

Proyecto Fin de Master:

**Estudio computacional del mecanismo
de unión-disociación de la flavodoxina de
Helicobacter pylori y su ligando.**

Autor: Lic. Guillermo Gutiérrez Teixeira

Director de Proyecto: Dr. Javier Sancho Sanz

Curso 2012-2013

Índice:

1.Introducción	1
2.Métodos y materiales	3
Plataformas de cálculo.....	3
Aplicaciones.....	3
Métodos	4
Dinámica molecular.....	4
Igualdad de Jarzynski.....	7
Método de medias de bloques.....	9
Experimentación.....	11
3.Resultados	17
Fase de preparación	17
Fase de producción.....	19
4.Discusión y conclusiones.....	25
5.Bibliografía y referencias	29

1. Introducción:

Helicobacter pylori es una bacteria que infecta el epitelio gástrico en el ser humano. Pese a que los primeros indicios de su existencia se encontraron hace más de un siglo no fue hasta las décadas finales del siglo XX cuando se consiguió cultivar y caracterizar ^[1]. A día de hoy se estima que más de la mitad de la población mundial está infectada, variando la prevalencia en función de la localización geográfica, la edad, grupo étnico y otros factores, pudiendo llegar a alcanzar cotas cercanas al 90% en algunos países ^[2]. Este microorganismo ha sido reconocido, no sin cierta controversia en su momento, como agente causal de diversos trastornos gástricos como gastritis, úlceras o cáncer de estómago ^[3]. En la mayoría de los casos la infección transcurre de forma asintomática, y de hecho se estudia la posibilidad de que la presencia de *H. pylori* en el estómago sea beneficiosa^[4], sin embargo, en los casos en los que se producen trastornos resulta claramente favorable para el paciente la eliminación de la bacteria de su tracto digestivo. Para ello es imperativo el uso de antibióticos (los casos en los que se han observado la erradicación espontánea son muy escasos y están asociados a un mal estado de la mucosa en el estómago), pero con su uso surgen cepas de *H. pylori* resistentes al tratamiento ^[5-7] y pueden aparecer trastornos asociados a la perturbación de la flora bacteriana beneficiosa y necesaria en caso de que la especificidad del antibiótico no sea lo suficientemente alta. Es por esto por lo que se buscan activamente nuevas dianas terapéuticas que sean propias de *H. pylori*, entre las que se encuentra la proteína sobre la que se centra este estudio^[8-10].

La flavodoxina de *H. pylori* es una proteína red-ox cuya función es acoplar la reacción de descarboxilación oxidativa del piruvato a la producción de NADPH mediante el transporte de electrones desde el complejo piruvato-ferredoxina oxidorreductasa al complejo flavodoxina-quinona reductasa^[11]. El buen funcionamiento de esta proteína es imprescindible para la supervivencia del microorganismo^[12]. En el centro activo se encuentra unida de forma no covalente una molécula de flavin-mononucleótido (FMN). Curiosamente el sitio de unión del ligando presenta una diferencia importante respecto de otras flavodoxinas semejantes, que es la sustitución de un residuo de triptófano por alanina, lo que hace que quede un

espacio descubierto cercano al cofactor. Esta particularidad hace que el cofactor sea *diferencialmente* accesible para desarrollar antibióticos específicos. En ese sentido ya se han desarrollado trabajos^[13-14], y otros están en plena investigación.

El propósito de este proyecto es profundizar en el conocimiento acerca de la interacción proteína-ligando orientado hacia el caso concreto de la flavodoxina de *H. pylori* completando aspectos que el trabajo previo de caracterización de la proteína en laboratorio no pudo llegar a resolver, entre ellos las interacciones que se dan entre el ligando y la proteína y aspectos del mecanismo de unión-disociación, para aportar información nueva que tal vez pueda ayudar de cara al desarrollo dirigido de antibióticos.

2. Métodos y materiales:

Todos los métodos aplicados en el desarrollo del proyecto son computacionales, y por tanto todos los materiales son o bien plataformas de computación de alto rendimiento o bien herramientas informáticas.

Plataformas de cálculo:

Las plataformas utilizadas han sido TERMINUS, y MEMENTO, alojados en las instalaciones del Instituto de Biocomputación y Física de Sistemas Complejos (BIFI), Minotauro, del Centro de Supercomputación de Barcelona (BSC) y adicionalmente, prometeo, una estación de mi propiedad y competitiva con las anteriores.

TERMINUS y MEMENTO son plataformas que basan su potencia de cálculo directamente en la CPU, (microprocesadores Intel® Xeon® y AMD Opteron™ respectivamente), mientras que Minotauro y prometeo, de diseño y tecnología más reciente y adecuado al ámbito del proyecto, realizan los cálculos a través de GPU (tarjetas gráficas Nvidia Tesla™ M2090 y GTX660, con tecnología Kepler™, respectivamente), aumentando la velocidad y potencia de cálculo a la vez que se reduce el tamaño del supercomputador y por tanto su consumo. Todos los centros de cálculo utilizan distribuciones de Linux como sistema operativo. El total de tiempo de cálculo consumido para llevar a cabo todos los experimentos y pruebas, y el tratamiento de datos ronda el medio millón de horas

Aplicaciones:

Para llevar a cabo las simulaciones se ha utilizado la aplicación NAMD^[15] en su versión 2.9, desarrollada por el grupo de Biofísica Teórica y Computacional en el Instituto Beckman de Ciencia Avanzada y Tecnología de la Universidad de Illinois. En las plataformas que así lo permiten se ha utilizado la versión CUDA de la aplicación para realizar los cálculos por medio de GPUs.

En la preparación del sistema, visionado de las simulaciones y la extracción de datos a partir de éstas, se ha utilizado la aplicación VMD^[16] en su versión 1.9.1, desarrollada también por el citado grupo, así como varios de sus plugins de análisis y modelado.

El análisis de componentes principales de las trayectorias se llevó a cabo a través de la suite de aplicaciones PCZtools en su versión 4.0, desarrollada por el grupo de Reconocimiento Molecular, dentro del grupo de Modelado Molecular y Bioinformática perteneciente al Instituto de Investigación Biomédica de Barcelona, al Parque Científico de Barcelona y a la Universidad de Barcelona.

Para la mayor parte de los análisis, configuración de las simulaciones, lanzamiento de simulaciones a través de los diversos supercomputadores, comprobación y mantenimiento de los archivos se han utilizado scripts y plantillas, todos ellos de desarrollo propio (incluidos los más importantes en los anexos) salvo el usado para calcular la energía libre del proceso de disociación por estimación de medias de bloques, que fue creado por los autores que describieron la técnica (ver apartado de métodos).

Las figuras han sido creadas a través de VMD (representaciones moleculares) o la herramienta de software libre gnuplot (para gráficas y funciones).

Métodos:

Dinámica molecular

Los datos sobre los que se realizan los cálculos para obtener resultados provienen todos en última instancia de simulaciones de dinámica molecular. Esta técnica consiste en aplicar las leyes de un modelo físico a escala atómica sobre los datos de la posición de cada átomo o partícula de una estructura molecular de partida para predecir su comportamiento a lo largo del tiempo. Habitualmente el volumen de cálculos que se deben realizar es inabordable de no usar ordenadores para resolverlos, y por tanto la técnica viene desarrollándose de forma efectiva desde poco después del origen de los computadores^[17], ampliándose al ritmo del crecimiento de la potencia de cálculo su ámbito de aplicación desde la química a un nivel muy básico en sus orígenes hasta la simulación de polímeros, nanomateriales y sistemas de interés biológico^[18], desde los más pequeños, abarcando pocas moléculas, hasta los más grandes, que incluyen decenas de millones de átomos, a día de hoy^[19-20].

La simulación de un sistema complejo en el ámbito de la mecánica cuántica, que sería lo adecuado a esta escala, resulta impracticable debido al alto coste computacional, que aumenta muy rápidamente con el número de electrones en el

sistema, por lo que se deben realizar aproximaciones más eficientes y que conduzcan a resultados razonables y verosímiles en un tiempo asequible. La primera y principal de estas aproximaciones consiste en aplicar las leyes de la dinámica de Newton, la mecánica clásica, lo que implica una serie de asunciones y consecuencias:

Los átomos son considerados corpuscularmente como esferas homogéneas y relativamente blandas dotadas de masa y carga, con posiciones bien definidas en el espacio y en el tiempo. También quedan excluidos del cálculo los fenómenos de polarización al igual que procesos que dependan directamente de fenómenos cuánticos como el efecto túnel, presente, por ejemplo, en la inversión del amoniaco y algunas aminas, o algunas reacciones de transferencia de hidruro.

Los enlaces se aproximan mediante potenciales armónicos modelados como si de muelles se tratase, con una distancia de equilibrio y constante de rigidez introducidas en el cálculo como parámetros. Esto implica que los enlaces no se pueden formar ni romper, impidiendo llevar a cabo simulaciones de reacciones, desde las que se puedan dar en el centro activo hasta fenómenos más ubicuos como la autoionización del agua u otras reacciones de equilibrio ácido-base.

La resolución del sistema de ecuaciones diferenciales que relaciona la fuerza sobre cada átomo y su posición a lo largo del tiempo no se puede llevar a cabo de forma analítica para sistemas de más de dos átomos (problema de N cuerpos), y por tanto se debe hacer una aproximación numérica en la que se calcula la fuerza sobre cada átomo y recursivamente las variaciones de aceleración, velocidad y posición para intervalos discretos de tiempo. El tamaño del incremento de tiempo debe ser tan aproximado a un diferencial como resulte posible. En la práctica suele tomar valores entre 0.5 y 2.0 femtosegundos. La fuerza para cada átomo se calcula como el gradiente de la energía potencial del sistema. Para poder hallarla se debe conocer la función de energía potencial, que se estima mediante un conjunto de funciones y parámetros al que se denomina campo de fuerzas. En este campo de fuerzas se clasifican las contribuciones energéticas entre las que se deben al enlace covalente (energía potencial del enlace, de rotación, de torsión y deformación de ángulos simples y diedros) y las que no (interacciones electrostáticas, interacciones débiles y otras). Pese a que las funciones en sí han sido muy estudiadas y se conocen bien, éstas requieren parámetros cuyos valores dependen del elemento y el entorno de enlaces en el que se

encuentre el átomo, y se han determinado por medios semiempíricos combinando resultados obtenidos de cálculos de mecánica cuántica con las observaciones experimentales.

Todo esto hace que la técnica pueda arrojar en ocasiones resultados poco fiables o que presenten una desviación importante de la realidad. Incluso si eso ocurre y los resultados cuantitativos no son realistas, en muchos casos se pueden extraer conclusiones en aspectos cualitativos.

Para paliar las carencias del método se puede utilizar una combinación de mecánica clásica (para la mayor parte del sistema) y mecánica cuántica (para una región concreta de interés con un tamaño reducido). Esta modificación resulta interesante para simular reacciones u otros procesos, aunque presenta sus dificultades y desventajas. Dado que en el proyecto no se ha utilizado esta técnica mixta de QM/MM, no se va a profundizar más en su funcionamiento.

Hay otras mejoras sobre la base de la dinámica molecular que sí se aplican en las simulaciones de este proyecto. Por una parte se utilizan condiciones de contorno periódico, es decir que la celda básica de simulación se repite infinitamente en todas las direcciones del espacio de forma que si una parte del sistema sale por uno de los lados de la celda aparece inmediatamente por el lado opuesto. En el cálculo de las interacciones, se aplica una corrección habitual en la que se disminuye el efecto de las interacciones entre átomos que distan una cierta distancia hasta eliminarse totalmente del cálculo a partir de una distancia límite, agilizando mucho los cálculos sin afectar de forma notable al desarrollo de la dinámica. Para las moléculas de agua se considera que los enlaces son rígidos evitando un volumen de cálculos de cierto peso. Se aplican también las mejoras ofrecidas por la dinámica de Langevin^[21-22] que introduciendo un componente estocástico en la simulación imita mejor el comportamiento del disolvente como fluido viscoso en sistemas reales. Esto permite además controlar la temperatura del sistema como si de un experimento termostatizado se tratase. La presión se mantiene constante mediante el llamado pistón de Langevin, o método de Nosé-Hoover^[23], consistente en re-escalar el tamaño de la celda unidad y las posiciones de los átomos en ella contenidos. Al reducir la celda la presión aumenta, y al agrandarla disminuye. Los cálculos de interacciones electrostáticas periódicas se llevan a cabo mediante el método de retícula de

partículas de Ewald (PME). Se utilizan también, en función del tipo de simulación, fuerzas externas para mantener de forma suave la proteína relativamente centrada en la caja, relajar el sistema por partes antes de llevar a cabo las simulaciones, o impedir el desplazamiento de la proteína en los experimentos de estiramiento.

Dentro de la dinámica molecular se ha hecho uso en este proyecto de las dinámicas de estiramiento o dinámicas forzadas (SMD por sus siglas en inglés). En ellas se aplica una fuerza sobre uno o más átomos del sistema en una dirección dada. Estas dinámicas pueden ser básicamente de dos tipos: a fuerza constante o a velocidad constante. En la primera, la fuerza que se aplica no varía en el tiempo, mientras que en la segunda la fuerza se calcula como si se debiese a un potencial armónico entre los átomos en cuestión y un átomo ficticio que se desplaza por la simulación a velocidad constante. Este segundo caso resulta más útil para los experimentos planteados y es el que se ha empleado. Para configurar este tipo de simulaciones es necesario introducir dos parámetros: la velocidad de desplazamiento del átomo ficticio y la constante elástica asociada al potencial armónico. Las dinámicas forzadas son una herramienta valiosa a la hora de estudiar sistemas fuera del equilibrio o dirigir la evolución del sistema para que siga preferentemente el desarrollo de un proceso. En este proyecto se utilizan fundamentalmente para obtener potenciales de fuerza media y diferencias de energía libre para el proceso de disociación del FMN a través de métodos derivados de la igualdad de Jarzyinski, sucintamente descrita más abajo. A la hora de dar valores a la constante elástica y la velocidad del átomo ficticio se debe llegar a un compromiso, puesto que valores altos de la velocidad alejan al sistema del equilibrio haciendo más lenta la convergencia de la igualdad de Jarzynski, mientras que valores altos de la constante afectan negativamente a las estimaciones por este mismo método.

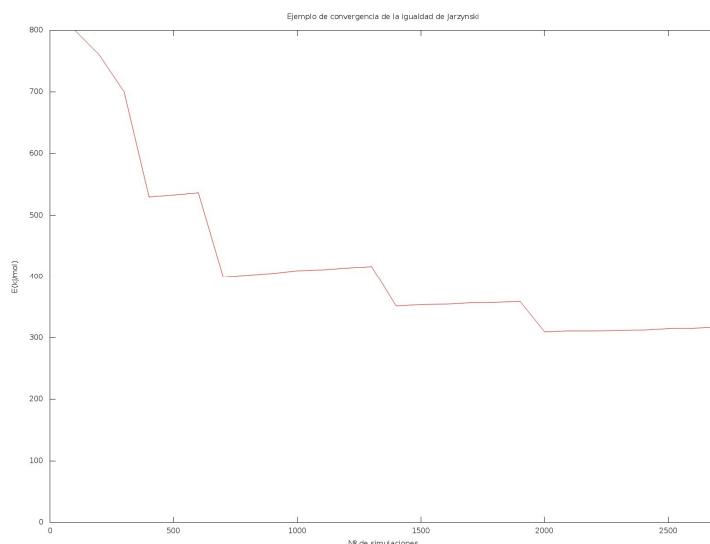
Igualdad de Jarzynski^[24]

Dado un sistema microscópico finito, cualquier alteración que produzca en él una variación de la energía libre está asociada a un trabajo realizado por o hacia el sistema. Este trabajo incluye contribuciones conservativas y disipativas, de forma que el trabajo que se realiza es siempre superior a la variación del trabajo “útil” que representa la energía libre, $W_T \geq \Delta F$. Las contribuciones no conservativas del trabajo se deben a fenómenos asociados al no equilibrio del proceso en cuestión, de forma que para un proceso que transcurre reversiblemente, es decir, que el sistema está en un

estado de quasi-equilibrio a lo largo del proceso, estas contribuciones desaparecen y la desigualdad anterior se convierte en una igualdad exacta. Para poder calcular el trabajo en condiciones de reversibilidad el proceso debe darse a una velocidad diferencial, haciendo que su duración sea infinita, lo que no es aplicable desde el punto de vista práctico. Jarzynski demostró en 1997 que se puede calcular el valor de la variación de energía libre de un proceso reversible desde aproximaciones no reversibles del mismo a través de la expresión:

$$e^{-\beta \Delta F} = \langle e^{-\beta W} \rangle$$

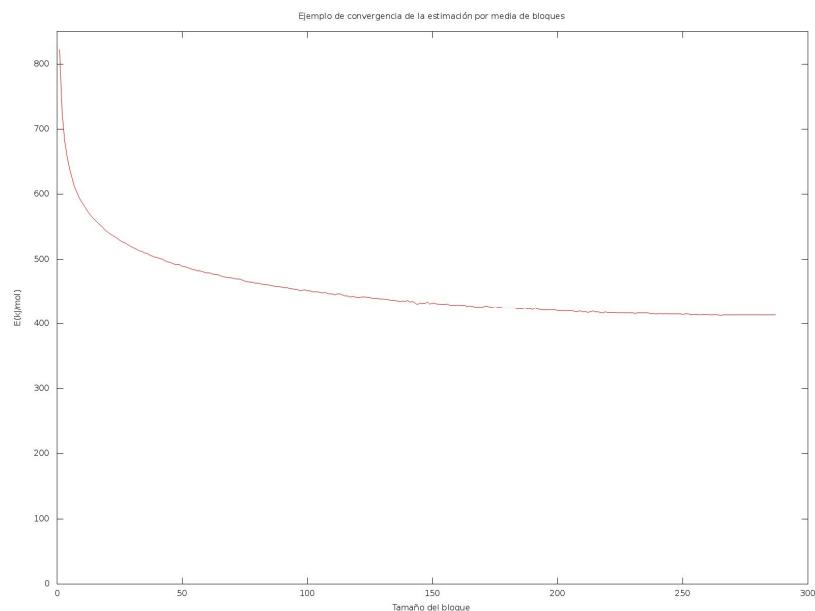
En ella β es $\frac{1}{k_B T}$, W el trabajo calculado para una trayectoria dada, y la notación ' $\langle \dots \rangle$ ' indica la media sobre el conjunto de trayectorias. A medida que el número de experimentos realizados crece, la energía libre calculada se aproxima al valor real, coincidiendo exactamente cuando el número de experimentos realizados es infinito. Matemáticamente, esta expresión hace que sea mucho más relevante el valor más bajo de energía encontrado hasta el momento, y el efecto se acentúa tanto más cuanto mayor sea la diferencia entre valores. La convergencia de la igualdad de Jarzynski es muy lenta por depender de la aparición de trayectorias con un valor muy pequeño del trabajo, que se dan con una probabilidad muy baja. Esto unido al hecho anterior implica que la velocidad de convergencia sea errática haciendo que una extrapolación en el infinito sea muy complicada mediante aplicación directa de la ecuación.



Ejemplo de convergencia de la igualdad de Jarzynski

Estimación por Medias de Bloques^[25]

La estimación por medias de bloques es un método de cálculo mediante el cual se puede estimar el trabajo aplicando la igualdad de Jarzynski a subconjuntos de datos de forma que los resultados converjan de forma monótona respecto al número de experimentos y se obtengan valores de la misma calidad con una cantidad mucho menor de repeticiones. El método se basa en aplicar la igualdad de Jarzynski a grupos de tamaño creciente de repeticiones ordenadas aleatoriamente y hallar la media de los resultados de los grupos. Al representar el trabajo frente al tamaño de los grupos se obtiene una gráfica que tiende asintóticamente al valor real del trabajo calculable.



Dada la problemática de una extrapolación en el infinito, se lleva a cabo un cambio de variable en la variable independiente, y se representa el trabajo frente a la inversa del tamaño de los grupos elevada a un parámetro de ajuste, τ : $W \text{ vs } \frac{1}{x^\tau}$

Así, los datos se pueden ajustar a una recta cuya ordenada en el origen es la extrapolación del trabajo calculable, dado que $\frac{1}{x^\tau} = 0 \rightarrow x = \infty$ para cualquier valor positivo de τ . Se sigue el consejo de los autores del método de obviar los puntos con bloques de menor tamaño. En esta representación son los puntos con los valores más altos para las abscisas y mayor error en las ordenadas, de forma que pueden afectar muy negativamente a la extrapolación sin alterar notablemente el coeficiente de regresión. Los grupos se forman eligiendo al azar sin repetición ni reemplazamiento los trabajos de las trayectorias a estudiar. En caso de que el número total de trayectorias no sea divisible entre el tamaño del bloque se forma el mayor número posible de

bloques de ese tamaño, quedando algunos trabajos calculados sin participar de la estimación para ese tamaño de grupo concreto. Cuando esto ocurre con grupos de tamaño grande (mayores que la mitad de los trabajos totales) puede suceder si hay trabajos de valor bajo y poco frecuentes que estos valores no se computen en alguno de los pasos del cálculo. Esto se refleja en oscilaciones en el trabajo calculado cuando n se hace grande, introduciendo ruido en la señal y por ello cuando los grupos son lo suficientemente grandes, estos valores se obvian para la regresión.

Como mejora al método original se ha desarrollado un método para estimar el valor óptimo de τ . Para ello se representan los valores del coeficiente de regresión al cuadrado frente al valor de τ , ajustando los puntos a una parábola. El valor de τ en el vértice de la parábola es el que proporciona una mejor regresión lineal para la gráfica de W vs $\frac{1}{x^2}$. El valor del parámetro queda acotado inferiormente por 0.3. Valores menores afectan negativamente a la estimación general, pese a que la regresión mejore en calidad.

Para ajustar a una parábola se ha utilizado el método de mínimos cuadrados por la aproximación matricial: Para el sistema sobredeterminado $A_{n,3} \cdot C_{3,1} = S_{n,1}$, en el

que A es la matriz formada por $\begin{pmatrix} x_1^2 & x_1 & 1 \\ \vdots & \vdots & \vdots \\ x_n^2 & x_n & 1 \end{pmatrix}$, C es la matriz de los coeficientes para

la parábola $\begin{pmatrix} a \\ b \\ c \end{pmatrix}$, y S la matriz con los valores de y $\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$, el ajuste por mínimos cuadrados se realiza así:

$$A_{3,n}^T \cdot A_{n,3} \cdot C_{3,1} = A_{3,n}^T \cdot S_{n,1}$$

La matriz A es idéntica para todos los experimentos realizados, de forma que el ajuste se puede calcular de una forma más rápida

$$(A_{3,n}^T \cdot A_{n,3})^{-1} A_{3,n}^T \cdot A_{n,3} \cdot C_{3,1} = (A_{3,n}^T \cdot A_{n,3})^{-1} A_{3,n}^T \cdot S_{n,1}$$

$$C_{3,1} = (A_{3,n}^T \cdot A_{n,3})^{-1} A_{3,n}^T \cdot S_{n,1}$$

Llamando $B_{3,n}$ a la matriz $(A_{3,n}^T \cdot A_{n,3})^{-1} A_{3,n}^T$, la última expresión resulta en esta otra más legible,

$$C_{3,1} = B_{3,n} \cdot S_{n,1}$$

Una vez obtenidos los coeficientes de la parábola se calcula la bondad del ajuste para comprobar la validez del mismo mediante la siguiente fórmula:

$$R^2 = 1 - \frac{\sum [(y_i - f(x_i))^2]}{\sum [(y_i - \bar{y})^2]}$$

Comprobada la bondad del ajuste, se obtiene el valor de τ en el vértice como $\frac{-b}{2a}$ (ó 0.3, el que sea mayor), y se lleva a cabo la regresión lineal con este valor óptimo del parámetro. Los límites de confianza se han calculado al 95% de la forma habitual:

$$y \pm t_{1-\alpha/2, n-2} \cdot \sqrt{\frac{1}{n} + \frac{(x - \hat{x})^2}{\sum [(x_i - \hat{x})^2]}} \cdot \sqrt{\frac{\sum [(y_i - f(x_i))^2]}{n-2}}$$

Experimentación:

Para la puesta a punto del sistema se ha utilizado un protocolo relativamente típico para la mayor parte (si no la totalidad) de las simulaciones por dinámica molecular. Tras obtener una estructura de partida, se debe completar ésta con los átomos que no figuren en ella (habitualmente los átomos de hidrógeno, cuya posición no suele poderse resolver mediante difracción de rayos X). A la estructura ya completa se le añade una 'caja de agua', una cierta cantidad de moléculas de agua que rodean a las moléculas de interés para que estén en disolución y se disponen en una geometría que permita el empaquetado compacto del sistema consigo mismo en todas las direcciones del espacio. Una condición que se debe satisfacer para que la simulación sea estable es la neutralidad de carga del sistema, de modo que se añaden iones (típicamente de sodio o de cloro) hasta contrarrestar la carga neta de la proteína y demás moléculas. Finalmente se genera un archivo que contenga la información de la posición de todos los átomos del sistema en el espacio y otro con la información de la conectividad de cada uno de los átomos y otra información relevante.

En este caso la estructura de partida es la estructura resuelta por difracción de rayos X de la flavodoxina de *H. pylori*, depositada en el Protein Data Bank bajo el identificador de 1FLV^[8]. En el archivo no figuran por tanto, los átomos de hidrógeno. Tampoco aparecen en él los dos primeros aminoácidos de la proteína pero, dada su movilidad y su lejanía tanto en secuencia como posición espacial respecto a la zona de interés, estos residuos no han sido reconstruidos para los experimentos.

La construcción de la estructura se hizo por separado para la proteína y el cofactor, por una cuestión de depuración de errores a la hora de crear y manejar los archivos de topología y parametrización asociados. Las coordenadas de los átomos que

no figuran en la estructura inicial, a excepción de los del primer residuo, fueron calculadas y los átomos creados en la nueva estructura mediante la extensión de VMD *guesscoord* de acuerdo a los archivos de topología (el básico de CHARMM en su versión 2.9 para la cadena peptídica y el descrito por Schulten et al. para la molécula de FMN, también en formato para CHARMM 2.6 o superior^[26]).

Al crear la caja de agua se eligió una geometría ortoédrica con suficientes moléculas de agua como para que la distancia mínima de cualquier átomo del complejo proteína-ligando con el límite de la caja fuera de 7 Å. El modelo de agua utilizado es TIP3P^[27-28] y las dimensiones finales de la caja fueron de 46.2 x 56.7 x 50.1 Å. La caja de agua fue generada a través de la extensión de VMD *waterbox*.

Para neutralizar la carga neta del sistema fue necesario añadir 19 átomos de sodio. La determinación de los emplazamientos óptimos de partida y el posicionamiento de los iones se llevaron a cabo a través de la extensión *autoionize*. Los archivos de topología del sistema fueron creados a través del paquete *psfgen* y la extensión *mergestruct* incluidos en la compilación de VMD. Una vez terminada la formación del constructo inicial, el sistema consta de un total de 11776 átomos, incluyendo la proteína, la molécula de FMN, las 3045 moléculas de agua y los 19 iones de sodio.

El campo de fuerzas utilizado en las simulaciones es un campo de fuerzas personalizado que incluye el campo de fuerzas básico de CHARMM en su versión 2.9 y la parametrización de la molécula de FMN descrita por Schulten et al.^[26].

Antes de poder llevar a cabo simulaciones de producción se debe minimizar la energía del sistema y calentar gradualmente hasta llegar a equilibrar el sistema a la temperatura de trabajo.

Primero se llevó a cabo la minimización de 5000 pasos de cálculo con los átomos de la proteína y el FMN fijos, de forma que la caja de agua quedó compactada y alcanzó un estado estable en un mínimo local de energía. Acto seguido se realizó una breve dinámica, al igual que en el paso anterior sólo para el agua, de 1ns con un tiempo de integración de 1 fs, en la que las moléculas del solvente quedaron relajadas. Posteriormente la fijación rígida de los átomos de la proteína y el FMN se sustituyó por potenciales armónicos con valores pares decrecientes para la constante asociada, desde 20 hasta 0, llegando a una dinámica libre. Para cada valor de la constante se hizo

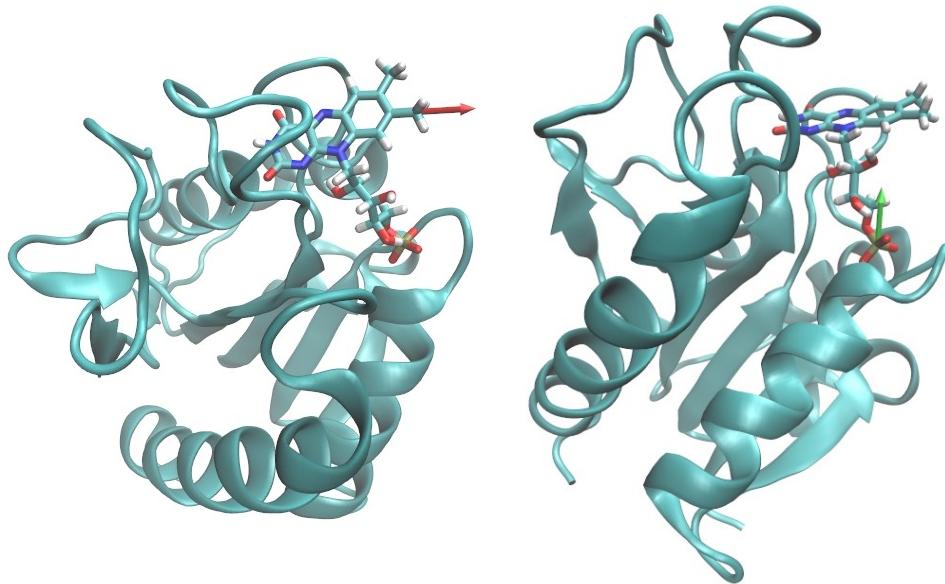
una simulación de 1 ps. Tras esto se llevó a cabo el calentamiento gradual del sistema, empezando desde 10 K y subiendo hasta los 310 K de trabajo en 30 pasos de 100 ps cada uno. Para finalizar el proceso de preparación se hizo una simulación sin fuerzas externas de 2 ns en las condiciones de trabajo.

Salvo que se indique lo contrario, en todas las simulaciones de producción se han utilizado dinámicas de Langevin sobre la proteína y el FMN con un coeficiente de 0.5, las distancias de corte para las interacciones ha sido de 14 Å y la de atenuación de 12 Å, el tamaño de gradilla para los cálculos del PME de 48x54x48 puntos sobre los vectores base de la celda, la temperatura de 310 K mantenida por la dinámica de Langevin y la presión de 1 atmósfera mantenida por el método de Nosé-Hoover.

A partir de la última estructura en la trayectoria previa se inició la primera simulación de producción, una dinámica de 100 ns con potenciales armónicos muy suaves sobre los carbonos carbonílicos de la tirosina 49 y de la lisina 115 para evitar la deriva de la proteína hacia los bordes de la caja de agua.

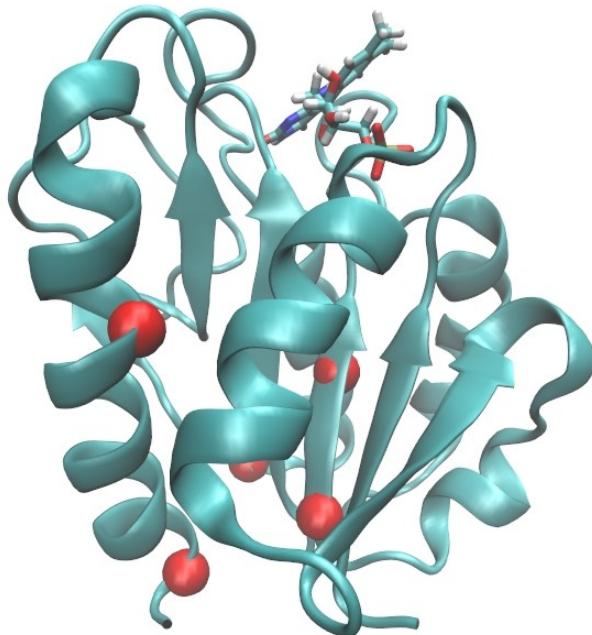
Sobre la trayectoria obtenida se llevó a cabo un análisis de componentes principales para evaluar el comportamiento de la proteína. En el cálculo se incluyeron sólo los átomos de la cadena principal de la proteína. De acuerdo a los resultados se eligieron aleatoriamente ocho estructuras equivalentes como punto de partida para llevar a cabo el resto de experimentos.

Para estudiar las posibles vías de salida del ligando se han utilizado experimentos de dinámica molecular forzada. Las experiencias se pueden dividir en dos grupos: Las que aplicaban la fuerza sobre el anillo de flavina del FMN y las que aplicaban la fuerza sobre el grupo fosfato, también del FMN. Dentro de cada grupo se eligió un conjunto de vectores (7 para la flavina y 11 para el fosfato) en cuya dirección que aplicar la fuerza. La elección de las direcciones fue arbitraria y sistemática, barriendo un espacio alrededor de la dirección aproximadamente más directa de salida del átomo. Las direcciones se calculan vectorialmente a partir de las posiciones relativas de varios átomos en la estructura, de forma que se puede automatizar la extracción del vector director minimizando las diferencias debidas a fluctuaciones conformacionales leves.



Ejemplos de vectores de salida. A la izquierda el vector Metil.2, a la derecha el vector Phos2.

Los valores de la constante elástica y la velocidad se ajustaron por prueba y error iterativamente para cada uno de los vectores. Los átomos concretos sobre los que se aplicó la fuerza fueron el carbono del grupo metilo del carbono 8 de la flavina y el átomo de fósforo. A las simulaciones sobre el metilo se les identifica en esta memoria por la partícula Metil, y a las realizadas sobre el fosfato por Phos. También se impusieron restricciones de movimiento a varios átomos de la cadena principal de la proteína para evitar la rotación y translación de la misma en respuesta a las fuerzas externas sobre el ligando. Los átomos fijados fueron el carbono carbonílico de la tirosina 49, el carbono alfa de la isoleucina 52, el carbono carbonílico de la lisina 115, el carbono alfa de la isoleucina 156 y el carbono carbonílico de la fenilalanina 167.



Posiciones de los átomos anclados durante las simulaciones SMD, marcadas con esferas rojas

Estas posiciones se eligieron atendiendo a motivos estructurales y topológicos, por una parte están situados lo suficientemente lejos de la zona de estudio como para no interferir en los experimentos, y por otra parte se encuentran en zonas relativamente robustas y rígidas estructuralmente de modo que al mantenerlos fijos en el espacio se evitan deformaciones de torsión de la estructura a gran escala. Se llevaron a cabo 56 repeticiones por cada vector para un total de 1008 simulaciones. En ellas el tiempo de integración se aumentó a 1.2 fs para reducir el tiempo de cálculo necesario. Las duraciones de estas simulaciones se puede clasificar en cuatro grupos (de 1.5 ns, de 1.86 ns de 2.34 ns y de 3 ns), debido a ampliaciones de algunas dinámicas en función de su comportamiento inicial.

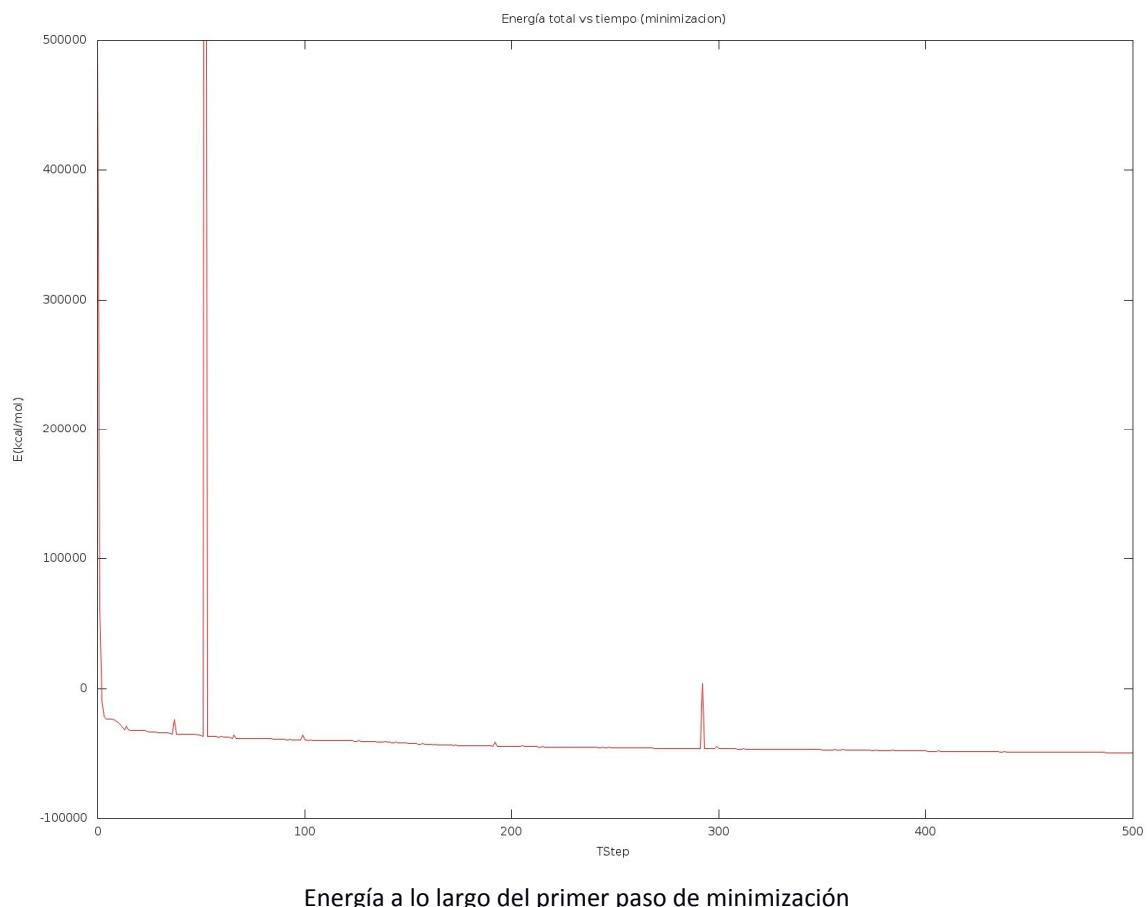
Tras seleccionar las trayectorias válidas y descartar las demás se procedió al cálculo de la energía libre de salida del ligando para cada uno de los vectores por separado y en conjunto, y para el proceso de salida del fosfato y del anillo de flavina teniendo en cuenta el orden de los sucesos y la vía de salida. El cribado de trayectorias se basa fundamentalmente en que el sistema haya sido capaz de llegar a tener una separación de 4 Å entre la proteína y el FMN. En algunos casos el FMN al salir por un lado de la caja y entrar por el opuesto se acercaba demasiado a la proteína o directamente entraba en contacto con ella antes de haberse separado lo suficiente

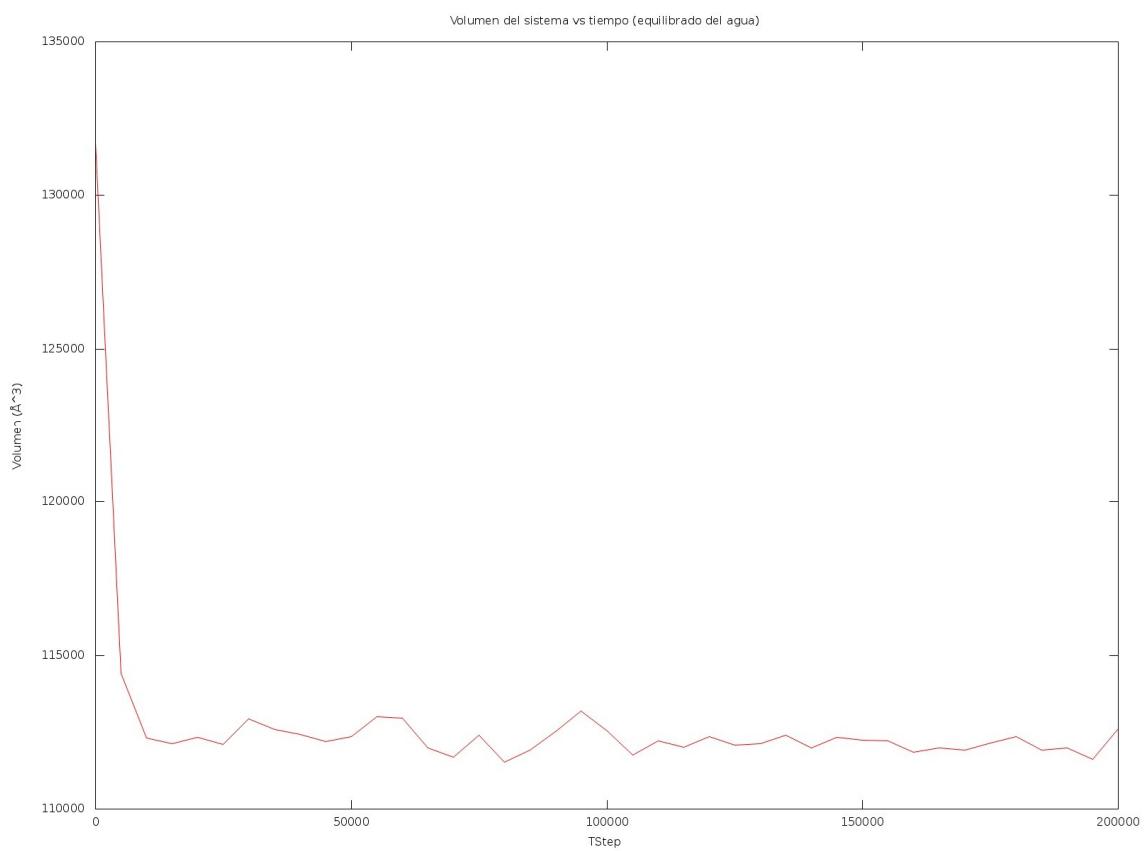
como para poder considerar que no hubiese interacciones intensas entre el FMN y los residuos de la zona circundante al sitio de unión. El trabajo fue calculado numéricamente como la integral de la fuerza a lo largo del desplazamiento del átomo en cuestión por el método de trapecios desde tiempo 0 hasta el momento en el que la molécula de FMN se encontrase a más de 4 Å de distancia de la proteína, medida entre los dos átomos más cercanos de ambas moléculas.

3.Resultados:

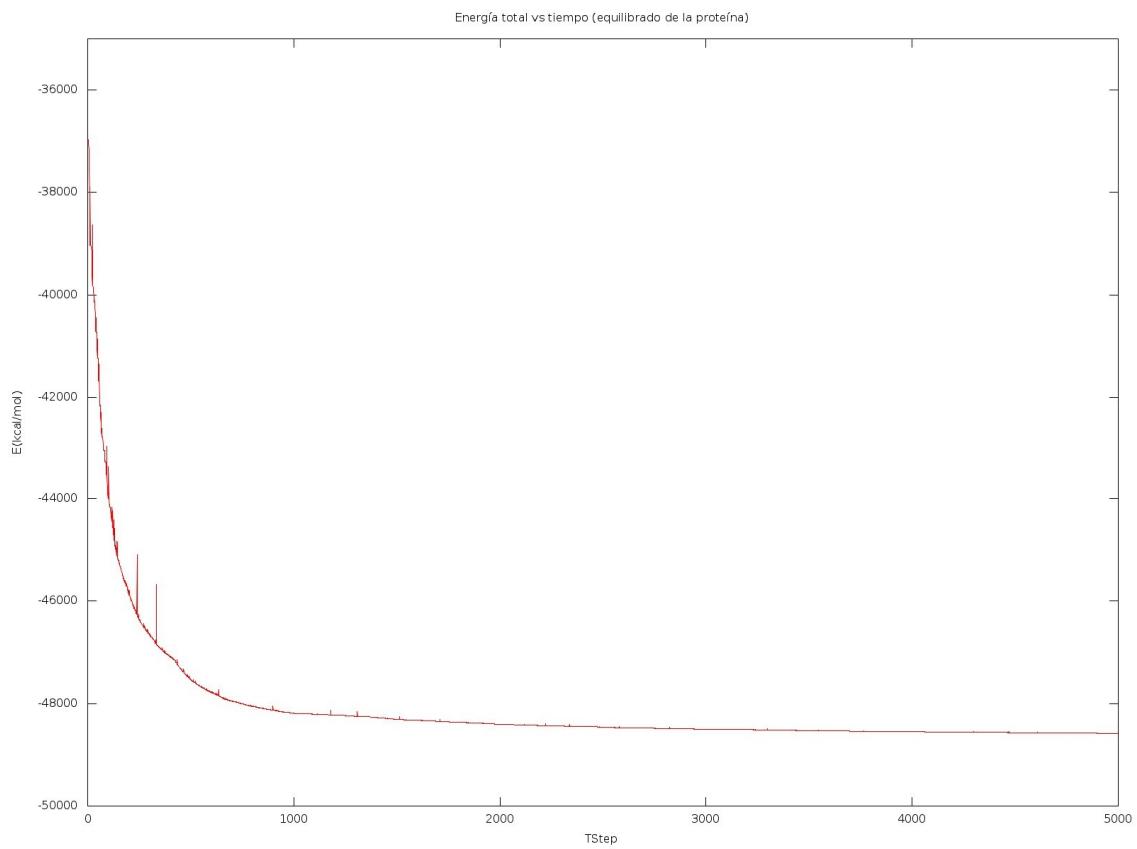
Fase de preparación

Tras la minimización se resolvieron 48 colisiones entre átomos y la energía global del sistema se estabilizó en torno a un valor mucho menor que el inicial. Durante el primer equilibrado de la caja de agua el sistema se compactó y aproximó su presión a la de trabajo. La fase de relajación de la proteína y el FMN también redujo la energía total del sistema como cabía esperar. El incremento de la temperatura se llevó a cabo sin contratiempos y la dinámica corta para terminar de equilibrar el sistema completo alcanzó un punto deseable para el inicio de las simulaciones.

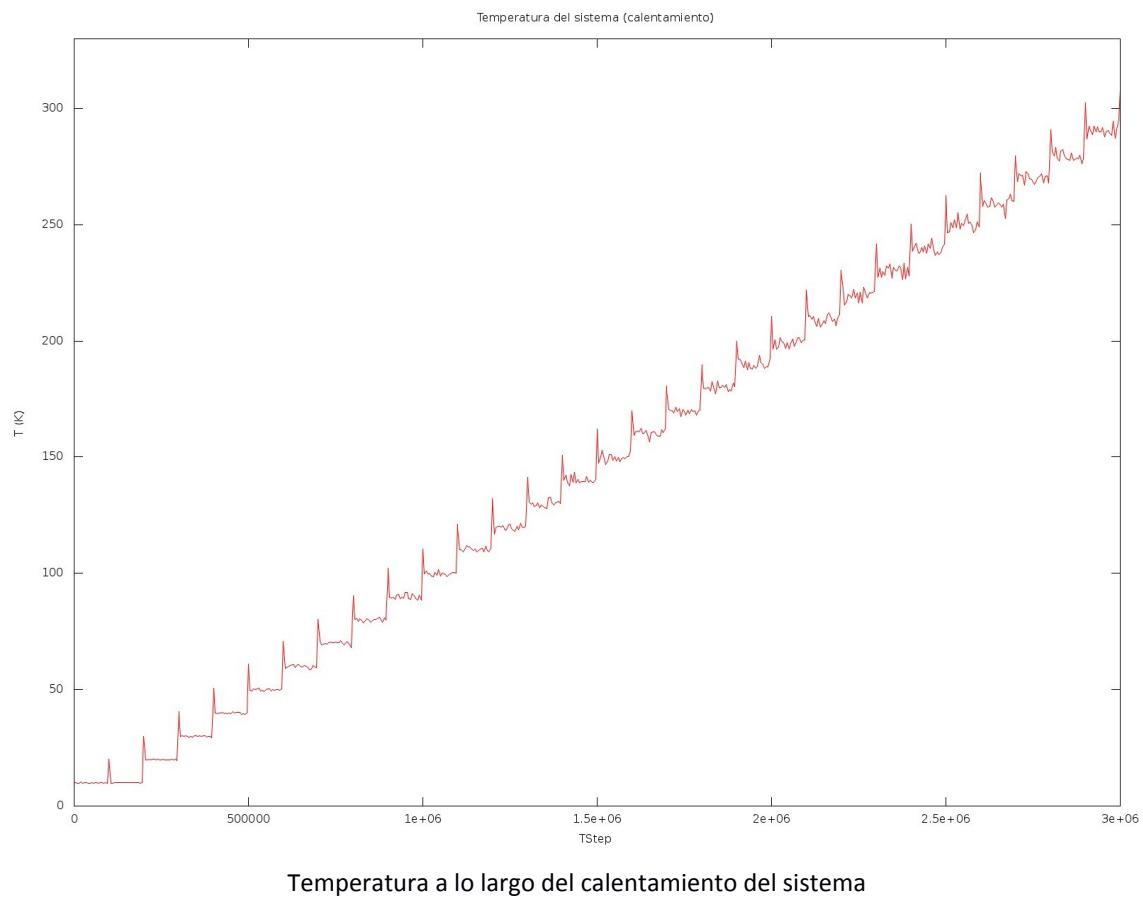




Volumen a lo largo del equilibrado del agua

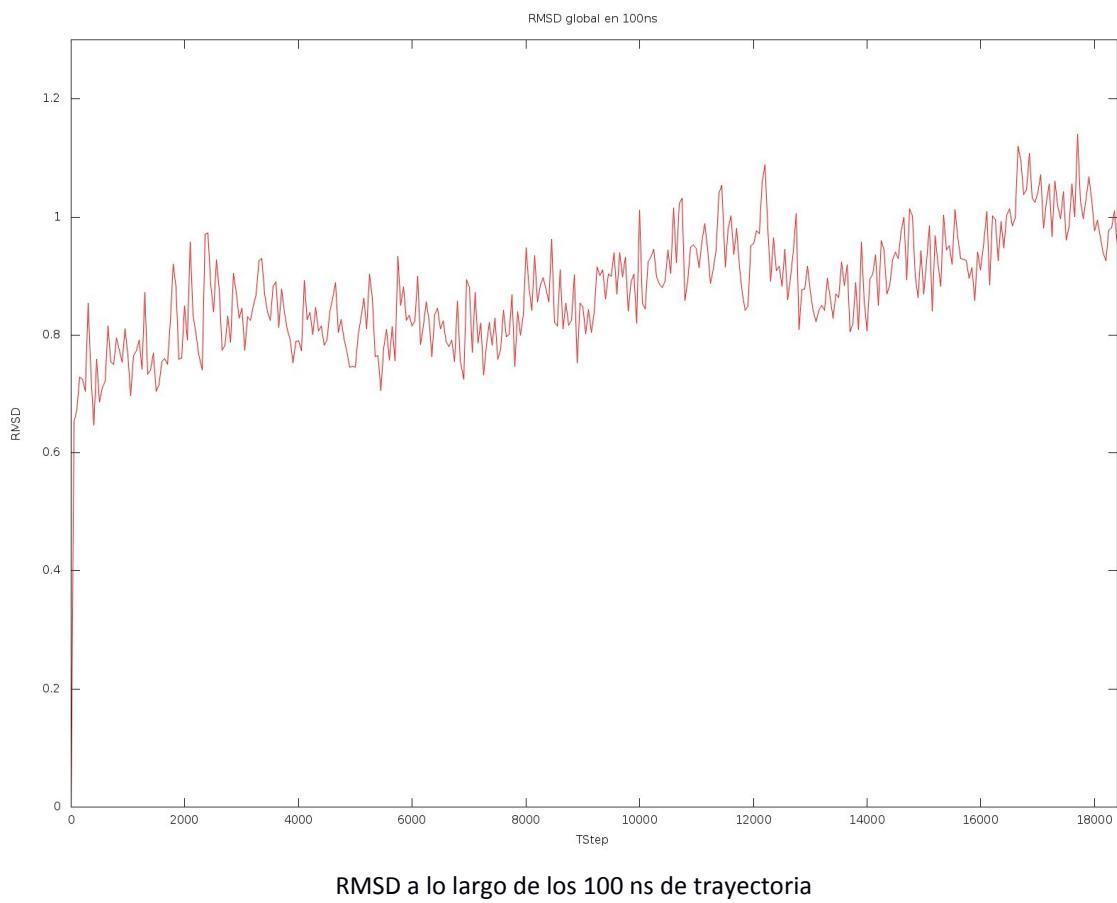


Energía a lo largo de la relajación de la proteína



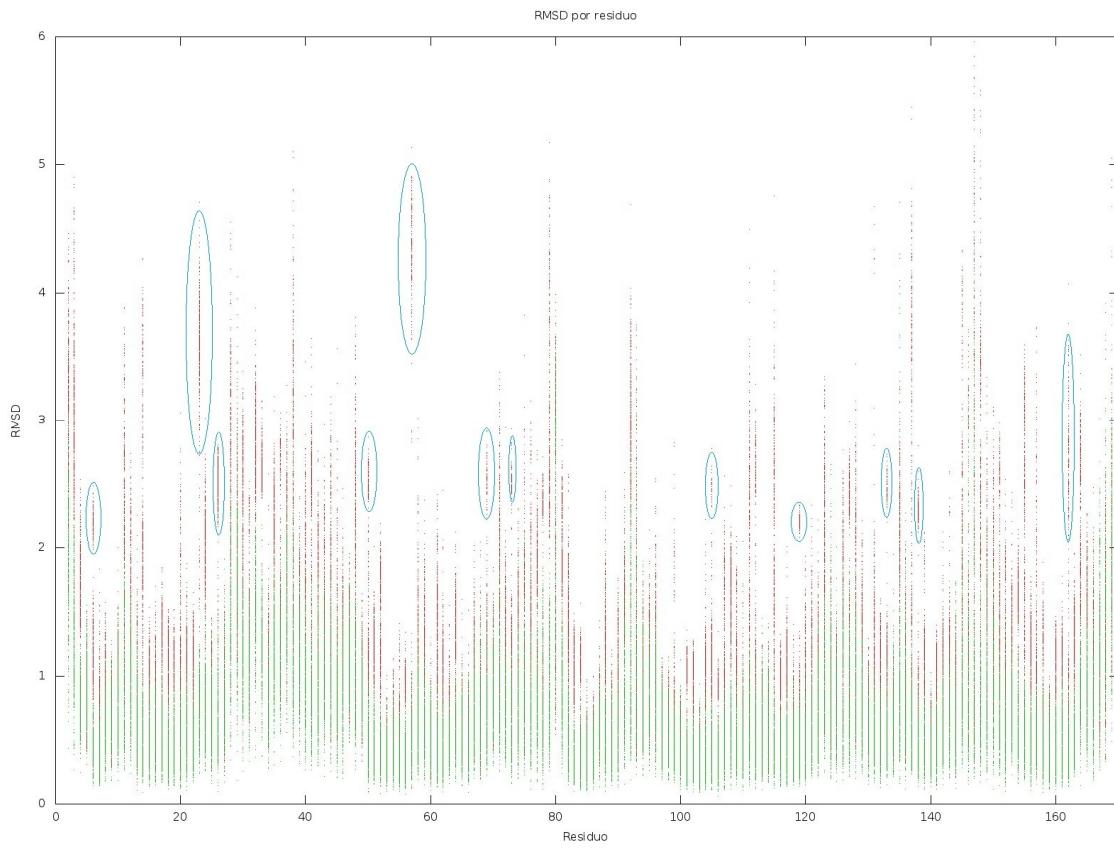
Fase de producción

El análisis de la variabilidad estructural a través del RMSD muestra que la trayectoria es estable, sin cambios conformacionales drásticos, aunque se pueden observar regiones diferenciables a lo largo de la secuencia temporal.

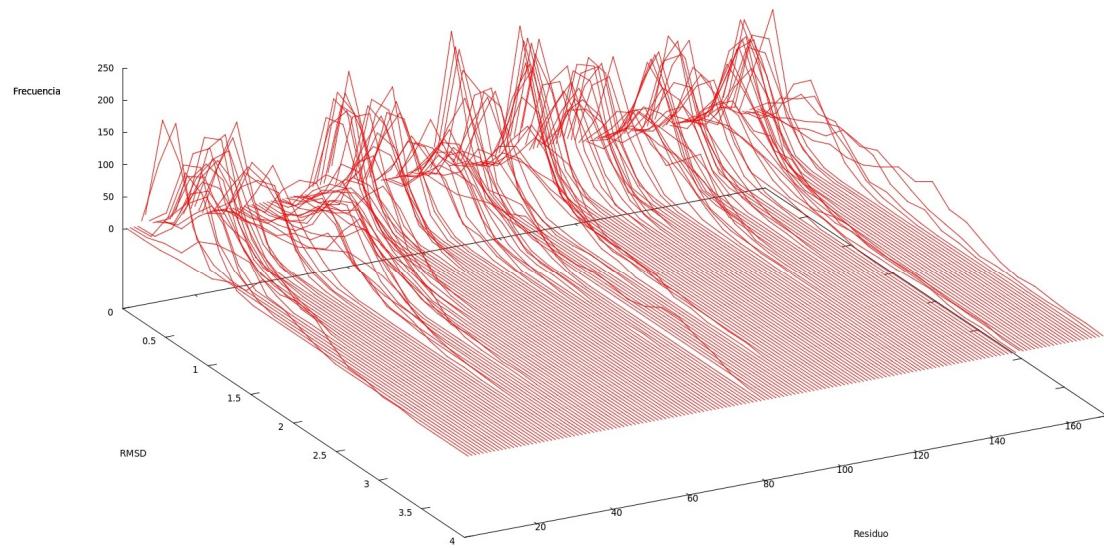


La temperatura se mantiene fluctuando alrededor de los 310K y la presión y el volumen se mantienen también aproximadamente constantes.

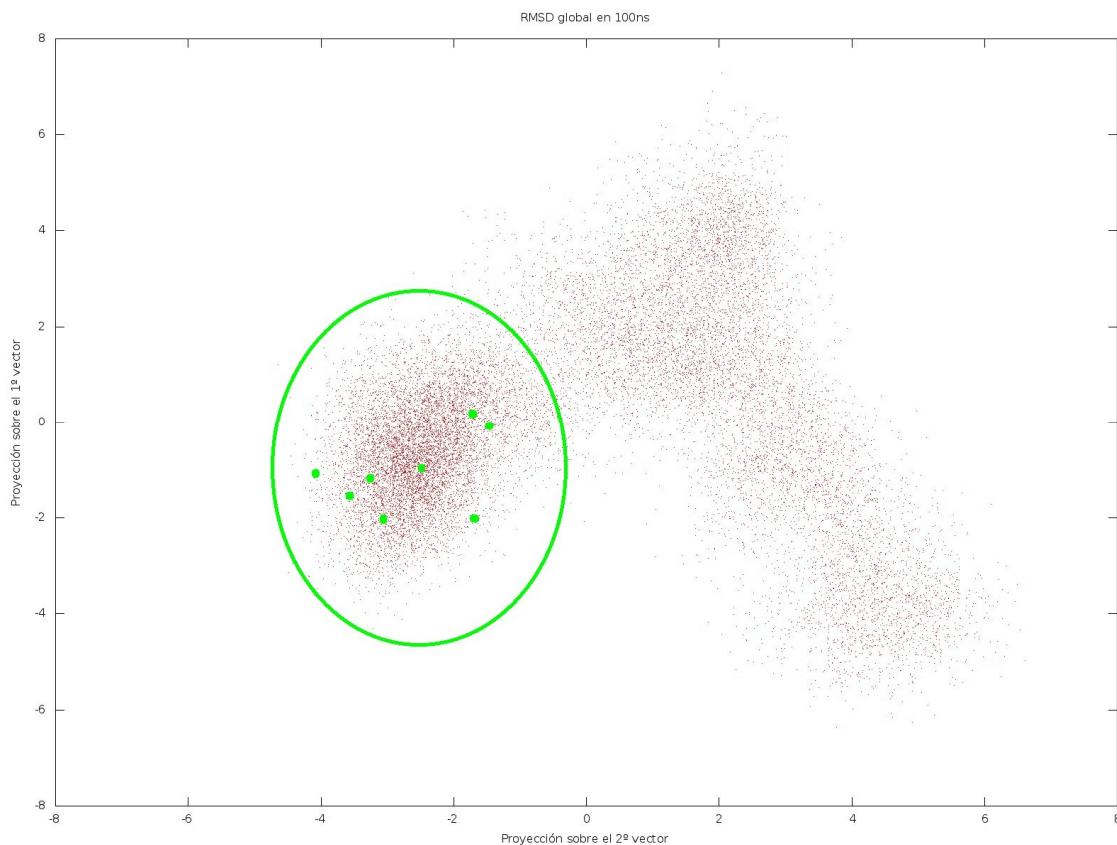
El análisis del RMSD por residuo hace intuir que pueden haberse visitado varios estados distintos a lo largo de la simulación. Al representar las proyecciones de las coordenadas de los átomos de cada instantánea de la trayectoria sobre los primeros vectores propios del análisis por componentes principales se confirma esta hipótesis observándose varios cúmulos que representan a estados distintos de la proteína. Las estructuras de partida para los experimentos siguientes se eligieron aleatoriamente del cúmulo más compacto, de forma que sus resultados fuesen comparables.



RMSD desglosado para cada residuo. En verde calculado sobre el esqueleto carbonado y en rojo para el aminoácido completo. Rodeados en azul varios de los grupos que presentan estados discernibles.



Curvas de frecuencias para intervalos de RMSD de amplitud 0.2 para cada residuo a lo largo de toda la simulación.



Representación de las proyecciones las estructuras de la trayectoria sobre los dos primeros vectores del análisis de componentes principales. Rodeado en verde el cúmulo del que se obtuvieron las estructuras de partida, cuyas proyecciones están indicadas con los puntos verdes.

Respecto a las simulaciones por SMD, varias fueron descartadas para el análisis, como ya se ha explicado antes. De las 1008 simulaciones iniciales se han podido analizar completamente 617. A partir de ellas se han realizado cálculos de la variación de energía libre del proceso de disociación utilizando el método de medias de bloques, tanto para todos los vectores en conjunto como para cada uno independientemente. Los cálculos estiman esta cantidad para el proceso reversible en torno a los 330 kJ/mol, un valor muy alejado de los 47.7 kJ/mol determinados experimentalmente. La tabla al final de esta sección recoge los valores calculados para cada vector y los respectivos intervalos de confianza al 95%. Las trayectorias del vector Phos.4 se descartaron debido a que se producían perturbaciones graves de la estructura de la proteína, desplegándose una región de 14 aminoácidos. Las simulaciones debían alargarse demasiado para llegar al límite de los 4 Å indicado anteriormente y además se alcanzaban unos valores del trabajo individual de cada simulación muy superiores a los del resto de vectores. Para los vectores Phos.6 y Phos.10 no había suficientes

trayectorias que pudiesen cumplir el requisito de la distancia debido a que el FMN, al salir por un lado de la caja y entrar por el otro se acercaba demasiado a la proteína y de prolongar la simulación habría acabado incrustándose en la estructura, de forma que también han quedado excluidos del cálculo y tampoco se presentan en la tabla.

Tabla resumen de los cálculos de variación de energía libre

Vector de Salida	Trabajo estimado (kJ/mol)	Intervalo de confianza (kJ/mol)	Número de valores	R ²
Metil.0	537.3	±4.3	25	0.98
Metil.1	528.3	±3.8	17	0.996
Metil.2	389.6 484.1	±1.3 ±0.1	20 20	0.991 0.996
Metil.3	540.9	±0.2	30	0.996
Metil.4	509.9	±0.2	30	0.996
Metil.5	552.7	±1.3	40	0.997
Metil.6	620.9	±0.4	40	0.998
Metil.Todos	379.2 482.5	±0.4 ±0.2	182 188	0.996 0.997
Phos.0	751.0	±0.2	40	0.998
Phos.1	618.8	<±0.1	40	0.997
Phos.2	402.0	±9.5	30	0.98
Phos.3	410.9	±1.7	40	0.97
Phos.5	663.5	±0.4	15	0.997
Phos.7	408.1	±7.9	20	0.98
Phos.8	495.5	±0.5	30	0.996
Phos.9	567.8	±2.5	20	0.997
Phos.Todos	501.9	±0.1	300	0.998
Global	333.3 483.3	±1.0 <±0.1	300 565	0.995 0.997

En el cálculo de la energía a escala global se excluyeron los valores iniciales como los autores de la técnica proponen, y también datos del final de la tabla debido a que introducían una cantidad notable de ruido y perturbaban el resultado haciendo la estimación poco consistente con el resto de estimaciones.

El intervalo de confianza incorpora sólo el error debido a la extrapolación lineal.

La estimación de medias de bloques produce resultados en los que el valor extrapolado suele ser siempre menor que cualquiera de los valores obtenidos en el experimento individualmente, de modo que no es sorprendente que los cómputos para conjuntos de simulaciones den valores inferiores a los de los vectores individuales.

Los resultados dobles se deben a que en la estimación Metil.2, cuyo valor se desmarca del resto claramente, el valor que más pesa y que hace que el resultado sea tan bajo comparativamente pertenece a un único experimento. No es sencillo distinguir si este experimento presenta algún tipo de problema en el desarrollo de la dinámica o si es uno de esos experimentos tan valiosos, infrecuentes e importantes para la convergencia de la igualdad de Jarzynski, de modo que se han realizado los cálculos tanto teniendo en cuenta ese valor y como descartándolo, en todos los conjuntos en los que haya participado del cálculo.

4.Discusión y conclusiones:

Las variaciones de energía libre calculadas no se aproximan bien a los valores medidos experimentalmente por un orden de magnitud aproximadamente, y la gráfica del potencial de fuerza media a lo largo de cada simulación, que debería corresponderse con la curva de energía libre del proceso, es siempre creciente en lugar de decaer tras superar el punto de energía máxima del proceso. El estudio de la técnica me lleva a creer que el origen de las desviaciones se basa en que el trabajo, al calcularse a través de la integral de la fuerza ejercida sobre un átomo concreto a lo largo del desplazamiento de este átomo, computa de forma directa el trabajo de una parte pequeña del sistema y de forma indirecta sólo las contribuciones positivas a la integral obviándose procesos estabilizantes como las nuevas interacciones que se forman en el sitio de unión del ligando entre la proteína y el agua o consigo misma, ya sean puentes de hidrógeno, interacciones hidrofóbicas o de cualquier otro tipo, o (probablemente la contribución más importante al error) los procesos de relajación de la estructura de la proteína al replegarse zonas que habían sido deformadas o arrastradas por el FMN en su movimiento hacia el exterior.

Reflexionando acerca de los resultados y especialmente el significado del cálculo en sí, se puede llegar a la conclusión de que el trabajo aquí estimado corresponde más bien a la suma de las energías de activación de los procesos que van teniendo lugar a lo largo de la trayectoria y que afectan también a los demás átomos y no sólo al que se le aplica la fuerza. Esto permite comparar de forma cualitativa las velocidades a las que se dan las disociaciones por dos vías distintas, siendo lo más probable en cada caso que una opción con menor diferencia de energía libre calculada represente un modo más rápido de reaccionar. En este sentido la dirección más rápida de salida sería la del vector Metil.2, que con 340 kJ/mol está 60 kJ/mol por debajo de experimentos con valores más cercanos, pero observando detenidamente el cálculo se puede ver que este resultado se ve tremadamente influido por una sola de las trayectorias que tuvo un resultado del trabajo comparativamente bajo, como ya se ha comentado anteriormente. La naturaleza caótica de las simulaciones hace que sea difícil distinguir si se trata simplemente de un artefacto o si se trata de una de esas

simulaciones de valor del trabajo pequeño de tanto interés para la buena convergencia de la igualdad de Jarzynski. Si no se incluye en el cálculo la réplica en cuestión, el valor estimado del trabajo aumenta hasta los 480 kJ/mol. Se puede apreciar como los valores estimados para Iso vectores Metil.i se agrupan entre los 510 y los 540 kJ/mol aproximadamente, siendo la estimación para todas las réplicas Metil muy cercana a los valores calculados para Metil.2. En los experimentos sobre el fosfato hay varios en torno a los 400 kJ/mol. Los que tienen una mayor diferencia de energía libre en la tanda Phos son aquellos en los que se produce una mayor perturbación de la estructura de la proteína. El cálculo para los vectores Phos en conjunto arroja un resultado muy superior al de los Metil incluyendo el dato extremo y ligeramente superior sin incluirlo. Es remarcable que la estimación para Phos.total es superior a las de Phos.2,3 y 7. Examinando las regresiones para estos vectores en el método de medias de bloques se aprecia como los valores asociados a un mayor tamaño de bloque no parecen converger adecuadamente y se puede estar subestimando el valor del trabajo extrapolado. Para contrastar esto se requeriría un mayor número de réplicas para estos vectores.

Asumiendo como cierta la reflexión acerca de la relación entre trabajo estimado y energías de activación, se llega a la conclusión de que, según la tabla de resultados los vectores preferentes de salida son Metil.2 y Phos.2,3 y 7 con una diferencia relativamente pequeña tomando el valor bajo de Metil.2. También a nivel global, los vectores Metil resultan más rápidos para la disociación que los Phos. Aplicando el principio de reversibilidad microscópica se puede expresar en términos de unión en lugar de disociación como que en los mecanismos más rápidos de unión el primer contacto lo establece el fosfato, y una vez esa parte del FMN está en el sitio de unión, la flavina accede al suyo y el bolsillo de la proteína se cierra sobre el ligando. Estos resultados entran en conflicto con la hipótesis planteada a la luz de los experimentos *in vitro*, pero dado que las estimaciones tienen un error relativamente grande el resultado no es lo suficientemente firme como para hacer decantarse y dar por válido cualquiera de los dos casos. Se debe seguir trabajando para conseguir disminuir este error y alcanzar una mejor concordancia con los resultados experimentales. En estos momentos se está trabajando para intentar comprobar la

relación entre el trabajo estimado y el valor de los coeficientes cinéticos del proceso de unión ya calculados experimentalmente^[29].

El análisis de la primera dinámica de producción ha puesto de manifiesto que hay más de un estado básico poblando el espacio conformacional observado. Para poder asegurar que los resultados obtenidos son representativos del comportamiento de la proteína nativa se debería ampliar el tamaño de la muestra de conformaciones, empezando por extender esta primera simulación de partida hasta un límite de tiempo razonablemente largo, en el orden de los microsegundos, o incluso decenas de microsegundos, ahora que estas escalas de tiempo están al alcance de la investigación en un tiempo razonable gracias a los avances tecnológicos y la paralelización masiva en el cálculo. Por otra parte, el número de vectores utilizado cubre una porción pequeña de las direcciones posibles, y de una forma realista no es probable que una molécula se mueva sólo en una dirección a través de todo un mecanismo en varios pasos para la unión-disociación.

Con más tiempo de cálculo y tiempo físico habría sido interesante aplicar técnicas RAMD (dinámicas de aceleración aleatoria, por sus siglas en inglés) sobre estructuras en los diversos estados de la proteína. Esta técnica resulta muy útil en la exploración de mecanismos en los que una molécula debe llegar a sitios de unión recónditos (o salir de ellos), o mecanismos en los que no se conocen los pasos en profundidad, pero resulta difícil de implementar, difícil también de analizar, arriesgada en el sentido de que se suele deber invertir mucho tiempo para obtener resultados útiles, y costosa en tiempo de cálculo. No resulta, por tanto, especialmente adecuada como primera aproximación a este tipo de análisis pero sí como un siguiente paso en el estudio del mecanismo del proceso.

Para terminar se debe remarcar el aporte que supone la biblioteca de scripts desarrollados para este proyecto, que permitirían realizar este proyecto u otros similares (o no tan similares) en menos de la cuarta parte de tiempo que se ha invertido (en el caso de este proyecto en concreto el tiempo de desarrollo se vería reducido en un 85% aproximadamente, disponiendo de los mismos medios). Esta colección, incluida en los anexos, resulta útil para tareas de tratamiento de estructuras en formato PDB, preparación de cualquier sistema de forma genérica para la simulación, paralelización masiva de simulaciones a través de distintas plataformas de

cálculo o análisis de gran cantidad de datos de forma automatizada entre otras y ha sido una herramienta tan poderosa como necesaria para el tratamiento de los más de 400 Gb de información generada a lo largo de esta investigación.

5. Bibliografía y Referencias:

- [1] KIDD, Mark; MODLIN, Irvin M. A century of *Helicobacter pylori*. *Digestion*, 1998, vol. 59, no 1, p. 1-15.
- [2] POUNDER, R. E.; NG, D. The prevalence of *Helicobacter pylori* infection in different countries. *Alimentary pharmacology and therapeutics*, 1995, vol. 9, no 2, p. 33-40.
- [3] KUSTERS, Johannes G.; VAN VLIET, Arnoud HM; KUIPERS, Ernst J. Pathogenesis of *Helicobacter pylori* infection. *Clinical microbiology reviews*, 2006, vol. 19, no 3, p. 449-490.
- [4] BLASER, Martin J. Who are we? Indigenous microbes and the ecology of human diseases. *EMBO reports*, 2006, vol. 7, no 10, p. 956.
- [5] MEGRAUD, F. *H pylori* antibiotic resistance: prevalence, importance, and advances in testing. *Gut*, 2004, vol. 53, no 9, p. 1374-1384.
- [6] WU, J.Y., et al. Tetracycline-resistant clinical *Helicobacter pylori* isolates with and without mutations in 16S rRNA-encoding genes. *Antimicrobial Agents Chemother.*, 2005, no 49, p.578-583.
- [7] GRAHAM, D.Y.; SHIOTANI, A. New concepts of resistance in the treatment of *Helicobacter pylori* infections. *Nat. Clin. Pract. Gastroenterol*, 2008, no 5, p.321-331.
- [8] FREIGANG, J. Et al, Crystal structure of oxidized flavodoxin, an essential protein in *Helicobacter pylori*. *Protein Sci.*, 2002, no 11, p.253-261.
- [9] CREMADES, N. et al. Towards a new therapeutic target: *Helicobacter pylori* infection. *Biophys. Chem.*, 2005, no 115, p.267-276.
- [10] SANCHO, J. Flavodoxins: sequence, folding, binding, function and beyond. *Cell. Mol. Life Sci.*, 2006, no 63, p.855-864.
- [11] HUGHES, Nicky J., et al. *Helicobacter pylori* porCDAB and oorDABCGenes Encode Distinct Pyruvate: Flavodoxin and 2-Oxoglutarate: Acceptor Oxidoreductases Which Mediate Electron Transport to NADP. *Journal of bacteriology*, 1998, vol. 180, no 5, p. 1119-1128.
- [12] FREIGANG, Jörg, et al. Crystal structure of oxidized flavodoxin, an essential protein in *Helicobacter pylori*. *Protein science*, 2002, vol. 11, no 2, p. 253-261.

- [13] CREMADES, Nunilo, et al. Towards a new therapeutic target: *Helicobacter pylori* flavodoxin. *Biophysical chemistry*, 2005, vol. 115, no 2, p. 267-276.
- [14] GALANO, J.J. Procedimiento in silico para la obtención de modelos QSAR en derivados de [(E)-2-R-vinil]-benceno. Aplicación al diseño de nuevas moléculas contra enfermedades producidas por *Helicobacter pylori*. 2012, Proyecto Fin de Master, disponible en el Repositorio Digital de la Facultad de Ciencias de la Universidad de Zaragoza.
- [15] PHILLIPS, James C., et al. Scalable molecular dynamics with NAMD. *Journal of computational chemistry*, 2005, vol. 26, no 16, p. 1781-1802.
- [16] HUMPHREY, William; DALKE, Andrew; SCHULTEN, Klaus. VMD: visual molecular dynamics. *Journal of molecular graphics*, 1996, vol. 14, no 1, p. 33-38.
- [17] ALDER, B. J.; T. E. WAINWRIGHT, "Studies in Molecular Dynamics. I. General Method". *J. Chem. Phys.* 1959, no 31 (2) p 459-567
- [18] KARPLUS, Martin; MCCAMMON, J. Andrew. Molecular dynamics simulations of biomolecules. *Nature Structural & Molecular Biology*, 2002, vol. 9, no 9, p. 646-652.
- [19] FREDDOLINO, Peter L., et al. Molecular dynamics simulations of the complete satellite tobacco mosaic virus. *Structure*, 2006, vol. 14, no 3, p. 437-449.
- [20] ZHAO, Gongpu, et al. Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. *Nature*, 2013, vol. 497, no 7451, p. 643-646.
- [21] ZWANZIG, Robert. Langevin theory of polymer dynamics in dilute solution. *Stochastic Processes in Chemical Physics*, 1969, p. 325-331.
- [22] SCHNEIDER, T.; STOLL, E. P.; MORF, R. Brownian motion of interacting and noninteracting particles subject to a periodic potential and driven by an external field. *Physical Review B*, 1978, vol. 18, no 3, p. 1417.
- [23] FELLER, Scott E., et al. Constant pressure molecular dynamics simulation: the Langevin piston method. *The Journal of Chemical Physics*, 1995, vol. 103, p. 4613.
- [24] JARZYNSKI, Christopher. A nonequilibrium equality for free energy differences. *arXiv preprint cond-mat/9610209*, 1996.
- [25] YTREBERG, F. Marty; ZUCKERMAN, Daniel M. Efficient use of nonequilibrium measurement to estimate free energy differences for molecular systems. *Journal of computational chemistry*, 2004, vol. 25, no 14, p. 1749-1759.

- [26] FREDDOLINO, Peter L.; GARDNER, Kevin H.; SCHULTEN, Klaus. Signaling mechanisms of LOV domains: new insights from molecular dynamics studies. *Photochem. Photobiol. Sci.*, 2013.
- [27] SUN, Yaxiong; KOLLMAN, Peter A. Hydrophobic solvation of methane and nonbond parameters of the TIP3P water model. *Journal of computational chemistry*, 1995, vol. 16, no 9, p. 1164-1169.
- [28] MARK, Pekka; NILSSON, Lennart. Structure and dynamics of the TIP3P, SPC, and SPC/E water models at 298 K. *The Journal of Physical Chemistry A*, 2001, vol. 105, no 43, p. 9954-9960.
- [29] AYUSO-TEJEDOR, S., et al. Mechanism of FMN Binding to the Apoflavodoxin from *Helicobacter pylori*. *Biochemistry*, 2011, vol. 50, no 40, p. 8703-8711.