://www.pnas.org/content/100/23/13280.abstract.

Zhu, H. & Pisabarro, M.T., 2011. MSPocket: An orientation-independent algorithm for the detection of ligand binding pockets. *Bioinformatics*, 27(3), pp.351–358.

Zimmerman, A.W. & Veerkamp, J.H., 2002. New insights into the structure and function of fatty acid-binding proteins. *Cellular and Molecular Life Sciences*, 59(7), pp.1096–1116.

Zonta, M.L., Oyhenart, E.E. & Navone, G.T., 2010. Nutritional status, body composition, and intestinal parasitism among the mbyá -guaraní communities of misiones, Argentina. *American Journal of Human Biology*, 22(2), pp.193–200.

Zuckerman, D.M., 2011. Equilibrium Sampling in Biomolecular Simulations. *Annual Review of Biophysics*, 40(1), pp.41–62.